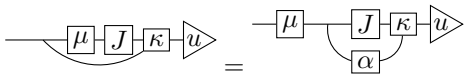
Sept 19: Identification and Estimation

1 The story at a high level

Take a causal theory $\mathcal{T}$ where we label each pair $\theta := (\kappa_\theta, \mu_\theta) \in \mathcal{T}$. Define the kernels $\kappa : \mathcal{T} \times D \to \mathcal{E}$ and $\mu : \mathcal{T} \to \mathcal{E}$ by $\kappa : (\theta, y; A) \mapsto \kappa_\theta(y; A)$ and $\mu : (\theta; A) \mapsto \mu_\theta(A)$.

Optmizibility: I make the claim (unproven) that it is possible to find a "universally optimal" decision function if the following identity holds for all decision functions $J : E \to \Delta(\mathcal{D})$:

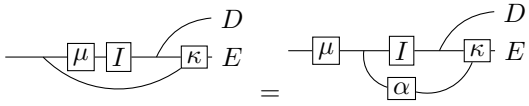


(1)

I have a proof that this is so in my notebook, but I still have to write it up and check it.

If the forward direction holds, the reverse direction does not hold - we can take a problem that respects 1 and introduce additional dominated decisions that break 1 without breaking the "universal optimizability" (i.e. decisions we know to be very bad, but exactly how bad depends on the state in a difficult-to-identify manner). Again in my notebook, I have an argument that if we only require 1 for any complete class of decision functions, then this requirement is necessary.

Necessary and Sufficient conditions for optimizibility: Setting aside the utility for now, if equation 1 holds for all decision functions dominated by I (that is, all J such that for all $x \in E$, $J(x; \cdot) \ll I(x; \cdot)$), this is equivalent to the following:

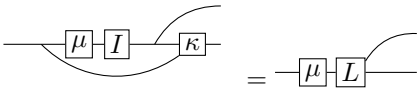


(2)

The extra copy map allows us to extend from equality "for I" to equality "for all kernels dominated by I".

We ignore the utility as introducing it moves us out of the world of Markov kernels; need to show that everything still works if we reintroduce it

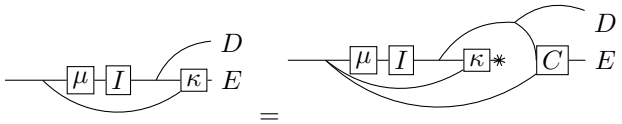
Immediate from Eq. 2 is the fact that there is some kernel L such that:



(3)

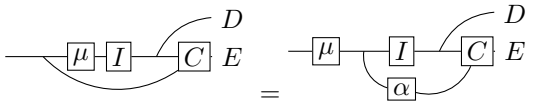
In addition, if we define $C : \mathcal{T} \times D \to \Delta(\mathcal{E})$ as a *generalised disintegration* of μL - i.e. a kernel with the following property:

Preprint. Under review.



(4)

Then we can see by the fact that $C = \kappa$ "A.S." that

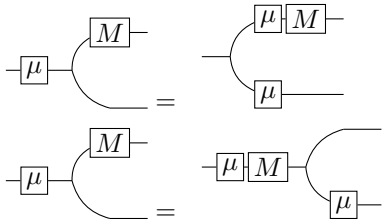


(5)

The reverse direction also holds: 3 and 5 imply 2. This decomposition is handy, because we can consider 3 to express "perfect identifiability of κ" and 5 to express "perfect estimability of μ", the former being a quintessentially "causal" notion and the latter being a "statistical" notion (note that this distinction is only based on *typical use* of the words causal and statistical).

The following assumptions are sufficient (but not necessary) for 5:

$\exists M$:



(6)

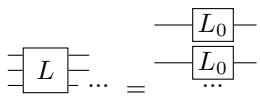
(7)

Sufficient for 6 is the condition that there is some function T with associated kernel F_{T} such that μF_{T} is deterministic and for 7 we require that T is sufficient for $\{\mu_\theta\}$. An example of this is where T is the mean of an infinite sequence of IID Bernoulli variables and μF_{T} is then deterministic via the strong law of large numbers.

Equation 7 is not necessary for 1, as observations may be "too informative". For example, if $\mathcal{T}$ contains many different μ_θ but only one κ_θ, then we can always perform 1 while we do not generally have 7.

1.1 IID variables and factorisibility

Equation 3 is quite general in the sense that, if μ maps from $\mathcal{T}$ to an infinite sequence of IID variables, we may satisfy it with an L that first estimates μ_θ from the sequence of variables and subsequently chooses an appropriate measure on $D \times E$. A particularly interesting case is where μ maps to an infinite sequence of IID variables and L factorises as follows:

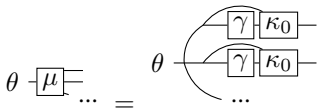


(8)

A special case of this is considered below. We can informally think of this case as "correlation *is* causation", as from *each* observed RV we can get an input-output pair via L_0. This is a generalisation of the usual case of "correlation is causation" as we allow the possibility that the output is randomised from observations rather than insisting it be distributed exactly as observations were. The possibility of post randomisation is helpful, for example, if the sequence of "result" variables we expect is of a different length to our sequence of observations.

1.2 Informal overview of sufficient conditions

First, we assume that the state-observations map μ sends a state to an infinite IID sequence generated by the one-shot state-consequence map κ_0 and some state-decision map γ:

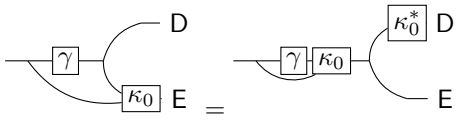


(9)

Infinite copy maps (indicated by ellipsis) are defined for distributions via the Kolmogorov extension theorem. I don't know if they always exist for kernels.

I think this assumption has a connection with the De Finetti representation theorem

Second we assume κ_0 is *globally invertible*. That is, there is a kernel $\kappa_0^* : E \to \Delta(\mathcal{D})$ such that



(10)

In the general, non-globally-invertible case we'd need a string from the "state" wire to κ_0^*.

Under these two conditions plus full support plus sufficient regularity for the strong law of large numbers, we have **??**, 7 and **??**.

I think a CBN is a causal theory where the consequence map is a decision randomised version of κ_0 in Eq. 9 (i.e. the true consequence $\kappa = M\kappa_0$) and these other conditions hold, and is therefore dominated by a theory of the form above. Global invertibility is related to variable setting/hard interventions, and I think it's also related to the "No Causes in No Causes out" theorems 4.9 and 4.10.

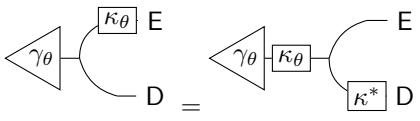
2 Recoverability

A natural assumption suggested by the notion of a CSDP is that of *recoverability* - that a causal theory $\mathcal{T} : E \times D \nrightarrow E$ permits some decision function that reproduces the distribution of the observed data. That is, we assume that for every $(\kappa_\theta, \mu_\theta) := \theta \in \mathcal{T}$ there exists $\gamma_\theta \in \Delta(\mathcal{D})$ such that

$$\gamma_\theta \kappa_\theta = \mu_\theta \tag{11}$$

"Traditional" causal inference doesn't have a strict equivalent of this assumption, though it corresponds roughly to the "easy" cases (for example, it is satisfied by a CBN where there are no backdoor paths between the "intervened" variable and the "target" variable). One reason I think it's interesting is that *randomised recoverability* may be quite a general assumption - that is, there is "in principle" a stochastic decision that recovers the observed distribution, but we are practically limited to taking mixed decisions that cannot necessarily accomplish this.

Suppose also that we have some κ^* that, for all $\theta \in \mathcal{T}$, is a Bayesian inversion of γ_θ and κ_θ; that is:



(12)

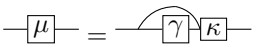
A sufficient condition for the existence of such a κ^* is the assumption that decisions correspond to *variable setting* - that is, there is some variable $\mathsf{X} : E \to X$ such that for all $a \in D$, $\theta \in \mathcal{T}$ we have $\delta_a \kappa_\theta F_\mathsf{X} = \delta_a$ (such an assumption arises in graphical models as hard interventions, and in potential outcomes as "potential-outcome identifiers"). Indeed F_X is in this case a candidate for κ^*. It is not

necessary that κ^* be deterministic, however - suppose every κ ignores D. Then choose $\gamma_\theta = \gamma$ for arbitrary $\gamma \in \Delta(\mathcal{D})$ and it can be verified that $\kappa^* : b \mapsto \gamma$ satisfies 12.

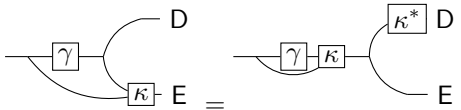
I believe a weaker sufficient condition for the existence of a universal κ^* is that every κ_θ factorises as $\kappa_\theta = h \curlyvee (\mathrm{Id}_F \otimes j_\theta)$ for some fixed $h : D \to \Delta(\mathcal{F})$, but I have not yet shown this.

We will proceed somewhat rashly: suppose that by defining $\gamma : \mathcal{T} \to \Delta(\mathcal{D})$, $\mu : \mathcal{T} \to \Delta(\mathcal{E})$ and $\kappa : \mathcal{T} \times D \to \Delta(\mathcal{E}$ by $\gamma : \theta \to \gamma_\theta$, $\mu : \theta \to \mu_\theta$ and $\kappa : (\theta, d) \to \kappa_\theta(d; \cdot)$ that all resulting objects are Markov kernels, and that $\mathcal{T}$ is a standard measurable space.

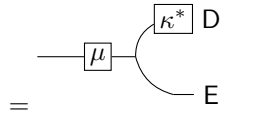
By previous assumptions, we have the following properties:



(13)

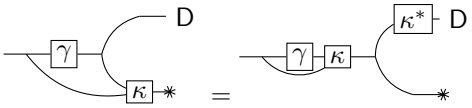


(14)

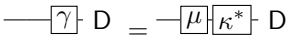


(15)

From 14 we also have



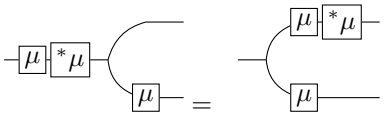
(16)



(17)

Where 17 follows from 11.

The following assumption is a formalisation of the notion that "we can determine μ precisely from observation" (alternatively, that we can find an optimal decision for a classical statistical decision problem). Suppose that μ is characterised by some kernel ${}^*\mu$. That is,



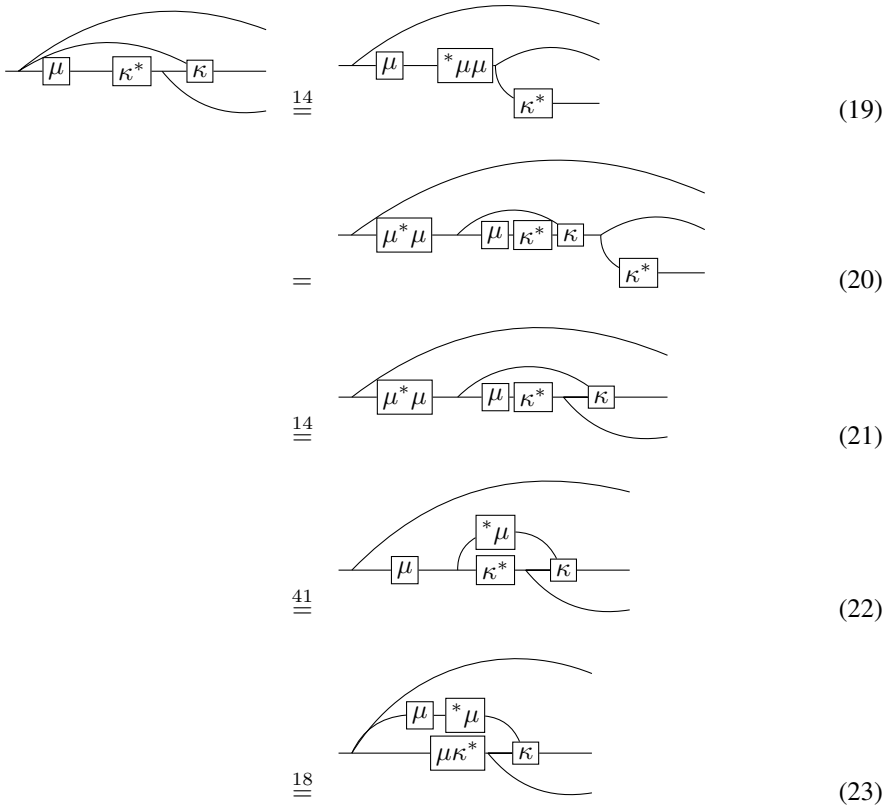
(18)

An equivalent condition to 18 is that for all $\theta, \theta' \in \mathcal{T}$, $A \in \mathcal{E}$, we have $\mu(\theta; A) = \mu(\theta'; A)$, $\mu^*\mu(\theta; \cdot)$-almost surely. More informally,the support of $\mu^*\mu$ for each input θ divides $\mathcal{T}$ into equivalence classes such that for all θ in a given equivalence class, μ maps to the same probability measure on $\mathcal{E}$.

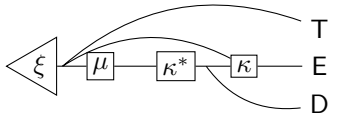
Note that as a result of 18 we also have $\mu^*\mu\mu = \mu$. This weaker condition is not sufficient for the following result.

There is a connection between equation 18 and the notion of a sufficient statistic

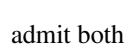
We then have



Equation 23 implies that, given any $\xi \in \Delta(\mathcal{T})$, all distributions of the form



(24)

admit both $\kappa :=$ and $\kappa_{\text{fac}} :=$  as disintegrations from $(\mathsf{D}, \mathsf{T}) \dashrightarrow \mathsf{E}$. Therefore these κ and κ_{fac} agree almost surely with respect to the distribution 24 for any prior ξ.

However, also by assumption 18, we have that for $\theta, \theta' \in \mathcal{T}$ either $\mu(\theta; A) = \mu(\theta'; A)$ for all $A \in \mathcal{E}$, or for any $A \in \mathcal{E}$ $\mu(\theta; A) = 0$ or $\mu(\theta'; A) = 0$. That is, any two states either have the same probability measure or probability measures with disjoint support. This is problematic, as the distribution 24 then has no support over much of the space $D \times E \times \mathcal{T}$. If μ were deterministic, for example, and hence associated with some function f, while 18 would be guaranteed via a left inverse, 24 would be supported on a subset of $D \times \{(\theta, f(\theta)) | \theta \in \mathcal{T}\}$. In particular, we have no guarantee that the desired equality of κ and κ_{fac} holds if we take any decision that doesn't reproduce the observed distribution. This isn't totally trivial: we may live in a world where most actions make things worse, in which case knowing how to keep things the same is valuable.

A stronger result can be found if we assume we have an infinite sequence of RVs $\mathsf{X}_i : E \to W$ and $\mathsf{D}_i : D \to V$ such that

- $W^{\mathbb{N}} = E$, $V^{\mathbb{N}} = D$ (i.e. the sequence of all X_i's is identified with E and the sequence of all D_i's is identified with D)
- $\mu = \curlyvee \otimes_{i \in \mathbb{N}} \mu F_{\mathsf{X}_i}$ (the X_i's are "IID conditional on θ")
- There exists κ_0 such that $\kappa = \curlyvee \otimes_{i \in \mathbb{N}} (F_{\mathsf{D}_i} \otimes \mathrm{Id}_{\mathcal{T}}) \kappa_0 F_{\mathsf{X}_i}$ (κ is "IID conditional on D, θ")

this might be closely related to exchangeability via de Finetti?

Here we define the "infinite copy map" $\curlyvee \otimes_{i\in\mathbb{N}} \mu F_{\mathsf{X}_i}$ to denote the kernel $\theta \mapsto \nu_\theta$ where ν_θ the unique distribution such that for all finite $A \subset \mathbb{N}$ and projections $\pi_A : E \to \Delta(W^{|A|})$, $\nu_\theta \pi_A = \otimes_{i\in A} \mu_\theta F_{\mathsf{X}_i}$. This distribution is unique via the Kolmogorov extension theorem (the symmetry of the copy map guarantees the required consistency conditions) [Tao, 2011].

I assume, for now, that measurability can be worked out in some cases; in particular, that there is a σ-algebra on infinite sequences that renders the above kernel measurable in the appropriate way.

Lemma 2.1 ("IID" kernels agree on truncations)**.** *For finite $A \subset \mathbb{N}$, $y, y' \in D$, if $\otimes_{i\in A} \mathsf{X}_i(y) = \otimes_{i\in A} \mathsf{X}_i(y')$ and $\kappa : \mathcal{T} \times D \to \Delta(\mathcal{E})$ is "IID" in the sense above then for all $\theta \in \mathcal{T}$, $B \in \mathcal{W}^{|A|}$, $\kappa(\theta, y; B)\pi_A = \kappa(\theta, y'; B)\pi_A$.*

Proof. By definition, we have

$$\kappa\pi_A(\theta, y; B) = \otimes_{i\in A} \kappa F_{\mathsf{X}_i}(\theta, \mathsf{D}_i(y); B) \tag{25}$$
$$= \otimes_{i\in A} \kappa F_{\mathsf{X}_i}(\theta, \mathsf{D}_i(y'); B) \tag{26}$$
$$= \kappa\pi_A(\theta, y'; B) \tag{27}$$

□

Suppose both X_i and D_i are binary, and that for each $\theta \in \mathcal{T}$ we have recoverability (Eq. 11) with $\mu_\theta = \gamma_\theta$ (we will conclude that X is "directly controlled" by D, but we will not assume this at the outset). κ^* is therefore trivial. For each θ, X_i are IID Bernoulli variables and so each μ_θ is characterised by a single parameter p; let p_θ be the value of this parameter for some given θ. Define $\overline{\mathsf{X}} := \lim_{n\to\infty} \frac{1}{m} \sum_{i\in[n]} \mathsf{X}_i$ and ${}^*\mu$ to be any kernel $E \to \Delta(\mathcal{T})$ such that the support of ${}^*\mu(x; \cdot)$ is a subset of $\{\theta | p_\theta = \overline{\mathsf{X}}(x)\}$. Note that for any $\theta, \theta' \in \mathcal{T}$ we have either $p_\theta = p_{\theta'}$ and so $\mu(\theta; A) = \mu(\theta'; A)$ for all A or θ' is not in the support of $\mu^*\mu(\theta; \cdot)$. Thus we have 18, and hence "almost sure" equality of κ and κ_{fac}.

However with the exception of states where $p_\theta = 0$ or 1, almost sure equality is enough for $\kappa_{\text{fac}}\pi_A(\theta, y; B) = \kappa\pi_A(\theta, y; B)$ for all $y \in D$, finite $A \subset \mathbb{N}$ and $B \in \mathcal{W}^{|A|}$. Then by the Kolmogorov extension theorem, we also have $\kappa_{\text{fac}}(\theta, y; B) = \kappa(\theta, y; B)$ for all $y \in D$ and "almost all" $\theta \in \mathcal{T}$.

This appears to have similarities to the general case where we are trying to identify a particular function from some set of possible functions and we know the output of that function for a subset of inputs. It still comes down to a question of whether or not the set of functions in question is small enough to be fully characterised by the set of inputs we're allowed to see.

3 Notes on category theoretic probability and string diagrams

Category theoretic treatments of probability theory often start with *probability monads* (for a good overview, see [Jacobs, 2018]). A monad on some category C is a functor $T : C \to C$ along with natural transformations called the unit $\eta : 1_C \to T$ and multiplication $\mu : T^2 \to T$. Roughly, functors are maps between categories that preserve identity and composition structure and natural transformations are "maps" between functors that also preserve composition structure. The monad unit is similar to the identity element of a monoid in that application of the identity followed by multiplication yields the identity transformation. The multiplication transformation is also (roughly speaking) associative.

An example of a probability monad is the discrete probability monad given by the functor $\mathcal{D} : \mathbf{Set} \to \mathbf{Set}$ which maps a countable set X to the set of functions from $X \to [0, 1]$ that are probability measures on X, denoted $\mathcal{D}(X)$. $\mathcal{D}$ maps a measurable function f to $\mathcal{D}f : X \to \mathcal{D}(X)$ given by $\mathcal{D}f : x \mapsto \delta_{f(x)}$. The unit of this monad is the map $\eta_X : X \to \mathcal{D}(X)$ given by $\eta_X : x \mapsto \delta_x$ (which is equivalent to $\mathcal{D}1_X$) and multiplication is $\mu_X : \mathcal{D}^2(X) \to \mathcal{D}(X)$ where $\mu_X : \Omega \mapsto \sum_\phi \Omega(\phi)\phi$.

For continuous distributions we have the Giry monad on the category **Meas** of mesurable spaces given by the functor $\mathcal{G}$ which maps a measurable space X to the set of probability measures on X,

denoted $\mathcal{G}(X)$. Other elements of the monad (unit, multiplication and map between morphisms) are the "continuous" version of the above.

Of particular interest is the Kleisli category of the monads above. The Kleisli C_T category of a monad T on category C is the category with the same objects and the morphisms $X \to Y$ in C_T is the set of morphisms $X \to TY$ in C. Thus the morphisms $X \to Y$ in the Kleisli category $\mathbf{Set}_{\mathcal{D}}$ are morphisms $X \to \mathcal{D}(Y)$ in **Set**, i.e. stochastic matrices, and in the Kleisli category $\mathbf{Meas}_{\mathcal{G}}$ we have Markov kernels. Composition of arrows in the Kleisli categories correspond to Matrix products and "kernel products" respectively.

Both $\mathcal{D}$ and $\mathcal{G}$ are known to be *commutative* monads, and the Kleisli category of a commutative monad is a symmetric monoidal category.

Diagrams for symmetric monoidal categories consist of wires with arrows, boxes and a couple of special symbols. The identity object (which we identify with the set $\{*\}$) is drawn as nothing at all $\{*\} :=$ and identity maps are drawn as bare wires:

$$\mathrm{Id}_X := \quad \overset{X}{\uparrow} \tag{28}$$

We draw Kleisli arrows from the unit (i.e. probability distributions) $\mu : \{*\} \to X$ as triangles and Kleisli arrows $\kappa : X \to Y$ (i.e. Markov kernels $X \to \Delta(\mathcal{Y})$) as boxes. We draw the Kleisli arrow $\mathbb{1}_X : X \to \{*\}$ (which is unique for each X) as below

$$\mu := \quad \overset{X}{\underset{\mu}{\triangleleft}} \qquad\qquad \kappa := \quad \overset{Y}{\boxed{\kappa}} \tag{29}$$

The product of objects in **Meas** is given by $(X, \mathcal{X}) \cdot (Y, \mathcal{Y}) = (X \times Y, \mathcal{X} \otimes \mathcal{Y})$, which we will often write as just $X \times Y$. Horizontal juxtaposition of wires indicates this product, and horizontal juxtaposition also indicates the tensor product of Kleisli arrows. Let $\kappa_1 : X \to W$ and $\kappa_2 : Y \to Z$:

$$(X \times Y, \mathcal{X} \otimes \mathcal{Y}) := \quad \overset{X}{\uparrow} \; \overset{Y}{\uparrow} \qquad\qquad \kappa_1 \otimes \kappa_2 := \quad \begin{array}{cc} \overset{W}{\uparrow} & \overset{Z}{\uparrow} \\ \boxed{\kappa_1} & \boxed{\kappa_2} \\ X & Y \end{array} \tag{30}$$

Composition of arrows is achieved by "wiring" boxes together. For $\kappa_1 : X \to Y$ and $\kappa_2 : Y \to Z$ we have

$$\kappa_1\kappa_2(x; A) = \int_Y \kappa_2(y; A)\kappa_1(x; dy) := \quad \begin{array}{c} \overset{Z}{\uparrow} \\ \boxed{\kappa_2} \\ \boxed{\kappa_1} \\ X \end{array} \tag{31}$$

Symmetric monoidal categoris have the following coherence theorem[Selinger, 2010]:

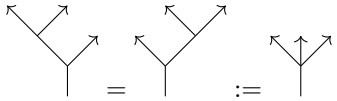
Theorem 3.1 (Coherence (symmetric monoidal))**.** *A well-formed equation between morphisms in the language of symmetric monoidal categories follows from the axioms of symmetric monoidal categories ifand only if it holds, up to isomorphism of diagrams, in the graphical language.*

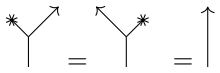
Isomorphism of diagrams for symmetric monoidal categories (somewhat informally) is any planar deformation of a diagram including deformations that cause wires to cross. We consider a diagram for a symmetric monoidal category to be well formed only if all wires point upwards.

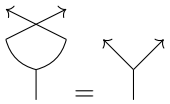
In fact the Kleisli categories of the probability monads above have (for each object) unique *copy*: $X \to X \times X$ and *erase*: $X \to \{*\}$ maps that satisfy the *commutative comonoid axioms* that (thanks to the coherence theorem above) can be stated graphically. These differ from the copy and erase

maps of *finite product* or *cartesian* categories in that they do not necessarily respect composition of morphisms.

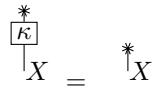
$$\text{Erase} = \mathbb{1}_X :=$$  $$\text{Copy} = x \mapsto \delta_{x,x} :=$$  (32)

 (33)

 (34)

 (35)

Finally, $\{*\}$ is a terminal object in the Kleisli categories of either probability monad. This means that the map $X \to \{*\}$ is unique for all objects X, and as a consequence for all objects X, Y and all $\kappa : X \to Y$ we have

 (36)

This is equivalent to requiring for all $x \in X$ $\int_Y \kappa(x; dy) = 1$. In the case of $\mathbf{Set}_{\mathcal{D}}$, this condition is what differentiates a stochastic matrix from a general positive matrix (which live in a larger category than $\mathbf{Set}_{\mathcal{D}}$).

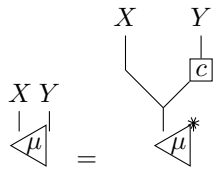
Thus when manipulating diagrams representing Markov kernels in particular (and, importantly, not more general symmetric monoidal categories) diagram isomorphism also includes applications of 33, 34, 35 and 36.

A particular property of the copy map in $\mathbf{Meas}_{\mathcal{G}}$ (and probably $\mathbf{Set}_{\mathcal{D}}$ as well) is that it commutes with Markov kernels iff the markov kernels are deterministic [Fong, 2013].

3.1 Disintegration and Bayesian inversion

Disintegration is a key operation on probability distributions (equivalently arrows $\{*\} \to X$) in the categories under discussion. It corresponds to "finding the conditional probability" (though conditional probability is usually formalised in a slightly different way).

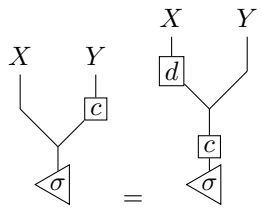
Given a distribution $\mu : \{*\} \to X \otimes Y$, a disintegration $c : X \to Y$ is a Markov kernel that satisfies

 (37)

Disintegrations always exist in $\mathbf{Set}_{\mathcal{D}}$ but not in $\mathbf{Meas}_{\mathcal{G}}$. The do exist in the latter if we restrict ourselves to standard measurable spaces. If c_1 and c_2 are disintegrations $X \to Y$ of μ, they are equal

μ-A.S. In fact, this equality can be strengthened somewhat - they are equal almost surely with respect to any distribution that shares the "X-marginal" of μ.

Given $\sigma : \{*\} \to X$ and a channel $c : X \to Y$, a Bayesian inversion of (σ, c) is a channel $d : Y \to X$ such that



(38)

We can obtain disintegrations from Bayesian inversions and vise-versa.

Clerc et al. [2017] offer an alternative view of Bayesian inversion which they claim doesn't depend on standard measurability conditions, but there is a step in their proof I didn't follow.

3.2 Generalisations

Cho and Jacobs [2019] make use of a larger "CD" category by dropping 36. I'm not completely clear whether you end up with arrows being "Markov kernels for general measures" or something else (can we have negative arrows?). This allows for the introduction of "observables" or "effects" of the form

.

Jacobs et al. [2019] make use of an embedding of $\mathbf{Set}_{\mathcal{D}}$ in $\mathbf{Mat}(\mathbb{R}^+)$ with morphisms all positive matrices (I'm not totally clear on the objects, or how they are self-dual - this doesn't seem to be exactly the same as the category of finite dimensional vector spaces). This latter category is compact closed, which - informally speaking - supports the same diagrams as symmetric monoidal categories with the addition of "upside down" wires.

3.3 Key questions for Causal Theories

We will first define *labeled diagrams*. Rather than labelling the wires of our diagrams with *spaces* (as is typical [Selinger, 2010]), we assign a unique label to each "wire segment" (with some qualifications). That is, we assign a unique label to each bare wire in the diagram with the following additonal qualifications:

- If we have a box in the diagram representing the identity map, the incoming and outgoing wires are given the same label
- If we have a wire crossing in the diagram, the diagonally opposite wires are given the same label
- The input wire and the *two* output wires of the copy map are given the same label

Given two diagrams G_1 and G_2 that are isomorphic under transformations licenced by the axioms of symmetric monoidal categories and commutative comonoid axioms, suppose we have a labelling of G_1. We can label G_2 using the following translation rule:

- For each box in G_2, we can identify a corresponding box in G_1 via labels on each box. For each such pair of boxes, we label the incoming wires of the G_2 box with the labels of the G_1 box preserving the left-right order. We do likewise for outgoing wires.

These rules will lead to a unique labelling of G_2 with all wire segments are labelled. We would like for these rules to yield the following:

- For any sequence of diagram isomorphisms beginning with G_1 and ending with G_2, we end up with the same set of labels
- If we label G_2 according to the rules above then relabel G_1 from G_2 according to the same rules we retrieve the original labels of G_1

I'm sure one of the papers I read mentioned labeled diagrams, I just couldn't find it when I looked for it

Since writing this, I found Kissinger [2014] as an example of a diagrammatic system with labeled wires, I will follow it up

We do not prove these properties here, but motivate them via the following considerations:

- These properties obviously hold for the wire segments into and out of boxes
- The only features a diagram may have apart from boxes and wires are wire crossings, copy maps and erase maps
- The labeling rule for wire crossings respects the symmetry of the swap map
- The labeling rule for copy maps respects the symmetry of the copy map and the property described in Equation 35

We will follow the convention whereby "internal" wire labels are omitted from diagrams.

Note also that each wire that terminates in a free end can be associated with a random variable. Suppose for $N \in \mathbb{N}$ we have a kernel $\kappa : A \to \Delta(\times_{i\in N} X_i)$. Define by p_j ($j \in [N]$) the projection map $p_j : \times_{i\in N} X_i \to X_j$ defined by $p_j : (x_0, ..., x_N) \mapsto x_j$. p_j is a measurable function, hence a random variable. Define by π_j the projection kernel $\mathcal{G}(\pi_j)$ (that is, $\pi_j : \mathbf{x} \mapsto \delta_{p_j(\mathbf{x})}$). Note that $\kappa(y; p_j^{-1}(A)) = \int_{X_j} \delta_{p_j(\mathbf{x})}(A)\kappa(y; d\mathbf{x}) = \kappa\pi_j$. Diagrammatically, π_j is the identity map on the j-th wire tensored with the erase map on every other wire. Thus the j-th wire carries the distribution associated with the random variable p_j. We will therefore consider the labels of the "outgoing" wires of a diagram to denote random varaibles (though there are obviously many random variables not represented by such wires). We will additionally distinguish wire labels from spaces by font - wire labels are sans serif $\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{X}, \mathsf{Y}, \mathsf{Z}$ while spaces are serif A, B, C, X, Y, Z.

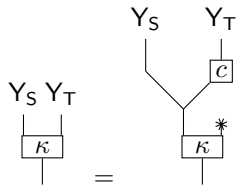
Wire labels appear to have a key advantage over random variables: they allow us to "forget" the sample space as the correct typing is handled automatically by composition and erasure of wires

generalised disintegrations : Of key importance to our work is generalising the notion of disintegration (and possibly Bayesian inversion) to general kernels $X \to Y$ rather than restricting ourselves to probability distributions $\{*\} \to Y$. We will define generalised disintegrations as a straightforward analogy regular disintegrations, but the conditions under which such disintegrations exist are more restrictive than for regular disintegraions.

Definition 3.2 (Label signatures)**.** If a kernel $\kappa : X \to \Delta(Y)$ can be represented by a diagram G with incoming wires $\mathsf{X}_1, ...\mathsf{X}_\mathsf{n}$ and outgoing wires $\mathsf{Y}_1, ..., \mathsf{Y}_\mathsf{m}$, we can assign the kernel a "label signature" $\kappa : \mathsf{X}_1 \otimes ... \otimes \mathsf{X}_\mathsf{n} \dashrightarrow \mathsf{Y}_1 \otimes ... \otimes \mathsf{Y}_m$ or, for short, $\kappa : \mathsf{X}_{[n]} \dashrightarrow \mathsf{Y}_{[m]}$. Note that this signature associates each label with a unique space - the space of X_1 is the space associated with the left-most wire of G and so forth. We will implicitly leverage this correspondence and write with X_1 the space associated with X_1 and so forth. Note that while X_1 is by construction always different from X_2 (or any other label), the space X_1 may coincide with X_2 - the fact that labels always maintain distinctions between wires is the fundamental reason for introducing them in the first place.

There might actually be some sensible way to consider κ to be transforming the measurable functions of a type similar to $\otimes_{i\in[n]}\mathsf{X}_i$ to functions of a type simlar to $\otimes_{i\in[m]}\mathsf{Y}_i$ (or vise versa - perhaps related to Clerc et al. [2017]), but wire labels are all we need at this point

Definition 3.3 (Generalised disintegration)**.** Given a kernel $\kappa : X \to \Delta(Y)$ with label signature $\kappa : \mathsf{X}_{[n]} \dashrightarrow \mathsf{Y}_{[m]}$ and disjoint subsets $S, T \subset [m]$ such that $S \cup T = [m]$, a kernel c is a *g-disintigration from S to T* if it's type is compatible with the label signature $c : \mathsf{Y}_\mathsf{S} \dashrightarrow \mathsf{Y}_\mathsf{T}$ and we have the identity (omitting incoming wire labels):



(39)

I have introduced without definition additional labeling operations here: first, each label has a particular space associated with it (in order to license the notion of "type compatible with label signature"), and we have supposed labels can be "bundled".

In contrast to regular disintegrations, generalised disintegrations "usually" do not exist. Consider $X = \{0,1\}$, $Y = \{0,1\}^2$ and κ has label signature $\mathsf{X}_1 \dashrightarrow \mathsf{Y}_{\{1,2\}}$ with

$$\kappa : \begin{cases} 1 \mapsto \delta_1 \otimes \delta_1 \\ 0 \mapsto \delta_1 \otimes \delta_0 \end{cases} \tag{40}$$

κ imposes contradictory requirements for any disintegration $c : \{0,1\} \to \{0,1\}$ from $\{1\}$ to $\{2\}$: equality for $\mathsf{X}_1 = 1$ requires $c(1;\cdot) = \delta_1$ while equality for $\mathsf{X}_1 = 0$ requires $c(1;\cdot) = \delta_0$. Subject to some regularity conditions (similar to standard Borel conditions for regular disintegrations), we can define g-disintegrations of a canonically related kernel that do generally exist; intuitively, g-disintegrations exist if they take the "input wires" of κ as input wires themselves.

Lemma 3.4. *Given $\kappa : X \to \Delta(Y)$, a kernel $\kappa^\dagger$ is a right inverse iff we have for all $x \in X$, $A \in \mathcal{X}$, $y \in Y$ $\kappa^\dagger(y;A) = \delta_x(A)$, $\kappa(x;\cdot)$-almost surely.*

Proof. Suppose $\kappa^\dagger$ satisfies the almost sure equality for all $x \in X$. Then for all $x \in X$, $A \in \mathcal{X}$ we have $\kappa\kappa^\dagger(x;A) = \int_Y \kappa^\dagger(y;A)\kappa(x;dy) = \int_Y \delta_x(A)\kappa(x;dy) = \delta_x(A)$; that is, $\kappa\kappa^\dagger = \mathrm{Id}_X$, so $\kappa^\dagger$ is a right inverse of κ.

Suppose we have a right inverse $\kappa^\dagger$. By definition, for all $x \in X$ and $A \in \mathcal{X}$ we have $\int_Y \kappa^\dagger(y;A)\kappa(x;dy) = \delta_x(A)$.

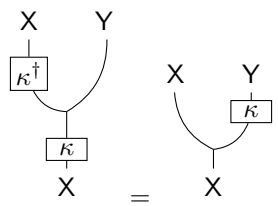
Suppose $x \notin A$ and let $B_\epsilon = \kappa_A^{\dagger -1}((\epsilon,1])$ for some $\epsilon > 0$. We have $\int_Y \kappa^\dagger(y;A)\kappa(x;dy) = 0 \geq \epsilon\kappa(x;B_\epsilon)$. Thus for any $\epsilon > 0$ we have $\kappa(x;B_\epsilon) = 0$. Consider the set $B_0 = \kappa_A^{\dagger -1}((0,1])$. For some sequence $\{\epsilon_i\}_{i\in\mathbb{N}}$ such that $\lim_{i\to\infty}\epsilon_i = 0$ we have $B_0 = \cup_{i\in\mathbb{N}} B_{\epsilon_i}$. By countable additivity, $\kappa(x;B_0) = 0$.

Suppose $x \in A$ and let $B^{1-\epsilon} = \kappa_A^{\dagger -1}([0,1-\epsilon))$. We have $\int_Y \kappa^\dagger(y;A)\kappa(x;dy) = 1 \leq (1-\epsilon)\kappa(x;B^{1-\epsilon}) + 1 - \kappa(x;B^{F.w1-\epsilon}) = 1 - \epsilon\kappa(x;B^{1-\epsilon})$. Thus $\kappa(x;B^{1-epsilon}) = 0$ for $\epsilon > 0$. By an argument analogous to the above, we also have $\kappa(x;B^1) = 0$. Thus the $\kappa(x;\cdot)$ measure of the set on which $\kappa^\dagger(y;A)$ disagrees with $\delta_x(A)$ is $\kappa(x;B_0) + \kappa(x;B^1) = 0$ and hence $\kappa^\dagger(y;A) = \delta_x(A)$ $\kappa(x;\cdot)$-almost surely. □

I haven't shown that any map inverting κ implies the existence of a Markov kernel that does so

I am using countable sets below to get my general argument in order without getting too hung up on measurability; I will try to lift it to standard measurable once it's all there

Lemma 3.5. *Given $\kappa : X \to \Delta(Y)$ and a right inverse $\kappa^\dagger$, we have*



(41)

Proof. Let the diagram on the left hand side be L and the diagram on the right hand side be R.

$$L(x; A \times B) = \int_Y \int_{Y \times Y} \mathrm{Id}_Y \otimes \kappa_S^\dagger(y, y'; A \times B)\delta_{(z,z)}(dy \times dy')\kappa\pi_S(x; dz) \tag{42}$$

$$= \int \mathrm{Id}_Y \otimes \kappa^\dagger(z, z; A \times B)\kappa\pi_S(x; dz) \tag{43}$$

$$= \int \delta_z(A)\kappa_S^\dagger(z; B)\kappa\pi_S(x; dz) \tag{44}$$

$$= \int_A \kappa_S^\dagger(z; B)\kappa\pi_S(x; dz) \tag{45}$$

$$= \delta_x(B)\kappa\pi_S(x; A) \tag{46}$$

Where 46 follows from Lemma 3.4.

$$R(x; A \times B) = \int \delta_{(x,x)}(dy \times dy')\kappa\pi_S \otimes \mathrm{Id}_X(y, y'; A \times B) \tag{47}$$

$$= \kappa\pi_S(x; A)\delta_x(B) \qquad = L \tag{48}$$

□

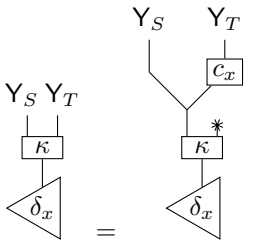
Theorem 3.6. *Given countable X and standard measurable Y, $n, m \in \mathbb{N}$, $S, T \subset [m]$, κ with label signature $\mathsf{X}_{[n]} \dashrightarrow \mathsf{Y}_{[m]}$ a g-disintegration exists from S to T if $\kappa\pi_S$ is right-invertible*

via a Markov kernel

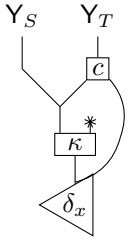
Proof. In addition, as R is a composition of Markov kernels, and hence a Markov kernel itself, L must also be a Markov kernel even if $\kappa^\dagger$ is not.

For all $x \in X$ we have a (regular) disintegration $c_x : Y_S \to \Delta(Y_T)$ of $\kappa(x; \cdot)$ by standard measurability of Y. Define $c : X \otimes Y_S \to \Delta(Y_T)$ by $c : (x, y_S) \mapsto c_x(y_S)$. Clearly, $c(x, y_S)$ is a probability distribution on Y_T for all $(x, y_S) \in X \otimes Y_S$. It remains to show $c(\cdot)^{-1}(B)$ is measurable for all $B \in \mathcal{B}([0, 1])$. But $c(\cdot)^{-1}(B) = \cap_{x \in X} c_y(\cdot)^{-1}(B)$. The right hand side is measurable by measurability of $c_y(\cdot)^{-1}(B)$ countability of X, so c is a Markov kernel.

By the definition of c_x, we have for all $x \in X$

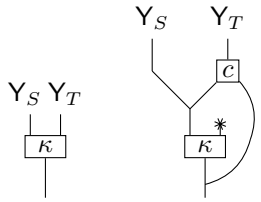


$=$ (49)



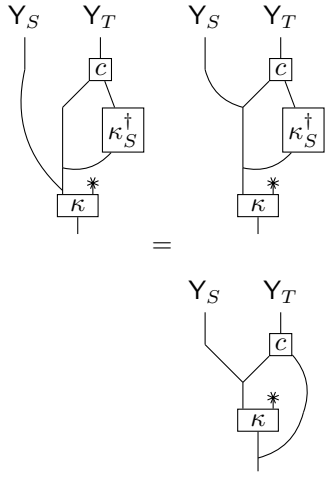
$=$ (50)

Which implies



(51)

Finally, we have



(52)

(53)

Where the first line follows from 34 and the second line from 41. If $\kappa_S^\dagger$ is a Markov kernel, then $\curlyvee(\mathrm{Id}_{Y_S} \otimes \kappa_S^\dagger)c$ is a g-disintegration. □

In the reverse direction, suppose κ is such that $\kappa\pi_T = \mathrm{Id}_X$; that is, π_T is a right inverse of κ. If $\kappa\pi_S$ is not right invertible then, by definition, there is no d such that $\kappa\pi_S d\pi_T = \mathrm{Id}_X$. However, if a g-disintegration of κ exists then there is a d such that $\kappa\pi_S d = \kappa$, a contradiction. Thus if $\kappa\pi_S$ is not right invertible then there is *in general* no g-disintegration from S to T.

4 No Free Lunch and Causal Assumptions

Call any causal theory $\mathfrak{T}$ where $\mathfrak{T} = \Delta(\mathcal{F})^D \times \mathcal{H}$ for any $\mathcal{H} \subset \Delta(\mathcal{E})$ a *skeptical theory* (this name is chosen because $\mathfrak{T}$ holds that every consequence is possible, no matter what is observed).

Theorem 4.1 (Minimax no free lunch)**.** *For any CSDP where $\mathfrak{T}$ is a skeptical theory and D, E and F are denumerable, all decision functions have the same minimax risk.*

Proof. Consider an arbitrary pair of stochastic decisions $\gamma, \gamma' \in \Delta(\mathcal{D})$. One can straightforwardly construct a kernel $A : D \to \Delta(\mathcal{D})$ such that $\gamma A = \gamma'$. Thus for any decision functions J and J', there is some Markov kernel $A : D \to \Delta(\mathcal{D})$ such that $\mu J A = \mu J'$. Noting that for all $(\kappa, \mu) \in \mathfrak{T}$ we also have $(A\kappa, \mu) \in \mathfrak{T}$, it follows that for each $(\kappa, \mu) \in \mathfrak{T}$ there is a corresponding $(A\kappa, \mu) \in \mathfrak{T}$ such that $R(J, \kappa, \mu) = R(J', A\kappa, \mu)$. □

Definition 4.2 (Constant on consequence permutations)**.** Given a causal theory $\mathfrak{T}$ with σ-algebra $\mathcal{T}$, a prior ξ on $\mathcal{T}$ is constant on consequence permutations if, given any invertible Markov kernel $S : D \to \Delta(\mathcal{D})$ (which must be a permutation map) and $A \in \mathcal{T}$ we have $\xi(A) = \xi(\{(S\kappa, \mu) | (\kappa, \mu) \in A\})$.

Lemma 4.3. *Given a set of consequences $\mathcal{K}$ where $\mathcal{K} \ni \kappa : D \to \Delta(\mathcal{F})$ with finite D and F, suppose we induce a measure on $\mathcal{K}$ via the vectorisation isomorphism vec from $\mathcal{K}$ to $\mathbb{R}^{|D||F|}$ endowed with the Borel algebra. Then the pushforward measure $\mathrm{vec}_\# \lambda$ where λ is the Lebesgue measure on $\mathbb{R}^{|D||F|}$ is constant with respect to consequence permutations.*

Proof. Let $\mathcal{K} = \{\text{vec}^{-1}(A) | A \in \mathcal{B}(\mathbb{R}^{|D||F|})\}$.

Given $A \in \mathcal{K}$ and some permutation kernel $P : D \to \Delta(\mathcal{D})$, define $PA = \{P\kappa | \kappa \in A\}$. We have

$$\text{vec}_{\#}\lambda(PA) = \lambda(\{\text{vec}(P\kappa) | \kappa \in A\}) \tag{54}$$
$$= \lambda(\{(I_{|F|} \otimes P)\text{vec}(\kappa) | \kappa \in A\}) \tag{55}$$

It is sufficient to show that $I_{|F|} \otimes P$ is a volume preserving transformation i.e. $|\det(I_{|F|} \otimes P)| = 1$. But $|\det(I_{|F|} \otimes P)| = |\det(I_{|F|})^{|D|}\det(P)^{|F|}| = |1^{|D|}(\pm 1)^{|F|}| = 1$. □

Theorem 4.4 (Bayes no free lunch)**.** *For any CSDP where $\mathcal{T} = \tilde{F}^D \times \mathcal{H}$ is a deterministic skeptical theory and D, E and F are finite, given a ξ that is constant on consequence permutations with respect on $\mathcal{T}$, all decision functions have the same Bayes risk.*

Proof. Given $y, y' \in D$, let $S_{yy'} : D \to \Delta(\mathcal{D})$ be the Markov kernel that swaps y and y' and leaves all other decisions in place: $S_{yy'}(y, A) := \delta_{y'}(A)$, $S_{yy'}(y', A) := \delta_y(A)$ and $S^2_{yy'} = I$. Clearly, for all consequences κ, $\delta_y S_{yy'}\kappa = \delta_{y'}\kappa$ and $\delta_{y'} S_{yy'}\kappa = \delta_y\kappa$.

In addition, for all κ, $f : \kappa \mapsto S_{yy'}\kappa$ is an invertible map from $\Delta(\mathcal{F})^D \to \Delta(\mathcal{F})^D$. By assumption then, for all $A \in \mathcal{T}$, $\xi(A) = \xi(\{(S_{yy'}\kappa, \mu) | (\kappa, \mu) \in A\})$. Therefore, defining the pushforward measure $f_{\#}\xi : A \mapsto \xi(f^{-1}(A))$, we have $f_{\#}\xi(A) = \xi(A)$.

Thus for any ordinary utility $u : F \to \mathbb{R}$:

$$\int_{\Delta(\mathcal{F})^D} \delta_y \kappa u d\xi = \int_{\Delta(\mathcal{F})^D} \delta_y \kappa u d f_{\#}\xi \tag{56}$$
$$= \int_{\Delta(\mathcal{F})^D} \delta_y S_{yy'}\kappa u d\xi \tag{57}$$
$$= \int_{\Delta(\mathcal{F})^D} \delta_{y'}\kappa u d\xi \tag{58}$$

Thus for any $\gamma \in \Delta(\mathcal{D})$:

$$\int_{\Delta(\mathcal{F})^D} \gamma\kappa u d\xi = \int_{\Delta(\mathcal{F})^D} \int_D \delta_y \kappa u d\gamma(y) d\xi \tag{59}$$
$$= \int_{\Delta(\mathcal{F})^D} \delta_{y_0}\kappa u \int_D d\gamma d\xi \tag{60}$$
$$= \int_{\Delta(\mathcal{F})^D} \delta_{y_0}\kappa u d\xi \tag{61}$$

For some $y_0 \in D$.

For all observational distributions $\mu \in \Delta(\mathcal{E})$ and stochastic decision functions J, $\mu J \in \Delta(\mathcal{D})$, so for all μ and J we have

$$\int_{\Delta(\mathcal{F})^D} \mu J \kappa u d\xi = \int_{\Delta(\mathcal{F})^D} \delta_{y_0}\kappa u d\xi \tag{62}$$

□

These no-free-lunch theorems are perhaps not very surprising - if we suppose that any consequence is equally likely not matter what we observe, it is not surprising that all decisions appear equally good.

4.1 Types of causal assumption

In light of the no free lunch theorems, it is necessary to introduce assumptions that yield causal theories smaller than skeptical theories, or to introduce nontrivial priors on a given skeptical theory. Because it is simpler than introducing priors, here we discuss assumptions that place hard restrictions on a causal theory.

Definition 4.5 (Statistical assumption)**.** A *statistical assumption* or *hypothesis class* is a pair of distribution classes $\mathcal{H} \subset \Delta(\mathcal{E})$, $\mathcal{I} \subset \Delta(\mathcal{F})$ that is assumed to contain both the observational distributions and outcome distributions. Specifically, a causal theory $\mathcal{T}$ on E, D and E is compatible with a hypothesis class $(\mathcal{H}, \mathcal{I})$ iff for all $(\kappa, \mu) \in \mathcal{T}$ we have $\mu \in \mathcal{H}$ and $\text{Conv}(\kappa) \subset \mathcal{I}$ where Conv denotes the convex hull.

The maximal causal theory compatible with $(\mathcal{H}, \mathcal{I})$ is the causal theory $\mathcal{T}^{\mathcal{H}}$ such that for all $\mathcal{T}$ compatible with $\mathcal{H}$, $\mathcal{T}^{\mathcal{H}} \supset \mathcal{T}$

Examples: Assuming that observations are IID and Gaussian fixes some $\mathcal{H}$. If we can additionally conduct hard interventions on some variables, $\mathcal{I}$ must must contain delta distributions on these variables, so $\mathcal{I}$ is not in general equal to $\mathcal{H}$, even if $E = F$.

Definition 4.6 (Consistency assumption)**.** Given $\mathcal{H} \subset \Delta(\mathcal{E})$ and $\mathcal{I} \subset \Delta(\mathcal{F})$, a *consistency assumption* is a set valued function $\mathcal{H} \to \mathcal{P}(\mathcal{I})$ (where $\mathcal{P}$ denotes the power set) that constrains the convex hull of the consequences given a particular observed distribution. Specifically, $\mathcal{T}$ is a causal theory compatible with a consistency assumption $C : \mathcal{H} \to \mathcal{P}(\mathcal{I})$ iff for all $(\kappa, \mu) \in \mathcal{T}$ we have $\mu \in \mathcal{H}$ and $\text{Conv}(\kappa) \subset C(\mu)$.

The maximal causal theory compatible with C is defined analogously to the maximal theory compatible with a hypothesis class.

Example: Invariance of conditional distributions: Suppose we have $E = F$ and random variables X and Y on E such that $\mu_{\mathsf{X}} F_{\mathsf{Y}} = (\gamma\kappa)_{\mathsf{X}} F_{\mathsf{Y}}$ for all $\gamma \in \Delta(\mathcal{D})$ and $(\kappa, \mu) \in \mathcal{T}$ (where μ_{X} denotes the conditional distribution on X and μF_{Y} denotes the marginal over Y). Then $\mathcal{T}$ is compatible with the consistency assumption $C : \mu \mapsto \{\nu | \nu_{\mathsf{X}} F_{\mathsf{Y}} = \mu_{\mathsf{X}} F_{\mathsf{Y}}\}$.

Every statistical assumption $(\mathcal{H}, \mathcal{I})$ can be associated with a consistency assumption $C : \mathcal{H} \to \mathcal{P}(\mathcal{I})$ where $C : \mu \mapsto \mathcal{I}$ for all $\mu \in \mathcal{H}$. Consistency assumptions can relate outcome distributions to input distributions but they cannot relate decisions to input distributions.

Definition 4.7 (Causal assumption)**.** A *causal assumption* is simply a causal theory. This is the most general type of assumption.

A special class of causal assumption is the *a priori causal assumption*, which is associated wtih a class of consequences $\mathcal{K} \subset \Delta(\mathcal{F})^D$. A theory $\mathcal{T}$ is compatible with an a priori assumption $\mathcal{K}$ iff for every $(\kappa, \mu) \in \mathcal{T}$ we have $\kappa \in \mathcal{K}$.

Example: Suppose we have a causal theory $\mathcal{T}$ that features hard interventions on a random variable X and nothing else; that is, for every $(\kappa, \mu) \in \mathcal{T}$ we have $\delta_y \kappa F_{\mathsf{X}}(A) = \delta_y(A)$. This is an a priori causal assumption defined by the set of consequences $\mathcal{K} = \{\kappa | \delta_y \kappa F_{\mathsf{X}}(A) = \delta_y(A)\}$.

Suppose we have a causal theory $\mathcal{T}$ on $D = \{0, 1\}$, $E = F = \{0, 1\}$ where we know only that decision 0 (which is our only available decision) does not increase the entropy of the observed variable. Thus $\mathcal{T} = \{(\kappa, \mu) | \|0.5 - \kappa(0; \{1\})\| \leq \|0.5 - \mu\|\}$.

As before, causal assumptions are strictly more general than consistency assumptions. This is not true for *a priori* causal assumptions.

The causal theory compatible some set of assumptions A is the intersection of maximal theories compatible with each $a \in A$. In order to obtain nontrivial results, at least one causal assumption is required.

Lemma 4.8. *Given an arbitrary set of consistency assumptions A, the causal theory $\mathcal{T}$ compatible with A is closed under left multiplication of the consequences.*

Proof. Suppose $\mathcal{T}$ is a causal theory on E, D and F. For any $L : D \to \Delta(\mathcal{D})$, $\text{Conv}(L\kappa) \subset \text{Conv}(\kappa)$.

For each $\mu \in \mathcal{H}$, $\mathcal{T}$ contains each consequence κ such that $\text{Conv}(\kappa) \subset \int_{c \in A} c(\mu)$. Thus for all $(\kappa, \mu) \in \mathcal{T}$, $(L\kappa, \mu) \in \mathcal{T}$. □

Theorem 4.9 ("No causes in no causes out" minimax)**.** *Given the causal theory $\mathcal{T}$ compatible with a set of assumptions A where each $a \in A$ is a statistical or consistency assumption, all decision functions have the same minimax risk.*

Proof. Without loss of generality, assume each $a \in A$ is a consistency assumption. We follow the proof for Theorem 4.1.

Consider an arbitrary pair of stochastic decisions $\gamma, \gamma' \in \Delta(\mathcal{D})$. One can straightforwardly construct a kernel $A : D \to \Delta(\mathcal{D})$ such that $\gamma A = \gamma'$. Thus for any decision functions J and J', there is some Markov kernel $A : D \to \Delta(\mathcal{D})$ such that $\mu J A = \mu J'$. By lemma 4.8, for all $(\kappa, \mu) \in \mathcal{T}$ we also have $(A\kappa, \mu) \in \mathcal{T}$, it follows that for each $(\kappa, \mu) \in \mathcal{T}$ there is a corresponding $(A\kappa, \mu) \in \mathcal{T}$ such that $R(J, \kappa, \mu) = R(J', A\kappa, \mu)$. □

Theorem 4.10 ("No causes in no causes out" Bayes)**.** *Given the causal theory $\mathcal{T}$ with σ-algebra $\mathcal{T}$ compatible with a set of assumptions A where each $a \in A$ is a statistical or consistency assumption, given a prior ξ that is constant on consequence permutations with respect on $\mathcal{T}$, all decision functions have the same Bayes risk.*

Proof. Without loss of generality, assume assume each $a \in A$ is a consistency assumption. Note that from lemma 4.8, for every $(\kappa, \mu) \in \mathcal{T}$ we have $(S_{yy'}\kappa, \mu) \in \mathcal{T}$ also, where $S_{yy'}$ was defined in the proof of Theorem 4.4.

The proof proceeds identically to the proof of Theorem 4.4. □

It is straightforward to show that the maximal causal theory for which assumptions a and b hold is the intersection of the maximal causal theory for which assumption a holds and the maximal causal theory for which assumption b holds. Similarly, the maximal causal theory for which assumptions a or b hold is the union of maximal theories for which assumptions a and b hold separately. Thus we can build complex causal theories from unions and intersections of simpler theories.

4.2 Causal Assumption Types in Graphical Models

At a high level, causal graphical models can *almost* be broken two sets of assumptions:

- Interventional assumptions, which are causal assumptions that specify which variables are known to be affected by decisions and how they are affected
- Invariance assumptions, which are consistency assumptions that hold certain conditionals to be invariant between the observed and outcome distributions

They don't quite break down in this manner due to the fact that causal graphical models typically admit a set of interventions that affect each variable separately. Thus no conditional distribution is *truly* invariant in a standard graphical model. We take the approach here of considering graphical models to encode consistency assumptions for subsets of the full decision set D. A desirable feature of this approach is that we can view the edges in graphical models as responsible only for specifying invariant conditional distributions, while interventions are handled by extra sets of assumptions. We make use of two-colour graphs (Definition 4.13) to distinguish between "intervened nodes" and "nodes that feature invariant conditionals".

Open question: can every causal theory be "factored" into a set of "restricted consistency assumptions" and *a priori* causal theories?

Subsequently, we show that the set of possible invariance assumptions for N variables is larger than the set of graphs on N variables.

Graphical models describe causal theories in terms of *sets* of IID random variables. For the following we assume that at the outset we are given $NM \in \mathbb{N}$ random variables $\mathsf{X}_0^0, ..., \mathsf{X}_0^M, ..., \mathsf{X}_N^M$ from $E \to \mathbb{R}$ such that the given information $\mathsf{X} = \curlyvee(\otimes_{i\in[N],j\in[M]}\mathsf{X}_i^j)$. We assume that for every theory $\mathcal{T}$ under consideration $E = F$ and we make the statistical assumption that all observed and outcome data are in the class of distributions $\mathcal{H} \subset \Delta(\mathcal{E})$ such that $\mathsf{X}_i := \curlyvee(\otimes_{j\in[M]}\mathsf{X}_i^j$ are IID. Henceforth we drop the subscript and take $\mathsf{X}^j := \mathsf{X}_0^j$ to be a "representative example" of the relevant random variable.

Definition 4.11 (Directed graph, walk, path, descendant, ancestor, parent, cycle)**.** A *directed graph* $\mathbf{G}$ of degree $N \in \mathbb{N}$ is a set of N vertices $\mathbf{V} := \{V_i | i \in [N]\}$ and $\leq N^2$ directed edges $\mathbf{E} := \{(V_i, V_j) | i, j \in [N]\}$.

A *walk* in a directed graph $\mathbf{G}$ is a sequence of edges from $\mathbf{E}$, $W := [(V_{i0}, V_{j0}), ...(V_{iM}, V_{jM})]$ such that for each adjacent pair of edges one of the following holds:

- $V_{ik} = V_{i(k+1)}$
- $V_{ik} = V_{j(k+1)}$
- $V_{jk} = V_{i(k+1)}$
- $V_{jk} = V_{j(k+1)}$

A *path* or *directed path* is a sequence of edges from $\mathbf{E}$ such that for each pair of adjacent edges we have $V_{jk} = V_{i(k+1)}$. We define paths to be forward directed (if a reverse directed path exists, then a forward directed path exists by reversing the order of edges).

If there is a path P in $\mathbf{G}$ such that, for some $k \in \mathbb{N}$, $n \in \mathbb{N}^+$, $V_i \in E_k \in P$ and $V_j \in E_{k+n} \in P$ then we say V_j is a descendant of degreen n of V_i and V_i is an ancestor of degree n of V_j.

An ancestor of degree 1 is a *parent*. The set of indices $S \subset [N]$ such that $\{V_j | j \in S\}$ are parents of V_i in $\mathbf{G}$ is defined as $\mathrm{Pa}_{\mathbf{G}}(i)$ or $\mathrm{Pa}(i)$ if the graph is clear from context.

A *cycle* is a path such that for some $k \in \mathbb{N}$, $n \in \mathbb{N}^+$ we have $V_{ik} = V_{i(k+n)}$. A graph $\mathbf{G}$ contains a cycle iff there is a node V_i that is its own ancestor.

A *directed acyclic graph (DAG)* $\mathbf{G} := (\mathbf{V}, \mathbf{E})$ of degree N is a directed graph of degree N with $\leq N(N-1)/2$ directed edges such that $\mathbf{E}$ contains no cycles.

Definition 4.12 (Completion)**.** Given a DAG $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ of degree M, a *completion* of $\mathbf{G}$, denoted $\overline{\mathbf{G}} := (\mathbf{V}, \overline{\mathbf{E}})$, is any DAG where $|\overline{\mathbf{E}}| = N(N-1)/2$ and $\overline{\mathbf{E}} \supset \mathbf{E}$. A completion of $\mathbf{G}$ is a fully connected graph that shares the edges of $\mathbf{G}$.

Definition 4.13 (Two colour directed graph)**.** A *two-colour directed graph* of degree $N \in \mathbb{N}$ is a directed graph $\mathbf{G}$ along with an additional set of "off-colour" vertex labels $S \subset [N]$. The "on-colour" label set is given by $[N] \setminus S$.

Definition 4.14 (Graphical Consistency Assumption)**.** Given a two-colour directed graph of degree N $\mathbf{G} = (\mathbf{V}, \mathbf{E}, S)$ and a set of N random variables $\mathsf{X}^j : E \to \mathbb{R}$ along with a hypothesis class $\mathcal{H} \subset \Delta(\mathcal{E})$, we can define the graphical consistency assumption $C_{\mathbf{G}} : \mathcal{H} \to \mathcal{P}(\mathcal{H})$ such that $C : \mu \mapsto \{\nu | \forall i \in [N] \setminus S : \nu_{\mathrm{Pa}(\mathsf{X}_i)} F_{\mathsf{X}_i} = \mu_{\mathrm{Pa}(\mathsf{X}_i)} F_{\mathsf{X}_i}\}$.

Thegraphical consistency assumption specifies that the marginal distributions of the "on-colour" random variables conditioned on their parents in $\mathbf{G}$ are the same for observations and outcomes.

We define an elementary graphical theory as a causal theory that satisfies only the graphical consistency assumption and the IID assumption. The former we allow to hold for only a subset of possible decisions.

Definition 4.15 (Elementary graphical theory)**.** Suppose we have a causal theory $\mathcal{T} : E \times D \nrightarrow F$ with a set of random variables $\mathsf{X}_i^j : E \to \mathbb{R}$, $j \in [M], i \in [N]$ such that $\curlyvee(\otimes_{j \in [M], i \in [N]} \mathsf{X}_i^j) = I_E$.

$\mathcal{T}$ is a $\mathbf{G}, D'$*-elementary graphical theory (EGT)* for some $D' \subset D$ and two-colour directed graph $\mathbf{G}$ of degree $[M]$ iff $\mathcal{T}$ is a maximal theory for which:

1. The statistical assumption that $\mathsf{X}_i := \curlyvee(\otimes_{j \in [M]} \mathsf{X}_i^j)$ holds for $\mathcal{T}$
2. The graphical consistency assumption $C_{\mathbf{G}}$ holds for the restriction $\mathcal{T}_{D'}$ of $\mathcal{T}$ to D'

Theorems 4.9 and 4.10 apply to any $\mathbf{G}, D$-EGTs (i.e. they apply if condition 2 holds for all decisions D). They do not apply to $\mathbf{G}, D'$-EGTs in general.

Graphical causal models are typically specified with the additional ingredient of *interventions*, which specify which variables are "targeted" by a particular decision and (possibly) how they are affected. **?** characterise graphical models with *general interventions* where it is known what variables are targeted by a given intervention but not what effect these interventions have.

Proposition 4.16 (Standard Graphical Models with General Interventions)**.** *The causal theory* $\mathcal{T}_{\mathbf{G}}$ *associated with a directed acyclic graph* $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ *of degree* M *with general interventions is*

equivalent to the intersection of M EGTs $\mathcal{T}_1, ...\mathcal{T}_M$ where $\mathcal{T}_i$ is $\overline{\mathbf{G}}_i, D_i$ elementary, $\{D_i | i \in [M]\}$ is a partition of D and $\overline{\mathbf{G}}_i := (\mathbf{V}, \overline{\mathbf{E}}, \{i\})$ is an arbitrary completion of $\mathbf{G}$ with the singleton off-colour index i.

This requires the definition of graphical models under general interventions and showing that the product rule given by **?** is equivalent to the invariances given by 4.15, which is likely require additional assumptions about the regularity of the causal theory.

Definition 4.17 (Hard Interventions)**.** Suppose we have a causal theory $\mathcal{T} : E \times D \nrightarrow F$ with a set of random variables $\mathsf{X}_i^j : E \to \mathbb{R}$, $j \in [M], i \in [N]$ such that $\curlyvee(\otimes_{j\in[M],i\in[N]}\mathsf{X}_i^j) = I_E$.

$\mathcal{T}$ is a $\mathbf{G}, D'$*-elementary hard intervention theory (EHIT)* for two-colour directed graph $\mathbf{G}$ and $D' \subset D$ of degree $[M]$ with off-colour set S iff $\mathcal{T}$ is a maximal theory for which:

1. The statistical assumption that $\mathsf{X}_i := \curlyvee(\otimes_{j\in[M]}\mathsf{X}_i^j)$ holds for $\mathcal{T}$
2. There exists a function $f : D' \to \mathbb{R}^{|}S|$ such that all $(\kappa, \mu) \in \mathcal{T}'_D$, $\kappa\curlyvee(\otimes_{i\in S}F_{\mathsf{X}_i}) : (y, A) \mapsto \delta_{f(y)}(A)$.

Proposition 4.18 (Standard Graphical Models with Hard Interventions)**.** *The causal theory $\mathcal{T}_\mathbf{G}$ associated with a directed acyclic graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ of degree M with hard interventions is equivalent to the intersection of M EGTs $\mathcal{T}_1, ...\mathcal{T}_M$ and M EHITs $\mathcal{T}'_1, ..., \mathcal{T}'_M$ where $\mathcal{T}_i$ is a $\mathbf{G}_i, D_i$ EGT, $\mathcal{T}'_i$ is a $\mathbf{G}_i, D_i$ EHIT, $\{D_i | i \in [M]\}$ is a partition of D and $\overline{\mathbf{G}}_i := (\mathbf{V}, \overline{\mathbf{E}}, \{i\})$ is an arbitrary completion of $\mathbf{G}$ with the singleton off-colour index i.*

Write up example of an ungraphable invariance assumption

Is there any relationship between the number of conditional independence stantements possible and the number of invariance statements possible on a given number of variables? Maybe via Dawid's extended conditional independence

Generalised invariance assumptions: conditionals invariances may be drawn from *all possible* RVs/functions of observed data; show that this is a strict generalisation of invariance

Can we do anything sensible WRT consistency assumptions besides conditional invariance? Can we show that there is nothing else?

References

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. URL https://www.cambridge.org/core/journals/mathematical-structures-in-computer-science/article/disintegration-and-bayesian-inversion-via-string-diagrams/0581C747DB5793756FE135C70B3B6D51.

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.

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.

Bart Jacobs. From probability monads to commutative effectuses. *Journal of Logical and Algebraic Methods in Programming*, 94:200–237, January 2018. ISSN 2352-2208. doi: 10.1016/j.jlamp.2016.11.006. URL http://www.sciencedirect.com/science/article/pii/S2352220816301122.

Bart Jacobs, Aleks Kissinger, and Fabio Zanasi. Causal Inference by String Diagram Surgery. In Mikołaj Bojańczyk and Alex Simpson, editors, *Foundations of Software Science and Computation Structures*, Lecture Notes in Computer Science, pages 313–329. Springer International Publishing, 2019. ISBN 978-3-030-17127-8.

Aleks Kissinger. Abstract Tensor Systems as Monoidal Categories. In Claudia Casadio, Bob Coecke, Michael Moortgat, and Philip Scott, editors, *Categories and Types in Logic, Language, and Physics: Essays Dedicated to Jim Lambek on the Occasion of His 90th Birthday*, Lecture Notes in Computer Science, pages 235–252. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. ISBN 978-3-642-54789-8. doi: 10.1007/978-3-642-54789-8_13. URL `https://doi.org/10.1007/978-3-642-54789-8_13`.

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.

Terence Tao. *An Introduction to Measure Theory*. American Mathematical Soc., September 2011. ISBN 978-0-8218-6919-2. Google-Books-ID: HoGDAwAAQBAJ.