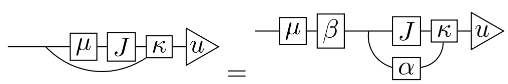
August 22: Exploring causal assumptions with string diagrams

Anonymous Author(s)
Affiliation
Address
email

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 $\mu : \mathcal{T} \to \mathcal{E}$.

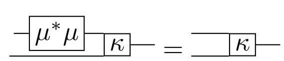
Optmizibility: I make the claim (unproven) that it is possible to find a "universally optimal" decision function if the following identity holds for all decision functions $J : E \to \Delta(\mathcal{D})$:



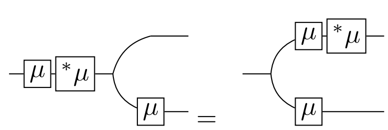
(1)

If the forward direction holds, the reverse direction does not hold - we can take a problem that respects **??** and introduce additional dominated decisions that break **??** 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). It is an open question whether the reverse direction might hold if we exclude such decisions.

Sufficient conditions for optimizibility: It is easy to show that **??** holds if there exists some kernel $^*\mu$ such that the following two identities hold:



(2)



(3)

The first condition says that κ is fixed on the support of $\mu^*\mu$.

The second is less obvious. It implies that if we "guess" the underlying state via $\mu^*\mu$ this is as good as having the actual underlying state for the purposes of determining the output of μ, but it is stronger than this. In particular, the *joint distribution* between the "guess" and the observations must be the same whether we use the guess or the true underlying state as input to μ.

Two sufficient conditions for **??** to obtain are 1) when μ is deterministic (as μ then has a left inverse) and 2) if observations are an infinite sequence of binary random variables where each μ_θ corresponds to a Bernoulli distribution for a particular parameter p_θ (via a $^*\mu$ that witnesses the strong law of large numbers).

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

A more general sufficient graphical condition is available, but it is not presently clear if it is also a necessary one.

These conditions are not necessary for **??**; 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 **??**, while we do not generally have **??**.

Below, I document additional assumptions that, along with **??** yield **??**.

I'm not sure how interesting the assumptions themselves are. One interesting point about the big picture story is that from one point of view the assumptions boil down to:

- We can characterise the input-output behaviour of κ for any given state and a small subset of available decisions

- κ is sufficiently regular that its behaviour on said subset of decisions characterises its complete behaviour

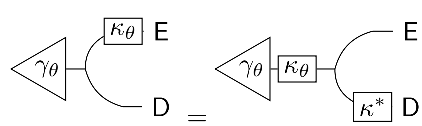
2 Recoverability

A natural assumption suggested by the notion of a CSDP is that of *recoverability* - that a causal theory $\mathcal{T} : E \times D \rightharpoonup E$ 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

$$\gamma_\theta \kappa_\theta = \mu_\theta \tag{4}$$

"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:



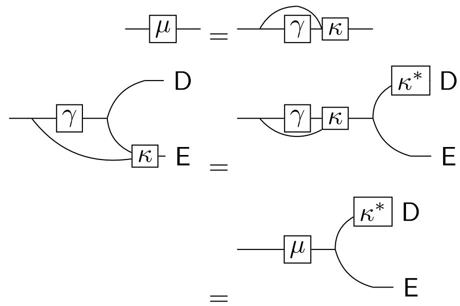
(5)

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 **??**.

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:

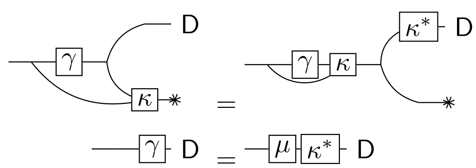


(6)

(7)

(8)

From **??** we also have

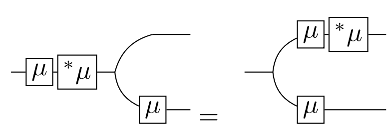


(9)

(10)

Where **??** follows from **??**.

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,



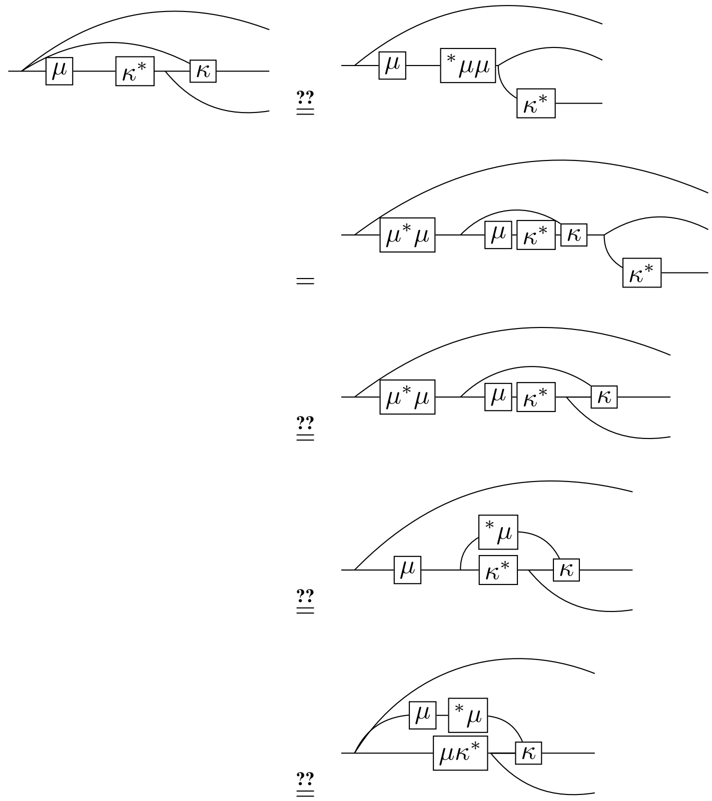
(11)

An equivalent condition to **??** 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 **??** we also have $\mu^*\mu\mu = \mu$. This weaker condition is not sufficient for the following result.

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

We then have



(12)

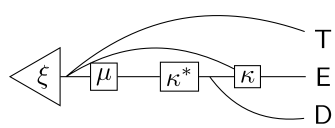
(13)

(14)

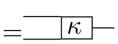
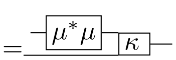
(15)

(16)

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



(17)

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 **??** for any prior ξ.

However, also by assumption **??**, 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 **??** 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 **??** would be guaranteed via a left inverse, **??** 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 $\kappa = \curlyvee \otimes_{i \in \mathbb{N}} (F_{\mathsf{D}_i} \otimes \text{Id}_{\mathcal{T}}) \kappa_0 F_{\mathsf{X}_i}$ (κ is "IID conditional on D, θ")

this might be closely related to exchange-ability 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) [**?**].

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

$$\begin{aligned}
\kappa\pi_A(\theta, y; B) &= \otimes_{i\in A}\kappa F_{\mathsf{X}_i}(\theta, \mathsf{D}_i(y); B) &(18)\\
&= \otimes_{i\in A}\kappa F_{\mathsf{X}_i}(\theta, \mathsf{D}_i(y'); B) &(19)\\
&= \kappa\pi_A(\theta, y'; B) &(20)
\end{aligned}$$

$\square$

Suppose both X_i and D_i are binary, and that for each $\theta \in \mathcal{T}$ we have recoverability (Eq. **??**) 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 **??**, 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

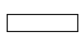
Category theoretic treatments of probability theory often start with *probability monads* (for a good overview, see [Jacobs, 2018]). A monad on some category C is a functor $T : C \to C$ along with natural transformations called the unit $\eta : 1_C \to T$ and multiplication $\mu : T^2 \to T$. Roughly, 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 $\mathcal{D}$: **Set** $\to$ **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.

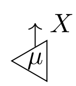
Both $\mathcal{D}$ and $\mathcal{G}$ are known to be *commutative* monads, and the Kleisli category of a commutative monad is a symmetric monoidal category.

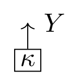
Diagrams for symmetric monoidal categories consist of wires with arrows, boxes and a couple of special symbols. The identity object (which we identify with the set $\{*\}$) is drawn as nothing at all $\{*\} :=$  and identity maps are drawn as bare wires:

$\mathrm{Id}_X :=$ 

(21)

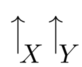
We draw Kleisli arrows from the unit (i.e. probability distributions) $\mu : \{*\} \to X$ as triangles and Kleisli arrows $\kappa : X \to Y$ (i.e. Markov kernels $X \to \Delta(\mathcal{Y})$) as boxes. We draw the Kleisli arrow $\mathbb{1}_X : X \to \{*\}$ (which is unique for each X) as below

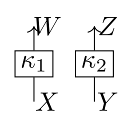
$\mu :=$ 

$\kappa :=$ 

(22)

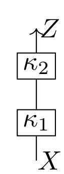
The product of objects in **Meas** is given by $(X, \mathcal{X}) \cdot (Y, \mathcal{Y}) = (X \times Y, \mathcal{X} \otimes \mathcal{Y})$, which we will often write as just $X \times Y$. Horizontal juxtaposition of wires indicates this product, and horizontal juxtaposition also indicates the tensor product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

$(X \times Y, \mathcal{X} \otimes \mathcal{Y}) :=$ 

$\kappa_1 \otimes \kappa_2 :=$ 

(23)

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

$$\kappa_1\kappa_2(x; A) = \int_Y \kappa_2(y; A)\kappa_1(x; dy) :=$$ 

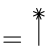
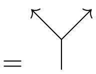
(24)

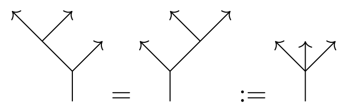
Symmetric monoidal categoris have the following coherence theorem[Selinger, 2010]:

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.*

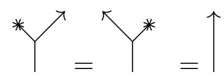
Isomorphism of diagrams for symmetric monoidal categories (somewhat informally) is any planar deformation of a diagram including deformations that cause wires to cross. We consider a diagram for a symmetric monoidal category to be well formed only if all wires point upwards.

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.

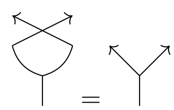
$\mathrm{Erase} = \mathbb{1}_X :=$  $\mathrm{Copy} = x \mapsto \delta_{x,x} :=$  (25)



(26)

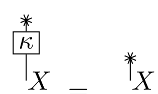


(27)



(28)

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



(29)

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 $\mathbf{Set}_\mathcal{D}$).

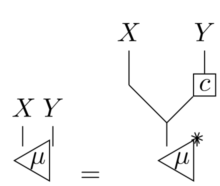
Thus when manipulating diagrams representing Markov kernels in particular (and, importantly, not more general symmetric monoidal categories) diagram isomorphism also includes applications of 6, 7, 8 and 9.

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

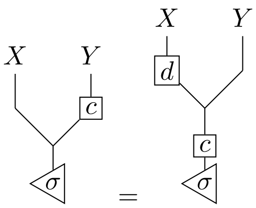


(30)

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



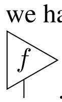
(31)

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 9. 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:

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

- 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

Since writing this, I found **?** as an example of a diagrammatic system with labeled wires, I will follow it up

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

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 8

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.

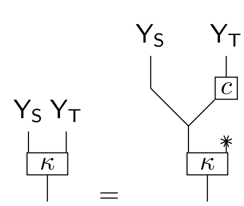
> Wire labels appear to have a key advantage over random variables: they allow us to "forget" the sample space as the correct typing is handled automatically by composition and erasure of wires

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 $\kappa : \mathsf{X}_{[\mathsf{n}]} \dashrightarrow \mathsf{Y}_{[\mathsf{m}]}$ 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):



(32)

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

$$\kappa : \begin{cases} 1 \mapsto \delta_1 \otimes \delta_1 \\ 0 \mapsto \delta_1 \otimes \delta_0 \end{cases} \tag{33}$$

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 $\int_Y \kappa^\dagger(y;A)\kappa(x;dy) = \delta_x(A)$.

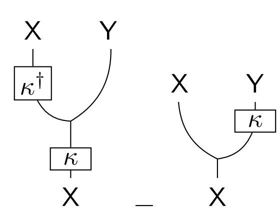
Suppose $x \notin A$ and let $B_\epsilon = \kappa_A^{\dagger -1}((\epsilon, 1])$ for some $\epsilon > 0$. We have $\int_Y \kappa^\dagger(y;A)\kappa(x;dy) = 0 \geq \epsilon\kappa(x;B_\epsilon)$. Thus for any $\epsilon > 0$ we have $\kappa(x;B_\epsilon) = 0$. Consider the set $B_0 = \kappa_A^{\dagger -1}((0,1])$. 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 $\int_Y \kappa^\dagger(y;A)\kappa(x;dy) = 1 \leq (1 - \epsilon)\kappa(x;B^{1-\epsilon}) + 1 - \kappa(x;B^{1-\epsilon}) = 1 - \epsilon\kappa(x;B^{1-\epsilon})$. 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. $\square$

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*



(34)

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

$$\begin{aligned}
L(x; A\times B) &= \int_Y \int_{Y\times Y} \mathrm{Id}_Y \otimes \kappa_S^\dagger(y, y'; A\times B)\delta_{(z,z)}(dy\times dy')\kappa\pi_S(x; dz) && (35)\\
&= \int \mathrm{Id}_Y \otimes \kappa^\dagger(z, z; A\times B)\kappa\pi_S(x; dz) && (36)\\
&= \int \delta_z(A)\kappa_S^\dagger(z; B)\kappa\pi_S(x; dz) && (37)\\
&= \int_A \kappa_S^\dagger(z; B)\kappa\pi_S(x; dz) && (38)\\
&= \delta_x(B)\kappa\pi_S(x; A) && (39)
\end{aligned}$$

Where **??** follows from Lemma **??**.

$$\begin{aligned}
R(x; A\times B) &= \int \delta_{(x,x)}(dy\times dy')\kappa\pi_S \otimes \mathrm{Id}_X(y, y'; A\times B) && (40)\\
&= \kappa\pi_S(x; A)\delta_x(B) \qquad\qquad = L && (41)
\end{aligned}$$

$\square$

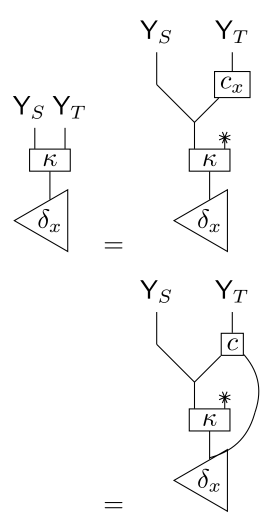
Theorem 3.6. *Given countable X and standard measurable Y, $n, m \in \mathbb{N}$, $S, T \subset [m]$, κ with label signature $\mathsf{X}_{[\mathsf{n}]} \dashrightarrow \mathsf{Y}_{[\mathsf{m}]}$ a g-disintegration exists from S to T if $\kappa\pi_S$ is right-invertible*

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.

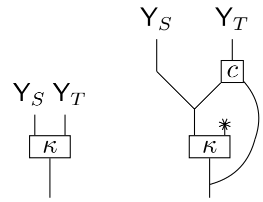
By the definition of c_x, we have for all $x \in X$



(42)

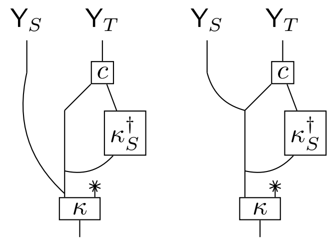
(43)

Which implies

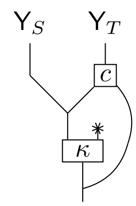


$$= \tag{44}$$

Finally, we have



$$= \tag{45}$$



$$= \tag{46}$$

Where the first line follows from 7 and the second line from **??**. 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.

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.

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.