Disintegration and Bayesian Inversion, Both Abstractly and Concretely[†]

KENTA CHO and BART JACOBS

Institute for Computing and Information Sciences Radboud University

P.O.Box 9010, 6500 GL Nijmegen, the Netherlands

Received 4 September 2017

The notions of disintegration and Bayesian inversion are fundamental in conditional probability theory. They produce channels, as conditional probabilities, from a joint state, or from an already given channel (in opposite direction). These notions exist in the literature, in concrete situations, but are presented here in abstract graphical formulations. The resulting abstract descriptions are used for proving basic results in conditional probability theory. The existence of disintegration and Bayesian inversion is discussed for discrete probability, and also for measure-theoretic probability — via standard Borel spaces and via likelihoods. Finally, the usefulness of disintegration and Bayesian inversion is illustrated in several non-trivial examples.

1. Introduction

The essence of conditional probability can be summarised informally in the following equation about probability distributions:

$$\textit{joint} = \textit{conditional} \cdot \textit{marginal}.$$

A bit more precisely, when we have joint probabilities $\mathrm{P}(x, y)$ for elements x, y ranging over two sample spaces, the above equation splits into two equations,

$$\mathrm{P}(y \mid x) \cdot \mathrm{P}(x) = \mathrm{P}(x, y) = \mathrm{P}(x \mid y) \cdot \mathrm{P}(y), \tag{1}$$

where $\mathrm{P}(x)$ and $\mathrm{P}(y)$ describe the marginals, which are obtained by discarding variables. We see that conditional probabilities $\mathrm{P}(y \mid x)$ and $\mathrm{P}(x \mid y)$ can be constructed in two directions, namely y given x, and x given y. We also see that we need to copy variables: x on the left-hand-side of the equations (1), and y on the right-hand-side.

Conditional probabilities play a crucial role in Bayesian probability theory. They form the nodes of Bayesian networks (Pearl, 1988; Bernardo and Smith, 2000; Barber, 2012),

[†] The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement nr. 320571.

which reflect the conditional independences of the underlying joint distribution via their graph structure. As part of our approach, we shall capture conditional independence in an abstract manner.

The main notion of this paper is *disintegration*. It is the process of extracting a conditional probability from a joint probability. Disintegration, as we shall formalise it here, gives a structural description of the above equation (1) in terms of *states* and *channels*. In general terms, a *state* is a probability distribution of some sort (discrete, measure-theoretic, or even quantum) and a *channel* is a map or morphism in a probabilistic setting, like $\mathrm{P}(y \mid x)$ and $\mathrm{P}(x \mid y)$ as used above. It can take the form of a stochastic matrix, probabilistic transition system, Markov kernel, conditional probability table (in a Bayesian network), or morphism in a Kleisli category of a 'probability monad' (Jacobs, 2017). A state is a special kind of channel, with trivial domain. Thus we can work in a monoidal category of channels, where we need discarding and copying — more formally, a comonoid structure on each object — in order to express the above conditional probability equations (1).

In this article we abstract away from interpretation details and will describe disintegration pictorially, in the language of string diagrams. This language can be seen as the internal language of symmetric monoidal categories (Selinger, 2010) — with comonoids in our case. The essence of disintegration becomes: extracting a conditional probability channel from a joint state.

A categorical approach to Bayesian conditioning has appeared for instance in (Culbertson and Sturtz, 2014; Staton et al., 2016; Clerc et al., 2017) and in (Jacobs et al., 2015; Jacobs and Zanasi, 2016; Jacobs, 2017). The latter references use effectus theory (Jacobs, 2015; Cho et al., 2015), a new comprehensive approach aimed at covering the logic of both quantum theory and probability theory, supported by a Python-based tool 'EfProb', for 'effectus probability'. This tool is used for the (computationally extensive) examples in this paper.

Disintegration, also known as regular conditional probability, is a notoriously difficult operation in measure-theoretic probability, see *e.g.* (Pollard, 2002; Panangaden, 2009; Chang and Pollard, 1997): it may not exist (Stoyanov, 2014); even if it exists it may be determined only up to negligible sets; and it may not be continuous or computable (Ackerman et al., 2011). Disintegration has been studied using categorical language in (Culbertson and Sturtz, 2014), which focuses on a specific category of probabilistic mappings. Our approach here is more axiomatic.

We thus describe disintegration as going from a joint state to a channel. A closely related concept is *Bayesian inversion*: it turns a channel (with a state) into a channel in opposite direction. We show how Bayesian inversion can be understood and expressed easily in terms of disintegration — and also how, in the other direction, disintegration can be obtained from Bayesian inversion. Bayesian inversion is taken as primitive notion in (Clerc et al., 2017). Here we start from disintegration. The difference is a matter of choice.

Bayesian inversion is crucial for backward inference. We explain it informally: let σ be a state of a domain/type X, and $c\colon X \to Y$ be a channel; Bayesian inversion yields a channel $d\colon Y \to X$. Informally, it produces for an element $y \in Y$, seen as singleton/point predicate $\{y\}$, the conditioning of the state σ with the pulled back evidence $c^{-1}(\{y\})$. A concrete example involving such 'point observations' will be described at the end of Section 8. More generally, disintegration and Bayesian inversion are used to structurally organise state updates in presence of new evidence in probabilistic programming, see *e.g.* (Gordon et al., 2014; Borgström et al., 2013; Staton et al., 2016; Katoen et al., 2015). See also (Shan and Ramsey, 2017), where disintegration is handled via symbolic manipulation.

Disintegration and Bayesian inversion are relatively easy to define in discrete probability theory. The situation is much more difficult in measure-theoretic probability theory, first of all because point predicates $\{y\}$ do not make much sense there, see also (Chang and Pollard, 1997). A common solution to the problem of the existence of disintegration / Bayesian inversion is to restrict ourselves to standard Borel spaces, as in (Clerc et al., 2017). We take this approach too. There is still an issue that disintegration is determined only up to negligible sets. We address this by defining 'almost equality' in our abstract pictorial formulation. This allows us to present a fundamental result from (Clerc et al., 2017) abstractly in our setting, see Section 5.

Another common, more concrete solution is to assume a *likelihood*, that is, a probabilistic relation $X \times Y \to \mathbb{R}_{\geq 0}$. Such a likelihood gives rise to probability density function (pdf), providing a good handle on the situation, see (Pawitan, 2001). The technical core of Section 8 is a generalisation of this likelihood-based approach.

The paper is organised as follows. It starts with a brief introduction to the graphical language that we shall be using, and to the underlying monoidal categories with discarding and copying. Then, Section 3 introduces both disintegration and Bayesian inversion in this graphical language, and relates the two notions. Subsequently, Section 4 contains an elaborated example, namely of naive Bayesian classification. A standard example from the literature (Witten et al., 2011) is redescribed in the current setting: first, channels are extracted via disintegration from a table with given data; next, Bayesian inversion is applied to the combined extracted channels, giving the required classification. This is illustrated in both the discrete and the continuous version of the example.

Next, Section 5 is more technical and elaborates the standard equality notion of 'equal almost everywhere' in the current setting. This is used for describing Bayesian inversion in a more formal way, following (Clerc et al., 2017). Section 6 uses our graphical approach to review conditional independence and to prove at an abstract level several known results, namely equivalence of various formulations of conditional independence, and the 'graphoid' axioms from (Verma and Pearl, 1988; Geiger et al., 1990). Section 7 relaxes the requirement that maps are causal, so that 'effects' can be used as the duals of states for validity and conditioning. The main result relates conditioning of joint states to forward and backward inference via the extracted channels, in the style of (Jacobs and Zanasi, 2016); it is illustrated in a concrete example, where a Bayesian network is seen as a graph in a Kleisli category — following (Fong, 2012). Finally, Section 8 gives the likelihood formulation of disintegration and inversion, as briefly described above.

2. Graphical language

The basic idea underlying this paper is to describe probability theory in terms of *channels*. A channel $f: X \to Y$ is a (stochastic) process from a system of type X into that of Y. Concretely, it may be a probability matrix or kernel. Our standing assumption is that types (as objects) and channels (as arrows) form a *symmetric monoidal category*. For the formal definition we refer to (Mac Lane, 1998). We informally summarise that we have the following constructions.

1 Sequential composition $g \circ f: X \to Z$ for appropriately typed channels $f: X \to Y$ and $g: Y \to Z$.

2 Parallel composition $f \otimes g: X \otimes Z \to Y \otimes W$ for $f: X \to Y$ and $g: Z \to W$. This involves composition of types $X \otimes Z$.

3 Identity channels $\mathrm{id}_X: X \to X$, which 'do nothing'. Thus $\mathrm{id} \circ f = f = \mathrm{id} \circ f$.

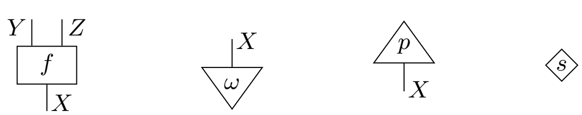
4 A unit type I, which represents 'no system'. Thus $I \otimes X \cong X \cong X \otimes I$.

5 Swap isomorphisms $X \otimes Y \cong Y \otimes X$ and associativity isomorphisms $(X \otimes Y) \otimes Z \cong X \otimes (Y \otimes Z)$, so the ordering in composed types does not matter.

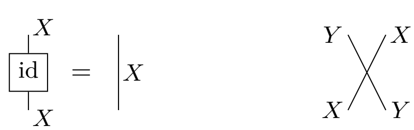
Representation of such channels in the ordinary 'formula' notation easily becomes complex and thus reasoning becomes hard to follow. A graphical language known as *string diagrams* offers a more convenient and intuitive way of reasoning in a symmetric monoidal category.

In string diagrams, types/objects are represented as wires '|', with information flowing bottom to top. The composition of types is depicted by juxtaposition of wires, and the unit type is 'no diagram' as below.

Channels/arrows are represented by boxes with an input wire(s) and an output wire(s), in upward direction. When a box does not have input or output, we write it as a triangle or diamond. For example, $f: X \to Y \otimes Z$, $\omega: I \to X$, $p: X \to I$, and $s: I \to I$ are respectively depicted as:



The identity channels are represented by 'no box', *i.e.* just wires, and the swap isomorphisms are represented by crossing of wires:

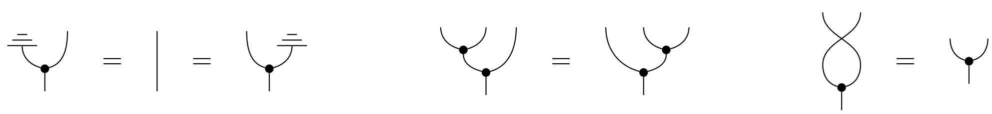


Finally, the sequential composition of channels is depicted by connecting the input and output wires, and the parallel composition is given by juxtaposition, respectively as below:

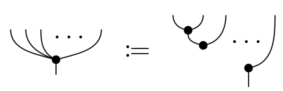
The use of string diagrams is justified by the following 'coherence' theorem, see (Selinger, 2010).

Theorem 2.1. A well-formed equation between composites of arrows in a symmetric monoidal category follows from the axioms of symmetric monoidal categories if and only if the string diagrams of both sides are equal up to isomorphism of diagrams. □

We further assume the following structure in our category. For each type X there are a discarder $\top_X \colon X \to I$ and a copier $\curlyvee_X \colon X \to X \otimes X$. They are required to satisfy the following equations:



This says that $(\curlyvee_X, \top_X)$ forms a commutative comonoid on X. By the associativity we may write:



Moreover we assume that the comonoid structures $(\curlyvee_X, \top_X)$ are compatible with the monoidal structure $(\otimes, I)$, in the sense that the following equations hold.

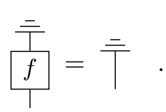
Note that we do *not* assume that these maps are natural. Explicitly, we do not necessarily have $\curlyvee \circ f = (f \otimes f) \circ \curlyvee$ or $\top \circ f = \top$.

We will use these symmetric monoidal categories throughout in the paper. For convenience, we introduce a term for them.

Definition 2.2. A *CD-category* is a symmetric monoidal category $(\mathbf{C}, \otimes, I)$ with a commutative comonoid $(\curlyvee_X, \top_X)$ for each $X \in \mathbf{C}$, suitably compatible as described above.

Here 'CD' stands for Copy/Discard.

Definition 2.3. An arrow $f \colon X \to Y$ in a CD-category is said to be *causal* if



A CD-category is *affine* if all the arrows are causal, or equivalently, the tensor unit I is a final object.

The term 'causal' comes from (categorical) quantum foundation (Coecke and Kissinger, 2017; D'Ariano et al., 2017), and is related to relativistic causality, see *e.g.* (Coecke, 2016).

We reserve the term 'channel' for causal arrows. Explicitly, causal arrows $c\colon X \to Y$ in a CD-category are called *channels*. A channel $\omega\colon I \to X$ with input type I is called a *state* (on X). For the time being we only consider affine CD-categories, where all arrows are channels.

Example 2.4. Our main examples of affine CD-categories are two Kleisli categories $\mathcal{K}\ell(\mathcal{D})$ and $\mathcal{K}\ell(\mathcal{G})$, respectively, for discrete probability, and more general, measure-theoretic probability.

1 What we call a *distribution* or a *state* over a set X is a finite subset $\{x_1, x_2, \dots, x_n\} \subseteq X$, called the support where each element x_i occurs with a multiplicity $r_i \in [0,1]$, such that $\sum_i r_i = 1$. Such a convex combination is often written as $r_1|x_1\rangle + \cdots + r_n|x_n\rangle$ with $r_i \in [0,1]$. The ket notation $|-\rangle$ is meaningless syntactic sugar that is used to distinguish elements $x \in X$ from occurrences in such formal sums. Notice that a distribution can also be written as a function $\omega\colon X \to [0,1]$ with finite support $\mathrm{supp}(\omega) = \{x \in X \mid \omega(x) \neq 0\}$. We shall write $\mathcal{D}(X)$ for the set of distributions over X. This $\mathcal{D}$ is a monad on the category **Set** of sets and functions.
A function $f\colon X \to \mathcal{D}(Y)$ is called a *Kleisli* map; it forms a channel $X \to Y$. Such maps can be composed as matrices, for which we use special notation $\odot$.

$$(g \odot f)(x)(z) \,=\, \textstyle\sum_{y\in Y} f(x)(y) \cdot g(y)(z) \qquad \text{for } g\colon Y \to \mathcal{D}(Z) \text{ with } x \in X, z \in Z.$$

We write $1 = \{*\}$ for a singleton set, and $2 = 1 + 1 = \{0,1\}$. Notice that $\mathcal{D}(1) \cong 1$ and $\mathcal{D}(2) \cong [0,1]$. We can identify a state on X with a channel $1 \to X$.
The monad $\mathcal{D}$ is known to be *commutative*. This implies that finite products of sets $X \times Y$ give rise to a symmetric monoidal structure on the Kleisli category $\mathcal{K}\ell(\mathcal{D})$. Specifically, for two maps $f\colon X \to \mathcal{D}(Y)$ and $g\colon Z \to \mathcal{D}(W)$, the tensor product / parallel composition $f \otimes g\colon X \times Z \to \mathcal{D}(Y \times W)$ is given by:

$$(f \otimes g)(x,z)(y,w) \,=\, f(x,y) \cdot g(z,w).$$

For each set X there are a copier $\curlyvee_X\colon X \to \mathcal{D}(X \times X)$ and a discarder $\bar{\top}_X\colon X \to \mathcal{D}(1)$ given by $\curlyvee_X(x) = 1|x,x\rangle$ and $\bar{\top}_X(x) = 1|*\rangle$, respectively. They come from the cartesian (finite product) structure of the base category **Set**, through the obvious functor $\mathbf{Set} \to \mathcal{K}\ell(\mathcal{D})$. Therefore $\mathcal{K}\ell(\mathcal{D})$ is a CD-category. It is moreover affine, since the monad is affine in the sense that $\mathcal{D}(1) \cong 1$.

2 Let $X = (X, \Sigma_X)$ be a measurable space, where Σ_X is a σ-algebra on X. A *probability measure*, also called a *state*, on X is a function $\omega\colon \Sigma_X \to [0,1]$ which is countably additive and satisfies $\omega(X) = 1$. We write $\mathcal{G}(X)$ for the collection of all such probability measures on X. This set $\mathcal{G}(X)$ is itself a measurable space. Notice that $\mathcal{G}(X) \cong \mathcal{D}(X)$ when X is a finite set (as discrete space). In particular, $\mathcal{G}(2) \cong \mathcal{D}(2) \cong [0,1]$. This $\mathcal{G}$ is a monad on the category **Meas** of measurable spaces, with measurable functions between them; it is called the Giry monad, after (Giry, 1982; Jacobs, 2017).
A Kleisli map, that is, a measurable function $f\colon X \to \mathcal{G}(Y)$ is a channel (or a *probability kernel*, see Example 7.1). These channels can be composed, via Kleisli composition $\odot$,

using integration:

$$(g \circ f)(x)(C) = \int_Y g(y)(C)\, f(x)(\mathrm{d}y) \quad \text{where } g\colon Y \to \mathcal{G}(Z) \text{ and } x \in X, C \in \Sigma_Z.$$

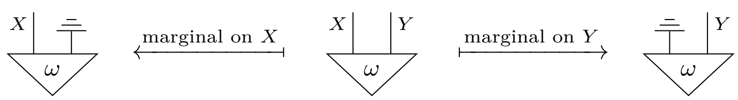
It is well-known that the monad $\mathcal{G}$ is commutative and affine, see also (Jacobs, 2017). Thus, in a similar manner to the previous example, the Kleisli category $\mathcal{K}\ell(\mathcal{G})$ is an affine CD-category. The parallel composition $f \otimes g\colon X \times Z \to \mathcal{G}(Y \times W)$ for $f\colon X \to \mathcal{G}(Y)$ and $g\colon Z \to \mathcal{G}(W)$ is given as:

$$(f \otimes g)(x, z)(B \times D) \;=\; f(x)(B) \cdot g(z)(D),$$

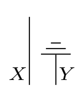
for $x \in X$, $z \in Z$, $B \in \Sigma_Y$, and $D \in \Sigma_W$. This indeed determines a unique measure $(f \otimes g)(x, z) \in \mathcal{G}(Y \times W)$, which is a product measure of probability measures $f(x)$ and $g(z)$.

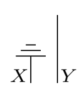
3. Marginalisation, integration and disintegration

Let $\mathbf{C}$ be an affine CD-category. We think of states $\omega\colon I \to X$ in $\mathbf{C}$ as abstract (probability) distributions on type X. States of the form $\omega\colon I \to X \otimes Y$, often called (bipartite) joint states, are seen as joint distributions on X and Y. Later on we shall also consider n-partite joint states, but for the time being we restrict ourselves to bipartite ones. For a joint distribution $\mathrm{P}(x, y)$ in ordinary discrete probability, we can calculate the marginal distribution on X by summing (or marginalising) Y out, as $\mathrm{P}(x) = \sum_y \mathrm{P}(x, y)$. The marginal distribution on Y is also calculated by $\mathrm{P}(y) = \sum_x \mathrm{P}(x, y)$. In our abstract setting, given a joint state $\omega\colon I \to X \otimes Y$, we can obtain marginal states simply by discarding wires, as in:



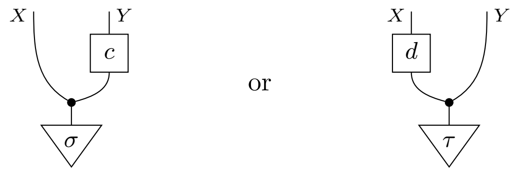
In other words, the marginal states are the state ω composed with the projection maps $\pi_1\colon X \otimes Y \to X$ and $\pi_2\colon X \otimes Y \to Y$, as below.

$\pi_1 :=$ 

$\pi_2 :=$ 

Example 3.1. For a joint state $\omega \in \mathcal{D}(X \times X)$ in $\mathcal{K}\ell(\mathcal{D})$, the first marginal $\omega_1 = \pi_1 \circ \omega$ is given by $\omega_1(x) = \sum_{y \in Y} \omega(x, y)$, as expected. For a joint state $\omega \in \mathcal{G}(X \times X)$ in $\mathcal{K}\ell(\mathcal{G})$, the first marginal is given by $\omega_1(A) = \omega(A \times Y)$ for $A \in \Sigma_X$. The second marginals are similar.

A channel $c\colon X \to Y$ is seen as an abstract *conditional* distribution $\mathrm{P}(y|x)$. In ordinary probability theory, we can calculate a joint distribution $\mathrm{P}(x, y)$ from a distribution $\mathrm{P}(x)$ and a conditional distribution $\mathrm{P}(y|x)$ by the formula $\mathrm{P}(x, y) = \mathrm{P}(y|x) \cdot \mathrm{P}(x)$, which is often called the *product rule*. Similarly we have $\mathrm{P}(x, y) = \mathrm{P}(x|y) \cdot \mathrm{P}(y)$. In our setting, starting from a state $\sigma\colon I \to X$ and a channel $c\colon X \to Y$, or a state $\tau\colon I \to Y$ and a channel $d\colon Y \to X$, we can 'integrate' them into a joint state on $X \otimes Y$ as follows, respectively:

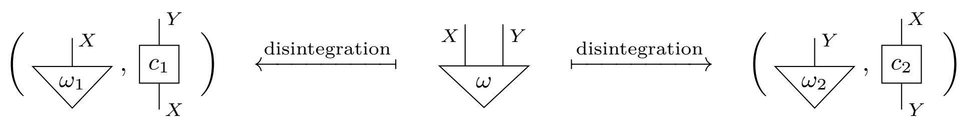


(2)

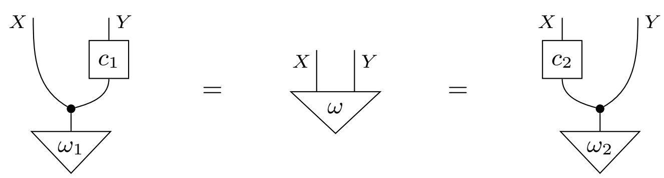
Example 3.2. Let $\sigma \in \mathcal{D}(X)$ and $c\colon X \to \mathcal{D}(Y)$ be a state and a channel in $\mathcal{K}\ell(\mathcal{D})$. An easy calculation verifies that $\omega = (\mathrm{id} \otimes c) \odot \curlyvee \odot \sigma$, the joint state on $X \times Y$ defined as in (2), satisfies $\omega(x, y) = c(x)(y) \cdot \sigma(x)$, as we expect from the product rule.

For a state $\sigma \in \mathcal{G}(X)$ and a channel $c\colon X \to \mathcal{G}(Y)$ in $\mathcal{K}\ell(\mathcal{G})$, the joint state $\omega = (\mathrm{id} \otimes c) \odot \curlyvee \odot \sigma$ is given by $\omega(A \times B) = \int_A c(x)(B)\, \sigma(\mathrm{d}x)$ for $A \in \Sigma_X$ and $B \in \Sigma_Y$. This 'integration' construction of a joint probability measure is standard, see *e.g.* (Pollard, 2002; Panangaden, 2009).

Disintegration is an inverse operation of the 'integration' of a state and a channel into a joint state, as in (2). More specifically, it starts from a joint state $\omega\colon I \to X \otimes Y$ and extracts either a state $\omega_1\colon I \to X$ and a channel $c_1\colon X \to Y$, or a state $\omega_2\colon I \to Y$ and a channel $c_2\colon Y \to X$ as below,

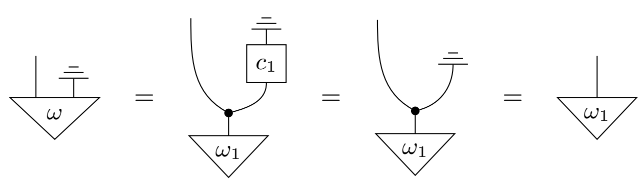


such that the equation on the left or right below holds, respectively.

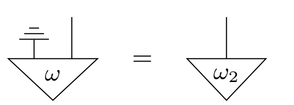


(3)

We immediately see from the equation that ω_1 and ω_2 must be marginals of ω:



and similarly



We are thus led to the following definition.

Definition 3.3. Let $\omega\colon I \to X \otimes Y$ be a joint state. A channel $c_1\colon X \to Y$ (or $c_2\colon Y \to X$) is called a *disintegration* of ω if it satisfies the equation (3) with ω_i the marginals of ω.

Example 3.4. Let $\omega \in \mathcal{D}(X \times Y)$ be a joint state in $\mathcal{K}\ell(\mathcal{D})$. We write $\omega_1 \in \mathcal{D}(X)$ for the first marginal, given by $\omega_1(x) = \sum_y \omega(x,y)$. Then a channel $c\colon X \to \mathcal{D}(Y)$ is a disintegration of ω if and only if $\omega(x,y) = c(x)(y) \cdot \omega_1(x)$ for all $x \in X$ and $y \in Y$. It turns out that there is always such a channel c. We define a channel c by:

$$c(x)(y) \coloneqq \frac{\omega(x,y)}{\omega_1(x)} \qquad \text{if} \quad \omega_1(x) \neq 0 \ ,$$

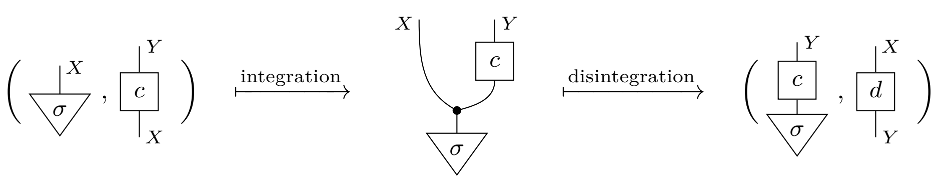
and $c(x) \coloneqq \tau$ if $\omega_1(x) = 0$, for an arbitrary state $\tau \in \mathcal{D}(Y)$. (We may assume that Y is nonempty.) This indeed defines a channel c satisfying the required equation. Roughly speaking, disintegration in discrete probability is nothing but the 'definition' of conditional probability: $\mathrm{P}(y|x) = \mathrm{P}(x,y)/\,\mathrm{P}(x)$. There is still some subtlety — disintegrations need not be unique, when there are $x \in X$ with $\omega_1(x) = 0$.

Disintegrations in measure-theoretic probability, in $\mathcal{K}\ell(\mathcal{G})$, are far more difficult. Let $\omega \in \mathcal{G}(X \times Y)$ be a joint state, with $\omega_1 \in \mathcal{G}(X)$ the first marginal. A channel $c\colon X \to \mathcal{G}(Y)$ is a disintegration of ω if and only if

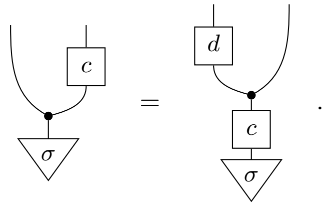
$$\omega(A \times B) = \int_A c(x)(B)\,\omega_1(\mathrm{d}x)$$

for all $A \in \Sigma_X$ and $B \in \Sigma_Y$. This is the ordinary notion of *disintegration* (of probability measures), also known as *regular conditional probability*; see *e.g.* (Faden, 1985; Pollard, 2002; Panangaden, 2009). We see that there is no obvious way to obtain a channel c here, unlike the discrete case. In fact, a disintegration may not exist (Stoyanov, 2014). There are, however, a number of results that guarantee the existence of a disintegration in certain situations. We will come back to this issue later in the section.

Bayesian inversion is a special form of disintegration, occurring frequently. We start from a state $\sigma\colon I \to X$ and a channel $c\colon X \to Y$. We then integrate them into a joint state on X and Y, and disintegrate it in the other direction, as below.



We call the disintegration $d\colon Y \to X$ a *Bayesian inversion* for $\sigma\colon I \to X$ along $c\colon X \to Y$. By unfolding the definitions, a channel $d\colon Y \to X$ is a Bayesian inversion if and only if



(4)

Example 3.5. Let $\sigma \in \mathcal{D}(X)$ and $c\colon X \to \mathcal{D}(Y)$ be a state and a channel in $\mathcal{K}\ell(\mathcal{D})$. Then a channel $d\colon Y \to \mathcal{D}(X)$ is a Bayesian inversion for σ along c if and only if $c(x)(y) \cdot \sigma(x) = d(y)(x) \cdot c_*(\sigma)(y)$, where $c_*(\sigma)(y) = \sum_{x'} c(x')(y) \cdot \sigma(x')$. In a similar manner to Example 3.4, we can obtain such a d by:

$$d(y)(x) := \frac{c(x)(y) \cdot \sigma(x)}{c_*(\sigma)(y)} = \frac{c(x)(y) \cdot \sigma(x)}{\sum_{x'} c(x')(y) \cdot \sigma(x')}$$

for $y \in Y$ with $c_*(\sigma)(y) \neq 0$. For $y \in Y$ with $c_*(\sigma)(y) = 0$, we may define $d(y)$ to be an arbitrary state in $\mathcal{D}(X)$. We can recognise the above formula as the Bayes formula:

$$\mathrm{P}(x|y) = \frac{\mathrm{P}(y|x) \cdot \mathrm{P}(x)}{\mathrm{P}(y)} = \frac{\mathrm{P}(y|x) \cdot \mathrm{P}(x)}{\sum_{x'} \mathrm{P}(y|x') \cdot \mathrm{P}(x')} \ .$$

Let $\sigma \in \mathcal{G}(X)$ and $c\colon X \to \mathcal{G}(Y)$ be a state and a channel in $\mathcal{K}\ell(\mathcal{G})$. A channel $d\colon Y \to \mathcal{G}(X)$ is a Bayesian inversion if and only if

$$\int_A c(x)(B)\, \sigma(\mathrm{d}x) = \int_B d(y)(A)\, c_*(\sigma)(\mathrm{d}y)$$

for all $A \in \Sigma_X$ and $B \in \Sigma_Y$. Here $c_*(\sigma) \in \mathcal{G}(Y)$ is the measure given by $c_*(\sigma)(B) = \int_X c(x)(B)\, \sigma(\mathrm{d}x)$. As we see below, Bayesian inversions are in some sense equivalent to disintegrations, and thus, they are as difficult as disintegrations. In particular, a Bayesian inversion need not exist.

In practice, however, the state σ and channel c are often given via density functions. This setting, so-called (absolutely) continuous probability, makes it easy to compute a Bayesian inversion. Suppose that X and Y are subspaces of $\mathbb{R}$, and that σ and c admit density functions as

$$\sigma(A) = \int_A f(x)\, \mathrm{d}x \qquad\qquad c(x)(B) = \int_B \ell(x,y)\, \mathrm{d}y$$

for measurable functions $f\colon X \to \mathbb{R}_{\geq 0}$ and $\ell\colon X \times Y \to \mathbb{R}_{\geq 0}$. The conditional probability density $\ell(x,y)$ of y given x is often called the *likelihood* of x given y. By the familiar Bayes formula for densities — see *e.g.* (Bernardo and Smith, 2000) — the conditional density of x given y is:

$$k(y,x) := \frac{\ell(x,y) \cdot f(x)}{\int_X \ell(x',y) \cdot f(x')\, \mathrm{d}x'} \ .$$

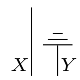
This k then gives a channel $d\colon Y \to \mathcal{G}(X)$ by

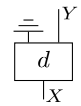
$$d(y)(A) = \int_A k(y,x)\, \mathrm{d}x$$

for each $y \in Y$ such that $\int_X \ell(x',y) \cdot f(x')\, \mathrm{d}x' \neq 0$. For the other y's we define $d(y)$ to be some fixed state in $\mathcal{G}(X)$. An elementary calculation verifies that d is indeed a Bayesian inversion for σ along c. Later, in Section 8, we generalise this calculation into our abstract setting.

Although Bayesian inversions are a special case of disintegrations, we can conversely obtain disintegrations from Bayesian inversions, as in the proposition below. Therefore, in some sense the two notions are equivalent.

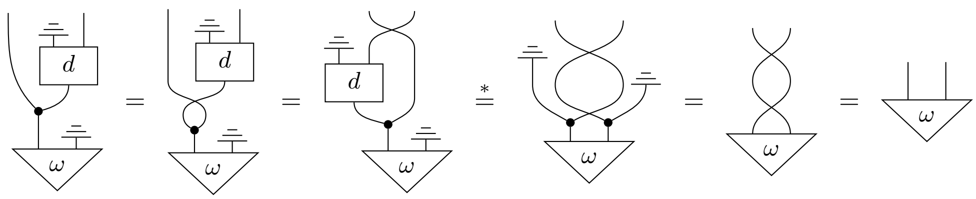
Proposition 3.6. Let ω be a state on $X \otimes Y$. Let $d\colon X \to X \otimes Y$ be a Bayesian inversion for ω along the first projection $\pi_1 \colon X \otimes Y \to X$ on the left below.

$\pi_1 =$ 

$\pi_2 \circ d =$ 

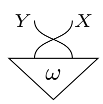
Then the composite $\pi_2 \circ d \colon X \to Y$ shown on the right above is a disintegration of ω.

Proof. We prove that the first equation in (3) holds for $c_1 = \pi_2 \circ d$, as follows.



For the marked equality $\overset{*}{=}$ we used the equation (4) for the Bayesian inversion d. □

We say that an affine CD-category $\mathbf{C}$ *admits disintegration* if for every bipartite state $\omega \colon I \to X \otimes Y$ there exist a disintegration $c_1 \colon X \to Y$ of ω. Note that in such categories there also exists a disintegration $c_2 \colon Y \to X$ of ω in the other direction, since it can be obtained as a disintegration of the following state:

 .

By Proposition 3.6, admitting disintegration is equivalent to admitting Bayesian inversion.

In Example 3.4, we have seen that $\mathcal{K}\ell(\mathcal{D})$ admits disintegration, but that in measure-theoretic probability, in $\mathcal{K}\ell(\mathcal{G})$, disintegrations may not exist. There are however a number of results that guarantee the existence of disintegrations in specific situations, see *e.g.* (Pachl, 1978; Faden, 1985). We here invoke one of these results and show that there is a subcategory of $\mathcal{K}\ell(\mathcal{G})$ that admits disintegration. A measurable space is called a *standard Borel space* if it is measurably isomorphic to a Polish space with its Borel σ-algebra, or equivalently, if it is measurably isomorphic to a Borel subspace of $\mathbb{R}$. Then the following theorem is standard, see *e.g.* (Pollard, 2002, §5.2) or (Faden, 1985, §5).

Theorem 3.7. Let X be any measurable space and Y be a standard Borel space. Then for any state (*i.e.* a probability measure) $\omega \in \mathcal{G}(X \times Y)$ in $\mathcal{K}\ell(\mathcal{G})$, there exists a disintegration $c_1 \colon X \to \mathcal{G}(Y)$ of ω. □

Let $\mathbf{pKrn}_{\mathrm{sb}}$ be the full subcategory of $\mathcal{K}\ell(\mathcal{G})$ consisting of standard Borel spaces as objects. It is easy to see that $\mathbf{pKrn}_{\mathrm{sb}}$ is an affine CD-category. Then the previous theorem immediately shows:

Corollary 3.8. The category $\mathbf{pKrn}_{\mathrm{sb}}$ admits disintegration. □

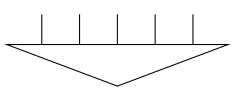
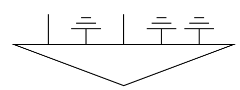
We note that $\mathbf{pKrn}_{\mathrm{sb}}$ can also be seen as the Kleisli category of the Giry monad restricted on the category of standard Borel spaces.

Since there are various ‘existence’ theorems like Theorem 3.7, there may be other subcategories of $\mathcal{K}\ell(\mathcal{G})$ that admit disintegration. A likely candidate is the category of perfect probabilistic mappings in (Culbertson and Sturtz, 2014). We do not go into this question here, since $\mathbf{pKrn}_{\mathrm{sb}}$ suffices for the present paper.

4. Example: naive Bayesian classifiers via inversion

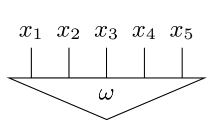
Bayesian classification is a well-known technique in machine learning that produces a distribution over data classifications, given certain sample data. The distribution describes the probability, for each data (classification) category, that the sample data is in that category. Here we consider an example of ‘naive’ Bayesian classification, where the features are assumed to be independent. We consider a standard classification example from the literature which forms an ideal setting to illustrate the use of both disintegration and Bayesian inversion. Disintegration is used to extract channels from a given table, and subsequently Bayesian inversion is applied to (the tuple of) these channels to obtain the actual classification. The use of channels and disintegration/inversion in this classification setting is new, as far as we know.

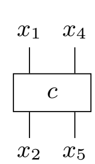
For the calculations in this example we use the *EfProb* library (Cho and Jacobs, 2017), of which we explain the notation for marginalisation and disintegration. There are many ways to marginalise an n-partite state, namely one for each subset of the wires $\{1, 2, \ldots, n\}$. In *EfProb* such a subset is described as a *mask*, consisting of a list of n zero’s or one’s, where a zero at position i means that the i-th wire/component is marginalised out, and a one at position i means that it remains. Such a mask $M = [b_1, \ldots, b_n]$ with $b_i \in \{0, 1\}$ is used in *EfProb* as a post-fix operation in $\omega \ \% \ M$ on an n-partite state ω. An example explains it all:

if $\omega =$  then $\omega \ \% \ [1, 0, 1, 0, 0] =$ 

In a similar way one can disintegrate an n-partite state in 2^n may ways, where a mask of length n is now used to describe which wires are used as input to the extracted channel and which ones as output. In *EfProb* this is written as $\omega \ /\!\!/ \ M$, where M is a mask, as above. A systematic description will be given in Section 6 below.

In practice it is often useful to be able to marginalise first, and disintegrate next. The general description in n-ary form is a bit complicated, so we use an example for $n = 5$. We shall label the wires with x_i, as on the left below. We seek the conditional probability written conventionally as $c = \omega[x_1, x_4 \mid x_2, x_5]$ on the right below.

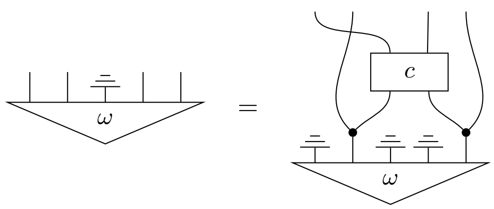




Outlook	**Temperature**	**Humidity**	**Windy**	**Play**
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

Fig. 1. Weather and play data, copied from (Witten et al., 2011).

This channel c must satisfy:



This picture shows how to obtain the channel c from ω: we first marginalise to restrict to the relevant wires x_1, x_2, x_4, x_5. This is written as ω % $[1,1,0,1,1]$. Subsequently we disintegrate with x_1, x_4 as output and x_2, x_5 as input. Hence:

$$\begin{aligned} c &\coloneqq \omega \;\%\; [1,1,0,1,1] \;/\!\!/\; [0,1,0,1] \\ &= \;\omega\big[[1,0,0,1,0] : [0,1,0,0,1]\big] \quad \text{in } \textit{EfProb} \text{ notation.} \end{aligned}$$

We see that the latter *EfProb* post-fix $\big[[1,0,0,1,0] : [0,1,0,0,1]\big]$ is a 'variable free' version of the traditional notation $[x_1, x_4 \mid x_2, x_5]$, selecting the relevant positions — with | replaced by :.

We have now prepared the ground and can turn to the classification example that we announced. It involves the classification of 'playing' (yes or no) for certain weather data, used in (Witten et al., 2011). We shall first go through the discrete example in some detail. The relevant data are in the table in Figure 1. The question is: given this table, what can be said about the probability of playing if the outlook is *Sunny*, the temperature is *Cold*, the humidity is *High* and it is Windy?

Our plan is to first organise these table data into four channels d_O, d_T, d_H, d_W in a network of the form:

$$\begin{tikzcd}
\boxed{\text{Outlook}} & \boxed{\text{Temperature}} & & \boxed{\text{Humidity}} & \boxed{\text{Windy}} \\
& & \boxed{\text{Play}} \arrow[ull, "d_O"] \arrow[ul, "d_T"'] \arrow[ur, "d_H"] \arrow[urr, "d_W"']
\end{tikzcd} \tag{5}$$

The abstraction of these channels works by disintegration. The representation in *EfProb* starts by defining the relevant domains for the categories in the table in Figure 1. We choose abbreviations for the entries in each of the categories.

```
>>> Outlook = ['S', 'O', 'R']
>>> Temp = ['H', 'M', 'C']
>>> Humidity = ['H', 'N']
>>> Windy = ['t', 'f']
>>> Play = ['y', 'n']
>>> D = [Outlook, Temp, Humidity, Windy, Play]
```

This last domain `D` combines the previous ones into a single domain. It is used for the representation of the table, where each of the 14 lines in Figure 1 gets a probability 1/14.

```
>>> table = 1/14  *  point_state(('S','H','H','f','n'), D) \
...         + 1/14 * point_state(('S','H','H','t','n'), D) \
...         + 1/14 * point_state(('O','H','H','f','y'), D) \
...         + 1/14 * point_state(('R','M','H','f','y'), D) \
...         + 1/14 * point_state(('R','C','N','f','y'), D) \
...         + 1/14 * point_state(('R','C','N','t','n'), D) \
...         + 1/14 * point_state(('O','C','N','t','y'), D) \
...         + 1/14 * point_state(('S','M','H','f','n'), D) \
...         + 1/14 * point_state(('S','C','N','f','y'), D) \
...         + 1/14 * point_state(('R','M','N','f','y'), D) \
...         + 1/14 * point_state(('S','M','N','t','y'), D) \
...         + 1/14 * point_state(('O','M','H','t','y'), D) \
...         + 1/14 * point_state(('O','H','N','f','y'), D) \
...         + 1/14 * point_state(('R','M','H','t','n'), D)
```

In this way the table is transformed into a joint probability distribution on the (5-ary) domain `D`. Here, the transformation from table to distribution was done by hand, but it is easy enough to automate this process.

We extract the four channels in Diagram (5) via appropriate disintegrations, from the Play column to the Outlook / Temperature / Humidity / Windy columns.

```
>>> dO = table[[1,0,0,0,0] : [0,0,0,0,1]]
>>> dT = table[[0,1,0,0,0] : [0,0,0,0,1]]
>>> dH = table[[0,0,1,0,0] : [0,0,0,0,1]]
```

```
>>> dW = table[[0,0,0,1,0] : [0,0,0,0,1]]
```

Thus, as described in the beginning of this section, the 'outlook' channel `dO` is extracted by first marginalising the table to the relevant wires, and then disintegrating. Explicitly, `dO` is `table % [1,0,0,0,1] // [0,1]`.

In a next step we combine these four channels into a single channel `d` via tupling. The answer that we are looking for will be obtained by Bayesian inversion of this channel `d`. But Bayesian inversion requires an additional initial state. For this we take the 'Play' marginal of the table, in the fifth position.

```
>>> d = (dO @ dT @ dH @ dW) * copy(Play,4)
>>> prior_play = table % [0,0,0,0,1]
>>> prior_play
0.643|y> + 0.357|n>
>>> posterior_play = d.inversion(prior_play)('S','C','H','t')
>>> posterior_play
0.205|y> + 0.795|n>
```

Notice that the assumptions — *Sunny* outlook, *Cold* temperature, *High* humidity, *true* windiness — are used as input to the inversion of `d`. The resulting classification probability of 0.205 coincides with the probability of 20.5% that is computed in (Witten et al., 2011) — in a rather *ad hoc* manner, without much of a theoretical basis.

One could complain that our approach is 'too' abstract, since it remains magical what these extracted channels do. We elaborate the outlook channel `dO`, going from the `Play` to the `Outlook` domain. We print the two probability distributions for the two values `y` and `n` of the `Play` domain:

```
>>> dO('y')
0.222|S> + 0.444|O> + 0.333|R>
>>> dO('n')
0.6|S> + 0|O> + 0.4|R>
```

The first distribution $\frac{2}{9}|S\rangle + \frac{4}{9}|O\rangle + \frac{3}{9}|R\rangle$ arises as follows. We need to concentrate on the 9 lines in Figure 1 for which Play is *yes*; in these lines, in the first Outlook column, 2 out of 9 entries are *Sunny*, 4 out of 9 are *Overcast*, and 3 out of 9 are *Rainy*. This corresponds to the first distribution `dO('y')`. Similarly, the second distribution captures the Outlook for the 5 lines where Play is *no*: 3 out of 5 are *Sunny* and 2 out of 5 are *Rainy*.

There is a 'continuous' variation of this example where numerical values are used for temperature and humidity. We shall not repeat the table and refer to (Witten et al., 2011) for details. We use Bayesian inversion, as before, for classification, but we need a slightly different approach to extract channels for the 'continuous' features. One first computes the mean and standard deviation of the given numerical values, separately when Play is *yes*, and when Play is *no*. These means and standard deviations are used as parameters for Gaussian (normal) distributions in the *EfProb* channel definitions below. The mean and standard deviation values are copied from (Witten et al., 2011). We only define continuous channels for temperature and humidity, and re-use the discrete channels for outlook and windiness to form a single 'hybrid' channel `c` that will be inverted for classification:

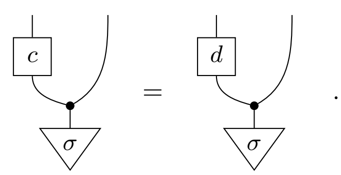
```
>>> cT = chan_from_states([gaussian_state(73, 6.2),
...                        gaussian_state(74.6, 7.9)], Play)
>>> cH = chan_from_states([gaussian_state(79.1, 10.2),
...                        gaussian_state(86.2, 9.7)], Play)
>>> c = (dO @ cT @ cH @ dW) * copy(Play,4)
>>> c.inversion(prior_play)('S',66,90,'t')
0.207|y> + 0.793|n>
```

The latter inversion computation produces the probability of 0.207 for playing when the outlook is *Sunny*, the temperature is *66* (Fahrenheit), the humidity is *90%* and the windiness is *true*. The value computed in (Witten et al., 2011) is 20.8%. The minor difference of 0.001 with our outcome can be attributed to (intermediate) rounding errors.

5. Almost equality of channels

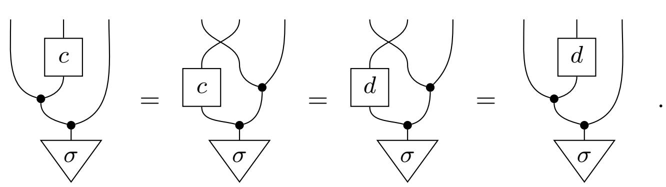
This section explains how the standard notion of 'equal up to negligible sets' or 'equal almost everywhere' (with respect to a measure) can be expressed abstractly using string diagrams. Via this equality relation Bayesian inversion can be characterised very neatly, following (Clerc et al., 2017). We consider an affine CD-category, continuing in the setting of Section 3.

Definition 5.1. Let $c, d\colon X \to Y$ be two parallel channels, and $\sigma\colon I \to X$ be a state on their domain. We say that c is *σ-almost equal* to d, written as $c \overset{\sigma}{\sim} d$ if

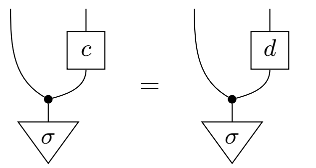


It is obvious that $\overset{\sigma}{\sim}$ is an equivalence relation on channels of type $X \to Y$. When S is a set of arrows of type $X \to Y$, we write S/σ for the quotient $S/\overset{\sigma}{\sim}$.

To put it more intuitively, we have $c \overset{\sigma}{\sim} d$ iff c and d can be identified whenever the input wires are connected to σ, possibly through copiers. For instance, using the associativity and commutativity of copiers, by $c \overset{\sigma}{\sim} d$ we may reason as:



In particular, $c \overset{\sigma}{\sim} d$ if and only if



Now the following is an obvious consequence from the definition.

Proposition 5.2. If both $c, d\colon X \to Y$ are disintegrations of a joint state $\omega\colon I \to X \otimes Y$, then $c \overset{\omega_1}{\sim} d$, where $\omega_1\colon I \to X$ is the first marginal of ω. $\square$

For channels $f, g\colon X \to \mathcal{D}(Y)$ and a state $\sigma \in \mathcal{D}(X)$ in $\mathcal{K}\ell(\mathcal{D})$, it is easy to see that $f \overset{\sigma}{\sim} g$ if and only if $f(x)(y) \cdot \sigma(x) = g(x)(y) \cdot \sigma(x)$ for all $x \in X$ and $y \in Y$ if and only if $f(x) = g(x)$ for any $x \in X$ with $\sigma(x) \neq 0$. Almost equality in $\mathcal{K}\ell(\mathcal{G})$ is less trivial but characterised in an expected way.

Proposition 5.3. Let $f, g\colon X \to \mathcal{G}(Y)$ be channels and $\mu \in \mathcal{G}(X)$ a state in $\mathcal{K}\ell(\mathcal{G})$. Then $f \overset{\mu}{\sim} g$ if and only if for any $B \in \Sigma_Y$, $f(-)(B) = g(-)(B)$ μ-almost everywhere.

Proof. By expanding the definition, $f \overset{\mu}{\sim} g$ if and only if

$$\int_A f(x)(B)\,\mu(\mathrm{d}x) = \int_A g(x)(B)\,\mu(\mathrm{d}x)$$

for all $A \in \Sigma_X$ and $B \in \Sigma_Y$. This is equivalent to $f(-)(B) = g(-)(B)$ μ-almost everywhere for all $B \in \Sigma_Y$, see (Fremlin, 2000, 131H). $\square$

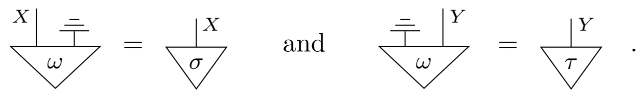
Almost-everywhere equality of probability kernels $f, g\colon X \to \mathcal{G}(Y)$ is often formulated by the stronger condition that $f = g$ μ-almost everywhere. The next proposition shows that the stronger variant is equivalent under a reasonable assumption (any standard Borel space is countably generated, for example).

Proposition 5.4. In the setting of the previous proposition, additionally assume that the measurable space Y is countably generated. Then $f \overset{\mu}{\sim} g$ if and only if $f = g$ μ-almost everywhere.

Proof. Let the σ-algebra Σ_Y on Y be generated by a countable family $(B_n)_n$. We may assume that $(B_n)_n$ is a π-system, *i.e.* a family closed under binary intersections. Let $A_n = \{x \in X \mid f(x)(B_n) = g(x)(B_n)\}$, and $A = \bigcap_n A_n$. Each A_n is μ-conegligible, and thus A is μ-conegligible. For each $x \in A$, we have $f(x)(B_n) = g(x)(B_n)$ for all n. By application of the Dynkin π-λ theorem, it follows that $f(x) = g(x)$. Therefore $f = g$ μ-almost everywhere. $\square$

We can now present a fundamental result from (Clerc et al., 2017, §3.3) in our abstract setting. Let $(I \downarrow \mathbf{C})$ be the comma (coslice) category. for an affine CD-category $\mathbf{C}$. Objects in $(I \downarrow \mathbf{C})$ are states in $\mathbf{C}$, formally pairs (X, σ) of objects $X \in \mathbf{C}$ and states $\sigma\colon I \to X$. Arrows from (X, σ) to (Y, τ) are *state-preserving* channels $c\colon X \to Y$ in $\mathbf{C}$

satisfying $c \circ \sigma = \tau$. A joint state $(X \otimes Y, \omega) \in (I \downarrow \mathbf{C})$ is called a *coupling* of two states $(X, \sigma), (Y, \tau) \in (I \downarrow \mathbf{C})$ if



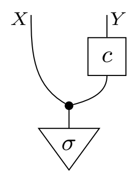
We write $\mathrm{Coupl}((X, \sigma), (Y, \tau))$ for the set of couplings of (X, σ) and (Y, τ).

Theorem 5.5. Let $\mathbf{C}$ be an affine CD-category that admits disintegration. For each pair of states $(X, \sigma), (Y, \tau) \in (I \downarrow \mathbf{C})$, there is the following bijection:

$$(I \downarrow \mathbf{C})\big((X, \sigma), (Y, \tau)\big)/\sigma \quad \cong \quad \mathrm{Coupl}((X, \sigma), (Y, \tau))$$

Proof. For each $c \in (I \downarrow \mathbf{C})\big((X, \sigma), (Y, \tau)\big)$, we define a joint state $I \to X \otimes Y$ to be the 'integration' of σ and c as below, for which we use the following *ad hoc* notation:

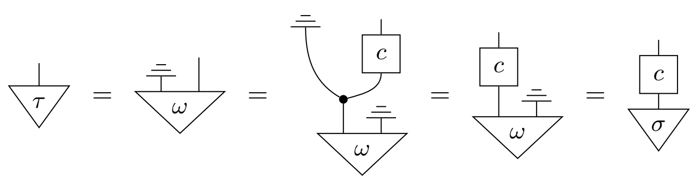
$$\sigma \ogreaterthan c :=$$



It is easy to check that $\sigma \ogreaterthan c$ is a coupling of σ and τ. For two channels $c, d \colon X \to Y$, we have $\sigma \ogreaterthan c = \sigma \ogreaterthan d$ if and only if $c \overset{\sigma}{\sim} d$, by the definition of $\overset{\sigma}{\sim}$. This means the mapping

$$c \longmapsto \sigma \ogreaterthan c\,, \quad (I \downarrow \mathbf{C})\big((X, \sigma), (Y, \tau)\big)/\sigma \longrightarrow \mathrm{Coupl}((X, \sigma), (Y, \tau))$$

is well-defined and injective. To prove the surjectivity let $(X \otimes Y, \omega) \in \mathrm{Coupl}((X, \sigma), (Y, \tau))$. Let $c \colon X \to Y$ be a disintegration of ω. Then c is state-preserving since



Moreover we have $\sigma \ogreaterthan c = \omega$, as desired. $\square$

Via the symmetry $X \otimes Y \overset{\cong}{\to} Y \otimes X$ we have the obvious bijection $\mathrm{Coupl}((X, \sigma), (Y, \tau)) \cong \mathrm{Coupl}((Y, \tau), (X, \sigma))$. This immediately gives the following corollary.

Corollary 5.6. Let $\mathbf{C}$ be an affine CD-category that admits disintegration. For any states $(X, \sigma), (Y, \tau) \in (I \downarrow \mathbf{C})$ we have

$$(I \downarrow \mathbf{C})\big((X, \sigma), (Y, \tau)\big)/\sigma \quad \cong \quad (I \downarrow \mathbf{C})\big((Y, \tau), (X, \sigma)\big)/\tau$$

The bijection sends a channel $c \colon X \to Y$ to a Bayesian inversion $d \colon Y \to X$ for σ along c. $\square$

Theorem 2 of (Clerc et al., 2017) is obtained as an instance, for the category $\mathbf{pKrn}_{\mathrm{sb}}$. This bijective correspondence yields a 'dagger' $(-)^\dagger$ functor on (a suitable quotient of) the comma category $(I \downarrow \mathbf{C})$ — as noted by the authors of (Clerc et al., 2017).

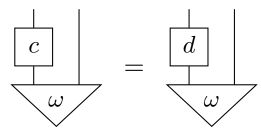
5.1. *Equality extension property*

The section concerns a property that allows us to extend almost-equality w.r.t. some state to a larger state which has the smaller state as marginal. This property is convenient for equational reasoning between string diagrams, and used later in the proof of Theorem 8.3.

Definition 5.7. We say an affine CD-category has the *equality extension property* if for any joint state $\omega\colon I \to X \otimes Z$ and for any channels $c, d\colon X \to Y$, $c \overset{\omega_1}{\sim} d$ implies $c \otimes \mathrm{id}_Z \overset{\omega}{\sim} d \otimes \mathrm{id}_Z$, where $\omega_1 = \pi_1 \circ \omega$ is the first marginal of ω.

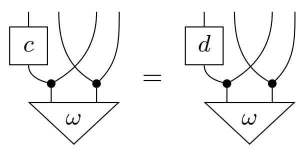
Lemma 5.8. An affine CD-category has the equality extension property if and only if for any joint state $\omega\colon I \to X \otimes Z$, and for any channels $c, d\colon X \to Y$,

$c \overset{\omega_1}{\sim} d$ implies

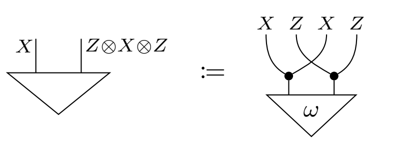


where $\omega_1 = \pi_1 \circ \omega$.

Proof. The 'only if' is obvious, since $c \otimes \mathrm{id}_Z \overset{\omega}{\sim} d \otimes \mathrm{id}_Z$ implies $(c \otimes \mathrm{id}_Z) \circ \omega = (d \otimes \mathrm{id}_Z) \circ \omega$. To prove the 'if', assume $c \overset{\omega_1}{\sim} d$ for a state $\omega\colon I \to X \otimes Z$ and for channels $c, d\colon X \to Y$. We have to prove $c \otimes \mathrm{id}_Z \overset{\omega}{\sim} d \otimes \mathrm{id}_Z$, *i.e.*



This follows from the latter condition applied to the following state:

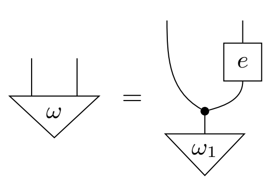


∎

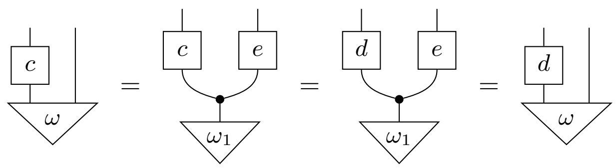
In fact, any category admitting disintegration has the equality extension property.

Proposition 5.9. If an affine CD-category admits disintegration, then it has the equality extension property.

Proof. Assume $c \overset{\omega_1}{\sim} d$ for a state $\omega\colon I \to X \otimes Z$ and for channels $c, d\colon X \to Y$, with $\omega_1 = \pi_1 \circ \omega$. Suppose that ω is disintegrated as:



Then



This concludes the proof by Lemma 5.8. □

Recall that $\mathcal{K}\ell(\mathcal{G})$ does not admit disintegration. Nevertheless, the category has the equality extension property.

Proposition 5.10. The category $\mathcal{K}\ell(\mathcal{G})$ satisfies the equality extension property.

Proof. We prove the claim using Proposition 5.3 and Lemma 5.8. Assume $c \overset{\omega_1}{\sim} d$ for $\omega \in \mathcal{G}(X \otimes Z)$ and $c, d\colon X \to \mathcal{G}(Y)$, with ω_1 the first marginal of ω. The equality $(c \otimes \eta_Z) \circledcirc \omega = (d \otimes \eta_Z) \circledcirc \omega$ is equivalent to:

$$\int_{X\times Z} c(x)(A)\mathbf{1}_C(z)\,\omega(\mathrm{d}(x,z)) = \int_{X\times Z} d(x)(A)\mathbf{1}_C(z)\,\omega(\mathrm{d}(x,z)) \tag{6}$$

for all $A \in \Sigma_X$ and $C \in \Sigma_Z$. Using

$$\left|c(x)(A)\mathbf{1}_C(z) - d(x)(A)\mathbf{1}_C(z)\right| = \left|c(x)(A) - d(x)(A)\right|\mathbf{1}_C(z) \le \left|c(x)(A) - d(x)(A)\right| \ ,$$

we have

$$\begin{aligned}
&\Big|\int_{X\times Z}\big(c(x)(A)\mathbf{1}_C(z) - d(x)(A)\mathbf{1}_C(z)\big)\,\omega(\mathrm{d}(x,z))\Big| \\
&\le \int_{X\times Z}\left|c(x)(A)\mathbf{1}_C(z) - d(x)(A)\mathbf{1}_C(z)\right|\omega(\mathrm{d}(x,z)) \\
&\le \int_{X\times Z}\left|c(x)(A) - d(x)(A)\right|\omega(\mathrm{d}(x,z)) \\
&= \int_{X}\left|c(x)(A) - d(x)(A)\right|(\pi_1)_*(\omega)(\mathrm{d}x) \\
&= \int_{X}\left|c(x)(A) - d(x)(A)\right|\omega_1(\mathrm{d}x) \\
&= 0 \ .
\end{aligned}$$

This proves the desired equality (6). □

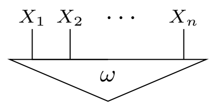
6. Conditional independence

Throughout this section, we consider an affine CD-category that admits disintegration.

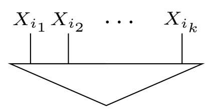
6.1. *Disintegration of multipartite states*

So far we have concentrated on bipartite states — except in the classification example in Section 4. In order to deal with a general n-partite state $\omega\colon I \to X_1 \otimes \cdots \otimes X_n$, we will introduce several notations and conventions in Definitions 6.1, 6.2 and 6.3 below; they are in line with standard practice in probability theory.

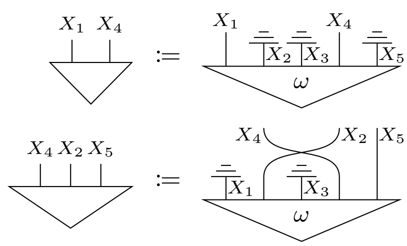
In the conventions, an n-partite state, as below, is fixed, and used implicitly.



Definition 6.1. When we write



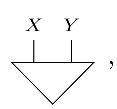
where $i_1, \ldots, i_k$ are distinct, it denotes the state $I \to X_{i_1} \otimes \cdots \otimes X_{i_k}$ obtained from ω by marginalisation and permutation of wires (if necessary). Let us give a couple of examples, for $n = 5$.



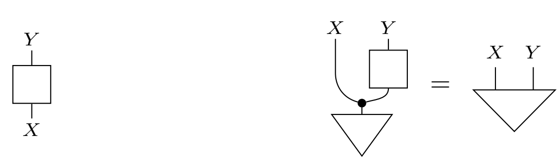
We permute wires via a combination of crossing. This is unambiguous by the coherence theorem.

Below we will use symbols $X, Y, Z, W, \ldots$ to denote not only a single wire X_i but also multiple wires $X_i \otimes X_j \otimes \cdots$. Disintegrations more general than in the bipartite case are now introduced as follows.

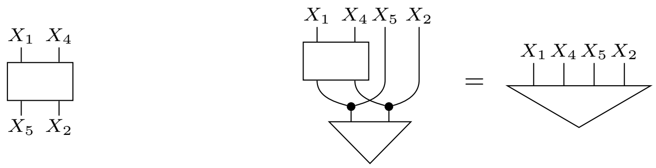
Definition 6.2. For $X = X_{i_1} \otimes \cdots \otimes X_{i_k}$ and $Y = X_{j_1} \otimes \cdots \otimes X_{j_l}$, with all $i_1, \ldots, i_k, j_1, \ldots, j_l$ distinct, a disintegration $X \to Y$ is defined to be a disintegration of



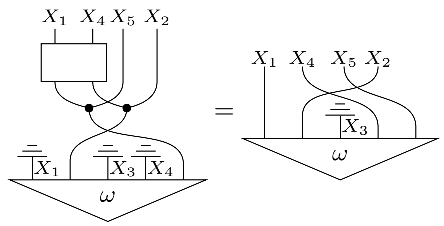
the marginal state given by the previous convention. We denote the disintegration simply as on the left below,

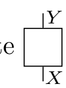


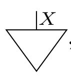
By definition, it must satisfy the equation on the right above. Let us give an example. The disintegration $X_1 \otimes X_4 \to X_5 \otimes X_2$ on the left below is defined by the equation on the right.



More specifically, assuming $n = 5$ and expanding the notation for marginals, the equation is:

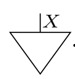


Note that disintegrations need not be unique. Thus when we write 

, we in fact *choose* one of them. Nevertheless, such disintegrations are unique up to almost-equality with respect to 

, which is good enough for our purpose.

Finally we make a convention about almost equality (Definition 5.1).

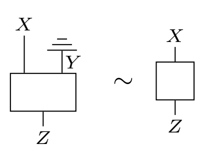
Definition 6.3. Let S and T be string diagrams of type $X \to Y$ that are made from marginals and disintegrations of ω as defined in Definitions 6.1 and 6.2. When we say S is almost equal to T (or write $S \sim T$) without reference to a state, it means that S is almost equal to T with respect to the state 

.

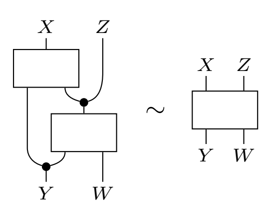
We shall make use of the following auxiliary equations involving discarding and composition of disintegrations.

Proposition 6.4. In the conventions and notations above, the following hold.

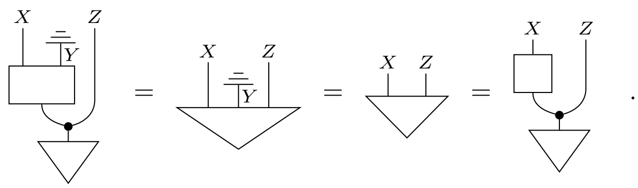
1



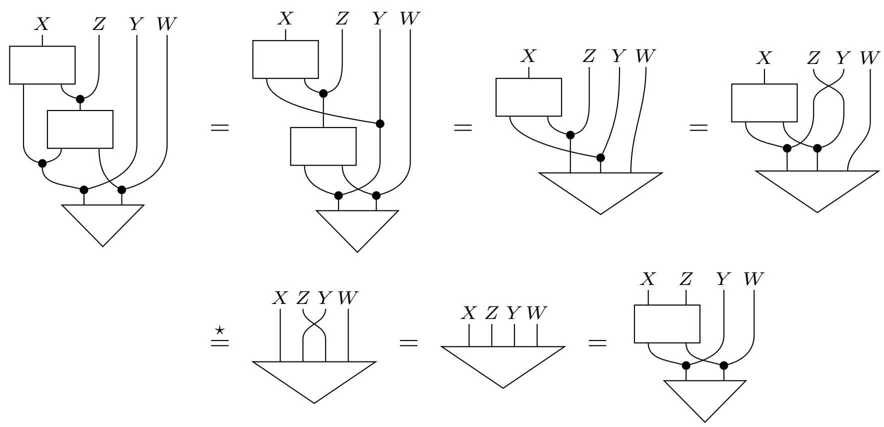
2



Proof. By the definition of almost equality, 1 is proved by:



Similarly, we prove 2 as follows.



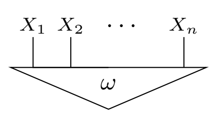
For the marked equality $\overset{\star}{=}$, we used the equality extension property, which is valid in a category with disintegration. □

The equations correspond respectively to $\sum_y \mathrm{P}(x, y|z) = \mathrm{P}(x|z)$ and $\mathrm{P}(x|y, z) \cdot \mathrm{P}(z|y, w) = \mathrm{P}(x, z|y, w)$ in discrete probability.

Remark 6.5. In this section, we use symbols $X_1, X_2, \ldots$ or $X, Y, Z, \ldots$ in order to specify wires in string diagrams. These symbols should not mean mere objects/types, since objects need not be distinct and thus we cannot distinguish wires by objects. As a consequence, our notations here are somewhat informal. One way to make our notations more formal is to introduce labels for wires (cf. (Kissinger, 2014)). For the present paper, however, our informal notations seem to be sufficient.

6.2. *Conditional independence*

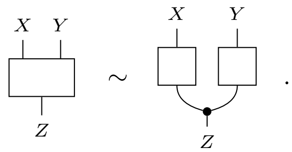
We continue using the notations in the previous subsection. Recall that we fix an n-partite state



and use symbols $X, Y, Z, W, \ldots$ to denote a wire X_i or multiple wires $X_i \otimes X_j \otimes \cdots$.

We now introduce the notion of conditional independence. Although it is defined with respect to the underlying state ω, we leave the state ω implicit, like an underlying probability space Ω in ordinary probability theory.

Definition 6.6. Let X, Y, Z denote distinct wires. Then we say X and Y are conditionally independent given Z, written as $X \perp\!\!\!\perp Y \mid Z$, if



The definition is analogous to the condition $\mathrm{P}(x, y|z) = \mathrm{P}(x|z)\,\mathrm{P}(y|z)$ in ordinary probability theory. Indeed our definition coincides with this ordinary one, as explained below.

Example 6.7. In $\mathcal{K}\ell(\mathcal{D})$, let $c_{X|Z} \colon Z \to \mathcal{D}(X)$, $c_{Y|Z} \colon Z \to \mathcal{D}(Y)$, $c_{XY|Z} \colon Z \to \mathcal{D}(X \times Y)$ be disintegrations of some joint state, say $\omega \in \mathcal{D}(X \times Y \times Z)$. Let $\omega_Z \in \mathcal{D}(Z)$ be the marginal on Z. Then $X \perp\!\!\!\perp Y \mid Z$ if and only if

$$c_{XY|Z}(z)(x, y) = c_{X|Z}(z)(x) \cdot c_{Y|Z}(z)(y) \quad \text{whenever} \quad \omega_Z(z) \neq 0$$

for all $x \in X$, $y \in Y$ and $z \in Z$. If we write $\mathrm{P}(x, y|z) = c_{XY|Z}(z)(x, y)$, $\mathrm{P}(x|z) = c_{X|Z}(z)(x)$, $\mathrm{P}(y|z) = c_{Y|Z}(z)(y)$, and $\mathrm{P}(z) = \omega_Z(z)$, then the condition will look more familiar:

$$\mathrm{P}(x, y|z) = \mathrm{P}(x|z) \cdot \mathrm{P}(y|z) \quad \text{whenever} \quad \mathrm{P}(z) \neq 0 \ .$$

Similarly, in $\mathcal{K}\ell(\mathcal{G})$, let $c_{X|Z} \colon Z \to \mathcal{G}(X)$, $c_{Y|Z} \colon Z \to \mathcal{G}(Y)$, $c_{XY|Z} \colon Z \to \mathcal{G}(X \times Y)$, and $\omega_Z \in \mathcal{G}(Z)$ be appropriate disintegrations and a marginal of some joint probability measure ω. Then $X \perp\!\!\!\perp Y \mid Z$ if and only if

$$c_{XY|Z}(z)(A \times B) = c_{X|Z}(z)(A) \cdot c_{Y|Z}(z)(B) \qquad \text{for } \omega_Z\text{-almost all } z \in Z$$

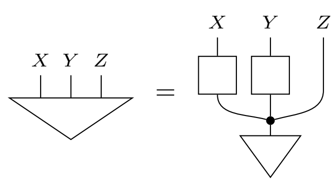
for all $A \in \Sigma_X$ and $B \in \Sigma_Y$.

The equivalences in the next result are well-known in conditional probability. Our contribution is that we formulate and prove them at an abstract, graphical level.

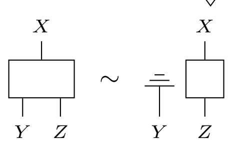
Proposition 6.8. The following are equivalent.

1 $X \perp\!\!\!\perp Y \mid Z$

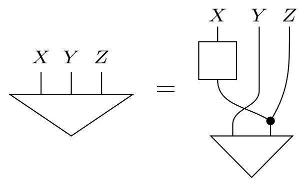
2



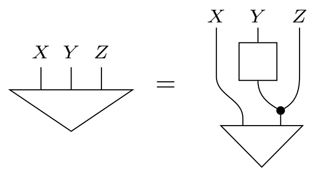
3



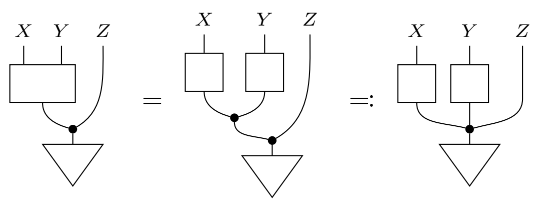
4



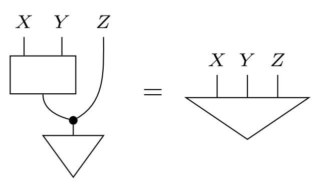
5



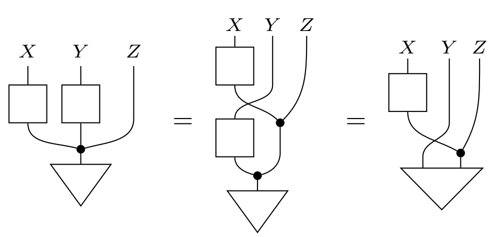
Proof. By definition of almost equality, 1 is equivalent to



We then have 1 ⇔ 2, since the identity below holds by the definition of disintegration.



Similarly, 3 ⇔ 4 follows by the definitions of almost equality and disintegration. We have 2 ⇔ 4 because



,

and similarly 2 ⇔ 5. □

Note that the condition 3 of the proposition is an analogue of $\mathrm{P}(x|y,z) = \mathrm{P}(x|z)$. The other conditions 2, 4 and 5 say that the joint state can be factorised in certain ways, corresponding to the following equations:

$$\mathrm{P}(x,y,z) \;=\; \mathrm{P}(x|z)\,\mathrm{P}(y|z)\,\mathrm{P}(z) \;=\; \mathrm{P}(x|z)\,\mathrm{P}(y,z) \;=\; \mathrm{P}(y|z)\,\mathrm{P}(x,z).$$

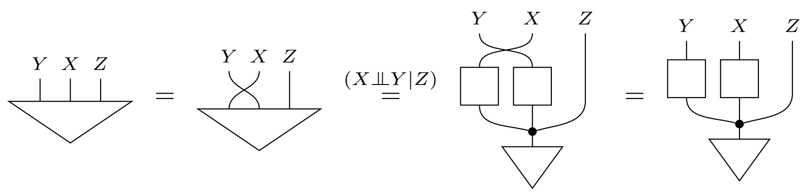
The proposition below shows that our abstract formulation of conditional independence does satisfy the basic 'rules' of conditional independence, which are known as *(semi-) graphoids axioms* (Verma and Pearl, 1988; Geiger et al., 1990).

Proposition 6.9. Conditional independence $(-) \perp\!\!\!\perp (-) \mid (-)$ satisfies:

1 (Symmetry) $X \perp\!\!\!\perp Y \mid Z$ if and only if $Y \perp\!\!\!\perp X \mid Z$.
2 (Decomposition) $X \perp\!\!\!\perp Y \otimes Z \mid W$ implies $X \perp\!\!\!\perp Y \mid W$ and $X \perp\!\!\!\perp Z \mid W$.
3 (Weak union) $X \perp\!\!\!\perp Y \otimes Z \mid W$ implies $X \perp\!\!\!\perp Y \mid Z \otimes W$.
4 (Contraction) $X \perp\!\!\!\perp Z \mid W$ and $X \perp\!\!\!\perp Y \mid Z \otimes W$ imply $X \perp\!\!\!\perp Y \otimes Z \mid W$.

Proof. We will freely use Proposition 6.8.

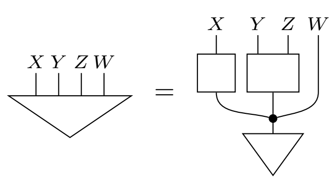
(1) Suppose $X \perp\!\!\!\perp Y \mid Z$. Then



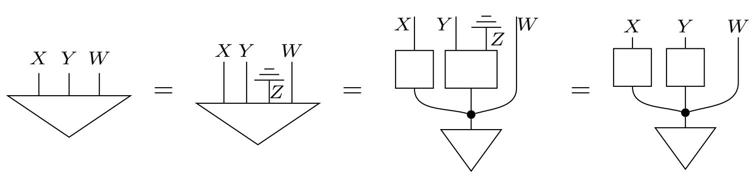
.

This means $Y \perp\!\!\!\perp X \mid Z$.

(2) Suppose $X \perp\!\!\!\perp Y \otimes Z \mid W$, namely:



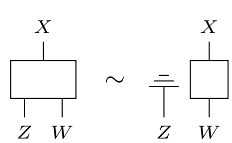
Marginalising Z, we obtain



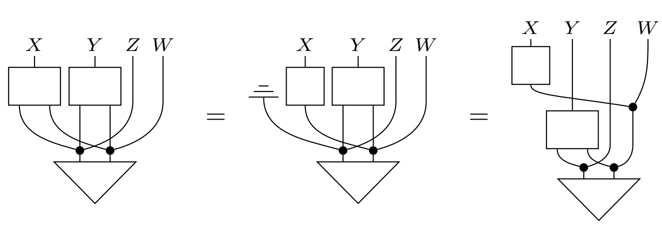
,

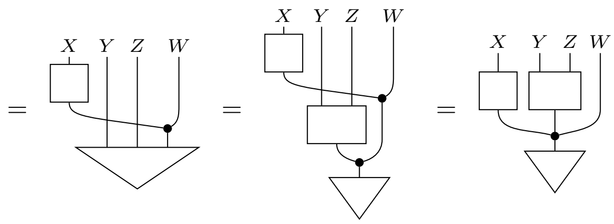
by Proposition 6.4.1. Thus $X \perp\!\!\!\perp Y \mid W$. Similarly we prove $X \perp\!\!\!\perp Z \mid W$.

Finally, we prove 3 and 4 at the same time. Note that $X \perp\!\!\!\perp Y \otimes Z \mid W$ implies $X \perp\!\!\!\perp Z \mid W$, as shown above. Therefore what we need to prove is that $X \perp\!\!\!\perp Y \otimes Z \mid W$ if and only if $X \perp\!\!\!\perp Y \mid Z \otimes W$, under $X \perp\!\!\!\perp Z \mid W$. Assume $X \perp\!\!\!\perp Z \mid W$, so we have



Then





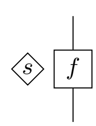
This proves $X \perp\!\!\!\perp Y \otimes Z \mid W$ if and only if $X \perp\!\!\!\perp Y \mid Z \otimes W$. $\square$

The four properties from the graphoid axioms are essential in reasoning of conditional independence with DAGs or Bayesian networks (Verma and Pearl, 1988; Geiger et al., 1990). We leave further details to future work.

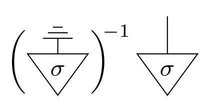
7. Beyond causal channels

All CD-categories $\mathbf{C}$ that we have considered so far are affine in the sense that all arrows $f\colon X \to Y$ are causal: $\bar{\top} \circ f = \bar{\top}$. We now drop the affineness, in order to enlarge our category to include 'non-causal' arrows. Essentially, we lose nothing by this change: all the arguments so far can still be applied to the subcategory $\mathrm{Caus}(\mathbf{C}) \subseteq \mathbf{C}$ containing all the objects and causal arrows. The category $\mathrm{Caus}(\mathbf{C})$ inherits the monoidal structure of $\mathbf{C}$ and the comonoid structures on each objects, so that $\mathrm{Caus}(\mathbf{C})$ is an affine CD-category.

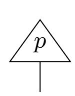
Recall that *channels* in $\mathbf{C}$ are causal arrows, *i.e.* arrows in $\mathrm{Caus}(\mathbf{C})$. *States* are channels of the form $\sigma\colon I \to X$. We call endomaps $I \to I$ on the tensor unit *scalars*. The set $\mathbf{C}(I, I)$ of scalars forms a monoid via the composition $s \cdot t = s \circ t$ and $1 = \mathrm{id}_I$. The monoid of scalars is always commutative — in fact, this is the case for any monoidal category, see *e.g.* (Abramsky and Coecke, 2009, §3.2). In string diagram scalars are written as $\diamond s$ or simply as s. We can multiply scalars s to any arrows $f\colon X \to Y$ by the parallel composition, or diagrammatically by juxtaposition:

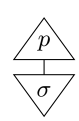


We call an arrow $\sigma\colon I \to X$ is *normalisable* if the scalar $\bar{\top} \circ \sigma\colon I \to I$ is (multiplicatively) invertible. In that case we can normalise σ into a proper state as follows.

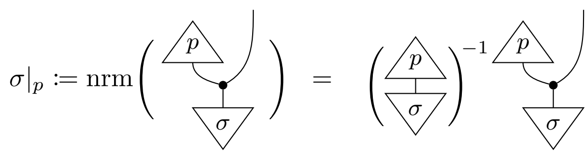
$\mathrm{nrm}(\sigma) :=$ 

Effects in $\mathbf{C}$ are arrows of the form $p\colon X \to I$; they correspond to observables, with predicates as special case. Diagrammatically they are written as on the left below.



$\omega \models p :=$ 

On the right the *validity* $\sigma \models p$ of a state $\sigma\colon I \to X$ and a effect $p\colon X \to I$ is defined. It is the scalar given by composition. Note that effects are not causal in general; by definition, only discarders $\bar{\top}$ are causal ones. States $\sigma\colon I \to X$ can be conditioned by effects $p\colon X \to I$ via normalisation, as follows.



The conditional state $\sigma|_p$ is defined if the validity $\sigma \models p$ is invertible.

Example 7.1. Recall that our previous examples $\mathcal{K}\ell(\mathcal{D})$ and $\mathcal{K}\ell(\mathcal{G})$ are both affine. We give two non-affine CD-categories that have $\mathcal{K}\ell(\mathcal{D})$ and $\mathcal{K}\ell(\mathcal{G})$ as subcategories, respectively.

1 For discrete probability, we use *multisets* (or *unnormalised distributions*) over nonnegative real numbers $\mathbb{R}_{\geq 0} = [0, \infty)$, such as

$$1|x\rangle + 0.5|y\rangle + 3|z\rangle \qquad \text{on a set } X = \{x, y, z, \dots\}$$

We denote by $\mathcal{M}(X)$ the set of multisets over $\mathbb{R}_{\geq 0}$ on X. More formally:

$$\mathcal{M}(X) = \{\phi\colon X \to \mathbb{R}_{\geq 0} \mid \phi \text{ has finite support}\} \ .$$

It extends to a commutative monad $\mathcal{M}\colon \mathbf{Set} \to \mathbf{Set}$, see (Coumans and Jacobs, 2013). In a similar way to the distribution monad $\mathcal{D}$, we can check that the Kleisli category $\mathcal{K}\ell(\mathcal{M})$ is a CD-category. For a Kleisli map $f\colon X \to \mathcal{M}(Y)$, causality $\bar{\top} \circ f = \bar{\top}$ amounts to the condition $\sum_y f(x)(y) = 1$ for all $x \in X$. It is thus easy to see that $\mathrm{Caus}(\mathcal{K}\ell(\mathcal{M})) \cong \mathcal{K}\ell(\mathcal{D})$. In fact, the distribution monad $\mathcal{D}$ can be obtained from $\mathcal{M}$ as its *affine submonad*, see (Jacobs, 2017).

An effect $p\colon X \to 1$ in $\mathcal{K}\ell(\mathcal{M})$ is a function $p\colon X \to \mathbb{R}_{\geq 0}$. Its validity $\sigma \models p$ in a state ω is given by the expected value $\sum_x \sigma(x) \cdot p(x)$. The state $\sigma|_p$ updated with 'evidence' p is defined as $\sigma|_p(x) = \frac{\sigma(x)\cdot p(x)}{\sigma \models p}$.

2 For general, measure-theoretic probability, we use *s-finite kernels* between measurable spaces (Kallenberg, 2017; Staton, 2017). Let X and Y be measurable spaces. A function $f\colon X \times \Sigma_Y \to [0, \infty]$ is called a *kernel* from X to Y if

— $f(x, -)\colon \Sigma_Y \to [0, \infty]$ is a measure for each X; and

— $f(-, B)\colon X \to [0, \infty]$ is measurable for each $B \in \Sigma_Y$.

We write $f\colon X \rightsquigarrow Y$ when f is a kernel from X to Y. A *probability kernel* is a kernel $f\colon X \rightsquigarrow Y$ with $f(x, Y) = 1$ for all $x \in X$. A kernel $f\colon X \rightsquigarrow Y$ is *finite* if there exists $r \in [0, \infty)$ such that for all $x \in X$, $f(x, Y) \leq r$. (Note that it must be 'uniformly' finite.) A kernel $f\colon X \rightsquigarrow Y$ is *s-finite* if $f = \sum_n f_n$ for some countable family $(f_n\colon X \to Y)_{n \in \mathbb{N}}$ of finite kernels.

For two s-finite kernels $f\colon X \rightsquigarrow Y$ and $g\colon Y \rightsquigarrow Z$, we define the (sequential) composite $g \circ f\colon X \rightsquigarrow Z$ by

$$(g \circ f)(x, C) = \int_Y g(y, C)\, f(x, \mathrm{d}y)$$

for $x \in X$ and $C \in \Sigma_Z$. There are identity kernels $\eta_X\colon X \rightsquigarrow X$ given by $\eta_X(x, A) = \mathbf{1}_A(x)$. With these data, measurable spaces and s-finite kernels form a category, which we denote by **sfKrn**. There is a monoidal structure on **sfKrn**. For measurable spaces X, Y we define the tensor product $X \otimes Y = X \times Y$ to be the cartesian product of measurable spaces. The tensor unit $I = 1$ is the singleton space. For s-finite kernels $f\colon X \rightsquigarrow Y$ and $g\colon Z \rightsquigarrow W$, we define $f \otimes g\colon X \times Z \rightsquigarrow Y \times W$ by

$$\begin{aligned}(f \otimes g)((x,z), E) &= \int_Y \Big(\int_W \mathbf{1}_E(y,w)\, g(z, \mathrm{d}w) \Big) f(x, \mathrm{d}y) \\ &= \int_W \Big(\int_Y \mathbf{1}_E(y,w)\, f(x, \mathrm{d}y) \Big) g(z, \mathrm{d}w)\end{aligned}$$

for $x \in X, z \in Z, E \in \Sigma_{Y \times W}$. The latter equality holds by the Fubini-Tonelli theorem for s-finite measures. These make the category **sfKrn** symmetric monoidal. Finally, for each measurable space X there is a 'copier' $\curlyvee\colon X \rightsquigarrow X \times X$ and a 'discarder' $\bar{\top}\colon X \rightsquigarrow 1$, given by $\curlyvee(x, E) = \mathbf{1}_E(x,x)$ and $\bar{\top}(x, 1) = 1$, so that **sfKrn** is a CD-category. For more technical details we refer to (Kallenberg, 2017; Staton, 2017).

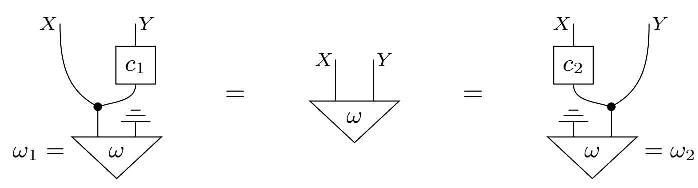
Note that an s-finite kernel $f\colon X \rightsquigarrow Y$ is causal if and only if it is a probability kernel, which is nothing but a Kleisli map $X \to \mathcal{G}(Y)$ for the Giry monad. Therefore the causal subcategory of **sfKrn** is the Kleisli category of the Giry monad: Caus(**sfKrn**) $\cong \mathcal{K}\ell(\mathcal{G})$. In particular, states in **sfKrn** are probability measures $\sigma \in \mathcal{G}(X)$.

An effect $p\colon X \rightsquigarrow 1$ in **sfKrn**, *i.e.* an s-finite kernel $p\colon X \times \Sigma_1 \to [0, \infty]$, can be identified with a measurable function $p\colon X \to [0, \infty]$. The validity $\sigma \models p$ is then the integral $\int_X p(x)\, \sigma(\mathrm{d}x)$, defined in $[0, \infty]$. The conditional state $\sigma|_p \in \mathcal{G}(X)$ is defined by:

$$\sigma|_p(A) = \frac{\int_A p(x)\, \sigma(\mathrm{d}x)}{\sigma \models p}$$

for $A \in \Sigma_X$, when the validity $\sigma \models p$ is neither 0 nor ∞.

With these notions in place we return to the original description of disintegration in Section 3. We assume a joint state ω with its two disintegrations c_1 and c_2 in:

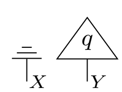


(7)

As indicated, we write ω_1 and ω_2 for the first and second marginals of ω.

Let q be an effect on Y. It can be extended to an effect $\mathbf{1} \otimes q$ on $X \otimes Y$, where:

$\mathbf{1} \otimes q \quad :=$



Then we can form the conditioned state $\omega|_{\mathbf{1} \otimes q}$. In a next step we take its first marginal, written as $\big(\omega|_{\mathbf{1} \otimes q}\big)_1$. It turns out that, in general, this first marginal is different from the original first marginal ω_1, even though the effect q only applies to the second coordinate. This is called 'crossover influence' in (Jacobs and Zanasi, 2017). It happens when the state ω is 'entwined', that is, when its two coordinates are correlated.

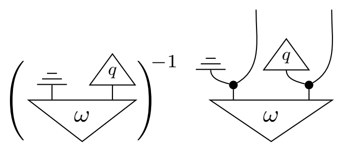
A fundamental result in this context is that this crossover influence can also be captured via the channels c_1, c_2 that are extracted from ω via disintegrations. This works via effect transformation $c^*(p) := p \circ c$ and state transformation $c_*(\sigma) := c \circ \sigma$ along a channel.

Theorem 7.2. In the above setting, assuming that the relevant conditioned states exist, there are equalities of states:

$$\omega_1|_{c_1^*(q)} \;=\; \big(\omega|_{\mathbf{1}\otimes q}\big)_1 \;=\; \big(c_2\big)_*(\omega_2|_q). \tag{8}$$

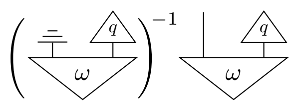
Following (Jacobs and Zanasi, 2016) we can say that the expression on the left in (8) uses *backward* inference, and the one on the right uses *forward* inference.

Proof. We first note that the state in the middle of (8) is the first marginal of:

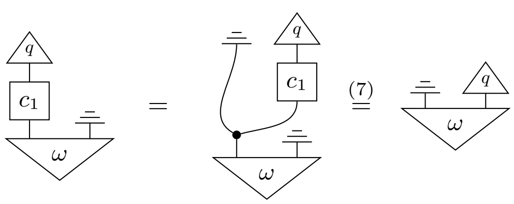


Hence:

$$\big(\omega|_{\mathbf{1}\otimes q}\big)_1 \quad = \tag{9}$$



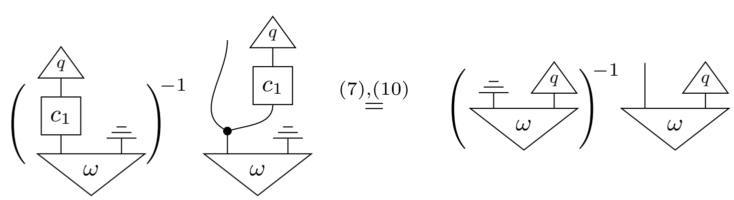
We note that the above scalar (that is inverted) can also be obtained as:



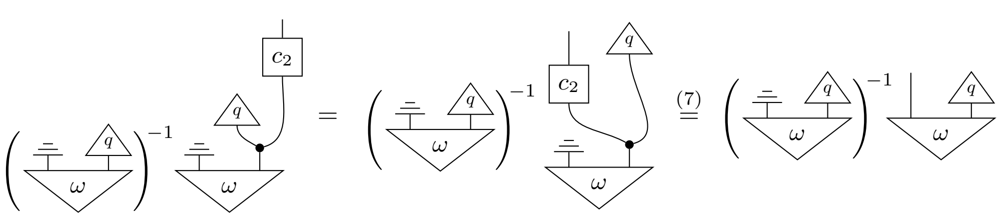
(10)

Hence we can prove the equation on the left in (8):

$$\omega_1|_{c_1^*(q)} \quad =$$



In a similar way we prove the equation on the right in (8), since $\big(c_2\big)_*(\omega_2|_q)$ equals:



□

The two equations in Theorem 8 will be illustrated in the 'disease and mood' example below, where a particular state (probability) will be calculated in three different ways.

Disease and mood example

We describe a non-trivial example of probabilistic (Bayesian) reasoning. The setting is the following. We consider a joint state about the occurrence and non-occurrence of a disease, written as D and $\sim D$, jointly with the occurrence and non-occurrence of a good mood, written as M and $\sim M$. The joint distribution that we start from is of the form:

$$0.05|M, D\rangle + 0.4|M, \sim D\rangle + 0.5|\sim M, D\rangle + 0.05|\sim M, \sim D\rangle. \tag{11}$$

Suppose there is a test for the disease, which is positive in 90% of all cases of people having the disease, and still 5% positive for people without the disease. Suppose the disease comes out positive. What is then the mood? It is expected that the mood will deteriorate, since the disease and the mood are 'entwined' (correlated) in the above joint state (11): a high likelihood of disease corresponds to a low mood.

We formalise this example in the *EfProb* language (Cho and Jacobs, 2017), see also Section 4. In *EfProb* one writes conditioning $s|_p$ as `s/p`, state transformation $c_*(s) = c \circ s$ as `c >> s` and predicate (effect) transformation $c^*(q) = q \circ c$ as `c << q`. We recall from Section 4 that marginalisation and disintegration of a state `s` are written as `s % M` and `s // M` respectively, where `M` is a mask of zeros and ones.

The above joint (prior) state (11) is defined in *EfProb* as follows, starting with the relevant domains (types).

```
>>> mood_dom = ['M', '~M']
>>> disease_dom = ['D', '~D']
>>> w = State([0.05, 0.4, 0.5, 0.05], [mood_dom, disease_dom])
>>> w
0.05|M,D> + 0.4|M,~D> + 0.5|~M,D> + 0.05|~M,~D>
```

We see that the latter state `w` in *EfProb* corresponds to the above state (11).

We can concentrate on the disease or mood separately, via marginalisations. In *EfProb* this is done as follows, via post-fix selection operations.

```
>>> w1 = w % [1,0]
>>> w2 = w % [0,1]
>>> w1
0.45|M> + 0.55|~M>
>>> w2
0.55|D> + 0.45|~D>
```

The sensitivity of the test is defined as a channel, called `sens` for 'sensitivity' below. If the disease is present, the test gives a positive outcome in 90% of the cases. But if the disease is absent, the test still has a 5% change of being positive. This is captured in the definition of the sensitivity channel below. It is applied to the first marginal of the prior state `w` to see what the a priori likelihood of a positive test is. This is done via state transformation, which is written in EfProb as >>.

```
>>> sens = chan_from_states([flip(9/10),flip(1/20)],disease_dom)
>>> sens >> w2
0.518|True> + 0.482|False>
```

As explained above, we are interested in the mood after a positive test. This requires updating the prior state. Below we first define the positive-test predicate, and then condition (update, revise) the state, via the *EfProb*-notation `/`. We introduce a positive-test predicate (effect) `pos_test` via predicate transformation, written as << in *EfProb*. In order to use it for conditioning the joint state `w`, we have to 'weaken' (extend) the predicate `pos_test` to the whole domain of `w`. This is done via parallel conjunction `@` with the truth predicate. Finally, the second marginal of the updated state `s` gives the new mood, after the test. This corresponds to the middle expression in (8).

```
>>> pos_test = sens << yes_pred
>>> pos_test
D: 0.9 | ~D: 0.05
>>> s = w / (pos_test @ truth(mood_dom))
>>> s % [0,1]
0.126|M> + 0.874|~M>
```

Clearly, a positive test leads to a lower mood: a reduction from 0.45 to 0.126.

Next we show how this result can also be obtained via disintegration of the prior joint state `w`. There are two ways to do this, via disintegration in the first component and predicate transformation (backward inference), or via disintegration in the second component and state transformation (forward inference). The first way corresponds to the expression on the left in (8):

```
>>> c1 = w // [1,0]
>>> w1 / (c1 << pos_test)
0.126|M> + 0.874|~M>
```

Via disintegration in the second component we get the same result, as on the right in (8):

```
>>> c2 = w // [0,1]
>>> c2 >> (w2 / pos_test)
0.126|M> + 0.874|~M>
```

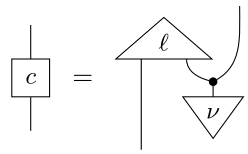
The fact that these three approaches lead to the same mood distribution follows from Theorem 7.2.

8. Disintegration via likelihoods

We continue in the setting of Section 7 in a CD-category that is not necessarily affine. The goal of this section is to present Theorem 8.3, which generalises a construction of Bayesian inversions using densities/likelihoods shown in Example 3.5.

We first introduce 'likelihoods' in our setting.

Definition 8.1. We say a channel $c\colon X \to Y$ is represented by a effect ℓ on $X \otimes Y$ with respect to an arrow $\nu\colon I \to Y$ if



(12)

We call ℓ a *likelihood relation* for the channel c with respect to ν.

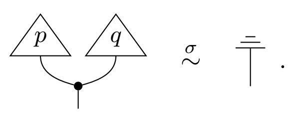
Interpreted in the category **sfKrn**, the definition says: a kernel $c\colon X \rightsquigarrow Y$ satisfies

$$c(x, B) = \int_B \ell(x, y)\, \nu(\mathrm{d}y)$$

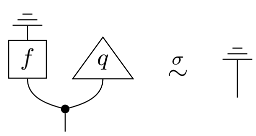
for a kernel $\ell\colon X \times Y \rightsquigarrow 1$ (identified with a measurable function $\ell\colon X \times Y \to [0, \infty]$) and a measure $\nu\colon 1 \rightsquigarrow Y$. This is basically the same as what we have in Example 3.5, but here ν is not necessarily the Lebesgue measure.

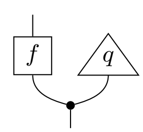
We use the σ-almost equality $\overset{\sigma}{\sim}$ also for non-causal arrows.

Definition 8.2. Let $\sigma\colon I \to X$ be a state. We say that an arrow $c\colon X \to Y$ is *σ-almost causal* if $\overset{=}{\top} \circ c \overset{\sigma}{\sim} \overset{=}{\top}$. An effect $p\colon X \to I$ is *σ-almost invertible* if there is an effect $q\colon X \to I$ such that:



The definition allows us to normalise an arrow $f\colon X \to Y$ into an almost causal one, as follows. If an effect $\overset{=}{\top} \circ f$ is σ-almost inverted by q, as on the left below,

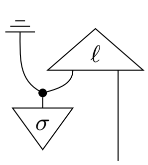




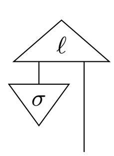
then clearly the arrow $X \to Y$ on the right is σ-almost causal.

We can now formulate and prove our main technical result.

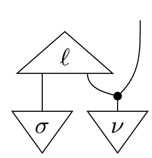
Theorem 8.3. Let σ be a state on X, and $c\colon X \to Y$ be a channel represented by a likelihood relation ℓ with respect to ν as in (12) above. Assume that the category has the equality extension property, and that the effect



$=$



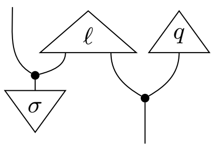
is almost invertible w.r.t. $c_*(\sigma) =$



.

Then, writing $q\colon Y \to I$ for an almost inverse to the effect, the channel

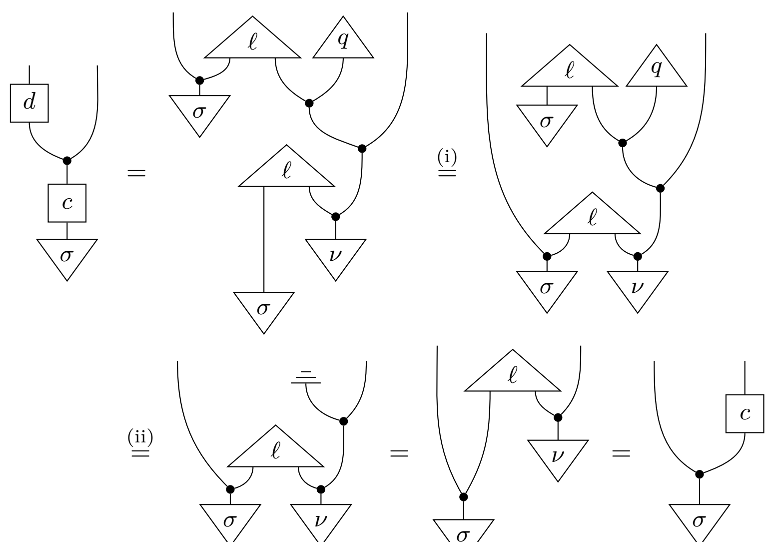
$$d\colon Y \to X \quad :=$$



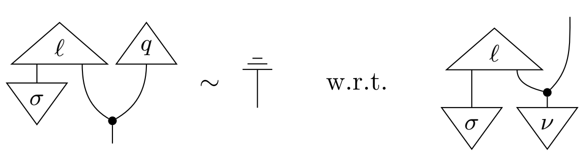
is a Bayesian inversion for σ along $c\colon X \to Y$. Namely, together they satisfy the equation (4).

Note that we now allow a non-causal arrow $d\colon Y \to X$ to be a Bayesian inversion. Nonetheless, it follows from the definition that any Bayesian inversion is $c_*(\sigma)$-almost causal. Similarly, we use the equality extension property for almost causal arrows.

Proof. We reason as follows.



For the equality $\overset{\text{(i)}}{=}$ we use associativity and commutativity of copiers $\curlyvee$. The equality $\overset{\text{(ii)}}{=}$ follows by



using the equality extension property. □

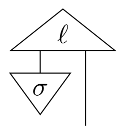
Example 8.4. We instantiate the Theorem 8.3 in **sfKrn**. Let $c\colon X \rightsquigarrow Y$ be a probability kernel represented by a likelihood relation $\ell\colon X \times Y \rightsquigarrow 1$ with respect to $\nu\colon 1 \rightsquigarrow Y$. The relation ℓ is identified with a measurable function $\ell\colon X \times Y \to [0, \infty]$ and ν with a measure $\nu\colon \Sigma_Y \to [0, \infty]$. The equation (12) amounts to

$$c(x, B) = \int_B \ell(x, y)\, \nu(\mathrm{d}y) \ .$$

In particular, each $\ell(x,-)$ is a probability density function, satisfying $\int_Y \ell(x,y)\,\nu(\mathrm{d}y) = 1$. Typically, we use the Lebesgue measure as ν, with Y a subspace of $\mathbb{R}$. Let $\sigma\colon 1 \rightsquigarrow X$ be a probability measure. Then $c_*(\sigma)\colon 1 \rightsquigarrow Y$ is given as:

$$c_*(\sigma)(B) = \int_X c(x,B)\,\sigma(\mathrm{d}x) = \int_X \int_B \ell(x,y)\,\nu(\mathrm{d}y)\,\sigma(\mathrm{d}x)$$

The effect

$p\colon Y \rightsquigarrow 1 \quad =$  is given as: $\quad p(y) = \int_X \ell(x,y)\,\sigma(\mathrm{d}x)$.

To define an inverse of p, we claim that $0 < p < \infty$, $c_*(\sigma)$-almost everywhere. We prove that $p^{-1}(\{0,\infty\}) = p^{-1}(0) \cup p^{-1}(\infty)$ is $c_*(\sigma)$-negligible, as:

$$\begin{aligned} c_*(\sigma)\big(p^{-1}(0)\big) &= \int_X \int_{p^{-1}(0)} \ell(x,y)\,\nu(\mathrm{d}y)\,\sigma(\mathrm{d}x) \\ &= \int_{p^{-1}(0)} p(x)\,\nu(\mathrm{d}y) \\ &= \int_{p^{-1}(0)} 0\,\nu(\mathrm{d}y) = 0 \end{aligned}$$

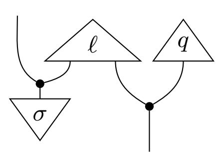
and, similarly we have

$$\int_{p^{-1}(\infty)} \infty\,\nu(\mathrm{d}y) = \int_{p^{-1}(\infty)} p(x)\,\nu(\mathrm{d}y) = c_*(\sigma)\big(p^{-1}(\infty)\big) \leq 1$$

but this is possible only when $\nu(p^{-1}(\infty)) = 0$, hence $c_*(\sigma)\big(p^{-1}(\infty)\big) = \int_{p^{-1}(\infty)} p(x)\,\nu(\mathrm{d}y) = 0$. Now define an effect $q\colon Y \rightsquigarrow 1$ by

$$q(y) = \begin{cases} p(y)^{-1} & \text{if } 0 < p(y) < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Then p is $c_*(\sigma)$-almost inverted by q. By Theorem 8.3, the Bayesian inversion for σ along c is given by

$d\colon Y \to X \quad \coloneqq$  ,

namely,

$$\begin{aligned} d(y,A) &= q(y) \int_A \ell(x,y)\,\sigma(\mathrm{d}x) \\ &= \frac{\int_A \ell(x,y)\,\sigma(\mathrm{d}x)}{\int_X \ell(x,y)\,\sigma(\mathrm{d}x)} \qquad \text{whenever} \quad 0 < \int_X \ell(x,y)\,\sigma(\mathrm{d}x) < \infty \end{aligned} \tag{13}$$

This may be seen as a variant of the Bayes formula. The calculation in Example 3.5 is reproduced when σ is also given via a density function.

We conclude with another example in which the likelihood-based calculation of Bayesian inversion, specifically the formula (13) above, is used to condition with respect to point observations.

Customers calling

Imagine a call centre that is open for 8 hours on each day of the week. The distribution of calls is different on weekends (Sat-Sun) from other days (Mon-Fri). What can we then learn from a single call at a given time of the day regarding whether it is weekend or not?

The formalisation in EfProb starts by defining a domain with label `W` for weekend and `~W` for non-weekend, together with a prior state that expresses a change of $\frac{2}{7}$ of being in a weekend.

```
>>> weekend_dom = ['W', '~W']
>>> prior = State([2/7, 5/7], weekend_dom)
>>> prior
0.286|W> + 0.714|~W>
```

Next we have a channel that assigns a different Gaussian distribution to `W` and to `~W`.

```
>>> hours_dom = R(0,8)
>>> c = chan_from_states([gaussian_state(5,4,hours_dom),
...                       gaussian_state(2,4,hours_dom)],
...                      weekend_dom)
```

The probability density functions of the two distributions look as follows.

In the weekend diagram on the left we see that the calls start coming in later. Now we ask ourselves the question: suppose we see one call at (hour) 6. How does this affect the prior distribution? Of course, the updated distribution should have a higher likelihood for 'weekend' since 6 is relatively late.

We construct an inversion `d` of the above channel `c`, together with the prior state, in order to compute the updated distribution. It yields the channel `d`, going from `hours_dom` to `weekend_dom`. The distribution at time 6 is obtained by applying channel `d` to the value 6.

```
>>> d = c.inversion(prior)
>>> d(6)
0.374|W> + 0.626|~W>
```

We see that the weekend probability has increased indeed.

References

Abramsky, S. and Coecke, B. (2009). Categorical quantum mechanics. In *Handbook of Quantum Logic and Quantum Structures: Quantum Logic*, pages 261–323. Elsevier.

Ackerman, N. L., Freer, C. E., and Roy, D. M. (2011). Noncomputable conditional distributions. In *Logic in Computer Science*, pages 107–116. IEEE.

Barber, D. (2012). *Bayesian Reasoning and Machine Learning.* Cambridge Univ. Press.

Bernardo, J. and Smith, A. (2000). *Bayesian Theory.* John Wiley & Sons.

Borgström, J., Gordon, A., Greenberg, M., Margetson, J., and Gael, J. V. (2013). Measure transformer semantics for Bayesian machine learning. *Logical Methods in Comp. Sci.*, 9(3):1–39.

Chang, J. T. and Pollard, D. (1997). Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317.

Cho, K. and Jacobs, B. (2017). The EfProb library for probabilistic calculations. In *Conference on Algebra and Coalgebra in Computer Science*, volume 72 of *LIPIcs*. Schloss Dagstuhl.

Cho, K., Jacobs, B., Westerbaan, A., and Westerbaan, B. (2015). An introduction to effectus theory. Preprint. arXiv:1512.05813 [cs.LO].

Clerc, F., Danos, V., Dahlqvist, F., and Garnier, I. (2017). Pointless learning. In *Foundations of Software Science and Computation Structures*, volume 10203 of *Lect. Notes Comp. Sci.*, pages 355–369. Springer.

Coecke, B. (2016). Terminality implies no-signalling ...and much more than that. *New Generation Computing*, 34(1):69–85.

Coecke, B. and Kissinger, A. (2017). *Picturing Quantum Processes: A First Course in Quantum Theory and Diagrammatic Reasoning.* Cambridge University Press.

Coumans, D. and Jacobs, B. (2013). Scalars, monads and categories. In Heunen, C., Sadrzadeh, M., and Grefenstette, E., editors, *Quantum Physics and Linguistics. A Compositional, Diagrammatic Discourse*, pages 184–216. Oxford Univ. Press.

Culbertson, J. and Sturtz, K. (2014). A categorical foundation for Bayesian probability. *Applied Categorical Structures*, 22(4):647–662.

D'Ariano, G. M., Chiribella, G., and Perinotti, P. (2017). *Quantum Theory from First Principles: An Informational Approach.* Cambridge University Press.

Faden, A. M. (1985). The existence of regular conditional probabilities: Necessary and sufficient conditions. *The Annals of Probability*, 13(1):288–298.

Fong, B. (2012). Causal theories: A categorical perspective on Bayesian networks. Master's thesis, Univ. of Oxford. arXiv:1301.6201 [math.PR].

Fremlin, D. H. (2000). *Measure Theory (5 volumes).* Torres Fremlin.

Geiger, D., Verma, T., and Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks*, 20:507–534.

Giry, M. (1982). A categorical approach to probability theory. In *Categorical Aspects of Topology and Analysis*, volume 915 of *Lecture Notes in Mathematics*, pages 68–85. Springer.

Gordon, A., Henzinger, T., Nori, A., and Rajamani, S. (2014). Probabilistic programming. In *Future of Software Engineering*, pages 167–181. ACM.

Jacobs, B. (2015). New directions in categorical logic, for classical, probabilistic and quantum logic. *Logical Methods in Comp. Sci.*, 11(3):1–76.

Jacobs, B. (2017). From probability monads to commutative effectuses. *Journ. of Logical and Algebraic Methods in Programming*, 156.

Jacobs, B., Westerbaan, B., and Westerbaan, A. (2015). States of convex sets. In *Foundations of Software Science and Computation Structures*, volume 9034 of *Lect. Notes Comp. Sci.*, pages 87–101. Springer.

Jacobs, B. and Zanasi, F. (2016). A predicate/state transformer semantics for Bayesian learning. In *Math. Found. of Programming Semantics*, volume 325 of *Elect. Notes in Theor. Comp. Sci.*, pages 185–200. Elsevier.

Jacobs, B. and Zanasi, F. (2017). A formal semantics of influence in Bayesian reasoning. In *Math. Found. of Computer Science*, volume 83 of *LIPIcs*. Schloss Dagstuhl.

Kallenberg, O. (2017). *Random Measures, Theory and Applications*. Springer.

Katoen, J.-P., Gretz, F., Jansen, N., Lucien Kaminski, B., and Olmedo, F. (2015). Understanding probabilistic programs. In *Correct System Design*, volume 9360 of *Lect. Notes Comp. Sci.*, pages 15–32. Springer.

Kissinger, A. (2014). Abstract tensor systems as monoidal categories. In *Categories and Types in Logic, Language, and Physics: Essays Dedicated to Jim Lambek on the Occasion of His 90th Birthday*, pages 235–252. Springer.

Mac Lane, S. (1998). *Categories for the Working Mathematician*. Springer, second edition.

Pachl, J. K. (1978). Disintegration and compact measures. *Mathematica Scandinavica*, 43:157–168.

Panangaden, P. (2009). *Labelled Markov Processes*. Imperial College Press.

Pawitan, Y. (2001). *In All Likelihood. Statistical Modelling and Inference Using Likelihood*. Clarendon Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

Selinger, P. (2010). A survey of graphical languages for monoidal categories. In *New Structures for Physics*, volume 813 of *Lecture Notes in Physics*, pages 289–355. Springer.

Shan, C.-c. and Ramsey, N. (2017). Exact Bayesian inference by symbolic disintegration. In *Princ. of Programming Languages*, pages 130–144. ACM.

Staton, S. (2017). Commutative semantics for probabilistic programming. In *European Symp. on Programming*, volume 10201 of *Lect. Notes Comp. Sci.*, pages 855–879. Springer.

Staton, S., Yang, H., Heunen, C., Kammar, O., and Wood, F. (2016). Semantics for probabilistic programming: higher-order functions, continuous distributions, and soft constraints. In *Logic in Computer Science*, pages 525–534. ACM.

Stoyanov, J. M. (2014). *Counterexamples in Probability*. Dover, third edition.

Verma, T. and Pearl, J. (1988). Causal networks: Semantics and expressiveness. In *Uncertainty in Artificial Intelligence*.

Witten, I., Frank, E., and Hall, M. (2011). *Data Mining Practical Machine Learning Tools and Techniques*. Elsevier, Amsterdam.