Thesis Proposal Review: How Hard is a Causal Inference Problem

David Johnston

December 10, 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. While the approach built in this thesis may have eventual applications to questions of responsibility and other causal questions, we take the attitude that in the worst case it will only be appliccable to decision problems and this is a large and important enough class of problems that a clearer understanding of just these problems will still be very valuable.

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
$\mathbf{F}_{\mathsf{X}}$	The Markov kernel associated with the function X: $\mathbf{F}_{\mathsf{X}} \equiv$
$\mathbf{N}_{\mathsf{A}\|\mathsf{B}}$	The conditional probability (disintegration) of A given B
$\nu\mathbf{F}_{\mathsf{X}}$	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 (which are only elements of the "extended" notation) $\mathsf{T} : Y \to T$ as triangles "closed on the right". We label all output wires with unique sans serif letters X, Y which can be understood as defining a random variable on some canonical space which corresponds to the labeled wire. We also adopt the convention that a random variable X_α (with optional index) takes values in the space X. Input wires are labeled only with the spaces X, Y, Z, and we don't associate random variables with input wires. See Paragraph 2.4 for a more detailed explanation of random variables in Causal Statistical Decision Theory.

For $\mathbf{A} : X \to \Delta(\mathcal{Y})$, $\mu \in \Delta(\mathcal{X})$ and $f : X \to W$:

(1)

Basic and extended notation 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 by analogy with the example probability measures and try to identify a measurable function $f : E \to \mathbb{R}$ with a Markov kernel $f' : E \times \{*\} \to \mathbb{R}$. For $x \in E$ we cannot generally have both $f'(x, *) = 1$ and $f'(x, *) = f(x)$, and so this attempt fails. This lack of normalisation is the reason we require an "extended" string diagram notation if we wish to incorporate functions and expectations which allows for the representation of scalars.

Elementary operations We can compose Markov kernels with appropriate spaces - the equivalent operation of the "matrix products" of product notation. Given $\mathbf{M} : E \to \Delta(\mathcal{F})$ and $\mathbf{N} : F \to \Delta(\mathcal{G})$, we have

$$\mathbf{MN} := \quad E \text{ —[M]—[N]— } \mathsf{G} \tag{2}$$

Probability measures are distinguished in that that they only admit "right composition" while functions only admit "left composition". For $\mu \in \Delta(\mathcal{E})$, $h : F \to X$:

$$\mu\mathbf{M} := \quad \langle\mu \text{—[M]— } \mathsf{G} \tag{3}$$

$$\mathbf{M}f := \quad E \text{ —[M]— } \rangle f \tag{4}$$

We can also combine Markov kernels using tensor products, which we represent with vertical juxtaposition. For $\mathbf{O} : G \to \Delta(\mathcal{H})$:

$$\mathbf{M} \otimes \mathbf{N} := \quad \begin{array}{l} E \text{ —[M]— } \mathsf{F} \\ G \text{ —[O]— } \mathsf{H} \end{array} \tag{5}$$

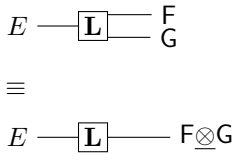
Product spaces can be represented either by two parallel wires or a single wire that "carries" the entire space:

$$X \times Y \cong \mathrm{Id}_X \otimes \mathrm{Id}_Y := \quad \begin{array}{l} X \text{ — } \mathsf{X} \\ Y \text{ — } \mathsf{Y} \end{array} \tag{6}$$

$$= \quad X \times Y \text{ —— } \mathsf{X} \underline{\otimes} \mathsf{Y} \tag{7}$$

The notation $\mathsf{X} \underline{\otimes} \mathsf{Y}$ will be explained in paragraph **??**.

Because a product space can be represented by parallel wires, a kernel $\mathbf{L} : E \to \Delta(\mathcal{F} \otimes \mathcal{G})$ can be written using either two parallel output wires or a single output wire:



(8)

$\equiv$ (9)

(10)

Markov kernels with special notation A number of Markov kernels are given special notation distinct from the generic "box" representation above. These special representations facilitate intuitive graphical interpretations.

The identity kernel $\mathbf{Id} : X \to \Delta(X)$ maps a point x to the measure δ_x that places all mass on the same point:

$$\mathbf{Id}_x : x \mapsto \delta_x \equiv \quad X \text{ — } \mathsf{X} \tag{11}$$

The copy map $\curlyvee : X \to \Delta(X \times X)$ maps a point x to two identical copies of x:

$$\curlyvee : x \mapsto \delta_{(x,x)} \equiv \quad X \text{ —(} \begin{matrix} \mathsf{X}_1 \\ \mathsf{X}_2 \end{matrix} \tag{12}$$

Note that we give output wires unique labels when they share a space.

The discard map $* : X \to \Delta(\{*\})$ maps every input to δ_* which is effectively mapping every input to 1

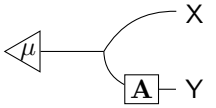
$$* : x \mapsto \delta_* \equiv \text{—}* \tag{13}$$

The swap map $\sigma : X \times Y \to \Delta(\mathcal{Y} \otimes \mathcal{X})$ swaps its inputs:

$$\sigma := (x, y) \to \delta_{(y,x)} \equiv \quad \begin{matrix} Y \\ X \end{matrix} \times \begin{matrix} \mathsf{X} \\ \mathsf{Y} \end{matrix} \tag{14}$$

Before introducing key rules of manipulation permitted by string diagrams, we will illustrate the correspondence between the three notations with a few simple examples. Given $\mu \in \Delta(X), \mathbf{A} : X \to \Delta(Y)$ and $A \in \mathcal{X}$, $B \in \mathcal{Y}$, the following correspondences hold, where we express the same object in elementary notation, product notation and string notation respectively:

$$\nu := A \times B \mapsto \int_A A(x; B) d\mu(x) \equiv \mu\curlyvee(\mathbf{Id}_X \otimes \mathbf{A}) \equiv \tag{15}$$

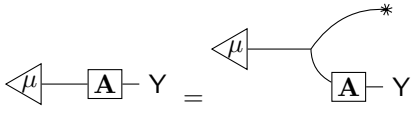


Where the resulting object is a probability measure $\nu \in \Delta(\mathcal{X} \otimes \mathcal{Y})$. Note that the elementary notation requires a function definition here, while the product and string notations can represent the measure without explicitly addressing its action on various inputs and outputs. Cho and Jacobs (2019) calls this construction "integrating $\mathbf{A}$ with respect to μ".

Define the marginal $\nu_Y \in \Delta(\mathcal{Y}) : B \mapsto \nu(X \times B)$ for $B \in \mathcal{Y}$ and similarly for ν_X. We can then express the result of marginalising 15 over X in our three separate notations as follows:

$$\nu_Y(B) = \nu(X \times B) = \int_X A(x; B) d\mu(x) \tag{16}$$

$$\nu_Y = \mu\mathbf{A} = \mu^{\curlyvee}(\mathbf{Id}_X \otimes \mathbf{A})(* \otimes \mathbf{Id}_Y) \tag{17}$$

$\nu_Y =$  (18)

The elementary notation 16 makes the relationship between ν_Y and ν explicit and, again, requires the action on each event to be defined. The product notation 17 is, in my view, the least transparent but also the most compact in the form $\mu\mathbf{A}$, and does not demand the explicit definition of how ν_Y treats every event. The graphical notation is the least compact in terms of space taken up on the page, but unlike the product notation it shows a clear relationship to the graphical construction in15, and displays a clear graphical logic whereby marginalisation corresponds to "cutting off branches". Like product notation, it also allows for the definition of derived measures such as ν_Y without explicit definition of the handling of all events. It also features a much smaller collection of symbols than does elementary notation.

String diagrams often achieve a good balance between interpretational transparency, expressive power and symbol economy.

Random Variables We take a slightly nonstandard view of random variables. Random variables are typically defined to be measurable functions on a probability space $\langle \Omega, \mathcal{F}, \mu \rangle$ (Çinlar, 2011). With this definition, a random variable $\mathsf{X} : \Omega \to E$ has a canonical probability measure given by the pushforward of μ: for all $A \in \mathcal{E}$ $\mathsf{X}_{\#}\mu(A) = \mu(\mathsf{X}^{-1}(A))$.

We take a random variable to be a measurable function on a *kernel space* $\langle F, \mathcal{F}, \mathbf{M}, G \rangle$ where $\mathbf{M} : G \to \Delta(\mathcal{F})$ is a Markov kernel. A random variable $\mathsf{X} : F \to X$ has a probability distribution only relative to some argument measure $\nu \in \Delta(\mathcal{G})$. Because of this, we cannot in general unambiguously talk about "the" distribution of a given random variable. We avoid $P(\mathsf{X})$ type notation to avoid ambiguity in this regard.

The reason for this choice is that we use random variables to model quantities with distributions that depend on the decision maker's choices. On the other hand, we don't regard the choices themselves to be modeled by random variables - a commitement to the notion that "choosing" is somehow different to "stochasticity".

Evidential decision theory, as defended by Jeffrey (1981), holds that is is proper to consider choices to be random variables, though doing so rigorously may necessitate a theory that allows for the assignment of probabilities to the outcomes of mathematical deliberations such as the theory of *logical induction* introduced in Garrabrant et al. (2017). Understanding the relationship between choices and stochastic processes is a deep, interesting and difficut question, and one we sidestep by presuming that we can address nearly all common decision problems while disregarding modelling whatever process gives rise to choices. The resulting decision theory is structurally similar to *causal decision theory* (Lewis, 1981).

This definition of random variables permits the convention of identifying every output wire of a string diagram with a random variable.

Definition 2.1 (Functional kernel)**.** Given a measurable function $\mathsf{X} : F \to X$, define the functional kernel $\mathbf{F}_{\mathsf{X}} : F \to \Delta(\mathcal{X})$ to the the Markov kernel $a \mapsto \delta_{\mathsf{X}(a)}$ for all $a \in X$.

Lemma 2.2 (Pushforward measures and functional kernels)**.** *Given a kernel space $\langle F, \mathcal{F}, \mathbf{M}, G\rangle$ and a random variable $\mathsf{X} : F \to X$, for any prior $\mu \in \Delta(\mathcal{G})$ the pushforward $\mathsf{X}_{\#}\mu\mathbf{A} = \mu\mathbf{A}\mathbf{F}_{\mathsf{X}}$.*

Proof. For all $B \in \mathcal{F}$:

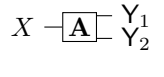
$$(\mathsf{X})_{\#}\mu\mathbf{A}(B) = \mu\mathbf{A}(\mathsf{X}^{-1}(B)) \tag{19}$$

$$= \int_F \delta_{\mathsf{X}(a)}(B) d\mu\mathbf{A}(a) \tag{20}$$

$$= \mu\mathbf{A}\mathbf{F}_{\mathsf{X}}(B) \tag{21}$$

□

Example 2.3 (Wire names to random variables)**.** Suppose we have a Markov kernel $\mathbf{A} : X \to \Delta(\mathcal{Y} \otimes \mathcal{Y})$:

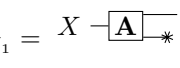


(22)

Define $\mathsf{Y}_1' : Y \times Y \to Y$ by the projection map $\mathsf{Y}_1' : (y_1, y_2) \mapsto y_1$ and $\mathsf{Y}_2' : Y \times Y \to Y$ by the projection $\mathsf{Y}_2' : (y_1, y_2) \mapsto y_2$. Given any prior $\mu \in \Delta(\mathcal{X})$, let $(\mathsf{Y}_1')_{\#}\mu\mathbf{A}$ be the pushforward of Y_1' by $\mu\mathbf{A}$. Then $\mathbf{F}_{\mathsf{Y}_1'} : Y \times Y \to Y$ will be given by $a \mapsto \delta_{\mathsf{Y}_1'(a)}$.

Define $\Pi_{\mathsf{Y}_1} : Y \times Y \to \Delta(\mathcal{Y})$ by $\Pi_{\mathsf{Y}_1} = \mathrm{Id}_Y \otimes *$. Π_{Y_1} is the Markov kernel that marginalises over the second argument; i.e. it marginalises over the wire named Y_2. Graphically:

$\mathbf{A}\Pi_{\mathsf{Y}_1} =$



(23)

Note that for all $(y_1, y_2) \in Y \times Y$, $(\Pi_1)_{y_1,y_2} = \delta_{y_1} = \delta_{\mathsf{Y}_1'(y_1,y_2)}$. That is, $\Pi_{\mathsf{Y}_1} = F_{\mathsf{Y}_1'}$ and so $(\mathsf{Y}_1')_{\#}\mu\mathbf{A} = \mu\mathbf{A}\Pi_{\mathsf{Y}_1}$.

Furthermore, define the joint distribution of Y'_1 and Y'_2 by $(\mathsf{Y}'_1 \underline{\otimes} \mathsf{Y}'_2)_{\#}\mu\mathbf{A}(B \times C) = \mu\mathbf{A}(\mathsf{Y}'^{-1}_1(B) \cap \mathsf{Y}'^{-1}_2(C))$ for all $B, C \in \mathcal{Y}$. Then, defining $\Pi_{\mathsf{Y}_1 \otimes \mathsf{Y}_2} = \mathrm{Id}_Y \otimes \mathrm{Id}_Y = \mathrm{Id}_{Y \times Y}$:

$$
\begin{aligned}
(\mathsf{Y}'_1 \underline{\otimes} \mathsf{Y}'_2)_{\#}\mu\mathbf{A}(B \times C) &= \mu\mathbf{A}(\mathsf{Y}'^{-1}_1(B) \cap \mathsf{Y}'^{-1}_2(C)) && (24)\\
&= \int_{Y \times Y} \delta_{\mathsf{Y}'_1(y_1,y_2)}(B)\delta_{\mathsf{Y}'_2(y_1,y_2)}(C) d\mu\mathbf{A}(y_1, y_2) && (25)\\
&= \int_{B \times C} d\mu\mathbf{A}(y_1, y_2) && (26)\\
&= \mu\mathbf{A}(B \times C) && (27)\\
&= \mu\mathbf{A}\Pi_{\mathsf{Y}_1 \otimes \mathsf{Y}_2}(B \times C) && (28)
\end{aligned}
$$

That is, for any prior μ, the joint distribution of Y'_1 and Y'_2 under $\mu\mathbf{A}$ is "carried" by the wires labeled Y_1 and Y_2, and the marginal distribution of Y_1 is "carried" by the wire labeled Y_1 alone. It's in this sense that we identify the random variable Y'_1 with Y_1. In general, given a Markov kernel with output space $\prod_{i \in [n]} X_i$, we can identify the j-th output wire with the random variable given by the projection map $\pi_j : (x_1, ...x_j, ...x_n) \mapsto x_j$.

Because of this identification, we treat wire names as random variables.

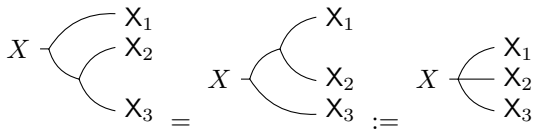
2.4.1 Rules for String Diagrams

todo:

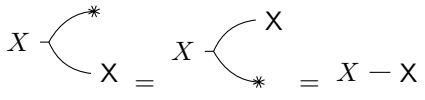
- Disintegration, Bayesian inversion
- Functional generalisation
- Conditioning
- Infinite copy map
- De Finetti's representation theorem

There are a relatively small number of manipulation rules and a number of special constructions that are useful for string diagrams.

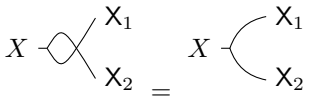
Axioms of Symmetric Monoidal Categories Recalling the unique Markov kernels defined above, the following equivalences, known as the *commutative comonoid axioms*, hold among string diagrams:



(29)

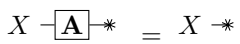


(30)



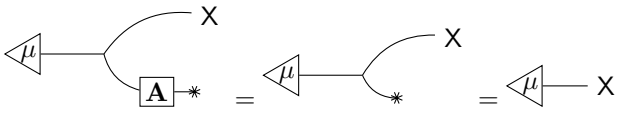
(31)

The discard map ⁎ can "fall through" any Markov kernel:



(32)

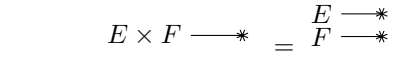
Combining 30 and 32 we can derive the following: integrating $\mathbf{A} : X \to \Delta(\mathcal{Y})$ with respect to $\mu \in \Delta(\mathcal{X})$ and then discarding the output of $\mathbf{A}$ leaves us with μ:



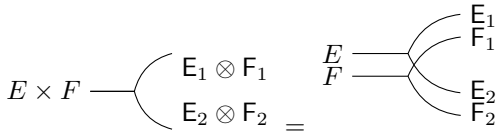
(33)

In elementary notation, this is equivalent to the fact that, for all $B \in \mathcal{X}$, $\int_B \mathbf{A}(x; B) d\mu(x) = \mu(B)$.

The following additional properties hold for ⁎ and ⋎:

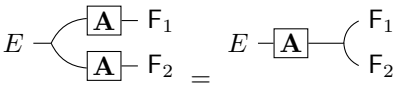


(34)



(35)

A key fact that *does not* hold in general is



(36)

In fact, it holds only when $\mathbf{A}$ is a *deterministic* kernel.

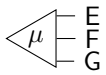
Definition 2.4 (Deterministic Markov kernel)**.** A *deterministic* Markov kernel $\mathbf{A} : E \to \Delta(\mathcal{F})$ is a kernel such that $\mathbf{A}_x(B) \in \{0, 1\}$ for all $x \in E$, $B \in \mathcal{F}$.

Theorem 2.5 (Copy map commutes for deterministic kernels (Fong, 2013))**.** *Equation 36 holds iff* $\mathbf{A}$ *is deterministic.*

Disintegraion and Bayesian Inversion We use *disintegration* to define a notion of conditional probability. It is not identical to the standard definition of conditional probability one can find in, for example, Çinlar (2011), but each can be recovered from the other.

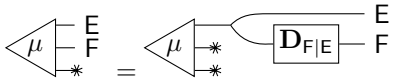
We'll proceed from an example to a general definition.

Example 2.6 (Disintegration)**.** Given a probability measure $\mu \in \Delta(\mathcal{E} \otimes \mathcal{F} \otimes \mathcal{G})$:



(37)

A Markov kernel $\mathbf{D}_{\mathsf{F}|\mathsf{E}}$ is a $\mathsf{F}|\mathsf{E}$ ("F on E")-disintegration of μ if



(38)

Equation 38 echoes the familiar property of conditional probability $P(A \cap B) = P(A|B)P(B)$; in elementary notation it states that for and disintegration $\mathbf{D}_{\mathsf{F}|\mathsf{E}}$ and all $A \in \mathcal{E}$, $B \in \mathcal{F}$, $\mu(A \times B) = \int_A \mathbf{D}_{\mathsf{F}|\mathsf{E}}(x; B) d\mu_{\mathsf{E}}(x)$ where $\mu_{\mathsf{E}} := \mu \mathbf{F}_{\mathsf{E}}$ is the marginal distribution of E under μ.

Example 2.6 defines disintegration given a probability measure μ and a pair of random variables E and F that are adapted to the product structure of the output space of μ, a product structure that allows us to draw a diagram for μ featuring two wires as outputs. There are three extensions to this definition that are desirable:

1. We would like to replace individual wires E and F with arbitrary sets of wires
2. We would like to be able to disintegrate a probability mesure with respect to arbitrary random variables, not just sets that are adapted to the product structure of the output space
3. We would like to define disintegration for arbitrary Markov kernels rather than probability measures only

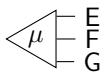
As we show, we can associate any set of wires with a random variable, so the first item is solved by a solution to the second.

Definition 2.7 (Random Variable Groups)**.** Given a measurable space $(E, \mathcal{E})$ and two random variables X, Y with shared domain E, the *random variable group* of X and Y is defined as the function $\curlyvee(\mathsf{X} \otimes \mathsf{Y})$, for which we use the shorthand $\mathsf{X} \underline{\otimes} \mathsf{Y}$.

For n random variables $\mathsf{X}_1, ...\mathsf{X}_n$, the random variable group $\underline{\otimes}_{i \in [n]} \mathsf{X}_i$ is given by $(\mathsf{X}_1 \underline{\otimes} \mathsf{X}_2) \underline{\otimes} \mathsf{X}_3$

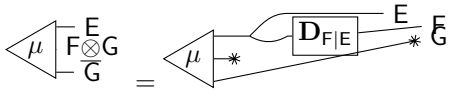
Lemma 2.8 ($\underline{\otimes}$ is associative)**.**

Example 2.9 (Disintegration - groups of variables)**.** Given a probability measure $\mu \in \Delta(\mathcal{E} \otimes \mathcal{F} \otimes \mathcal{G})$ as before:



(39)

Write
A Markov kernel $\mathbf{D}_{\mathsf{F}|\mathsf{E}\underline{\otimes}\mathsf{G}}$ is a $\mathsf{F}\underline{\otimes}\mathsf{G}|\mathsf{E}$-disintegration of μ if



(40)

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`.

Scott Garrabrant, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, and Jessica Taylor. Logical Induction. *arXiv:1609.03543 [cs, math]*, December 2017. URL http://arxiv.org/abs/1609.03543. arXiv: 1609.03543.

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.

Richard Jeffrey. The logic of decision defended. *Synthese*, 48(3):473–492, September 1981. ISSN 1573-0964. doi: 10.1007/BF01063989. URL https://doi.org/10.1007/BF01063989.

David Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59(1): 5–30, March 1981. ISSN 0004-8402. doi: 10.1080/00048408112340011. URL https://doi.org/10.1080/00048408112340011.

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: