Causal Statistical Decision Theory|What are interventions?

David Johnston

December 17, 2020

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. 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. Potential investors in cigarette manufacturers want to know if the product they are backing is likely to see limited adoption due to health concerns. People holding investments in cigarette manufacturering firms want the world to be such that cigarettes do not pose a substantial health risk, as this increases the value of their investment. 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 million in 2020, while (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 some causal questions is of tremendous importance to huge numbers of people. Second, even when large amounts of data show unambiguous associations between phenomena of interest, it is still difficult to know when a causal conclusion is justified.

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

> [...] 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, 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.

Theories of causal inference exist to enable formal rather than ad-hoc reasoning about causal questions. Instead of posing informal causal question and answering them based on ad-hoc reasoning, within a theory of causal inference we ask about properties of "causal models" (which are simply mathematical types defined by the theory) subject to certain assumptions we are willing to make. A successful theory of causal inference should enable causal models that "adequately represent" the original informal question, and the assumptions we invoke should be more accessible to scrutiny than ad-hoc assertions made in the course of answering the informal question.

As well as defining causal models, which represent *claims about causation*, theories of causal inference also formalise the problem of *inferring the correct causal model* - this is the problem of taking some observational data and concluding which causal models are "possible" or "appropriate to use for the given purpose".

Defining causal models is difficult. In general, applied theories of causal inference posit that:

1. "X_i causes X_j" means that there exist different *ideal interventions* that result in different values of of X_i, hold other "causally sufficient" variables constant, do not directly affect X_j but nonetheless entail different values of X_j

2. "X_i causes X_j" means that the *counterfactual value* of X_j would be different "if X_i had taken a different value"

In practice, most theories of causal inference seem to be based on the notion of *ideal interventions*. Even "counterfactual" theories of causal inference (such as the theory based on "potential outcomes" notation) tend to define counterfactual values as "values that a variable would have taken were it exposed to an ideal intervention", if they are defined at all (Morgan and Winship, 2014; Rubin, 2005; Richardson and Robins, 2013). Alternative definitions of counterfactual values do exist, however, such as Lewis' closest world semantics (Lewis, 1986).

"Ideal interventions" themselves are difficult to define. The structural model approach of Pearl (2009) defines ideal interventions in terms of "causally sufficient models". However, most attempts to formalise this definition end up being circular. For example:

- 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 triple $((\Omega, \mathcal{F}, \mathbb{Q}), \mathcal{G}, \boldsymbol{U})$

- A triple $((\Omega, \mathcal{F}, \mathbb{Q}), \mathcal{G}, \boldsymbol{U})$ is $[\mathsf{X}_i, \mathsf{X}_j]$-causally sufficient if U contains X_i, X_j and "all intervenable variables" that *cause* (definition (1)) both X_i and X_j [1]

Circularity is a recognised problem with interventional definitions of causation (Woodward, 2016). In Section **??**, I further show that assuming ideal interventions always exist leads to counterintuitive conclusions. An alternative approach is to designate certain real-world events – such as flipping coins, querying random number generators and so forth – as prototypical "ideal interventions". This approach is rather inflexible, and refuses to offer answers to

[1] Weaker conditions for causal sufficiency are possible, but they are still premised on causal relationships, so circularity stands (Shpitser and Pearl, 2008).

causal questions that don't happen to have involve just the right kinds of real world events, typically randomised experiments. However, many causal questions do have apparent answers (Pearl, 2018), and even when gold-standard randomised experimental data is available, it may not permit answers to the original questions of interest (Deaton and Cartwright, 2018; Heckman, 1991).

The difficulty in defining "ideal interventions" is not unprecedented. It is also difficult to provide an account of what it means for data to be "distributed according to probability distribution $\mathbb{P}$"(Hájek, 2019), but the usefulness of using probability distributions to model data is widely accepted.

Causal statistical decision theory (CSDT) is a theory of "causal questions" that does not depend on an underlying theory of causation. 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 show that it is also possible to formalise the problem of determining *counterfactual consequences* without appealing to an underlying theory of causation.

A key feature of CSDT is the importance of the *option set*, which is the set of decisions, acts or counterfactual actions under consideration in a given problem. A great deal of work on causal inference defines with the option set implicitly, possibly also relying on default choices such as that of "hard intervention". We argue that:

- 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 can affect the difficulty of causal questions
- Hard interventions are not a good choice for default option sets

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 $\ast : E \to \Delta(\{\{*\}, \emptyset\})$ 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.

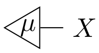
Indiscrete spaces play a key role in string diagrams. An indiscrete space is any one element measurable space $(\{*\}, \{\emptyset, \{*\}\})$ which admits the unique probability measure $\mu : \{\emptyset, \{*\}\} \to (0, 1)$ given by $\mu(\emptyset) = 0$, $\mu(\{*\}) = 1$. Any probability measure $\mu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

A Markov kernel $\mathbb{M} : X \to \varDelta(\mathcal{Y})$ 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 $\mu \in \varDelta(\mathcal{X})$ can be written as triangles:



(7)

Note that Eq 7 technically represents a Markov kernel $\mu' : \{*\} \to \varDelta(\mathcal{X})$, 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:

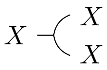
$$\mathbf{Id} : x \mapsto \delta_x \equiv \; X \text{ --- } X \tag{17}$$

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:

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

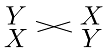


The copy map "copies" its arguments to kernels or under the right product:

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

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 $\mathbb{L} : Z \to \Delta(\mathcal{X} \otimes \mathcal{Y})$, $A \in \mathcal{X}$, $B \in \mathcal{Y}$:

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

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

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

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

Definition 0.1.1 (Kernel associated with a function)**.** Given a measurable function $g : W \to X$, define the function induced kernel $\mathbb{F}^g : W \to \Delta(\mathcal{X})$ to be the the Markov kernel $w \mapsto \delta_{g(w)}$ for all $w \in W$.

Definition 0.1.2 (Pushforward kernel)**.** Given a kernel $\mathbb{M} : V \to \Delta(\mathcal{W})$ and a measurable function $g : W \to X$, the *pushforward kernel* $g_{\#}\mathbb{M} : V \to \Delta(\mathcal{X})$ is the kernel $g_{\#}\mathbb{M}$ such that $(g_{\#}\mathbb{M})_a(B) = \mathbb{M}_a(g^{-1}(B))$ for all $a \in V$, $B \in \mathcal{X}$.

Lemma 0.1.3 (Pushforward kernels are functional kernel products)**.** *Given a kernel* $\mathbb{M} : V \to \Delta(\mathcal{W})$ *and a measurable function* $g : W \to X$*,* $g_{\#}\mathbb{M} = \mathbb{M}\mathbb{F}^g$*.*

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

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

□

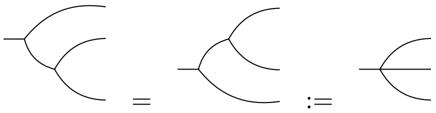
Working With String Diagrams

todo:

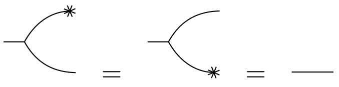
- Infinite copy map
- De Finetti's representation theorem

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.

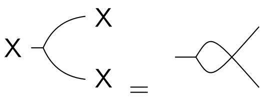
Axioms of Symmetric Monoidal Categories For the following, we either omit labels or label diagrams with their domain and codomain spaces, as we are discussing identities of kernels rather than identities of components of a condtional probability space. Recalling the unique Markov kernels defined above,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



(39)

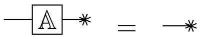


(40)



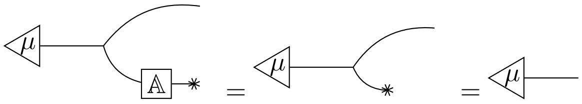
(41)

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



(42)

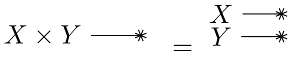
Combining 40 and 42 we can derive the following: integrating $\mathbb{A} : X \to \Delta(\mathcal{Y})$ with respect to $\mu \in \Delta(\mathcal{X})$ and then discarding the output of $\mathbb{A}$ leaves us with μ:



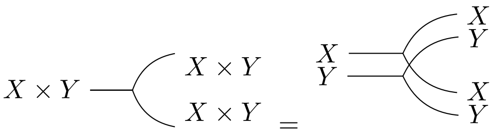
(43)

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 $\curlyvee$:

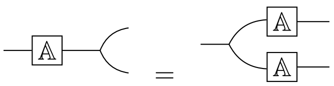


(44)



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

Theorem 0.1.5 (Copy map commutes for deterministic kernels (Fong, 2013))**.** *Equation 46 holds iff $\mathbb{A}$ is deterministic.*

Examples

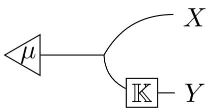
Given $\mu \in \Delta(X), \mathbb{K} : X \to \Delta(Y)$, $A \in \mathcal{X}$ and $B \in \mathcal{Y}$:

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

$$\equiv \tag{48}$$

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

$$\equiv \tag{50}$$

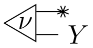


(51)

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

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

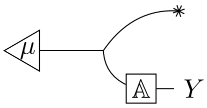


$$\begin{aligned}
\nu(\ast \otimes \mathrm{Id}^Y)(B) &:= \nu(\ast \otimes \mathrm{Id}^Y)(B \times \{*\}) && (53)\\
&= \int_{X\times Y} \mathrm{Id}^Y_y(B)\ast_x(\{*\})d\nu(x,y) && (54)\\
&= \int_{X\times Y} \delta_y(B)\delta_*(\{*\})d\nu(x,y) && (55)\\
&= \int_{X\times B} d\nu(x,y) && (56)\\
&= \nu(X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

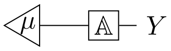
Thus the action of the erasing wire "X" is equivalent to marginalising over "X".

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:

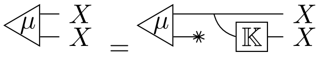
- Random variables are usually defined as measurable functions on a *probability space*
- It's possible to define them as measurable functions on a *Markov kernel space* instead
- It is useful to label wires with random variable names instead of names of spaces

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.

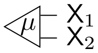
Suppose that given a particular distribution over decision variables, a probability distribution over the decision variables and random variables is obtained. Such a model is described by a Markov kernel rather than a probability distribution. We therefore investigate *Markov kernel spaces.*

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:

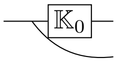


(62)

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 0.1.6 (Probability space, Markov kernel space)**.** A *Markov kernel space* $(\mathbb{K}, \Omega, \mathcal{F}, D, \mathcal{D})$ 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: $(\mathbb{K}, \Omega, D)$.

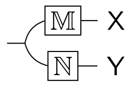
A *probability space* $(\mathbb{P}, \Omega, \mathcal{F})$ is a probability measure $\mathbb{P} : \Delta(\Omega)$, which we call the *ambient measure*, along with the *sample space* Ω and the *events* $\mathcal{F}$. A probability space is equivalent to a Markov kernel space with domain $D = \{*\}$ - note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

Definition 0.1.7 (Random variable)**.** Given a Markov kernel space $(\mathbb{K}, \Omega, D)$, a random variable X is a measurable function $\Omega \times D \to E$ for arbitrary measurable E.

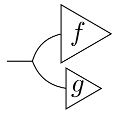
Definition 0.1.8 (Domain variable)**.** Given a Markov kernel space $(\mathbb{K}, \Omega, D)$, the *domain variable* $\mathsf{D} : \Omega \times D \to D$ is the distinguished random variable $\mathsf{D} : (x, d) \mapsto d$.

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



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



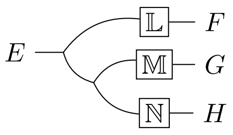
$$f\underline{\otimes}g := \tag{65}$$

The operation denoted by $\underline{\otimes}$ is associative (Lemma 0.1.10), so we can without ambiguity write $f\underline{\otimes}g\underline{\otimes}h = (f\underline{\otimes}g)\underline{\otimes}h = f\underline{\otimes}(g\underline{\otimes}h)$ for finite groups of functions or Markov kernels sharing a domain.

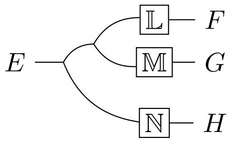
The notation $\underline{\bigotimes}_{i\in[N]} f_i$ is taken to mean $f_1\underline{\otimes}f_2\underline{\otimes}...\underline{\otimes}f_N$.

Lemma 0.1.10 ($\underline{\otimes}$ is associative)**.** *For Markov kernels* $\mathbb{L} : E \to \delta(\mathcal{F})$*,* $\mathbb{M} : E \to \delta(\mathcal{G})$ *and* $\mathbb{N} : E \to \delta(\mathcal{H})$*,* $(\mathbb{L}\underline{\otimes}\mathbb{M})\underline{\otimes}\mathbb{N} = \mathbb{L}\underline{\otimes}(\mathbb{M}\underline{\otimes}\mathbb{N})$*.*

Proof.



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



$$= \tag{67}$$

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

This follows directly from Equation 39. □

Definition 0.1.11 (Marginal distribution, marginal kernel)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ and the random variable $\mathsf{X} : \Omega \to G$ the *marginal distribution* of X is the probability measure $\mathbb{P}^{\mathsf{X}} := \mathbb{P}\mathbb{F}^{\mathsf{X}}$.

See Lemma 0.1.3 for the proof that this matches the usual definition of marginal distribution.

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

Definition 0.1.12 (Joint distribution, joint kernel)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ and the random variables $\mathsf{X} : \Omega \to G$ and $\mathsf{Y} : \Omega \to H$, the *joint distribution* of 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 0.1.13 which follows.

Lemma 0.1.13 (Product marginalisation interchange)**.** *Given two functions, the kernel associated with their coupled product is equal to the coupled product of the kernels associated with each function.*

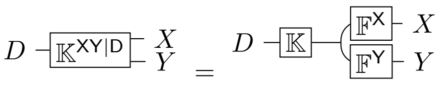
Given $\mathsf{X} : \Omega \to G$ *and* $\mathsf{Y} : \Omega \to H$, $\mathbb{F}^{\mathsf{X}\underline{\otimes}\mathsf{Y}} = \mathbb{F}^{\mathsf{X}}\underline{\otimes}\mathbb{F}^{\mathsf{Y}}$

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

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

Equality follows from the monotone class theorem. □

Corollary 0.1.14. *Given a Markov kernel space* $(\mathbb{K}, \Omega, D)$ *and random variables* $\mathsf{X} : \Omega \times D \to X$, $\mathsf{Y} : \Omega \times D \to Y$*, the following holds:*

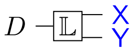


D
$\mathbb{K}^{\mathsf{XY|D}}$
X
Y
$=$
D
$\mathbb{K}$
$\mathbb{F}^{\mathsf{X}}$
X
$\mathbb{F}^{\mathsf{Y}}$
Y

(73)

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

Definition 0.1.15 (Wire labels - joint kernels)**.** Suppose we have a Markov kernel space $(\mathbb{K}, \Omega, D)$, 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 $\mathbf{L}$:

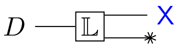


(74)

is *valid* iff

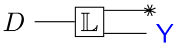
$$\mathbb{L} = \mathbb{K}_{\mathsf{XY|D}} \tag{75}$$

and



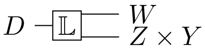
$$= \mathbb{K}^{\mathsf{X}|\mathsf{D}} \tag{76}$$

and



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

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

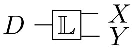


(78)

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

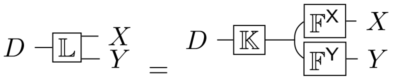
Lemma 0.1.16 (Output label assignments exist)**.** *Given Markov kernel space* $(\mathbb{K}, \Omega, D)$*, random variables* $\mathsf{X} : \Omega \times D \to X$ *and* $\mathsf{Y} : \Omega \times D \to Y$ *then there exists a diagram of* $\mathbb{L} := \mathbb{K}^{\mathsf{XY}|\mathsf{D}}$ *with a valid 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



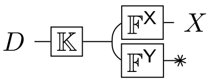
(79)

By Corollary 0.1.14, we have

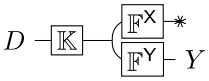


(80)

Therefore



$$= \mathbb{K}\mathbb{F}^{\mathsf{X}} \tag{81}$$
$$= \mathbb{K}^{\mathsf{X}|\mathsf{D}} \tag{82}$$



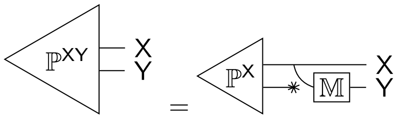
$$= \mathbb{K}\mathbb{F}^{\mathsf{Y}} \tag{83}$$

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

□

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

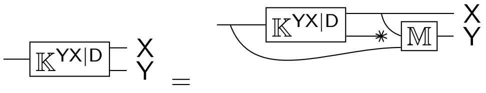
Definition 0.1.17 (Disintegration)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$, and random variables X and Y, we say that $\mathbb{M} : E \to \Delta(\mathcal{F})$ is a Y *on* X *disintegration* of $\mathbb{P}$ iff



(85)

$\mathbb{M}$ is a version of $\mathbb{P}^{\mathsf{Y|X}}$, "the probability of Y given X". Let $\mathbb{P}^{\{\mathsf{Y|X}\}}$ be the set of all kernels that satisfy 85 and $\mathbb{P}^{\mathsf{Y|X}}$ an arbitrary member of $\mathbb{P}^{\mathsf{Y|X}}$.

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



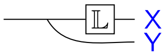
(86)

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

Definition 0.1.18 (Wire labels – input)**.** An input wire is *connected* to an output wire if it is possible to trace a path from the start of the input wire to the end of the output wire without passing through any boxes, erase maps or right facing triangles.

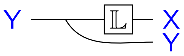
If an input wire is connected to an output wire and that output wire has a valid label X, then it is valid to label the input wire with X.

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



(87)

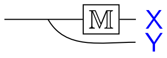
i.e. if $\mathbb{L} \in \mathbb{P}^{\{\mathsf{XY|Y}\}}$, then the following is a valid input label:



(88)

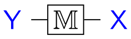
An input wire in a diagram for $\mathbb{M}$ may be labeled X *if and only if* copy and identity maps can be inserted to yield a diagram in which the input wire labeled X is connected to an output wire with valid label X.

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



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

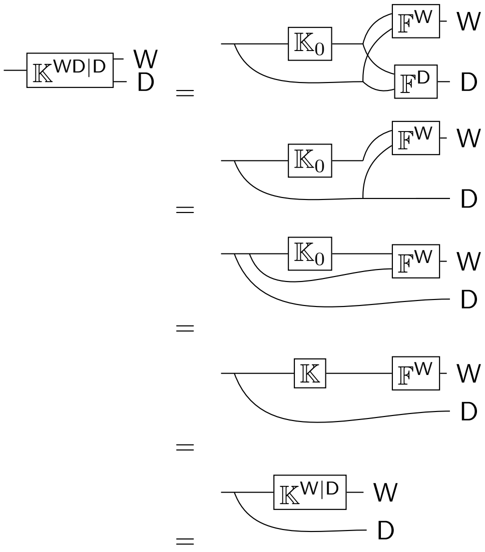
and hence the output labels are valid. Diagram 89 is constructed by taking the product of the copy map with $\mathbb{M} \otimes \mathbf{Id}$. Thus it is valid to label $\mathbb{M}$ with



(90)

Lemma 0.1.19 (Labeling of disintegrations)**.** *Given a kernel space* $(\mathbb{K}, \Omega, D)$*, 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)

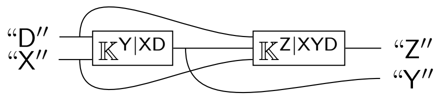
(94)

(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 0.1.20 (Iterated disintegration)**.** *Given a kernel space* $(\mathbb{K}, \Omega, D)$*, 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″

$$\in \mathbb{K}^{\{\mathsf{ZY}|\mathsf{XD}\}} \tag{96}$$

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

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

Proof.

write this up

□

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

Theorem 0.1.21 (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 $(\mathbb{K}, \Omega, D)$ 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 0.1.22 (Relative probability space)**.**

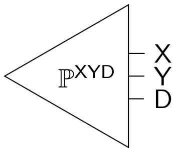
better name

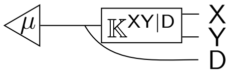
Given a Markov kernel space $(\mathbb{K}, \Omega, D)$ and a positive definite measure $\mu \in \Delta(\mathcal{D})$, $(\mu\mathbb{K}, \Omega \times D)$ is a *relative* probability space.

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

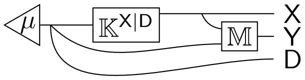
Lemma 0.1.23 (Agreement of disintegrations)**.** *Given a Markov kernel space* $(\mathbb{K}, \Omega, D)$*, any relative probability space* $(\mu\mathbb{K}, \Omega \times D)$ *and any random variables* $\mathsf{X} : \Omega \times D \to X$*,* $\mathsf{Y} : \Omega \times D \to Y$*,* $\mathbb{K}^{\{\mathsf{Y}|\mathsf{XD}\}} = (\mu\mathbb{K})^{\{\mathsf{Y}|\mathsf{XD}\}}$ *(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

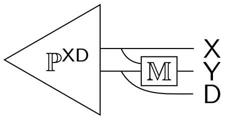




$$= \tag{98}$$



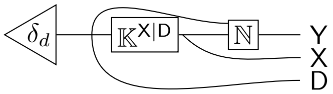
$$= \tag{99}$$



$$= \tag{100}$$

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 $d \in D$



$$\mathbb{Q} := \tag{101}$$

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

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:

$$\begin{aligned}
\mathbb{Q}(A\times B\times\{d\}) &= \int_{X^2}\int_X\int_{D^3}\mathbb{N}_{d'',x'}(A)\mathbf{Id}_{x''}(B)\mathbf{Id}_{d'''}(\{d\})d\curlyvee_d(d',d'',d''')d\mathbb{K}^{\mathsf{X|D}}_{d'}(x)d\curlyvee_x(x',x'') \tag{103}\\
&= \delta_d(\{d\})\int_X \mathbb{N}_{d,x}(A)\delta_x(B)d\mathbb{K}^{\mathsf{X|D}}_d(x) \tag{104}\\
&= \frac{1}{\mu(\{d\})}\int_{\{d\}}d\mu(d')\int_X \mathbb{N}_{d,x}(A)\delta_x(B)d\mathbb{K}^{\mathsf{X|D}}_d(x) \tag{105}\\
&= \frac{1}{\mu(\{d\})}\int_D\int_X \mathbb{N}_{d,x}(A)\delta_{d'}(\{d\})\delta_x(B)d\mathbb{K}^{\mathsf{X|D}}_d(a)d\mu(d') \tag{106}\\
&= \frac{1}{\mu(\{d\})}\int_D\int_X \mathbb{N}_{d,x}(A)\delta_{d'}(\{d\})\delta_x(B)d\mathbb{K}^{\mathsf{X|D}}_{d'}(a)d\mu(d') \tag{107}\\
&= \frac{1}{\mu(\{d\})}\mathbb{P}^{\mathsf{XYD}}(A\times B\times\{d\}) \tag{108}\\
&= \frac{1}{\mu(\{d\})}\int_D \mathbb{K}^{\mathsf{XYD|D}}_{d'}(A\times B\times\{d\})d\mu(d') \tag{109}\\
&= \frac{1}{\mu(\{d\})}\int_D \mathbb{K}_{d'}\mathsf{XY|D}(A\times B)\delta_{d'}(\{d\})d\mu(d') \tag{110}\\
&= \mathbb{K}^{\mathsf{XY|D}}_d(A\times B) \tag{111}\\
&= \mathbb{K}^{\mathsf{XY|D}}_d(A\times B)\delta_d(\{d\}) \tag{112}\\
&= \int_D \mathbb{K}^{\mathsf{XY}}_{d'}(A\times B)\delta_{d''}(\{d\})d\curlyvee_d(d',d'') \tag{113}\\
&= \mathbb{K}^{\mathsf{XYD|D}}_d(A\times B\times\{d\}) \tag{114}
\end{aligned}$$

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 $(\mathbb{K},\Omega,D)$.

Conditional Independence

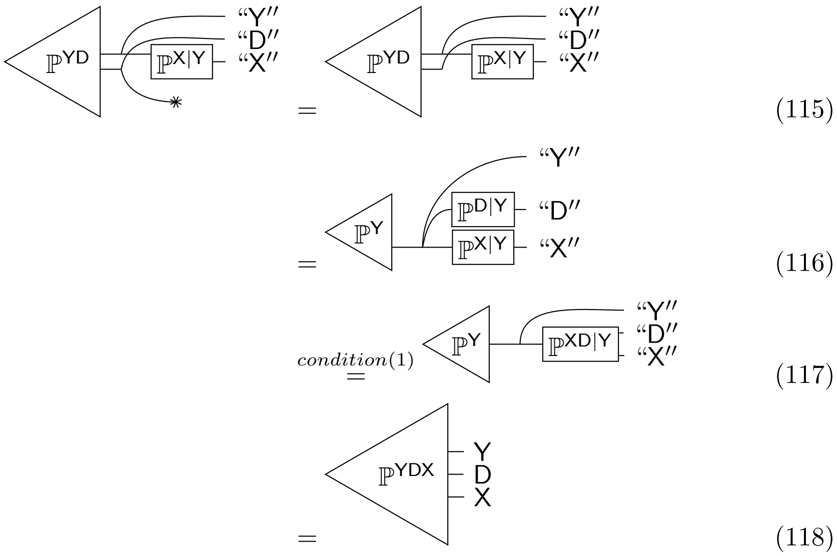
Definition 0.1.24 (Kernels constant in an argument)**.** Given a kernel $(\mathbb{K},\Omega,D)$ and random variables Y and X, we say a verstion of the disintegration $\mathbb{K}^{\mathsf{Y|XD}}$ is constant in D if for all $x\in X$, $d,d'\in D$, $\mathbb{K}^{\mathsf{Y|XD}}_{(x,d)}=\mathbb{K}^{\mathsf{Y|XD}}_{(x,d')}$.

Definition 0.1.25 (Domain Conditional Independence)**.** Given a kernel space $(\mathbb{K},\Omega,D)$, 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

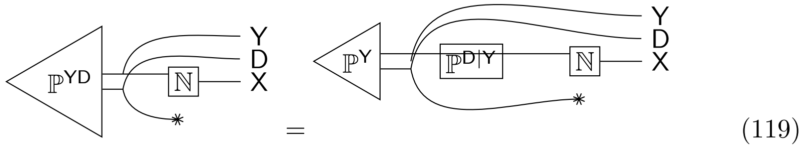
- $\mathbb{P}^{\mathsf{XD|Y}} \sim \mathbb{P}^{\mathsf{X|Y}} \underline{\otimes} \mathbb{P}^{\mathsf{D|Y}}$
- For any version of $\mathbb{P}^{\{\mathsf{X|Y}\}}$, $\mathbb{P}^{\mathsf{X|Y}} \otimes *_D$ is a version of $\mathbb{K}^{\{\mathsf{X|YD}\}}$
- There exists a version of $\mathbb{K}^{\{\mathsf{X|YD}\}}$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

Theorem 0.1.26 (Definitions are equivalent)**.** *(1) $\implies$ (2): By Lemma 0.1.23, $\mathbb{P}^{\{\mathsf{Y|XD}\}} = \mathbb{K}^{\{\mathsf{Y|XD}\}}$. Thus it is sufficient to show that $\mathbb{P}^{\mathsf{X|Y}} \otimes *$ is a version of $\mathbb{P}^{\{\mathsf{X|YD}\}}$.*


$\mathbb{P}^{\mathsf{YD}}$
$\mathbb{P}^{\mathsf{X|Y}}$
"Y"
"D"
"X"
*
=
$\mathbb{P}^{\mathsf{YD}}$
$\mathbb{P}^{\mathsf{X|Y}}$
"Y"
"D"
"X"
(115)
"Y"
$\mathbb{P}^{\mathsf{Y}}$
$\mathbb{P}^{\mathsf{D|Y}}$
"D"
$\mathbb{P}^{\mathsf{X|Y}}$
"X"
=
(116)
condition(1)
=
$\mathbb{P}^{\mathsf{Y}}$
$\mathbb{P}^{\mathsf{XD|Y}}$
"Y"
"D"
"X"
(117)
$\mathbb{P}^{\mathsf{YDX}}$
Y
D
X
=
(118)

(2) $\implies$ (3)
$\mathbb{P}^{\mathsf{X|Y}} \otimes *_D$ is a version of $\mathbb{K}^{\{\mathsf{X|YD}\}}$ by assumption, and is clearly constant in D.

(3) $\implies$ (1)
By lemma 0.1.23, there also exists a version of $\mathbb{P}^{\{\mathsf{X|YD}\}}$ 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 $\mathbb{P}^{\mathsf{X|Y}} \underline{\otimes} \mathbb{P}^{\mathsf{D|Y}} \in \mathbb{P}^{\{\mathsf{XD|Y}\}}$. By Theorem 0.1.20, we have*


$\mathbb{P}^{\mathsf{YD}}$
$\mathbb{N}$
Y
D
X
*
=
$\mathbb{P}^{\mathsf{Y}}$
$\mathbb{P}^{\mathsf{D|Y}}$
$\mathbb{N}$
Y
D
X
*
(119)

Definition 0.1.27 (Conditional probability existence)**.** Given a kernel space $(\mathbb{K}, \Omega, D)$ 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 0.1.28 (Conditional Independence)**.** Given a kernel space $(\mathbb{K}, \Omega, D)$, 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

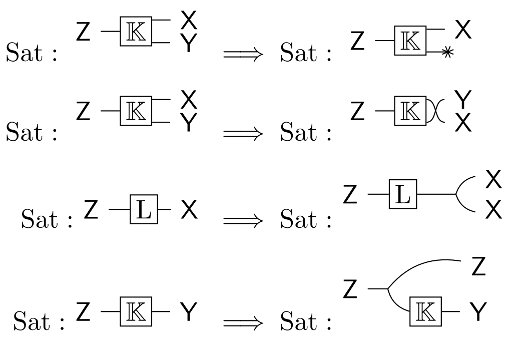
- $\mathbb{P}^{\mathsf{XZ}|\mathsf{Y}} \sim \mathbb{P}^{\mathsf{X}|\mathsf{Y}} \underline{\otimes} \mathbb{P}^{\mathsf{Z}|\mathsf{Y}}$
- For any version of $\mathbb{P}^{\{\mathsf{X}|\mathsf{Y}\}}$, $\mathbb{P}^{\mathsf{X}|\mathsf{Y}} \otimes *_Z$ is a version of $\mathbb{K}^{\{\mathsf{X}|\mathsf{YZ}\}}$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

Lemma 0.1.29 (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 $(\mathbb{K}, \Omega, D)$ *and random variables* X, Y, Z *taking values in* X, Y, Z *respectively.* Sat : *indicates that a labeled diagram satisfies definitions 0.1.15 and 0.1.18 with respect to* $(\mathcal{K}, \Omega, D)$ *and* X, Y, Z. *The following always holds:*

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

and the following implications hold:



(121)

(122)

(123)

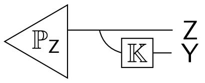
(124)

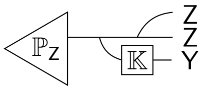
Proof.

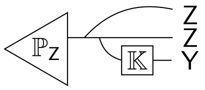
- Id_X is a version of $\mathbb{P}_{\mathsf{X}|\mathsf{X}}$ for all $\mathbb{P}$; $\mathbb{P}_{\mathsf{X}}\mathrm{Id}_X = \mathbb{P}_{\mathsf{X}}$
- $\mathbb{K}\mathrm{Id} \otimes *)(w; A) = \int_{X \times Y} \delta_x(A) \mathbb{1}_Y(y) d\mathbb{K}_w(x, y) = \mathbb{K}_w(A \times Y) = \mathbb{P}_{\mathsf{X}|\mathsf{Z}}(w;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

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

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

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

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

$=$  (127)

Therefore $\curlyvee(\mathrm{Id}_X \otimes \mathbb{K})$ is a version of $\mathbb{P}_{\mathsf{ZY|Z}}$ 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 $\mathbb{P}_{\mathsf{Y|X}} = x \mapsto \text{Bernoulli}(0.5)$ for all $z \in Z$. Then $\mathbb{P}_{\mathsf{Z|Y}} = y \mapsto \mathbb{P}_{\mathsf{Z}}$, $\forall y \in Y$ and therefore $\mathbb{P}_{\mathsf{Y|Z}}\mathbb{P}_{\mathsf{Z|Y}} = x \mapsto \mathbb{P}_{\mathsf{Z}}$ but $\mathbb{P}_{\mathsf{Z|X}} = x \mapsto \delta_x \neq \mathbb{P}_{\mathsf{Y|Z}}\mathbb{P}_{\mathsf{Z|Y}}$.

Chapter 1

Chapter 2: See-do models

A *statistical model* is an indexed set of probability distributions $\{\mathbb{P}_\theta|\theta \in \Theta\}$ where each distribution shares a sample space $\mathbb{P}_\theta \in \Delta(\mathcal{E})$, $\forall\theta$. Equivalently, a statistical model is a map $\mathbb{P} : \Theta \to \Delta(\mathcal{E})$. Statistical models are ubiquitous – they are found in statistical decision theory where the elements of Θ are typically called "states"(Wald, 1950), in Bayesian inference where the elements of Θ may be called "parameters" (Freedman, 1963) and in frequentist inference where elements of Θ they may be called "hypotheses" (Fisher, 1992).

Though the precise terminology might differ, in all cases described a statistical model can be interpreted as formalising the assumption that, if the true hypothesis (or state, or parameter) is θ, then the observations we receive are distributed according to $\mathbb{P}_\theta$. A further piece of interpretation – which we are not compelled to accept, but typically do – is that we have no ability to choose whether or not a particular hypothesis θ is true, no matter how much we might want it to be. Given a collection of ways the world could function, Θ, statistical models tell us what we would be likely to see for each possibility $\theta \in \Theta$. It is up to Mother Nature to determine which θ actually describes the world, but if we are lucky we might be able to infer something about it from our observations.

Sometimes, though, we do get to make some choices, and these choices which affect what we are likely to see. I can choose to turn the light on or off, which will affect whether or not I have a bright room or a dim room.

Here we are interested in a class of model where "we" (specifically, the model's user) *can* make a choice that influences whatever happens in the end. Not all outcomes can be influenced – some may be *observations* collected prior to making a choice, for example. In these models, *hypotheses* serve two roles: firstly, given a hypothesis θ we know that *observations* are distributed according to $\mathbb{P}_\theta$ for some $\mathbb{P} : \Theta \to \Delta(\mathcal{E})$, just as with statistical models. Secondly, a hypothesis θ *also* tells us that the consequences of each choice $d \in D$, given by $\mathbb{C}_{\theta,d} \in \Delta(\mathcal{F})$ for some $\mathbb{C} : \Theta \to \Delta(\mathcal{F})$. Thus the model tells us, for each hypothesis θ, what we expect to *see* and what we expect to happen as a result of *doing* something, hence the name "see-do model".

See-do models are a straightforward extension of statistical models, adding a "do model" to the "see model" that statistical models already provide. We shall also show that they are closely related to *statistical decision problems*, which also consider the possibility of making a choice after observing a given dataset, but unlike see-do models statistical decision problems do not allow for general consequences of decisions.

See-do models can formalise causal questions, whether of the "associational", "interventional" or "counterfactual" types, in the terminology of Pearl and Mackenzie (2018). In 2, I show that see-do models overlap significantly with existing approaches to formalising questions of these types and argue that where see-do models differ, they are favoured by the difference.

1.1 Definition

The primitives of

- *Option sets*; the set of options I have available
- *Observations*; the observations I might be given
- *Consequences*; what might arise as a result of my having chosen a particular option
- *Preferences*; which consequences I want, and which I do not want
- *Uncertainty*; I am uncertain about which consequences are brought about by my available options, and I may be able to become less uncertain after considering the data I have been given

See-Do Models formalise the above notions. See-Do Models use expected utility theory to formalise preferences and probability theory to model noisy observations and uncertainty over consequences. We make these choices because these are well understood and widely accepted tools for modelling preferences and uncertainty respectively. For a given CSDP, See-Do Models regard the *option set*, the *observation space* and the *consequence space* to be fixed.

Definition 1.1.1 (Option set)**.** An *option set* D is a finite or countable set of options. A decision maker may select any option from D and in addition may select any mixture of options from $\Delta(\mathcal{D})$.

Definition 1.1.2 (Observation space)**.** The decision maker receives an *observation*, which is an element of a standard measurable set $(X, \mathcal{X})$.

Definition 1.1.3 (Consequence space)**.** The decision maker's choice of option results in a *consequence*, which is an element of a standard measurable set $(Y, \mathcal{Y})$.

We allow for two types of uncertainty over which consequences will actually take place given the selection of a particular option. Consequences may be stochastic functions of the option selected and of the given data. Secondly, a decision maker may entertain a number of different *hypotheses* about the relationship between data, decision and consequences. A decision maker may select a particular distribution of hypotheses called a *prior* so that the only remaining uncertainty is stochastic, but we avoid assuming that a canonical prior is available at the outset.

Definition 1.1.4 (See-Do Model)**.** A See-Do Model takes a hypothesis $\theta \in \Theta$ and an option $d \in D$ and returns a joint distribution over observations and consequences with the restriction that, given a particular hypothesis, observations are independent of options. Formally, a See-Do Model is a kernel space $(\mathbb{T}, X \times Y, \Theta \times D)$ where $\mathbb{T} : \Theta \times D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ where Θ is the hypothesis space, D is the option set, X is the observation space and Y is the consequence space. Defining $\mathsf{X}, \mathsf{Y}, \mathsf{D}, \Theta$ to be projection maps to the associated spaces, a see-do model has the property $\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\Theta$. That is, for each hypothesis $\theta \in \Theta$, $\mathbb{T}$ holds that the observations X are independent of the choice of option D.

It is therefore possible to specify a see-do model $\mathbb{T}$ with the following elements:

- Hypothesis space Θ, options D, observations X and consequences Y
- Observation map $\mathbb{T}^{\mathsf{X}|\Theta}$, which exists by virtue of the independence of observations from consequences
- Consequence map $\mathbb{T}^{\mathsf{Y}|\Theta\mathsf{X}\mathsf{D}}$

Definition 1.1.5 (Hypothesis sufficiency)**.** According to the general definition of a see-do model, observations provide evidence about which hypothesis θ is correct *and also* may directly affect the consequences. See-do models may be simplified if only the hypothesis and decision affects the consequence. The hypotheses are *sufficient* for a See-Do Model $(\mathbb{T}, X \times Y, \Theta \times D)$ if $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{X}|\Theta\mathsf{D}$.

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

- Hypothesis space Θ, options D, observations X and consequences Y
- Observation map $\mathbb{T}^{\mathsf{X}|\Theta}$, which exists by virtue of the independence of observations from consequences
- Consequence map $\mathbb{T}^{\mathsf{Y}|\Theta\mathsf{D}}$

Definition 1.1.6 (Bayesian See-Do Model)**.** A Bayesian See-Do Model $\mathbb{U}$ is a Markov kernel $D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ with the property $\mathsf{X} \perp\!\!\!\perp_{\mathbb{U}} \mathsf{D}$.

A Bayesian See-Do Model can be constructed from a see-do model $\mathbb{T}$ by choosing an arbitrary prior $\gamma \in \Delta(\Theta)$ and taking the product:

$$\mathbb{U} = (\gamma \otimes \mathrm{Id}^D)\mathbb{T} \tag{1.1}$$

For all $A \in \mathcal{X}$, $B \in \mathcal{Y}$, $d \in D$:

$$\mathbb{U}_d(A \times B) = \int_\Theta \int_A \mathbb{T}^{\mathsf{Y}|\mathsf{X}\mathsf{D}\Theta}_{\theta,x,d}(B) d\mathbb{T}^{\mathsf{X}|\Theta}_\theta(x) d\gamma(\theta) \tag{1.2}$$

In this case, $\mathbb{U}^{\mathsf{Y}|\mathsf{X}\mathsf{D}\Theta} \overset{a.s.}{=} \mathbb{T}^{\mathsf{Y}|\mathsf{X}\mathsf{D}\Theta}$ and $\mathbb{U}^{\mathsf{X}|\Theta} \overset{a.s.}{=} \mathbb{T}^{\mathsf{X}|\Theta}$

Examples of see-do models

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 $\mathbb{B} : (0,1) \to \mathit{\Delta}(\{0,1\})$ by $\mathbb{B} : \theta \mapsto \text{Bernoulli}(\theta)$. Then define $\mathbb{T}^{(1)}$ by:

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

In this model, the chance θ 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 sneak a 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 $\mathbb{T}^{(2)}$:

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

In this case, the observed data tells us more about how the bet will work out than the hypothesis alone.

It is also 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 $\mathit{\Theta}' = (0,1) \times \{0,1\}$ and define $\mathbb{T}^{(3)}$ by:

- $D = \{0,1\}$
- $X = \{0,1\}^{N+1}$
- $Y = \{0,1\}$
- $\mathit{\Theta}' = (0,1) \times \{0,1\}$

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

However, X_{N+1} is related to the previous flips $\boldsymbol{\mathsf{X}}_{<N}$. In particular, given $\theta \in \Theta$, X_{N+1} should be distributed according to Bernoulli(θ). That is, defining $\Theta : \Theta' \times D \times X \times Y \to (0,1)$ to be the projection map that yields the parameter θ, any Bayesian model $\mathbb{U}$ should have the property $\mathbb{U}^{\mathsf{X}_{N+1}|\Theta} = \mathbb{B}$. Then for any $A \in \sigma(\{0,1\}^N), B \in \sigma(\{0,1\}), C \in \sigma(\{0,1\})$

$$\mathbb{U}_d^{(3)}(A \times B \times C) = \int_\Theta \int_A \int_B \mathbb{T}_{\theta',\mathbf{x},d}^{\mathsf{Y}|\mathsf{XD}\Theta'(3)}(C) d\mathbb{U}_\theta^{\mathsf{X}_{N+1}|\Theta}(x_{N+1}) d\mathbb{T}_\theta^{\mathsf{X}|\Theta'(3)}(\mathbf{x}) d\mathbb{U}^\Theta(\theta) \tag{1.3}$$

$$= \int_\Theta \int_A \int_B \mathbb{T}_{\theta',\mathbf{x},d}^{\mathsf{Y}|\mathsf{XD}\Theta'(3)}(C) d\mathbb{B}_\theta(x_{N+1}) d(\curlyvee^N \mathbb{B})(\mathbf{x}) d\mathbb{U}^\Theta(\theta) \tag{1.4}$$

$$= \int_\Theta \int_{A\times B} \mathbb{T}_{\theta',\mathbf{x},d}^{\mathsf{Y}|\mathsf{XD}\Theta'(2)}(C) d(\curlyvee^{N+1} \mathbb{B})(\mathbf{x}) d\mathbb{U}^\Theta(\theta) \tag{1.5}$$

$$\tag{1.6}$$

Which is equivalent to $\mathbb{T}^{(2)}$ equipped with the prior $\mathbb{U}^\Theta \in \Delta(\Theta)$. Thus the difference between $\mathbb{T}^{(2)}$ and $\mathbb{T}^{(3)}$ is whether a constraint is expressed in the model (as in $\mathbb{T}^{(2)}$) or in over the class of allowable priors (as in $\mathbb{T}^{(3)}$).

I don't have a theory of stochastic vs non-stochastic uncertainty, but it is the case in general that a hypothesis sufficient model with additional restrictions on the prior can be replaced by a hypothesis insufficient model with no restrictions on the prior. This is mainly relevant with regard to counterfactuals

- *Option sets*; the set of options I have available
- *Observations*; the observations I might be given
- *Consequences*; what might arise as a result of my having chosen a particular option
- *Preferences*; which consequences I want, and which I do not want
- *Uncertainty*; I am uncertain about which consequences are brought about by my available options, and I may be able to become less uncertain after considering the data I have been given

See-Do Models formalise the above notions. See-Do Models use expected utility theory to formalise preferences and probability theory to model noisy observations and uncertainty over consequences. We make these choices because these are well understood and widely accepted tools for modelling preferences and uncertainty respectively. For a given CSDP, See-Do Models regard the *option set*, the *observation space* and the *consequence space* to be fixed.

Definition 1.1.7 (Option set)**.** An *option set* D is a finite or countable set of options. A decision maker may select any option from D and in addition may select any mixture of options from $\Delta(\mathcal{D})$.

Definition 1.1.8 (Observation space)**.** The decision maker receives an *observation*, which is an element of a standard measurable set $(X, \mathcal{X})$.

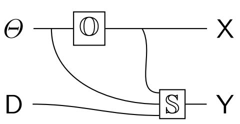
Definition 1.1.9 (Consequence space)**.** The decision maker's choice of option results in a *consequence*, which is an element of a standard measurable set $(Y, \mathcal{Y})$.

We allow for two types of uncertainty over which consequences will actually take place given the selection of a particular option. Consequences may be stochastic functions of the option selected and of the given data. Secondly, a decision maker may entertain a number of different *hypotheses* about the relationship between data, decision and consequences. A decision maker may select a particular distribution of hypotheses called a *prior* so that the only remaining uncertainty is stochastic, but we avoid assuming that a canonical prior is available at the outset.

Definition 1.1.10 (See-Do Model)**.** A See-Do Model takes a hypothesis $\theta \in \Theta$ and an option $d \in D$ and returns a joint distribution over observations and consequences with the restriction that, given a particular hypothesis, observations are independent of options. Formally, a See-Do Model is a Markov kernel $\mathbb{T} : \Theta \times D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ where, given the obvious definitions of observations X, consequences Y, options D and hypothesis Θ, $\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\Theta$.

Letting $\mathbb{O}$ be a version of $\mathbb{T}^{\mathsf{X}|\Theta}$ and letting $\mathbb{S}$ be a version of $\mathbb{T}^{\mathsf{Y}|\mathsf{D}\mathsf{X}\Theta}$ we can write

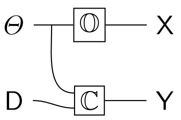
$\mathbb{T} =$



(1.7)

Definition 1.1.11 (Hypothesis sufficiency)**.** According to the general definition, observations provide evidence about which hypothesis θ is correct *and also* may directly affect the consequences. The hypotheses are *sufficient* for a See-Do Model $\mathbb{T} : \Theta \times D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ if $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{X}|\Theta\mathsf{D}$. In this case, letting $\mathbb{O}$ be a version of $\mathbb{T}^{\mathsf{X}|\Theta}$ and $\mathbb{C}$ be a version of $\mathbb{T}^{\mathsf{Y}|\mathsf{D}\Theta}$ we can write

$\mathbb{T} =$

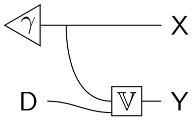


(1.8)

To specify a see-do model with sufficient hypotheses we require an observation model $\mathbb{O} : \Theta \to \Delta(\mathcal{X})$ and a consequence map $\mathbb{C} : \Theta \times D \to \Delta(\mathcal{Y})$. To specify an insufficient model, we require an observation model (as before) and a *state-dependent* consequence map $\mathbb{S} : \Theta \times X \times D \to \Delta(\mathcal{Y})$.

Definition 1.1.12 (Bayesian See-Do Model)**.** A Bayesian See-Do Model $\mathbb{T}$ is a Markov kernel $D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ with the property $\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}$. Letting $\mathbb{V} \in \mathbb{T}^{\mathsf{Y}|\mathsf{X}\mathsf{D}}$, it can be written in the form

$\mathbb{T} =$

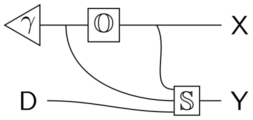


(1.9)

A Bayesian See-Do Model can be constructed from a See-Do Model $\mathbb{T}'$ by choosing an arbitrary prior $\gamma \in \Delta(\Theta)$ and taking the product:

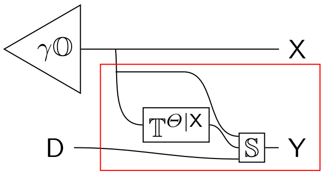
$$\mathbb{T} = (\gamma \otimes \mathrm{Id}^D)\mathbb{T}' \tag{1.10}$$

$=$



(1.11)

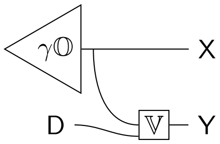
$=$



(1.12)

The existence of $\mathbb{T}^{\Theta|\mathsf{X}}$ follows from the fact that Θ is independent of D in 1.11. Define $\mathbb{V}$ as the kernel in the red box. Then

$\mathbb{T} =$



(1.13)

Examples of see-do models

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 $\mathbb{B} : \theta \mapsto \text{Bernoulli}(\theta)$. Then define $\mathbb{T}^1$ by:

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

- $\Theta = (0,1)$
- $\mathbb{O} : \curlyvee^N \mathbb{B}$
- $\mathbb{C} : (\theta, d) \mapsto \text{Bernoulli}(1 - |d - \theta|)$

Where $\otimes^N$ indicates the tensor product copied N times. The chance θ 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 sneak a 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 $\mathbb{T}^2$:

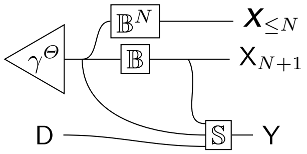
- $D = \{0,1\}$
- $X = \{0,1\}^{N+1}$
- $Y = \{0,1\}$
- $\Theta = (0,1)$
- $\mathbb{O} : \curlyvee^{N+1} \mathbb{B}$
- $\mathbb{S} : (\theta, \mathbf{x}, d) \mapsto \delta_{1-|d-x_{N+1}|}$

It is also possible to model the second situation with a hypothesis sufficient model if we include the result of the $N+1$th flip in the hypothesis. Define $\mathbb{T}^3$ by the elements:

- $D = \{0,1\}$
- $X = \{0,1\}^{N+1}$
- $Y = \{0,1\}$
- $\Theta = (0,1) \times \{0,1\}$
- $\mathbb{O} : (\curlyvee^N \mathbb{B} \otimes \delta_{x_{N+1}}$
- $\mathbb{S} : (\theta, x_{N+1}, d) \mapsto \delta_{1-|d-x_{N+1}|}$

However, X_{N+1} is related to the prior flips $\boldsymbol{\mathsf{X}}_{<N}$. In particular, given $\theta \in \Theta$, X_{N+1} should be distributed according to Bernoulli(θ). If we require that any prior γ over $\Theta \times \{0,1\}$ have this property, then definining $\mathbb{B}^N := \curlyvee^N \mathbb{B}$, the model will factorise as

$(\gamma \otimes \mathrm{Id}^D)\mathbb{T}^3 =$



(1.14)

$$= (\gamma^\Theta \otimes \mathrm{Id}^D)\mathbb{T}^2 \tag{1.15}$$

The only real choice that can be made about the prior is γ^{Θ}, and adding the requirement that X_{N+1} is distributed as Bernoulli(θ) to $\mathbb{T}^3$ yields $\mathbb{T}^2$.

I don't have a theory of stochastic vs non-stochastic uncertainty, but it is the case in general that a hypothesis sufficient model with additional restrictions on the prior can be replaced by a hypothesis insufficient model with no restrictions on the prior. This is mainly relevant with regard to counterfactuals

1.2 Causal Questions

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

1. Association: How are X and Y related? How would observing X change my beliefs about Y?
2. Intervention: What would happen if I do ... ? How can I make E happen?
3. Counterfactual: What if I had done ... instead of what I actually did?

I will initially focus on the second question type: "How can I make E happen?", and later show how the approach for this type of question can be generalised to handle questions of the third type. Call this kind of question a *causal decision problem*:

> Given available options, which ones are most likely to lead to a desirable result?

Causal *statistical* causal decision problems extend causal decision problems by introducing data:

> Given my available options and data, which options are likely to lead to a desirable result?

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

Definition 1.2.2 (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 1.2.3 (Decision rule)**.** Given a see-to map $\mathbb{T} : \Theta \times D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$, 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 1.2.4 (Risk)**.** Given a see-to map $\mathbb{T} : \Theta \times D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$, a utility $u : Y \to \mathbb{R}$ and the set of decision rules $\mathcal{U}$, the *risk* is a function $l : \Theta \times \mathcal{U} \to \mathbb{R}$ given by

$$R(\theta, \mathbb{U}) := -\int_X \mathbb{U}_x \mathbb{T}^{\mathsf{Y|DX}\Theta}_{\cdot,x,\theta} u d\mathbb{T}^{\mathsf{X}|\Theta}_{\theta}(x) \tag{1.16}$$

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

The loss induces a partial order on decision rules. If for all θ, $l(\theta, \mathbb{U}) \leq l(\theta, \mathbb{U}')$ then $\mathbb{U}$ is at least as good as $\mathbb{U}'$. If, furthermore, there is some θ_0 such that $l(\theta_0, \mathbb{U}) < l(\theta_0, \mathbb{U}')$ then $\mathbb{U}$ is preferred to $\mathbb{U}'$.

Definition 1.2.5 (Induced statistical decision problem)**.** A see-do model $\mathbb{T} : \Theta \times D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ along with a utility u induces the *statistical decision problem* $(\Theta, \mathcal{U}, R)$ with states Θ, 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

An alternative method of converting hypothesis insufficient to hypothesis sufficient models involves expanding the decision set; this is not applicable to counterfactual models.

A key difference between CSDT and other approaches to causal inference is that diagrams in CSDT feature two coupled maps $\mathbb{O}$ and $\mathbb{C}$, while most other approaches to causal inference represent both $\mathbb{O}$ and $\mathbb{C}$ in one diagram. Lattimore and Rohde (2019) is the only other example I am aware of that represents both $\mathbb{O}$ and $\mathbb{C}$. Nevertheless, "one-picture" causal models such as Causal Bayesian Networks, Single World Intervention Graphs *do* represent observational distributions and interventional maps, and the two differ (see Section **??**)

A causal hypothesis class Θ induces a binary relation between observed probability distributions $\mathbb{O}_\theta$ and consequence maps $\mathbb{C}_\theta$. This approach is very agnostic about the actual relation induced – we do not even insist that the range of the observed data X is the same as the range of possible consequences Y (though we will generally limit our attention to cases where the two coincide).

In common with Heckerman and Shachter (1995), decisions (or “acts”) are primitive elements of See-Do Models. In contrast to our work, Heckerman and Shachter (1995) only discuss deterministic *consequence maps*, while See-Do Models represent relations between consequence maps and observed probability.

Decisions are similar to the “regime indicators” found in Dawid (2020). They coincide precisely if we suppose that the observation and consequence spaces coincide ($X = Y$) and there exists an “idle” decision $d^* \in D$ such that $\mathbb{C}_{(\cdot,d^*)} = \mathbb{O}$.. However, in general we don’t require that $\mathbb{O}$ and $\mathbb{C}$ are related in this manner. This assumption will be revisited in

A section I haven’t written yet

.

1.2.1 D-causation

While we take D to be a primitive element of causal decision problems, and therefore a primitive of See-Do Models. Causes are not primitive, but we can offer a secondary notion of causation. We call this D-causation to stress the fact that it arises in a theory of causal inference in which the set D of available decisions is primitive. A similar idea is discussed extensively in Heckerman and Shachter (1995). The main differences are that what we call “consequence maps” map decisions to probability distributions over possible consequences while Heckerman and Shachter work with “states” that map decisions deterministically to consequences. In addition, while we define D-causation relative to a particular consequence map $\mathbb{C}_\theta$, Heckerman and Shachter define it with respect to a *set* of states.

Section **??** explores the difficulty of defining “objective causation” without reference to a set of basic decisions, acts or operations. D need not be interpreted as the set of decisions an agent may make, but whatever interpretation it is assigned, all existing examples of causal models seem to require a “domain set”.

See Section 0.1.4 for the definition of random variables.

Add definition of conditional independence, revise wire label definitions

One way to motivate the notion of D-causation is to observe that for many decision problems, the full set D may be extremely large. Suppose I aim to have my light switched on, and there is a switch that controls the light. Often, the relevant choice of acts for such a problem would appear to be $D_0 = \{\text{flip the switch}, \text{don't flip the switch}\}$. However, in principle I have a much larger range of options to choose from. 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 they have the same ultimate impact on the position of the light switch. D_0 is a quotient over D_1 under this equivalence relation.

If I hypothesize that, relative to D_1, the ultimate state of the light switch is all that matters to determine the ultimate state of the light, I can say that the light switch D_1-causes the state of the light. Given this D_1-causation, the D_1 decision problem can (subject to my hypothesis) be reduced to a D_0 decision between states of the light switch.

If I consider an even larger set of possible acts D_2, I might not accept the hypothesis of D_2-causation. Let D_2 be the following acts:

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

In this case, it would be unreasonable to hypothesize 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:

Definition 1.2.6 (D-causation)**.** Given a consequence map $\mathbb{C}_\theta : D \to \Delta(\mathcal{Y})$, random variables $\mathsf{Y}_1 : Y \times D \to Y_1$, $\mathsf{Y}_2 : Y \times D \to Y_2$ and domain variable $\mathsf{D} : Y \times D \to D$ (Definition 0.1.8), Y_1 D-causes Y_2 iff $\mathsf{Y}_2 \perp\!\!\!\perp_{\mathbb{C}_\theta} \mathsf{D}|\mathsf{Y}_1$.

1.2.2 D-causation vs Heckerman and Shachter

Heckerman and Shachter study deterministic “consequence maps”. Furthermore, what we call hypotheses $\theta \in \Theta$, Heckerman and Schachter call states $s \in S$. One could consider a state to be a hypothesis that is specific enough to yield a deterministic map from decisions to outcomes. 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 1.2.7 (Limited unresponsiveness)**.** Given states S, deterministic consequence maps $\mathbb{C}_s : D \to \Delta(F)$ for each $s \in A$ and a random variables $\mathsf{X} : F \to X$, $\mathsf{Y} : F \to Y$, Y is unresponsive to D in states limited by X if $\mathbb{C}^{\mathsf{X}|\mathsf{D}}_{(s,d)} = \mathbb{C}^{\mathsf{X}|\mathsf{DS}}_{(s,d')} \implies \mathbb{C}^{\mathsf{Y}|\mathsf{DS}}_{(s,d)} = \mathbb{C}^{\mathsf{Y}|\mathsf{DS}}_{(s,d')}$ for all $d, d' \in D$, $s \in S$. Write $\mathsf{Y} \not\hookrightarrow_{\mathsf{X}} \mathsf{D}$

Lemma 1.2.8 (Limited unresponsiveness implies D-causation)**.** *For deterministic consequence maps,* $\mathsf{Y} \not\hookrightarrow_{\mathsf{X}} \mathsf{D}$ *implies* X *D-causes* Y *in every state* $s \in S$.

Proof. By the assumption of determinism, for each $s \in S$ and $d \in D$ there exists $x(s, d)$ and $y(s, d)$ such that $\mathbb{C}^{\mathsf{XY}|\mathsf{DS}}_{d,s} = \delta_{x(s,d)} \otimes \delta_{y(s,d)}$.

By the assumption of limited unresponsiveness, for all d, d' such that $x(s, d) = x(s, d')$, $y(s, d) = y(s, d')$ also. Define $f : X \times S \to Y$ by $(s, x) \mapsto y(s, [x(s, \cdot)]^{-1}(x(s, d)))$ where $[x(s, \cdot)]^{-1}(a)$ is an arbitrary element of $\{d|x(s, d) = a\}$. For all s, d,

$f(x(s,d),s) = y(s,d)$. Define $\mathbb{M} : X \times D \times S \to \Delta(\mathcal{Y})$ by $(x,d,s) \mapsto \delta_{f(x,s)}$. $\mathbb{M}$ is a version of $\mathbb{C}^{\mathsf{Y}|\mathsf{X},\mathsf{D},\mathsf{S}}$ because, for all $A \in \mathcal{X}$, $B \in \mathcal{Y}$, $s \in S$, $d \in D$:

$$\mathbb{C}^{\mathsf{X}|\mathsf{DS}}_{(d,s)} \curlyvee (\mathbb{M} \otimes \mathrm{Id}) = \int_A \mathbb{M}(x', d, s; B) d\delta_{x(s,d)}(x') \tag{1.17}$$

$$= \int_A \delta_{f(x',s)}(B) d\delta_{x(s,d)}(x') \tag{1.18}$$

$$= \delta_{f(x(s,d),s)}(B)\delta_{x(s,d)}(A) \tag{1.19}$$

$$= \delta_{y(s,d)}(B)\delta_{x(s,d)}(A) \tag{1.20}$$

$$= \delta_{x(s,d)} \otimes \delta_{y(s,d)}(A \times B) \tag{1.21}$$

$\mathbb{M}$ is also independent of D, given the obvious labeling of inputs. Therefore $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{C}_s} \mathsf{D}|\mathsf{X}$. □

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

$$\mathbb{C}^{\mathsf{LF}|\mathsf{DS}}_{d,[i,m]} = \delta_{(d \text{ XOR } i) \text{ AND } m} \otimes \delta_{(d \text{ XOR } i) \text{ AND } m} \tag{1.22}$$

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

$$\mathbb{C}_\gamma = \frac{1}{4}\mathbb{C}_{\cdot,[0,0]} + \frac{1}{4}\mathbb{C}_{\cdot,[0,1]} + \frac{1}{2}\mathbb{C}_{\cdot,[1,1]} \tag{1.23}$$

$$\mathbb{C}^{\mathsf{FL}|\mathsf{D}}_\gamma = \frac{1}{4}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \otimes \begin{bmatrix}1 & 0\\ 1 & 0\end{bmatrix} + \frac{1}{4}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \otimes \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} + \frac{1}{2}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} \otimes \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} \tag{1.24}$$

Then

$$[1,0]\mathbb{C}^{\mathsf{FL}|\mathsf{D}}_\gamma = \frac{1}{4}[0,1] \otimes [1,0] + \frac{1}{4}[0,1] \otimes [0,1] + \frac{1}{2}[1,0] \otimes [1,0] \tag{1.25}$$

$$[1,0] \curlyvee (\mathbb{C}^{\mathsf{F}|\mathsf{D}}_\gamma \otimes \mathbb{C}^{\mathsf{L}|\mathsf{D}}_\gamma) = (\frac{1}{2}[0,1] + \frac{1}{2}[1,0]) \otimes (\frac{1}{4}[0,1] + \frac{3}{4}[1,0]) \tag{1.26}$$

$$\implies [1,0]\mathbb{C}^{\mathsf{FL}|\mathsf{D}}_\gamma \neq [1,0] \curlyvee (\mathbb{C}^{\mathsf{F}|\mathsf{D}}_\gamma \otimes \mathbb{C}^{\mathsf{L}|\mathsf{D}}_\gamma) \tag{1.27}$$

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 $\theta \in S$, 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 $\theta \in S$, and in this way we differ substantially from Heckerman and Shachter (1995).

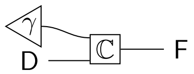
Instead, we simply extend the definition of D-causation to mixtures of hypotheses: if $\gamma \in \Delta(\Theta)$ is a mixture of hypotheses, define $\mathbb{C}_\gamma := (\gamma \otimes \mathbf{Id})\mathbb{C}$. Then X D-causes Y relative to γ iff $\mathsf{Y} \perp\!\!\!\perp_{\mathbb{C}_\gamma} \mathsf{D}|\mathsf{X}$.

Theorem 1.2.9 shows that under some conditions, D-causation can hold for arbitrary mixtures over subsets of the hypothesis class Θ.

Theorem 1.2.9 (Universal D-causation)**.** *If* $\mathbb{C}_\theta^{\mathsf{X}|\mathsf{D}} = \mathbb{C}_{\theta'}^{\mathsf{X}|\mathsf{D}}$ *for all* $\theta, \theta' \in S \subset \Theta$ *and* X *D-causes* Y *in all* $\theta \in S$*, then* X *D-causes* Y *with respect to all mixed consequence maps* $\mathbb{C}_\gamma$ *for all* $\gamma \in \Delta(\Theta)$ *with* $\gamma(S) = 1$*.*

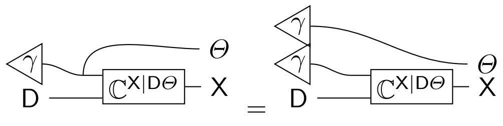
Proof. For $\gamma \in \Delta(\Theta)$, define the mixture

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



(1.28)

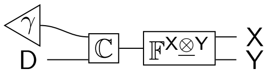
Because $\mathbb{C}_\theta^{\mathsf{X}|\mathsf{D}} = \mathbb{C}_{\theta'}^{\mathsf{X}|\mathsf{D}}$ for all $\theta, \theta' \in \Theta$, we have



(1.29)

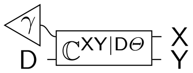
Also

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



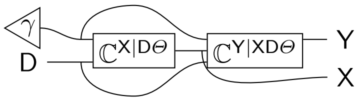
(1.30)

$=$



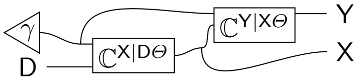
(1.31)

$=$



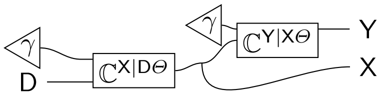
(1.32)

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



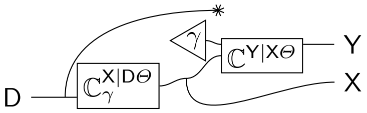
(1.33)

$\overset{1.29}{=}$



(1.34)

$= \overset{1.29}{=}$



(1.35)

Equation 1.35 establishes that $(\gamma \otimes \mathbf{Id}_X \otimes *_D)\mathbb{C}^{\mathsf{Y|X}\Theta}$ 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:

$$\begin{aligned}
\Theta \perp\!\!\!\perp \mathsf{D} \wedge \Theta \perp\!\!\!\perp \mathsf{X|D} &\implies \Theta \perp\!\!\!\perp \mathsf{XD} && (1.36)\\
&\implies \Theta \perp\!\!\!\perp \mathsf{D|X} && (1.37)\\
\mathsf{D} \perp\!\!\!\perp \Theta\mathsf{|X} \wedge \mathsf{D} \perp\!\!\!\perp \mathsf{Y|X}\Theta &\implies \mathsf{D} \perp\!\!\!\perp \mathsf{Y|X} && (1.38)\\
&\implies \mathsf{Y} \perp\!\!\!\perp \mathsf{D|X} && (1.39)
\end{aligned}$$

□

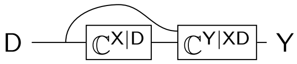
1.2.3 Properties of D-causation

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

$$\mathbb{C}_\theta^{\mathsf{X|D}} = \mathsf{D} - \boxed{\mathbb{C}^{\mathsf{X|D}}} - \boxed{\mathbb{C}^{\mathsf{Y|X}}} - \mathsf{Y} \qquad (1.40)$$

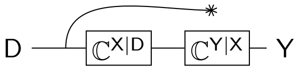
This follows from version (2) of Definition 0.1.28:

$\mathbb{C}_\theta^{\mathsf{X|D}} =$



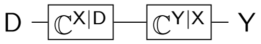
(1.41)

$=$



(1.42)

$=$



(1.43)

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

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. 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.

Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210:2–21, August 2018. ISSN 0277-9536. doi: 10.1016/j.socscimed.2017.12.005. URL `http://www.sciencedirect.com/science/article/pii/S0277953617307359`.

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.

James J. Heckman. Randomization and Social Policy Evaluation. SSRN Scholarly Paper ID 995151, Social Science Research Network, Rochester, NY, July 1991. URL https://papers.ssrn.com/abstract=995151.

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.

Alan Hájek. Interpretations of Probability. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019. URL https://plato.stanford.edu/archives/fall2019/entries/probability-interpret/.

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

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

David K Lewis. Causation. *Journal of Philosophy*, 1986.

Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* Cambridge University Press, New York, NY, 2 edition edition, November 2014. ISBN 978-1-107-69416-3.

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. Challenging the hegemony of randomized controlled trials: A commentary on Deaton and Cartwright. *Social Science & Medicine*, 2018.

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`.

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`.

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: