Thesis Proposal Review: How Hard is a Causal Inference Problem

David Johnston

December 2, 2019

1 Introduction: Consequences of Decisions

This thesis is concerned with understanding a particular kind of decision problem: we are given a set of feasible decisions and a set of observed data, we know the potential consequences these decisions may have and we know how desirable these consequences are. We wish to develop strategies for selecting decisions that are likely to lead to favourable consequences. For example, the decisions may be a set of possible medical treatments, consequences are states of health and data are from published medical trials; we also assume that some states of health are known to be more desirable than others.

This general kind of problem seems to me to be a reasonable description of a type of problem that people often face (allowing that it may be somewhat simplified). But I need not rely only on an appeal to intuition to argue that this is an important class of problem, as decision problems of this type have a long and extensive history of study: Von Neumann and Morgenstern (1944) considers the problem of choosing between consequences directly with some means of evaluating their desirability, Weirich (2016) discusses decision problems featuring decisions, consequences and desirability but no explicit consideration of data. Wald (1950) considers the problem of selecting a favourable decision given a set of data and a desirability function, though he eschews explicitly considering consequences, and Savage (1972) develops Wald's theory to also include consequences of decisions, yielding a class of decision problems very similar to those discussed here. Many of the solutions presented by these authors have "entered the water supply" - in particular, the expected utility theory of Von Neumann and Morgenstern (1944) underpins an enormous amount of the work on decision problems of any type, and the risk functionals of Wald (1950) are fundamental to much of statistics and machine learning. Even theories that reject the particulars proposed by these authors build on the foundations laid by them - in short, the type of problem studied here is widely accepted to be a very important class of problem.

This type of problem has particular practical relevance to the field of *causal inference*. A Google Scholar search for "causal inference" found, in the top five results:

- Holland (1986) and Frangakis and Rubin (2002) discuss causal inference as the project of relating *treatments* to *responses* via *observations*. If we postulate an implicit desirability of responses, we have a decision problem of the type outlined
- Morgan and Winship (2014) provide in their opening paragraph three examples of causal problems. Two of them have clear interpretations as decision problems where decisions involve funding of charter schools and engaging in or encouraging college study, while the third is perhaps more concerned with *responsibility* and *remedy*:
  - Do charter schools increase test scores?
  - Does obtaining a college degree increase an individual's labor market earnings?
  - Did the use of a butterfly ballot in some Florida counties in the 2000 presidential election cost Al Gore votes?
- Pearl (2009a) begins with four examples of causal questions. The first appears to be part of a decision problem, while the second to fourth are questions of responsibility and remedy:
  - What is the efficacy of a given drug in a given population?
  - Whether data can prove an employer guilty of hiring discrimination?
  - What fraction of past crimes could have been avoided by a given policy?
  - What was the cause of death of a given individual, in a specific incident?
- Robins et al. (2000) is again concerned with estimating responses to treatments via observations

From this informal survey we have six out of ten example problems that correspond directly to the type of decision problem studied here. While decision problems are a substantial class of causal inference problems, we find that questions of responsibility also figure prominently. It is OK that there are other interesting causal questions; the focus on decision problems is justified by the fact that decision problems are an important class of problem in general, and also a large and important class of problems within causality in particular. We do not require that they are the *only* class of problems that causal researchers may be interested in.

One key difference between CSDT and existing popular approaches to causal inference is that we stipulate that *the set of decisions is a feature of the problem*, and does not depend in any way on how we choose to analyse the problem. Existing approaches provide "standard" objects (e.g. counterfactual random variables) or operations (e.g. intervening on the value of some random variable) which, if they are to be interpreted as decisions, impose some presuppositions on

the nature of the decisions available. Even if these presuppositions correspond to very common regularities of decision problems, we take the view that such regularities should be included as assumptions rather than be part of the language used to express the problem.

This difference is illustrated by the question of *external validity*. Given a randomised controlled trial (RCT), under ideal conditions existing causal inference approaches agree that certain causal effects can be consistently estimated. However, as reported by Deaton and Cartwright (2018):

> Trials, as is widely noted, often take place in artificial environments which raises well recognized problems for extrapolation. For instance, with respect to economic development, Drèze (J. Drèze, personal communications, November 8, 2017) notes, based on extensive experience in India, that "when a foreign agency comes in with its heavy boots and deep pockets to administer a 'treatment,' whether through a local NGO or government or whatever, there tends to be a lot going on other than the treatment." There is also the suspicion that a treatment that works does so because of the presence of the 'treators,' often from abroad, and may not do so with the people who will work it in practice.

Here, Drèze is describing the problem of determining the consequences of the "treatment in practice", and why these may differ from the "causal effects of treatment in the trial" - the question of external validity is, loosely, the question of how informative the latter are about the former. The usual approach of causal inference is to determine conditions under which the latter can be estimated and then, maybe, consider some additional assumptions that might allow for the latter estimate to inform the former. CSDT inverts the priority of these questions: the question of treatment in practice is primary and the question of causal effects in the trial may be a subproblem of interest under particular conditions.

Bareinboim and Pearl (2012) have claimed to have a complete solution to the problem of "[identifying] conditions under which causal information learned from experiments can be reused in a different environment where only passive observations can be collected", a claim made with more force in Pearl (2018). A complete solution to the transportability of causal information is *not* a claim of a complete solution to the problem of determining the effects of "treatment in practice" or the problem of making decisions with causal information. These latter problems ask when causal effects are informative about the consequences of decisions in the given problem, a question that doesn't even make sense without our insistence that decisions are a feature of the problem.

Key features (/aims - not all are realised yet) of CSDT are:

- Conceptual clarity:
  - CSDT separates of those aspects of a problem that are fixed by non-causal considerations (objectives, feasible decisions) and causal assumptions

- Unification and extension of existing approaches to causal inference for decision problems
  - Faithful translation from any existing approach to CSDT (including the derivation of key results)
  - Exact and approximate comparison of arbitrary causal theories
  - Quantification of the *difficulty* of a causal problem
  - Necessary conditions for key results
  - Novel approaches/assumptions for causal inference

the following seems like a reasonable point, but not sure where to put it right now

The core features of CSDT are that it is a new approach to causality that is strictly more capable of representing decision problems than existing approaches, and that it allows for novel and fundamental questions to be asked. However, a secondary feature of CSDT is that its statements can be clearly resolved to statements in the underlying theory of probability. This may also be true of some counterfactual approaches, but I don't think it is true of interventional graphical models. For example, Causal Bayesian Networks feature an elementary operation notated $P(\cdot|do(\mathsf{X}_k = a))$ where X_k is a random variable on some implicit sample space E. We can ask: what does $P(\cdot|do(\mathsf{X}_k = a))$ mean in more elementary terms? $do(\mathsf{X}_k = a)$ itself *looks* like a function, and the conventional interpretation of $\mathsf{X}_k = a$ is the preimage of a under X_k. Thus, $do()$ appears to be a function typed like a measure on $\mathcal{E}$ with the domain being the sigma algebra generated by all statements $\mathsf{X}_i = a$ for all X_i associated with some graph $\mathcal{G}$, which we will denote $\sigma(\underline{\otimes}_{i\in\mathcal{G}}\mathsf{X}_i)$. We might surmise that the "conditional probability" $P(\cdot|do(\mathsf{X}_k = \cdot))$ might then be the conditional probability on $\sigma(\underline{\otimes}_{i\in\mathcal{G}}\mathsf{X}_i)$. However, CBNs in general support models where $P(\cdot|do(\mathsf{X}_k = \cdot))$ is not equal to $P(\cdot|\mathsf{A})$ for any $\mathsf{A} \in \sigma(\underline{\otimes}_{\mathcal{G}}\mathsf{X}_i)$, so our attempt to parse this notation by "conventional reading" has failed.

In fact, the situation is even more dire: we may view $do(\mathsf{X}_k = a)$ as a relation between probability measures on E which is not, in general, functional – an interpretation compatible with the definitions in Pearl (2009b). If $do()$ were functional, we could define $P(\cdot|(\mathsf{X}_k = a)$ to be the element of $\Delta(\mathcal{E})$ related to P by $(\mathsf{X}_k = a)$. However, because $do(\mathsf{X}_k = a)$ is not functional, "conditioning" on $do(\mathsf{X}_k = \cdot)$ is ambiguous - does $P(\cdot|do(\mathsf{X}_k = a))$ refer to the set of probability measures related to P? A distinguished member of this set? In contrast to regular conditioning, where a similar ambiguity prevails but the ambient measure guarantees that disagreement can only happen on sets of measure zero, $P(\cdot|do(\mathsf{X}_k = a))$ can under different interpretations assign different measures to the same set. Causal Bayesian Network notational conventions suggest interpretations that do not make sense, and their meaning may be ambiguous even if we dig more deeply into the matter.

2 Definitions and key notation

We use three notations for working with probability theory. The "elementary" notation makes use of regular symbolic conventions (functions, products, sums, integrals, unions etc.) along with the expectation operator $\mathbb{E}$. This is the most flexible notation which comes at the cost of being verbose and difficult to read. Secondly, we use a semi-formal string diagram notation extending the formal diagram notation for symmetric monoidal categories Selinger (2010). Objects in this diagram refer to stochastic maps, and by interpreting diagrams as symbols we can, in theory, be just as flexible as the purely symbolic approach. However, we avoid complex mixtures of symbols and diagrams elements, and fall back to symbolic representations if it is called for. Finally, we use a matrix-vector product convention that isn't particularly expressive but can compactly express some common operations.

2.1 Standard Symbols

Symbol	Meaning
$[n]$	The natural numbers $\{1, ..., n\}$
$f : a \mapsto b$	Function definition, equivalent to $f(a) := b$
Dots appearing in function arguments: $f(\cdot, \cdot, z)$	The "curried" function $(x, y) \mapsto f(x, y, z)$
Capital letters: A, B, X	sets
Script letters: $\mathcal{A}, \mathcal{B}, \mathcal{X}$	σ-algebras on the sets A, B, X respectively
Script $\mathcal{G}$	A directed acyclic graph made up of nodes V and edg
Greek letters μ, ξ, γ	Probability measures
δ_x	The Dirac delta measure: $\delta_x(A) = 1$ if $x \in A$ and 0 oth
Capital delta: $\Delta(\mathcal{E})$	The set of all probability measures on $\mathcal{E}$
Bold capitals: $\mathbf{A}$	Markov kernel $\mathbf{A} : X \times \mathcal{Y} \to [0, 1]$ (stochastic maps
Subscripted bold capitals: $\mathbf{A}_x$	The probability measure given by the curried Markov kern
$A \to \Delta(\mathcal{B})$	Markov kernel signature, treated as equivalent to $A \times \mathcal{B}$
$\mathbf{A} : x \mapsto \nu$	Markov kernel definition, equivalent to $\mathbf{A}(x, B) = \nu(B)$ f
Sans serif capitals: A, X	Measurable functions; we will also call them random va
Π_{X}	The Markov kernel associated with the function X: $\Pi_{\mathsf{X}} \equiv$
$[\mathsf{A}\|\mathsf{B}]_\nu$	The conditional probability (disintegration) of A given B
$\nu\Pi_{\mathsf{X}}$	The marginal distribution of X under ν

2.2 Probability Theory

Given a set A, a σ-algebra $\mathcal{A}$ is a collection of subsets of A where

- $A \in \mathcal{A}$ and $\emptyset \in \mathcal{A}$
- $B \in \mathcal{A} \implies B^C \in \mathcal{A}$
- $\mathcal{A}$ is closed under countable unions: For any countable collection $\{B_i | i \in Z \subset \mathbb{N}\}$ of elements of $\mathcal{A}$, $\cup_{i \in Z} B_i \in \mathcal{A}$

A measurable space $(A, \mathcal{A})$ is a set A along with a σ-algebra $\mathcal{A}$. Sometimes the sigma algebra will be left implicit, in which case A will just be introduced as a measurable space.

Common σ algebras For any A, $\{\emptyset, A\}$ is a σ-algebra. In particular, it is the only sigma algebra for any one element set $\{*\}$.

For countable A, the power set $\mathcal{P}(A)$ is known as the discrete σ-algebra.

Given A and a collection of subsets of $B \subset \mathcal{P}(A)$, $\sigma(B)$ is the smallest σ-algebra containing all the elements of B.

Let T be all the open subsets of $\mathbb{R}$. Then $\mathcal{B}(\mathbb{R}) := \sigma(T)$ is the *Borel σ-algebra* on the reals. This definition extends to an arbitrary topological space A with topology T.

A *standard measurable set* is a measurable set A that is isomorphic either to a discrete measurable space A or $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For any A that is a complete separable metric space, $(A, \mathcal{B}(A))$ is standard measurable.

Given a measurable space $(E, \mathcal{E})$, a map $\mu : \mathcal{E} \to [0, 1]$ is a *probability measure* if

- $\mu(E) = 1$, $\mu(\emptyset) = 0$
- Given countable collection $\{A_i\} \subset \mathcal{E}$, $\mu(\cup_i A_i) = \sum_i \mu(A_i)$

Write by $\Delta(\mathcal{E})$ the set of all probability measures on $\mathcal{E}$.

Given a second measurable space $(F, \mathcal{F})$, a *stochastic map* or *Markov kernel* is a map $\mathbf{M} : E \times \mathcal{F} \to [0, 1]$ such that

- The map $\mathbf{M}(\cdot; A) : x \mapsto \mathbf{M}(x; A)$ is $\mathcal{E}$-measurable for all $A \in \mathcal{F}$
- The map $\mathbf{M}_x : A \mapsto \mathbf{M}(x; A)$ is a probability measure on F for all $x \in E$

Extending the subscript notation above, for $\mathbf{C} : X \times Y \to \Delta(\mathcal{Z})$ and $x \in X$ we will write $\mathbf{C}_x$ for the "curried" map $y \mapsto \mathbf{C}_{x,y}$.

The map $x \mapsto \mathbf{M}_x$ is of type $E \to \Delta(\mathcal{F})$. We will abuse notation somewhat to write $\mathbf{M} : E \to \Delta(\mathcal{F})$, which captures the intuition that a Markov kernel maps from elements of E to probability measures on $\mathcal{F}$. Note that we "reverse" this idea and consider Markov kernels to map from elements of $\mathcal{F}$ to measurable functions $E \to [0, 1]$, an interpretation found in Clerc et al. (2017), but (at this stage) we don't make use of this interpretation here.

Given an indiscrete measurable space $(\{*\}, \{\{*\}, \emptyset\})$, we identify Markov kernels $\mathbf{N} : \{*\} \to \Delta(\mathcal{E})$ with the probability measure $\mathbf{N}_*$. In addition, there is a unique Markov kernel $* : E \to \Delta(\{\{*\}, \emptyset\})$ given by $x \mapsto \delta_*$ for all $x \in E$ which we will call the "discard" map

2.3 Product Notation

We can use a notation similar to matrix-vector products to represent operations with Markov kernels. Probability measures $\mu \in \Delta(\mathcal{X})$ can be read as row vectors, Markov kernels as matrices and measurable functions $\mathsf{T} : Y \to T$ as column

vectors. Defining $\mathbf{M} : X \to \Delta(\mathcal{Y})$ and $\mathbf{N} : Y \to \Delta(\mathcal{Z})$, the measure-kernel product $\mu\mathbf{A}(G) := \int \mathbf{A}_x(G)d\mu(x)$ yields a probability measure $\mu\mathbf{A}$ on $\mathcal{Z}$, the kernel-kernel product $\mathbf{MN}(x; H) = \int_Y \mathbf{B}(y; H)d\mathbf{A}_x$ yields a kernel $\mathbf{MN} : X \to \Delta(\mathcal{Z})$ and the kernel-function product $\mathbf{A}\mathsf{T}(x) := \int_Y \mathsf{T}(y)d\mathbf{A}_x$ yields a measurable function $X \to T$. Kernel products are associative (Çinlar, 2011).

The tensor product $(\mathbf{M} \otimes \mathbf{N})(x, y; G, H) := \mathbf{M}(x; G)\mathbf{N}(y; H)$ yields a kernel $(\mathbf{M} \otimes \mathbf{N}) : X \times Y \to \Delta(\mathcal{Y} \otimes \mathcal{Z})$.

2.4 String Diagrams

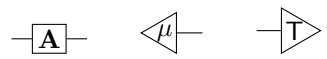
Some constructions are unwieldly in product notation; for example, given $\mu \in \Delta(\mathcal{E})$ and $\mathbf{M} : E \to (\mathcal{F})$, it is not straightforward to construct a measure $\nu \in \Delta(\mathcal{E} \otimes \mathcal{F})$ that captures the "joint distribution" given by $A \times B \mapsto \int_A \mathbf{M}(x; B)d\mu$.

Such constructions can, however, be straightforwardly captured with string diagrams, a notation developed for category theoretic probability. Cho and Jacobs (2019) also provides an extensive introduction to the notation discussed here.

Some key ideas of string diagrams:

- Basic string diagrams can always be interpreted as a mixture of kernel-kernel products and tensor products of Markov kernels
  - Extended string diagrams can be interepreted as a mixture of kernel-kernel products, kernel-function products, tensor products of kernels and functions and scalar products
- String diagrams are the subject of a coherence theorem: taking a string diagram and applying a planar deformation yields a string diagram that represents the same kernel (Selinger, 2010). This also holds for a number of additional transformations detailed below

A kernel $\mathbf{A} : X \to \Delta(\mathcal{Y})$ is written as a box with input and output wires, probability measures $\mu \in \Delta(\mathcal{X})$ are written as triangles "closed on the left" and measurable functions $\mathsf{T} : Y \to T$ as triangles "closed on the right".



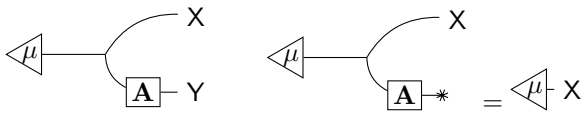
(1)

We canonically regard a probability measure $\mu \in \Delta(\mathcal{E})$ to be a Markov kernel $\mu : \{*\} \to \Delta(\mathcal{E})$. This allows for the definition of "basic" string diagrams for which Markov kernels are the only building blocks. Such a definition isn't possible for measurable functions. Suppose we try supposing a measurable function $f : E \to \mathbb{R}$ is "really" a Markov kernel $f : E \to \Delta(\{*\})$. The problem is, depending on the input $x \in E$, $f(x)$ is not in general equal to 1, and so this supposition fails. This is the reason we require an "extended" string diagram notation if we wish to incorporate functions and expectations.

The identity $\mathbf{Id} : X \to \Delta(X)$ is the Markov kernel $x \mapsto \delta_x$, which we represent with a bare wire. The copy map $\curlyvee : X \to \Delta(X \times X)$ is the Markov kernel

$x \mapsto \delta_{(x,x)}$. For $\mathbf{A} : X \to \Delta(Y)$ and $\mathbf{B} : X \to \Delta(Z)$, $\curlyvee(A \otimes B)_x = A_x \otimes B_x$. The discard map $*$ is the Markov kernel $X \to \{\#\}$ given by $x \mapsto \delta_\#$, where $\#$ is some one element set. Placing boxes side by side with connected wires corresponds to taking kernel products as defined above.

We will apply these notions to a couple of example constructions. Given $\mu \in \Delta(X), \mathbf{A} : X \to \Delta(Y)$ as before, the joint distribution on $X \times Y$ given by $\nu(A \times B) = \int_A A(x; B) d\mu(x)$ is given in string diagram on the left of 2. Marginalisation is accomplished with the discard map $*$; hence $\mu\curlyvee(\mathbf{Id} \otimes \mathbf{A}*) = \mu$; this is shown on the right of 2



(2)

2.5 Category theoretic probability and string diagrams

We now turn to the category theoretic underpinnings of string diagrams and introduce a number of additional axioms useful in proofs using the notation.

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

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

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

Of particular interest is the Kleisli category of the monads above. The Kleisli C_T category of a monad T on category C is the category with the same objects and the morphisms $X \to Y$ in C_T is the set of morphisms $X \to TY$ in C. Thus the morphisms $X \to Y$ in the Kleisli category $\mathbf{Set}_\mathcal{D}$ are morphisms $X \to \mathcal{D}(Y)$

in **Set**, i.e. stochastic matrices, and in the Kleisli category $\mathbf{Meas}_{\mathcal{G}}$ we have Markov kernels. Composition of arrows in the Kleisli categories correspond to Matrix products and "kernel products" respectively.

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

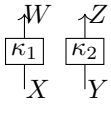
Diagrams for symmetric monoidal categories consist of wires with arrows, boxes and a couple of special symbols. The identity object (which we identify with the set $\{*\}$) is drawn as nothing at all $\{*\} :=$ and identity maps are drawn as bare wires:

$\mathrm{Id}_X :=$  (3)

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

$\mu :=$  $\kappa :=$  (4)

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 $\kappa_2 : Y \to Z$:

$(X \times Y, \mathcal{X} \otimes \mathcal{Y}) :=$ $\kappa_1 \otimes \kappa_2 :=$  (5)

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

$\kappa_1\kappa_2(x; A) = \int_Y \kappa_2(y; A)\kappa_1(x; dy) :=$  (6)

Symmetric monoidal categories have the following coherence theorem(Selinger, 2010):

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

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

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

$$\text{Erase} = \mathbb{1}_X := \quad \text{Copy} = x \mapsto \delta_{x,x} := \tag{7}$$

$$= \quad := \tag{8}$$

$$= \quad = \tag{9}$$

$$= \tag{10}$$

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

$$\kappa \quad X \quad = \quad X \tag{11}$$

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

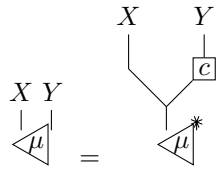
Thus when manipulating diagrams representing Markov kernels in particular (and, importantly, not more general symmetric monoidal categories) diagram isomorphism also includes applications of 8, 9, 10 and 11.

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).

2.6 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).

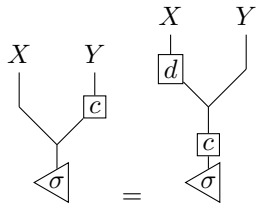
Given a distribution $\mu : \{*\} \to X \otimes Y$, a disintegration $c : X \to Y$ is a Markov kernel that satisfies



(12)

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



(13)

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.

2.7 Generalisations

Cho and Jacobs (2019) make use of a larger "CD" category by dropping 11. 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.

2.8 Key questions for Causal Theories

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

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

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

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

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

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

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 10

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

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

We 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

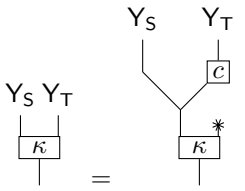
generalised disintegrations : Of key importance to our work is generalising the notion of disintegration (and possibly Bayesian inversion) to general kernels $X \to Y$ rather than restricting ourselves to probability distributions $\{*\} \to Y$. We will define generalised disintegrations as a straightforward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

Definition 2.2 (Label signatures)**.** If a kernel $\kappa : X \to \Delta(Y)$ can be represented by a diagram G with incoming wires $\mathsf{X_1}, ...\mathsf{X_n}$ and outgoing wires $\mathsf{Y_1}, ..., \mathsf{Y_m}$, 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 2.3 (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 $c : \mathsf{Y_S} \dashrightarrow \mathsf{Y_T}$ and we have the identity (omitting incoming wire labels):



(14)

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

$$\kappa : \begin{cases} 1 \mapsto \delta_1 \otimes \delta_1 \\ 0 \mapsto \delta_1 \otimes \delta_0 \end{cases} \tag{15}$$

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

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

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

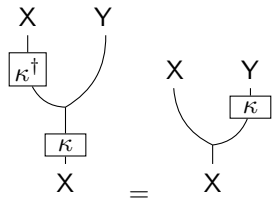
Suppose $x \notin A$ and let $B_\epsilon = \kappa_A^{\dagger -1}((\epsilon,1])$ for some $\epsilon > 0$. We have $\int_Y \kappa^\dagger(y;A)\kappa(x;dy) = 0 \geq \epsilon\kappa(x;B_\epsilon)$. Thus for any $\epsilon > 0$ we have $\kappa(x;B_\epsilon) = 0$. Consider the set $B_0 = \kappa_A^{\dagger -1}((0,1])$.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

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

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

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

Lemma 2.5. *Given $\kappa : X \to \Delta(Y)$ and a right inverse $\kappa^\dagger$, we have*



(16)

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

$$L(x; A \times B) = \int_Y \int_{Y \times Y} \mathrm{Id}_Y \otimes \kappa_S^\dagger(y, y'; A \times B)\delta_{(z,z)}(dy \times dy')\kappa\pi_S(x; dz) \tag{17}$$

$$= \int \mathrm{Id}_Y \otimes \kappa^\dagger(z, z; A \times B)\kappa\pi_S(x; dz) \tag{18}$$

$$= \int \delta_z(A)\kappa_S^\dagger(z; B)\kappa\pi_S(x; dz) \tag{19}$$

$$= \int_A \kappa_S^\dagger(z; B)\kappa\pi_S(x; dz) \tag{20}$$

$$= \delta_x(B)\kappa\pi_S(x; A) \tag{21}$$

Where 21 follows from Lemma 2.4.

$$R(x; A \times B) = \int \delta_{(x,x)}(dy \times dy')\kappa\pi_S \otimes \mathrm{Id}_X(y, y'; A \times B) \tag{22}$$

$$= \kappa\pi_S(x; A)\delta_x(B) \qquad = L \tag{23}$$

□

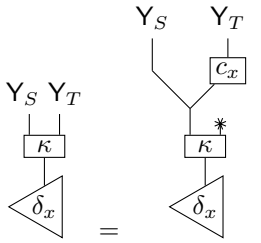
Theorem 2.6. *Given countable X and standard measurable Y, $n, m \in \mathbb{N}$, $S, T \subset [m]$, κ with label signature $\mathsf{X}_{[n]} \dashrightarrow \mathsf{Y}_{[m]}$ a g-disintegration exists from S to T if $\kappa\pi_S$ is right-invertible*

via a Markov kernel

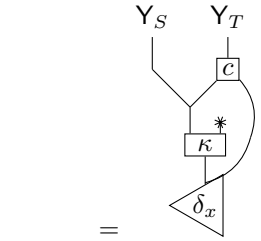
Proof. In addition, as R is a composition of Markov kernels, and hence a Markov kernel itself, L must also be a Markov kernel even if $\kappa^\dagger$ is not.

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

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

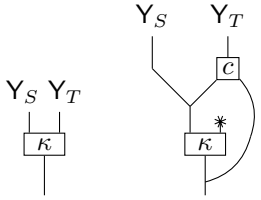


(24)



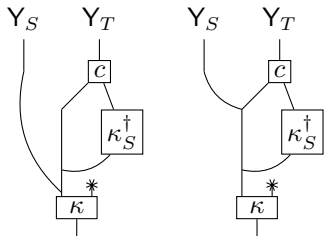
(25)

Which implies

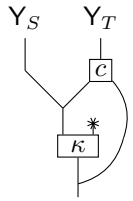


(26)

Finally, we have



$$=$$ (27)



$$=$$ (28)

Where the first line follows from 9 and the second line from 16. 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

Elias Bareinboim and Judea Pearl. Transportability of Causal Effects: Completeness Results. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, July 2012. URL `https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5188`.

Erhan Çinlar. *Probability and Stochastics*. Springer, 2011.

Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, August 2019. ISSN 0960-1295, 1469-8072. doi: 10.1017/S0960129518000488.

Florence Clerc, Fredrik Dahlqvist, Vincent Danos, and Ilias Garnier. Pointless learning. *20th International Conference on Foundations of Software Science and Computation Structures (FoSSaCS 2017)*, March 2017. doi: 10.1007/978-3-662-54458-7_21. URL `https://www.research.ed.ac.uk/portal/en/publications/pointless-learning(694fb610-69c5-469c-9793-825df4f8ddec).html`.

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

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.

Constantine E. Frangakis and Donald B. Rubin. Principal Stratification in Causal Inference. *Biometrics*, 58(1):21–29, 2002. ISSN 1541-0420. doi: 10.1111/j.0006-341X.2002.00021.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2002.00021.x.

Paul W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960, December 1986. ISSN 0162-1459. doi: 10.1080/01621459.1986.10478354. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478354.

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.

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

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009a. ISSN 1935-7516. doi: 10.1214/09-SS057.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009b.

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

James M. Robins, Miguel Ángel Hernán, and Babette Brumback. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, 11(5):550, September 2000. ISSN 1044-3983. URL `https://journals.lww.com/epidem/Fulltext/2000/09000/Marginal_Structural_Models_and_Causal_Inference_in.11.aspx/`.

Leonard J. Savage. *Foundations of Statistics*. Dover Publications, New York, revised edition edition, June 1972. ISBN 978-0-486-62349-8.

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.

J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Theory of games and economic behavior. Princeton University Press, Princeton, NJ, US, 1944.

Abraham Wald. *Statistical decision functions*. Statistical decision functions. Wiley, Oxford, England, 1950.

Paul Weirich. Causal Decision Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016. URL `https://plato.stanford.edu/archives/win2016/entries/decision-causal/`.

Appendix: