

August 22: Exploring causal assumptions with string diagrams

Anonymous Author(s)

Affiliation

Address

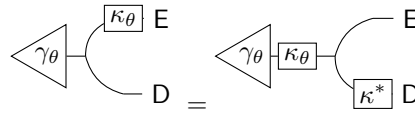
email

1 Recoverability

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

$$\gamma_\theta \kappa_\theta = \mu_\theta \quad (1)$$

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:



$$\quad (2)$$

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 $X : E \rightarrow X$ such that for all $a \in D$, $\theta \in \mathcal{T}$ we have $\delta_a \kappa_\theta F_X = \delta_a$ (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 2.

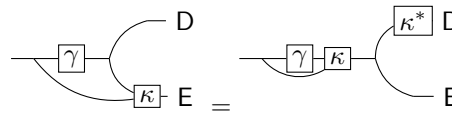
I believe a weaker sufficient condition for the existence of a universal κ^* is that every κ_θ factorises as $\kappa_\theta = h \vee (\text{Id}_F \otimes j_\theta)$ for some fixed $h : D \rightarrow \Delta(\mathcal{F})$, but I have not yet shown this.

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

By previous assumptions, we have the following properties:



$$\quad (3)$$



$$\quad (4)$$



$$\quad (5)$$

18 From 4 we also have

$$\begin{array}{c} \text{---} \end{array} \begin{array}{|c|} \hline \gamma \\ \hline \end{array} \begin{array}{c} \text{---} \text{D} \\ \text{---} \kappa^* \end{array} = \begin{array}{c} \text{---} \end{array} \begin{array}{|c|} \hline \gamma \kappa \\ \hline \end{array} \begin{array}{c} \text{---} \kappa^* \text{D} \\ \text{---} \end{array} \quad (6)$$

$$\text{---} \boxed{\gamma} \text{---} D = \text{---} \boxed{\mu} \boxed{\kappa^*} \text{---} D \quad (7)$$

19 Where 7 follows from 1.

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

$$\begin{array}{c} \boxed{\mu} \boxed{*} \boxed{\mu} \end{array} \begin{array}{c} \diagup \\ \diagdown \end{array} \begin{array}{c} \boxed{\mu} \end{array} = \begin{array}{c} \begin{array}{c} \boxed{\mu} \boxed{*} \boxed{\mu} \\ \boxed{\mu} \end{array} \\ \begin{array}{c} \boxed{\mu} \end{array} \end{array} \quad (8)$$

23 An equivalent condition to 8 is that for all $\theta, \theta' \in \mathcal{T}$, $A \in \mathcal{E}$, we have $\mu(\theta; A) = \mu(\theta'; A)$, $\mu^* \mu(\theta; \cdot)$
24 almost surely. More informally, the support of $\mu^* \mu$ for each input θ divides \mathcal{T} into equivalence classes
25 such that for all θ in a given equivalence class, μ maps to the same probability measure on \mathcal{E} .

26 Note that as a result of 8 we also have $\mu^* \mu \mu = \mu$. This weaker condition is not sufficient for the
27 following result.

There is a connection between equation 8 and the notion of a sufficient statistic

29 We then have

$$\begin{array}{c}
\text{Diagram 1: } \mu \rightarrow \kappa^* \rightarrow \kappa \\
\text{Diagram 2: } \mu \rightarrow \mu\mu^* \rightarrow \kappa^* \\
\hline
(9)
\end{array}$$

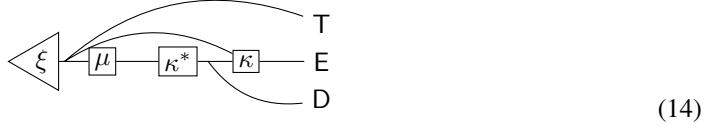
$$= \text{Diagram (10)} \quad (10)$$

$$\begin{array}{c} \text{---} \mu^* \mu \text{---} \mu \kappa^* \text{---} \kappa \text{---} \\ \text{---} \end{array} \quad (11)$$

$$\begin{array}{c} \text{---} \mu L \text{---} \mu^* L \text{---} K^* \text{---} K \text{---} \\ \text{---} \end{array} \quad (12)$$

$$\begin{array}{c} \text{---} \mu \text{---} \mu^* \text{---} \\ \text{---} \mu \kappa^* \text{---} \end{array} \quad \begin{array}{c} \text{---} \kappa \text{---} \end{array} \quad (13)$$

