9

The Logical Essentials of Bayesian Reasoning

Bart Jacobs
Radboud Universiteit, Nijmegen
Fabio Zanasi
University College London

Abstract: This chapter offers an accessible introduction to the channel-based approach to Bayesian probability theory. This framework rests on algebraic and logical foundations, inspired by the methodologies of programming language semantics. It offers a uniform, structured and expressive language for describing Bayesian phenomena in terms of familiar programming concepts, like channel, predicate transformation and state transformation. The introduction also covers inference in Bayesian networks, which will be modelled by a suitable calculus of string diagrams.

9.1 Introduction

In traditional imperative programming one interprets a program as a function that changes states. Intuitively, the notion of 'state' captures the state of affairs in a computer, as given for instance by the contents of the relevant parts of the computer's memory. More abstractly, a program is interpreted as a *state transformer*. An alternative, logical perspective is to interpret a program as a *predicate transformer*. In that case the program turns one predicate into a new predicate. This works in opposite direction: the program turns a predicate on the 'post-state' into a predicate on the 'pre-state', for instance via the weakest precondition computation. As discovered in the early days of programming semantics, basic relations exists between state transformation and predicate transformation, see for instance Dijkstra and Scholten (1990), Dijkstra (1997), and Proposition 9.11 below.

A similar theory of state and predicate transformation has been developed for probabilistic programming, see Kozen (1981, 1985). This approach has been generalised and re-formulated in recent years in categorical terms, typically using so-called Kleisli categories, in Jacobs (2017), and more generally via the notion of

[a] From *Foundations of Probabilistic Programming*, edited by Gilles Barthe, Joost-Pieter Katoen and Alexandra Silva published 2020 by Cambridge University Press.

effectus, in Cho et al. (2015). Category theory provides a fundamental language for the semantics of programming languages. This is clear in approaches based on domain theory. For instance, many constructions for types in programming languages have categorical counterparts, like (co)products, exponentials, and initial algebras (and final coalgebras) – where these (co)algebras are used for fixed points. These categorical notions come with universal properties that guide the design (syntax) and rules of programming languages.

This use of category theory is well-established in functional programming languages. However, it is less established in probabilistic programming. The description of some of the basic notions of probability theory in categorical terms goes back to the early 1980s (see Giry, 1982) and has seen a steady stream of activities since – see *e.g.* Jones and Plotkin (1989), Jung and Tix (1998), de Vink and Rutten (1999), Bartels et al. (2004), Tix et al. (2005), Varacca and Winskel (2006), Keimel (2008), Keimel and Plotkin (2009), Panangaden (2009), Sokolova (2011), Mislove (2012), Fong (2012), Culbertson and Sturtz (2014), Ścibior et al. (2015), Staton et al. (2016), Ścibior et al. (2018). This categorical perspective is not a goal in itself, but it does offer a structural, implementation-independent way of thinking which is natural for systematic programmers.

This chapter offers an introduction to this principled perspective on probability theory, especially for Bayesian probabilistic programming, based on earlier work of the authors' in this direction, see *e.g.* Jacobs (2011, 2013, 2018b), Jacobs and Zanasi (2016, 2017,?). Even though it is categorically-driven, our exposition does not require any categorical prerequisite. The reader interested in an explicitly categorical description of the framework may consult Jacobs and Zanasi (2016) or Cho and Jacobs (2019).

The fundamental concept will be called *channel*: all the basics of Bayesian probability theory (event, belief revision, Bayesian network, disintegration, . . .) will be derived from this single primitive. In analogy with approaches in programming language semantics, channels are formally definable as arrows of a certain Kleisli category: depending on the category of choice, the derived notions instantiate to discrete or to continuous probability theory – and even to quantum probability too, although the quantum world is out of scope here (but see Jacobs and Zanasi, 2016 and Jacobs, 2018b). This setting does not only provide a completely *uniform* mathematical description for a variety of phenomena, but also introduces in Bayesian reasoning fundamental principles of programming theory, such as *compositionality*: channels are composable in a variety of ways, resulting in a structured and modular theory. Furthermore, we argue that the channel-based perspective improves traditional approaches. We will study scenarios in which the established language for describing probabilistic phenomena lacks in flexibility, expressiveness and rigour, while the

new foundations disclose the underlying logical structure of phenomena, leading to new insights.

- Section 9.2 gives an informal overview of the channel-based view on probability theory in terms of a number of perspectives. These perspectives are stated explicitly, in imperative form, imposing how things are – or should be – seen from a channel-based perspective, in contrast to traditional approaches. These perspectives will already informally use the notions of 'state', 'predicate' and 'channel'.
- Section 9.3 commences the formal presentation of the ingredients of the channel-based framework, illustrating the concepts of state and predicate, as special forms of channels.
- Section 9.4 is devoted to conditioning, a key concept of Bayesian probability.
- Sections 9.5 and 9.6 are devoted to channel-based Bayesian inference. First, Section 9.5 explains the use of channels as predicate and state transformers. Then, Section 9.6 illustrates this setup to model inference in a Bayesian network, for the standard 'student' example from Koller and Friedman (2009). This section concludes with a clash of interpretations in an example taken from Barber (2012).
- Section 9.7 introduces a graphical calculus for channels – formally justified by their definition as arrows of a monoidal category. The calculus encompasses and enhances the diagrammatic language of Bayesian networks. It offers an intuitive, yet mathematically rigorous, description of basic phenomena of probability, including conditional independency.
- Section 9.8 uses the tools developed in the previous sections to study the relationship between joint distributions and their representation as Bayesian networks. First, we use the graphical calculus to give a channel-based account of disintegration. Second, we prove the equivalence of inference as performed on joint distributions and as performed in Bayesian networks (Theorem 9.16). The channel perspective explains the different dynamics at work in the two forms of inference, justifying our terminology of *crossover inference* and *transformer inference* respectively.

9.2 Perspectives

This section gives a first, informal view of the channel-based approach, through a series of perspectives. Each of them contains a prescription on how Bayesian phenomena appear in the channel-based perspective, and motivates how the channel language improves more traditional descriptions. These perspectives will informally use the notions of 'state', 'predicate' and 'channel'. To begin, we briefly explain what they are. A more systematic description is given later on.

- What is usually called a discrete probability distribution, we call a *state*. This terminology emphasises the role that distributions play in our programming-oriented framework: they express knowledge of a certain configuration (a state of affairs) that may be transformed by program execution (channels).

  A state/distribution is represented by a convex combination of elements from a set. For instance, on a set $A = \{a, b, c\}$ one can have a state $\frac{1}{3}|a\rangle + \frac{1}{2}|b\rangle + \frac{1}{6}|c\rangle$. The 'ket' notation $|\cdot\rangle$ is syntactic sugar: it has no mathematical meaning, but echoes how states are represented in quantum theory, where our theory may be also instantiated, see Jacobs and Zanasi (2016).
- A 'predicate' on a set A is a function $p\colon A \to 0, 1$. It assigns a probability $pa \in 0, 1$ to each element $a \in A$. Such predicates are often called 'fuzzy'. When $pa \in \{0, 1\}$, so that either $pa = 0$ or $pa = 1$, for each $a \in A$ the predicate is called sharp. A sharp predicate is traditionally called an event, and corresponds to a subset of A. For such a subset $E \subseteq A$ we write $\mathbf{1}_E\colon A \to 0, 1$ for the corresponding characteristic function (or sharp predicate), given by $\mathbf{1}_E a = 1$ if $a \in E$ and $\mathbf{1}_E a = 0$ if $a \notin E$. For the special case of a singleton set/event we write $\mathbf{1}_a$ instead of $\mathbf{1}_{\{a\}}$.

  Similarly to the case of states, our terminology draws an analogy with programming language semantics. There is a duality between states and predicates, which goes beyond the scope of this introduction – so the interested reader is referred to Jacobs (2017).
- A 'channel' $A \mathrel{-\!\!\circ\!\!\to} B$ from a set A to another set B is an A-indexed collection $(\omega_a)_{a\in A}$ of states ω_a on the set B. Alternatively, it is a function $a \mapsto \omega_a$ that sends each element $a \in A$ to a distribution on B. For A and B finite, yet another equivalent description is as a stochastic matrix with $|A|$ columns and $|B|$ rows.

  Channels are the pivot of our theory: states, predicates, and – as we shall see in Section 9.6 – also Bayesian networks can be seen as particular cases of a channel. More specifically, a state ω on B can be seen as a channel $f\colon \{\star\} \mathrel{-\!\!\circ\!\!\to} B$ with source the one-element set $\{\star\}$, defined by $f\star = \omega$. A predicate $p\colon A \to 0, 1$ can be seen as a channel $A \mathrel{-\!\!\circ\!\!\to} \{0, 1\}$ that assigns to $a \in A$ the state $pa|1\rangle + 1 - pa|0\rangle$.

9.2.1 The state is made explicit

Our first perspective elaborates on the observation that, in traditional probability, it is custom to leave the probability distribution implicit, for instance in describing the probability $\Pr E$ of an event $E = \{a, c\} \subseteq A$. This is justified because this distribution, say $\omega = \frac{1}{3}|a\rangle + \frac{1}{2}|b\rangle + \frac{1}{6}|c\rangle$, is typically fixed, so that carrying it around explicitly, as in $\Pr_\omega E$, burdens the notation. In contrast, in probabilistic programming, programs act on distributions (states) and change them with every

step. Hence in our framework it makes sense to use a richer notation, where states/distributions have a more prominent role.

First, pursuing a more abstract, logical viewpoint, we introduce notation $\models$ in place of Pr. For an arbitrary state ω on a set A and a predicate $p\colon A \to 0,1$ on the same set A, the *validity* $\omega \models p$ of p in ω is the number in $0,1$ given by:

$$\omega \models p \;\coloneqq \sum_{a \in A} \omega a \cdot pa. \tag{9.1}$$

When we identify an event (sharp predicate) $E \subseteq A$ with its characteristic function $\mathbf{1}_E\colon A \to 0,1$, we have $\omega \models \mathbf{1}_E = \mathrm{Pr}_\omega E = \frac{1}{2}$. The enhanced notation allows to distinguish this from the probability of E wrt. an alternative state $\psi = \frac{1}{4}|b\rangle + \frac{3}{4}|c\rangle$, written $\psi \models \mathbf{1}_E = \frac{3}{4}$.

Once we start treating states as explicit entities, we can give proper attention to basic operations on states, like parallel composition $\otimes$, marginalisation, and convex combination. These operations will be elaborated below in Section 9.3.

9.2.2 Conditional probability is state update with a predicate

Traditionally, conditional probability is described as $\mathrm{Pr}B \mid A$, capturing the probability of event B given event A. This notation is unfortunate, certainly in combination with the notation $\mathrm{Pr}B$ for the probability of event B. It suggests that conditioning $\mid$ is an operation on events, and that the probability $\mathrm{Pr}\cdot$ of the resulting event $B \mid A$ is computed. This perspective is sometimes called 'measure-free conditioning', see Dubois and Prade (1990). The fact that states are left implicit, see the previous point 9.2.1, further contributes to the confusion.

In the view advocated here, conditioning is an operation that updates a state ω in the light of evidence in the form of a predicate p. This is well-defined when ω and p have the same underlying set A, and when the validity $\omega \models p$ is non-zero. We shall then write $\omega|_p$ for the state "ω given p", see Section 9.4 for more details. We emphasise that the validity $\mathrm{Pr}B \mid A$ in state ω can now be expressed as $\omega|_{\mathbf{1}_A} \models \mathbf{1}_B$. It is the validity of B in the state where the evidence A is incorporated.

9.2.3 State/predicate transformation are basic operations

The following notation $\mathrm{Pr}X = a$ often occurs in traditional probability theory. What does it mean, and what is assumed? On close reading we find that the following data are involved.

- A set, often called sample space, Ω with a state/distribution ω on it; please note that ω is not an element of Ω but a probability distribution over elements of Ω;
- A stochastic (or random) variable, $X\colon \Omega \to A$, for some set A of outcomes;

- An element $a \in A$ with associated event $X^{-1}a = \{z \in \Omega \mid Xz = a\} \subseteq \Omega$;
- The probability $\Pr X = a$ is then the validity of the latter event in the state ω, that is, it is $\omega \models \mathbf{1}_{X^{-1}a}$.

A stochast is a special kind of channel (namely a deterministic one). The operation $X^{-1}a$ will be described more systematically as 'predicate transformation' $X \ll \mathbf{1}_a$ along the channel X. It turns the (singleton, sharp) predicate $\mathbf{1}_a$ on A into a predicate on Ω. In fact, $X \ll \mathbf{1}_a$ can be seen as just function composition $\Omega \to A \to 0, 1$. Since $X \ll \mathbf{1}_a$ is now a predicate on Ω, the probability $\Pr X = a$ can be described more explicitly as validity: $\omega \models X \ll \mathbf{1}_a$. More generally, for an event E on A we would then determine the probability $\Pr X \in E$ as $\omega \models X \ll E$.

One can use the channel X also for 'state transformation'. In this way one transforms the state ω on Ω into a state $X \gg \omega$ on A. This operation $\gg$ is sometimes (aptly) called pushforward, and $X \gg \omega$ is the pushforward distribution. The probability $\Pr X = a$ can equivalently be described as validity $X \gg \omega \models \mathbf{1}_a$.

In Section 9.5 we elaborate on channels. One of our findings will be that the probabilities $\omega \models c \ll p$ and $c \gg \omega \models p$ are always the same – for a channel c from A to B, a state ω on A, and a predicate p on B.

Moreover, we can profitably combine predicate transformation $\ll$ and state transformation $\gg$ with conditioning of states from point 9.2.2. As will be elaborated later on, we can distinguish the following two basic combinations of conditioning and transformation, with the associated terminology.

notation	action	terminology
$\omega\big\vert_{c \ll q}$	first do predicate transformation $\ll$, then update the state	evidential reasoning, or explanation, or backward inference
$c \gg (\omega\big\vert_p)$	first update the state, then do state transformation $\gg$,	causal reasoning, or prediction, or forward inference

9.2.4 Channels are used as probabilistic functions

We have already mentioned the notation $c \colon A \nrightarrow B$ to describe a channel c from A to B. Recall that such a channel produces a state ca on B for each element $a \in A$. It turns out that there is a special way to compose channels: for $c \colon A \nrightarrow B$ and $d \colon B \nrightarrow C$ we can form a composite channel $d \circ c \colon A \nrightarrow C$, understood as "$d$ after c". We can define it via state transformation as $d \circ ca = d \gg ca$. It is not hard to

check that $\circ$ is associative, and that there are identity maps $\mathrm{id} \colon A \nrightarrow A$, given by $\mathrm{id}a = 1|a\rangle$. They form unit elements for channel composition $\circ$.

Abstractly, channels form morphisms in a 'category'. The concept of a category generalises the idea of sets and functions between them, to objects and morphisms between them. These morphisms in a category need not be actual functions, but they must be composable (and have units). Such morphisms can be used to capture different forms of computation, like non-deterministic, or probabilistic (via channels). Here we shall not use categorical machinery, but use the relevant properties in more concrete form. For instance, composition $\circ$ of channels interacts appropriately with state transformation and with predicate transformation, as in:

$$d \circ c \gg \omega \;=\; d \gg c \gg \omega \qquad \text{and} \qquad d \circ c \ll p \;=\; c \ll d \ll p.$$

In addition to sequential composition $\circ$ we shall also use parallel composition $\otimes$ of channels, with an associated calculus for combinining $\circ$ and $\otimes$.

9.2.5 Predicates are generally fuzzy

In the points above we have used *fuzzy* predicates, with outcomes in the unit interval $0, 1$, instead of the more usual *sharp* predicates, with outcomes in the two-element set $\{0, 1\}$ of Booleans. Why?

- The main technical reason is that fuzzy predicates are closed under probabilistic predicate transformation $\ll$, whereas sharp predicates are not. Thus, if we wish to do evidential (backward) reasoning $\omega|_{c \ll q}$, as described in point 9.2.3, we are forced to use fuzzy predicates.
- Fuzzy predicates are also closed under another operation, namely *scaling*: for each $p \colon A \to 0, 1$ and $s \in 0, 1$ we have a new, scaled predicate $s \cdot p \colon A \to 0, 1$, given by $s \cdot pa = s \cdot pa$. This scaling is less important than predicate transformation, but still it is a useful operation.
- Fuzzy predicates naturally fit in a probabilistic setting, where uncertainty is a leading concept. It thus makes sense to use this uncertainty also for evidence.
- Fuzzy predicates are simply more general than sharp predicates. Sharp predicates p can be recognised logically among all fuzzy predicates via the property $p \mathbin{\&} p = p$, where conjunction $\&$ is pointwise multiplication.

The traditional approach in probability theory focuses on sharp predicates, in the form of events. This is part of the notation, for instance in expressions like $\Pr X \in E$, as used earlier in point 9.2.2. It does not make much sense to replace this sharp E with a fuzzy p when writing $\Pr X \in E$. That is one more reason why we write validity via $\models$ and not via $\Pr$. Fuzzy predicates have actually surfaced in more recent

research in Bayesian probability, see *e.g.* the concepts of 'soft' evidence in Valtorta et al. (2002) and 'uncertain' evidence in Mrad et al. (2015), see also Barber (2012).

Fuzzy predicates have a different algebraic structure than sharp predicates. The latter form Boolean algebras. Fuzzy predicates however form effect modules (see *e.g.* Jacobs, 2013). However, these algebraic/logical structures will not play a role in the current setting.

We shall later sketch how a fuzzy predicate can be replaced by an additional node in a Bayesian network, see Remark 9.4.

9.2.6 Marginalisation and weakening are operations

Marginalisation is the operation of turning a joint distribution ω on a product domain $X \times Y$ into a distribution on one of the components, say on X. Traditionally marginalisation is indicated by omitting one of the variables: if $\omega x, y$ is written for the joint distribution on $X \times Y$, then ωx is its (first) marginal, as a distribution on X. It is defined as $\omega x = \sum_y \omega x, y$.

We prefer to write marginalisation as an explicit operation, so that $\mathsf{M}_1\omega$ is the first marginal (on X), and $\mathsf{M}_2\omega$ is the second marginal (on Y). More generally, marginalisation can be performed on a state σ on a domain $X_1 \times \cdots \times X_n$ in 2^n many ways.

A seemingly different but closely related operation is weakening of predicates. If $p \in [0, 1]^X$ is a predicate on a domain X, we may want to use it on a larger domain $X \times Y$ where we ignore the Y-part. In logic this called weakening; it involves moving a predicate to a larger context. One could also indicate this via variables, writing px for the predicate on X, and px, y for its extension to $X \times Y$, where y in px, y is a spurious variable. Instead we write $\mathsf{W}_1 p \in [0, 1]^{X \times Y}$ for this weakened predicate. It maps x, y to px.

Marginalisation M and weakening W are each other's 'cousins'. As we shall see, they can both be expressed via projection maps $\pi_1 \colon X \times Y \to X$, namely as state transformation $\mathsf{M}_1\omega = \pi_1 \gg \omega$ and as predicate transformation $\mathsf{W}_1 p = \pi_1 \ll p$. As a result, the symbols M and W can be moved accross validity $\models$, as in (9.6) below. In what we call crossover inference later on, the combination of marginalisation and weakening plays a crucial role.

9.2.7 States and predicates are clearly distinguished

As just argued, marginalisation is an operation on states, whereas weakening acts on predicates (evidence). In general, certain operations only make sense on states (like convex sum) and others on predicates. This reflects the fact that states and predicates form very different algebraic structures: states on a given domain form a convex set

(see *e.g.* Jacobs, 2013), whereas, as already mentioned in Section 9.2.5, predicates on a given domain form an effect module.

Despite the important conceptual differences, states and predicates are easily confused, also in the literature (see *e.g.* Example 9.14 below). The general rule of thumb is that states involve finitely many probabilities that add up to one – unlike for predicates. We elaborate formally on this distinction in Remark 9.3 below.

On a more conceptual level, one could spell out the difference by saying that states have an ontological flavour, whereas predicates play an epistemological role. That means, states describe factual reality, although in probabilistic form, via convex combinations of combined facts. In contrast, predicates capture just the likelihoods of individual facts as perceived by an agent. Thus probabilities in predicates do not need to add up to one, because our perception of reality (contrary to reality itself) is possibly inconsistent or incomplete.[1] We shall elaborate more on this perspective at the end of Example 9.14 below.

9.3 States and predicates

Section 9.2.1 claimed that states (finite probability distributions) and fuzzy predicates – and their different roles – should be given more prominence in probability theory. We now elaborate this point in greater detail. We thus retell the same story as in the beginning, but this time with more mathematical details, and with more examples.

States

A *state* (probability distribution) over a 'sample' set A is a formal weighted combination $r_1|a_1\rangle + \cdots + r_n|a_n\rangle$, where the a_i are elements of A and the r_i are elements of $0, 1$ with $\sum_i r_i = 1$. We shall write $\mathcal{D}A$ for the set of states/distributions on a set A. We will sometimes treat $\omega \in \mathcal{D}A$ equivalently as a 'probability mass' function $\omega\colon A \to 0, 1$ with finite support $\mathrm{supp}\omega = \{a \in A \mid \omega a \neq 0\}$ and with $\sum_{a\in A} \omega a = 1$. More explicitly, the formal convex combination $\sum_i r_i|a_i\rangle$ corresponds to the function $\omega\colon A \to 0, 1$ with $\omega a_i = r_i$ and $\omega a = 0$ if $a \notin \{a_1, \ldots, a_n\}$. Then $\mathrm{supp}\omega = \{a_1, \ldots, a_n\}$, by construction.

For two states $\sigma_1 \in \mathcal{D}A_1$ and $\sigma_2 \in \mathcal{D}A_2$, we can form the joint 'product' state $\sigma_1 \otimes \sigma_2 \in \mathcal{D}A_1 \times A_2$ on the cartesian product $A_1 \times A_2$ of the underlying sets, namely as:

$$\sigma_1 \otimes \sigma_2 a_1, a_2 \;\coloneqq\; \sigma_1 a_1 \cdot \sigma_2 a_2. \tag{9.2}$$

[1] Within this perspective, it is intriguing to read conditioning of a state by a predicate as adapting the facts according to the agent's beliefs. In Philosophy one would say that our notion of conditioning forces an "idealistic" view of reality; in more mundane terms, it yields the possibility of "alternative facts".

For instance, if $\sigma_1 = \frac{1}{3}|a\rangle + \frac{2}{3}|b\rangle$ and $\sigma_2 = \frac{1}{8}|1\rangle + \frac{5}{8}|2\rangle + \frac{1}{4}|3\rangle$, then their product is written with ket-notation as:

$$\sigma_1 \otimes \sigma_2 \;=\; \tfrac{1}{24}|a,1\rangle + \tfrac{5}{24}|a,2\rangle + \tfrac{1}{12}|a,3\rangle + \tfrac{1}{12}|b,1\rangle + \tfrac{5}{12}|b,2\rangle + \tfrac{1}{6}|b,3\rangle.$$

Marginalisation works in the opposite direction: it moves a 'joint' state on a product set to one of the components: for a state $\omega \in \mathcal{D}A_1 \times A_2$ we have first and second marginalisation $\mathsf{M}_i\omega \in \mathcal{D}A_i$ determined as:

$$\mathsf{M}_1\omega a_1 \;=\; \sum_{a_2 \in A_2} \omega a_1, a_2 \qquad \mathsf{M}_2\omega a_2 \;=\; \sum_{a_a \in A_1} \omega a_1, a_2. \tag{9.3}$$

Here we use explicit operations M_1 and M_2 for taking the first and second marginal. The traditional way to write a marginal is to drop a variable: a joint distribution is written as $\Pr x, y$, and its marginals as $\Pr x$ and $\Pr y$, where $\Pr x = \sum_y \Pr x, y$ and $\Pr y = \sum_x \Pr x, y$.

The two original states σ_1 and σ_2 in a product state $\sigma_1 \otimes \sigma_2$ can be recovered as marginals of this product state: $\mathsf{M}_1\sigma_1 \otimes \sigma_2 = \sigma_1$ and $\mathsf{M}_2\sigma_1 \otimes \sigma_2 = \sigma_2$.

In general a joint state $\omega \in \mathcal{D}A_1 \times A_2$ does *not* equal the product $\mathsf{M}_1\omega \otimes \mathsf{M}_2\omega$ of its marginals, making the whole more than the sum of its parts. When we do have $\omega = \mathsf{M}_1\omega \otimes \mathsf{M}_2\omega$, we call ω *non-entwined*. Otherwise it is called *entwined*.

Example 9.1 Given sets $X = \{x, y\}$ and $A = \{a, b\}$, one can prove that a state $\omega = r_1|x,a\rangle + r_2|x,b\rangle + r_3|y,a\rangle + r_4|y,b\rangle \in \mathcal{D}X \times A$, where $r_1 + r_2 + r_3 + r_4 = 1$, is non-entwined if and only if $r_1 \cdot r_4 = r_2 \cdot r_3$. This fact also holds in the quantum case, see *e.g.* Mermin (2007, §1.5).

For instance, the following joint state is entwined:

$$\omega \;=\; \tfrac{1}{8}|x,a\rangle + \tfrac{1}{4}|x,b\rangle + \tfrac{3}{8}|y,a\rangle + \tfrac{1}{4}|y,b\rangle.$$

Indeed, ω has marginals $\mathsf{M}_1\omega \in \mathcal{D}X$ and $\mathsf{M}_2\omega \in \mathcal{D}A$, namely:

$$\mathsf{M}_1\omega \;=\; \tfrac{3}{8}|x\rangle + \tfrac{5}{8}|y\rangle \qquad \text{and} \qquad \mathsf{M}_2\omega \;=\; \tfrac{1}{2}|a\rangle + \tfrac{1}{2}|b\rangle.$$

The original state ω differs from the product of its marginals:

$$\mathsf{M}_1\omega \otimes \mathsf{M}_2\omega \;=\; \tfrac{3}{16}|x,a\rangle + \tfrac{3}{16}|x,b\rangle + \tfrac{5}{16}|y,a\rangle + \tfrac{5}{16}|y,b\rangle.$$

There is one more operation on states that occurs frequently, namely convex sum: if we have n states $\omega_i \in \mathcal{D}A$ on the same sets and n probabilities $r_i \in 0, 1$ with $\sum r_i = 1$, then $\sum r_i\omega_i$ is a state again.

Predicates

A *predicate* on a sample space (set) A is a function $p\colon A \to 0, 1$, taking values in the unit interval $0, 1$. We shall use the exponent notation $0, 1^A$ for the set of

predicates on A. What in probability theory are usually called events (subsets of A) can be identified with *sharp* predicates, taking values in the subset of booleans $\{0,1\} \subseteq 0,1$. We write $\mathbf{1}_E \in 0,1^A$ for the sharp predicate associated with the event $E \subseteq A$, defined by $\mathbf{1}_E a = 1$ if $a \in E$ and $\mathbf{1}_E a = 0$ if $a \notin E$, where we recall that we simply write $\mathbf{1}_a$ for $\mathbf{1}_{\{a\}}$. Thus predicates are a more general, 'fuzzy' notion of event, which we prefer to work with for the reasons explained in Section 9.2.5. We write $\mathbf{1} = \mathbf{1}_A$, $\mathbf{0} = \mathbf{1}_\emptyset$ for the truth and falsity predicates. They are the top and bottom elements in the set of predicates $0,1^A$, with pointwise order. As special case, for an element $a \in A$ we write $\mathbf{1}_a$ for the 'singleton' or 'point' predicate on A that is 1 only on input $a \in A$.

For predicates $p, q \in 0,1^A$ and scalar $r \in 0,1$ we define $p \mathbin{\&} q \in 0,1^A$ as $a \mapsto pa \cdot qa$ and $r \cdot p \in 0,1^A$ as $a \mapsto r \cdot pa$. Moreover, there is an orthosupplement predicate $p^\perp \in 0,1^A$ given by $p^\perp a = 1 - pa$. Then $p^{\perp\perp} = p$. Notice that $\mathbf{1}_E \mathbin{\&} \mathbf{1}_D = \mathbf{1}_{E \cap D}$ and ${\mathbf{1}_E}^\perp = \mathbf{1}_{\neg E}$, where $\neg E \subseteq A$ is the set-theoretic complement of E.

Definition 9.2 Let $\omega \in \mathcal{D}A$ be a state and $p \in 0,1^A$ be a predicate, both on the same set A. We write $\omega \models p$ for the *validity* or *expected value* of p in state ω. This validity is a number in the unit interval $0,1$. We recall its definition from (9.1):

$$\omega \models p \;\coloneqq\; \sum_{a \in A} \omega a \cdot pa. \tag{9.4}$$

For an event (sharp predicate) E, the probability $\Pr E$ wrt. a state ω is defined as $\sum_{a \in E} \omega a$. Using the above validity notation (9.4) we write $\omega \models \mathbf{1}_E$ instead. As special case we have $\omega \models \mathbf{1}_x = \omega x$.

Notice that the validity $\omega \models \mathbf{1}$ of the truth predicate $\mathbf{1}$ is 1 in any state ω. Similarly, $\omega \models \mathbf{0} = 0$. Additionally, $\omega \models p^\perp = 1 - \omega \models p$ and $\omega \models r \cdot p = r \cdot \omega \models p$.

There is also a parallel product $\otimes$ of predicates, like for states. Given two predicates $p_1 \in 0,1^{A_1}$ and $p_2 \in 0,1^{A_2}$ on sets A_1, A_2 we form the product predicate $p_1 \otimes p_2$ on $A_1 \times A_2$ via: $p_1 \otimes p_2 a_1, a_2 = p_1 a_1 \cdot p_2 a_2$. It is not hard to see that:

$$\omega_1 \otimes \omega_2 \models p_1 \otimes p_2 \;=\; (\omega_1 \models p_1) \cdot (\omega_2 \models p_2).$$

A product $p \otimes \mathbf{1}$ or $\mathbf{1} \otimes p$ with the truth predicate $\mathbf{1}$ corresponds to *weakening*, that is to moving a predicate p to a bigger set (or context). We also write:

$$\mathsf{W}_1 p \;\coloneqq\; p \otimes \mathbf{1} \qquad \text{and} \qquad \mathsf{W}_2 p \;\coloneqq\; \mathbf{1} \otimes p \tag{9.5}$$

for these first and second weakening operations, like in Section 9.2.6. We deliberately use 'dual' notation for marginalisation M and weakening W because these operations are closely related, as expressed by the following equations.

$$\mathsf{M}_1 \omega \models p \;=\; \omega \models \mathsf{W}_1 p \qquad \text{and} \qquad \mathsf{M}_2 \omega \models q \;=\; \omega \models \mathsf{W}_2 q. \tag{9.6}$$

As a result, $\sigma_1 \otimes \sigma_2 \models \mathsf{W}_1 p \;=\; \sigma_1 \models p$ and similarly $\sigma_1 \otimes \sigma_2 \models \mathsf{W}_2 q \;=\; \sigma_2 \models q$.

Remark 9.3 As already mentioned in Section 9.2.7, conceptually, it is important to keep states and predicates apart. They play different roles, but mathematically it is easy to confuse them. States describe a state of affairs, whereas predicates capture evidence. We explicitly emphasise the differences between a state $\omega \in \mathcal{D}A$ and a predicate $p\colon A \to 0,1$ in several points.

(i) A state has finite support. Considered as function $\omega\colon A \to 0,1$, there are only finitely many elements $a \in A$ with $\omega a \neq 0$. In contrast, there may be infinitely many elements $a \in A$ with $pa \neq 0$. This difference only makes sense when the underlying set A has infinitely many elements.

(ii) The finite sum $\sum_{a \in A} \omega a$ equals 1, since states involve a convex sum. In contrast there are no requirements about the sum of the probabilities $pa \in 0,1$ for a predicate p. In fact, such a sum may not exist when A is an infinite set. We thus see that each state ω on A forms a predicate, when considered as a function $A \to 0,1$. But a predicate in general does not form a state.

(iii) States and predicates are closed under completely different operations. As we have seen, for states we have parallel products $\otimes$, marginalisation M_i, and convex sum. In contrast, predicates are closed under orthosupplement $-^{\perp}$, conjunction $\&$, scalar multiplication $s \cdot -$ and parallel product $\otimes$ (with weakening as special case). The algebraic structures of states and of predicates is completely different: each set of states $\mathcal{D}A$ forms a convex set whereas each set of predicates $0,1^A$ is an effect module, see *e.g.* Jacobs (2018b) for more details.

(iv) State transformation (along a channel) happens in a forward direction, whereas predicate transformation (along a channel) works in a backward direction. These directions are described with respect the direction of the channel. This will be elaborated in Section 9.5.

Remark 9.4 One possible reason why fuzzy predicates are not so common in (Bayesian) probability theory is that they can be mimicked via an extra node in a Bayesian network, together with a sharp predicate. We sketch how this works. Assume we have a set $X = \{a, b, c\}$ and we wish to consider a fuzzy predicate $p\colon X \to 0,1$ on X, say with $pa = \frac{2}{3}$, $pb = \frac{1}{2}$ and $pc = \frac{1}{4}$. Then we can introduce an extra node $2 = \{t, f\}$ with a channel $h\colon X \nrightarrow 2$ given by:

$$ha = \tfrac{2}{3}|t\rangle + \tfrac{1}{3}|f\rangle \qquad hb = \tfrac{1}{2}|t\rangle + \tfrac{1}{2}|f\rangle \qquad hc = \tfrac{1}{4}|t\rangle + \tfrac{3}{4}|f\rangle.$$

The original predicate p on $X = \{a, b, c\}$ can now be reconstructed via predicate transformation along h as $h \ll \mathbf{1}_t$, where, recall, $\mathbf{1}_t$ is the sharp predicate on 2 which is 1 at t and 0 at f.

As an aside: we have spelled out the general isomorphism between predicates on a set A and channels $A \nrightarrow 2$. Conceptually this is pleasant, but in practice we do not wish to extend our Bayesian network every time a fuzzy predicate pops up.

What this example also illustrates is that sharpness of predicates is not closed under predicate transformation.

9.4 Conditioning

Conditioning is one of the most fundamental operations in probability theory. It is the operation that updates a state in the light of certain evidence. This evidence is thus incorporated in a new, updated state, that reflects the new insight. For this reason conditioning is sometime called belief update or belief revision. It forms the basis of learning, training and inference, see also Section 9.6.

A conditional probability is usually written as $\Pr E \mid D$. It describes the probability of event E, given event D. In the current context we follow a more general path, using fuzzy predicates instead of events. Also, we explicitly carry the state around. From this perspective, the update of a state ω with a predicate p, leading to an updated state $\omega|_p$, is the fundamental operation. It allows us to retrieve probabilities $\Pr E \mid D$ as special case, as will be shown at the end of this section.

Definition 9.5 Let $\omega \in \mathcal{D}A$ be a state and $p \in 0,1^A$ be a predicate, both on the same set A. If the validity $\omega \models p$ is non-zero, we write $\omega|_p$ for the conditional state "ω given p", defined as formal convex sum:

$$\omega|_p \;:=\; \sum_{a \in A} \frac{\omega a \cdot pa}{\omega \models p} |a\rangle. \tag{9.7}$$

Example 9.6 Let's take the numbers of a dice as sample space: $\mathsf{pips} = \{1,2,3,4,5,6\}$, with a fair/uniform dice distribution $\mathsf{dice} \in \mathcal{D}\mathsf{pips}$.

$$\mathsf{dice} \;=\; \tfrac{1}{6}|1\rangle + \tfrac{1}{6}|2\rangle + \tfrac{1}{6}|3\rangle + \tfrac{1}{6}|4\rangle + \tfrac{1}{6}|5\rangle + \tfrac{1}{6}|6\rangle.$$

We consider the predicate $\mathsf{evenish} \in 0,1^{\mathsf{pips}}$ expressing that we are fairly certain of pips being even:

$$\begin{array}{lll} \mathsf{evenish}1 \;=\; \frac{1}{5} & \mathsf{evenish}3 \;=\; \frac{1}{10} & \mathsf{evenish}5 \;=\; \frac{1}{10} \\ \mathsf{evenish}2 \;=\; \frac{9}{10} & \mathsf{evenish}4 \;=\; \frac{9}{10} & \mathsf{evenish}6 \;=\; \frac{4}{5} \end{array}$$

We first compute the validity of evenish for our fair dice:

$$\begin{aligned} \mathsf{dice} \models \mathsf{evenish} \;&=\; \sum_x \mathsf{dice}x \cdot \mathsf{evenish}x \\ &=\; \tfrac{1}{6}\cdot\tfrac{1}{5} + \tfrac{1}{6}\cdot\tfrac{9}{10} + \tfrac{1}{6}\cdot\tfrac{1}{10} + \tfrac{1}{6}\cdot\tfrac{9}{10} + \tfrac{1}{6}\cdot\tfrac{1}{10} + \tfrac{1}{6}\cdot\tfrac{4}{5} \\ &=\; \tfrac{2+9+1+9+1+8}{60} \;=\; \tfrac{1}{2}. \end{aligned}$$

If we take evenish as evidence, we can update our state and get:

$$\begin{aligned}
&\mathsf{dice}|_{\mathsf{evenish}} \\
&= \sum_x \frac{\mathsf{dice}x \cdot \mathsf{evenish}x}{\mathsf{dice} \models \mathsf{evenish}} |x\rangle \\
&= \tfrac{\frac{1}{6}\cdot\frac{1}{5}}{\frac{1}{2}}|1\rangle + \tfrac{\frac{1}{6}\cdot\frac{9}{10}}{\frac{1}{2}}|2\rangle + \tfrac{\frac{1}{6}\cdot\frac{1}{10}}{\frac{1}{2}}|3\rangle + \tfrac{\frac{1}{6}\cdot\frac{9}{10}}{\frac{1}{2}}|4\rangle + \tfrac{\frac{1}{6}\cdot\frac{1}{10}}{\frac{1}{2}}|5\rangle + \tfrac{\frac{1}{6}\cdot\frac{5}{6}}{\frac{1}{2}}|6\rangle \\
&= \tfrac{1}{15}|1\rangle + \tfrac{3}{10}|2\rangle + \tfrac{1}{30}|3\rangle + \tfrac{3}{10}|4\rangle + \tfrac{1}{30}|5\rangle + \tfrac{4}{15}|6\rangle.
\end{aligned}$$

As expected, the probability of the even pips is now higher than the odd ones. The evidence has been factored into the state.

We collect some basic properties of conditioning.

Lemma 9.7 *Let $\omega \in \mathcal{D}A$ and $p, q \in [0,1]^A$ be a state with predicates on the same set A.*

(i) *Conditioning with truth does nothing: $\omega|_{\mathbf{1}} = \omega$.*

(ii) *Conditioning with a conjunction amounts to separate conditionings, that is: $\omega|_{p\&q} = (\omega|_p)|_q$.*

(iii) *Conditioning with scalar product has no effect, when the scalar is non-zero: $\omega|_{r\cdot p} = \omega|_p$ when $r \neq 0$.*

(iv) *Conditioning with a point predicate yields a point state: $\omega|_{\mathbf{1}_x} = 1|x\rangle$, when $\omega x \neq 0$.*

Now let $\sigma_i \in \mathcal{D}A_i$ and $p_i \in [0,1]^{A_i}$.

(v) $\sigma_1 \otimes \sigma_2|_{p_1 \otimes p_2} = \sigma_1|_{p_1} \otimes \sigma_2|_{p_2}$.

(vi) $\mathsf{M}_1(\sigma \otimes \tau|_{\mathsf{M}_1 p_1}) = \sigma|_{p_1}$ *and* $\mathsf{M}_2(\sigma \otimes \tau|_{\mathsf{M}_2 p_2}) = \tau|_{p_2}$.

Proof All these properties follow via straightforward computation. We shall do (ii) and (v).

For (ii) we use:

$$\begin{aligned}
(\omega|_p|_q)a \;=\; \frac{\omega|_p a \cdot qa}{\omega|_p \models q} \;&=\; \frac{\frac{\omega a \cdot pa}{\omega \models p} \cdot qa}{\sum_b \omega|_p b \cdot qb} \\
&= \frac{\frac{\omega a \cdot pa}{\omega \models p} \cdot qa}{\sum_b \frac{\omega b \cdot pb}{\omega \models p} \cdot qb} \\
&= \frac{\omega a \cdot pa \cdot qa}{\sum_b \omega b \cdot pb \cdot qb} \\
&= \frac{\omega a \cdot p \,\&\, qa}{\omega \models p \,\&\, q} \\
&= (\omega|_{p\&q})a.
\end{aligned}$$

Similarly, for (v) we use:

$$\begin{aligned}\sigma_1 \otimes \sigma_2\big|_{p_1 \otimes p_2} a_1, a_2 &= \frac{\sigma_1 \otimes \sigma_2 a_1, a_2 \cdot p_1 \otimes p_2 a_1, a_2}{\sigma_1 \otimes \sigma_2 \models p_1 \otimes p_2} \\ &= \frac{\sigma_1 a_1 \cdot \sigma_2 a_2 \cdot p_1 a_1 \cdot p_2 a}{\sigma_1 \models p_1 \cdot \sigma_2 \models p_2} \\ &= \frac{\sigma_1 a_1 \cdot p_1 a_1}{\sigma_1 \models p_1} \cdot \frac{\sigma_2 a_2 \cdot p_2 a}{\sigma_2 \models p_2} \\ &= \sigma_1|_{p_1} \otimes \sigma_2|_{p_2}. \qquad \square\end{aligned}$$

The following result gives the generalisation of Bayes' rule to the current setting with states and predicates.

Theorem 9.8 *Let $\omega \in \mathcal{D}A$ and $p, q \in 0,1^A$ be a state and two predicates on the set A.*
(i) *The* product rule *holds:*

$$\omega|_p \models q \;=\; \frac{\omega \models p \,\&\, q}{\omega \models p} \tag{9.8}$$

(ii) Bayes' rule *holds:*

$$\omega|_p \models q \;=\; \frac{\omega|_q \models p \cdot \omega \models q}{\omega \models p} \tag{9.9}$$

Proof Point (ii) follows directly from (i) by using that $p \,\&\, q = q \,\&\, p$, so we concentrate on (i).

$$\begin{aligned}\omega|_p \models q \;=\; {}_a\omega|_p a \cdot qa \;&=\; {}_a \frac{\omega a \cdot pa}{\omega \models p} \cdot qa \\ &= {}_a \frac{\omega a \cdot p \,\&\, qa}{\omega \models p} \\ &= \frac{\omega \models p \,\&\, q}{\omega \models p}. \qquad \square\end{aligned}$$

We now relate our state-and-predicate based approach to conditioning to the traditional one. Recall that for events $E, D \subseteq A$ one has, by definition:

$$\Pr E \mid D \;=\; \frac{\Pr E \cap D}{\Pr D}.$$

If these probabilities $\Pr\cdot$ are computed wrt. a distribution $\omega \in \mathcal{D}A$, we can continue as follows.

$$\Pr E \mid D \;=\; \frac{\Pr E \cap D}{\Pr D} \;=\; \frac{\omega \models \mathbf{1}_{E \cap D}}{\omega \models \mathbf{1}_D} \;=\; \frac{\omega \models \mathbf{1}_E \,\&\, \mathbf{1}_D}{\omega \models \mathbf{1}_D} \overset{(9.8)}{=} \omega|_{\mathbf{1}_D} \models \mathbf{1}_E.$$

Thus the probability $\Pr E \mid D$ can be expressed in our framework as the validity

of the sharp predicate E in the state updated with the sharp predicate D. This is precisely the intended meaning.

9.5 Bayesian inference via state/predicate transformation

As mentioned in Section 9.2.4, a channel $c\colon A \nrightarrow B$ between two sets A, B is a probabilistic function from A to B. It maps an an element $a \in A$ to a state $ca \in \mathcal{D}B$ of B. Hence it is an actual function of the form $A \to \mathcal{D}B$. Such functions are often described as conditional probabilities $a \mapsto \Pr b \mid a$, or as stochastic matrices. We repeat that channels are fundamental – more so than states and predicates – since a state $\omega \in \mathcal{D}A$ can be identified with a channel $\omega\colon 1 \nrightarrow A$ for the singleton set $1 = \{0\}$. Similarly, a predicate $p \in [0,1]^A$ can be identified with a channel $p\colon A \nrightarrow 2$, where $2 = \{0,1\}$; this uses that $\mathcal{D}2 \cong [0,1]$.

Channels are used for probabilistic state transformation $\gg$ and predicate transformation $\ll$, in the following manner.

Definition 9.9 Let $c\colon A \nrightarrow B$ be a channel, with a state $\omega \in \mathcal{D}A$ on its domain A and a predicate $q \in [0,1]^B$ on its codomain B.
(i) State transformation yields a state $c \gg \omega$ on B defined by:

$$(c \gg \omega)b \;\coloneqq\; \sum_{a \in A} \omega a \cdot cab. \tag{9.10}$$

(ii) Predicate transformation gives a predicate $c \ll q$ on A defined by:

$$(c \ll q)a \;\coloneqq\; \sum_{b \in B} cab \cdot qb. \tag{9.11}$$

The next example illustrates how state and predicate transformation can be used systematically to reason about probabilistic questions.

Example 9.10 In a medical context we distinguish patients with low (L), medium (M), and high (H) blood pressure. We thus use as 'blood' sample space $B = \{L, M, H\}$, say with initial ('prior' or 'base rate') distribution $\beta \in \mathcal{D}B$:

$$\beta \;=\; \tfrac{1}{8}|L\rangle + \tfrac{1}{2}|M\rangle + \tfrac{3}{8}|H\rangle.$$

We consider a particular disease, whose a priori occurrence in the population depends on the blood pressure, as given by the following table.

blood pressure	disease likelihood
Low	5%
Medium	10%
High	15%

We choose as sample space for the disease $D = \{d, d^\perp\}$ where the element d represents presence of the disease and $d^\perp$ represents absence. The above table is now naturally described as a 'sickness' channel $s\colon B \nrightarrow D$, given by:

$$\begin{aligned} sL &= 0.05|\,d\rangle + 0.95|\,d^\perp\rangle \\ sM &= 0.1|\,d\rangle + 0.9|\,d^\perp\rangle \\ sH &= 0.15|\,d\rangle + 0.85|\,d^\perp\rangle. \end{aligned}$$

We ask ourselves two basic questions.

(i) **What is the a priori probability of the disease?** The answer to this question is obtained by state transformation, namely by transforming the blood pressure distribution β on B to a disease distribution $s \gg \beta$ on D along the sickness channel s. Concretely:

$$\begin{aligned} \big(s \gg \beta d &\overset{(9.10)}{=} \textstyle\sum_{x\in B} \beta x \cdot sxd \\ &= \beta L \cdot sLd + \beta M \cdot sMd + \beta H \cdot sHd \\ &= \tfrac{1}{8} \cdot \tfrac{1}{20} + \tfrac{1}{2} \cdot \tfrac{1}{10} + \tfrac{3}{8} \cdot \tfrac{3}{20} \\ &= \tfrac{9}{80} \\ \big(s \gg \beta d^\perp &\overset{(9.10)}{=} \textstyle\sum_{x\in B} \beta x \cdot sxd^\perp \\ &= \beta L \cdot sLd^\perp + \beta M \cdot sMd^\perp + \beta H \cdot sHd^\perp \\ &= \tfrac{1}{8} \cdot \tfrac{19}{20} + \tfrac{1}{2} \cdot \tfrac{9}{10} + \tfrac{3}{8} \cdot \tfrac{17}{20} \\ &= \tfrac{71}{80}. \end{aligned}$$

Thus we obtain as a priori disease distribution $c \gg \beta = \frac{9}{80}|\,d\rangle + \frac{71}{80}|\,d^\perp\rangle = 0.1125|\,d\rangle + 0.8875|\,d^\perp\rangle$. A bit more than 11% of the population has the disease at hand.

(ii) **What is the likely blood pressure for people without the disease?** Before we calculate the updated ('a posteriori') blood pressure distribution, we reason intuitively. Since non-occurrence of the disease is most likely for people with low blood pressure, we expect that the updated blood pressure – after taking the

evidence 'absence of disease' into account – will have a higher probability of low blood pressure than the orignal (a priori) value of $\frac{1}{8}$ in β.

The evidence that we have is the point predicate $\mathbf{1}_{d^\perp}$ on D, representing absence of the disease. In order to update $\beta \in \mathcal{D}B$ we first apply predicate transformation $s \ll \mathbf{1}_{d^\perp}$ to obtain a predicate on B. This transformed predicate in $0,1^B$ is computed as follows.

$$\begin{array}{rcl} \big(s \ll \mathbf{1}_{d^\perp} L & \overset{(9.11)}{=} & \sum_{y \in D} sLy \cdot \mathbf{1}_{d^\perp} y \;=\; sLd^\perp \;=\; 0.95 \\ \big(s \ll \mathbf{1}_{d^\perp} M & \overset{(9.11)}{=} & \sum_{y \in D} sMy \cdot \mathbf{1}_{d^\perp} y \;=\; sMd^\perp \;=\; 0.9 \\ \big(s \ll \mathbf{1}_{d^\perp} H & \overset{(9.11)}{=} & \sum_{y \in D} sHy \cdot \mathbf{1}_{d^\perp} y \;=\; sHd^\perp \;=\; 0.85. \end{array}$$

Notice that although $\mathbf{1}_{d^\perp}$ is a sharp predicate, the transformed predicate $s \ll \mathbf{1}_{d^\perp}$ is not sharp. This shows that sharp predicates are not closed under predicate transformation – as mentioned earlier in Section 9.2.5.

We can now update the original blood pressure distribution β with the transformed evicence $s \ll \mathbf{1}_{d^\perp}$. We first compute validity, and then perform conditioning:

$$\begin{array}{rcl} \beta \models s \ll \mathbf{1}_{d^\perp} & \overset{(9.4)}{=} & \sum_{x \in B} \beta x \cdot s \ll \mathbf{1}_{d^\perp} x \\ & = & \beta L \cdot s \ll \mathbf{1}_{d^\perp} L \\ & & \quad + \beta M \cdot s \ll \mathbf{1}_{d^\perp} M \\ & & \quad + \beta H \cdot s \ll \mathbf{1}_{d^\perp} H \\ & = & \frac{1}{8} \cdot \frac{19}{20} + \frac{1}{2} \cdot \frac{9}{10} + \frac{3}{8} \cdot \frac{17}{20} \\ & = & \frac{71}{80} \\ \beta|_{s \ll \mathbf{1}_{d^\perp}} & \overset{(9.7)}{=} & \sum_{x \in B} \dfrac{\beta x \cdot s \ll \mathbf{1}_{d^\perp} x}{\beta \models s \ll \mathbf{1}_{d^\perp}} |x\rangle \\ & = & \dfrac{\frac{1}{8} \cdot \frac{19}{20}}{\frac{71}{80}} |L\rangle + \dfrac{\frac{1}{2} \cdot \frac{9}{10}}{\frac{71}{80}} |M\rangle + \dfrac{\frac{3}{8} \cdot \frac{17}{20}}{\frac{71}{80}} |H\rangle \\ & = & \frac{19}{142}|L\rangle + \frac{36}{71}|M\rangle + \frac{51}{142}|H\rangle \\ & \sim & 0.134|L\rangle + 0.507|M\rangle + 0.359|H\rangle. \end{array}$$

As intuitively expected, a posteriori the probability of low blood pressure is higher than in the a priori distribution β – and the probability of high blood pressure is lower too.

These calculations with probabilities are relatively easy but may grow out of hand quickly. Therefore a library has been developed, called EfProb see Cho and Jacobs (2017), that provides the relevant functions, for validity, state update, state and predicate transformation, *etc.*

It is natural to see a state β and a channel s, as used above, as stochastic matrices

M_β and M_s, of the form:

$$M_\beta = \begin{pmatrix} \frac{3}{8} \\ \frac{1}{2} \\ \frac{3}{8} \end{pmatrix} \qquad M_s = \begin{pmatrix} 0.05 & 0.1 & 0.15 \\ 0.95 & 0.9 & 0.85 \end{pmatrix}$$

These matrices are called stochastic because the columns add up to 1. The matrix of the state $s \gg \beta$ is then obtained by matrix multiplication $M_s M_\beta$. For predicate transformation $s \ll \mathbf{1}_{d^\perp}$ with $M_{\mathbf{1}_{d^\perp}} = \begin{pmatrix} 0 & 1 \end{pmatrix}$ one uses matrix multiplication in a different order:

$$M_{\mathbf{1}_{d^\perp}} M_s = \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 0.05 & 0.1 & 0.15 \\ 0.95 & 0.9 & 0.85 \end{pmatrix} = \begin{pmatrix} 0.95 & 0.9 & 0.85 \end{pmatrix}.$$

The diligent reader may have noticed in this example that the probability $s \gg \beta d^\perp = s \gg \beta \models \mathbf{1}_{d^\perp} = \frac{71}{80}$ in Example 9.10 coincides with the probability $\beta \models s \ll \mathbf{1}_{d^\perp} = \frac{71}{80}$. This in fact in an instance of the following general result, relating validity and transformations.

Proposition 9.11 *Let $c \colon A \nrightarrow B$ be a channel, $\omega \in \mathcal{D}A$ be a state on its domain, and $q \in [0,1]^B$ a predicate on its codomain. Then:*

$$c \gg \omega \models q \;=\; \omega \models c \ll q. \tag{9.12}$$

Proof The result follows from a simple calculation:

$$\begin{aligned} c \gg \omega \models q &\overset{(9.4)}{=} \sum_{b \in B} c \gg \omega b \cdot qb \\ &\overset{(9.10)}{=} \sum_{b \in B} \left(\sum_{a \in A} \omega a \cdot cab \right) \cdot qb \\ &= \sum_{a \in A, b \in B} \omega a \cdot cab \cdot qb \\ &= \sum_{a \in A} \omega a \cdot \left(\sum_{b \in B} cab \cdot qb \right) \\ &\overset{(9.11)}{=} \sum_{a \in A} \omega a \cdot c \ll qa \\ &\overset{(9.4)}{=} \omega \models c \ll q. \end{aligned}$$

□

There are two more operations on channels that we need to consider, namely sequential composition $\odot$ and parallel composition $\otimes$.

Definition 9.12 Consider channels $f \colon A \nrightarrow B$, $g \colon C \nrightarrow D$ and $h \colon X \nrightarrow Y$. These channels can be composed sequentially and in parallel, yielding new channels:

$$g \odot f \colon A \nrightarrow C \qquad \text{and} \qquad f \otimes h \colon A \times X \nrightarrow B \times Y,$$

via the following definitions.

$$g \circ f a \;\coloneqq\; g \gg f a \qquad \text{so that} \qquad g \circ f ac \;=\; \sum_{b \in B} fab \cdot gbc.$$

The latter formula shows that channel composition is essentially matrix multiplication.

Next,

$$\begin{aligned} f \otimes ha, x \;&\coloneqq\; fa \otimes hx \qquad \text{so that} \\ f \otimes ha, xb, y \;&=\; fab \cdot hxy. \end{aligned}$$

The product $\otimes$ on the right of $\coloneqq$ is the product of states, as described in (9.2). In terms of matrices, parallel composition of channels is given by the Kronecker product.

It is not hard to see that $\circ$ and $\otimes$ are well-behaved operations, satisfying for instance:

$$(g \otimes k) \circ (f \otimes h) \;=\; (g \circ f) \otimes (k \circ h).$$

They interact nicely with state and predicate transformation:

$$\begin{aligned} (g \circ f \gg \omega \;&=\; g \gg (f \gg \omega \\ (g \circ f \ll q \;&=\; f \ll (g \ll q \\ (f \otimes h \gg \sigma \otimes \tau \;&=\; (f \gg \sigma) \otimes (h \gg \tau) \\ (f \otimes h \ll p \otimes q \;&=\; (f \ll p) \otimes (g \ll q). \end{aligned}$$

Moreover, for the identity channel id given by $\mathrm{id}x = 1|x\rangle$ we have:

$$\mathrm{id} \circ f \;=\; f \;=\; f \circ \mathrm{id} \qquad\qquad \mathrm{id} \otimes \mathrm{id} \;=\; \mathrm{id}.$$

We will see examples of parallel composition of channels in Section 9.7 when we discuss (the semantics of) Bayesian networks.

Remark 9.13 An ordinary function $f\colon A \to B$ can be turned into a 'deterministic' channel $\langle f \rangle\colon A \nrightarrow B$ via:

$$\langle f \rangle a \;\coloneqq\; 1|fa\rangle. \tag{9.13}$$

This operation $\langle \cdot \rangle$ sends function composition to channel composition: $\langle g \circ f \rangle = \langle g \rangle \circ \langle f \rangle$. The random variable $X\colon \Omega \to A$ that we used in Section 9.2.3 is an example of such a deterministic channel. Formally, we should now write $X^{-1}a = \langle X \rangle \ll \mathbf{1}_a$ for the event $X^{-1}a$ on Ω.

There are some further special cases of deterministic channels that we mention explicitly.

(i) For two sets A_1, A_2 we can form the cartesian product $A_1 \times A_2$ with its two projection functions $\pi_1 \colon A_1 \times A_2 \to A_1$ and $\pi_2 \colon A_1 \times A_2 \to A_2$. They can be turned into (deterministic) channels $\langle \pi_i \rangle \colon A_1 \times A_2 \nrightarrow A_i$. One can then see that marginalisation and weakening are state transformation and predicate transformation along these projection channels:

$$\langle \pi_1 \rangle \gg \omega \;=\; \mathsf{M}_1\omega \qquad \langle \pi_1 \rangle \ll p \;=\; \mathsf{W}_1 p \;=\; p \otimes \mathbf{1}$$
$$\langle \pi_2 \rangle \gg \omega \;=\; \mathsf{M}_2\omega \qquad \langle \pi_2 \rangle \ll q \;=\; \mathsf{W}_2 q \;=\; \mathbf{1} \otimes q$$

As a result, equation (9.6) is a special case of (9.12).

Moreover, these projection channels commute with parallel composition $\otimes$ of channels, in the sense that:

$$\langle \pi_1 \rangle \mathbin{\odot} \big(f \otimes h\big) \;=\; f \mathbin{\odot} \langle \pi_1 \rangle \qquad \langle \pi_2 \rangle \mathbin{\odot} \big(f \otimes h\big) \;=\; h \mathbin{\odot} \langle \pi_2 \rangle$$

(ii) For each set A there is a diagonal (or 'copy') function $\Delta \colon A \to A \times A$ with $\Delta a = a, a$. It can be turned into a channel too, as $\langle \Delta \rangle \colon A \nrightarrow A \times A$. However, this copy channel does *not* interact well with parallel composition of channels, in the sense that in general:

$$\langle \Delta \rangle \mathbin{\odot} f \;\neq\; \big(f \otimes f\big) \mathbin{\odot} \langle \Delta \rangle.$$

This equation does hold when the channel f is deterministic. Via diagonals we can relate parallel products $\otimes$ and conjunctions $\&$ of predicates:

$$\langle \Delta \rangle \ll \big(p \otimes q\big) \;=\; p \mathbin{\&} q.$$

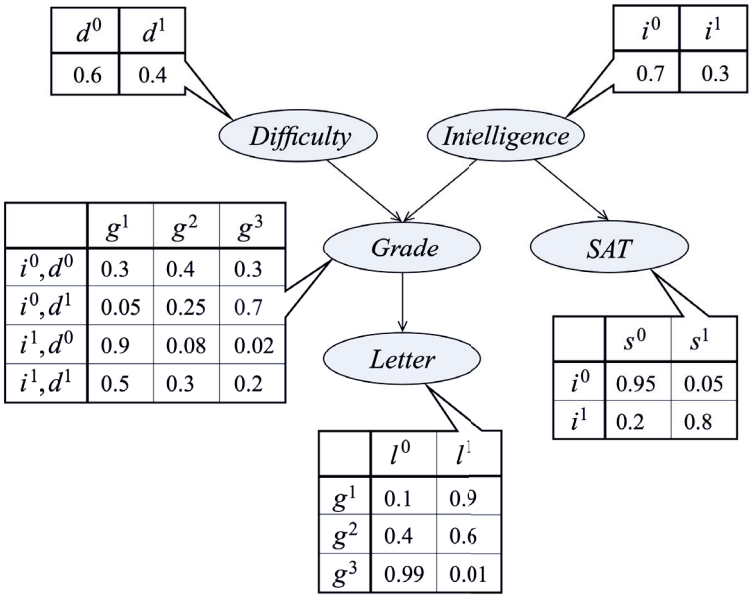
In what follows we often omit the braces $\langle \cdot \rangle$ around projections and diagonals, and simply write projection and copy channels as $\pi_i \colon A_1 \times A_2 \nrightarrow A_i$ and $\Delta \colon A \nrightarrow A \times A$.

9.6 Inference in Bayesian networks

In this section we illustrate how channels can be used both to model Bayesian networks and to reason about them. We shall use a standard example from the literature, namely the 'student' network from Koller and Friedman (2009). The graph of the student network is described in original form in Figure 9.1. We see how a student's grade depends on the difficulty of a test and the student's intelligence. The SAT score only depends on intelligence; whether or not the student gets a strong (l^1) or weak (l^0) recommendation letter depends on the grade.

With each of the five nodes in the network a sample space is associated, namely:

$$D = \{d^0, d^1\}, \quad I = \{i^0, i^1\}, \quad G = \{g^1, g^2, g^3\}, \quad S = \{s^0, s^1\}, \quad L = \{l^0, l^1\}.$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Figure 9.1 Picture of the student Bayesian network, copied from Koller and Friedman (2009), with conditional probability tables.

For the two inital nodes Difficulty (D) and Intelligence (I) we obtain two distributions/states ω_D and ω_I, whose probabilities are given in the two upper tables in Figure 9.1:

$$\omega_D \;=\; 0.6|\,d^0\,\rangle + 0.4|\,d^1\,\rangle \qquad \omega_I \;=\; 0.7|\,i^0\,\rangle + 0.3|\,i^1\,\rangle.$$

They capture the a priori state of affairs, with a 0.4 likelihood of a difficult test (d^1), and a 0.3 likelihood of an intelligent student (i^1).

The remaining three nodes Grade (G), Letter (L) and SAT (S) have incoming arrows from parent nodes, and are thus not initial. They correspond to three channels:

$$c_G \colon D \times I \nrightarrow G, \qquad c_S \colon I \nrightarrow S, \qquad c_L \colon G \nrightarrow L.$$

The definitions of these channels can be read directly from the tables. The SAT channel $c_S \colon I \nrightarrow S$ and the Letter channel $c_L \colon G \nrightarrow L$ are thus of the form:

$$\begin{aligned} c_S i^0 \;&=\; 0.95|\,s^0\,\rangle + 0.05|\,s^1\,\rangle \\ c_S i^1 \;&=\; 0.2|\,s^0\,\rangle + 0.8|\,s^1\,\rangle \end{aligned} \qquad \begin{aligned} c_L g^1 \;&=\; 0.1|\,l^0\,\rangle + 0.9|\,l^1\,\rangle \\ c_L g^2 \;&=\; 0.4|\,l^0\,\rangle + 0.6|\,l^1\,\rangle \\ c_L g^3 \;&=\; 0.99|\,l^0\,\rangle + 0.01|\,l^1\,\rangle \end{aligned}$$

The Grade channel $c_G \colon D \times I \nrightarrow G$ looks as follows.

$$\begin{aligned} c_G d^0, i^0 &= 0.3|g^1\rangle + 0.4|g^2\rangle + 0.3|g^3\rangle \\ c_G d^0, i^1 &= 0.9|g^1\rangle + 0.08|g^2\rangle + 0.02|g^3\rangle \\ c_G d^1, i^0 &= 0.05|g^1\rangle + 0.25|g^2\rangle + 0.7|g^3\rangle \\ c_G d^1, i^1 &= 0.5|g^1\rangle + 0.3|g^2\rangle + 0.2|g^3\rangle \end{aligned}$$

(Notice that we switched the order of i and d wrt. the tables in Figure 9.1; we have done so in order to remain consistent with the order of the inputs D and I as suggested in the network in Figure 9.1. This is actually a subtle issue, because usually in graphs there is no order on the parents of a node, that is, the parents form a *set* and not a *list*.)

We now discuss a number of inference questions from Koller and Friedman (2009) and illustrate how they are answered systematically using our perspective with states, predicates and channels.

(i) **What are the a priori probabilities for the recommendation?** To answer this question we follow the graph in Figure 9.1 and see that the answer is given by twice using state transformation, namely:

$$\begin{aligned} c_L \gg (c_G \gg \omega_D \otimes \omega_I) &= 0.498|l^0\rangle + 0.502|l^1\rangle, \qquad \text{or, equivalently,} \\ &= (c_L \circ c_G) \gg \omega_D \otimes \omega_I. \end{aligned}$$

(ii) **What if we know that the student is not intelligent?** The non-intelligence translates into the point predicate $\mathbf{1}_{i^0}$ on the set I, which we use to update the intelligence state ω_I before doing the same state transformations:

$$c_L \gg (c_G \gg \omega_D \otimes \omega_I|_{\mathbf{1}_{i^0}}) = 0.611|l^0\rangle + 0.389|l^1\rangle.$$

(iii) **What if we additionally know that the test is easy?** The easiness evidence translates into the predicate $\mathbf{1}_{d^0}$ on D, which is used for updating the difficulty state:

$$\begin{aligned} & c_L \gg (c_G \gg \omega_D|_{\mathbf{1}_{d^0}} \otimes \omega_I|_{\mathbf{1}_{i^0}}) \\ &= 0.487|l^0\rangle + 0.513|l^1\rangle \\ &= (c_L \circ c_G) \gg (\omega_D \otimes \omega_I|_{\mathbf{1}_{d^0} \otimes \mathbf{1}_{i^0}}). \end{aligned}$$

The previous two outcomes are obtained by what is called 'causal reasoning' or 'prediction' or 'forward inference', see the table at the end of Section 9.2.3. We continue with 'backward inference', also called 'evidential reasoning' or 'explanation'.

(iv) **What is the intelligence given a C-grade (g^3)?** The evidence predicate $\mathbf{1}_{g^3}$ is a predicate on the set G. We like to learn about the revised intelligence. This is done as follows. Via predicate transformation we obtain a predicate $c_G \ll \mathbf{1}_{g^3}$ on $D \times I$. We can use it to update the product state $\omega_D \otimes \omega_I$. We then get the update intelligence by taking the second marginal, as in:

$$\mathsf{M}_2\big(\omega_D \otimes \omega_I|_{c_G \ll \mathbf{1}_{g^3}}\big) \;=\; 0.921|\,i^0\,\rangle + 0.0789|\,i^1\,\rangle.$$

We see that the new intelligence (i^1) is significantly lower than the a priori value of 0.3, once a low grade is observed. The updated difficulty (d^1) probability is higher than the original 0.4; it is obtained by taking the first marginal:

$$\mathsf{M}_1\big(\omega_D \otimes \omega_I|_{c_G \ll \mathbf{1}_{g^3}}\big) \;=\; 0.371|\,d^0\,\rangle + 0.629|\,d^1\,\rangle.$$

(v) **What is the intelligence given a weak recommendation?** We now have a point predicate $\mathbf{1}_{l^0}$ on the set L. Hence we have to do predicate transformation twice, along the channels c_L and c_G, in order to reach the initial states. This is done as:

$$\begin{aligned}\mathsf{M}_2\big(\omega_D \otimes \omega_I|_{c_G \ll c_L \ll \mathbf{1}_{l^0}}\big) \;&=\; 0.86|\,i^0\,\rangle + 0.14|\,i^1\,\rangle, \qquad \text{or, equivalently,}\\ &=\; \mathsf{M}_2\big(\omega_D \otimes \omega_I|_{c_L \circ c_G \ll \mathbf{1}_{l^0}}\big).\end{aligned}$$

(vi) **What is the intelligence given a C-grade but a high SAT score?** We now have two forms of evidence, namely the point predicate $\mathbf{1}_{g^3}$ on G for the C-grade, and the point predicate $\mathbf{1}_{s^1}$ on S for the high SAT score. We can transform the latter to a predicate $c_S \ll \mathbf{1}_{s^1}$ on the set I and update the state ω_I with it. Then we can procede as in question (iv):

$$\mathsf{M}_2\big(\omega_D \otimes \omega_I|_{c_S \ll \mathbf{1}_{s^1}}|_{c_G \ll \mathbf{1}_{g^3}}\big) \;=\; 0.422|\,i^0\,\rangle + 0.578|\,i^1\,\rangle.$$

Thus the probability of high intelligence is 57.8% under these circumstances.

Using earlier calculation rules, see notably in Lemma 9.7, this intelligence distribution can also be computed by weakening the predicate $c_S \ll \mathbf{1}_{s^1}$ to $\mathsf{W}_2 c_S \ll \mathbf{1}_{s^1}$ on $D \times I$. Then we can take the conjunction with $c_G \ll \mathbf{1}_{g^3}$ and perform a single update, as in:

$$\mathsf{M}_2\big(\omega_D \otimes \omega_I|_{\mathsf{W}_2 c_S \ll \mathbf{1}_{s^1} \,\&\, c_G \ll \mathbf{1}_{g^3}}\big)$$

But one can also do the update with $c_S \ll \mathbf{1}_{s^1}$ at the very end, after the marginalisation, as in:

$$\mathsf{M}_2\big(\omega_D \otimes \omega_I|_{c_G \ll \mathbf{1}_{g^3}}\big)|_{c_S \ll \mathbf{1}_{s^1}}$$

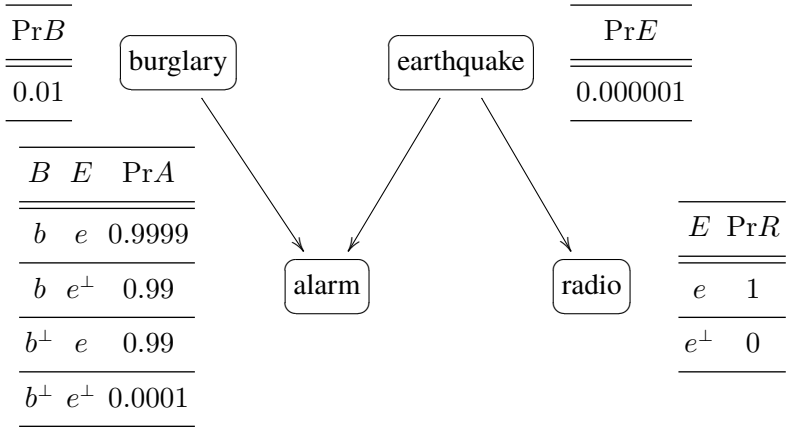
The associated difficulty level is the first marginal:

$$\mathsf{M}_1\big(\omega_D \otimes \omega_I|_{\mathbf{1}_{s^1}}|_{c_G \ll \mathbf{1}_{g^3}}\big) \;=\; 0.24|\,d^0\,\rangle + 0.76|\,d^1\,\rangle.$$

The answers to the above questions hopefully convey the systematic thinking that is behind the use of channels – in forward or backward manner, following the network structure – in order to capture the essence of Bayesian networks. This systematics is elaborated further in subsequent sections. In the above 'student' example we have obtained the same outcomes as in traditional approaches. We conclude with an illustration where things differ.

Example 9.14 The power of the channel-based approach is that it provides a 'logic' for Bayesian inference, giving high-level expressions $c \gg \omega|_p$ and $\omega|_{c \ll q}$ for forward and backward inference. We include an illustration from Barber (2012) where our method produces a different outcome. The logical description may help to clarify the differences.

Consider the following Bayesian network, with nodes for burglary (B), earthquake (E), alarm (A) and radio (R).



$\Pr B$
0.01

$\Pr E$
0.000001

B	E	$\Pr A$
b	e	0.9999
b	$e^\perp$	0.99
$b^\perp$	e	0.99
$b^\perp$	$e^\perp$	0.0001

E	$\Pr R$
e	1
$e^\perp$	0

In this case we have binary sets $B = \{b, b^\perp\}$, $E = \{e, e^\perp\}$, $A = \{a, a^\perp\}$ and $R = \{r, r^\perp\}$ with initial states $\omega_B = 0.01|\,b\,\rangle + 0.99|\,b^\perp\,\rangle$ and $\omega_E = 0.000001|\,e\,\rangle + 0.999999|\,e^\perp\,\rangle$. There are two channels $c_A \colon B \times E \nrightarrow A$ and $c_R \colon E \nrightarrow R$ based on the above (two lower) tables. The predicted (a priori) alarm probability is 1%; it is computed as $c_A \gg \omega_B \otimes \omega_E = 0.01|\,a\,\rangle + 0.99|\,a^\perp\,\rangle$.

The following questions are asked in Barber (2012, Example 3.1 and 3.2).

(i) **What is the probability of a burglary given that the alarm sounds?** In this case we have evidence $\mathbf{1}_a$ on the set A, we pull it back to $B \times E$ along the channel c_A, and we update the joint state $\omega_B \otimes \omega_E$ and take the first marginal:

$$\mathsf{M}_1\big(\omega_B \otimes \omega_E|_{c_A \ll \mathbf{1}_a}\big) \;=\; 0.99000198|\,b\,\rangle + 0.00999802|\,b^\perp\,\rangle.$$

(ii) **What is this probability if we additionally hear a warning on the radio?** In that case we have additional evidence $\mathbf{1}_r$ on R, which is pulled back along the channel c_R and used to update the state ω_E. Then:

$$\mathsf{M}_1\big(\omega_B \otimes \omega_E|_{c_R \ll \mathbf{1}_r}|_{c_A \ll \mathbf{1}_a}\big) \;=\; 0.010099|\,b\rangle + 0.989901|\,b^\perp\rangle.$$

(iii) ... **"imagine that we are only 70% sure we heard the burglar alarm sounding"** In this situation we have a fuzzy predicate $q\colon A \to 0,1$ with $qa = 0.7$ and $qa^\perp = 0.3$. We perform the same computation as in question (i), but now with evidence q instead of $\mathbf{1}_a$. This yields:

$$\mathsf{M}_1\big(\omega_B \otimes \omega_E|_{c_A \ll q}\big) \;=\; 0.0229|\,b\rangle + 0.9771|\,b^\perp\rangle. \tag{9.14}$$

However, in Barber (2012) a completely different computation is performed, following Jeffrey's rule. The assumption about the alarm is not interpreted as a predicate, but as a state $\sigma = 0.7|\,a\rangle + 0.3|\,a^\perp\rangle$. The result is computed via a corresponding convex combination of states:

$$\begin{aligned} &0.7 \cdot \big(\text{update with evidence } a\big) \;+\; 0.3 \cdot \big(\text{update with evidence } a^\perp\big) \\ &=\; 0.7 \cdot \mathsf{M}_1\big(\omega_B \otimes \omega_E|_{c_A \ll \mathbf{1}_a}\big) \;+\; 0.3 \cdot \mathsf{M}_1\big(\omega_B \otimes \omega_E|_{c_A \ll \mathbf{1}_{a^\perp}}\big) \\ &=\; 0.693|\,b\rangle + 0.307|\,b^\perp\rangle. \end{aligned} \tag{9.15}$$

For questions (i) and (ii) our calculations coincide with the ones in Barber (2012), but for question (iii) the answers clearly differ. We briefly analyse the situation. For a more extensive analysis in terms of Jeffrey's rule versus Pearl's for updating with soft/fuzzy evidence we refer to Jacobs (2019).

- The above computation (9.14) and the one (9.15) from Barber (2012) are based on different ways of understanding what soft evidence actually means. In Barber (2012) this notion, even though it is not made mathematically precise, appears to have an ontological interpretation: "the alarm was heard" is a new state of affairs, with 70% alarm probability, and is therefor used as a state (distribution). On the other hand, our fuzzy predicate interpretation has an epistemological flavour: it is new information about an agent's belief.
- In line with the previous point, different intuitive descriptions are developed in Jacobs (2019): the approach in (9.14) factors the soft evidence in, as improvement, using Bayesian rules. The approach (9.15) interpretes the soft evidence as a 'surprising' state of affairs, that one has to adjust to – as correction – following Jeffrey's rule.

  The above formulation, quoted from Barber (2012), does not seem to suggest that the 70% certainty is a new, surprising state of affairs that we have to adjust to. Instead, it seems to be more like an improvement, so that the calculation (9.14) seems most appropriate.

We shall briefly return to these different was of computation in Example 9.17 where we show that the outcome in (9.14) also appears via 'crossover inference'.

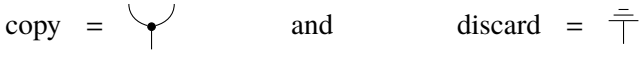
9.7 String diagrams for Bayesian probability

Abstractly, channels are arrows of a category, which is *symmetric monoidal*: it has sequential $\circ$ and parallel $\otimes$ composition. This categorical structure enables the use of a graphical (yet completely formal) notation for channels in terms of *string diagrams* (Selinger, 2011). We have no intention of giving a complete account of the string diagrammatic calculus here, and refer instead to Fong (2012), Jacobs and Zanasi (2016, Sec. 3), and Jacobs et al. (2019, Sec. 3) for details. Nonetheless, it is worthwhile pointing similarities and differences between the graphical representation of channels as string diagrams and the usual Bayesian network notation, see also Jacobs et al. (2019). We shall also use string diagrams to give a pictorial account of the important notion of disintegration (in the next section).

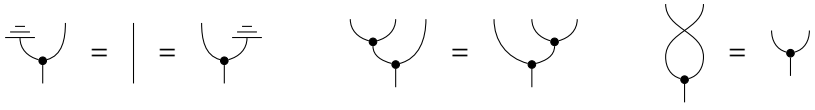
Informally speaking, string diagrams for channels are similar to the kind of graphs that is used for Bayesian networks, see Figure 9.1, but there are important differences.

(i) Whereas flow in Bayesian networks is top-down, we will adopt the convention that in string diagrams the flow is bottom-up. This is an non-essential, but useful difference, because it makes immediately clear in the current context whether we are dealing with a Bayesian network or with a string diagram. Also, it makes our presentation uniform with previous work, see *e.g.* Cho and Jacobs (2019).

(ii) The category where channels are arrows has extra structure, which allows for the use of "special" string diagrams representing certain elementary operations. We will have explicit string diagrams for *copying* and *discarding* variables, namely:



There are some 'obvious' equations between diagrams involving such copy and discard, such as:



These equations represent the fact that copy is the multiplication and discard is the unit of a commutative monoid.

(iii) With string diagrams one can clearly express joint states, on product domains like $X_1 \times X_2$, or $X_1 \times \cdots \times X_n$. This is done by using multiple outgoing pins, coming out of a triangle shape – used for states – as for $\omega \in \mathcal{D}X_1 \times X_2$ and

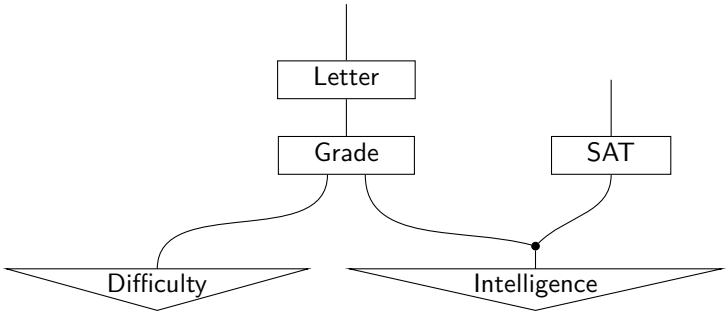
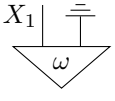
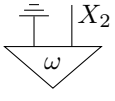


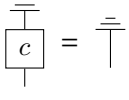
Figure 9.2 Student network from Figure 9.1 expressed as string diagram. Note that the two arrows coming out of Intelligence in Figure 9.1 are translated here into a single wire coming out the Intelligence state, followed by a copy.

$\sigma \in \mathcal{D}X_1 \times \cdots \times X_n$ in:

With this notation in place we can graphically express the marginals via discarding $\mathrel{\overline{\top}}$ of wires:

$\mathsf{M}_1\omega$ =  and $\mathsf{M}_2\omega$ = 

(iv) Channels are *causal* or *unitary* in the sense that discarding their output is the same as discarding their input:



The Intelligence node in Figure 9.1 has two outgoing arrows, but this does not mean that Intelligence is a joint state. Instead, these two arrows indicate that the outgoing wire should be copied, with one copy going to the Grade node and one to the SAT node. In string diagram notation this copying is written explicitly as in the string-diagrammatic analogue of the student network in Figure 9.2.

Recall that we wrote $\omega_D = 0.6|\,d^0\,\rangle + 0.4|\,d^1\,\rangle$ and $\omega_I = 0.7|\,i^0\,\rangle + 0.3|\,i^1\,\rangle$ for the initial states of the student network. The product state

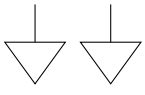
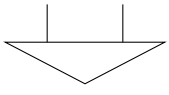
$$\omega_D \otimes \omega_I \;=\; 0.42|\,d^0, i^0\,\rangle + 0.18|\,d^0, i^1\,\rangle + 0.28|\,d^1, i^0\,\rangle + 0.12|\,d^1, i^1\,\rangle$$

is non-entwined, since it equals the product of its marginals ω_D and ω_I. A basic fact in probability is that conditioning can create entwinedness, see *e.g.* Jacobs and

Zanasi (2017) for more information. We can see this concretely when the above product state $\omega_D \otimes \omega_I$ is conditioned as in the fourth question in the previous section:

$$\begin{aligned}&\omega_D \otimes \omega_I\big|_{c_G \ll \mathbf{1}_{g^3}} \\ &= \tfrac{12600}{34636}|\,d^0, i^0\,\rangle + \tfrac{36}{34636}|\,d^0, i^1\,\rangle + \tfrac{19600}{34636}|\,d^1, i^0\,\rangle + \tfrac{2400}{34636}|\,d^1, i^1\,\rangle.\end{aligned}$$

With some effort one can show that this state is *not* the product of its marginals: it is entwined. In the language of string diagrams we can express this difference by writing:

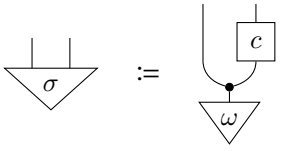
$\omega_D \otimes \omega_I \;=$  $\qquad\qquad \omega_D \otimes \omega_I\big|_{c_G \ll \mathbf{1}_{g^3}} \;=$ 

9.8 From joint states to Bayesian networks

Our framework allows to express states/distributions and Bayesian networks as entities of the same kind, namely as channels. It is natural to ask how the process of forming a Bayesian network from a distribution can be integrated in the picture.

In traditional probability theory, this procedure forms one of the original motivations for developing the notion of Bayesian network in the first place. Such networks allow for more efficient representation of probabilistic information (via probability tables, as in Figure 9.1) than joint states, which quickly become unmanageable via an exponential explosion. We quote from Koller and Friedman (2009): "... the explicit representation of the joint distribution is unmanageable from every perspective. Computationally, it is very expensive to manipulate and generally too large to store in memory" and from Russell and Norvig (2003): "... a Bayesian network can often be far more *compact* than the full joint distribution".

The procedure of forming a Bayesian network from a given state usually goes through a sub-routine called *disintegration*. For a channel-based definition of disintegration, suppose we have a state $\omega \in \mathcal{D}X$ and a channel $c\colon X \to Y$. Then we can form a joint state $\sigma \in \mathcal{D}X \times Y$ as described by the following string diagram:



that is $\qquad \sigma x, y \;=\; \omega x \cdot c x y. \qquad (9.16)$

The state ω is determined as the first marginal $\omega = \mathsf{M}_1\sigma$ of σ. This can be seen by discarding $\top$ the second wire – on the left and on the right in the above equation – and using that channels are causal, and that discarding one wire of a copy is the identity wire.

Disintegration is the process in the other direction, from a joint state to a channel.

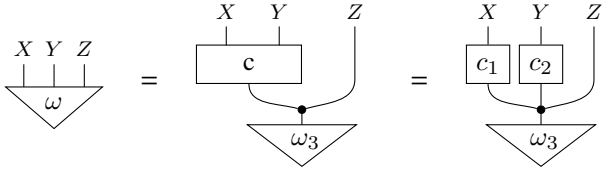
Definition 9.15 Let $\sigma \in \mathcal{D}X \times Y$ be a joint state. A *disintegration* of σ is a channel $c\colon X \nrightarrow Y$ for which the equation (9.16) holds, where $\omega = \mathsf{M}_1\sigma$.

There is a standard formula for disintegration of a state $\sigma \in \mathcal{D}X \times Y$, namely:

$$cx \coloneqq \mathsf{M}_2(\sigma|_{\mathbf{1}_x \otimes \mathbf{1}}) = \sum_y \frac{\sigma x, y}{\mathsf{M}_1\sigma x}|y\rangle. \tag{9.17}$$

We shall say that the channel c is 'extracted' from σ, or also that σ 'factorises' via c as in (9.16). Intuitively, channel c captures the conditional probabilities expressed in traditional notation as $\Pr_\sigma y \mid x$ via a distribution on Y indexed by elements $x \in X$.

Definition 9.15 gives the basic form of disintegration. There are several variations, which are explored in Cho and Jacobs (2019) as part of a more abstract account of this notion. For instance, by swapping the domains one can also extract a channel $Y \nrightarrow X$, in the other direction. Also, if σ is a joint state on n domains, there are in principle 2^n ways of extracting a channel, depending on which pins are marginalised out, and which (other) ones are reconstructed via the channel. For instance, a disintegration of $\omega \in \mathcal{D}X \times Y \times Z$ can also be a channel $c\colon Z \nrightarrow X \times Y$. This example suggests a digression on a channel-based definition of conditional independence: X and Y are conditionally independent in ω given Z, written as $X \perp Y \mid Z$, if any such disintegration c for ω can actually be decomposed into channels $c_1\colon Z \nrightarrow X$ and $c_2\colon Z \nrightarrow Y$. In string diagrams:



(9.18)

where $\omega_3 = \mathsf{M}_3\omega = \Pr_\omega z$ is the third marginal. These channels c_1, c_2 may also be obtained by disintegration from the state $\mathsf{M}_{1,2}\omega = \Pr_\omega x, y$ obtained by marginalising out the third variable. In more traditional notation, one can intuitively read (9.18) as saying that $\Pr_\omega x, y, z = \Pr_\omega x \mid z \cdot \Pr_\omega y \mid z \cdot \Pr_\omega z$. We refer to Cho and Jacobs (2019) for the adequacy of this definition of conditional independence and its properties.

Another interesting observation is that disintegration forms a *modular* procedure. The formula (9.16) shows that disintegration yields a new decomposition of a given state: such a decomposition being a state itself, disintegration may be applied again. In fact, this repeated application is how a joint state on multiple domains gets represented as a Bayesian network. The channel-based approach understands this process uniformly as a step-by-step transformation of a given channel (a state) into another, equivalent channel (a Bayesian network). Once again, string diagrams are a useful formalism for visualising such correspondence. For instance, the joint state

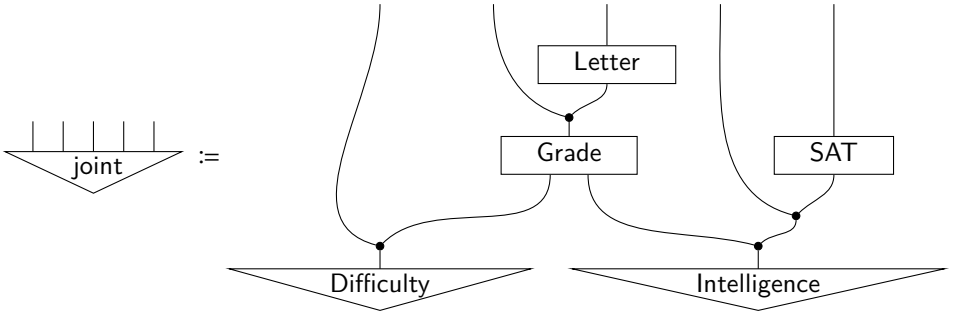


Figure 9.3 The joint distribution (9.19) for the student network from Figure 9.1 obtained as string diagram with additional copiers for non-final nodes.

associated with the Student network from Figures 9.1 and 9.2 can be expressed as in Figure 9.3. Notice that the diagram in Figure 9.3 is just the one in Figure 9.2 where each non-final node has been made externally accessible via additional copiers $\curlyvee$. Figure 9.3 has the type of a joint state. Its value can then be calculated via state transformation, via a composite channel that can be "read off" directly from the graph in Figure 9.3, namely:

$$\begin{aligned} &\mathsf{joint} \\ &:= \big(\mathrm{id}\otimes\mathrm{id}\otimes c_L\otimes\mathrm{id}\otimes\mathrm{id}\circ\mathrm{id}\otimes\Delta\otimes\mathrm{id}\otimes c_S\circ\mathrm{id}\otimes c_G\otimes\Delta\circ\Delta\otimes\Delta\big) \\ &\qquad \gg \big(\omega_D\otimes\omega_I\big). \end{aligned} \tag{9.19}$$

The tool EfProb, see Cho and Jacobs (2017), has been designed precisely to evaluate such systematic expressions.

Now that we have a formal description of the relationship between joint states and Bayesian networks, we turn to comparing Bayesian inference in these two settings. For reasons of simplicity, we concentrate on the binary case. Suppose we have a joint state $\sigma \in \mathcal{D}X \times Y$, now with *evidence* on X. In the present setting this evidence can be an arbitrary predicate $p \in 0,1^X$ and not only a point predicate $\mathbf{1}_x$, as usual. We like to find out the distribution on Y, given the evidence p. Informally, this may be written as $\Pr Y \mid p$. More precisely, it is the second marginal of the state obtained by updating with the weakened version $\mathsf{W}_1 p = p \otimes \mathbf{1}$, as in:

$$\mathsf{M}_2\big(\sigma|_{\mathsf{W}_1 p}\big).$$

Now suppose we have factorised the joint state σ as a (mini) network (9.16) via the extracted state $c\colon X \mathrel{-\!\!\circ\!\!\to} Y$. We can also perfom causal reasoning – *i.e.* forward inference – and obtain the state:

$$c \gg \omega|_p \qquad \text{where} \qquad \omega \;=\; \mathsf{M}_1\sigma.$$

The *Bayesian inference theorem* says that these outcomes are the same, not only for forward reasoning, but also for backward reasoning.

Theorem 9.16 *Let $\sigma \in \mathcal{D}X \times Y$ be a joint state with extracted channel $c\colon X \nrightarrow Y$ as in* (9.16)*. For predicates $p \in [0,1]^X$ and $q \in [0,1]^Y$ one has:*

$$\mathsf{M}_2(\sigma|_{\mathsf{W}_1 p}) \;=\; c \gg (\mathsf{M}_1\sigma|_p) \qquad \textit{and} \qquad \mathsf{M}_1(\sigma|_{\mathsf{W}_2 q}) \;=\; \mathsf{M}_1\sigma|_{c \ll q}.$$

Before giving a proof, we comment on the significance of the statement. Inference with joint states, as on the left-hand-side of the equations in Theorem 9.16, involves weakening W of evidence in one coordinate and marginalisation M in another coordinate. It uses the entwinedness of the joint state σ, so that one coordinate can influence the other, see Jacobs and Zanasi (2017) where this is called *crossover* influence. Therefor we like to call this form of inference via joint states *crossover inference*.

In contrast, inference on the right-hand-side of the equations in Theorem 9.16 essentially uses state and predicate transformation $\gg$ and $\ll$. Therefor we refer to this form of inference as *transformer inference*. It consists of what we have called backward and forward inference in the table at the end of Section 9.2.3.

Thus the Bayesian inference theorem states the equivalence of crossover inference and transformer inference. Whereas crossover inference works well with small samples (see the examples below), it does not scale to larger networks, where transformations inference is preferable. The equivalence is widely known at some implicit level, but its formulation in this explicit form only arises within the current channel-based perspective on Bayesian networks.

We now provide a proof of the theorem. A purely diagrammatic argument is given in Cho and Jacobs (2019).

Proof of Theorem 9.16 We confine ourselves to proving the first equation in concrete form, using the definition of extracted channel from (9.17):

$$\begin{aligned}
&(c \gg \mathsf{M}_1\sigma|_p)y \\
&\quad = \sum_x cxy \cdot \mathsf{M}_1\sigma|_p x \\
&\quad \overset{(9.17)}{=} \sum_x \frac{\sigma x, y}{\mathsf{M}_1\sigma x} \cdot \frac{\mathsf{M}_1\sigma x \cdot px}{\mathsf{M}_1\sigma \models p} \\
&\quad \overset{(9.6)}{=} \sum_x \frac{\sigma x, y \cdot px}{\sigma \models \mathsf{W}_1 p} \\
&\quad = \sum_x \frac{\sigma x, y \cdot \mathsf{W}_1 px, y}{\sigma \models \mathsf{W}_1 p} \qquad \text{since } \mathsf{W}_1 px, y = p \otimes \mathbf{1}x, y = px \\
&\quad = \sum_x \sigma|_{\mathsf{W}_1 p} x, y \\
&\quad = \mathsf{M}_2(\sigma|_{\mathsf{W}_1 p})y.
\end{aligned}$$

□

We conclude this section by giving two demonstrations of the equivalence stated in Theorem 9.16. First, we answer once again the six questions about the student network in Section 9.6: whereas therein we applied transformer inference, we now compute using crossover inference. We shall write:

$$\mathsf{joint} \in \mathcal{D}\big(D \times G \times L \times I \times S\big)$$

for the joint state associated with the student network, obtained in formula (9.19), following Figure 9.3. We write M_i for the i-th marginal, obtained by summing out all domains which are not in the i-th position. For the sake of clarity we do not use the notation W for weakening, but use parallel product with the truth predicate $\mathbf{1}$ instead. In agreement with Theorem 9.16, the outcomes are the same as in Section 9.6, but they have been computed separately (in EfProb).

(i) **What are the a priori probabilities for the recommendation?**

$$\mathsf{M}_3\mathsf{joint} \;=\; 0.498|\,l^0\,\rangle + 0.502|\,l^1\,\rangle.$$

(ii) **What if we know that the student is not intelligent?**

$$\mathsf{M}_3\big(\mathsf{joint}|_{\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}_{i^0}\otimes\mathbf{1}}\big) \;=\; 0.611|\,l^0\,\rangle + 0.389|\,l^1\,\rangle.$$

(iii) **What if we additionally know that the test is easy?**

$$\mathsf{M}_3\big(\mathsf{joint}|_{\mathbf{1}_{d^0}\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}_{i^0}\otimes\mathbf{1}}\big) \;=\; 0.487|\,l^0\,\rangle + 0.513|\,l^1\,\rangle.$$

(iv) **What is the intelligence given a C-grade (g^3)?**

$$\mathsf{M}_4\big(\mathsf{joint}|_{\mathbf{1}\otimes\mathbf{1}_{g^3}\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}}\big) \;=\; 0.921|\,i^0\,\rangle + 0.0789|\,i^1\,\rangle.$$

(v) **What is the intelligence given a weak recommendation?**

$$\mathsf{M}_4\big(\mathsf{joint}|_{\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}_{l^0}\otimes\mathbf{1}\otimes\mathbf{1}}\big) \;=\; 0.86|\,i^0\,\rangle + 0.14|\,i^1\,\rangle.$$

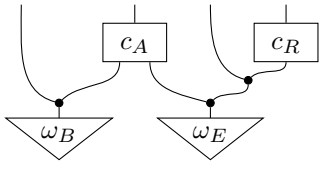
(vi) **What is the intelligence given a C-grade but a high SAT score?**

$$\mathsf{M}_4\big(\mathsf{joint}|_{\mathbf{1}\otimes\mathbf{1}_{g^3}\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}_{s^1}}\big) \;=\; 0.422|\,i^0\,\rangle + 0.578|\,i^1\,\rangle.$$

As a second demonstration of the theorem, we briefly return to the controversy around inference with soft predicates in Example 9.14.

Example 9.17 We first re-arrange the Bayesian network from Example 9.14 in string diagrammatic form so that we can compute the joint state $\omega \in \mathcal{D}B \times A \times E \times R$

as:

$\omega \;:=$ 

i.e.

$$\begin{aligned}\omega \;=\; \big(&\mathrm{id} \otimes \mathrm{id} \otimes c_R \otimes \mathrm{id} \\ &\circ \mathrm{id} \otimes c_A \otimes \Delta \\ &\circ \Delta \otimes \Delta\big) \gg \omega_B \otimes \omega_E.\end{aligned}$$

Recall that we have soft/fuzzy evidence $qa = 0.7, qa^{\perp} = 0.3$ on A. Given this evidence, we want to know the burglar probability. Using crossover inference it is computed as:

$$\mathsf{M}_1\big(\omega|_{\mathbf{1}\otimes q\otimes \mathbf{1}\otimes \mathbf{1}}\big) \;=\; 0.0229|b\rangle + 0.9771|b^{\perp}\rangle.$$

We obtain the same outcome as via transformer inference in (9.14). Of course, the Bayesian Inference Theorem 9.16 tells that the outcomes should coincide in general. This additional computation just provides further support for the appropriateness of doing inference via forward and backward transformations along channels.

9.9 Conclusions

This chapter provides an introduction to an emerging area of channel-based probability theory. It uses standard compositional techniques from programming semantics in the area of Bayesian inference, giving a conceptual connection between forward and backward inference (or: causal and evidential reasoning) on the one hand, and crossover influence on the other.

Promising research directions within this framework include the development of channel-based algorithms for Bayesian reasoning, see Jacobs (2018a). Moreover, the abstract perspective offered by the channel approach may apply to probabilistic graphical models other than Bayesian networks, including models for machine learning such as neural networks (see Jacobs and Sprunger, 2019 for first steps). Paired with the mathematical language of string diagrams, this framework may eventually offer a unifying compositional perspective on the many different pictorial notations for probabilistic reasoning.

Acknowledgements

Fabio Zanasi acknowledges support from EPSRC grant nr. EP/R020604/1. Bart Jacobs' research has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement nr. 320571.

References

Barber, D. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge Univ. Press. publicly available via http://web4.cs.ucl.ac.uk/staff/D.Barber/pmwiki/pmwiki.php?n=Brml.HomePage.

Bartels, F., Sokolova, A., and de Vink, E. 2004. A hierarchy of probabilistic system types. *Theoretical Computer Science*, **327(1-2)**, 3–22.

Cho, K., and Jacobs, B. 2017. The EfProb Library for Probabilistic Calculations. In: Bonchi, F., and König, B. (eds), *Conference on Algebra and Coalgebra in Computer Science (CALCO 2017)*. LIPIcs, vol. 72. Schloss Dagstuhl.

Cho, K., and Jacobs, B. 2019. Disintegration and Bayesian inversion via string diagrams. *Math. Struct. in Comp. Sci.*, **29(7)**, 938–971.

Cho, K., Jacobs, B., Westerbaan, A., and Westerbaan, B. 2015. *An Introduction to Effectus Theory*. see arxiv.org/abs/1512.05813.

Culbertson, J., and Sturtz, K. 2014. A Categorical Foundation for Bayesian Probability. *Appl. Categorical Struct.*, **22(4)**, 647–662.

de Vink, E., and Rutten, J. 1999. Bisimulation for probabilistic transition systems: a coalgebraic approach. *Theoretical Computer Science*, **221**, 271–293.

Dijkstra, E., and Scholten, C. 1990. *Predicate Calculus and Program Semantics*. Berlin: Springer.

Dijkstra, E. W. 1997. *A Discipline of Programming*. 1st edn. Upper Saddle River, NJ, USA: Prentice Hall PTR.

Dubois, D., and Prade, H. 1990. The Logical View of Conditioning and Its Application to Possibility and Evidence Theories. *Int. Journ. of Approximate Reasoning*, **4**, 23–46.

Fong, B. 2012. *Causal Theories: A Categorical Perspective on Bayesian Networks*. M.Phil. thesis, Univ. of Oxford. see arxiv.org/abs/1301.6201.

Giry, M. 1982. A categorical approach to probability theory. Pages 68–85 of: Banaschewski, B. (ed), *Categorical Aspects of Topology and Analysis*. Lect. Notes Math., no. 915. Springer, Berlin.

Jacobs, B. 2011. Probabilities, Distribution Monads, and Convex Categories. *Theoretical Computer Science*, **412(28)**, 3323–3336.

Jacobs, B. 2013. Measurable Spaces and their Effect Logic. In: *Logic in Computer Science*Computer Science Press, for IEEE.

Jacobs, B. 2017. A recipe for State and Effect Triangles. *Logical Methods in Comp. Sci.*, **13(2)**. See https://lmcs.episciences.org/3660.

Jacobs, B. 2018a. *A Channel-based Exact Inference Algorithm for Bayesian Networks*. See arxiv.org/abs/1804.08032.

Jacobs, B. 2018b. From Probability Monads to Commutative Effectuses. *Journ. of Logical and Algebraic Methods in Programming*, **94**, 200–237.

Jacobs, B. 2019. The Mathematics of Changing one's Mind, via Jeffrey's or via Pearl's update rule. *Journ. of Artif. Intelligence Res.*, **65**, 783–806.

Jacobs, B., and Sprunger, D. 2019, to appear. Neural Nets via Forward State Transformation and Backward Loss Transformation. In: König, B. (ed), *Math. Found. of Programming Semantics*. Elect. Notes in Theor. Comp. Sci. Elsevier, Amsterdam. See arxiv.org/abs/1803.09356.

Jacobs, B., and Zanasi, F. 2016. A predicate/state transformer semantics for Bayesian learning. Pages 185–200 of: Birkedal, L. (ed), *Math. Found. of Programming Semantics*. Elect. Notes in Theor. Comp. Sci., no. 325. Elsevier, Amsterdam.

Jacobs, B., and Zanasi, F. 2017. A Formal Semantics of Influence in Bayesian Reasoning. In: Larsen, K., Bodlaender, H., and Raskin, J.-F. (eds), *Math. Found. of Computer Science*. LIPIcs, vol. 83. Schloss Dagstuhl.

Jacobs, B., Kissinger, A., and Zanasi, F. 2019. Causal Inference by String Diagram Surgery. Pages 313–329 of: Bojańczyk, M., and Simpson, A. (eds), *Foundations of Software Science and Computation Structures*. Lect. Notes Comp. Sci., no. 11425. Springer, Berlin.

Jones, C., and Plotkin, G. 1989. A probabilistic powerdomain of evaluations. Pages 186–195 of: *Logic in Computer Science*Computer Science Press, for IEEE.

Jung, A., and Tix, R. 1998. The Troublesome Probabilistic Powerdomain. Pages 70–91 of: Edalat, A., Jung, A., Keimel, K., and Kwiatkowska, M. (eds), *Comprox III, Third Workshop on Computation and Approximation*. Elect. Notes in Theor. Comp. Sci., no. 13. Elsevier, Amsterdam.

Keimel, K. 2008. The monad of probability measures over compact ordered spaces and its Eilenberg-Moore algebras. *Topology and its Applications*, **156**, 227–239.

Keimel, K., and Plotkin, G. 2009. Predicate transformers for extended probability and non-determinism. *Math. Struct. in Comp. Sci.*, **19(3)**, 501–539.

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models. Principles and Techniques*. Cambridge, MA: MIT Press.

Kozen, D. 1981. Semantics of probabilistic programs. *Journ. Comp. Syst. Sci*, **22(3)**, 328–350.

Kozen, D. 1985. A probabilistic PDL. *Journ. Comp. Syst. Sci*, **30(2)**, 162–178.

Mermin, N.D. 2007. *Quantum Computer Science: An Introduction*. Cambridge Univ. Press.

Mislove, M. 2012. Probabilistic Monads, Domains and Classical Information. Pages 87–100 of: Kashefi, E., Krivine, J., and van Raamsdonk, F. (eds), *Developments of Computational Methods (DCM 2011)*. Elect. Proc. in Theor. Comp. Sci., no. 88.

Mrad, A. Ben, Delcroix, V., Piechowiak, S., Leicester, P., and Abid, M. 2015. An explication of uncertain evidence in Bayesian networks: likelihood evidence and probabilistic evidence. *Applied Intelligence*, **23(4)**, 802–824.

Panangaden, P. 2009. *Labelled Markov Processes*. London: Imperial College Press.

Russell, S., and Norvig, P. 2003. *Artificial Intelligence. A Modern Approach*. Prentice Hall.

Ścibior, A., Ghahramani, Z., and Gordon, A. 2015. Practical Probabilistic Programming with Monads. Pages 165–176 of: *Proc. 2015 ACM SIGPLAN Symp. on Haskell*. ACM.

Ścibior, A., Kammar, O., Vákár, M., Staton, S., Yang, H., Cai, Y., Ostermann, K., Moss, S., Heunen, C., and Ghahramani, Z. 2018. Denotational Validation of Higher-order Bayesian Inference. Pages 60:1–60:29 of: *Princ. of Programming Languages*. ACM Press.

Selinger, P. 2011. A survey of graphical languages for monoidal categories. *Springer Lecture Notes in Physics*, **13**(813), 289–355.

Sokolova, A. 2011. Probabilistic systems coalgebraically: A survey. *Theoretical Computer Science*, **412(38)**, 5095–5110.

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* Computer Science Press, for IEEE.

Tix, R., Keimel, K., and Plotkin, G. 2005. *Semantic Domains for Combining Probability and Non-Determinism*. Elect. Notes in Theor. Comp. Sci., no. 129. Elsevier, Amsterdam.

Valtorta, M., Kim, Y.-G., and Vomlel, J. 2002. Soft evidential update for probabilistic multiagent systems. *Int. Journ. of Approximate Reasoning*, **29(1)**, 71–106.

Varacca, D., and Winskel, G. 2006. Distributing probability over non-determinism. *Math. Struct. in Comp. Sci.*, **16**, 87–113.