Thesis Proposal Review: How Hard is a Causal Inference Problem

David Johnston

February 11, 2020

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 the standard notation for 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 $\mathbf{M} : X \to \Delta(\mathcal{Y})$ 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". For this introduction we will label wires with the names of their corresponding spaces, but in practice we will usually name them with corresponding *random variables*, though additional care is required when using random variables as labels (see paragraph 2.4.2).

For $\mathbf{M} : X \to \Delta(\mathcal{Y})$, $\mu \in \Delta(\mathcal{X})$ 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 $\mathbf{M} : X \to \Delta(\mathcal{Y})$ and $\mathbf{N} : Y \to \Delta(\mathcal{Z})$, we have

$$\mathbf{MN} := \quad X \text{ —[M]—[N]— } Z \tag{2}$$

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

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

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

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

$$\mathbf{M} \otimes \mathbf{N} := \quad \begin{array}{l} X \text{ —[M]— } Y \\ Z \text{ —[O]— } W \end{array} \tag{5}$$

Product spaces can be represented either by two parallel wires or a single wire:

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

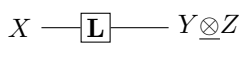
$$= \quad X \underline{\otimes} Y \text{ —— } X \underline{\otimes} Y \tag{7}$$

The notation $\mathsf{X}\underline{\otimes}\mathsf{Y}$ will be explained in paragraph **??** - $\mathsf{X}\underline{\otimes}\mathsf{Y}$ is a meta variable taking values in in the product space $X \times Y$.

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

(8)

≡ (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:

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

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

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

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

$$\sigma := (x, y) \to \delta_{(y,x)} \equiv \quad \begin{matrix} Y & X \\ X & Y \end{matrix} \tag{13}$$

Apart from identity, copy and swap maps, we assign different names to the input and output wires of Markov kernels.

The discard map $* : X \to \Delta(\{*\})$ maps every input to δ_*. Note that the only non-empty event in $\{\emptyset, \{*\}\}$ must have probability 1.

$$* : x \mapsto \delta_* \equiv \quad X \text{ ——}* \tag{14}$$

We can associate a Markov kernel $F \to \Delta(\mathcal{X})$ with any measurable function $F \to X$. A useful property of functional kernels is that products with functional kernels induce push-forward measures.

Definition 2.1 (Function induced kernel)**.** Given a measurable function $g : F \to X$, define the function induced kernel $\mathbf{F}_g : F \to \Delta(\mathcal{X})$ to be the the Markov kernel $a \mapsto \delta_{g(a)}$ for all $a \in X$.

Definition 2.2 (Pushforward kernel)**.** Given a kernel $\mathbf{M} : E \to \Delta(\mathcal{F})$ and a measurable function $g : F \to X$, the *pushforward kernel* $g_\# \mathbf{M} : E \to \Delta(\mathcal{X})$ is the kernel such that $g_\# \mathbf{M}(a; B) = \mathbf{M}(a; g^{-1}(B))$.

If E is the indiscrete space $\{*\}$, then $\mathbf{M}$ can be identified with the probability measure $\mu := \mathbf{M}_*$ and the pushforwark kernel $g_\# \mathbf{M}$ identified with the pushforward measure $g_\# \mu$, so pushforward kernels reduce to pushforward measures.

Lemma 2.3 (Pushforward kernels are functional kernel products)**.** *Given a kernel* $\mathbf{M} : E \to \Delta(\mathcal{F})$ *and a measurable function* $g : F \to X$*,the pushforward* $g_{\#}\mathbf{M} = \mathbf{M}\mathbf{F}_g$*.*

Proof.

$$\mathbf{M}\mathbf{F}_g(a; B) = \int_F \delta_{g(y)}(B) d\mathbf{M}_a(y) \tag{15}$$

$$= \int_F \delta_y(g^{-1}(B)) d\mathbf{M}_a(y) \tag{16}$$

$$= \int_{g^{-1}(B)} d\mathbf{M}_a(y) \tag{17}$$

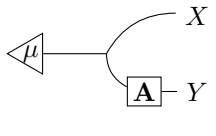
$$= g_{\#}\mathbf{M}(a; B) \tag{18}$$

□

2.4.1 Comparison of notations

We are in a position to compare the three introduced notations using a few 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:

$$\nu := A \times B \mapsto \int_A A(x; B) d\mu(x) \equiv \mu\curlyvee(\mathbf{Id}_X \otimes \mathbf{A}) \equiv \tag{19}$$



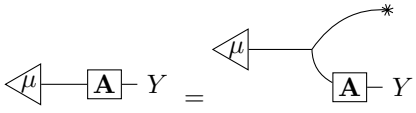
Where the resulting object is a probability measure $\nu \in \Delta(\mathcal{X} \otimes \mathcal{Y})$. Note that the elementary notation requires a function definition here, while the product and string notations can represent the measure without explicitly addressing its action on various inputs and 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 19 over X in our three separate notations as follows:

$$\nu_Y(B) = \nu(X \times B) = \int_X A(x; B) d\mu(x) \tag{20}$$

$$\nu_Y = \mu\mathbf{A} = \mu\curlyvee(\mathbf{Id}_X \otimes \mathbf{A})(* \otimes \mathbf{Id}_Y) \tag{21}$$

$$\nu_Y = \tag{22}$$



The elementary notation 20 makes the relationship between ν_Y and ν explicit and, again, requires the action on each event to be defined. The product

notation 21 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 in19, 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. Downsides of string diagrams are that they can be time consuming to typeset, and formal reasoning with them takes some practice.

2.4.2 Random Variables

The summary of this section is:

- We label wires with the names of random variables
- Diagrams with random variable labeled wires correspond to conditional/marginal distributions of those random variables in the obvious way
- We work with *conditional probability spaces* which are like probability spaces except some random variables don't have marginal distributions

Probability theory is primarily concerned with the behaviour of *random variables*. This behaviour can be analysed via a collection of probability measures and Markov kernels representing joint, marginal and conditional distributions of random variables of interest. In the framework developed by Kolmogorov, this collection of joint, marginal and conditional distributions is modeled by a single underlying *probability space*, and random variables by measurable functions on the probability space.

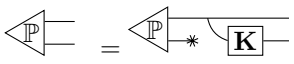
We use the same approach here, with a couple of additions.

1. First, we are interested in variables whose outcomes depend both on random processes and decisions. These variables are better modelled by a Markov kernels than probability measure - *given* a particular decision, they inherit a particular probability distribution. Thus, variables in our work are modeled by an underlying Markov kernel rather than a probability measure; we call this a *conditional probability space*
2. Secondly, we show how Markov kernel diagrams representing joint, marginal and conditional distributions within a conditional probability space can be identified by labels on the wires, analogously to the argument labels in $\mathbb{P}(\mathsf{X}, \mathsf{Y})$ identifying the joint distribution of X and Y

With regard to the first point, Hájek (2003) notes that there are many Markov kernels that cannot be uniquely specified by conditionals of probability

measures. Thus, in general, conditional probability spaces cannot be identified with probability spaces. However, rather than dealing with the issues raised by this possibility, we limit ourselves to conditional probability spaces that can be identified with probability spaces.

As a general motivation, suppose the following identity holding for some μ, $\mathbf{K}$:



(23)

This implies, roughly, that $\mathbf{K}$ is the probability of the lower wire of μ conditional on the upper (it is a *disintegration* of μ, defined later). We will want to deal with conditional probabilities such as $\mathbf{K}$ regularly; it would be nice to be able to define it as such without having to explicitly write out an equation like 23. Instead, we will adopt a system whereby the wires of a distinguished probability measure (or general Markov kernel) have names:

(24)

And then we adopt the convention that any kernel labelled as

(25)

satisfies Equation 23 when substituted for $\mathbf{K}$. These names stand for random variables on an appropriately defined *conditional probability space*.

Definition 2.4 (Probability space, conditional probability space)**.** A *probability space* $(\mathbb{P}, \Omega, \mathcal{F})$ is a probability measure $\mathbb{P}$, which we call the *ambient measure*, along with the *sample space* Ω and the *events* $\mathcal{F}$.

A *conditional probability space* $(\mathcal{K}, \Omega, \mathcal{F}, D, \mathcal{D})$ is a Markov kernel $\mathcal{K}$, called the *ambient kernel*, along with the sample space Ω and the *global conditioning space* D.

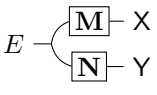
Note that we use the blackboard and script fonts to distinguish ambient measures and kernels. They are formally the same thing as ordinary probability measures and kernels, but it is useful to distinguish them.

Definition 2.5 (Random variable)**.** Given a sample space Ω and an input space D, a random variable X is a measurable function $\Omega \times D \to E$ for arbitrary measurable E. If the input space is trivial, it is simply a measurable function $\Omega \to X$.

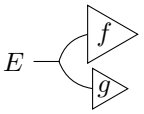
We call the random variable $\mathsf{D} : \Omega \times D \to D$ given by $(w, d) \mapsto d$ to be the *global conditioning variable*. D does not have a marginal distribution on a nontrivial conditional probability space (i.e. we don't have D isomorphic to the indiscrete set $\{*\}$).

Definition 2.6 (Coupled tensor product $\underline{\otimes}$)**.** Given two Markov kernels $\mathbf{M}$ and $\mathbf{N}$ or functions f and g with shared domain E, let $\mathbf{M}\underline{\otimes}\mathbf{N} := \curlyvee(\mathbf{M} \otimes \mathbf{N})$ and $f\underline{\otimes}g := \curlyvee(f \otimes g)$ where these expressions are interpreted using standard product notation. Graphically:

$$\mathbf{M}\underline{\otimes}\mathbf{N} := \tag{26}$$



$$f\underline{\otimes}g := \tag{27}$$



The operation denoted by $\underline{\otimes}$ is associative (Lemma 2.7), so we can without ambiguity write $f\underline{\otimes}g\underline{\otimes}...\underline{\otimes}h$ for finite groups of functions or Markov kernels sharing a domain.

Lemma 2.7 ($\underline{\otimes}$ is associative)**.** *For Markov kernels* $\mathbf{L}$*,* $\mathbf{M}$ *and* $\mathbf{N}$ *sharing a domain* E*,* $(\mathbf{L}\underline{\otimes}\mathbf{M})\underline{\otimes}\mathbf{N} = \mathbf{L}\underline{\otimes}(\mathbf{M}\underline{\otimes}\mathbf{N})$.

Definition 2.8 (Marginal distribution, marginal kernel)**.** Given $\mathbb{P} \in \Delta(\mathcal{F})$, random variable $\mathsf{X} : \Omega \to G$ the *marginal distribution* of X $\mathbb{P}_\mathsf{X} \in \Delta(\mathcal{G})$ of X is the product measure $\mathbb{P}\mathbf{F}_\mathsf{X}$.

See Lemma 2.3 for the proof that this matches the usual definition of marginal distribution.

Following this, given $\mathcal{K} : D \to \Delta(\mathcal{F})$ and random variable $\mathsf{X} : \Omega \to G$, the *marginal kernel* is $\mathcal{K}_{\mathsf{X}|\mathsf{D}} := \mathcal{K}\mathbf{F}_\mathsf{X}$. Note that we include the global conditioning variable D in this notation.

Definition 2.9 (Joint distribution, joint kernel)**.** Given $\mathbb{P} \in \Delta(\mathcal{F})$, $\mathsf{X} : \Omega \to G$ and $\mathsf{Y} : \Omega \to H$, the *joint distribution* $\mathbb{P}_{\mathsf{XY}} \in \Delta(\mathcal{G} \otimes \mathcal{H})$ of X and Y is the marginal distribution of $\mathsf{X}\underline{\otimes}\mathsf{Y}$.

This is identical to the definition in, for example, Çinlar (2011) if we note that the random variable $(\mathsf{X}, \mathsf{Y}) : \omega \mapsto (\mathsf{X}(\omega), \mathsf{Y}(\omega))$ (Çinlar's definition) is the same thing as $\mathsf{X}\underline{\otimes}\mathsf{Y}$.

Analogously, the joint kernel $\mathcal{K}_{\mathsf{YX}|\mathsf{D}}$ is the product $\mathcal{K}\mathbf{F}_{\mathsf{X}\underline{\otimes}\mathsf{Y}}$.

This is just an aside

Joint distributions and kernels have a nice visual representation, as a result of Lemma 2.10 which follows.

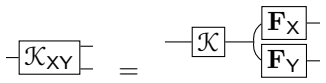
Lemma 2.10 (Joint distributions and coupled products)**.** *Given* $\mathsf{X} : \Omega \to G$ *and* $\mathsf{Y} : \Omega \to H$*,* $\mathbf{F}_{\mathsf{X}\underline{\otimes}\mathsf{Y}} = \mathbf{F}_\mathsf{X}\underline{\otimes}\mathbf{F}_\mathsf{Y}$

Proof. For $a \in \Omega$, $B \in \mathcal{G}$, $C \in \mathcal{H}$,

$$\begin{aligned}
\mathbf{F}_{\mathsf{X}\underline{\otimes}\mathsf{Y}}(a; B \times C) &= \delta_{\mathsf{X}(a),\mathsf{Y}(a)}(B \times C) && (28)\\
&= \delta_{\mathsf{X}(a)}(B)\delta_{\mathsf{Y}(a)}(C) && (29)\\
&= (\delta_{\mathsf{X}(a)} \otimes \delta_{\mathsf{Y}(a)})(B \times C) && (30)\\
&= \mathbf{F}_{\mathsf{X}}\underline{\otimes}\mathbf{F}_{\mathsf{Y}} && (31)
\end{aligned}$$

Equality follows from the monotone class theorem. □

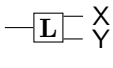
Therefore the following holds:



(32)

We are now in a position to define wire labels for "output" wires.

Definition 2.11 (Wire labels - joint probabilities)**.** Given a conditional probability space with ambient kernel $\mathcal{K} : D \to \Delta(\mathcal{F})$ (or a probability space with measure $\mathbb{P}$) and some collection of random variables X, Y, ... on $\Omega \times D$, any diagram representing a kernel $\mathbf{L} : D \to \Delta(\mathcal{E})$ with *output* wires labeled X, Y,... represents the corresponding marginal of $\mathcal{K}$ (we assume that the space E factorises appropriately). For example



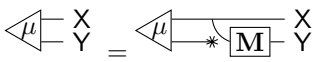
(33)

asserts that $\mathbf{L} = \mathbf{K}_{\mathsf{XY}|\mathsf{D}}$, the joint kernel of X and Y.
If we have an ambient measure $\mathbb{P}$, then $D = \{*\}$ and the diagram

(34)

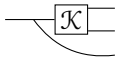
asserts that $\mu = \mathbb{P}_{\mathsf{XY}}$.

Definition 2.12 (Disintegration)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$, random variables X and Y and joint probability measure $\mu := \mathbb{P}_{\mathsf{XY}} \in \Delta(\mathcal{E} \otimes \mathcal{F})$, we say that $\mathbf{M} : E \to \Delta(\mathcal{F})$ is a disintegration of μ if



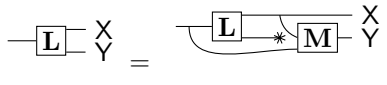
(35)

Notationally, $\mathbf{N}$ is a version of $\mathbb{P}_{\mathsf{Y}|\mathsf{X}}$, "the probability of Y given X", or $\mathbf{M} \in \mathbb{P}_{\mathsf{Y}|\mathsf{X}}$.
Given a conditional probability space $(\mathcal{K}, \Omega, D)$, define $\mathcal{K}^*$ to be the kernel



(36)

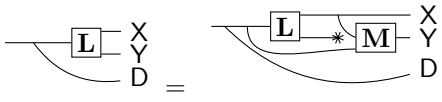
Given random variables X, Y on $\Omega \times D$ and kernel $\mathbf{L} := \mathcal{K}^*_{\mathsf{X},\mathsf{Y}} : D \to \Delta(\mathcal{E} \otimes \mathcal{F})$, we say that $\mathbf{M} : D \times E \to \Delta(\mathcal{F})$ is a disintegration of $\mathbf{L}$ if



(37)

Similarly, recalling that D is the global conditioning variable, we can say $\mathbf{M} \in \mathcal{K}_{\mathsf{Y}|\mathsf{X}\mathsf{D}}$. We require disintegrations of kernel spaces to be conditional on the global conditioning variable, as this along with certain other conditions guarantees the existence of a disintegration.

Note that Eq. 38 also implies



(38)

Definition 2.13 (Wire labels - disintegrations)**.** Given a conditional probability space with ambient kernel $\mathcal{K} : D \to \Delta(\mathcal{F})$ (or a probability space with measure $\mathbb{P}$),

Note that $\mathbb{P}^*$ is simply $\mathbb{P}$ for a probability space

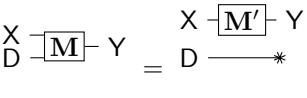
Recall that D is the global conditioning variable. Given two collections of random variables $c_1 = [\mathsf{X}_1, \mathsf{X}_2, ...]$ and $c_2 = [\mathsf{Y}_1, \mathsf{Y}_2]$, we adopt the convention that any diagram with the input wires labeled with c_1 and the output wires labeled with c_2 is an element of $\mathcal{K}^*_{\mathsf{Y}_1\mathsf{Y}_2...|\mathsf{X}_1\mathsf{X}_2...}$.

That is, by this convention, the diagram

(39)

implies that $\mathbf{M} \in \mathcal{K}_{\mathsf{Y}|\mathsf{X}\mathsf{D}}$. Note further that by Theorem 2.15, we can rely on the existence of disintegrations such as $\mathbf{M}$ that are conditional on the global conditioning variable D provided we have countable D and standard measurable $(Y, \mathcal{Y})$.

If we have some version $\mathbf{M}$ of $\mathcal{K}_{\mathsf{Y}|\mathsf{X}\mathsf{D}}$ that does not depend on the value of D - i.e. $\mathbf{M}_{(x,d)} = \mathbf{M}_{(x,d')}$ for all $x \in X$, $d, d' \in D$, then there exists some $\mathbf{M}'$ such that:



(40)

Under these circumstances, we will abuse notation to say $\mathbf{M}' = \mathcal{K}_{\mathsf{Y}|\mathsf{X}}$.

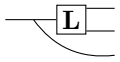
We can't expect Equation 40 to hold in an arbitrary conditional probability spaces. For a very simple example, take $\mathcal{K} : \{0, 1\} \to \Delta(\{0, 1\})$ where $\mathcal{K}_0 =$

$\mathcal{K}_1 = \text{Bernoulli}(0.5)$, and let $\mathsf{X} : (x, d) \mapsto x$ - i.e. the random variable projecting the output of $\mathcal{K}$. Then there is no disintegration $\mathcal{K}_{\mathsf{D}|\mathsf{X}}$ - we can't recover the input D from X.

Under some (strong) regularity conditions, disintegrations of conditional probability spaces do exist.

Theorem 2.14 (Disintegration existence - probability space)**.** *Given a probability measure $\mu \in \Delta(\mathcal{E} \otimes \mathcal{F})$, if $(F, \mathcal{F})$ is standard then a disintegration $\mathbf{K} : E \to \Delta(\mathcal{F})$ exists (Çinlar, 2011).*

Theorem 2.15 (Disintegration existence - conditional probability space)**.** *Given a kernel $\mathbf{L} : D \to \Delta(\mathcal{E} \otimes \mathcal{F})$, define $\mathbf{L}^*$:*



(41)

If D is countable and $(F, \mathcal{F})$ is standard, then there is a disintegration $\mathbf{M} : D \times E \to \Delta(\mathcal{F})$ of $\mathbf{L}^*$.

Proof. By Theorem 2.14, for each $d \in D$ we have a disintegration $\mathbf{K}^{(d)} : E \to \Delta(\mathcal{F})$ of $\mathbf{L}_d$. Define $\mathbf{M} : D \times E \to \Delta(\mathcal{F})$ by $\mathbf{M}(d, e; A) = \mathbf{K}^{(d)}(e; A)$ for $d \in D$, $e \in E$, $A \in \mathcal{F}$. Clearly $\mathbf{M}_{(d,e)}$ is a probability measure. Furthermore, for $B \in \mathcal{B}(\mathbb{R})$, $\mathbf{M}^{-1}(\cdot; A)(B) = \cup_{d \in D}\{d\} \times \mathbf{K}^{(d)-1}(\cdot; A)(B)$, which is a countable union of measurable sets and therefore measurable. □

From here on out, we will assume whether explicitly stated or not that any global conditioning space is countable and any other measureable space is standard, guaranteeing the existence of disintegrations.

In general, we don't want to spent time explicitly setting up conditional probability spaces. Rather, we will specify key marginals and disintegrations from which a conditional probability space can be constructed - call these marginals and conditional "components". Clearly we cannot build a conditional probability space from two kernels that represent the same component but disagree with each other on a non-negligible set. Also, in general, for an arbitrary collection of components there may be many ambient kernels from which we can extract these components. There is no particular problem if we have multiple ambient kernels over undefined random variable; if we are only interested in X then the possibility of many joint kernels over X and Y is no cause for concern. We do, however, want to avoid ambient kernels supporting non-negligibly distinct marginals or disintegrations over the random variables that have been defined.

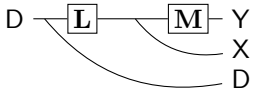
Example 2.16 (Implicit conditional probability space)**.** Suppose we have labeled Markov kernels

(42)

We want to define a conditional probability space $(\mathcal{K}, \Omega, D)$ supporting random variables D, X and Y yielding the above kernels as the relevant marginals and disintegrations. Strictly:

- $\mathbf{L} = \mathcal{K}_{\mathsf{X}|\mathsf{D}}$
- $\mathbf{M} \otimes {}^{*}\!\!\uparrow_D \in \mathcal{K}_{\mathsf{Y}|\mathsf{X}D}$ ("informally", $\mathbf{M} \in \mathcal{K}_{\mathsf{Y}|\mathsf{X}}$)

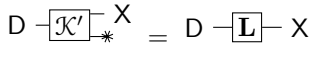
Take $\Omega = W \times X \times Y \times Z$ and define $\mathcal{K}$ such that

$\mathcal{K}^* =$ 

(43)

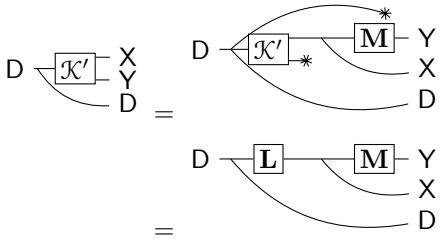
Where $\mathcal{K}^*$ is the copy map composed with $\mathcal{K}$ as in previous definitions. $\mathcal{K}$ is the unique Markov kernel $D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ supporting the two criteria above, assuming finite D and standard measurable X, Y.

Proof. By assumption, for any suitable $\mathcal{K}' : D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ we have



(44)

and by the fact that $\mathbf{M} \otimes {}^{*}\!\!\uparrow_D$ is by assumption a disintegration of $\mathcal{K}'^*$:

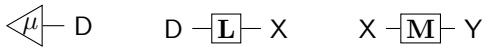


(45)

(46)

Finally, if $\mathcal{K}^* = \mathcal{K}'^*$, then at least if D is countable we must have $\mathcal{K} = \mathcal{K}'$ as they must agree on all points in D. □

This example was chosen to illustrate a peculiarity of our notation of conditional probability spaces. Consider a problem that appears similar: find an ambient measure $\mathbb{P}$ decomposing into the following marginal and conditionals:



(47)

Here there are many choices of $\mathbb{P}$ that satisfy our conditions arising from different choices of $\mathbb{P}_{\mathsf{Y}|\mathsf{X}\mathsf{D}}$. This is not possible in the conditional probability space because $\mathcal{K}_{\mathsf{Y}|\mathsf{X}}$ only exists if $\mathcal{K}_{\mathsf{Y}|\mathsf{X}\mathsf{D}}$ is independent of D. That is, in a conditional probability space every disintegration is conditional on D, but we may not explicitly write this if it does not actually depend on D.

A sufficient condition for the construction of a unique ambient kernel from a collection of components $\{C_1, ...C_n\}$ is if there is some ordering of components $\{i_1, i_2, ...i_n\}$ such that the input labels of $C_{i_{k+1}}$ is the union of the inputs and outputs of $C_{i_1}, ..., C_{i_j}$. This can be shown by repeated application of Theorem 2.15.

In general, diagram labels are "well behaved" with regard to the application of any of the special Markov kernels: identities 11, swaps 13, discards 14 and copies 12 as well as with respect to the coherence theorem of the CD category. They are not "well behaved" with respect to composition.

Lemma 2.17 (Diagrammatic consequences of labels)**.** *Fix some conditional probability space* $(\mathcal{K}, \Omega, D)$ *and random variables* X, Y, Z *taking values in arbitrary spaces.* Sat : *indicates that a labeled diagram satisfies definitions 2.11 and 2.13 with respect to* $(\mathcal{K}, \Omega, D)$ *and* X, Y, Z*. The following always holds:*

$$\text{Sat} : \mathsf{X} \text{ — } \mathsf{X} \tag{48}$$

and the following implications hold:

$$\text{Sat} : \mathsf{Z} \text{ — } \boxed{\text{K}} \text{ — } \begin{matrix}\mathsf{X}\\ \mathsf{Y}\end{matrix} \implies \text{Sat} : \mathsf{Z} \text{ — } \boxed{\text{K}} \text{ — } \mathsf{X} \quad (\ast) \tag{49}$$

$$\text{Sat} : \mathsf{Z} \text{ — } \boxed{\text{K}} \text{ — } \begin{matrix}\mathsf{X}\\ \mathsf{Y}\end{matrix} \implies \text{Sat} : \mathsf{Z} \text{ — } \boxed{\text{K}} \text{ — } \begin{matrix}\mathsf{Y}\\ \mathsf{X}\end{matrix} \tag{50}$$

$$\text{Sat} : \mathsf{Z} \text{ — } \boxed{\text{L}} \text{ — } \mathsf{X} \implies \text{Sat} : \mathsf{Z} \text{ — } \boxed{\text{L}} \text{ — } \begin{matrix}\mathsf{X}\\ \mathsf{X}\end{matrix} \tag{51}$$

$$\text{Sat} : \mathsf{Z} \text{ — } \boxed{\mathbf{K}} \text{ — } \mathsf{Y} \implies \text{Sat} : \begin{matrix}\mathsf{Z}\\ \mathsf{Z}\end{matrix} \text{ — } \boxed{\mathbf{K}} \text{ — } \mathsf{Y} \tag{52}$$

Proof.

- Id_X is a version of $\mathbb{P}_{\mathsf{X}|\mathsf{X}}$ for all $\mathbb{P}$; $\mathbb{P}_{\mathsf{X}}\mathrm{Id}_X = \mathbb{P}_{\mathsf{X}}$
- $\mathbf{K}\mathrm{Id} \otimes \ast)(w; A) = \int_{X\times Y} \delta_x(A)\mathbb{1}_Y(y)d\mathbf{K}_w(x, y) = \mathbf{K}_w(A \times Y) = \mathbb{P}_{\mathsf{X}|\mathsf{Z}}(w; A)$
- $\int_{X\times Y} \delta_{\mathrm{swap(x,y)}}(A \times B)d\mathbf{K}_w(x, y) = \mathbb{P}_{\mathsf{YX}|\mathsf{Z}}(w; A \times B)$
- $\mathbf{K}\curlyvee(w; A \times B) = \int_X \delta_{x,x}(A \times B)d\mathbf{K}_w(x) = \mathbb{P}_{\mathsf{XX}|\mathsf{Z}}(w; A \times B)$

52: Suppose $\mathbf{K}$ is a version of $\mathbb{P}_{\mathsf{Y}|\mathsf{Z}}$. Then

$$\mathbb{P}_{\mathsf{ZY}} = \tag{53}$$

$$\mathbb{P}_{\mathsf{ZZY}} = \tag{54}$$

$$= \tag{55}$$

Therefore $\curlyvee(\mathrm{Id}_X \otimes \mathbf{K})$ is a version of $\mathbb{P}_{\mathsf{ZY}|\mathsf{Z}}$ by **??** □

The following property, on the other hand, does *not* generally hold:

$$\mathrm{Sat} : \mathsf{Z} \text{ —}\boxed{\mathbf{K}}\text{— } \mathsf{Y} \ , \ \mathsf{Y} \text{ —}\boxed{\mathbf{L}}\text{— } \mathsf{X} \implies \mathrm{Sat} : \mathsf{Z} \text{ —}\boxed{\mathbf{K}}\text{—}\boxed{\mathbf{L}}\text{— } \mathsf{X} \tag{56}$$

Consider some ambient measure $\mathbb{P}$ with $\mathsf{Z} = \mathsf{X}$ and $\mathbb{P}_{\mathsf{Y}|\mathsf{X}} = x \mapsto \text{Bernouli}(0.5)$ for all $z \in Z$. Then $\mathbb{P}_{\mathsf{Z}|\mathsf{Y}} = y \mapsto \mathbb{P}_{\mathsf{Z}}$, $\forall y \in Y$ and therefore $\mathbb{P}_{\mathsf{Y}|\mathsf{Z}}\mathbb{P}_{\mathsf{Z}|\mathsf{Y}} = x \mapsto \mathbb{P}_{\mathsf{Z}}$ but $\mathbb{P}_{\mathsf{Z}|\mathsf{X}} = x \mapsto \delta_x \neq \mathbb{P}_{\mathsf{Y}|\mathsf{Z}}\mathbb{P}_{\mathsf{Z}|\mathsf{Y}}$.

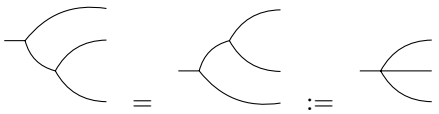
2.4.3 Working With String Diagrams

todo:

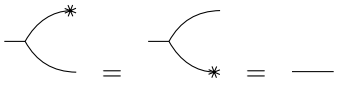
- Functional generalisation
- Conditioning
- Infinite copy map
- De Finetti's representation theorem

There are a relatively small number of manipulation rules that are useful for string diagrams. In addition, we will define graphically analogues of the standard notions of *conditional probability*, *conditioning*, and infinite sequences of exchangeable random variables.

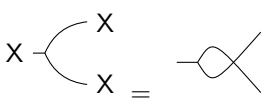
Axioms of Symmetric Monoidal Categories For the following, we either omit labels or label diagrams with their domain and codomain spaces, as we are discussing identities of kernels rather than identities of components of a conditional probability space. Recalling the unique Markov kernels defined above, the following equivalences, known as the *commutative comonoid axioms*, hold among string diagrams:



(57)

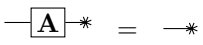


(58)



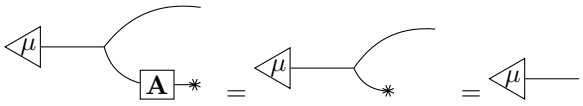
(59)

The discard map $*$ can "fall through" any Markov kernel:



(60)

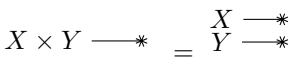
Combining 58 and 60 we can derive the following: integrating $\mathbf{A} : X \to \Delta(\mathcal{Y})$ with respect to $\mu \in \Delta(\mathcal{X})$ and then discarding the output of $\mathbf{A}$ leaves us with μ:



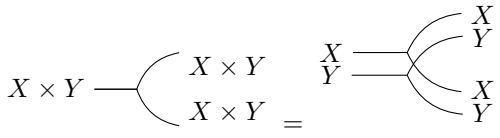
(61)

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 $\curlyvee$:

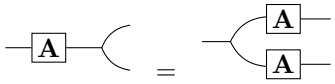


(62)



(63)

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



(64)

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

Definition 2.18 (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 2.19 (Copy map commutes for deterministic kernels (Fong, 2013))**.** *Equation 64 holds iff* $\mathbf{A}$ *is deterministic.*

Causal statistical decision theory offers a different set of basic assumptions to causal reasoning based on causal Bayesian networks, and also offers an identification theorem that applies to situations where the backdoor criterion of CBNs applies, but makes a different set of furhter assumptions to derive this result.

In particular, CSDT avoids the need for a *do-scheme*, an assumption that I find particularly hard to evaluate as I explain below. One of the costs of

adopting the CSDT view is the lack of a formal definition of "causal effect". Under the right conditions CSDT features stable relationships between random variables that may be informally interpreted as causal effects, but this is only at the level of informal interpretation.

Whether the assumptions underlying CSDT are better than those underlying the CBN for any class of questions is not easy to say. Addressing it, I would imagine, should proceed from adopting an established criteria for judging assumptions and comparing assumptions that differ under each framework according to these criteria. There would be a substantial amount of work involved in doing this and it's not presently in scope.

2.5 Evaluating Assumptions (philosophy is hard)

Tentatively, the following properties of assumptions seem desirable:

1. People who are solving problems should be able to evaluate assumptions with a low rate of false positives (i.e. problems where further assumptions are judged to hold but in fact they do not)

2. False negatives should be minimised subject to the requirement of a very low number of false positives (i.e. as long as it doesn't increase the false positive rate, it is desirable to find more problems that can be solved by a particular theorem)

3. Assumptions and results should exhibit continuity - where an assumption holds approximately, approximate results follow

Knowing how well assumptions satisfy these criteria is usually difficult. For example, we may not have any ground truth answers at all for whether some assumptions hold or not, and so whatever we can say about false positive rates is limited. This raises the possibility of a further desirable property of an assumption: we have some information about how well it satisfies 1-3. This addition invokes a meta-assumption that assumptions with a track record tend to be better than those without.

Actually, these desiderata don't seem quite right. Suppose A1 is an appealing but usually false assumption - i.e. it is subject to a high false positive rate. The desiderate imply that A1 is inferior A1 + F, where F stands for an assumption universally judged to be a contradiction. This violates the common notion that it is always better to work with fewer assumptions (though it does have a degree of plausibility - "everyone believes A1 is true, but everyone would be better off believing that A1 is false").

It's not obvious to me whether the CBN or CSDT approach is unequivocally better with regard to the criteria above.

2.6 The switch and lightbulb

The two approaches will be compared using the test case of "the switch and lightbulb". This is intentionally chosen as a very simple problem of causal

inference.

> There is a lightbulb and a switch (maybe there are other things as well). We have a set D of actions available, some of which we know turn the switch on and the rest of which we know turn the switch off, but the effect of any action on the lightbulb is uncertain. We have a very large number N of IID observations of the joint state of the lightbulb and switch. We are interested in choosing decisions that control the light.

This problem exhibits a number of features which we take to be common to any statistical causal decision problem:

- There is a set of decisions, known in advance, that may be chosen
- We possess some incomplete prior knowledge about the effects of decisions
- Data is available (observations of the switch and light state)
- The objective is to control some feature of the environment

This problem supports many different modes of operation. For example, we might have:

- The switch always turns the light on both in observations and as a result of our actions (more precisely, the switch and light always share the same state in observations and for all actions)
- In observations the light is on iff the switch is on, some actions preserve this relationship ("just turn the light on") while some break it ("cut power to the house, turn the light on")
- There is another switch ("switch 2") in the house which also controls the light which is randomly on or off in observations. Thus in observations the light is uncorrelated with switch 1 but there are nevertheless actions available that control the light: "turn switch 2 on, then turn switch 1 on"

Consider this problem setup with two additional assumptions: (1) the light operates according to the same rules in both observations and interaction and (2) there is only 1 way to switch the light off and 1 way to switch it on (i.e. $|D| = 2$). In this case, however we observe the light to behave when the switch is off is how it must behave when we turn the switch off, and similarly how we observe the light to behave when the switch is on must be how it behaves when we turn the switch on as the observed behaviour in each case must be the behaviour of the light under the unique "off" and "on" actions. Suppose we instead assume (2') there is only 1 way to switch the light off and there are 2 ways to switch it on (let's call them d_1 and d_2). In this case we can still determine the behaviour of the light from observations when we turn the switch off, but we cannot determine as much about it when we switch it on - the best

that we can say is that some mixture of d_1 and d_2 will reproduce the observed behaviour of the light when the switch is on. This doesn't limit the behaviour of the light under d_1 - if we observe the light was on every time the switch was on, but this was always the result of taking d_2, then it is perfectly possible that d_1 never switches the light on. On the other hand, if we always observe the light is on when the switch is on, then a random action that takes d_1 half the time and d_2 half the time must leave the light on at least half the time.

2.6.1 Using Causal Theories

To model this situation with a causal theory, we need to define the Markov kernel representing the coupled observation model $\mathcal{H} : \Theta \to \Delta(\{0,1\}^2)$ and consequence model $\mathcal{C} : \Theta \times D \to \Delta(\{0,1\}^2)$. The latent space Θ represents the set of coupled models we consider to be possible.

If we make no assumptions beyond what was specified in the problem, our observatonal model should admit every possible distribution in $\Delta(\{0,1\})^2$. Because we've made not assumptions about how observations go with consequences, each possible observational distribution will go with each possible consequence.

We have assumed that some decisions are known to turn the switch on and some turn the switch off. Let decisions be enumerated with an integer index j, and suppose $d_j \in D$ is a decision that turns the switch off iff $j <= 0$. That is, we have

$$\mathcal{C}_{\mathsf{S}_c|\mathsf{D}} = \begin{cases} d_j \mapsto \delta_1 & j > 0 \\ d_j \mapsto \delta_0 & j <= 0 \end{cases} \tag{65}$$

The "post-decision" state of the light L_c conditioned on the switch S_c and the decision D is given by $\mathcal{C}_{\mathsf{L}_c|\mathsf{S}_c\mathsf{D}}$ and like the observations is unrestricted.

Thus Θ must represent the cartesian product of the set of observational distributions $\Delta(\{0,1\}^2)$ and set of possible conditionals $\mathcal{C}_{\mathsf{L}_c|\mathsf{S}_c\mathsf{D}}$. We can represent the latter set as $[0,1]^{2|D|}$ - the distribution of the binary L_c can be parametrised by $p \in [0,1]$, and each possible model assigns a particular parameter $p_{d_j} \in [0,1]$ to each $d_j \in \mathsf{D}$, giving $\Theta = \Delta(\{0,1\}^2) \times [0,1]^{2|D|}$.

This general setup adds to the given information in the problem in that it supposes that the unknown things - namely, what the light will do given the various decisions available to us - should be represented as probability distributions. In addition, we add the assumption that D is countable. Otherwise, we have as far as possible represented just the information given, and we conclude (as we ought to from the original setup) that we do not possess sufficient information to control the light.

Under some assumptions, we may be able to control the light. For example, suppose $|D| = 2$, with one decision that turns the switch off and one that turns it on (decisions may still have side effects, but this rules out the possibility that there may be two decisions that turn the switch off each with *different* side effects). Suppose also that the problem is *reproducible* - that is, there is

some stochastic action $\gamma \in \Delta(\mathcal{D})$ such that $\gamma \mathcal{C}_{\mathsf{L}_c\mathsf{S}_c|\mathsf{D}} = \mathcal{H}_0$. Then it follows that $\mathcal{C}_{\mathsf{L}_c|\mathsf{S}_c\mathsf{D}} = \mathcal{H}_{\mathsf{L}_o|\mathsf{S}_o} \otimes \mathbb{1}_D$ and so the light may be controlled to the extent permitted by $\mathcal{H}_{\mathsf{L}_o|\mathsf{S}_o}$. This is a special case of Theorem **??**, and is a formalisation of the informal result above.

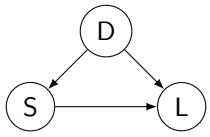
If we suppose that there are two decisions that turn the switch on - d_1 and d_2 - then we can derive inequality constraints on the consequences of mixtures of d_1 and d_2. Fix a state $\theta^* \in \Theta$ and let $p_{(on)} := \mathcal{H}^{\theta^*}_{\mathsf{L}_o|\mathsf{S}_o}(1; \{1\})$ be the observed probability that the light is on given that the switch is on. Then given the mixed decision $\gamma = a\delta_{d_1} + (1-a)\delta_{d_2}$ and supposing $a > 0.5$, we can say that

$$\min(0, p_{(on)} - a) \leq (\gamma \mathcal{C}^{\theta^*})_{\mathsf{L}_c|\mathsf{S}_c} \leq \max(1, p_{(on)} + a) \tag{66}$$

2.6.2 Using Causal Bayesian Networks

In order to approach this problem using causal Bayesian networks we must represent our assumptions in a graph $\mathcal{G}$ over some set of variables, and we must furthermore provide a *do-scheme*, which is a map from the set of decisions D provided by the problem. I am not aware of a set of instructions anywhere about how this ought to be done, and as we will see the process of doing so is less straightforward than one might expect.

To begin with, it is very difficult to model this problem with a CBN without adding more assumptions than are necessary to model the problem with a causal theory. If we use causal language informally, we have posited a decision that can influence the state of a switch and a light. If we propose causal random variables D, S and L, then we might turn this sentence into the following diagram:



(67)

If we suppose that this diagram represents both the observational data generating process (with censored D) and the consequence map, then it corresponds to a theory that implies *reproducibility*. While it is possible to construct alternative diagrams that don't imply reproducibility, they typically introduce their own additional assumptions. The fact that it is hard to draw a CBN without making assumptions about the relationship between the data generating process and the consequence map is related to the fact that a CBN represents both objects with one diagram.

Diagram 67 is also a non-standard example of a CBN, as the ostensibly "intervenable" variable D is censored in the data generating process. The vast majority of identifiability theorems concerning CBNs feature *observed* "intervenables". We also need to somehow incorporate the knowledge of how D is already known to influence S.

Observe also that in the controllable case above (where $|D| = 2$), the causal theory can be shown to correspond precisely to the CBN given by

(68)

when equipped with the *do-scheme* $f : D \to \mathrm{Do}(S \times L)$ given by

$$f = \begin{cases} d_0 \mapsto \mathrm{do}(\mathsf{S} = 0) \\ d_1 \mapsto \mathrm{do}(\mathsf{S} = 1) \end{cases} \tag{69}$$

Once again, the example was setup to suggest this interpretation. I also belive that in the controllable case, veterans of the CBN approach are likely to agree that

1. The DAG 68 is an appropriate representation of the causal structure of the switch-light example
2. The two options available do indeed correspond to $do(\mathsf{S} = 0)$ and $do(\mathsf{S} = 1)$; in other words, the do-scheme 69 is appropriate. I give the informal version here as formal do-schemes aren't a component of the normal DAG approach, but identification of actions available with $do()$ statements is

This second approach to drawing a causal graph becomes more complicated if we suppose that there are two ways to turn the light on and only one way to turn it off. In this case there is no do-scheme on diagram 68 that yields the inequality 66 as the sole constraint on the results of mixtures of d_1 and d_2. This is true even if we allow:

- Compound $do(..)$ statements like $do(\mathsf{S} = 1, \mathsf{L} = 0)$
- The passive $do()$ statement that yields the same distribution over S and L as was observed
- Conditional $do(..)$ operations; e.g. define $g : \{0, 1\} \to \{0, 1\}$ and let $do(\mathsf{S} = 1, \mathsf{L} = g(\mathsf{S}))$ be defined as the operation that sets $\mathsf{S} = 1$ and $\mathsf{L} = \mathsf{S}$
- Stochastic do-schemes e.g. $f(d_0) = 0.25(\delta_{do(\mathsf{S}=1)} + \delta_{do(\mathsf{S}=1,\mathsf{L}=0)} + \delta_{do()} + \delta_{do(\mathsf{S}=1,\mathsf{L}=g(\mathsf{S}))})$

In fact this is easy to see: holding observations fixed, the causal theory for three decisions permits any consequence map satisfying 66, while fixing any do-scheme on 68 will yield a unique consequence map. The situation may be resolved by allowing for observation-dependent uncertainty over do-schemes, but a do scheme with observation-dependent uncertainty can represent any causal theory independent of the graph on which it is defined (this follows from the fact that any consequence map can be represented as a map from actions to mixtures of compound do-statements on every random variable).

It seems reasonable to me to conclude in light of this that diagram 68 is not an appropriate diagram for the case of three actions. This is already an

interesting result - the notion that diagram 68 "correctly represents the causal relationships of the problem" can hold when $|D| = 2$ but not when $|D| = 3$, and the nature of the set D of available actions is rarely given any thought at all in the field of causal graphical models. In defense of the potential outcomes crowd, at least they recognise that this is an assumption that they have to make:

> SUTVA [...] assumes that there are no hidden versions of treatments; no matter how unit i received treatment 1, the outcome that would be observed would be Y i (1) and similarly for treatment 0 (Rubin, 2005)

What kind of graph *can* we use when $|D| >= 3$?

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.

Alan Hájek. What Conditional Probability Could Not Be. *Synthese*, 137(3): 273–323, December 2003. ISSN 1573-0964. doi: 10.1023/B:SYNT.0000004904.91112.16. URL https://doi.org/10.1023/B:SYNT.0000004904.91112.16.

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/.

Donald B. Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, March 2005. ISSN 0162-1459. doi: 10.1198/016214504000001880. URL https://doi.org/10.1198/016214504000001880.

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: