Causal Statistical Decision Problems

Anonymous Author(s)
Affiliation
Address
`email`

1 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 **Set**$_\mathcal{D}$ are morphisms $X \to \mathcal{D}(Y)$ in **Set**, i.e. stochastic matrices, and in the Kleisli category **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.

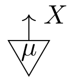
Diagrams for symmetric monoidal categories consist of wires with arrows, boxes and a couple of special symbols. The identity object (which we identify with the set $\{*\}$) is drawn as nothing at all $\{*\} := \boxed{}$ and identity maps are drawn as bare wires:

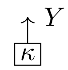
$$\mathrm{Id}_X := \;\; \uparrow_X \tag{1}$$

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

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

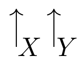
$\mathbb{1}_X : X \to \{*\}$ (which is unique for each X) as below

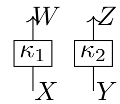
$\mu :=$ 

$\kappa :=$ 

(2)

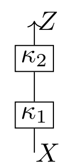
The product of objects in **Meas** is given by $(X, \mathcal{X}) \cdot (Y, \mathcal{Y}) = (X \times Y, \mathcal{X} \otimes \mathcal{Y})$, which we will often write as just $X \times Y$. Horizontal juxtaposition of wires indicates this product, and horizontal juxtaposition also indicates the tensor product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

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

$\kappa_1 \otimes \kappa_2 :=$ 

(3)

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

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

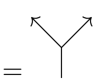
(4)

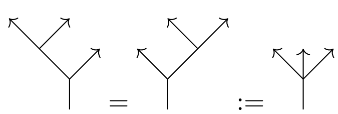
Symmetric monoidal categoris have the following coherence theorem[Selinger, 2010]:

Theorem 1.1 (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.*

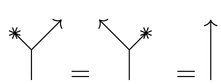
Isomorphism of diagrams for symmetric monoidal categories (somewhat informally) is any planar deformation of a diagram including deformations that cause wires to cross. We consider a diagram for a symmetric monoidal category to be well formed only if all wires point upwards.

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.

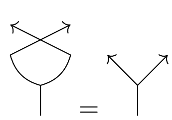
Erase $= \mathbb{1}_X :=$  Copy $= x \mapsto \delta_{x,x} :=$  (5)



(6)

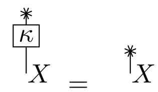


(7)



(8)

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



(9)

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

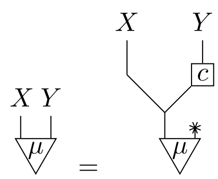
Thus when manipulating diagrams representing Markov kernels in particular (and, importantly, not more general symmetric monoidal categories) diagram isomorphism also includes applications of 6, 7, 8 and 9.

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].

1.1 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).

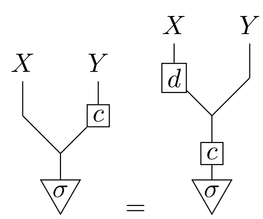
Given a distribution $\mu : \{*\} \to X \otimes Y$, a disintegration $c : X \to Y$ is a Markov kernel that satisfies



(10)

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



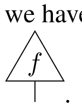
(11)

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.

1.2 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.

1.3 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

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.

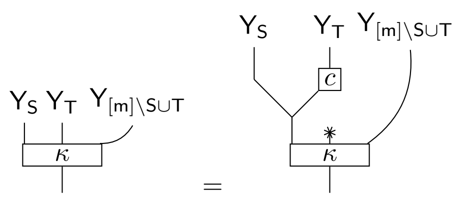
Wire labels appear to have a key advantage over random variables: they allow us to "forget" the sample space as the correct typing is handled automatically by composition and erasure of wires

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 1.2 (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 1.3 (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 $S, T \subset [m]$, 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):



(12)

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

$$\kappa : \begin{cases} 1 \mapsto \delta_1 \otimes \delta_1 \\ 0 \mapsto \delta_1 \otimes \delta_0 \end{cases} \tag{13}$$

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.

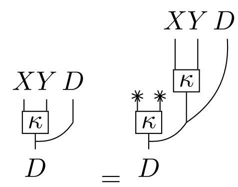
Lemma 1.4 (Right inverses are deterministic kernels)**.** *Given $\kappa : X \to \Delta(Y)$ for countable X and standard measurable Y. A right inverse $\kappa^\dagger$ exists iff there exists a measurable function $f : Y \to X$ such that for all $x \in X$, $y \in Y$, $f(y) = x$ κ_x-almost surely. Moreover, in such a case $\kappa^\dagger(y; A) = \delta_{f(y)}(A)$.*

Proof. Suppose we have such a function f. We immediately have measurability: $f^{-1}(A)$ for all measurable A is a countable union of supports which are themselves measurable. Furthermore, $\int_Y \kappa(x; dy)\delta_{f(y)}(A) = \delta_x(A)$ so $\delta_{f(y)}(A)$ is a right inverse of κ as needed.

Suppose we have a right inverse $\kappa^\dagger$. By definition, $\int_Y \kappa(x; dy)\kappa^\dagger(y; A) = \delta_x(A) \overset{\kappa_x-A.S.}{=} \int_Y \kappa(x; dy)\delta_x(A)$. $\square$

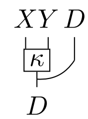
Theorem 1.5. *Given countable X, standard measurable Y, $n, m \in \mathbb{N}$, $S, T \subset [m]$ a g-disintegration exists from S to T for all $\kappa : X \to Y$ with label signature $\mathsf{X}_{[\mathsf{n}]} \dashrightarrow \mathsf{Y}_{[\mathsf{m}]}$ iff $\kappa\pi_S$ is right-invertible.*

Proof. For $D \to X \times Y$, we note that κ is always a disintegration:



(14)

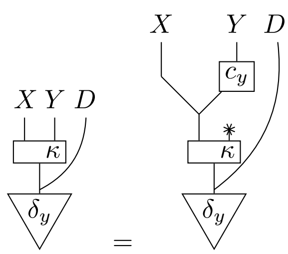
Given $\kappa : D \to X \times Y$, define the *canonical extension* κ_{id}:

$\kappa_{id} :=$ 

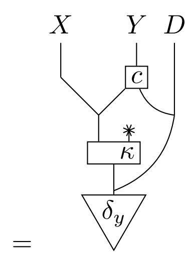
(15)

The canonical extension takes a copy of the input and maps it to the output.

$D \times X \to Y$ and $D \times Y \to X$ are symmetric directions, so we will argue only for $D \times X \to Y$. For all $y \in D$ we have a disintegration $c_y : X \to Y$ of $\delta_y \kappa$ by standard measurability of $X \times Y$. Define $c : D \times X \to Y$ by $c : (y, x) \mapsto c_y(x)$. Clearly, $c(y, x)$ is a probability distribution on Y for all $(y, x) \in D \times X$. It remains to show $c(\cdot)^{-1}(B)$ is measurable for all $B \in \mathcal{B}([0, 1])$. But $c(\cdot)^{-1}(B) = \cap_{y \in D} c_y(\cdot)^{-1}(B)$. The right hand side is measurable by measurability of $c_y(\cdot)^{-1}(B)$ and the properties of a σ-algebra, so c is a Markov kernel. By the definition of c_y, we have for all $y \in D$

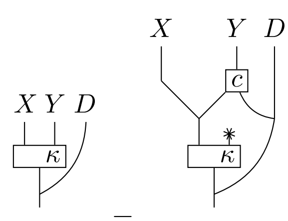


(16)



(17)

Which implies



(18)

Note that the only other (non-trivial) disintegrations do not feature D as an "input wire". All such

□

Conjecture: This can be generalised to any κ that is determined by its values on a countable set of points along with some notion of continuity. This seems likely to be true. In a more general setting, I think I could find a counterexample, but the converse also seems unlikely.

The extension of *conditional independence* to g-disintegrations becomes a directional relationship. Suppose we have $\kappa : D \to X \times Y$ and a disintegration $c : D \times X \to Y$. We say Y is directionally conditionally independent (DCI) of D given X if

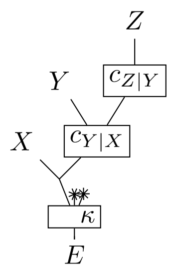
Generalised disintegrations facilitate the following construction of a "graphical model":

Suppose we have two causal theories, $\mathcal{T}^*$ and $\mathcal{T}$ both with signature $E \times D \rightarrow\!\!\triangleright E$, and $\mathcal{T}$ is a decision randomised version of $\mathcal{T}^*$ (i.e. $\mathcal{T} = \{(\lambda\kappa, \mu) | (\kappa, \mu) \in \mathcal{T}^*\}$ for some $\lambda : D \to D$. We will construct a graphical model from $\mathcal{T}^*$ and $\mathcal{T}$ in three steps:

First, we assume *reproducibility* in the stronger theory $\mathcal{T}^*$. That is, for all $(\kappa, \mu) \in \mathcal{T}^*$ we suppose there exists $\gamma \in \Delta(\mathcal{D})$ such that $\gamma\kappa = \mu$.

I don't think reproducibility is quite the right assumption, but it is good enough for now

Second, we will assume certain *generalised conditional independences* hold for the stronger theory $\mathcal{T}^*$ (we have not defined these, but they are the obvious generalisation of standard conditional independence lifted to g-disintegrations). Because we're constructing a graphical model, we will assume these are a "DAG-compatible" set, though we are under no obligation to do so. I conjecture we can illustrate these independences graphically. Suppose we have random variables $\mathsf{X} : E \to X$, $\mathsf{Y} : E \to Y$ and $\mathsf{Z} : E \to Z$, and we assume we have at least the generalised CIs implied by the following diagram for all $(\kappa, \mu) \in \mathcal{T}^*$:

$\kappa =$ 

(19)

The above diagram is typed incorrectly, but we can always construct a kernel κ_{XYZ} that maps to $X \times Y \times Z$.

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.