Equation 13 implies that, given any $\xi \in \Delta(\mathcal{T})$, all distributions of the form



admit both $\boxed{\kappa}$ and $\boxed{\mu^* \mu} \boxed{\kappa}$ as disintegrations from $(D, T) \dashrightarrow E$. Therefore these two kernels agree almost surely with respect to the distribution 14 for any prior ξ .

However, also by assumption 8, 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 14 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 8 would be guaranteed via a left inverse, 14 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 kernels 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 $X_i : E \rightarrow W$ and $D_i : D \rightarrow V$ 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)
- For $A := A_0 \times A_1 \times \dots \in E, \mu = \mu \curlyvee \otimes_{i \in \mathbb{N}} F_{X_i} = \curlyvee \otimes_{i \in \mathbb{N}} \mu F_{X_i}$ (the X_i 's are IID)
- For $y := (y_0, y_1, \dots) \in D, A \in E, \kappa = \curlyvee \otimes_{i \in \mathbb{N}} F_{D_i} \kappa F_{X_i}$ (κ is "IID")

the sequence is IID with respect to $\mu(\theta; \cdot)$ for all θ .

We appear to require additional assumptions in order to support a nontrivial result, however. Given that μ is deterministic, we might suppose κ^* is also likely to be deterministic (or, if it is not, the nondeterminism is not carried through by κ). Then we will have only a single pair in $D \times \mathcal{T}$ having nonzero measure for each $\theta \in \mathcal{T}$, so the "almost surely" condition rules out almost all feasible decision functions. There appear to be two competing demands - we want κ^* stochastic in order to determine the results of a wide variety of decision functions, but we want μ deterministic in order to support statistical inference. One option might be to relax the assumption of determinism on μ slightly and hope that we "gain more than we lose". I suspect, at this point, that this does not work.

An alternative nontrivial case of "optimisability" requires the additional assumption of *double exchangeability*. This is exchangeability in the standard statistical sense, not in the sense of the Rubin causal model; a doubly exchangeable kernel is a kernel that remains the same if inputs and outputs are permuted in the same way.

Definition 1.1 (Double exchangeability). A kernel $\kappa : X \rightarrow \Delta(\mathcal{Y})$ is *doubly exchangeable* with respect to random variable sets $\{X_i\}_{i \in A}, \{Y_i\}_{i \in A}$ where $A = [n]$ or $A = \mathbb{N}$ and $X_i : X \rightarrow X_i, Y_i : Y \rightarrow Y_i$ if, given any finite permutation σ and its inverse σ^{-1} we have both

- There exists $\sigma_X : X \rightarrow X$ and $\sigma_Y^{-1} : Y \rightarrow Y$ such that $F_{\sigma_X} \curlyvee (\otimes_{i \in A} X_{a_i}) = \curlyvee (\otimes_{i \in A} X_{\sigma(a_i)})$ and similarly for $F_{\sigma_Y^{-1}}$ and
- $F_{\sigma_X} \kappa F_{\sigma_Y^{-1}} = \kappa$

Double exchangeability is similar to exchangeability for probability distributions, but there is no possible analogue of the set $\{X_i\}$ in that case.

An example follows. We identify $\mathcal{T} \cong [0, 1] \times T, E \cong [0, 1]^2$ and $D \cong \{0, 1\}^{\mathbb{N}}$. For $(\theta, \phi) \in \mathcal{T}$ let $\mu : (\theta; A \times B) \mapsto \delta_{0.5}(A) \delta_{\frac{\theta}{2}}(B)$.

71 Define $\bar{D}^n : \{0, 1\}^n \rightarrow [0, 1]$ by $\bar{D}^n : (y_0, \dots, y_n) \mapsto \frac{1}{n} \sum_{i \in [n]} y_i$ and let $\bar{D} : D \rightarrow [0, 1]$ be the limit
 72 $\bar{D} = \lim_{n \rightarrow \infty} \bar{D}^n$. Let γ be for all (θ, ϕ) the unique distribution such that $\gamma F_{\bar{D}^n}(\theta, \phi; A) = \delta_{0.5}(A)$
 73 (i.e. the distribution of an infinite sequence of IID RVs with success probability 0.5) and assert that
 74 $\delta_{(\theta, \phi)} \Upsilon (\text{Id}_{\mathcal{T}} \otimes \gamma) \kappa = \delta_{\theta} \mu$ almost surely for all (θ, ϕ) . We note that $\kappa^* : (\cdot; A) \mapsto \gamma(\cdot; A)$ satisfies 4
 75 for arbitrary κ .

76 2 Notes on category theoretic probability and string diagrams

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

85 An example of a probability monad is the discrete probability monad given by the functor $\mathcal{D} : \mathbf{Set} \rightarrow$
 86 \mathbf{Set} which maps a countable set X to the set of functions from $X \rightarrow [0, 1]$ that are probability
 87 measures on X , denoted $\mathcal{D}(X)$. \mathcal{D} maps a measurable function f to $\mathcal{D}f : X \rightarrow \mathcal{D}(X)$ given by
 88 $\mathcal{D}f : x \mapsto \delta_{f(x)}$. The unit of this monad is the map $\eta_X : X \rightarrow \mathcal{D}(X)$ given by $\eta_X : x \mapsto \delta_x$ (which
 89 is equivalent to $\mathcal{D}1_X$) and multiplication is $\mu_X : \mathcal{D}^2(X) \rightarrow \mathcal{D}(X)$ where $\mu_X : \Omega \mapsto \sum_{\phi} \Omega(\phi) \phi$.

90 For continuous distributions we have the Giry monad on the category \mathbf{Meas} of measurable spaces
 91 given by the functor \mathcal{G} which maps a measurable space X to the set of probability measures on X ,
 92 denoted $\mathcal{G}(X)$. Other elements of the monad (unit, multiplication and map between morphisms) are
 93 the "continuous" version of the above.

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

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

102 Diagrams for symmetric monoidal categories consist of wires with arrows, boxes and a couple of
 103 special symbols. The identity object (which we identify with the set $\{*\}$) is drawn as nothing at all
 104 $\{*\} := \square$ and identity maps are drawn as bare wires:

$$\text{Id}_X := \uparrow_X \quad (15)$$

105 We draw Kleisli arrows from the unit (i.e. probability distributions) $\mu : \{*\} \rightarrow X$ as triangles and
 106 Kleisli arrows $\kappa : X \rightarrow Y$ (i.e. Markov kernels $X \rightarrow \Delta(\mathcal{Y})$) as boxes. We draw the Kleisli arrow
 107 $\mathbb{1}_X : X \rightarrow \{*\}$ (which is unique for each X) as below

$$\mu := \begin{array}{c} \uparrow X \\ \triangle \\ \mu \end{array} \quad \kappa := \begin{array}{c} \uparrow Y \\ \boxed{\kappa} \end{array} \quad (16)$$

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

$$(X \times Y, \mathcal{X} \otimes \mathcal{Y}) := \begin{array}{c} \uparrow X \quad \uparrow Y \end{array} \quad \kappa_1 \otimes \kappa_2 := \begin{array}{cc} \begin{array}{c} \uparrow W \\ \boxed{\kappa_1} \end{array} & \begin{array}{c} \uparrow Z \\ \boxed{\kappa_2} \end{array} \\ \downarrow X & \downarrow Y \end{array} \quad (17)$$

111 Composition of arrows is achieved by “wiring” boxes together. For $\kappa_1 : X \rightarrow Y$ and $\kappa_2 : Y \rightarrow Z$
 112 we have

$$\kappa_1 \kappa_2(x; A) = \int_Y \kappa_2(y; A) \kappa_1(x; dy) := \begin{array}{c} \uparrow Z \\ \boxed{\kappa_2} \\ \downarrow \\ \boxed{\kappa_1} \\ \downarrow X \end{array} \quad (18)$$

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

114 **Theorem 2.1** (Coherence (symmetric monoidal)). *A well-formed equation between morphisms in*
 115 *the language of symmetric monoidal categories follows from the axioms of symmetric monoidal*
 116 *categories if and only if it holds, up to isomorphism of diagrams, in the graphical language.*

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

120 In fact the Kleisli categories of the probability monads above have (for each object) unique *copy*:
 121 $X \rightarrow X \times X$ and *erase*: $X \rightarrow \{*\}$ maps that satisfy the *commutative comonoid axioms* that (thanks
 122 to the coherence theorem above) can be stated graphically. These differ from the copy and erase
 123 maps of *finite product* or *cartesian* categories in that they do not necessarily respect composition of
 124 morphisms.

$$\text{Erase} = \mathbb{1}_X := \begin{array}{c} * \\ \downarrow \end{array} \quad \text{Copy} = x \mapsto \delta_{x,x} := \begin{array}{c} \swarrow \quad \searrow \\ \downarrow \end{array} \quad (19)$$

$$\begin{array}{c} \swarrow \quad \searrow \\ \downarrow \end{array} = \begin{array}{c} \swarrow \quad \searrow \\ \downarrow \end{array} := \begin{array}{c} \swarrow \quad \searrow \\ \downarrow \end{array} \quad (20)$$

$$\begin{array}{c} * \\ \downarrow \end{array} = \begin{array}{c} * \\ \downarrow \end{array} = \begin{array}{c} \uparrow \end{array} \quad (21)$$

$$\begin{array}{c} \swarrow \quad \searrow \\ \downarrow \end{array} = \begin{array}{c} \swarrow \quad \searrow \\ \downarrow \end{array} \quad (22)$$

125 Finally, $\{*\}$ is a terminal object in the Kleisli categories of either probability monad. This means
 126 that the map $X \rightarrow \{*\}$ is unique for all objects X , and as a consequence for all objects X, Y and all
 127 $\kappa : X \rightarrow Y$ we have

$$\begin{array}{c} * \\ \boxed{\kappa} \\ \downarrow X \end{array} = \begin{array}{c} * \\ \downarrow X \end{array} \quad (23)$$

128 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
 129 what differentiates a stochastic matrix from a general positive matrix (which live in a larger category
 130 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 20, 21, 22 and 23.

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].

2.1 Disintegration and Bayesian inversion

Disintegration is a key operation on probability distributions (equivalently arrows $\{*\} \rightarrow X$) 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 $\mu : \{*\} \rightarrow X \otimes Y$, a disintegration $c : X \rightarrow Y$ is a Markov kernel that satisfies

$$\begin{array}{c} X \quad Y \\ | \quad | \\ \triangleleft \mu \end{array} = \begin{array}{c} X \quad Y \\ | \quad | \\ \triangleleft \mu \end{array} \begin{array}{c} \square c \\ \square \mu^* \end{array} \quad (24)$$

Disintegrations always exist in $\mathbf{Set}_{\mathcal{D}}$ but not in $\mathbf{Meas}_{\mathcal{G}}$. They do exist in the latter if we restrict ourselves to standard measurable spaces. If c_1 and c_2 are disintegrations $X \rightarrow Y$ 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 $\sigma : \{*\} \rightarrow X$ and a channel $c : X \rightarrow Y$, a Bayesian inversion of (σ, c) is a channel $d : Y \rightarrow X$ such that

$$\begin{array}{c} X \quad Y \\ | \quad | \\ \triangleleft \sigma \end{array} \begin{array}{c} \square c \\ \square \sigma \end{array} = \begin{array}{c} X \quad Y \\ | \quad | \\ \triangleleft \sigma \end{array} \begin{array}{c} \square d \\ \square c \end{array} \quad (25)$$

We can obtain disintegrations from Bayesian inversions and vice-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.

2.2 Generalisations

Cho and Jacobs [2019] make use of a larger “CD” category by dropping 23. 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.

2.3 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).

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

163 That is, we assign a unique label to each bare wire in the diagram with the following additional
 164 qualifications:

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

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

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

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

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

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

- 183 • These properties obviously hold for the wire segments into and out of boxes
- 184 • The only features a diagram may have apart from boxes and wires are wire crossings, copy
 185 maps and erase maps
- 186 • The labeling rule for wire crossings respects the symmetry of the swap map
- 187 • The labeling rule for copy maps respects the symmetry of the copy map and the property
 188 described in Equation 22

189 We will follow the convention whereby “internal” wire labels are omitted from diagrams.

190 Note also that each wire that terminates in a free end can be associated with a random variable.
 191 Suppose for $N \in \mathbb{N}$ we have a kernel $\kappa : A \rightarrow \Delta(\times_{i \in N} X_i)$. Define by p_j ($j \in [N]$) the projection
 192 map $p_j : \times_{i \in N} X_i \rightarrow X_j$ defined by $p_j : (x_0, \dots, x_N) \mapsto x_j$. p_j is a measurable function, hence
 193 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
 194 $\kappa(y; p_j^{-1}(A)) = \int_{X_j} \delta_{p_j(\mathbf{x})}(A) \kappa(y; d\mathbf{x}) = \kappa \pi_j$. Diagrammatically, π_j is the identity map on the j -th
 195 wire tensored with the erase map on every other wire. Thus the j -th wire carries the distribution
 196 associated with the random variable p_j . We will therefore consider the labels of the “outgoing” wires
 197 of a diagram to denote random variables (though there are obviously many random variables not
 198 represented by such wires). We will additionally distinguish wire labels from spaces by font - wire
 199 labels are sans serif A, B, C, X, Y, 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

200

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

206 **Definition 2.2** (Label signatures). If a kernel $\kappa : X \rightarrow \Delta(Y)$ can be represented by a diagram
 207 G with incoming wires X_1, \dots, X_n and outgoing wires Y_1, \dots, Y_m , we can assign the kernel a “label

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

signature” $\kappa : X_1 \otimes \dots \otimes X_n \dashrightarrow Y_1 \otimes \dots \otimes Y_m$ or, for short, $\kappa : X_{[n]} \dashrightarrow Y_{[m]}$. 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 $\otimes_{i \in [n]} X_i$ to functions of a type similar to $\otimes_{i \in [m]} Y_i$ (or vice versa - perhaps related to Clerc et al. [2017]), but wire labels are all we need at this point

214

Definition 2.3 (Generalised disintegration). Given a kernel $\kappa : X \rightarrow \Delta(Y)$ with label signature $\kappa : X_{[n]} \dashrightarrow Y_{[m]}$ and disjoint subsets $S, T \subset [m]$ such that $S \cup T = [m]$, a kernel c is a *g-disintegration from S to T* if it’s type is compatible with the label signature $c : Y_S \dashrightarrow Y_T$ and we have the identity (omitting incoming wire labels):

$$\text{Diagram 1} = \text{Diagram 2} \quad (26)$$

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”.

219

In contrast to regular disintegrations, generalised disintegrations “usually” do not exist. Consider $X = \{0, 1\}$, $Y = \{0, 1\}^2$ and κ has label signature $X_1 \dashrightarrow Y_{\{1,2\}}$ with

$$\kappa : \begin{cases} 1 \mapsto \delta_1 \otimes \delta_1 \\ 0 \mapsto \delta_1 \otimes \delta_0 \end{cases} \quad (27)$$

κ imposes contradictory requirements for any disintegration $c : \{0, 1\} \rightarrow \{0, 1\}$ from $\{1\}$ to $\{2\}$: equality for $X_1 = 1$ requires $c(1; \cdot) = \delta_1$ while equality for $X_1 = 0$ 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 2.4. Given $\kappa : X \rightarrow \Delta(Y)$, a kernel κ^\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 κ^\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, $\kappa \kappa^\dagger = \text{Id}_X$, so κ^\dagger is a right inverse of κ .

Suppose we have a right inverse κ^\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 $\lim_{i \rightarrow \infty} \epsilon_i = 0$ 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 2.5. Given $\kappa : X \rightarrow \Delta(Y)$ and a right inverse κ^\dagger , we have

$$(28)$$

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

$$L(x; A \times B) = \int_Y \int_{Y \times Y} \text{Id}_Y \otimes \kappa_S^\dagger(y, y'; A \times B) \delta_{(z, z)}(dy \times dy') \kappa \pi_S(x; dz) \quad (29)$$

$$= \int \text{Id}_Y \otimes \kappa^\dagger(z, z; A \times B) \kappa \pi_S(x; dz) \quad (30)$$

$$= \int \delta_z(A) \kappa_S^\dagger(z; B) \kappa \pi_S(x; dz) \quad (31)$$

$$= \int_A \kappa_S^\dagger(z; B) \kappa \pi_S(x; dz) \quad (32)$$

$$= \delta_x(B) \kappa \pi_S(x; A) \quad (33)$$

Where 33 follows from Lemma 2.4.

$$R(x; A \times B) = \int \delta_{(x, x)}(dy \times dy') \kappa \pi_S \otimes \text{Id}_X(y, y'; A \times B) \quad (34)$$

$$= \kappa \pi_S(x; A) \delta_x(B) = L \quad (35)$$

□

Theorem 2.6. Given countable X and standard measurable Y , $n, m \in \mathbb{N}$, $S, T \subset [m]$, κ with label signature $X_{[n]} \dashrightarrow Y_{[m]}$ 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 κ^\dagger is not.

For all $x \in X$ we have a (regular) disintegration $c_x : Y_S \rightarrow \Delta(Y_T)$ of $\kappa(x; \cdot)$ by standard measurability of Y . Define $c : X \otimes Y_S \rightarrow \Delta(Y_T)$ 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 $x \in X$

$$\begin{array}{c}
Y_S \quad Y_T \\
\downarrow \quad \downarrow \\
\boxed{\kappa} \\
\downarrow \\
\triangle \delta_x
\end{array}
=
\begin{array}{c}
Y_S \quad Y_T \\
\downarrow \quad \downarrow \\
\boxed{C_x} \\
\downarrow \\
\boxed{\kappa}^* \\
\downarrow \\
\triangle \delta_x
\end{array}
\quad (36)$$

$$=
\begin{array}{c}
Y_S \quad Y_T \\
\downarrow \quad \downarrow \\
\boxed{C} \\
\downarrow \\
\boxed{\kappa}^* \\
\downarrow \\
\triangle \delta_x
\end{array}
\quad (37)$$

260 Which implies

$$\begin{array}{c}
Y_S \quad Y_T \\
\downarrow \quad \downarrow \\
\boxed{\kappa}
\end{array}
=
\begin{array}{c}
Y_S \quad Y_T \\
\downarrow \quad \downarrow \\
\boxed{C} \\
\downarrow \\
\boxed{\kappa}^* \\
\downarrow \\
\triangle \delta_x
\end{array}
\quad (38)$$

261 Finally, we have

$$\begin{array}{c}
Y_S \quad Y_T \\
\downarrow \quad \downarrow \\
\boxed{C} \\
\downarrow \\
\boxed{\kappa_S^\dagger} \\
\downarrow \\
\boxed{\kappa}^* \\
\downarrow \\
\triangle \delta_x
\end{array}
=
\begin{array}{c}
Y_S \quad Y_T \\
\downarrow \quad \downarrow \\
\boxed{C} \\
\downarrow \\
\boxed{\kappa_S^\dagger} \\
\downarrow \\
\boxed{\kappa}^* \\
\downarrow \\
\triangle \delta_x
\end{array}
\quad (39)$$

$$=
\begin{array}{c}
Y_S \quad Y_T \\
\downarrow \quad \downarrow \\
\boxed{C} \\
\downarrow \\
\boxed{\kappa}^* \\
\downarrow \\
\triangle \delta_x
\end{array}
\quad (40)$$

262 Where the first line follows from 21 and the second line from 28. If κ_S^\dagger is a Markov kernel, then
263 $\forall (\text{Id}_{Y_S} \otimes \kappa_S^\dagger) c$ is a g-disintegration. \square

264 In the reverse direction, suppose κ is such that $\kappa \pi_T = \text{Id}_X$; that is, π_T is a right inverse of κ . If
265 $\kappa \pi_S$ is not right invertible then, by definition, there is no d such that $\kappa \pi_S d \pi_T = \text{Id}_X$. However, if a
266 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
267 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](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.