Thesis Proposal Review: How Hard is a Causal Inference Problem

David Johnston

February 4, 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 $\sigma(\underline{\bigotimes}_{i\in\mathcal{G}}\mathsf{X}_i)$. We might surmise that the "conditional probability" $P(\cdot|do(\mathsf{X}_k = \cdot))$ might then be the conditional probability on $\sigma(\underline{\bigotimes}_{i\in\mathcal{G}}\mathsf{X}_i)$. However, CBNs in general support models where $P(\cdot|do(\mathsf{X}_k = \cdot))$ is not equal to $P(\cdot|\mathsf{A})$ for any $\mathsf{A} \in \sigma(\underline{\bigotimes}_{\mathcal{G}}\mathsf{X}_i)$, 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 map
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: $\mathbf{F}_{\mathsf{X}} \equiv$
$\mathbf{N}_{\mathsf{A}\vert\mathsf{B}}$	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 unwieldy 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 $A \times B \mapsto \int_A \mathbf{M}(x; B) d\mu$.

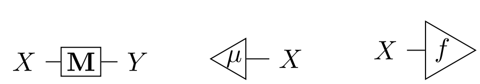
Such constructions can, however, be straightforwardly captured with string diagrams, a notation developed for category theoretic probability. Cho and Jacobs (2019) also provides an extensive introduction to the notation discussed here.

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

$$\mathbf{MN} := X \;—\boxed{\mathbf{M}}—\boxed{\mathbf{N}}—\; Z \tag{2}$$

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

$$\mu\mathbf{M} := \mu —\boxed{\mathbf{M}}—\; Z \tag{3}$$

$$\mathbf{M}f := X \;—\boxed{\mathbf{M}}— f \tag{4}$$

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

$$\mathbf{M} \otimes \mathbf{N} := \begin{array}{l} X \;—\boxed{\mathbf{M}}—\; Y \\ Z \;—\boxed{\mathbf{O}}—\; W \end{array} \tag{5}$$

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

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

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

The notation $\mathsf{X}\underline{\otimes}\mathsf{Y}$ will be explained in paragraph 2.4.4 - $\mathsf{X}\underline{\otimes}\mathsf{Y}$ 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:

$$X \;\text{—}\boxed{\mathbf{L}}\text{=}\; \begin{matrix} Y \\ Z \end{matrix} \tag{8}$$

$$\equiv \tag{9}$$

$$X \;\text{—}\boxed{\mathbf{L}}\text{—}\; Y \underline{\otimes} Z \tag{10}$$

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:

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

The identity map preserves the name of a wire.

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

$$\curlyvee : x \mapsto \delta_{(x,x)} \equiv \;\; X \prec \begin{matrix} X \\ X \end{matrix} \tag{12}$$

Copy maps *copy* the name of a wire.

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

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

The swap map preserves the names of visually connected wires.

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.

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

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.

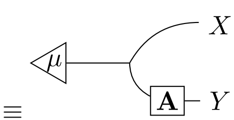
$$\mathbf{MF}_g(a; B) = \int_F \delta_{g(y)}(B) d\mathbf{M}_a(y) \tag{15}$$
$$= \int_F \delta_y(g^{-1}(B)) d\mathbf{M}_a(y) \tag{16}$$
$$= \int_{g^{-1}(B)} d\mathbf{M}_a(y) \tag{17}$$
$$= g_\#\mathbf{M}(a; B) \tag{18}$$

$\square$

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:

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

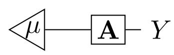
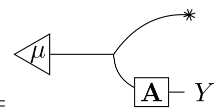


Where the resulting object is a probability measure $\nu \in \Delta(\mathcal{X} \otimes \mathcal{Y})$. Note that the elementary notation requires a function definition here, while the product and string notations can represent the measure without explicitly addressing its action on various inputs and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

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

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

$\nu_Y =$  $=$  (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.

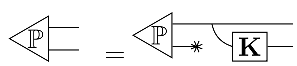
1. First, we are interested in variables whose outcomes depend both on random processes and decisions. These variables are better modelled by a

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

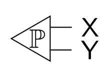
With regard to the first point, Hájek (2003) notes that there are many Markov kernels that cannot be uniquely specified by conditionals of probability measures. Thus, in general, conditional probability spaces cannot be identified with probability spaces. However, rather than dealing with the issues 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; presently, we need to explicitly introduce $\mathbf{K}$ in a diagram such as 23 to define it as such. Doing this every time we want a conditional probability makes the text much longer, and introduces mental overhead to the reader. 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

$$\mathsf{X} - \boxed{\mathbb{P}_{\mathsf{Y}|\mathsf{X}}} - \mathsf{Y} \tag{25}$$

satisfies Equation 23 when substituted for $\mathbf{K}$.

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* $\mathcal{F}$.

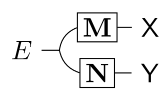
A *conditional probability space* $(\mathcal{K},\Omega,\mathcal{F},D,\mathcal{D})$ is a Markov kernel $\mathcal{K}$, called the *ambient kernel*, along with the sample space Ω and the *input space* D.

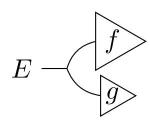
Note that we use the blackboard and script fonts to distinguish ambient measures and kernels. They are formally the same thing as ordinary probability measures and kernels, but it is useful to distinguish them to clarify the conventions we introduce in this section.

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:

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

$f\underline{\otimes}g :=$  (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 $\mathcal{K}\mathbf{F}_{\mathsf{X}}$.

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 $\mathcal{K}_{\mathsf{Y}|\mathsf{X}}$ is the product $\mathcal{K}\mathbb{F}_{\mathsf{X}\underline{\otimes}\mathsf{Y}}$.

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

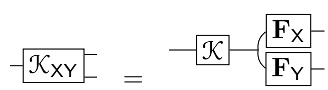
Lemma 2.10 (Joint distributions and coupled products)**.** *Given* $\mathsf{X} : \Omega \to G$ *and* $\mathsf{Y} : \Omega \to H$*,* $\mathbf{F}_{\mathsf{X}\underline{\otimes}\mathsf{Y}} = \mathbf{F}_{\mathsf{X}}\underline{\otimes}\mathbf{F}_{\mathsf{Y}}$

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

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

Equality follows from the monotone class theorem. □

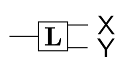
Therefore the following holds:



(32)

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

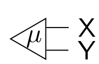
Definition 2.11 (Wire labels - joint probabilities)**.** Given a conditional probability space with ambient kernel $\mathcal{K} : D \to \Delta(\mathcal{F})$ (or a probability space with measure $\mathbb{P}$) and some collection of random variables X, Y, ... on $\Omega \times D$, any diagram representing 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 $\mathbf{L} = \mathbf{K}_{\mathsf{XY}}$, the joint kernel of X and Y.

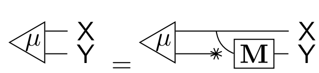
If we have an ambient measure $\mathbb{P}$, then $D = \{*\}$ and the diagram



(34)

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

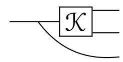
Definition 2.12 (Disintegration)**.** Given a probability space $(\mathbb{P}, \Omega, \mathcal{F})$, random variables X and Y and joint probability measure $\mu := \mathbb{P}_{\mathsf{XY}} \in \Delta(\mathcal{E} \otimes \mathcal{F})$, we say that $\mathbf{M} : E \to \Delta(\mathcal{F})$ is a disintegration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



(35)

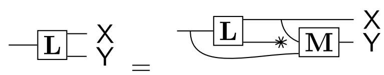
Notationally, $\mathbf{N}$ is a version of $\mathbb{P}_{\mathsf{Y}|\mathsf{X}}$, "the probability of Y given X", or $\mathbf{M} \in \mathbb{P}_{\mathsf{Y}|\mathsf{X}}$.

Given a conditional probability space $(\mathcal{K}, \Omega, D)$, define $\mathcal{K}^*$ to be the kernel



(36)

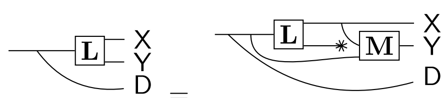
Given random variables X, Y on $\Omega \times D$ and kernel $\mathbf{L} := \mathcal{K}^*_{\mathsf{X},\mathsf{Y}} : D \to \Delta(\mathcal{E} \otimes \mathcal{F})$, we say that $\mathbf{M} : D \times E \to \Delta(\mathcal{F})$ is a disintegration of $\mathbf{L}$ if



(37)

Similarly, recalling that D is the global conditioning variable, we can say $\mathbf{M} \in \mathcal{K}_{\mathsf{Y}|\mathsf{XD}}$. 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



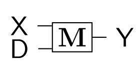
(38)

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 $\mathcal{K}^*_{\mathsf{Y}_1\mathsf{Y}_2...|\mathsf{X}_1\mathsf{X}_2...}$.

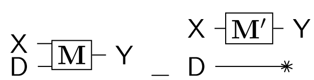
That is, by this convention, the diagram



(39)

implies that $\mathbf{M} \in \mathcal{K}_{\mathsf{Y}|\mathsf{XD}}$. 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 $\mathcal{K}_{\mathsf{Y}|\mathsf{XD}}$ 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)

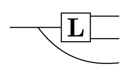
Under these circumstances, we will abuse notation to say $\mathbf{M}' = \mathcal{K}_{\mathsf{Y}|\mathsf{X}}$.

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. $\square$

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.

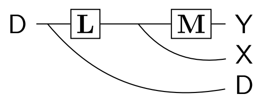
Example 2.16 (Implicit conditional probability space)**.** Suppose we have labeled Markov kernels

$$\mathsf{D} - \boxed{\mathbf{L}} - \mathsf{X} \qquad \mathsf{X} - \boxed{\mathbf{M}} - \mathsf{Y} \tag{42}$$

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:

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

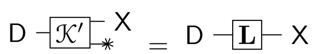
Take $\Omega = W \times X \times Y \times Z$ and define $\mathcal{K}$ such that

$\mathcal{K}^* =$ 

(43)

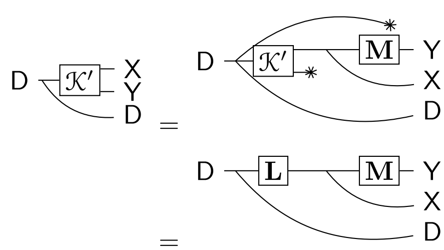
Where $\mathcal{K}^*$ is the copy map composed with $\mathcal{K}$ as in previous definitions. $\mathcal{K}$ is the unique Markov kernel $D \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ supporting the two criteria above, assuming finite D and standard measurable X, Y.

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



(44)

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

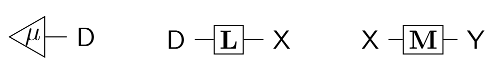


(45)

(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 $\mathbb{P}_{\mathsf{Y|XD}}$. This is not possible in the conditional probability space because $\mathcal{K}_{\mathsf{Y|X}}$ only exists if $\mathcal{K}_{\mathsf{Y|XD}}$ 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.

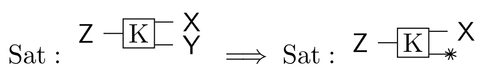
A sufficient condition for the construction of a unique ambient kernel from a collection of components $\{C_1, ...C_n\}$ is if there is some ordering of components $\{i_1, i_2, ...i_n\}$ such that the input labels of $C_{i_{k+1}}$ is the union of the inputs and outputs of $C_{i_1}, ..., C_{i_j}$. This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

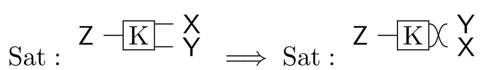
Lemma 2.17 (Diagrammatic consequences of labels)**.** *Fix some conditional probability space* $(\mathcal{K}, \Omega, D)$ *and random variables* X, Y, Z *taking values in arbitrary spaces.* Sat : *indicates that a labeled diagram satisfies definitions 2.11 and 2.13 with respect to* $(\mathcal{K}, \Omega,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

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

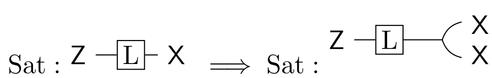
and the following implications hold:



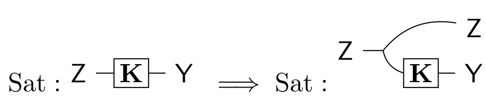
(49)



(50)



(51)

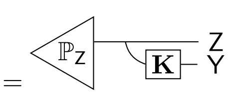


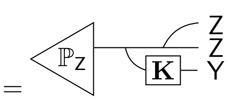
(52)

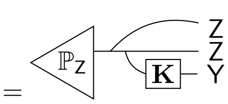
Proof.

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

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

$\mathbb{P}_{\mathsf{ZY}} =$  (53)

$\mathbb{P}_{\mathsf{ZZY}} =$  (54)

$=$  (55)

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

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

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

Consider $\mathsf{Z} = \mathsf{X}$ and $\mathbf{K} = z \mapsto \text{Bernouli}(0.5)$ for all z. Then $\mathbf{L} := x \mapsto \mathbb{P}_{\mathsf{X}} \in \mathbb{P}_{\mathsf{X}|\mathsf{X}}$ and $\mathbf{KL} = z \mapsto \mathbb{P}_{\mathsf{X}}$ but $\mathbb{P}_{\mathsf{X}|\mathsf{Z}} = z \mapsto \delta_z \neq \mathbf{KL}$.

Finally, we have the challenge that our causal theories take *Markov kernels* as basic, rather than probability measures. That is, we suppose we have some set of decisions D and a map $\mathbf{C} : D \to \Delta(\mathcal{E})$ encoding the consequences of taking available decisions, and the decision maker my select a strategy $\gamma \in \Delta(\mathcal{D})$. We solve this problem in the following somewhat inelegant manner: we supppose that the set of decisions D available is always at most countable, and so there is always some $\gamma^* \in \Delta(\mathcal{D})$ that dominates all available strategies. We then take $\gamma^*(\mathrm{Id}_D \underline{\otimes} \mathbf{C})$ to be the joint distribution of D and E (being D and E valued random variables respectively, defined in the obvious way) on some probability space $(\mathbb{P}, \Omega, \mathcal{F})$, and $\mathbf{C}$ is thus the unique conditional $\mathbb{P}_{\mathsf{E}|\mathsf{D}}$. Note that joint distributions of D and anything with respect to $\mathbb{P}$ are not "physically meaningful", they just exist to allow us to embed $\mathbf{C}$ in a probability space.

2.4.3 Alternative approach

It seems like there should be a more direct way of giving meaning to wire labels than going via a probability space as above, particularly as for our purposes it necessitates the limitation of D to a countable set and the introduction of γ^* to support embedding consequence mappings in probability spaces.

An alternative approach could be to begin by defining coherence rules for diagram labels similar to **??** – 52. We could then

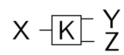
This should probably be a category somehow, but I don't think categories of random variables have actually been worked out

Definition 2.18 (Labelled diagram)**.** Given a measurable space E, suppose it has some factorisation $E = A \times B$. An assignment of *labels* is a map from a countable label set L to an enumeration the chosen factorisation of E, $F := [|\{A, B\}|]$.

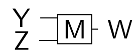
Given a Markov kernel $\mathbf{K} : E \to \Delta(\mathcal{F})$, an abstract labelled diagram $\mathfrak{D}$ is the triple $(\mathrm{Dia}(\mathfrak{D}), \mathrm{In}(\mathfrak{D}), \mathrm{Out}(\mathfrak{D})$ where $\mathrm{Dia}(\mathfrak{D})$ is a string diagram encoding $\mathbf{K}$, along with "input label assignments" $\mathrm{In}(\mathfrak{D})$ and "output label assignments" $\mathrm{Out}(\mathfrak{D})$. We represent $\mathfrak{D}$ with a diagram for $\mathbf{K}$ featuring $|\mathrm{In}(\mathfrak{D})|$ input wires and $|\mathrm{Out}(\mathfrak{D})|$ output wires, with each input and output wire labelled with a label from L. Define $\mathrm{Ker}(\mathfrak{D}) := \mathbf{K}$ to be the .

Given two diagrams $\mathfrak{D}_1$ and $\mathfrak{D}_2$ with $|\mathrm{In}(\mathfrak{D}_2)| = |\mathrm{Out}(\mathfrak{D}_1)|$ (i.e. the number of of output wires of $\mathfrak{D}_1$ is the same as the number of input wires of $\mathfrak{D}_2$) then write $\mathfrak{D}_1$–$\mathfrak{D}_2$ for the diagram formed by connecting corresponding wires of $\mathfrak{D}_1$ and $\mathfrak{D}_2$ preserving the relevant labels.

Example 2.19 (Labelled diagrams)**.** If we have

$\mathfrak{D}_1 :=$ 

(57)

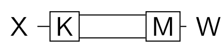
$\mathfrak{D}_2 :=$ 

(58)

then

$$\mathrm{In}(\mathfrak{D}_2) = \begin{cases} \mathsf{V} \mapsto \text{“input wire 1”} \\ \mathsf{W} \mapsto \text{“input wire 2”} \end{cases} \tag{59}$$

and $\mathfrak{D}_1$–$\mathfrak{D}_2$ is the diagram



(60)

Note that we

Definition 2.20 (Namespace)**.** A *labelled kernel space* N is a collection of labelled string diagrams $\{\mathfrak{A}, \mathfrak{B}, \mathfrak{C}, ...\}$ such that the following rules hold:

1. **Composition:** $\mathfrak{A}$–$\mathfrak{B} \in N$ iff $\mathfrak{A} \in N$ and $\mathfrak{B} \in N$ and $\mathrm{Out}(\mathfrak{A}) = \mathrm{In}(\mathfrak{B})$
2. **Unison:** for $\mathfrak{A}, \mathfrak{B} \in N$, if $\mathrm{In}(\mathfrak{A}) = \mathrm{In}(\mathfrak{B})$ and $\mathrm{Out}(\mathfrak{A}) = \mathrm{Out}(\mathfrak{B})$ then $\mathrm{Ker}(\mathfrak{A}) = \mathrm{Ker}(\mathfrak{B})$

In addition,

Copied inputs: if $\mathfrak{A} \in N$ then we also have $\mathfrak{B} \in N$ such that $\mathrm{Ker}(\mathfrak{B}) = \mathrm{Id}_{\mathrm{In}(\mathfrak{A})} \underline{\otimes} \mathrm{Ker}(\mathfrak{A})$

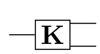
have to deal with the case where we want to work with some Markov kernel $\mathbf{M} : D \to \Delta(\mathcal{E})$ but are uncommitted as to whether this is a

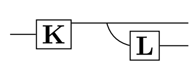
We will begin by defining wire names in the context of *joint probability distributions*, which will then yield an unambiguous meaning for wire names in the context of string diagrams representing probability measaures. We then add a number of coherence rules to derive wire names for general Markov kernels. We first define analogues of 2.9 and 2.12 for Markov kernels.

Definition 2.21 (Pushforward map, joint map)**.** Given a kernel space $(\mathbf{K}, \Omega, E, \mathcal{F}, \mathcal{A})$ and a random variable $\mathsf{X} : \Omega \to G$, the pushforward map is $\mathbf{KF}_{\mathsf{X}}$.

Given $\mathsf{Y} : \Omega \to H$ in addition, the joint map of X and Y is $\mathbf{KF}_{\mathsf{X} \underline{\otimes} \mathsf{Y}}$.

Definition 2.22 (Kernel disintegration)**.** Given a markov kernel $\mathbf{K} : E \to \Delta(\mathcal{F})$, we say that $\mathbf{L}$ is a disintegration of $\mathbf{K}$ if



$=$


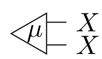
(61)

Lemma 2.23 (Joint distributions from coupled tensor products)**.** *Given a probability space $\langle E, \mathcal{E}, \mu \rangle$ and a finite set of random variables $G = \{\mathsf{X}_i | i \in [n]\}$, the joint distribution of G is given by $\mu(\underline{\bigotimes}_{i \in [n]} \mathbf{F}_{\mathsf{X}_i})$.*

Proof. This follows directly from Definition 2.9 and Lemma 2.3. $\square$

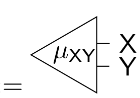
When we define a joint probability distribution μ_{XY} on some produce space $F \times G$, we implicitly define an association between the random variable X and the first factor of the product space $F \times G$, and similarly for Y and the second factor. Consider two random variables $\mathsf{X}_1 : E \to X$ and $\mathsf{X}_2 : E \to X$ and a "unusually named" joint distribution $\mu_{??}$ over $X \times X$. I cannot uniquely associate the elements of a tuple $(a, b) \in X \times X$ with values of X_1 or X_2 - to define this association, I need to either specify it separately to μ or use a standard joint probability notation such as $\mu_{\mathsf{X}_1\mathsf{X}_2}$ or $\mu(\mathsf{X}_1, \mathsf{X}_2)$.

The first purpose of wire names is to unambiguously refer to particular wires in a given diagram. For example, suppose we have some $\mu \in \Delta(\mathcal{X} \times \mathcal{X})$, and we label wires with *spaces* rather than names:



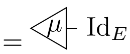
(62)

Given just only the diagram 64, we cannot easily refer to "the top wire" or "the bottom wire". Trying to say something like "the probability of the bottom wire conditional on the top wire" is very confusing, and we really do need to be able to talk about such conditional probabilities (see 2.4.4). Giving wires unique names solves this, but (as suggested by this example), there is another desirable property of wire names: they should function as de-facto random variables, so that "the probability of the bottom wire conditional on the top wire" actually refers to a conditional probability defined in terms of random variables. Suppose we have a probability space $\langle E, \mathcal{E}, \mu \rangle$, and for arbitrary random variables $\mathsf{X} : E \to X$ and $\mathsf{Y} : E \to Y$ write the joint distribution μ_{XY}. We want wire labels to correspond to random variables in the sense that

$\mu_{\mathsf{XY}} :=$ 

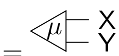
(63)

That is, the correspondence between the product space $X \times Y$ and the values taken by the random variables X and Y implicit in the definition of μ_{XY} is reflected by the wire names on the right hand side of 63. Given this definition, if we have some finite set of random variables $S = \{\mathsf{X}, \mathsf{Y}, ..., \mathsf{Z}\}$ such that $\mu_{\mathsf{XY...Z}} = \mu$, then we must be able to represent μ in a diagram with $|S|$ output wires (as the joint distribution is by definition on an appropriate product space), and we should label the wires of this diagram with $\mathsf{X}, \mathsf{Y}, ...\mathsf{Z}$. Note that $\{\mathrm{Id}_E\}$ always satisfies this criterion, thus we can always draw μ with a single output wire labeled Id_E.

$\mu =$ 

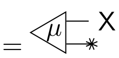
(64)

In general, if we have $E = X \times Y$, we can define $\mathsf{X} : X \times Y \to X$ and $\mathsf{Y} : X \times Y \to Y$ by the projection maps $\mathsf{X} : (x, y) \mapsto x$, $\mathsf{Y} : (x, y) \mapsto y$ and $\mu_{\mathsf{XY}} = \mu$ and we can write (see Lemma 2.26):

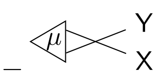
$\mu =$ 

(65)

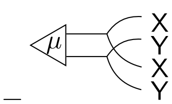
If we take 64 to *define* X and Y, then Equation 63 compels the following labels for products involving μ:

$\mu_{\mathsf{X}} =$ 

(66)

$\mu_{\mathsf{YX}} =$ 

(67)

$\mu_{\mathsf{XYXY}} =$ 

(68)

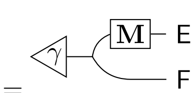
This illustrates the logic of the representations of the identity 11, the swap 13 and the copy maps 12: their representations visually preserve the identities of wires that should be identified according to Equation 63. Note that the presence of a copy map as in 68 is the *only* time when a diagram will feature identical labels on wires.

Free Random Variables We are interested in working with general Markov kernels, not just probability measures, and we will thus not always have an ambient probability space as in 63 to ground our wire labels.

Definition 2.24 (Free Random Variables)**.** Given an ambient Markov kernel $\mathbf{M} : E \to \Delta(\mathcal{F})$, a free random variable X is a measurable function on $\mathcal{E} \otimes \mathcal{F}$.

The equivalent of marginal and joint distributions for free random variables are marginal and joint *maps*.

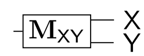
Definition 2.25 (Marginal, joint maps)**.** Given a Markov kernel $\mathbf{M} : E \to \Delta(\mathcal{F})$, note that given $\gamma \in \Delta(\mathcal{E})$, by our definitions, $\gamma \prec \mathbf{M} := \gamma \curlyvee (\mathrm{Id}_E \otimes \mathbf{M})$ is a probability measure on $\Delta(\mathcal{E} \otimes \mathcal{F})$ and

$\gamma \prec \mathbf{M} =$ 

(69)

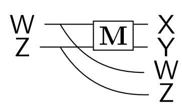
Given $\mathbf{M} : E \to \Delta(\mathcal{F})$ and a free random variable X, the marginal map $\mathbf{M}_{\mathsf{X}} : E \to \Delta(\mathcal{X})$ is the unique Markov kernel such that for all $\gamma \in \Delta(\mathcal{E})$, $\gamma(\mathbf{M}_{\mathsf{X}}) = (\gamma \prec \mathbf{M})_{\mathsf{X}}$ where the right hand side is an ordinary marginal distribution. Similarly, given free random variables X, Y, the joint map $\mathbf{M}_{\mathsf{XY}} : E \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ is the unique Markov kernel such that $\gamma \in \Delta(\mathcal{E})$, $\gamma(\mathbf{M}_{\mathsf{XY}}) = (\gamma \prec \mathbf{M})_{\mathsf{XY}}$.

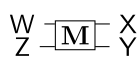
We are now placed to impose a criterion on wire labels equivalent to 63: given the ambient Markov kernel $\mathbf{M} : E \to \Delta(\mathcal{F})$, wire labels in diagrams representing joint distributions must correspond in the sense of 63:

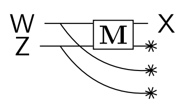
$\mathbf{M}_{\mathsf{XY}} :=$  (70)

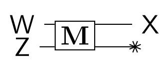
In addition, we impose the requirement of identity 11, swap 13 and copy maps 12 - "output" wires that are connected to "input" wires with no boxes in between share names.

Wire labels for general kernels behave similarly to those for probability measures. Suppose $E = X \times Y$ and $F = W \times Z$ and define $\mathsf{X}, \mathsf{Y}, \mathsf{W}, \mathsf{Z}$ as projection maps from $X \times Y \times W \times Z$ to their respective spaces with $\mathbf{M}$ as before. Then the following labels are compelled:

$\mathrm{Id}_E \prec \mathbf{M} =$  (71)

$\mathbf{M} =$  (72)

$\mathbf{M}_{\mathsf{X}} =$  (73)

$=$  (74)

As well as properties analogous to Equations 68 and 67.

Derivations of label properties

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

$$
\begin{aligned}
(\mathsf{X})_{\#}\mu\mathbf{A}(B) &= \mu\mathbf{A}(\mathsf{X}^{-1}(B)) &(75)\\
&= \int_F \delta_{\mathsf{X}(a)}(B) d\mu\mathbf{A}(a) &(76)\\
&= \mu\mathbf{A}\mathbf{F}_{\mathsf{X}}(B) &(77)
\end{aligned}
$$

□

Lemma 2.26 (Coupled projection maps are equal to the identity)**.** *Suppose E is a finite Cartesian product: $E = \prod_{i\in[n]} A_i$. Let $\pi_i : E \to A_i$ be the projection map $(a_1, .., a_i, .., a_n) \mapsto a_i$. Then $\underline{\bigotimes}_{i\in[n]}\pi_i = \mathrm{Id}_E$ where Id_E is the identity function on E.*

Proof. Define $\pi_{[m]} : E \to \prod_{i\in[m]} A_i$ by $(a_1, ..., a_m, ..., a_n) \mapsto (a_1, ..., a_m)$. Suppose $\underline{\bigotimes}_{i\in[n-1]}\pi_i = \pi_{[n-1]}$. Then by associativity of $\underline{\otimes}$, $\underline{\bigotimes}_{i\in[n]}\pi_i = \pi_{[n-1]}\underline{\otimes}\pi_n$ and for all $(a_1, ..., a_n) \in E$, $\pi_{[n-1]}\underline{\otimes}\pi_n(a_1, ..., a_n) = (\pi_{[n-1]}(a_1, ..., a_n), \pi_n(a_1, ..., a_n)) = (a_1, ..., a_{n-1}, a_n) = \pi_{[n]}(a_1, ..., a_n)$.

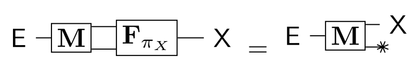
Also, $\underline{\bigotimes}_{i\in[1]}\pi_i = \pi_1$, thus $\underline{\bigotimes}_{i\in[n]}\pi_i = \pi_{[n]}$. But $\pi_{[n]} = \mathrm{Id}_E$. $\square$

Corollary 2.27. *If we have a probability space $\langle E, \mathcal{E}, \mu\rangle$ where $E = \prod_{i\in[n]} A_i$ and $\mathsf{X}_i := \pi_i$, then $\mu_{\underline{\bigotimes}_{i\in[n]}\mathsf{X}_i} = \mu$.*

Lemma 2.28 (A projection is the identity tensored with the erase map)**.** *Let $\pi_X : X \times Y \to X$ be the projection $\pi_X : (x, y) \mapsto x$. Then $\mathbf{F}_{\pi_x} = \mathbf{F}_{\mathrm{Id}_X} \otimes *$*

Proof. $\mathbf{F}_{\pi_X}$ is, by definition, the Markov kernel $(x, y) \mapsto \delta_x$, which is equivalent to $\mathbf{F}_{\mathrm{Id}_X} \otimes *$. $\square$

Corollary 2.29. *For any $\mathbf{M} : E \to \Delta(\mathcal{X} \otimes \mathcal{Y})$,*



(78)

We take a random variable to be a measurable function on a *kernel space* $\langle F, G, \mathcal{F} \otimes \mathcal{G}, \mathbf{M}\rangle$ 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; in general we have only conditional probabilities (which we define in Paragraph 2.4.4). This approach mirrors to some extent the approach suggested by Hájek (2003), which also takes conditional probability to be fundamental.

This choice is largely pragmatic - it is helpful to make statements about the properties of random variables on a kernel space without quantifying over prior distributions. However, there is a connection between this choice and the philosophical field of decision theory. In particular, we use random variables to model both stochastic quantities and quantities that depend on the decision maker’s choices. We treat the uncertainty associated with choosing and that associated with stochasticity as different – we do *not* suppose that uncertainty over which choice will be made should itself be modeled using probability. *Evidential decision theory*, as defended by Jeffrey (1981), proposes that is is proper to consider choices to be random variables, though it 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.

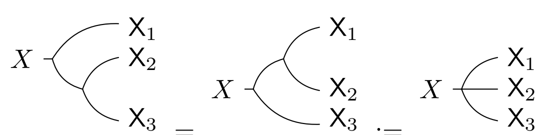
2.4.4 Working With String Diagrams

todo:

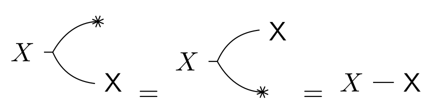
- Functional generalisation
- Conditioning
- Infinite copy map
- De Finetti's representation theorem

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.

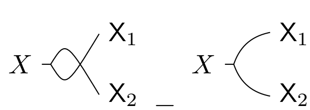
Axioms of Symmetric Monoidal Categories Recalling the unique Markov kernels defined above, the following equivalences, known as the *commutative comonoid axioms*, hold among string diagrams:



(79)



(80)

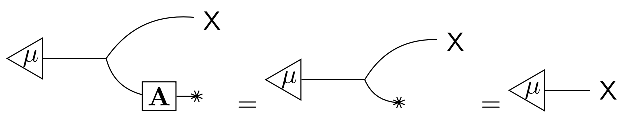


(81)

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

(82)

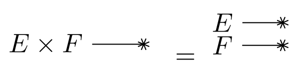
Combining 80 and 82 we can derive the following: integrating $\mathbf{A} : X \to \Delta(\mathcal{Y})$ with respect to $\mu \in \Delta(\mathcal{X})$ and then discarding the output of $\mathbf{A}$ leaves us with μ:



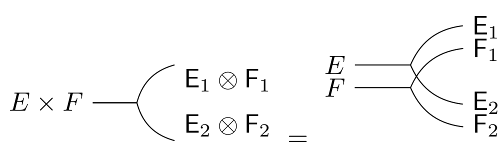
(83)

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 $\curlyvee$:

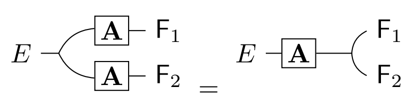


(84)



(85)

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



(86)

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

Definition 2.30 (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 2.31 (Copy map commutes for deterministic kernels (Fong, 2013))**.** *Equation 86 holds iff* $\mathbf{A}$ *is deterministic.*

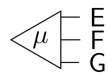
This needs updating in light of the more thorough treatment of variables and wire labels

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.

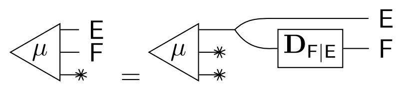
We'll proceed from an example to a general definition.

Example 2.32 (Disintegration with respect to "convenient" random variables)**.** Given a probability measure $\mu \in \Delta(\mathcal{E} \otimes \mathcal{F} \otimes \mathcal{G})$:



(87)

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



(88)

Equation 88 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 2.32 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.

Lemma 2.33. *For measurable functions* $\mathsf{X} : E \to X$ *and* $\mathsf{Y} : E \to Y$*,* $\mathbf{F}_{\mathsf{X}\underline{\otimes}\mathsf{Y}} = \mathbf{F}_{\mathsf{X}}\underline{\otimes}\mathbf{F}_{\mathsf{Y}}$*.*

Proof. $\mathcal{X} \otimes \mathcal{Y}$ is by definition generated by the rectangles $A \times B$ for $A \in \mathcal{X}$, $B \in \mathcal{Y}$. To show equivalence of kernels $E \to \Delta(\mathcal{X} \otimes \mathcal{Y})$ it is sufficient to show agreement on all $s \in E$ and rectangles $A \times B$, $A \in \mathcal{X}$, $B \in \mathcal{Y}$.

For all $q \in X$, $r \in Y$, $A \in \mathcal{X}$ and $B \in \mathcal{Y}$, we have

$$\delta_{(q,r)}(A \times B) = \delta_q(A)\delta_r(B) \tag{89}$$

This can be verified by checking the four combinations of q in or not in A and r in or not in B.

For all $s \in E$, $A \in \mathcal{X}, B \in \mathcal{Y}$,

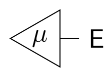
$$\mathbf{F}_{\mathsf{X}\underline{\otimes}\mathsf{Y}}(s; A \times B) = \delta_{\mathsf{X}(s),\mathsf{Y}(s)}(A \times B) \tag{90}$$

$$= \delta_{\mathsf{X}(s)}(A)\delta_{\mathsf{Y}(s)}(B) \tag{91}$$

$$= (\mathbf{F}_{\mathsf{X}}\underline{\otimes}\mathbf{F}_{\mathsf{Y}})(s; A \times B) \tag{92}$$

□

Definition 2.34 (Disintegration and Conditional Probability)**.** Given a probability measure $\mu \in \Delta(\mathcal{E})$:

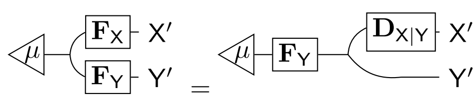


(93)

and two groups of random variables $G_X = \{\mathsf{X}_i | i \in [n]\}$ and $G_Y = \{\mathsf{Y}_i | i \in [m]\}$, define $\mathsf{X} = \underline{\otimes}_{i\in[n]}\mathsf{X}_i$, $\mathsf{Y} := \underline{\otimes}_{i\in[m]}\mathsf{Y}_i$ and $\mathsf{W} = \mathsf{X}\underline{\otimes}\mathsf{Y}$. Define the $\mathsf{X}|\mathsf{Y}$ disintegration of μ to be the $\mathsf{X}'|\mathsf{Y}'$ disintegration of $\mu\mathbf{F}_{\mathsf{W}}$, where $\mathsf{X}' : \prod_{i\in[n]} X_i \times \prod_{j\in[m]} Y_j \to \prod_{i\in[n]} X_i$ and $\mathsf{Y}' : \prod_{i\in[n]} X_i \times \prod_{j\in[m]} Y_j \to \prod_{i\in[n]} Y_i$ are the respective projection maps $\mathsf{X}' : (x_0, ..., x_[n], y_0, ..., y_[m]) \mapsto (x_0, ..., x_m)$ and $\mathsf{Y}' : (x_0, ..., x_[n], y_0, ..., y_[m]) \mapsto (y_0, ..., y_[m])$.

By construction, X' and Y' are such that $\mathsf{X}' \circ \mathsf{W} = \mathsf{X}$ and $\mathsf{Y}' \circ \mathsf{W} = \mathsf{Y}$.

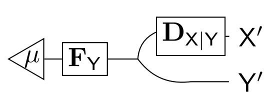
By Lemma 2.33, $\mathbf{D}_{\mathsf{X}|\mathsf{Y}}$ is a $\mathsf{X}|\mathsf{Y}$ disintegration of μ if and only if



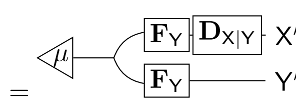
(94)

Equation 94 generalises 88 beyond disintegrations by "convenience" wire names to disintegrations by arbitrary sets of random variables. Note that where we *can* construct a diagram for μ with convenient labels for X and Y, due to the identification between random variables and wires, 94 reduces to 88, and we will use the simpler definition where possible as it yields less cluttered diagrams.

From the fact that $\mathbf{F}_{\mathsf{Y}}$ is a deterministic kernel, we also have:



(95)



(96)

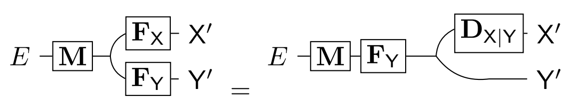
We note, without proof, that: $\mathbf{F}_{\mathsf{Y}}\mathbf{D}_{\mathsf{X}|\mathsf{Y}}$ is an ordinary conditional probability $\mu(\mathsf{X}| \otimes_{i\in[m]} \mathcal{Y}_i)$, and where such an ordinary conditional probability exists we can find a disintegration $\mathbf{D}_{\mathsf{X}|\mathsf{Y}}$ (Çinlar, 2011).

In addition, it is well known that disintegrations of μ are non-unique, and different disintegrations are equal only up to sets of μ-measure 0. This ambiguity turns out to be more problematic in our causal work as sets of μ-measure 0 may meaningfully impact the consequences of decisions in fairly ordinary circumstances.

We are also interested in an analogue of disintegration that applies to an arbitrary Markov kernel $\mathbf{M} : E \to \Delta(\mathcal{F})$ rather than only to probability measures. A generalised notion of disintegraion will allow for a formal definition of Dawid (2010)'s definition of "extended conditional independence".

There are a number of choices we could make here, and we have made a particular choice that leads to generalised disintegrations usually failing to exist. This choices allows for a much cleaner treatment of conditional independence, as we will explain below.

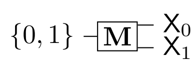
Definition 2.35 (Generalised Disintegration)**.** Given a Markov kernel $\mathbf{M} : E \to \Delta(\mathcal{F})$ and random variables $\mathsf{X} : F \to X$, $\mathsf{Y} : F \to Y$, a $\mathsf{X}|\mathsf{Y}$ disintegration of $\mathbf{M}$ is any kernel $\mathbf{D}_{\mathsf{X}|\mathsf{Y}}$ such that



(97)

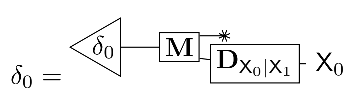
Similarly to the standard definition, this reduces in the case that $\mathbf{M}$ can be drawn with X and Y labelling output wires.

Example 2.36 (Generalised disintegrations usually do not exist)**.** Let $\mathbf{M} : \{0,1\} \to \Delta(\{0,1\}^2)$ be defined by $\mathbf{M} : q \mapsto \delta_q \otimes \delta_1$ and let $\mathsf{X}_0, \mathsf{X}_1 : \{0,1\}^2 \to \{0,1\}$ be defined by $\mathsf{X}_0 : (r,s) \mapsto r$ and $\mathsf{X}_1 : (r,s) \mapsto s$. Note that we can write:

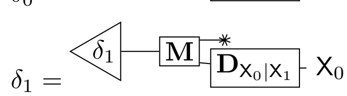


(98)

Suppose we have some $\mathbf{D}_{\mathsf{X}_0|\mathsf{X}_1}$. Then we must have

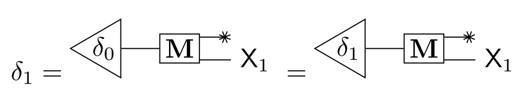


(99)



(100)

but by assumption



(101)

Thus no such $\mathbf{D}_{\mathsf{X}_0|\mathsf{X}_1}$ can exist.

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`.

A. Philip Dawid. Beware of the DAG! In *Causality: Objectives and Assessment*, pages 59–86, February 2010. URL `http://proceedings.mlr.press/v6/dawid10a.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`.

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`.

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: