Multisets and Distributions, in Drawing and Learning [⋆]

Bart Jacobs

Institute for Computing and Information Sciences,
Radboud University, Nijmegen, The Netherlands
bart@cs.ru.nl

March 12, 2021

Abstract. Multisets are 'sets' in which elements may occur multiple times. Discrete probability distributions capture states in which elements may occur with probabilities that add up to one. This paper describes how the interaction between multisets and distributions lies at the heart of some basic constructions in probability theory, especially in distributions arising from drawing from an urn with multiple balls and in learning distributions from multiple occurrences of data. Drawing multiple balls from an urn is described uniformly in terms of Kleisli iteration for a monad, covering the four standard distinctions of ordered/unordered draws, with/without replacement. In probabilistic learning the paper distinguishes two forms of likelihood, based on also on iteration, with corresponding forms of learning. Both of these forms occur in the literature, but they are not clearly distinguished, even though they lead to different outcomes.

1 Introduction

When we wish to combine elements from a certain set there are several mechanisms to do so, depending on whether or not the order of the elements is relevant and on how to count these elements. Concretely, one can use:

- *subsets*, in which neither the order nor the multiplicity of the elements matters;
- *lists*, in which the order matters, and in which elements may occur multiple times;
- *multisets*, in which the order of elements is irrelevant, but elements may occur multiple times;
- *distributions*, in which the order does not matter, but where elements may occur with a certain probability, taken from the unit interval $[0, 1]$, in such a way that all probabilities add up to one.

All these collection mechanisms can be described in terms of monads on the category of sets.

This article concentrates on the latter two collection types, namely multisets and distributions, and in particular on their interaction. This interaction will be studied in the following two typical situations.

[⋆] To appear in: A. Palmigiano and M. Sadrzadeh (eds), *Samson Abramsky on Logic and Structure in Computer Science and Beyond*, Springer 2021.

1. Suppose we have an urn with three red (R) and two green (G) balls, represented as a multiset $3|R\rangle + 2|G\rangle$. Then the probability of drawing a red ball is $\frac{3}{5}$. Thus, one can sample a multiset by drawing elements — as one draws elements from an urn — with a certain probability determined by the multiplicities in the multiset. Such drawing may be repeated, where the distinction is important whether or not a drawn ball is replaced to the urn, and whether or not the order of multiple draws matters.
2. Suppose we have a coin with unknown bias, and flipping it three times gives outcomes head (H), tail (T), and head again. What is then the probability that the next flip is head? In this situation the data form a multiset $2|H\rangle + 1|T\rangle$, which can be used to update the prior bias distribution — which we assume to be uniform, since there is no prior knowledge. Then one can calculate the probability of head in the updated, newly learned distribution — which, in this case, is $\frac{3}{5}$.

The two topics of this paper are thus: drawing and learning, both starting from a multiset, and both yielding a probability distribution. These two topics form the two main parts of the paper: Section 3 is on drawing from an urn, and Sections 6 – 9 are on learning. Section 5 forms the glue between drawing and learning. In between, Sections 2 and 4 provide background information about multisets and distributions and about predicates and probabilistic conditioning/updating.

In many textbooks on probability, see *e.g.* [33,34], one finds the physical model of an urn with coloured balls, from which balls can be drawn with a certain probability. This can be done in four different ways, depending on whether the order matters in a draw of multiple balls, and whether withdrawn balls are replaced into the urn or deleted. The first part of this article on drawing introduces four 'transition' operations for drawing from an urn, corresponding to these four distinctions. All the transitions form Kleisli (endo)maps for the monad $\mathcal{D}(M \times -)$, combining the distribution monad $\mathcal{D}$ with the writer monad $M \times -$, for a monoid M. This M is the (non-commutative) list monoid for ordered draws, and the (commutative) multiset monoid for unordered draws. By iterating the transition map, using Kleisli composition for the combined monad, and then taking the first marginal, one obtains appropriate distributions on draws. In this way we reconstruct the familiar multinomial and hypergeometric distributions, for unordered draws, and two more distributions for ordered draws. This part reorganises existing material into a canonical form. It forms a topic of its own that could be used for instance in a course on the use of categorical methods, especially in probability theory.

The second part on probabilistic learning elaborates the idea that learning is about finding a probability distribution that best fits given data. In general, such learning is described as consisting of small steps that need to be repeated in order to reach a certain optimum. These steps can be used to increase the likelihood of data or to decrease errors. The latter approach is generally based on gradient descent and occurs for instance in logistic regression (see *e.g.* [11]). Here we concentrate on the first approach, increasing likelihood, but we do relate it to decreasing divergence at some point (name in Proposition 20).

We organise the data from which we learn as multisets. Learning involves finding the distribution, possibly in an iterative process, that gives highest likelihood to the data. Our approach leads to *two* forms of likelihood, called 'external' and 'internal'. It is shown how both forms of likelihood arise from repeated transitions, like for drawing. Associated with these two likelihoods there are two techniques for learning. Both forms of learning occur in the literature, but they are not clearly distinguished, even though they can lead to quite different outcomes. In times where learning from huge amounts of data has become common, proper understanding of the basic concepts is not only scientifically but also practically (societally) urgent. Here, the difference is illustrated in several examples (from the literature), including coin bias learning (internal), and parameter learning and Expectation-Maximisation (both external). In the end we do not offer a mathematical criterion for when to use internal / external likelihood (and learning); for now we only have an intuitive perspective, see Section 9.

This second part extends material from an earlier conference publication [24], for instance with the new descriptions of conjugate priorship in Corollary 25 and of Expectation-Maximisation in Theorem 26, with several examples, and with the discussion about 'external' and 'internal' in Section 9.

The formalisations and results in this paper demonstrate that at a very elementary level there is categorical (esp. monadic) structure in probability theory. Of course, this observation is well-known by now, starting with the early work of Lawvere and Giry [20] and of Kozen [29,30] in the 1980s. The research contributions of Samson Abramsky fit in this line of work, as inspiration in unveiling fundamental mathematical (often categorical) structure in many areas, including *e.g.* physics and economy. Abramsky has not worked so much in (easy) classical probability theory; his work concentrates on the much more difficult field of quantum probability, with a focus on its categorical structure [6,8,2,9,5], on its inherent limitations [1,4] and especially on contextuality [7,3,12]. Abramsky has been influential for my own ERC advanced grant (2012-2017) in this area, which laid the foundation for the current work.

2 Multisets and distributions

This section briefly introduces (finite) multisets and (finite discrete probability) distributions. They both are collection types that can have elements occurring multiple times or with certain probabilities.

First we like to fix our notation for lists/sequences. We write $\mathcal{L}(X)$ for the set of finite sequences $[x_1, \ldots, x_n]$ of elements $x_i \in X$, of length n. This set forms a monoid, with concatenation $+\!\!+$ as binary operation and empty list $[]$ as neutral element. As is well-known, the operation $\mathcal{L}$ forms a monad on the category of sets.

2.1 Multisets

A multiset (or bag) is a 'set' in which (finitely many) elements may occur multiple times, with natural numbers as multiplicities. We write $\mathcal{M}(X)$ for the set of such multisets over a set X, defined as:

$$\mathcal{M}(X) \coloneqq \{\phi\colon X \to \mathbb{N} \mid \mathrm{supp}(\phi) \text{ is finite}\},$$

where $\mathrm{supp}(\phi) \subseteq X$ is the support of ϕ, *i.e.* the subset $\{x \in X \mid \phi(x) \neq 0\}$. We often write concrete multisets as finite formal sums, using a 'ket' notation: $\phi = \sum_i n_i|x\rangle$, where $\mathrm{supp}(\phi) = \{x_1, \ldots, x_n\}$ and $n_i = \phi(x_i) \in \mathbb{N}$. Taking multisets on a set is functorial: for $f\colon X \to Y$ we get $\mathcal{M}(f)\colon \mathcal{M}(X) \to \mathcal{M}(Y)$ via $\mathcal{M}(f)(\phi)(y) = \sum_{x \in f^{-1}(y)} \phi(x)$. Alternatively, in terms of formal sums: $\mathcal{M}(f)(\sum_i n_i|x_i\rangle) = \sum_i n_i|f(x_i)\rangle$. In fact, $\mathcal{M}$ is a monad on the category of sets. Moreover, $\mathcal{M}(X)$ with pointwise addition and empty multiset $\mathbf{0}$, is the free commutative monoid on X. This monoid is ordered: for $\phi, \psi \in \mathcal{M}(X)$ we write $\phi \leq \psi$ if $\phi(x) \leq \psi(x)$ for all $x \in X$. This implies $\mathrm{supp}(\phi) \subseteq \mathrm{supp}(\psi)$. In that case we write $\psi - \phi \in \mathcal{M}(X)$ for the obvious multiset, with multiplicities $(\psi - \phi)(x) = \psi(x) - \phi(x)$. The situation $\phi \leq \psi$ arises for instance when ψ is an urn with balls, and ϕ is a handful of balls drawn from the urn.

One can associate several numbers with a multiset. The next definition gives an overview.

Definition 1. *Let $\phi \in \mathcal{M}(X)$ be a multiset on a set X.*

1. *The* size $\|\phi\| \in \mathbb{N}$ *is the total number of elements occurring in a multiset, taking multiplicities into account:*

$$\|\phi\| \coloneqq \sum_{x \in X} \phi(x).$$

2. *The* factorial $\phi\,!\!! \in \mathbb{N}$ *the product of factorials of multiplicities:*

$$\phi\,!\!! \coloneqq \prod_{x \in X} \phi(x)!.$$

3. *The* multiset coefficient*, or simply* coefficient $\boldsymbol{(}\phi\boldsymbol{)}$ *of ϕ is a multinomial coefficient:*

$$\boldsymbol{(}\phi\boldsymbol{)} \coloneqq \frac{\|\phi\|!}{\phi\,!\!!} = \frac{\|\phi\|!}{\prod_x \phi(x)!} = \binom{\|\phi\|}{\phi(x_1) \;\cdots\; \phi(x_n)},$$

The latter multinomial coefficient formulation assumes that ϕ's support is $\{x_1, \ldots, x_n\}$.

4. *Finally, when $\phi \leq \psi \in \mathcal{M}(X)$ we use a binomial coefficient of multisets as product of binomial coefficients of multiplicities:*

$$\begin{aligned} \binom{\psi}{\varphi} &\coloneqq \frac{\psi\,!\!!}{\varphi\,!\!! \cdot (\psi - \varphi)\,!\!!} \\ &= \frac{\prod_x \psi(x)!}{\big(\prod_x \varphi(x)!\big) \cdot \big(\prod_x (\psi(x) - \varphi(x))!\big)} = \prod_{x \in X} \binom{\psi(x)}{\varphi(x)}. \end{aligned}$$

Frequently we like to have some grip on the total number of elements occurring in a multiset, taking multiplicities into account. We write for $K \in \mathbb{N}$,

$$\mathcal{M}[K](X) \coloneqq \{\phi \in \mathcal{M}(X) \mid \|\phi\| = K\}.$$

Clearly, $\mathcal{M}[0](X)$ is a singleton, containing only the empty multiset $\mathbf{0}$. This $\mathcal{M}[K]$ is a functor, but not a monad. However, it has the structure of a *graded* monad with respect to the monoid of natural numbers with multiplication.

There is an accumulation map $acc\colon \mathcal{L} \Rightarrow \mathcal{M}$, turning lists into multisets, given by $acc([x_1, \ldots, x_n]) = 1|x_1\rangle + \cdots + 1|x_n\rangle$. Thus, *e.g.*, $acc([a, a, b, b, a] = 3|a\rangle + 2|b\rangle$. This accumulation forms a map of monads. We often use accumulation for a fixed size $K \in \mathbb{N}$, and then write it as $acc[K]\colon X^K \to \mathcal{M}[K](X)$. The parameter $K \in \mathbb{N}$ is omitted when it is clear for the context.

A basic question is: how many lists accumulate to a given multiset ϕ? The (standard) answer is: $\boldsymbol{(}\phi\boldsymbol{)}$. For instance, for $\phi = 2|a\rangle + 3|b\rangle$ there are $\boldsymbol{(}\phi\boldsymbol{)} = \frac{5!}{2!\cdot 3!} = 10$ sequences with length 5 of a's and b's that accumulate to ϕ. It is not hard to see that $\boldsymbol{(}-\boldsymbol{)}$ satisfies the following recurrence equation:

$$\boldsymbol{(}\phi\boldsymbol{)} = \sum_{y \in \mathrm{supp}(\phi)} \boldsymbol{(}\phi - 1|y\rangle\boldsymbol{)}. \tag{1}$$

The following result is a generalisation of Vandermonde's formula. We include a proof, for convenience.

Lemma 2. *Let $\psi \in \mathcal{M}(X)$ be a multiset of size $L = \|\psi\|$, with a number $K \leq L$. We write $\phi \leq_K \psi$ as short-hand for: $\phi \in \mathcal{M}[K](X)$ with $\phi \leq \psi$. Then:*

$$\sum_{\phi \leq_K \psi} \binom{\psi}{\phi} = \binom{L}{K} \qquad \textit{so that} \qquad \sum_{\phi \leq_K \psi} \frac{\binom{\psi}{\phi}}{\binom{L}{K}} = 1.$$

Proof. We use induction on the number of elements in $\mathrm{supp}(\psi)$. We go through some initial values explicitly. If the number is 0, then $\psi = \mathbf{0}$ and so $L = 0 = K$ and $\phi \leq_K \psi$ means $\phi = \mathbf{0}$, so that the result holds. Similarly, if $\mathrm{supp}(\psi)$ is a singleton, say $\{x\}$, then $L = \psi(x)$. For $K \leq L$ and $\phi \leq_K \psi$ we get $\mathrm{supp}(\phi) = \{x\}$ and $K = \phi(x)$. The result then obviously holds.

The case where $\mathrm{supp}(\psi) = \{x, y\}$ captures the ordinary form of Vandermonde's formula. We reformulate it for numbers $B, G \in \mathbb{N}$ and $K \leq B + G$. Then:

$$\tbinom{B+G}{K} = \sum_{b \leq B,\, g \leq G,\, b+g=K} \tbinom{B}{b} \cdot \tbinom{G}{g}. \tag{2}$$

Intuitively: if you select K children out of B boys and G girls, the number of options is given by the sum over the options for $b \leq B$ boys times the options for $g \leq G$ girls, with $b + g = K$.

The equation (2) can be proven by induction on G. When $G = 0$ both sides amount to $\binom{B}{K}$ so we proceed to the induction step. The case $K = 0$ is trivial, so we may assume $K > 0$. We use what's called Pascal's rule $\binom{n+1}{m} = \binom{n}{m} + \binom{n}{m-1}$ for binomials.

$$\begin{array}{rl}
& \displaystyle\sum_{b\leq B,\, g\leq G+1,\, b+g=K} \tbinom{B}{b}\cdot\tbinom{G+1}{g} \\
= & \tbinom{B}{K}\cdot\tbinom{G+1}{0} + \tbinom{B}{K-1}\cdot\tbinom{G+1}{1} + \cdots + \tbinom{B}{K-G}\cdot\tbinom{G+1}{G} + \tbinom{B}{K-G-1}\cdot\tbinom{G+1}{G+1} \\
= & \tbinom{B}{K}\cdot\tbinom{G}{0} + \tbinom{B}{K-1}\cdot\tbinom{G}{1} + \tbinom{B}{K-1}\cdot\tbinom{G}{0} \\
& \qquad + \cdots + \tbinom{B}{K-G}\cdot\tbinom{G}{G} + \tbinom{B}{K-G}\cdot\tbinom{G}{G-1} + \tbinom{B}{K-G-1}\cdot\tbinom{G}{G} \\
= & \displaystyle\sum_{b\leq B,\, g\leq G,\, b+g=K} \tbinom{B}{b}\cdot\tbinom{G}{g} + \sum_{b\leq B,\, g\leq G,\, b+g=K-1} \tbinom{B}{b}\cdot\tbinom{G}{g} \\
\overset{\text{(IH)}}{=} & \tbinom{B+G}{K} + \tbinom{B+G}{K-1} \\
= & \tbinom{B+G+1}{K}.
\end{array}$$

We now turn to the (first) equation in Lemma 2. For the induction step, let $\mathrm{supp}(\psi) = \{x_1, \ldots, x_n, y\}$, for $n \geq 2$. Writing $\ell = \psi(y)$, $L' = L - \ell$ and $\psi' = \psi - \ell|y\rangle \in \mathcal{M}[L'](X)$ gives:

$$\begin{array}{rcl}
\displaystyle\sum_{\phi\leq_K\psi} \tbinom{\psi}{\phi} \;=\; \sum_{\phi\leq_K\psi} \prod\nolimits_x \tbinom{\psi(x)}{\phi(x)} & = & \displaystyle\sum_{n\leq\ell}\;\sum_{\phi\leq_{K-n}\psi'} \tbinom{\ell}{n}\cdot\prod\nolimits_i \tbinom{\psi(x_i)}{\phi(x_i)} \\
& \overset{\text{(IH)}}{=} & \displaystyle\sum_{n\leq\ell,\, K-n\leq L-\ell} \tbinom{\ell}{n}\cdot\tbinom{L-\ell}{K-n} \overset{(2)}{=} \tbinom{L}{K}. \qquad \square
\end{array}$$

For completeness we also include the *multinomial theorem*, without proof.

Lemma 3. *For $K \in \mathbb{N}$ and $a_1, \ldots, a_n \in \mathbb{R}$,*

$$\big(a_1 + \cdots + a_n\big)^K \;=\; \sum_{\phi\in\mathcal{M}[K](\{1,\ldots,n\})} \boldsymbol{(}\phi\boldsymbol{)}\cdot a_1^{\phi(1)}\cdot\ldots\cdot a_n^{\phi(n)}. \qquad \square$$

2.2 Distributions

A distribution (or a state, or multinomial) is like a multiset but where its multiplicities are taken from the unit interval $[0, 1]$ and add up to one. We thus define the set $\mathcal{D}(X)$ of distribution on a set X as:

$$\mathcal{D}(X) \coloneqq \{\phi\colon X \to [0,1] \mid \mathrm{supp}(\phi) \text{ is finite, and } \textstyle\sum_x \phi(x) = 1\}.$$

This $\mathcal{D}$ is also monad on the category of sets.

A *channel* $f\colon X \rightsquigarrow Y$ is a probabilistic computation from X to Y. Notice that it is written with a small circle on the shaft of the arrow. A channel can be understood as an X-indexed collection of states on Y, or alternatively, as a conditional probability $f(y \mid x)$. We look at a channel more categorically, as a 'Kleisli' map $f\colon X \to \mathcal{D}(Y)$. Such a channel can 'push' a state $\omega \in \mathcal{D}(X)$ forward to a state $f \gg \omega \in \mathcal{D}(Y)$, via 'Kleisli extension' or 'state transformation', where

$(f \gg \omega)(y) = \sum_x \omega(x) \cdot f(x)(y)$. Via $\gg$ we can define composition $g \circ f$ of channels as $(g \circ f)(x) = g \gg f(x)$.

An example of a channel is what we call *arrangement* $arr\colon \mathcal{M}(X) \mathrel{\circ\!\!\!\to} \mathcal{L}(X)$. It maps a multiset ϕ to a (uniform) distribution over the sequences $\vec{x}$ that accumulate to ϕ. As we have seen before, there are $\boldsymbol{(}\phi\boldsymbol{)}$-many such sequences. Hence:

$$arr(\phi) \coloneqq \sum_{\vec{x} \in acc^{-1}(\phi)} \frac{1}{\boldsymbol{(}\phi\boldsymbol{)}} \big| \vec{x} \big\rangle. \tag{3}$$

This channel restricts to $arr\colon \mathcal{M}[K](X) \mathrel{\circ\!\!\!\to} X^K$. The composite $acc \circ arr$ is the identity channel $\mathcal{M}[K](X) \mathrel{\circ\!\!\!\to} \mathcal{M}[K](X)$. In the other direction, $arr \circ acc\colon X^K \mathrel{\circ\!\!\!\to} X^K$ sends a sequence to the uniform distribution of all its permutations.

We shall use the parallel product $\sigma \otimes \tau$ of distributions $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Y)$. It is a distribution on the product space $X \times Y$ defined as:

$$\big(\sigma \otimes \tau\big)(x, y) \,=\, \sigma(x) \cdot \tau(y).$$

We write $iid[K]\colon \mathcal{D}(X) \mathrel{\circ\!\!\!\to} X^K$ for the channel that maps a state ω to the K-fold tensor: $iid(\omega) = \omega^K = \omega \otimes \cdots \otimes \omega \in \mathcal{D}(X^K)$. This gives the so-called identical and independent distribution.

Similarly, for channels $f\colon A \mathrel{\circ\!\!\!\to} X$ and $g\colon B \mathrel{\circ\!\!\!\to} Y$ we get $f \otimes g\colon A \times B \mathrel{\circ\!\!\!\to} X \times Y$ via $(f \otimes g)(a, b) = f(a) \otimes g(b)$. This makes the Kleisli category $\mathcal{K}\ell(\mathcal{D})$ of the distribution monad symmetric monoidal.

An ordinary function $f\colon X \to Y$ is often implicitly promoted to a channel $f\colon X \mathrel{\circ\!\!\!\to} Y$ via $x \mapsto 1|f(x)\rangle$. This is used in particular for diagonals $\Delta = \langle \mathrm{id}, \mathrm{id}\rangle\colon X \to X \times X$ and projections $\pi_i\colon X_1 \times X_2 \to X_i$. Marginalisation of $\omega \in \mathcal{D}(X_1 \times X_2)$ can then be described as $\pi_i \gg \omega \in \mathcal{D}(X_i)$. One has $\pi_i \gg (\sigma_1 \otimes \sigma_2) = \sigma_i$, but in general:

$$(\pi_1 \gg \omega) \otimes (\pi_2 \gg \omega) \,\neq\, \omega \qquad \text{and} \qquad \Delta \gg \sigma \,\neq\, \sigma \otimes \sigma.$$

For two channels $c\colon A \mathrel{\circ\!\!\!\to} X$ and $d\colon A \mathrel{\circ\!\!\!\to} Y$ we shall write $\langle c, d\rangle = (c \otimes d) \circ \Delta\colon A \mathrel{\circ\!\!\!\to} X \times Y$ for their tuple.

As is well-known in probability theory, from a joint state $\tau \in \mathcal{D}(X \times Y)$ one can extract a channel (conditional probability) $c\colon X \mathrel{\circ\!\!\!\to} Y$, given by:

$$c(x)(y) \,=\, \frac{\omega(x, y)}{(\pi_1 \gg \omega)(x)} \qquad \text{so that} \qquad \langle \mathrm{id}, c\rangle \gg (\pi_1 \gg \omega) \,=\, \omega. \tag{4}$$

Clearly, this channel extraction is a partial operation, since the first marginal needs to be non-zero. The latter equation is commonly written as $\omega(y \mid x) \cdot \omega(x) = \omega(x, y)$. This extraction of a channel is called 'disintegration', see [14,19,15].

We include two classic examples, that play an important role later on: multinomial and hypergeometric distributions. Informally, they assign a probability to taking a handful of coloured balls from an urn. These distributions are most common in *binary form*, for an urn with two colours only. Here we look at the *multivariate* form, with an arbitrary set X of colours. We shall describe this “handful”, say with K balls, as a multiset of size K. Thus, multinomial and hypergeometric distributions produce outcomes in the set $\mathcal{D}\big(\mathcal{M}[K](X)\big)$ of distributions on multisets of size K. The difference between multinomial and hypergeometric distributions lies in whether drawn balls are replaced into the urn or not. When the balls are replaced, in the case of multinomials, the urn itself may be represented abstractly as a distribution. When the drawn balls are actually deleted, the urn changes with every draw, and is represented as a multiset. We shall describe the multinomial and geometric distributions as channels, of the form:

$$\mathcal{D}(X) \xrightarrow{\mathit{mulnom}[K]} \mathcal{M}[K](X) \qquad\qquad \mathcal{M}[L](X) \xrightarrow{\mathit{hypgeom}[K]} \mathcal{M}[K](X), \tag{5}$$

where K is the number of drawn balls. In the hypergeometric case one needs $K \leq L$, where L is the number of balls in the urn. The channels are defined on a distribution $\omega \in \mathcal{D}(X)$ and multiset $\psi \in \mathcal{M}[L](X)$ as:

$$\begin{aligned} \mathit{mulnom}[K](\omega) &\coloneqq \sum_{\phi\in\mathcal{M}[K](X)} (\phi) \cdot \prod\nolimits_x \omega(x)^{\phi(x)} \,\big|\phi\big\rangle \\ \mathit{hypgeom}[K](\psi) &\coloneqq \sum_{\phi\leq_K\psi} \frac{\binom{\psi}{\phi}}{\binom{L}{K}} \,\big|\phi\big\rangle. \end{aligned} \tag{6}$$

Recall that we write $\phi \leq_K \psi$ as shorthand for: $\phi \in \mathcal{M}[K](X)$ with $\phi \leq \psi$. The multinomial definition yields a distribution via Lemma 3. In the hypergeometric case we use Lemma 2.

2.3 Frequentist learning

In general in probabilistic learning, one learns from ‘data’. A perspective that underlies this paper is that such data are naturally organised as multisets. For instance, if we wish to learn about the bias of an arbitrary coin, we need data in the form of coin flips. If we have seen 10 heads and 9 tails, we will organise these flips as a single multiset of the form $10|H\rangle + 9|T\rangle$ over the set $\{H, T\}$, whose elements represent head and tail. In a multiset the order of elements does not matter. This corresponds to the fact that the order of data elements does not matter in probabilistic learning.

One basic form of learning starts by counting. This is what we call frequentist learning *Flrn*; it amounts to normalisation. For instance $\mathit{Flrn}(10|H\rangle + 9|T\rangle) = \frac{10}{19}|H\rangle + \frac{9}{19}|T\rangle$. In general, for a non-empty multiset $\phi \in \mathcal{M}_*(X)$,

$$\mathit{Flrn}(\phi) \coloneqq \sum_{x\in\mathrm{supp}(\phi)} \frac{\phi(x)}{\|\phi\|}\big|x\big\rangle \quad \text{where, recall,} \quad \|\phi\| = \textstyle\sum_x \phi(x) > 0. \tag{7}$$

This result $\mathit{Flrn}(\phi) \in \mathcal{D}(X)$ is often called the empirical distribution. It is typical of such frequentist learning that learning from more of the same does not have any effect. We can make this precise via the equation:

$$Flrn\big(K \cdot \phi\big) \;=\; Flrn(\phi) \qquad \text{for } K > 0. \tag{8}$$

It is not hard to see that *Flrn* is a natural transformation $\mathcal{M}_* \Rightarrow \mathcal{D}$. This means in particular that it commutes with marginalisation. Thus, if one applies frequentist learning *Flrn* to a multi-dimensional table $\tau \in \mathcal{M}_*(X_1 \times \cdots \times X_n)$ it does not matter if one learns from the entire table first and then marginalises to $\mathcal{D}(X_i)$, or if one first adds up totals in columns $\mathcal{M}(X_i)$ and then applies frequentist learning.

When one has already learned the distribution $Flrn(\phi)$ and a new batch of data ψ arrives, all probabilities have to be re-adjusted, as in the convex sum of distributions:

$$Flrn\big(\phi + \psi\big) \;=\; \frac{\|\phi\|}{\|\phi\| + \|\psi\|} \cdot Flrn(\phi) + \frac{\|\psi\|}{\|\phi\| + \|\psi\|} \cdot Flrn(\psi).$$

3 Drawing from an urn

The very basic concepts of probability theory are often explained in terms of urns: containers of objects of a certain kind, typically coloured balls. One can then draw a ball from the urn, whose colour probability is determined by the different numbers of the various balls in the urn. Such drawing can be repeated, where drawn balls are either replaced, or not. Also, the order of drawn balls may be taken into account, or not. The four cases are commonly described in terms ordered / unordered draws with / without replacement. They can be represented in a 2 × 2 table, see (11) below.

Our aim is to describe these four cases in a principled manner via probabilistic channels. In order to do so we first look at single-draw transition mappings, which may be described informally as:

$$Urn \longmapsto \Big(single\text{-}draw,\; Urn'\Big) \tag{9}$$

We use the *ad hoc* notation Urn' of an urn with an accent, to describe the urn after the draw. It may be the same urn as before, in case of a draw with replacement, or it may be a different urn, with one ball missing, namely the original urn without the single ball that was drawn.

The above transition arrow will be described as a probabilistic channel. It gives for each single draw the associated probability. In this description we shall combine multisets and distributions. For instance, an urn with three red balls and two blue ones will be described as a multiset $3|R\rangle + 2|B\rangle$. The transition associated with drawing a single ball *without* replacement gives a mapping:

$$3|R\rangle + 2|B\rangle \longmapsto \tfrac{3}{5}\Big|R,\, 2|R\rangle + 2|B\rangle\Big\rangle + \tfrac{2}{5}\Big|B,\, 3|R\rangle + 1|B\rangle\Big\rangle$$

It gives the $\frac{3}{5}$ probability of drawing a red ball, together with the remaining urn, and a $\frac{2}{5}$ probability of drawing a blue one, with a different new urn.

The situation *with* replacement is given by:

$$3|R\rangle + 2|B\rangle \longmapsto \tfrac{3}{5}\big|R,\, 3|R\rangle + 2|B\rangle\big\rangle + \tfrac{2}{5}\big|B,\, 3|R\rangle + 2|B\rangle\big\rangle$$

Here we see that the urn/multiset does not change. An important first observation is that in that case we may as well use a distribution as urn, instead of a multiset. The distribution represents an abstract urn. In the above example we would use the distribution $\frac{3}{5}|R\rangle + \frac{2}{5}|B\rangle$ as abstract urn, when we draw with replacement. The distribution contains all the relevant information. Clearly, it is obtained via frequentist learning from the original multiset. Using distributions instead of multisets gives more flexibility, since not all distributions are obtained via frequentist learning — in particular when the probabilities are proper real numbers and not fractions.

We formulate this approach explicitly.

- In a situation *without* replacement, an urn is a (non-empty, natural) multiset, which changes with every draw, via removal of the drawn ball. This no-replacement scenario will also be described in terms of deletion.
- In a situation *with* replacement, an urn is a probability distribution; it does not change when balls are drawn.

This covers the first distinction, between draws with and without replacement. The second distinction between ordered and unordered draws cannot be made for single draw transitions. Hence we need to suitably iterate the single-draw transition (9) to:

$$\mathit{Urn} \longmapsto \Big(\mathit{multiple\text{-}draws},\, \mathit{Urn}'\Big) \tag{10}$$

Now we can make the distinction between ordered and unordered draws explicit. Let X be the set of colours, for the balls in the urn — so $X = \{R, B\}$ in the above illustration.

- An *ordered* draw of multiple balls, say K many, is represented via a list $X^K = X \times \cdots \times X$ of length K.
- An *unordered* draw of K-many balls is represented as a K-sized multiset, in $\mathcal{M}[K](X)$.

Thus, in the latter case, both the urn and the handful of balls drawn from it, are represented as a multiset.

In the end we are interested in assigning probabilities to draws, ordered or not, with replacement or without. These probabilities on draws are obtained by taking the first marginal/projection of the iterated transition map (10). It yields a mapping from an urn to multiple draws. The following table gives an overview of the types of these operations, where X is the set of colours of the balls.

K-sized draws	**with replacement**	**with deletion**
ordered	$\mathcal{D}(X) \xrightarrow{\mathit{OdR}}\!\!\circ\; X^K$	$\mathcal{M}[L](X) \xrightarrow{\mathit{OdD}}\!\!\circ\; X^K$
unordered	$\mathcal{D}(X) \xrightarrow{\mathit{UdR}}\!\!\circ\; \mathcal{M}[K](X)$	$\mathcal{M}[L](X) \xrightarrow{\mathit{UdD}}\!\!\circ\; \mathcal{M}[K](X)$

(11)

We see that in the replacement scenario the inputs of these channels are distributions in $\mathcal{D}(X)$, as abstract urns. In the deletion scenario (without replacements) the input (urns) are multisets in $\mathcal{M}[L](X)$, of size L. In the ordered case the outputs are tuples in X^K of length K and in the unordered case they are multisets in $\mathcal{M}[K](X)$ of size K. Implicitly in this table we assume that $L \geq K$, so that the urn is full enough for K single draws.

We see that the table (11) combines the basic data types of lists, multisets and distributions. The names of the channels in the table reflect the two distinctions Below we explain these short names and relate them to commonly used names.

- *UdR* = unordered-draw-with-replacement; we will show that it is the multinomial channel, on the left in (5);
- *UdD* = unordered-draw-with-deletion; this will turn out to be the hypergeometric channel, on the right in (5);
- *OdR* = ordered-draw-with-replacement; it is the identical and independent (iid) channel;
- *OdD* = ordered-draw-with-deletion.

In the last situation there is no established name, so we shall simply use the short name *OdD* from Table 11.

Below we elaborate how the channels in Table 11 actually arise. It makes the earlier informal descriptions in (9) and (10) mathematically precise. We use that for any monoid M, the mapping $X \mapsto M \times X$ is a monad, called the writer monad. This can be combined with the distribution monad $\mathcal{D}$, giving a combined monad $X \mapsto \mathcal{D}(M \times X)$. It comes with an associated Kleisli composition $\odot$. It is precisely this composition that we use for iterating a single draw. Moreover, for ordered draws we use the monoid $M = \mathcal{L}(X)$ of lists, and for unordered draws we use the monoid $M = \mathcal{M}(X)$ of multisets. It is rewarding, from a formal perspective, to see that from this abstract principled approach, common distributions for different sorts of drawing arise, including the well-known multinomial and hypergeometric distributions.

Lemma 4. *Let $M = (M, 0, +)$ be a monoid. The mapping $X \mapsto \mathcal{D}\big(M \times X\big)$ is a monad on* **Sets**, *with unit $\eta \colon X \to \mathcal{D}\big(M \times X\big)$ given by:*

$$\eta(x) \,=\, 1\big|\, 0, x\,\big\rangle \qquad \textit{where } 0 \in M \textit{ is the zero element.}$$

For Kleisli maps $f \colon A \to \mathcal{D}(M \times B)$ and $g \colon B \to \mathcal{D}(M \times C)$ there is the Kleisli composition $g \odot f \colon A \to \mathcal{D}(M \times C)$ given by:

$$\big(g \odot f\big)(a) \,=\, \sum_{m,m',c} \Big(\textstyle\sum_b f(a)(m,b) \cdot g(b)(m',c)\Big)\,\big|\, m + m', c\,\big\rangle. \tag{12}$$

Notice the occurrence of the sum $+$ of the monoid M in the first component of the ket $\big|-,-\big\rangle$ in (12). When M is the list monoid, this sum is the (non-commutative) concatenation $+\!\!+$ of lists, producing an ordered list of drawn elements. When M is the multiset monoid, this sum is the (commutative) $+$ of multisets, so that the accumulation of drawn elements yields a multiset, in which the order of elements is irrelevant.

If we have an 'endo' Kleisli map for the combined monad of Lemma 4, of the form $t\colon A \to \mathcal{D}(M \times A)$, we can iterate it K times, giving $t^K\colon A \to \mathcal{D}(M \times A)$. This iteration is defined via the above unit and Kleisli composition:

$$t^0 = \eta \qquad \text{and} \qquad t^{K+1} = t^K \odot t = t \odot t^K.$$

Below in (13) we define the four transition channels for drawing a single element from an urn. In the "with replacement" column on the left the distribution ω acts as abstract urn and remains unchanged. In the "without replacement" column on the right, the drawn element x is actually removed from the urn/multiset ψ via subtraction $\psi - 1|x\rangle$. Implicitly it is assumed that the multiset ψ is non-empty.

$$\begin{array}{ll}
\mathcal{D}(X) \xrightarrow{OtR} \mathcal{D}\big(\mathcal{L}(X) \times \mathcal{D}(X)\big) & \mathcal{M}(X) \xrightarrow{OtD} \mathcal{D}\big(\mathcal{L}(X) \times \mathcal{M}(X)\big) \\
\quad \omega \longmapsto \displaystyle\sum_{x \in \mathrm{supp}(\omega)} \omega(x)\big|[x], \omega\big\rangle & \quad \psi \longmapsto \displaystyle\sum_{x \in \mathrm{supp}(\psi)} \frac{\psi(x)}{\|\psi\|}\big|[x], \psi - 1|x\rangle\big\rangle \\
\mathcal{D}(X) \xrightarrow{UtR} \mathcal{D}\big(\mathcal{M}(X) \times \mathcal{D}(X)\big) & \mathcal{M}(X) \xrightarrow{UtD} \mathcal{D}\big(\mathcal{M}(X) \times \mathcal{M}(X)\big) \\
\quad \omega \longmapsto \displaystyle\sum_{x \in \mathrm{supp}(\omega)} \omega(x)\big|1|x\rangle, \omega\big\rangle & \quad \psi \longmapsto \displaystyle\sum_{x \in \mathrm{supp}(\psi)} \frac{\psi(x)}{\|\psi\|}\big|1|x\rangle, \psi - 1|x\rangle\big\rangle
\end{array} \tag{13}$$

In the subsections below we analyse what iteration means for these four channels. Subsequently, we can describe the associated K-sized draw channels, as first projection $\pi_1 \odot t^K$, going from urns to drawn elements. Notice that we use a letter 't' in a name like OtR to denote the *transition* channel $\mathcal{D}(X) \mathrel{\circ\!\!\!\rightarrow} \mathcal{L}(X) \times \mathcal{D}(X)$, for Ordered transitions with Replacement. Similarly, we use the letter 'd' for the associated K-fold *draw* channel $OdR[K]\colon \mathcal{D}(X) \mathrel{\circ\!\!\!\rightarrow} X^K$, in Table 11, where $OdR[K] = \pi_1 \odot OtR^K$. The same convention is used for the other forms of drawing.

3.1 Ordered draws from an urn

We start to look at the upper two 'ordered' transition channels $OtR\colon \mathcal{D}(X) \to \mathcal{D}\big(\mathcal{L}(X) \times \mathcal{D}(X)\big)$ and $OtD\colon \mathcal{M}(X) \to \mathcal{D}\big(\mathcal{L}(X) \times \mathcal{M}(X)\big)$ in (13). Towards a general formula for their iteration, let's look first at the easiest case, namely ordered transitions with replacement. By definition we have as first iteration.

$$OtR^1(\omega) = OtR(\omega) = \sum_{x_1 \in \mathrm{supp}(\omega)} \omega(x_1)\,\big|[x_1], \omega\big\rangle.$$

Accumulation of drawn elements in the first coordinate of $\big|-,-\big\rangle$ starts in the second iteration:

$$\begin{aligned}
\mathit{OtR}^2(\omega) &= \mathit{OtR} \gg \mathit{OtR}(\omega) \\
&= \sum_{\ell\in\mathcal{L}(X),\, x_1\in\mathrm{supp}(\omega)} \omega(x_1)\cdot \mathit{OtR}(\omega)(\ell,\omega)\,\big|\,[x_1] +\!\!+\, \ell,\omega\,\big\rangle \\
&= \sum_{x_1,x_2\in\mathrm{supp}(\omega)} \omega(x_1)\cdot\omega(x_2)\,\big|\,[x_1] +\!\!+\, [x_2],\omega\,\big\rangle \\
&= \sum_{x_1,x_2\in\mathrm{supp}(\omega)} (\omega\otimes\omega)(x_1,x_2)\,\big|\,[x_1,x_2],\omega\,\big\rangle.
\end{aligned}$$

The formula for subsequent iterations is beginning to appear.

Theorem 5. *Consider in* (13) *the ordered-transition-with-replacement channel* $\mathit{OtR}\colon \mathcal{D}(X) \mathrel{\circ\!\!\rightarrow} \mathcal{L}(X)\times\mathcal{D}(X)$*, with distribution* $\omega\in\mathcal{D}(X)$*.*

1. *Iterating* $K\in\mathbb{N}$ *times yields:*
$$\mathit{OtR}^K(\omega) \,=\, \sum_{\vec{x}\in X^K} \omega^K(\vec{x})\,\big|\,\vec{x},\omega\,\big\rangle.$$

2. *The associated* K*-draw channel* $\mathit{OdR}[K] \coloneqq \pi_1 \odot \mathit{OtR}^K : \mathcal{D}(X) \mathrel{\circ\!\!\rightarrow} X^K$ *satisfies*
$$\mathit{OdR}[K](\omega) \,=\, \omega^K \,=\, \mathit{iid}[K](\omega),$$
where iid is the identical and independent channel. □

The situation for ordered transition with deletion is less straightforward. We look at two iterations explicitly, starting from a multiset $\psi\in\mathcal{M}(X)$.

$$\begin{aligned}
\mathit{OtD}^1(\psi) &= \sum_{x_1\in\mathrm{supp}(\psi)} \frac{\psi(x_1)}{\|\psi\|}\,\big|\,x_1, \psi{-}1|\,x_1\,\rangle\,\big\rangle \\
\mathit{OtD}^2(\psi) &= \mathit{OtD} \gg \mathit{OtD}(\psi) \\
&= \sum_{\substack{x_1\in\mathrm{supp}(\psi),\\ x_2\in\mathrm{supp}(\psi-1|\,x_1\,\rangle}} \frac{\psi(x_1)}{\|\psi\|}\cdot\frac{(\psi{-}1|\,x_1\,\rangle)(x_2)}{\|\psi\|-1}\,\big|\,x_1, x_2, \psi{-}1|\,x_1\,\rangle{-}1|\,x_2\,\rangle\,\big\rangle.
\end{aligned}$$

Etcetera. We first collect some basic observations in an auxiliary result.

Lemma 6. *Let* $\psi\in\mathcal{M}[L](X)$ *be a multiset/urn of size* $L = \|\psi\|$*.*

1. *Iterating* $K\leq L$ *times satisfies:*
$$\mathit{OtD}^K(\psi) \,=\, \sum_{\vec{x}\in X^K,\,\mathrm{acc}(\vec{x})\leq\psi}\;\prod_{0\leq i<K} \frac{\big(\psi-\mathrm{acc}(x_1,\ldots,x_i)\big)(x_{i+1})}{L-i}\,\big|\,\vec{x},\psi{-}\mathrm{acc}(\vec{x})\,\big\rangle.$$

2. For $\vec{x} \in X^K$ write $\phi = acc(\vec{x})$. Then:

$$\prod_{0 \leq i < K} \Big(\psi - \mathrm{acc}(x_1, \ldots, x_i)\Big)(x_{i+1}) \;=\; \prod\nolimits_y \frac{\psi(y)!}{(\psi(y) - \phi(y))!} \;=\; \frac{\psi\text{❕}}{(\psi - \phi)\text{❕}}.$$

The right-hand-side is thus independent of the sequence $\vec{x}$.

This independence means that any order of the elements of the same multiset of balls gets the same (draw) probability. This is not entirely trivial.

Proof. 1. Directly from the definition of the transition channel OtD, using Kleisli composition (12).

2. Write $\phi = acc(\vec{x})$ as $\phi = \sum_j n_j | y_j \rangle$. Then each element $y_j \in X$ occurs n_j times in the sequence $\vec{x}$. The product

$$\prod_{0 \leq i < K} \Big(\psi - \mathrm{acc}(x_1, \ldots, x_i)\Big)(x_{i+1})$$

does not depend on the order of the elements in $\vec{x}$: each element y_j occurs n_j times in this product, with multiplicities $\psi(y_j), \ldots, \psi(y_j) - n_j + 1$, independently of the exact occurrences of the y_j in $\vec{x}$. Thus:

$$\begin{aligned}
\prod_{0 \leq i < K} \Big(\psi - \mathrm{acc}(x_1, \ldots, x_i)\Big)(x_{i+1}) \;&=\; \prod\nolimits_j \psi(y_j) \cdot \ldots \cdot (\psi(y_j) - n_j + 1) \\
&=\; \prod\nolimits_j \psi(y_j) \cdot \ldots \cdot (\psi(y_j) - \phi(y_j) + 1) \\
&=\; \prod\nolimits_j \frac{\psi(y_j)!}{(\psi(y_j) - \phi(y_j))!} \\
&=\; \prod_{y \in X} \frac{\psi(y)!}{(\psi(y) - \phi(y))!}.
\end{aligned}$$

We can extend the product over j to a product over all $y \in X$ since if $y \notin \mathrm{supp}(\phi)$, then, even if $\psi(y) = 0$,

$$\frac{\psi(y)!}{(\psi(y) - \phi(y))!} \;=\; \frac{\psi(y)!}{\psi(y)!} \;=\; 1. \qquad \square$$

Theorem 7. *Consider the ordered-transition-with-deletion channel OtD on $\psi \in \mathcal{M}[L](X)$.*

1. For $K \leq L$,

$$OtD^K(\psi) \;=\; \sum_{\phi \leq_K \psi} \; \sum_{\vec{x} \in \mathrm{acc}^{-1}(\phi)} \frac{\boldsymbol{(}\psi - \phi\boldsymbol{)}}{\boldsymbol{(}\psi\boldsymbol{)}} \, \big| \vec{x}, \psi - \phi \big\rangle.$$

2. *The associated K-draw channel $OdD[K] \coloneqq \pi_1 \circ OtD^K \colon \mathcal{M}[L](X) \rightsquigarrow X^K$ satisfies:*

$$\begin{aligned} OdD[K](\psi) &= \sum_{\phi \leq_K \psi} \sum_{\vec{x} \in \mathrm{acc}^{-1}(\phi)} \frac{\boldsymbol{(}\psi - \phi\boldsymbol{)}}{\boldsymbol{(}\psi\boldsymbol{)}} \big|\vec{x}\big\rangle \\ &= \sum_{\vec{x} \in X^K,\, \mathrm{acc}(\vec{x}) \leq \psi} \frac{\boldsymbol{(}\psi - \mathrm{acc}(\vec{x})\boldsymbol{)}}{\boldsymbol{(}\psi\boldsymbol{)}} \big|\vec{x}\big\rangle. \end{aligned}$$

As mentioned in the beginning of this section, the latter ordered-draw-deletion distribution does not have its own name.

Proof. 1. By combining the two points of Lemma 6 and using:

$$\prod_{0 \leq i < K} (L - i) = L \cdot (L-1) \cdot \ldots \cdot (L - K + 1) = \frac{L!}{(L-K)!},$$

we get: we get:

$$\begin{aligned} OtD^K(\psi) &= \sum_{\varphi \leq_K \psi} \sum_{\vec{x} \in \mathrm{acc}^{-1}(\varphi)} \frac{(L-K)!}{L!} \cdot \prod\nolimits_y \frac{\psi(y)!}{(\psi(y)-\varphi(y))!} \big|\vec{x}, \psi-\varphi\big\rangle \\ &= \sum_{\varphi \leq_K \psi} \sum_{\vec{x} \in \mathrm{acc}^{-1}(\varphi)} \frac{(L-K)!}{\prod_y (\psi(y)-\varphi(y))!} \cdot \frac{\prod_y \psi(y)!}{L!} \big|\vec{x}, \psi-\varphi\big\rangle \\ &= \sum_{\varphi \leq_K \psi} \sum_{\vec{x} \in \mathrm{acc}^{-1}(\varphi)} \frac{(L-K)!}{(\psi-\varphi)\mathbf{!}} \cdot \frac{\psi\mathbf{!}}{L!} \big|\vec{x}, \psi-\varphi\big\rangle \\ &= \sum_{\varphi \leq_K \psi} \sum_{\vec{x} \in \mathrm{acc}^{-1}(\varphi)} \frac{\boldsymbol{(}\psi - \varphi\boldsymbol{)}}{\boldsymbol{(}\psi\boldsymbol{)}} \big|\vec{x}, \psi-\varphi\big\rangle. \end{aligned}$$

2. Directly by the previous point. □

3.2 Unordered draws from an urn

We now concentrate on the transition channels $UtR \colon \mathcal{D}(X) \rightsquigarrow \mathcal{M}(X) \times \mathcal{D}(X)$ and $UtD \colon \mathcal{M}(X) \rightsquigarrow \mathcal{M}(X) \times \mathcal{M}(X)$ in (13), for unordered draws. Notice that we are now using $M = \mathcal{M}(X)$ as commutative monoid in the setting of Lemma 4. We immediately formulate a characterisation of iteration. We immediately recognise the resemblance with multinomial and hypergeometric distributions.

Lemma 8. *1. For $\omega \in \mathcal{D}(X)$ and $K \in \mathbb{N}$,*

$$UtR^K(\omega) = \sum_{\phi \in \mathcal{M}[K](X)} \boldsymbol{(}\phi\boldsymbol{)} \cdot \prod\nolimits_x \omega(x)^{\phi(x)} \big|\phi, \omega\big\rangle.$$

2. For $\psi \in \mathcal{M}[L{+}K](X)$,

$$UtD^K(\psi) = \sum_{\phi \leq_K \psi} \frac{\prod_x \binom{\psi(x)}{\phi(x)}}{\binom{L+K}{K}} \big|\phi, \psi - \phi\big\rangle = \sum_{\phi \leq_K \psi} \frac{\binom{\psi}{\phi}}{\binom{L+K}{K}} \big|\phi, \psi - \phi\big\rangle.$$

This result shows how the relatively complicated expressions with binomial coefficients $\binom{x}{y}$ in the multinomial and hypergeometric distributions arise from the structure of the monad in Lemma 4.

Proof. 1. We use induction on $K \in \mathbb{N}$. For $K = 0$ we have $\mathcal{M}[K](X) = \{\mathbf{0}\}$ and so:

$$UtR^0(\omega) \,=\, \eta(\omega) \,=\, 1\big|\mathbf{0}, \omega\big\rangle \,=\, \sum_{\phi\in\mathcal{M}[0](X)} \boldsymbol{(}\phi\boldsymbol{)} \cdot \prod\nolimits_x \omega(x)^{\phi(x)} \,\big|\phi, \omega\big\rangle.$$

For the induction step:

$$\begin{array}{rl}
& UtR^{K+1}(\omega) \\
= & \big(UtR^K \odot UtR\big)(\omega) \\
\overset{(12)}{=} & \displaystyle\sum_{\psi\in\mathcal{M}[1](X),\, \phi\in\mathcal{M}[K](X)} UtR^K(\omega)(\phi,\omega) \cdot UtR(\omega)(\psi,\omega)\,\big|\psi{+}\phi, \omega\big\rangle \\
\overset{(\mathrm{IH})}{=} & \displaystyle\sum_{y\in X,\, \phi\in\mathcal{M}[K](X)} \boldsymbol{(}\phi\boldsymbol{)} \cdot \Big(\prod\nolimits_x \omega(x)^{\phi(x)}\Big) \cdot \omega(y)\,\big|1|y\rangle{+}\phi, \omega\big\rangle \\
= & \displaystyle\sum_{\psi\in\mathcal{M}[K+1](X)} \Big(\sum\nolimits_y \boldsymbol{(}\psi - 1|y\rangle\boldsymbol{)}\Big) \cdot \prod\nolimits_x \omega(x)^{\psi(x)}\,\big|\psi, \omega\big\rangle \\
\overset{(1)}{=} & \displaystyle\sum_{\psi\in\mathcal{M}[K+1](X)} \boldsymbol{(}\psi\boldsymbol{)} \cdot \prod\nolimits_x \omega(x)^{\psi(x)}\,\big|\psi, \omega\big\rangle.
\end{array}$$

2. For $K = 0$ both sides are equal to the empty multiset $\mathbf{0}$. Next, for a multiset $\psi \in \mathcal{M}[L{+}K{+}1](X)$ we have:

$$\begin{array}{rl}
& UtD^{K+1}(\psi) \\
= & \big(UtD^K \odot UtD\big)(\psi) \\
\overset{(12)}{=} & \displaystyle\sum_{\substack{y\in\mathrm{supp}(\psi),\, \chi\in\mathcal{M}[L](X), \\ \phi\leq_K \psi-1|\,y\,\rangle}} UtD^K(\psi - 1|y\rangle)(\phi,\chi) \cdot \frac{\psi(y)}{L{+}K{+}1}\,\big|\phi + 1|y\rangle, \chi\big\rangle \\
\overset{(\mathrm{IH})}{=} & \displaystyle\sum_{\substack{y\in\mathrm{supp}(\psi), \\ \phi\leq_K \psi-1|\,y\,\rangle}} \frac{\binom{\psi-1|\,y\,\rangle}{\phi}}{\binom{L+K}{K}} \cdot \frac{\psi(y)}{L{+}K{+}1}\,\big|\phi + 1|y\rangle, \psi - 1|y\rangle - \phi\big\rangle \\
\overset{(*)}{=} & \displaystyle\sum_{\substack{y\in\mathrm{supp}(\psi), \\ \phi\leq_K \psi-1|\,y\,\rangle}} \frac{\phi(y) + 1}{K{+}1} \cdot \frac{\binom{\psi}{\phi+1|\,y\,\rangle}}{\binom{L+K+1}{K+1}}\,\big|\phi + 1|y\rangle, \psi - (\phi + 1|y\rangle)\big\rangle \\
= & \displaystyle\sum_{\chi\leq_{K+1}\psi,\, y} \frac{\chi(y)}{K{+}1} \cdot \frac{\binom{\psi}{\chi}}{\binom{L+K+1}{K+1}}\,\big|\chi, \psi - \chi\big\rangle \\
= & \displaystyle\sum_{\chi\leq_{K+1}\psi} \frac{\binom{\psi}{\chi}}{\binom{L+K+1}{K+1}}\,\big|\chi, \psi - \chi\big\rangle.
\end{array}$$

The equation marked (*) holds, firstly because:

$$(n+1) \cdot \binom{n}{m} \;=\; (m+1) \cdot \binom{n+1}{m+1},$$

and thus:

$$\psi(y) \cdot \binom{\psi - 1|y\rangle}{\phi} \;=\; (\phi(y)+1) \cdot \binom{\psi}{\phi + 1|y\rangle}. \qquad \square$$

We are now in a position to describe the multinomial and hypergeometric distributions (6) using iterations of the *UtR* and *UtD* maps.

Theorem 9. *1. The K-draw multinomial is the first marginal of the K-iteration of the unordered-with-replacement transition:*

$$\mathit{mulnom}[K] \;=\; \pi_1 \circledcirc \mathit{UtR}^K \;=:\; \mathit{UdR}[K].$$

2. Similarly the hypergeometric distribution arises from iterated unordered-with-deletion:

$$\mathit{hypgeom}[K] \;=\; \pi_1 \circledcirc \mathit{UtD}^K \;=:\; \mathit{UdD}[K].$$

Proof. Directly by Lemma 8, see the definitions of multinomial and hypergeometric distribution in (6). □

Theorems 5, 7 and 9 provide a principled account of the four drawing operations in Table 11. This concludes the first part of this paper, on drawing balls from urns.

4 Intermezzo on predicates, validity and conditioning

This section recalls the basic constructions associated with (fuzzy) predicates. Predicates play a role as *evidence*, notably in updating and learning.

4.1 Predicates

In the current setting of discrete probability we define a predicate on an arbitrary set X to be a function $p\colon X \to [0,1]$. Thus, predicates are fuzzy, taking values in the unit interval $[0,1]$. Such a predicate is called *sharp* if it restricts to $X \to \{0,1\}$, that is, if 0 and 1 are the only possible outcomes. Sharp predicates can be identified with subsets of X and are often called *events*. In general, for a subset $U \subseteq X$ we write $\mathbf{1}_U\colon X \to \{0,1\}$ for the sharp predicate with $\mathbf{1}_U(x) = 1$ iff $x \in U$. We simply write $\mathbf{1}_x$ for $\mathbf{1}_{\{x\}}$ and call $\mathbf{1}_x$ a point predicate.

The set $\mathrm{Pred}(X) \coloneqq [0,1]^X$ of predicates on X carries a pointwise order. We shall write $\mathbf{0} = \mathbf{1}_\emptyset$ and $\mathbf{1} = \mathbf{1}_X$ for the least and greatest predicates (falsum and truth), with $\mathbf{0}(x) = 0$ and $\mathbf{1}(x) = 1$ for each $x \in X$. Predicates form a commutative monoid via truth $\mathbf{1}$ and conjunction &, where $(p \;\&\; q)(x) = p(x) \cdot q(x)$ involves pointwise multiplication. We have $\mathbf{1}_U \;\&\; \mathbf{1}_V = \mathbf{1}_{U \cap V}$ and thus in particular $\mathbf{1}_U \mathbin{\&} \mathbf{1}_U = \mathbf{1}_U$. However, for properly fuzzy (*i.e.* non-sharp) predicates p one has $p \mathbin{\&} p \neq p$.

There is also scalar multiplication $r \cdot p$, for $r \in [0,1]$, with $(r \cdot p)(x) = r \cdot p(x)$, and orthocomplement (negation) $p^\perp$ with $p^\perp(x) = 1 - p(x)$. Then: $p^{\perp\perp} = p$ and $\mathbf{1}^\perp = \mathbf{0}$, so that $\mathbf{0}^\perp = \mathbf{1}$. In addition there is a partial sum operation written as $\ovee$. For predicates $p, q \in \mathrm{Pred}(X)$ with $p(x) + q(x) \leq 1$ for all x, one has $p \ovee q \in \mathrm{Pred}(X)$ given by $(p \ovee q)(x) = p(x) + q(x)$. Then, for instance, $p \ovee p^\perp = \mathbf{1}$ and $\mathbf{1}_U \ovee \mathbf{1}_V = \mathbf{1}_{U \cup V}$ if U, V are disjoint subsets. Thus, on a finite set X one can write a predicate $p \in \mathrm{Pred}(X)$ in a normal form as $p = \bigovee_x p(x) \cdot \mathbf{1}_x$. All this structure makes the set of predicates $\mathrm{Pred}(X)$ an effect module $(\mathbf{0}, \ovee, (-)^\perp)$ with a commutative (non-idempotent) monoid structure $(\mathbf{1}, \&)$, see *e.g.* [21,22] for more details.

4.2 Validity and conditioning

For a state $\omega \in \mathcal{D}(X)$ and a predicate $p \in \mathrm{Pred}(X)$ on the same set X we write $\omega \models p$ for the *validity* of p in ω. It is defined as the expected value:

$$\omega \models p \;\coloneqq\; \sum_{x \in X} \omega(x) \cdot p(x).$$

Then, for instance, $\omega \models \mathbf{1} = 1$ and $\omega \models \mathbf{0} = 0$.

When this validity $\omega \models p$ is non-zero, we can *update* or *condition* the state $\omega \in \mathcal{D}(X)$ to a new state $\omega|_p \in \mathcal{D}(X)$, in the light of the evidence p. This $\omega|_p$ is a normalised inner product:

$$\omega|_p \;\coloneqq\; \sum_{x \in X} \frac{\omega(x) \cdot p(x)}{\omega \models p} \big| x \big\rangle.$$

We shall use the following basic properties, see [21,22,24] for more details.

Lemma 10. *Assuming the relevant validities are non-zero, one has:*

1. $\omega|_{\mathbf{1}} = \omega$ *and* $\omega|_p|_q = \omega|_{p \& q}$*;*
2. *Bayes' rule holds:*

$$\omega|_p \models q \;=\; \frac{\omega \models p \mathbin{\&} q}{\omega \models p} \;=\; \frac{(\omega|_q \models p) \cdot (\omega \models q)}{\omega \models p}. \qquad \square$$

A consequence of the first point is that the order of conditioning is irrelevant: $\omega|_p|_q = \omega|_{p\&q} = \omega|_{q\&p} = \omega|_q|_p$. It is for this reason that data, as the material to learn from, are best organised as multisets — where, recall, the order of elements is irrelevant, but not there multiplicity.

Remark 11. There are two points we like to make about conditioning and drawing.

1. One obvious thought is to try and describe a draw from an urn via conditioning. What would this mean? If the urn is a multiset $\psi \in \mathcal{M}(X)$ we can turn it into a distribution $\mathit{Flrn}(\psi) \in \mathcal{D}(X)$ via frequentist learning. Then the thought can be reformulated as a question: can we write:

$$\mathit{Flrn}(\psi - 1|x\rangle) \;=\; \mathit{Flrn}(\psi)\big|_p$$

for a suitable predicate p, depending on the element x that is drawn from the urn ψ?

This does not work, as we will illustrate. Take $X = \{a, b\}$ and $\psi = 3|a\rangle + 2|b\rangle$. Then:

$$\mathit{Flrn}(\psi - 1|b\rangle) \;=\; \mathit{Flrn}(3|a\rangle + 1|b\rangle) \;=\; \tfrac{3}{4}|a\rangle + \tfrac{1}{4}|b\rangle.$$

Now assume that $p\colon \{a, b\} \to [0, 1]$ satisfies:

$$\mathit{Flrn}(\psi)|_p \;=\; \mathit{Flrn}(\psi - 1|b\rangle) \;=\; \tfrac{3}{4}|a\rangle + \tfrac{1}{4}|b\rangle.$$

This would mean:

$$\frac{{}^3\!/\!_5 \cdot p(a)}{{}^3\!/\!_5 \cdot p(a) + {}^2\!/\!_5 \cdot p(b)} \;=\; \tfrac{3}{4} \qquad\text{and}\qquad \frac{{}^1\!/\!_5 \cdot p(b)}{{}^3\!/\!_5 \cdot p(a) + {}^2\!/\!_5 \cdot p(b)} \;=\; \tfrac{1}{4}.$$

This gives two equations $12 \cdot p(a) = 9 \cdot p(a) + 6 \cdot p(b)$ and $4 \cdot p(b) = 3 \cdot p(a) + 2 \cdot p(b)$. The only solution is $p(a) = p(b) = 0$, so that $p = \mathbf{0}$. But conditioning with falsum $\mathbf{0}$ is not well-defined, since it involves a zero validity.

2. The probabilities in the ordered/unordered transitions with deletion in (13) can be described in terms of frequentist learning *Flrn*, as in:

$$\begin{aligned} \mathit{OtD}(\psi) \;&=\; \textstyle\sum_x \mathit{Flrn}(\psi)(x)\big|\, [x], \psi - 1|x\rangle \,\big\rangle \\ \mathit{UtD}(\psi) \;&=\; \textstyle\sum_x \mathit{Flrn}(\psi)(x)\big|\, 1|x\rangle, \psi - 1|x\rangle \,\big\rangle. \end{aligned}$$

where ψ is a (non-empty) multiset over X. In case we have a predicate p on X, we can use it to describe a 'biased' (or 'non-central') draw by updating this distribution $\mathit{Flrn}(\psi)$ with p in the above expressions, as in:

$$\begin{aligned} \mathit{OtD}_p(\psi) \;&=\; \textstyle\sum_x \mathit{Flrn}(\psi)\big|_p(x)\big|\, [x], \psi - 1|x\rangle \,\big\rangle \\ \mathit{UtD}_p(\psi) \;&=\; \textstyle\sum_x \mathit{Flrn}(\psi)\big|_p(x)\big|\, 1|x\rangle, \psi \,\big\rangle. \end{aligned}$$

Taking such a bias into account may be useful, for instance in a poll-by-phone, where working people may be under-represented (since they are less often at home).

4.3 Predicate transformation

We have seen in Subsection 2.2 that a channel $c\colon X \mathrel{\circ\!\!\to} Y$, that is, a function $c\colon X \to \mathcal{D}(Y)$, gives rise to a state transformation function $c \gg (-)\colon \mathcal{D}(X) \to \mathcal{D}(Y)$. It pushes a state forward. One can also pull a predicate backward, along a channel. This is done via a predicate transformation function $\mathrm{Pred}(Y) \to \mathrm{Pred}(X)$, acting in the opposite direction. On $q \in \mathrm{Pred}(Y) = [0,1]^Y$ it is written as $c \ll q \in \mathrm{Pred}(X) = [0,1]^X$, defined by:

$$\big(c \ll q\big)(x) \;\coloneqq\; \sum_{y\in Y} c(x)(y)\cdot q(y).$$

It is not hard to see that predicate transformation preserves $\mathbf{0}, \mathbf{1}, \ovee, (-)^{\perp}$ and scalar multiplication $r \cdot (-)$. However, it does *not* preserve conjunction &. Predicate transformation is functorial, in the sense that $\mathrm{id} \ll q = q$ and $(d \odot c) \ll q = c \ll (d \ll q)$.

State transformation $\gg$, predicate transformation $\ll$, and validity $\models$ are connected via the following fundamental relationship:

$$c \gg \omega \models q \;=\; \omega \models c \ll q. \tag{14}$$

A function $f\colon X \to Y$ is often implicitly promoted to a channel $f\colon X \mathrel{\circ\!\!\to} Y$ via $x \mapsto 1|\,f(x)\,\rangle$. Then $f \ll q$ is simply $q \circ f$. Predicate transformation $\pi_i \ll q$ along a projection is *weakening*, that is moving a predicate to a bigger context. For $p_i \in \mathrm{Pred}(X_i)$ we define parallel conjunction $p_1 \otimes p_2 \in \mathrm{Pred}(X_1 \times X_2)$ as:

$$p_1 \otimes p_2 \;\coloneqq\; (\pi_1 \ll p_1) \mathbin{\&} (\pi_2 \ll p_2).$$

Thus, $(p_1 \otimes p_2)(x,y) = p_1(x)\cdot p_2(y)$. Hence weakening can also be expressed as parallel conjunction with truth: $\pi_1 \ll p = p \otimes \mathbf{1}$ and $\pi_2 \ll q = \mathbf{1} \otimes q$. Also, $q_1 \mathbin{\&} q_2 = \Delta \ll (q_1 \otimes q_2)$. Further, $(\omega_1 \otimes \omega_2) \models (q_1 \otimes q_2) = (\omega_1 \models q_1)\cdot(\omega_2 \models q_2)$.

4.4 Daggers of channels

Let $c\colon X \mathrel{\circ\!\!\to} Y$ be a channel and $\omega \in \mathcal{D}(X)$ be a state on its domain. Under a certain side-condition (see below), one can turn this channel around to obtain a 'dagger' channel $c^{\dagger}_{\omega}\colon Y \mathrel{\circ\!\!\to} X$ in the other direction. This new channel is defined via conditioning, as:

$$c^{\dagger}_{\omega}(y) \coloneqq \omega|_{c \ll \mathbf{1}_y} = \textstyle\sum_x \dfrac{\omega(x)\cdot c(x)(y)}{(c \gg \omega)(y)}\big|\,x\,\big\rangle. \tag{15}$$

This formulation reveals the side-condition for existence of the dagger: the state $c \gg \omega$ must have full support.

Almost by construction one has:

$$c^{\dagger}_{\omega} \gg (c \gg \omega) = \omega \qquad \text{and} \qquad \big(c^{\dagger}_{\omega}\big)^{\dagger}_{c \gg \omega} = c. \tag{16}$$

This dagger is the probabilistic analogue of the conjugate transpose of a bounded map between Hilbert spaces. It can be shown that this dagger is functorial, when channels are organised in a suitable category with states as objects, see [15] and [14,19] for more details. This dagger channel is called Bayesian inversion in [15] and is related to learning. It will show up later on in Proposition 23, for learning along a channel.

4.5 Learning as likelihood increase

Consider a validity expression:

$$\underset{\text{distribution / state}}{\omega} \models \underset{\text{predicate / evidence}}{p} \tag{17}$$

One form of learning involves increasing this validity, by changing the state ω into a new state ω' such that $\omega' \models p \;\geq\; \omega \models p$. Thus, in this validity-based learning one takes the evidence p as a given, fixed datum that one needs to adjust to. Learning happens by changing the state ω so that it better fits the evidence. Informally, this is learning by increasing what's right, in contrast to learning by decreasing what's wrong.

The above validity expression $\omega \models p$ in 17 may be reorganised as a function $\mathit{val}\colon \mathrm{Pred}(X) \to \mathrm{Pred}\big(\mathcal{D}(X)\big)$, namely $\mathit{val}(p)(\omega) := (\omega \models p)$. It then becomes an instance of a *likelihood* function $\mathcal{L}$, which is typically of the form:

$$\mathit{Data} \xrightarrow{\quad\mathcal{L}\quad} \mathrm{Pred}\big(\mathcal{D}(X)\big) = [0,1]^{\mathcal{D}(X)}.$$

The predicate $\mathcal{L}(d)\colon \mathcal{D}(X) \to [0,1]$ sends a state $\omega \in \mathcal{D}(X)$ to the likelihood of the data d in that state. The idea in learning is to find a maximum for $\mathcal{L}(d)$, that is, to find the state that makes the data most likely. In (17) we use a single predicate as data. Below we shall generalise this to multisets of predicates. This corresponds to the idea that data may come in volumes of separable, possibly identical units, where the order does not matter. The predicates that we use as data/evidence may be point predicates, corresponding to data points.

One important way to learn in the above situation is to update (condition) the state ω with the evidence p, as introduced in Subsection 4.2.

Proposition 12. *There is an inequality:*

$$\omega|_p \models p \;\geq\; \omega \models p.$$

This result captures an important intuition behind conditioning with p: it changes the state so that evidence p becomes 'more true'.

Proof. We first show that it suffices to prove an inequality:

$$(\omega \models p^2) \;\geq\; (\omega \models p)^2, \tag{*}$$

where $p^2 = p \mathbin{\&} p$. Indeed, with (*) we are done, since by Bayes' rule (Lemma 10),

$$\omega_p \models p \;=\; \frac{\omega \models p \mathbin{\&} p}{\omega \models p} \;\geq\; \frac{\big(\omega \models p\big)^2}{\omega \models p} \;=\; \omega \models p.$$

In order to prove the inequality (*) we use the standard notion of variance $\mathit{Var}(\omega, p)$ of predicate p in state ω as validity:

$$\mathit{Var}(\omega, p) \;\coloneqq\; \omega \models \big(p - (\omega \models p) \cdot \mathbf{1}\big)^2.$$

This number is non-negative since the predicate on right-hand-side of $\models$ is defined as square: $x \mapsto (p(x) - (\omega \models p))^2$. The inequality (*) follows from the (also standard) equation:

$$\big(\omega \models p^2\big) - (\omega \models p)^2 \;=\; \mathit{Var}(\omega, p) \;\geq\; 0.$$

We show how this equation is obtained in the current setting:

$$\begin{array}{l}
\mathit{Var}(\omega, p) \\
= \omega \models \big(p - (\omega \models p) \cdot \mathbf{1}\big)^2 \\
= \sum_x \omega(x)\big(p(x) - (\omega \models p)\big)^2 \\
= \sum_x \omega(x)\Big(p(x)^2 - 2(\omega \models p)p(x) + (\omega \models p)^2\Big) \\
= \Big(\sum_x \omega(x)p^2(x)\Big) - 2(\omega \models p)\Big(\sum_x \omega(x)p(x)\Big) + \Big(\sum_x \omega(x)(\omega \models p)^2\Big) \\
= \big(\omega \models p^2\big) - 2\big(\omega \models p)(\omega \models p) + (\omega \models p)^2 \\
= \big(\omega \models p^2\big) - (\omega \models p)^2.
\end{array}$$

□

5 Evaluating instead of drawing

The two main topics of this paper are drawing from an urn and learning. This section glues these two topics together. It replaces drawing a ball from an urn, as analysed in Section 3, by evaluating a predicate via validity $\models$. Both drawing and evaluating are seen as experiments that assign probabilities to (multiple) draws/evaluations.

Our starting point is a single transition step, as used for drawing one ball from an urn, via a channel of the form $\mathcal{D}(X) \to \mathcal{D}(M \times \mathcal{D}(X))$. It can be iterated via the combined monad $\mathcal{D}(M \times -)$ from Lemma 4. Now that we wish to use predicates for making observations, we need to decide what M is in this situation. For convenience we concentrate on the unordered case, and ignore ordered scenarios.

We fix a set $P = \{p_1, \ldots, p_n\}$ of predicates and take $M = \mathcal{M}(P)$ as monoid, containing multisets $\sum_i n_i | p_i \rangle$ of predicates. We require that the predicates in P form a *test*, that is: $p_1 \ovee \cdots \ovee p_n = \mathbf{1}$. If needed, we can always force such predicates to be a test, by switching to $p'_i = \frac{p_i}{p}$, where $p = \sum_i p_i$.

The scenarios that we consider are denoted as unordered-transition-update (UtU) and unordered-transition-continue (UtC). They are described by the following two transition maps.

$$\begin{array}{rcl} \mathcal{D}(X) \xrightarrow{UtC} \mathcal{D}\big(\mathcal{M}(P) \times \mathcal{D}(X)\big) & \qquad & \mathcal{D}(X) \xrightarrow{UtU} \mathcal{D}\big(\mathcal{M}(P) \times \mathcal{D}(X)\big) \\ \omega \longmapsto \sum_i (\omega \models p_i) \big| 1|p_i\rangle, \omega \big\rangle & & \omega \longmapsto \sum_i (\omega \models p_i) \big| 1|p_i\rangle, \omega/p_i \big\rangle \end{array} \tag{18}$$

We see that in the 'continue' case UtC the state ω remains the same, whereas in the 'update' case UtU it is updated with each occurring predicate p_i. In this description we ignore undefinedness of updating, when validities are zero.

The approach of Section 3 involves iterating (similar) transitions maps and then taking the first projection, for marginalisation. This is what we shall do here as well — without once again elaborating all the details.

Lemma 13. *Fix a state $\omega \in \mathcal{D}(X)$ with set of predicates $P = \{p_1, \ldots, p_n\}$ on X, forming a test. Then for $K \in \mathbb{N}$,*

1. *By iterating the 'continue' map in* (18) *one gets:*
$$\big(\pi_1 \circ UtC^K\big)(\omega) \;=\; \sum_{\phi \in \mathcal{M}[K](P)} \boldsymbol{(}\phi\boldsymbol{)} \cdot \prod_i (\omega \models p_i)^{\phi(i)} \,\big|\phi\big\rangle.$$

2. *The 'update' case in* (18) *yields:*
$$\big(\pi_1 \circ UtU^K\big)(\omega) \;=\; \sum_{\phi \in \mathcal{M}[K](P)} \boldsymbol{(}\phi\boldsymbol{)} \cdot \big(\omega \models \&_i \; p_i^{\phi(i)}\big) \,\big|\phi\big\rangle.$$

Proof. 1. This works very much like in Lemma 8 (1).

2. The fact that the state ω is updated in the UtU transition in (18) introduces new dynamics, where Bayes' rule, see Lemma 10 (2), starts to play a role. For instance, after two steps we get:
$$\begin{aligned} \big(\pi_1 \circ UtU^2\big)(\omega) &= \sum_{i,j} (\omega|_{p_i} \models p_j) \cdot (\omega \models p_i) \,\big| 1|p_i\rangle + 1|p_j\rangle \big\rangle \\ &= \sum_{i,j} (\omega \models p_i \,\&\, p_j) \,\big| 1|p_i\rangle + 1|p_j\rangle \big\rangle. \end{aligned}$$

Continuing yields the above formula in point (2). □

One interesting thing to note is that different validity expressions arise for a multiset ϕ, namely $\prod_i (\omega \models p_i)^{\phi(i)}$ with a product on the outside, and $\omega \models \&_i \; p_i^{\phi(i)}$ with a product (conjunction) on the inside. This difference will be explored further in the remainder of this article.

Also it is noteworthy that the probabilities in the two formulas in Lemma (13) add up to one because the predicates p_i form a test. For point (1) we use the Multinomial Theorem (Lemma 3) in the obvious way:

$$\begin{aligned}\sum_{\phi\in\mathcal{M}[K](P)} \boldsymbol{(}\phi\boldsymbol{)}\cdot\prod\nolimits_i(\omega\models p_i)^{\phi(i)} &= \big(\sum\nolimits_i \omega\models p_i\big)^K \\ &= \big(\omega\models \textstyle\bigovee_i p_i\big)^K = \big(\omega\models \mathbf{1}\big)^K = 1^K = 1.\end{aligned}$$

Again by the Multinomial Theorem, but now in slightly different form, the probabilities in point (2) add up to one:

$$\begin{aligned}\sum_{\phi\in\mathcal{M}[K](P)} \boldsymbol{(}\phi\boldsymbol{)}\cdot\big(\omega\models \&_i\ p_i^{\phi(i)}\big) &= \omega\models \bigovee_{\phi\in\mathcal{M}[K](P)} \boldsymbol{(}\phi\boldsymbol{)}\cdot\big(\&_i\ p_i^{\phi(i)}\big) \\ &= \omega\models \big(\textstyle\bigovee_i p_i\big)^K = \omega\models \mathbf{1}^K = \omega\models\mathbf{1} = 1.\end{aligned}$$

The two expressions $\prod_i(\omega\models p_i)^{\phi(i)}$ and $\omega\models \&_i\, p_i^{\phi(i)}$ give, in general, different outcomes — even though when multiplied with $\boldsymbol{(}\phi\boldsymbol{)}$ and summed they both add up to one. This difference will be illustrated next.

Example 14. Let's consider a political party that has to decide on its future policies. We simplify these options to left (L), centre (C), and right (R) in a space $X = \{L, C, R\}$. The party leadership leans to the right. Its position is captured by the following distribution, giving a convex combination of the three directions.

$$\omega \coloneqq \tfrac{1}{5}|L\rangle + \tfrac{3}{10}|C\rangle + \tfrac{1}{2}|R\rangle.$$

The party has four factions; their positions on the three options L, C, R are expressed via the following percentages.

	faction 1	**faction 2**	**faction 3**	**faction 4**
left	30%	10%	50%	10%
centre	20%	30%	20%	30%
right	30%	20%	30%	20%

We can read these columns as four predicates p_1, p_2, p_3, p_4 on the space X, where $p_4 = p_2$. Explicitly:

$$\begin{aligned} p_1 &= \tfrac{3}{10}\cdot\mathbf{1}_L + \tfrac{2}{10}\cdot\mathbf{1}_L + \tfrac{3}{10}\cdot\mathbf{1}_L \\ p_2 &= \tfrac{1}{10}\cdot\mathbf{1}_L + \tfrac{3}{10}\cdot\mathbf{1}_L + \tfrac{2}{10}\cdot\mathbf{1}_L \\ p_3 &= \tfrac{5}{10}\cdot\mathbf{1}_L + \tfrac{2}{10}\cdot\mathbf{1}_L + \tfrac{3}{10}\cdot\mathbf{1}_L. \end{aligned}$$

The four predicates of the table form a test: $p_1 \ovee p_2 \ovee p_3 \ovee p_4 = \mathbf{1}$. The table can be described as a multiset $\phi = 1|p_1\rangle + 2|p_2\rangle + 1|p_3\rangle$ of predicates.

The validities $\omega \models p_i$ can be interpreted as the level of support for the leadership's position within the corresponding faction. It is not hard to see that:

$$\omega\models p_1 = \tfrac{27}{100} \qquad \omega\models p_2 = \tfrac{21}{100} \qquad \omega\models p_3 = \tfrac{31}{100}. \tag{19}$$

How should the party's leadership measure the total support for its position ω within all factions?

1. In one scenario, four secretaries of the leadership visit the four factions separately and collect their separate support values (19). The total support can then be computed as product of the individual support values:

$$(\omega \models p_1) \cdot (\omega \models p_2)^2 \cdot (\omega \models p_3) \;=\; \tfrac{27 \cdot 21^2 \cdot 31}{10^8} \;=\; \tfrac{369.117}{10^8} \;\approx\; 0.0037.$$

   This involves computing $\prod_i (\omega \models p_i)^{\phi(i)}$, as in Lemma 13 (1).

2. Alternatively the party may hold a congress where each faction expresses its position, as percentages in the above table, in order. A mathematical savvy secretary of the leadership may then quickly start computing, after hearing the intermediary results. After announcement of the first faction's position, in the form of predicate p_1, this secretary calculates the validity $\omega \models p_1 = \frac{27}{100}$ and the updated distribution:

$$\begin{aligned} \omega|_{p_1} \;&=\; \frac{1/5 \cdot 3/10}{27/100} |\,L\,\rangle + \frac{3/10 \cdot 2/10}{27/100} |\,C\,\rangle + \frac{1/2 \cdot 3/10}{27/100} |\,R\,\rangle \\ &=\; \tfrac{2}{9} |\,L\,\rangle + \tfrac{2}{19} |\,C\,\rangle + \tfrac{5}{9} |\,R\,\rangle. \end{aligned}$$

   Next, after faction 2 announces its percentages p_2 the secretary computes the validity in the latest state, and also the next update:

$$\omega|_{p_1} \models p_2 \;=\; \tfrac{1}{5} \qquad\qquad \omega|_{p_1}|_{p_2} \;=\; \omega|_{p_1 \& p_2} \;=\; \tfrac{1}{9} |\,L\,\rangle + \tfrac{1}{3} |\,C\,\rangle + \tfrac{5}{9} |\,R\,\rangle.$$

   Continuing like this we get:

$$\omega|_{p_1}|_{p_2} \models p_3 \;=\; \tfrac{13}{45} \qquad\qquad \omega|_{p_1}|_{p_2}|_{p_3} \;=\; \tfrac{5}{26} |\,L\,\rangle + \tfrac{3}{13} |\,C\,\rangle + \tfrac{15}{26} |\,R\,\rangle.$$

   After hearing the percentages of faction 4 the secretary calculates the remaining validity $\omega|_{p_1}|_{p_2}|_{p_3} \models p_2 = \frac{53}{260}$, and then also the product of all these validities:

$$\begin{aligned} &(\omega \models p_1) \cdot (\omega|_{p_1} \models p_2) \cdot (\omega|_{p_1}|_{p_2} \models p_3) \cdot (\omega|_{p_1}|_{p_2}|_{p_3} \models p_2) \\ &=\; \tfrac{18.603}{5.850.000} \;\approx\; 0.0032. \end{aligned}$$

   We see that this second calculation shows slightly less support.
   Finally, via Bayes' rule we may also compute this second outcome as validity $\omega \models p_1 \;\&\; p_2 \;\&\; p_3 \;\&\; p_2 = \omega \models \&_i \, p_i^{\phi(i)}$. The latter expression occurs in Lemma 13 (2). We conclude that the two approaches of Lemma 13 are really different.

6 Internal and external likelihood and learning, with multiset data

The straightforward way to describe collections of data "of type X" is as multisets over X, that is, as elements of $\mathcal{M}(X)$. In line with the previous section we are going to push things to a slightly higher level of abstraction: we will use *multisets of predicates on X* as data of type X. This will include point data of type X via point predicates $\mathbf{1}_x$ for $x \in X$, giving an inclusion $\mathcal{M}(X) \hookrightarrow \mathcal{M}\big(\mathrm{Pred}(X)\big)$. Using the more general predicates instead of points is useful, as we will illustrate below, for instance when we deal with incomplete information — caused for instance by measurement or transmission errors. We can handle such situations *e.g.* via uniform predicates, giving each element the same probability. We shall also see that learning along a channel can be handled via multisets of predicates, even if we start from point data.

The first step that we need to take is to formulate likelihood for such data, as multisets of predicates. After all, learning is about increasing likelihood. As we have already seen in Section 5, especially in Lemma 13, there are two forms of likelihood that make sense. We call them *external* and *internal* and write them as $\models_{\mathrm{E}}$ and $\models_{\mathrm{I}}$. The distinction is first made in [23], but it is not explicit elsewhere — as far as we are aware. Both forms of likelihood make sense, and also the associated learning methods. We shall discuss the non-trivial, unsolved issue of when to use which likelihood (and learning method) in Section 9.

We fix the general formulation of these two forms of likelihood[1].

Definition 15. *Let $\omega \in \mathcal{D}(X)$ be a state and $\psi \in \mathcal{M}\big(\mathrm{Pred}(X)\big)$ be a multiset of* data*.*

1. *The* external likelihood *of the data in this state is defined as:*

$$\omega \models_{E} \psi \coloneqq \prod\nolimits_{p} (\omega \models p)^{\psi(p)}.$$

2. *The* internal likelihood *is:*

$$\omega \models_{I} \psi \coloneqq \omega \models \&_p \, p^{\psi(p)}.$$

The log-likelihood is the (natural) logarithm of these expressions. In the external case it can be computed simply as sum $\sum_p \psi(p) \cdot \log(\omega \models p)$. In the internal case we can compute the log-likelihood as an iterated sum, using Bayes' rule.

This log-likelihood is useful since these likelihoods can be become very small in the presence of lots of data.

One could argue that external and internal likelihood require an additional multinomial coefficient (ϕ), like in Lemma 13, in order to accommodate all possible orderings of the data items. However, when considering likelihood, it is usually omitted. Learning aims at increasing likelihood and a constant factor is then irrelevant.

The following result is standard, see *e.g.* [28, Ex. 17.5]. A proof is in the appendix.

[1] In [23] the phrases 'multiple state' and 'copied state' are used for what we here started calling 'external' and 'internal'.

Proposition 16. *For point data $\phi \in \mathcal{M}_*(X)$ the predicate "external likelihood of ϕ"*

$$\mathcal{D}(X) \xrightarrow{(-) \models_{\mathrm{E}} \phi} [0,1]$$

takes its maximum at the distribution $\mathit{Flrn}(\phi) \in \mathcal{D}(X)$ that is obtained by frequentist learning. □

The next observation is of interest mostly for categorical aficionados.

Remark 17. As briefly mentioned in Subsection 2.1, the set of multisets $\mathcal{M}(X)$ is the free commutative monoid on X. Both forms of likelihood $\models_{\mathrm{E}}$ and $\models_{\mathrm{I}}$ in Definition 15 can be understood as maps $\mathcal{L}_E, \mathcal{L}_I \colon \mathcal{M}\big(\mathrm{Pred}(X)\big) \to \mathrm{Pred}\big(\mathcal{D}(X)\big)$, arising via this freeness, but in different ways:

$$\begin{tikzcd}
\mathcal{M}\big(\mathrm{Pred}(X)\big) \arrow[rr, dashed, "\overline{\mathit{val}}"] \arrow[dr, dashed, "\overline{\mathrm{id}}"'] & & \mathrm{Pred}\big(\mathcal{D}(X)\big) \\
& \mathrm{Pred}(X) \arrow[ur, "\mathit{val}"'] &
\end{tikzcd}$$

as

$$\begin{cases} \mathcal{L}_E \;=\; \overline{\mathit{val}} \\ \mathcal{L}_I \;=\; \mathit{val} \circ \overline{\mathrm{id}}. \end{cases}$$

where $\mathit{val} \colon \mathrm{Pred}(X) \to \mathrm{Pred}(\mathcal{D}(X))$ is $\mathit{val}(p)(\omega) \coloneqq \omega \models p$.

1. Predicates with conjunction $(\mathbf{1}, \&)$ form a commutative monoid. Hence the above validity map *val* can be extended uniquely to a homomorphism of monoids $\overline{\mathit{val}} \colon \mathcal{M}(\mathrm{Pred}(X)) \to \mathrm{Pred}(\mathcal{D}(X))$, given by:

$$\begin{aligned}
\overline{\mathit{val}}\Big(\textstyle\sum_i n_i|p_i\rangle\Big)(\omega) &= \big(\mathit{val}(p_1)^{n_1} \,\&\, \cdots \,\&\, \mathit{val}(p_k)^{n_k}\big)(\omega) \\
&= \mathit{val}(p_1)(\omega)^{n_1} \cdot \ldots \cdot \mathit{val}(p_k)(\omega)^{n_k} \\
&= (\omega \models p_1)^{n_1} \cdot \ldots \cdot (\omega \models p_k)^{n_k} \\
&= \omega \models_{\mathrm{E}} \textstyle\sum_i n_i|p_i\rangle.
\end{aligned}$$

   We use that conjunction & of predicates is defined via pointwise multiplication.
2. The identity map $\mathrm{id} \colon \mathrm{Pred}(X) \to \mathrm{Pred}(X)$ can also be extended to a homomorphims of monoids $\overline{\mathrm{id}} \colon \mathcal{M}(\mathrm{Pred}(X)) \to \mathrm{Pred}(X)$, via:

$$\overline{\mathrm{id}}\Big(\textstyle\sum_i n_i|p_i\rangle\Big) \;=\; p_1^{n_1} \,\&\, \cdots \,\&\, p_k^{n_k}.$$

   Hence, $\big(\mathit{val} \circ \overline{\mathrm{id}}\big)(\sum_i n_i|p_i\rangle)(\omega) \;=\; \omega \models p_1^{n_1} \,\&\, \cdots \,\&\, p_k^{n_k} \;=\; \omega \models_{\mathrm{I}} \sum_i n_i|p_i\rangle$.

We have described (a form of) learning in Subsection 4.5 as validity increase, whereby we immediately mentioned that validity is really used as a likelihood function. For the explicitly defined likelihood functions $\models_{\mathrm{E}}$ and $\models_{\mathrm{I}}$ introduced in this section there are also associated (different) learning steps, both as 'likelihood increase'.

Theorem 18. *For a state $\omega \in \mathcal{D}(X)$ with a data $\psi = \sum_i n_i|p_i\rangle \in \mathcal{M}_*(\mathrm{Pred}(X))$, one has:*

1. $\omega \models_{E} \psi \;\leq\; \mathit{Elrn}(\omega, \psi) \models_{E} \psi$, *where:*

$$\mathit{Elrn}(\omega, \psi) \coloneqq \sum\nolimits_i \frac{n_i}{n} \cdot \omega|_{p_i} \;=\; \sum\nolimits_p \mathit{Flrn}(\psi)(p) \cdot \omega|_p,$$

   for $n = \|\psi\| = \sum_i n_i > 0$.

2. $\omega \models_{I} \psi \;\leq\; \mathit{Ilrn}(\omega, \psi) \models_{I} \psi$, *where:*

$$\mathit{Ilrn}(\omega, \psi) \coloneqq \omega|_{\&_i p_i^{n_i}}.$$

Proof. The second point is an immediate consequence of Proposition 12. A proof of the first point is given in the appendix. □

We notice that frequentist learning $\mathit{Flrn}(\phi)$ for $\phi \in \mathcal{M}_*(X)$ is a special case of external learning form the uniform state υ with point data ϕ, namely $\mathit{Elrn}(\upsilon, \phi)$, since $\upsilon|_{\mathbf{1}_x} = 1|x\rangle$. In fact, one can take instead of υ any state with full support. External learning from point data immediately jumps to the maximal likelihood, see Proposition 16.

External learning also satisfies, like frequentist learning, the more-is-the-same property (8), namely:

$$\mathit{Elrn}(\omega, K \cdot \psi) \;=\; \mathit{Elrn}(\omega, \psi), \quad \text{for } K > 0. \tag{20}$$

This external learning thus combines frequentist and Bayesian learning, via the convex combination of (Bayesian) updated states, see the formulation of *Elrn* in Theorem 18 (1).

An important advantage of *internal* learning is that it can be done incrementally: when new data arrives, one can continue learning with what has been learned so far, simply by performing a new conditioning. In particular, $\mathit{Ilrn}(\omega, K \cdot \phi)$ is not the same as $\mathit{Ilrn}(\omega, \phi)$, but involves K iterations of learning from ϕ.

Proposition 19. $\mathit{Ilrn}(\omega, \phi + \psi) = \mathit{Ilrn}\big(\mathit{Ilrn}(\omega, \phi), \psi\big)$.

Proof. Since:

$$\begin{aligned}
\mathit{Ilrn}(\omega, \phi + \psi) &= \omega\big|_{\&_p\, p^{(\phi+\psi)(p)}} \\
&= \omega\big|_{\&_p\, p^{\phi(p)+\psi(p)}} \\
&= \omega\big|_{\&_p\, p^{\phi(p)} \,\&\, p^{\psi(p)}} \\
&= \omega\big|_{(\&_p\, p^{\phi(p)}) \,\&\, (\&_p\, p^{\psi(p)})} \\
&= \omega\big|_{\&_p\, p^{\phi(p)}}\big|_{\&_p\, p^{\psi(p)}} && \text{by Lemma 10 (1)} \\
&= \mathit{Ilrn}\big(\mathit{Ilrn}(\omega, \phi), \psi\big).
\end{aligned}$$

□

More technically, this result says that the multiset monoid $\mathcal{M}(\mathrm{Pred}(X))$ acts on $\mathcal{D}(X)$ via internal learning. Indeed, we also have $\mathit{Ilrn}(\omega, \mathbf{0}) = \omega|_{\mathbf{1}} = \omega$.

Interestingly, in the *external* case we can express an increase in likelihood equivalently as a decrease in divergence. Informally, this means that learning from what's right coincides with learning from what's wrong. The divergence is the familiar Kullback-Leibler divergence, which is defined for two states $\sigma, \tau \in \mathcal{D}(X)$ as:

$$D_{KL}(\sigma, \tau) \coloneqq \sum_{x \in X} \sigma(x) \cdot \log\left(\frac{\sigma(x)}{\tau(x)}\right).$$

The logarithm log is typically the 2-log.

For simplicity we shall assume, like in Section 5, that the predicates $P = \{p_1, \ldots, p_n\}$ at hand form a test, *i.e.* satisfy $\bigovee_i p_i = \mathbf{1}$. In this way we can define an evaluation channel:

$$\mathcal{D}(X) \xrightarrow{\text{ev}} P \qquad \text{by} \qquad \text{ev}(\omega) \coloneqq \textstyle\sum_i (\omega \models p_i) \big| p_i \big\rangle. \tag{21}$$

Proposition 20. *Let $\phi \in \mathcal{M}(P)$ be a non-empty multiset of data, over a set of predicates $P = \{p_1, \ldots, p_n\} \subseteq \text{Pred}(X)$ forming a test. Then, for two states ω, ω',*

$$\big(\omega \models_{\scriptscriptstyle E} \phi\big) \;\leq\; \big(\omega' \models_{\scriptscriptstyle E} \phi\big) \iff D_{KL}\big(\mathit{Flrn}(\phi), \text{ev}(\omega)\big) \;\geq\; D_{KL}\big(\mathit{Flrn}(\phi), \text{ev}(\omega')\big).$$

Thus, increasing likelihood of data ϕ in state ω corresponds to decreasing divergence between the distributions $\mathit{Flrn}(\phi)$ and $\text{ev}(\omega)$.

Proof. Let $\phi = \sum_i n_i | p_i \rangle$ and $n = \|\phi\| = \sum_i n_i$. We first notice that:

$$\begin{aligned}
D_{KL}\big(\mathit{Flrn}(\phi), \text{ev}(\omega)\big) &= \textstyle\sum_i \dfrac{n_i}{n} \cdot \log\left(\dfrac{n_i/n}{\omega \models p_i}\right) \\
&= \textstyle\sum_i \dfrac{n_i}{n} \cdot \log\left(\dfrac{n_i}{n}\right) - \textstyle\sum_i \dfrac{n_i}{n} \cdot \log\big(\omega \models p_i\big) \\
&= \Big(\textstyle\sum_i \dfrac{n_i}{n} \cdot \log\left(\dfrac{n_i}{n}\right)\Big) - \dfrac{1}{n} \cdot \log\big(\textstyle\prod_i (\omega \models p_i)^{n_i}\big) \\
&= -H(\phi) - \dfrac{1}{n} \cdot \log\big(\omega \models_{\scriptscriptstyle E} \phi\big).
\end{aligned}$$

where $H(\phi)$ is the entropy of ϕ. The result now follows easily, using that log preserves and reflects the order:

$$\begin{aligned}
& D_{KL}\big(\mathit{Flrn}(\phi), \text{ev}(\omega)\big) \;\geq\; D_{KL}\big(\mathit{Flrn}(\phi), \text{ev}(\omega')\big) \\
& \iff -H(\phi) - \frac{1}{n} \cdot \log\big(\omega \models_{\scriptscriptstyle E} \phi\big) \;\geq\; -H(\phi) - \frac{1}{n} \cdot \log\big(\omega' \models_{\scriptscriptstyle E} \phi\big) \\
& \iff \log\big(\omega \models_{\scriptscriptstyle E} \phi\big) \;\leq\; \log\big(\omega' \models_{\scriptscriptstyle E} \phi\big) \\
& \iff \big(\omega \models_{\scriptscriptstyle E} \phi\big) \;\leq\; \big(\omega' \models_{\scriptscriptstyle E} \phi\big).
\end{aligned}$$

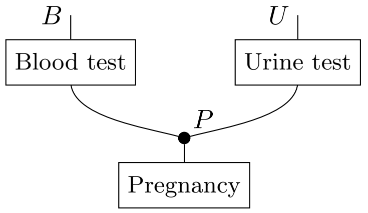
□

In the remainder of this section we consider illustrations of external and internal learning.

6.1 External learning with complete and missing data

We examine an example from [27, §6.2.1], first with complete and then with missing data. The goal is two-fold: to illustrate the external learning step of Theorem 18 (1), and also to show why it pays off to have multisets of *predicates* as data, instead of multisets of elements. These predicates will be used to deal with the uncertainty given by missing data.

The example involves pregnancy of cows, which can be deduced from a urine test and a blood test. A simple Bayesian network structure is assumed, which we write as string diagram with explicit copy:



$$\text{with sets} \qquad \begin{cases} P = \{p, p^{\perp}\} \\ B = \{b, b^{\perp}\} \\ U = \{u, u^{\perp}\}. \end{cases} \qquad (22)$$

The elements p and $p^{\perp}$ represent 'pregnancy' and 'no pregnancy', respectively. Similarly, $b, b^{\perp}$ and $u, u^{\perp}$ represent a positive and negative blood/urine test.

We have two tables with data in Figure 1: the one on the left below contains 'complete' data that can be used directly for learning. The table on the right (copied from [27]) uses a question mark for a missing item. In both cases the aim is to learn an interpretation of the above Bayesian network. This is commonly called parameter learning. It involves learning a state on P and two channels $P \mathrel{\circ\!\!\rightarrow} B$ and $P \mathrel{\circ\!\!\rightarrow} U$. These channels correspond to conditional probability tables in Bayesian networks, see [26] for more details. The state and two channels can be obtained from a joint state on $P \times B \times U$ by marginalisation and disintegration (channel extraction). Our aim is thus to first learn such a joint state from the tables.

case	Pregn	Blood	Urine
1	$p^{\perp}$	b	u
2	p	$b^{\perp}$	u
3	p	b	$u^{\perp}$
4	p	b	$u^{\perp}$
5	$p^{\perp}$	$b^{\perp}$	u
6	$p^{\perp}$	$b^{\perp}$	$u^{\perp}$
7	$p^{\perp}$	b	u
8	p	b	$u^{\perp}$

case	Pregn	Blood	Urine
1	?	b	u
2	p	$b^{\perp}$	u
3	p	b	?
4	p	b	$u^{\perp}$
5	?	$b^{\perp}$	?

Fig. 1. Two tables with data to learn an interpretation of the Bayesian network (22), with 'complete' data on the left and with 'missing' data on the right.

We start with the table on the left. It is translated into a multiset ϕ of point predicates on the product space $P \times B \times U$. The table translates directly into:

$$\phi = 2\big|\mathbf{1}_{(p^{\perp},b,u)}\big\rangle + 1\big|\mathbf{1}_{(p,b^{\perp},u)}\big\rangle + 3\big|\mathbf{1}_{(p,b,u^{\perp})}\big\rangle + 1\big|\mathbf{1}_{(p^{\perp},b^{\perp},u)}\big\rangle + 1\big|\mathbf{1}_{(p^{\perp},b^{\perp},u^{\perp})}\big\rangle.$$

Since there is no prior knowledge we use the uniform state $\upsilon \in \mathcal{D}(P \times B \times U)$ in external learning, giving, according to Theorem 18 (1):

$$\begin{aligned}
\omega &\coloneqq \mathit{Elrn}(\upsilon, \phi) \\
&= \tfrac{2}{8}\upsilon|_{\mathbf{1}_{(p^\perp,b,u)}} + \tfrac{1}{8}\upsilon|_{\mathbf{1}_{(p,b^\perp,u)}} + \tfrac{3}{8}\upsilon|_{\mathbf{1}_{(p,b,u^\perp)}} + \tfrac{1}{8}\upsilon|_{\mathbf{1}_{(p^\perp,b^\perp,u)}} + \tfrac{1}{8}\upsilon|_{\mathbf{1}_{(p^\perp,b^\perp,u^\perp)}} \\
&= \tfrac{2}{8}|p^\perp, b, u\rangle + \tfrac{1}{8}|p, b^\perp, u\rangle + \tfrac{3}{8}|p, b, u^\perp\rangle + \tfrac{1}{8}|p^\perp, b^\perp, u\rangle + \tfrac{1}{8}|p^\perp, b^\perp, u^\perp\rangle \\
&= \mathit{Flrn}(\phi).
\end{aligned}$$

Notice that internal learning would not work in this situation because the conjunction & of these point predicates is falsum $\mathbf{0}$.

This learned joint state $\omega \in \mathcal{D}(P \times B \times U)$ has first marginal $\pi_1 \gg \omega = \frac{1}{2}|p\rangle + \frac{1}{2}|p^\perp\rangle \in \mathcal{D}(P)$, which is used as interpretation of the Pregnancy state in (22). Channels $c\colon P \mathrel{\circ\!\!\to} B$ and $d\colon P \mathrel{\circ\!\!\to} U$ are extracted from ω as conditional probabilities, via disintegration (see Subsection 2.2):

$$\begin{aligned}
c(p) &= \frac{\omega(p,b,u) + \omega(p,b,u^\perp)}{\omega(p,b,u) + \omega(p,b,u^\perp) + \omega(p,b^\perp,u) + \omega(p,b^\perp,u^\perp)}\big|\,b\,\big\rangle \\
&\qquad + \frac{\omega(p,b^\perp,u) + \omega(p,b^\perp,u^\perp)}{\omega(p,b,u) + \omega(p,b,u^\perp) + \omega(p,b^\perp,u) + \omega(p,b^\perp,u^\perp)}\big|\,b^\perp\,\big\rangle \\
&= \frac{3/8}{3/8 + 1/8}|\,b\,\rangle + \frac{1/8}{3/8 + 1/8}|\,b^\perp\,\rangle \;=\; \tfrac{3}{4}|\,b\,\rangle + \tfrac{1}{4}|\,b^\perp\,\rangle \\
c(p^\perp) &= \frac{1/4}{1/4 + 1/8 + 1/8}|\,b\,\rangle + \frac{1/8 + 1/8}{1/4 + 1/8 + 2/8}|\,b^\perp\,\rangle \;=\; \tfrac{1}{2}|\,b\,\rangle + \tfrac{1}{2}|\,b^\perp\,\rangle
\end{aligned}$$

In the same way one gets $d(p) = \frac{1}{4}|\,u\,\rangle + \frac{3}{4}|\,u^\perp\,\rangle$ and $d(p^\perp) = \frac{3}{4}|\,u\,\rangle + \frac{1}{4}|\,u^\perp\,\rangle$. The table on the left in Figure 1 thus gives us an interpretation of the Bayesian network in (22).

We now turn to the table on the right in Figure 1 with missing data. For cases 1,3,5 we don't use point predicates, like above, but predicates $p_1, p_3, p_5\colon P \times B \times U \to [0,1]$ given by:

$$\begin{gathered}
p_1(p,b,u) \;=\; p_1(p^\perp,b,u) \;=\; 1 \qquad\qquad p_3(p,b,u) \;=\; p_3(p,b,u^\perp) \;=\; 1 \\
p_5(p,b^\perp,u) \;=\; p_5(p^\perp,b^\perp,u) \;=\; p_5(p,b^\perp,u^\perp) \;=\; p_5(p^\perp,b^\perp,u^\perp) \;=\; 1.
\end{gathered}$$

These predicate are zero elsewhere — and are thus sharp. We thus translate the table on the right in Figure 1 to the multiset of predicates:

$$\phi \;=\; 1|\,p_1\,\rangle + 1|\,\mathbf{1}_{(p,b^\perp,u)}\,\rangle + 1|\,p_3\,\rangle + 1|\,\mathbf{1}_{(p,b,u^\perp)}\,\rangle + 1|\,p_5\,\rangle.$$

We again follow Theorem 18 (1):

$$\begin{aligned}
\rho \;&\coloneqq\; \mathit{Elrn}\big(\upsilon, \phi\big) \\
&=\; \tfrac{1}{5}\upsilon|_{p_1} + \tfrac{1}{5}\upsilon|_{\mathbf{1}_{(p,b^{\perp},u)}} + \tfrac{1}{5}\upsilon|_{p_3} + \tfrac{1}{5}\upsilon|_{\mathbf{1}_{(p,b,u^{\perp})}} + \tfrac{1}{5}\upsilon|_{p_5} \\
&=\; \tfrac{1}{10}|\,p,b,u\,\rangle + \tfrac{1}{10}|\,p^{\perp},b,u\,\rangle + \tfrac{1}{5}|\,p,b^{\perp},u\,\rangle + \tfrac{1}{10}|\,p,b,u\,\rangle + \tfrac{1}{10}|\,p,b,u^{\perp}\,\rangle \\
&\qquad + \tfrac{1}{5}|\,p,b,u^{\perp}\,\rangle + \tfrac{1}{20}|\,p,b^{\perp},u\,\rangle + \tfrac{1}{20}|\,p^{\perp},b^{\perp},u\,\rangle + \tfrac{1}{20}|\,p,b^{\perp},u^{\perp}\,\rangle + \tfrac{1}{20}|\,p^{\perp},b^{\perp},u^{\perp}\,\rangle. \\
&=\; \tfrac{1}{5}|\,p,b,u\,\rangle + \tfrac{3}{10}|\,p,b,u^{\perp}\,\rangle + \tfrac{1}{4}|\,p,b^{\perp},u\,\rangle + \tfrac{1}{20}|\,p,b^{\perp},u^{\perp}\,\rangle \\
&\qquad + \tfrac{1}{10}|\,p^{\perp},b,u\,\rangle + \tfrac{1}{20}|\,p^{\perp},b^{\perp},u\,\rangle + \tfrac{1}{20}|\,p^{\perp},b^{\perp},u^{\perp}\,\rangle.
\end{aligned}$$

This yields a different interpretation for the Bayesian network (string diagram) in (22): the first marginal of ρ is $\frac{5}{8}|\,b\,\rangle + \frac{3}{8}|\,b^{\perp}\,\rangle$. The extracted channels $c\colon P \rightsquigarrow B$ and $d\colon P \rightsquigarrow U$ from ρ are obtained as before:

$$\begin{aligned}
c(p) \;&=\; \frac{{}^{1}/_{5} + {}^{3}/_{10}}{{}^{1}/_{5} + {}^{3}/_{10} + {}^{1}/_{4} + {}^{1}/_{20}}|\,b\,\rangle + \frac{{}^{1}/_{4} + {}^{1}/_{20}}{{}^{1}/_{5} + {}^{3}/_{10} + {}^{1}/_{4} + {}^{1}/_{20}}|\,b^{\perp}\,\rangle \;=\; \tfrac{5}{8}|\,b\,\rangle + \tfrac{3}{8}|\,b^{\perp}\,\rangle. \\
c(p^{\perp}) \;&=\; \frac{{}^{1}/_{10}}{{}^{1}/_{10} + {}^{1}/_{20} + {}^{1}/_{20}}|\,b\,\rangle + \frac{{}^{1}/_{20} + {}^{1}/_{20}}{{}^{1}/_{10} + {}^{1}/_{20} + {}^{1}/_{20}}|\,b^{\perp}\,\rangle \;=\; \tfrac{1}{2}|\,b\,\rangle + \tfrac{1}{2}|\,b^{\perp}\,\rangle.
\end{aligned}$$

Similarly $d(p) = \frac{9}{16}|\,u\,\rangle + \frac{7}{16}|\,u^{\perp}\,\rangle$ and $d(p^{\perp}) = \frac{3}{4}|\,u\,\rangle + \frac{1}{4}|\,u^{\perp}\,\rangle$. These outcomes are precisely as described in [27, §6.2.1]; there, the computation is presented as an instance of the (E-part of the) EM-algorithm (see Section 8 below).

7 Learning coin bias, along a channel

So far we have looked at likelihood of data in a state and at how to increase this likelihood by adapting the state. We have considered the situation where the state and data are on *the same* set X. In practice, it often happens that there is a difference, like in:

$$\begin{tikzcd}[row sep=small]
& X \arrow[r, "e"] & Y & \\
\text{state to be learned} \arrow[ur, dotted] & & & \text{data} \arrow[ul, dotted]
\end{tikzcd} \tag{23}$$

We will assume that there is a channel between the two spaces — as in the above picture — that can be used to mediate between the given data and the state that we wish to learn. This is what we call 'learning along a channel'. This learning challenge is often described in terms of 'hidden' or 'latent' variables, since the elements of the space X are not directly accessible, but only indirectly via the 'emission' channel e. This forms the E-part of what is called Expectation-Maximisation (EM), see Section 8, where, in the M-part, the channel e becomes a learning goal in itself. In Expectation-Maximisation these E- and M-parts are alternated. But here we first concentrate on the E-part only and assume that the channel e is given and remains fixed. This E-part typically uses what we call external learning.

Now suppose, in the setting (23) we have data $\psi \in \mathcal{M}(\mathrm{Pred}(Y))$ in the form of multiset of predicates on the codomain Y of the channel. We can easily turn this multiset on Pred(Y) into a multiset on Pred(X), via predicate transformation (and functoriality of $\mathcal{M}$). Then we can both externally and internally learn 'along a channel', using the formulations of Theorem 18:

$$\begin{aligned} \mathit{Elrn}(\omega, e, \psi) &\coloneqq \mathit{Elrn}\big(\omega, \textstyle\sum_q \psi(q)\big| e \ll q\big\rangle\big) \;=\; \textstyle\sum_q \frac{\psi(q)}{\|\psi\|} \cdot \omega|_{e \ll q} \\ \mathit{Ilrn}(\omega, e, \psi) &\coloneqq \mathit{Ilrn}\big(\omega, \textstyle\sum_q \psi(q)\big| e \ll q\big\rangle\big) \;=\; \omega\big|_{\&_q (e \ll q)^{\psi(q)}}. \end{aligned} \tag{24}$$

Notice that we overload the notation *Elrn* / *Ilrn*, since on the left of $\coloneqq$ it is used with three arguments, for learning along a channel, which is defined in terms of the original notion, on the right of $\coloneqq$, with two arguments.

Proposition 21. *The above definitions* (24) *give the following likelihood increases. For* $\omega' \coloneqq \mathit{Elrn}(\omega, e, \psi)$ *one gets:*

$$e \gg \omega' \models_{E} \psi \;\;\geq\;\; e \gg \omega \models_{E} \psi.$$

And for $\omega' \coloneqq \mathit{Ilrn}(\omega, e, \psi)$ *one simply has:*

$$\omega' \models_{I} \textstyle\sum_q \psi(q)\big| e \ll q\big\rangle \;\;\geq\;\; \omega \models_{I} \textstyle\sum_q \psi(q)\big| e \ll q\big\rangle.$$

Proof. Both likelihood inequalities follow from Theorem 18. The first one also involves (14). □

Internal learning along a channel also works incrementally, analogously to Proposition 19. We now use the terminology of actions, as already briefly mentioned after the proof of Proposition 19.

Proposition 22. *For a fixed channel* $e \colon X \mathrel{\circ\!\!\!\rightarrow} Y$*, internal learning along* e *forms an action of the multiset monoid of data on states:*

$$\mathcal{D}(X) \times \mathcal{M}(\mathrm{Pred}(Y)) \xrightarrow{\mathit{Ilrn}(-,e,-)} \mathcal{D}(X)$$

The same works for point data $\mathcal{M}(Y)$ *instead of predicates* $\mathcal{M}(\mathrm{Pred}(Y))$*.*

Alternatively, one can say that internal learning forms an algebra for the writer monad $(-) \times \mathcal{M}(\mathrm{Pred}(Y))$ on the category of sets.

Proof. As before we have $\mathit{Ilrn}(\omega, e, \mathbf{0}) = \mathit{Ilrn}(\omega, \mathbf{0}) = \omega$ and:

$$\begin{aligned} \mathit{Ilrn}\big(\omega, e, \phi + \psi\big) &= \mathit{Ilrn}\big(\omega, \textstyle\sum_q (\phi + \psi)(q)\big| e \ll q\big\rangle\big) \\ &= \mathit{Ilrn}\big(\omega, \textstyle\sum_q (\phi(q) + \psi(q))\big| e \ll q\big\rangle\big) \\ &= \mathit{Ilrn}\big(\omega, (\textstyle\sum_q \phi(q)\big| e \ll q\big\rangle) + (\textstyle\sum_q \psi(q)\big| e \ll q\big\rangle)\big) \\ &= \mathit{Ilrn}\big(\mathit{Ilrn}\big(\omega, \textstyle\sum_q \phi(q)\big| e \ll q\big\rangle\big), \textstyle\sum_q \psi(q)\big| e \ll q\big\rangle\big) \\ &\qquad \text{by Proposition 19} \\ &= \mathit{Ilrn}\big(\mathit{Ilrn}(\omega, e, \phi), e, \psi\big). \end{aligned}$$

□

External learning with point data can also be captured via the dagger of a channel, see Subsection 4.4.

Proposition 23. *When the data in the above situation consists of point data $\phi \in \mathcal{M}(Y)$, then external learning along channel e can be described via the dagger of e, as in:*

$$\mathit{Elrn}(\omega, e, \psi) \,=\, e^{\dagger}_{\omega} \gg \mathit{Flrn}(\phi).$$

Proof. Since:

$$\begin{aligned}
\mathit{Elrn}(\omega, e, \phi) \overset{(24)}{=} \sum\nolimits_x \frac{\phi(x)}{\|\phi\|} \cdot \omega|_{e \ll \mathbf{1}_x} &\overset{(16)}{=} \sum\nolimits_x \frac{\phi(x)}{\|\phi\|} \cdot e^{\dagger}_{\omega}(x) \\
&= \sum\nolimits_x \mathit{Flrn}(\phi)(x) \cdot e^{\dagger}_{\omega}(x) \\
&= e^{\dagger}_{\omega} \gg \mathit{Flrn}(\phi). \qquad \square
\end{aligned}$$

Notice that in this result we start with point data $\phi \in \mathcal{M}(Y)$. But the actual learning happens via transformed data $\sum_y \frac{\phi}{\|\phi\|} \big| e \ll \mathbf{1}_y \big\rangle$. The latter multiset no longer involves sharp point predicates, but fuzzy predicates $e \ll \mathbf{1}_y$. This is another reason why we have formulated data as multisets of *predicates* and not simply as multisets of *points*.

In the remainder of this section we illustrate learning along a channel in the classical situation where one wishes to learn the bias of an unknown coin form a given number of coin flips. In this situation one typically uses the flip channel describing a biased coin:

$$[0,1] \xrightarrow{\mathit{flip}} \{H, T\} \qquad \text{with} \qquad \mathit{flip}(r) \,=\, r|H\rangle + (1-r)|T\rangle.$$

In order to keep things simple we avoid continuous probability on the unit interval $[0,1]$. Instead, we discretise it and use instead the 21-point domain:

$$D \coloneqq \{0, \tfrac{1}{20}, \tfrac{2}{20}, \ldots, \tfrac{19}{20}, 1\} \,\subseteq\, [0,1] \qquad \text{with} \qquad D \xrightarrow{\mathit{flip}} \{H, T\}.$$

The codomain $\{H, T\}$ of the flip channel carries two sharp point predicates $\mathbf{1}_H$ and $\mathbf{1}_T$ describing head and tail evidence. Predicate transformation turns them into two fuzzy predicates on D, namely:

$$\mathit{flip} \ll \mathbf{1}_H,\, \mathit{flip} \ll \mathbf{1}_T \,\in\, \mathrm{Pred}(D) \qquad \text{with} \qquad \begin{cases} (\mathit{flip} \ll \mathbf{1}_H)(r) \,=\, r \\ (\mathit{flip} \ll \mathbf{1}_T)(r) \,=\, 1 - r. \end{cases}$$

If we start from the uniform state $\upsilon = \sum_{0 \leq i \leq 20} \frac{1}{21} \big| \frac{i}{20} \big\rangle$ on D, then the probability of getting head is:

$$\upsilon \models \mathit{flip} \ll \mathbf{1}_H \,=\, \sum_{0 \leq i \leq 20} \tfrac{1}{21} \cdot \tfrac{i}{20} \,=\, \tfrac{1}{21} \cdot \tfrac{1}{20} \cdot \Big(\sum_{0 \leq i \leq 20} i \Big) \,=\, \tfrac{1}{21} \cdot \tfrac{1}{20} \cdot \tfrac{20 \cdot 21}{2} \,=\, \tfrac{1}{2}.$$

Similarly $\upsilon \models \mathit{flip} \ll \mathbf{1}_T = \frac{1}{2}$.

The questions we now ask ourselves are:

> What is the probability of seeing head after observing n heads and m tails, that is after learning from the multiset of predicates $n\big|\,\mathit{flip} \ll \mathbf{1}_H\,\big\rangle + m\big|\,\mathit{flip} \ll \mathbf{1}_T\,\big\rangle$. In which updated/learned state should the predicate $c \ll \mathbf{1}_H$ be evaluated to answer this question? Should we use external or internal learning along the flip channel?

Concretely, should we use the external variant or the internal version below, as newly learned state on D:

$$\begin{aligned}
\mathit{Elrn}\big(\upsilon, \mathit{flip}, n|\,H\,\rangle + m|\,T\,\rangle\big) &= \tfrac{n}{n+m}\upsilon|_{\mathit{flip}\ll\mathbf{1}_H} + \tfrac{m}{n+m}\upsilon|_{\mathit{flip}\ll\mathbf{1}_T} \\
&= \mathit{flip}^{\dagger}_{\upsilon} \gg \big(\tfrac{n}{n+m}|\,H\,\rangle + \tfrac{m}{n+m}|\,T\,\rangle\big) \\
\mathit{Ilrn}\big(\upsilon, \mathit{flip}, n|\,H\,\rangle + m|\,T\,\rangle\big) &= \upsilon|_{(\mathit{flip}\ll\mathbf{1}_H)^n \,\&\, (\mathit{flip}\ll\mathbf{1}_T)^m} \\
&= \upsilon\,\underbrace{|_{\mathit{flip}\ll\mathbf{1}_H}\cdots|_{\mathit{flip}\ll\mathbf{1}_H}}_{n \text{ times}}\underbrace{|_{\mathit{flip}\ll\mathbf{1}_T}\cdots|_{\mathit{flip}\ll\mathbf{1}_T}}_{m \text{ times}}.
\end{aligned}$$

The order of the single updates in the last line does not matter.

Figure 2 contains bar charts describing these learned distributions, for various numbers n, m. We see, from these charts, and from the above formula, that the externally learned distribution for n, m is the same as for $K \cdot n, K \cdot m$. As we notices before, this is characteristic for external/frequentist learning. In contrast, internal learning is truly Bayesian: in the charts for the internally learned distribution one can recognise the a discretised version of the continuous beta distribution — to be precise, $\beta(n+1, m+1)$. Its variance becomes small with rising n, m, so that a higher precision is reached. It is well-known that these distributions have $\frac{n+1}{(n+1)+(m+1)}$ as mean. This is at the same time the validity $\mathit{Ilrn}(\upsilon, \mathit{flip}, n|\,H\,\rangle + m|\,T\,\rangle) \models c \ll \mathbf{1}_H$.

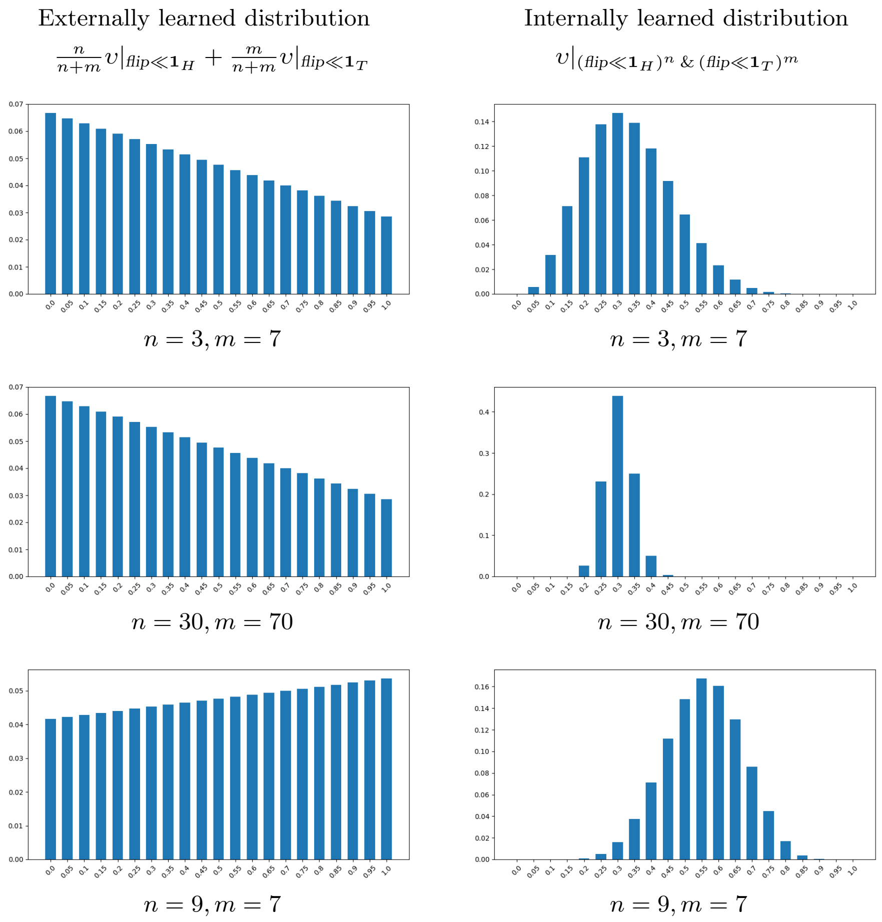
The interested reader may wish to check that the mean of the externally learned distribution $\frac{n}{n+m}\upsilon|_{\mathit{flip}\ll\mathbf{1}_H} + \frac{m}{n+m}\upsilon|_{\mathit{flip}\ll\mathbf{1}_T}$ is $\frac{41\cdot n+19\cdot m}{60\cdot(n+m)}$.

What to make of this? In every textbook treatment of coin bias learning one finds the internal approach. It presents an intuitively clear picture, with decreasing variance as the numbers n, m of heads and tails go up, and the 'expected' expected value $\frac{n+1}{(n+1)(m+1)}$. Is there an intrinsic reason why external learning is appropriate in Subsection 6.1 (and in the next section) and not here? See Section 9 for a perspective.

We like to conclude our description of coin bias learning with the conjugate prior property for internal learning. Our presentation here is different from traditional descriptions in two ways:

- it works in discrete, not continuous, probability, with a discretised version of the standard β distribution on $[0, 1]$;
- it formulates conjugate priorship in terms of a homomorphism of actions, building on Proposition 22.

Recall that we use $D = \{0, \frac{1}{20}, \ldots, 1\} \subseteq [0, 1]$ as sample space, with uniform distribution υ on D. For $n, m \in \mathbb{N}$ we define on D the discretised beta



Externally learned distribution

$\frac{n}{n+m}\upsilon|_{\mathit{flip}\ll\mathbf{1}_H} + \frac{m}{n+m}\upsilon|_{\mathit{flip}\ll\mathbf{1}_T}$

Internally learned distribution

$\upsilon|_{(\mathit{flip}\ll\mathbf{1}_H)^n \,\&\, (\mathit{flip}\ll\mathbf{1}_T)^m}$

$n = 3, m = 7$ $n = 3, m = 7$

$n = 30, m = 70$ $n = 30, m = 70$

$n = 9, m = 7$ $n = 9, m = 7$

Fig. 2. Bar charts of learned distributions for coin bias, for different numbers n (of heads) and m (of tails).

distribution:

$$\beta_D(n,m) \coloneqq \sum_{r\in D} \frac{r^n\cdot(1-r)^m}{\sum_{s\in D} s^n\cdot(1-s)^m}\big|\, r\,\big\rangle.$$

Proposition 24. *1. For $n,m\in\mathbb{N}$ the above distribution $\beta_D(n,m)$ satisfies:*

$$\begin{aligned}\beta_D(n,m) &= \upsilon\big|_{(\mathit{flip}\ll\mathbf{1}_H)^n\,\&\,(\mathit{flip}\ll\mathbf{1}_T)^m} \\ &= \mathit{Ilrn}(\upsilon,\mathit{flip}, n|H\rangle + m|T\rangle).\end{aligned}$$

2. For additional numbers $n',m'\in\mathbb{N}$ one has:

$$\beta_D(n,m)\big|_{(\mathit{flip}\ll\mathbf{1}_H)^{n'}\,\&\,(\mathit{flip}\ll\mathbf{1}_T)^{m'}} = \beta_D(n+n',m+m').$$

Proof. 1. For $r\in D$ we have:

$$\begin{aligned}\upsilon\big|_{(\mathit{flip}\ll\mathbf{1}_H)^n\,\&\,(\mathit{flip}\ll\mathbf{1}_T)^m}(r) &= \frac{\upsilon(r)\cdot(\mathit{flip}\ll\mathbf{1}_H)^n(r)\cdot(\mathit{flip}\ll\mathbf{1}_T)^m(r)}{\upsilon\models(\mathit{flip}\ll\mathbf{1}_H)^n\ \&\ (\mathit{flip}\ll\mathbf{1}_T)^m} \\ &= \frac{1/21\cdot r^n\cdot(1-r)^m}{\sum_{s\in D} 1/21\cdot s^n\cdot(1-s)^m} \\ &= \beta_D(n,m)(r).\end{aligned}$$

2. We use this result and Lemma 10 (1) in:

$$\begin{aligned}&\beta_D(n,m)\big|_{(\mathit{flip}\ll\mathbf{1}_H)^{n'}\&(\mathit{flip}\ll\mathbf{1}_T)^{m'}} \\ &= \upsilon\big|_{(\mathit{flip}\ll\mathbf{1}_H)^n\&(\mathit{flip}\ll\mathbf{1}_T)^m}\big|_{(\mathit{flip}\ll\mathbf{1}_H)^{n'}\&(\mathit{flip}\ll\mathbf{1}_T)^{m'}} \\ &= \upsilon\big|_{(\mathit{flip}\ll\mathbf{1}_H)^n\&(\mathit{flip}\ll\mathbf{1}_T)^m\&(\mathit{flip}\ll\mathbf{1}_H)^{n'}\&(\mathit{flip}\ll\mathbf{1}_T)^{m'}} \\ &= \upsilon\big|_{(\mathit{flip}\ll\mathbf{1}_H)^{n+n'}\&(\mathit{flip}\ll\mathbf{1}_T)^{m+m'}} \\ &= \beta_D(n+n',m+m').\end{aligned}$$

□

The equation in the above second point shows that β_D is closed under updating with point predicates transformed along *flip*. It is the reason for calling β_D *conjugate prior* to *flip*. This is convenient because it means that we don't have to perform all the state updates explicitly; instead we can just adapt the inputs n,m of the channel β_D. These inputs are often called hyperparameters.

This conjugate priorship property is described at an abstract level in [25]. Below we give a novel alternative description in terms of monoid actions — or equivalently, algebras of the writer monad. It again expresses that internal learning can be done incrementally.

Corollary 25. *The discretised beta channel $\beta_D\colon\mathbb{N}\times\mathbb{N}\to\mathcal{D}(X)$ forms a map of monoid actions in:*

$$\begin{tikzcd}(\mathbb{N}\times\mathbb{N})\times\mathcal{M}(\{H,T\}) \arrow[r, "\beta_D\times\mathrm{id}"] \arrow[d, "\mathit{add}"'] & \mathcal{D}(D)\times\mathcal{M}(\{H,T\}) \arrow[d, "\mathit{Ilrn}(-,\mathit{flip},-)"] \\ \mathbb{N}\times\mathbb{N} \arrow[r, "\beta_D"] & \mathcal{D}(D)\end{tikzcd}$$

The monoid action add on the left is given by:

$$add\big(n,\, m,\, n'|H\rangle + m'|T\rangle\big) \,=\, (n+n', m+m').$$

The action on the right comes for internal learning along the channel flip: $D \rightsquigarrow \{H, T\}$, *see Proposition 22.* □

8 Expectation-Maximisation

Recall the situation (23) where we have a channel $X \rightsquigarrow Y$ and data on Y. The goal we have considered in the previous section is learning a state on X 'along the channel'. Within the so-called Expectation Maximisation (EM) algorithm, see [17] (and also [31]) this is called the E-step. There is an additional M-step which involves learning a better channel, so as to increase the (external) likelihood of the data. This section contains a fresh description of the EM-algorithm in which the two steps (E and M) are combined in a single learning step. This alternative approach uses a combination of the state on X and the channel $X \rightsquigarrow Y$ into a joint state on $X \times Y$, which is improved via external learning; subsequently, a new state on X and channel $X \rightsquigarrow Y$ are extracted. This re-description of the EM mechanism is applied to a standard EM example from the literature.

We first recall that a state $\omega \in \mathcal{D}(X)$ and a channel $e\colon X \rightsquigarrow Y$ can be combined into a joint state $\tau = \langle \mathrm{id}, e\rangle \gg \omega$, where $\langle \mathrm{id}, e\rangle = (\mathrm{id} \otimes e) \circ \varDelta\colon X \rightsquigarrow X \times Y$. Then: $\tau(x,y) = \omega(x) \cdot e(x)(y)$. The marginals of τ are:

$$\pi_1 \gg \tau \,=\, \omega \qquad \text{and} \qquad \pi_2 \gg \tau \,=\, e \gg \omega.$$

When we extract a channel $X \rightsquigarrow Y$ from τ we rediscover the original channel $e\colon X \rightsquigarrow Y$, via the formula (4).

Now assume we have data $\psi \in \mathcal{M}(\mathrm{Pred}(Y))$ on Y. We can transform (weaken) ψ to data on $X \times Y$, written as:

$$\mathbf{1} \otimes \psi \,:=\, \sum\nolimits_q \psi(q)\big|\,\pi_2 \ll q\,\big\rangle \,=\, \sum\nolimits_q \psi(q)\big|\,\mathbf{1} \otimes q\,\big\rangle.$$

Then $\tau \models_{\mathrm{E}} \mathbf{1} \otimes \psi \,=\, e \gg \omega \models_{\mathrm{E}} \psi$.

The next result gives our combined description of the E- and M-steps of the EM-algorithm via a single external learning step on a joint state. It used the conditioning $e|_q$ of a channel, which is defined pointwise as: $e|_q(x) = e(x)|_q$.

Theorem 26. *Let* $\omega \in \mathcal{D}(X)$ *be state with a channel* $e\colon X \rightsquigarrow Y$, *and with data* $\psi \in \mathcal{M}(\mathrm{Pred}(Y))$. *Write:*

$$\tau \,:=\, \langle \mathrm{id}, e\rangle \gg \omega \,\in\, \mathcal{D}(X \times Y) \qquad \textit{and} \qquad \tau' \,:=\, \mathit{Elrn}(\tau, \mathbf{1} \otimes \psi).$$

1. *The first marginal* $\omega' = \pi_1 \gg \tau'$ *is then the outcome of external learning from the data* ψ *along* e*:*

$$\omega' \,=\, \mathit{Elrn}(\omega, e, \psi) \overset{(24)}{=} \sum\nolimits_q \frac{\psi(q)}{\|\psi\|} \cdot \omega|_{e \ll q}.$$

2. *The channel* $e' \colon X \mathrel{\circ\!\!\!\rightarrow} Y$ *extracted from* $\tau' \in \mathcal{D}(X \times Y)$ *by disintegration is:*

$$e'(x) \;=\; \sum\nolimits_q \frac{\psi(q)}{\|\psi\|} \cdot \frac{\omega|_{e \ll q}(x)}{\omega'(x)} \cdot e|_q(x).$$

   Then $e' \gg \omega' \models_{\scriptscriptstyle E} \psi \;\geq\; e \gg \omega \models_{\scriptscriptstyle E} \psi$.

3. *In the special case where the data is given by points, so* $\psi \in \mathcal{M}(Y)$, *we know from Proposition 23 that the newly learned state* $\omega' = \pi_1 \gg \tau'$ *can be expressed via a dagger, as:* $\omega' = e^{\dagger}_{\omega} \gg \mathit{Flrn}(\psi)$; *the newly learned channel* e' *is then a double dagger:*

$$e' \;=\; \big(e^{\dagger}_{\omega}\big)^{\dagger}_{\mathit{Flrn}(\psi)} \;\colon\; X \mathrel{\circ\!\!\!\rightarrow} Y.$$

   It satisfies $e' \gg \omega' = \mathit{Flrn}(\psi)$ *by* (16), *so that a second channel-learning step with the same data has no effect:* $\big(e'^{\dagger}_{\omega'}\big)^{\dagger}_{\mathit{Flrn}(\psi)} = e'$.

Proof. 1. We get as first marginal of the newly learned joint state τ',

$$\begin{aligned}
\big(\pi_1 \gg \tau'\big)(x) &\;=\; \sum\nolimits_y \tau'(x,y) \\
&\overset{(24)}{=} \sum\nolimits_y \sum\nolimits_q \frac{\psi(q)}{\|\psi\|} \cdot \tau|_{\mathbf{1}\otimes q}(x,y) \\
&\;=\; \sum\nolimits_q \frac{\psi(q)}{\|\psi\|} \cdot \sum\nolimits_y \frac{\tau(x,y) \cdot q(y)}{\tau \models \mathbf{1} \otimes q} \\
&\;=\; \sum\nolimits_q \frac{\psi(q)}{\|\psi\|} \cdot \sum\nolimits_y \frac{\omega(x) \cdot e(x)(y) \cdot q(y)}{e \gg \omega \models q} \\
&\;=\; \sum\nolimits_q \frac{\psi(q)}{\|\psi\|} \cdot \frac{\omega(x) \cdot (e \ll q)(x)}{\omega \models e \ll q} \\
&\;=\; \sum\nolimits_q \frac{\psi(q)}{\|\psi\|} \cdot \omega|_{e \ll q}(x) \\
&\overset{(24)}{=} \mathit{Elrn}(\omega, e, \psi)(x).
\end{aligned}$$

2. The channel $e' \colon X \mathrel{\circ\!\!\!\rightarrow} Y$ extracted from $\tau' \in \mathcal{D}(X \times Y)$ is:

$$\begin{aligned}
e'(x)(y) &\overset{(4)}{=} \frac{\tau'(x,y)}{(\pi_1 \gg \tau')(x)} \\
&= \sum\nolimits_q \frac{\psi(q)}{\|\psi\|} \cdot \frac{\tau|_{\mathbf{1}\otimes q}(x,y)}{\omega'(x)} \\
&= \sum\nolimits_q \frac{\psi(q)}{\|\psi\|} \cdot \frac{1}{\omega'(x)} \cdot \frac{\omega(x) \cdot e(x)(y) \cdot q(y)}{\omega \models e \ll q} \\
&= \sum\nolimits_q \frac{\psi(q)}{\|\psi\|} \cdot \frac{1}{\omega'(x)} \cdot \frac{\omega(x) \cdot (e \ll q)(x)}{\omega \models e \ll q} \cdot \frac{e(x)(y) \cdot q(y)}{e(x) \models q} \\
&= \sum\nolimits_q \frac{\psi(q)}{\|\psi\|} \cdot \frac{1}{\omega'(x)} \cdot \omega|_{e \ll q}(x) \cdot e|_q(x)(y).
\end{aligned}$$

3. Since $e|_{\mathbf{1}_z}(x) = e(x)|_{\mathbf{1}_z} = 1|\,z\,\rangle$ we get:

$$\begin{aligned}
e'(x)(y) \;&=\; \textstyle\sum_z \dfrac{\psi(z)}{\|\psi\|} \cdot \dfrac{\omega|_{e \ll \mathbf{1}_z}(x)}{\omega'(x)} \cdot e|_{\mathbf{1}_z}(x)(y) && \text{by point (2)} \\
&=\; \textstyle\sum_z \dfrac{\mathit{Flrn}(\psi)(z) \cdot e^{\dagger}_{\omega}(z)(x)}{(e^{\dagger}_{\omega} \gg \mathit{Flrn}(\psi))(x)} \cdot 1\big|\,y\,\big\rangle(z) && \text{by (15) and Proposition 23} \\
&=\; \dfrac{\mathit{Flrn}(\psi)(y) \cdot e^{\dagger}_{\omega}(y)(x)}{(e^{\dagger}_{\omega} \gg \mathit{Flrn}(\psi))(x)} \\
&=\; \big(e^{\dagger}_{\omega}\big)^{\dagger}_{\mathit{Flrn}(\psi)}(x)(y) && \text{by (15).} \qquad \square
\end{aligned}$$

The joint-state learning approach, followed by marginalisation and extraction, of Theorem 26 can in principle also be used for internal learning. However, then we don't get a correspondence with learning along a channel — like in Theorem 26 (1). Hence the internal approach fails at this point.

8.1 Candy examples

The textbook [35] contains a chapter titled *Statistical learning methods*, with candy examples in two forms.

First, there is a situation with five different bags, numbered 1, ..., 5, each containing its own mixture of cherry (C) and lime (L) candies. This situation can be described via a candy channel:

$$B \xrightarrow{c} \{C, L\} \quad \text{where} \quad B \,=\, \{1,2,3,4,5\} \quad \text{and} \quad \begin{cases} c(1) \,=\, 1|\,C\,\rangle \\ c(2) \,=\, \frac{3}{4}|\,C\,\rangle + \frac{1}{4}|\,L\,\rangle \\ c(3) \,=\, \frac{1}{2}|\,C\,\rangle + \frac{1}{2}|\,L\,\rangle \\ c(4) \,=\, \frac{1}{4}|\,C\,\rangle + \frac{3}{4}|\,L\,\rangle \\ c(5) \,=\, 1|\,L\,\rangle. \end{cases}$$

The initial bag distribution is $\omega = \frac{1}{10}|\,1\,\rangle + \frac{1}{5}|\,2\,\rangle + \frac{2}{5}|\,3\,\rangle + \frac{1}{5}|\,4\,\rangle + \frac{1}{10}|\,5\,\rangle$.

In the situation described in [35, §20.1] the space of bags B is regarded as hidden (not directly observable), in a scenario where a new bag $i \in B$ is given and candies are drawn from it. It turns out that 10 consecutive draws yield a lime candy[2]. Transforming the lime point predicate along channel c yields the fuzzy predicate $c \ll \mathbf{1}_L \colon B \to [0,1]$ given by:

$$c \ll \mathbf{1}_L \;=\; \textstyle\bigovee_i c(i)(L) \cdot \mathbf{1}_i \;=\; \frac{1}{4} \cdot \mathbf{1}_2 \ovee \frac{1}{2} \cdot \mathbf{1}_3 \ovee \frac{3}{4} \cdot \mathbf{1}_4 \ovee 1 \cdot \mathbf{1}_5.$$

The question is what we learn about the bag distribution after observing this predicate 10 consecutive times? Figure 20.1 in [35] gives a plot of (Bayesian) internal learning along the channel c; it is reconstructed in Figure 3 via internal learning.

[2] The bags are described in [35] as very large, so that withdrawing one candy does not change the distribution of candies in the bag. This amounts to replacing the drawn candy.

$$
\begin{aligned}
&\mathit{Ilrn}(\omega, c, n|\,L\,\rangle) \\
&\quad = \omega\big|_{(c \ll \mathbf{1}_L)^n}
\end{aligned}
$$

for $n = 0, 1, \ldots, 10$.

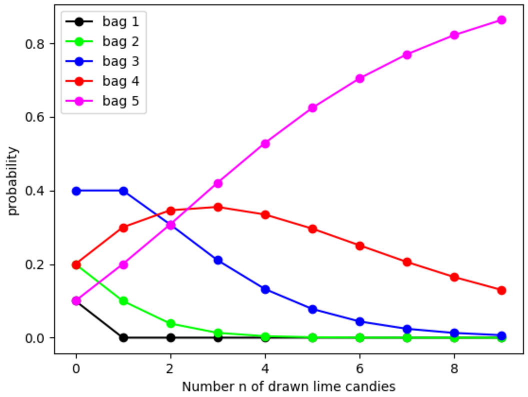


Fig. 3. Bag distributions, aligned vertically, after multiple lime candy draws.

This leads for $n = 1, 2, 3$ to distributions of bags:

$$
\begin{aligned}
\mathit{Ilrn}(\omega, c, 1|\,L\,\rangle) &= \tfrac{1}{10}|\,2\,\rangle + \tfrac{2}{5}|\,3\,\rangle + \tfrac{3}{10}|\,4\,\rangle + \tfrac{1}{5}|\,5\,\rangle \\
\mathit{Ilrn}(\omega, c, 2|\,L\,\rangle) &= \tfrac{1}{26}|\,2\,\rangle + \tfrac{4}{13}|\,3\,\rangle + \tfrac{9}{26}|\,4\,\rangle + \tfrac{4}{13}|\,5\,\rangle \\
&\approx 0.0385|\,2\,\rangle + 0.308|\,3\,\rangle + 0.346|\,4\,\rangle + 0.308|\,5\,\rangle \\
\mathit{Ilrn}(\omega, c, 3|\,L\,\rangle) &= \tfrac{1}{76}|\,2\,\rangle + \tfrac{4}{19}|\,3\,\rangle + \tfrac{27}{76}|\,4\,\rangle + \tfrac{8}{19}|\,5\,\rangle \\
&\approx 0.0132|\,2\,\rangle + 0.211|\,3\,\rangle + 0.355|\,4\,\rangle + 0.421|\,5\,\rangle.
\end{aligned}
$$

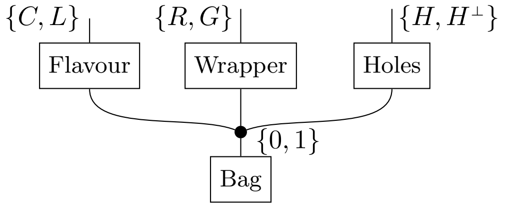
We see, here and in Figure 3, that bag 5 quickly becomes more likely — as expected because it contains most lime candies — and that bag 1 is impossible after drawing the first lime.

As an aside, applying external learning in this candy situation gives, for $n \geq 1$,

$$
\mathit{Elrn}(\omega, c, n|\,L\,\rangle) = \mathit{Elrn}(\omega, c, 1|\,L\,\rangle) = \omega|_{c \ll \mathbf{1}_L}.
$$

The outcome is then the same for each number $n > 0$ of drawn lime candies: multiple lime-draws give no further information, as in (20). This shows that external learning is not appropriate here, in the first candy example.

The second candy example in [35, §20.3] is used as an illustration of the Expectation-Maximisation (EM) algorithm. Therefor it now uses *external* learning. It involves the Bayesian network described below, with (two) bags of candies, named 0 and 1, each described by three features, namely their flavour (cherry or lime), their wrapper (red or green), and whether or not they have holes. The interpretation of the Bayesian network in terms of channels (conditional probability tables) $f\colon \{0,1\} \rightsquigarrow \{C, L\}$, $w\colon \{0,1\} \rightsquigarrow \{R, G\}$, $h\colon \{0,1\} \rightsquigarrow \{H, H^{\perp}\}$, and initial state $\rho \in \mathcal{D}(\{0,1\})$ is on the left.



with

$$\begin{cases} f(0) \;=\; \frac{6}{10}|C\rangle + \frac{4}{10}|L\rangle \\ f(1) \;=\; \frac{4}{10}|C\rangle + \frac{6}{10}|L\rangle \\ w(0) \;=\; \frac{6}{10}|R\rangle + \frac{4}{10}|G\rangle \\ w(1) \;=\; \frac{4}{10}|R\rangle + \frac{6}{10}|G\rangle \\ h(0) \;=\; \frac{6}{10}|H\rangle + \frac{4}{10}|H^{\perp}\rangle \\ h(1) \;=\; \frac{4}{10}|H\rangle + \frac{6}{10}|H^{\perp}\rangle \\ \rho \;=\; \frac{6}{10}|0\rangle + \frac{4}{10}|1\rangle. \end{cases}$$

The three channels f, w, h are combined into a single (three-)tuple channel $\langle f, w, h\rangle \colon \{0,1\} \rightsquigarrow \{C, L\} \times \{R, G\} \times \{H, H^{\perp}\}$. At 0 it is:

$$\begin{array}{rl} \langle f, w, h\rangle(0) & = \; f(0) \otimes w(0) \otimes h(0) \\ = & \frac{216}{1000}|C, R, H\rangle + \frac{144}{1000}|C, R, H^{\perp}\rangle + \frac{144}{1000}|C, G, H\rangle + \frac{96}{1000}|C, G, H^{\perp}\rangle \\ & \quad + \frac{144}{1000}|L, R, H\rangle + \frac{96}{1000}|L, R, H^{\perp}\rangle + \frac{96}{1000}|L, G, H\rangle + \frac{64}{1000}|L, G, H^{\perp}\rangle. \end{array}$$

The point-data $\psi \in \mathcal{M}\big(\{C, L\} \times \{R, G\} \times \{H, H^{\perp}\}\big)$ is given by the multiset:

$$\begin{array}{rl} \psi \;=\; & 273|C, R, H\rangle + 93|C, R, H^{\perp}\rangle + 104|C, G, H\rangle + 90|C, G, H^{\perp}\rangle \\ & \quad + 79|L, R, H\rangle + 100|L, R, H^{\perp}\rangle + 94|L, G, H\rangle + 167|L, G, H^{\perp}\rangle, \end{array}$$

containing $\|\psi\| = 1000$ items. We are now set to learn a better state and channel, via EM as described in [35]. Here we use the description of external learning with a dagger channel, from Proposition 23. The newly learned distribution on $\{0,1\}$ is:

$$\begin{array}{rcl} \mathit{Elrn}(\rho, \langle f, w, h\rangle, \psi) & = & \langle f, w, h\rangle^{\dagger}_{\rho} \gg \mathit{Flrn}(\psi) \\ & = & \frac{273}{1000} \cdot \rho|_{\langle f,w,h\rangle \ll \mathbf{1}_{(C,R,H)}} + \cdots + \frac{167}{1000} \cdot \rho|_{\langle f,w,h\rangle \ll \mathbf{1}_{(L,G,H^{\perp})}} \\ & = & \frac{30891}{50440}|0\rangle + \frac{19549}{50440}|1\rangle \\ & \approx & 0.6124|0\rangle + 0.3876|1\rangle. \end{array}$$

This probability 0.6124 is exactly as computed in [35, §20.3]. The newly learned channel is obtained like in Theorem 26 (3) as a 'double dagger', which we abbreviate as:

$$dd \coloneqq \big(\langle f, w, h\rangle^{\dagger}_{\rho}\big)^{\dagger}_{\mathit{Flrn}(\psi)} \,:\, \{0,1\} \rightsquigarrow \{C, L\} \times \{R, G\} \times \{H, H^{\perp}\}.$$

We then obtain the individually learned channels f', w', h' via marginalisation of the channels:

$$\begin{array}{rcl} f' & \coloneqq & \pi_1 \circ dd \;:\; \{0,1\} \rightsquigarrow \{C, L\} \\ w' & \coloneqq & \pi_2 \circ dd \;:\; \{0,1\} \rightsquigarrow \{R, G\} \\ h' & \coloneqq & \pi_3 \circ dd \;:\; \{0,1\} \rightsquigarrow \{H, H^{\perp}\} \end{array}$$

This yields precisely the values reported in [35]:

$$\begin{aligned}
f'(0) &= 0.6684|C\rangle + 0.3316|L\rangle & \quad f'(1) &= 0.3887|C\rangle + 0.6113|L\rangle \\
w'(0) &= 0.6483|R\rangle + 0.3517|G\rangle & \quad w'(1) &= 0.3817|R\rangle + 0.6183|G\rangle \\
h'(0) &= 0.6558|H\rangle + 0.3442|H^{\perp}\rangle & \quad h'(1) &= 0.3827|H\rangle + 0.6173|H^{\perp}\rangle.
\end{aligned}$$

In two adjacent sections on learning, §20.1 and §20.3, in the same textbook [35], two different learning methods are used, for similar examples (bags of candies). The book makes neither that difference explicit, nor what is actually improved (like increase of some form of likelihood) by these different forms of learning.

9 Discussion about likelihood and learning

In (likelihood-based) probabilistic learning one seeks a distribution (state) that better fits given data. In this paper we have argued that such data form multisets, of points, or, more generally, of predicates. These data give rise to a likelihood function on states, assigning a numerical value in $[0, 1]$, to a state. Learning may happen in multiple steps, where each step increases the likelihood of the data, by changing a given state ω to ω' which fits better, in the sense that it gives higher likelihood to the data.

This paper has described two likelihood functions, namely the 'external' version $\models_{\mathrm{E}}$ and the 'internal' version $\models_{\mathrm{I}}$, with two associated learning methods. In Section 5 it is shown that both forms of likelihood arise naturally from repeated transitions on states, with or without updates. Both forms of learning are used in the literature, but implicitly: the difference is not made explicit — as far as we have seen.

This final section tries to develop a perspective on this matter, with as underlying question: when, under which circumstances, should we use external likelihood and external learning and when internal likelihood and internal learning? No mathematically precise answer is formulated. Instead, an intuition is developed, see esp. points (6) and (7) below, in terms of a combination of external and internal, using batches of data that can be handled separately externally, and jointly internally.

It is unlikely that this admittedly vague answer will settle the matter. Therefor the points below are best seen as a first step in further research and debate.

1. In our approach we have consistently used fuzzy predicates, taking values in $[0, 1]$. This is unusual in probability theory (with exceptions *e.g.* in [13,16,32,36]), where people standardly use sharp predicates (with values in $\{0, 1\}$), also called events. Conjunction of events is simply intersection: $\mathbf{1}_U \;\&\; \mathbf{1}_V = \mathbf{1}_{U \cap V}$ and taking powers of events has no effect: $(\mathbf{1}_U)^n = \mathbf{1}_U \;\&\; \cdots \;\&\; \mathbf{1}_U = \mathbf{1}_U$. Thus, the internal likelihood formulation — $\omega \models_{\mathrm{I}} \sum_i n_i|p_i\rangle = \omega \models \&_i \, p_i^{n_i}$ — only really makes sense in a context with fuzzy predicates p_i. This might explain why internal likelihood has

not been made explicit before, and then also why the distinction between external and internal likelihood is absent in the literature.

2. Let's make things concrete and recall the coin bias learning situation in Section 7, with uniform state υ on the discretised unit interval D. The probability of seeing head (or tail) is:

$$\upsilon \models \mathit{flip} \ll \mathbf{1}_H \;=\; \upsilon \models \mathit{flip} \ll \mathbf{1}_T \;=\; \tfrac{1}{2}.$$

Now suppose we have data saying: both head and tail. How should this be interpreted? What is the likelihood of these data? We formalise it as a multiset of predicates $\psi = 1|\mathit{flip} \ll \mathbf{1}_H\rangle + 1|\mathit{flip} \ll \mathbf{1}_T\rangle$. Then:

$$\begin{aligned} \upsilon \models_{\scriptscriptstyle E} \psi &= (\upsilon \models \mathit{flip} \ll \mathbf{1}_H) \cdot (\upsilon \models \mathit{flip} \ll \mathbf{1}_T) \;=\; \tfrac{1}{2} \cdot \tfrac{1}{2} \;=\; 0.25 \\ \upsilon \models_{\scriptscriptstyle I} \psi &= \upsilon \models (\mathit{flip} \ll \mathbf{1}_H) \;\&\; (\mathit{flip} \ll \mathbf{1}_T) \\ &= \sum_{0 \leq i \leq 20} \tfrac{1}{21} \cdot \tfrac{i}{20} \cdot (1 - \tfrac{i}{20}) \;=\; \tfrac{1}{2} - \tfrac{41}{120} \;=\; \tfrac{19}{120} \;\approx\; 0.16. \end{aligned}$$

What is now the 'right' likelihood of seeing both head and tail: 25% or 16%? This question challenges our basic probabilistic intuitions. The internal perspective offers a reasonable interpretation by Bayes' rule: both head and tail means, first seeing head, and updating, and then seeing tail (or the other way around):

$$\begin{aligned} &\big(\upsilon \models \mathit{flip} \ll \mathbf{1}_H\big) \cdot \big(\upsilon|_{\mathit{flip} \ll \mathbf{1}_H} \models \mathit{flip} \ll \mathbf{1}_T\big) \\ &\quad= \upsilon \models (\mathit{flip} \ll \mathbf{1}_H) \;\&\; (\mathit{flip} \ll \mathbf{1}_T) \\ &\quad= \big(\upsilon \models \mathit{flip} \ll \mathbf{1}_T\big) \cdot \big(\upsilon|_{\mathit{flip} \ll \mathbf{1}_T} \models \mathit{flip} \ll \mathbf{1}_H\big). \end{aligned}$$

3. Maybe the fuzzy predicates in the previous example over-complicate the situation. So let's move to sharp predicates: we take a fair dice $\omega = \frac{1}{6}|1\rangle + \frac{1}{6}|2\rangle + \frac{1}{6}|3\rangle + \frac{1}{6}|4\rangle + \frac{1}{6}|5\rangle + \frac{1}{6}|6\rangle$, with events $E = \{2, 4, 6\}$ for 'even' and $H = \{4, 5, 6\}$ for 'high', corresponding to sharp predicates $\mathbf{1}_E$ and $\mathbf{1}_H$. Clearly, $\omega \models \mathbf{1}_E = \omega \models \mathbf{1}_H = \frac{1}{2}$. We take as data $\phi = 1|\mathbf{1}_E\rangle + 2|\mathbf{1}_H\rangle$, representing that we observe 'even' once and 'high' twice. What is ϕ's likelihood in state ω?

$$\begin{aligned} \omega \models_{\scriptscriptstyle E} \phi &= (\omega \models \mathbf{1}_E) \cdot (\omega \models \mathbf{1}_H) \cdot (\omega \models \mathbf{1}_H) \;=\; \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} \;=\; \tfrac{1}{8} \\ \omega \models_{\scriptscriptstyle I} \phi &= \omega \models \mathbf{1}_E \;\&\; \mathbf{1}_H \;\&\; \mathbf{1}_H \;=\; \omega \models \mathbf{1}_{E \cap H} \;=\; \tfrac{1}{3}. \end{aligned}$$

What is now the right likelihood? It depends ... But on what?

4. We can try to explain the difference between $\models_{\scriptscriptstyle E}$ and $\models_{\scriptscriptstyle I}$ in terms of different observers, who operate *separately* or *jointly*, like in Example 14. Suppose we have a state ω and data in the form of a multiset of predicates $\psi = \sum_i n_i|p_i\rangle$, with $n = \|\psi\| = \sum_i n_i$. We assume that there are n individual observers, and each predicate occurring in the multiset ψ is assigned to one observer. Hence n_1 of them obtain p_1, n_2 observers have p_2 etc.

- In the external likelihood perspective each observer, say with predicate p, gets an instance of the state ω, via some 'external' copy mechanism. These observers now ask: what is the probability that we are all right *separately*? Each of them determines the validity $\omega \models p$ of their own predicate. Then, all observers get together and multiply their validities, giving the external likelihood $\omega \models_{\mathrm{E}} \psi$.
- In the internal likelihood perspective each observer is looking at the same state ω. These observers ask: what is the probability that we are *jointly* right? This joint view is obtained by putting all their predicates together in a single conjunction $\&_i\, p_i^{n_i}$. The validity of this conjunction predicate gives internal likelihood $\omega \models_{\mathrm{I}} \psi$.

5. In the approach to quantum logic described in [21] the operation & on predicates is called *sequential* conjunction, whereas $\otimes$ is *parallel* conjunction. This phrase 'sequential' makes sense there, since & is not commutative in a quantum setting. One could use the term 'sequential' also in the current setting of classical (non-quantum) probability, for instance by understanding 'joint' in the previous point in a sequential manner — one after the other — but then in such a way that the order does not matter.

6. One can combine external and internal likelihood by moving one more step up the abstraction ladder and introduce multisets of multisets of predicates $\Psi \in \mathcal{M}(\mathcal{M}(\mathrm{Pred}(X)))$ as data. This may be useful when data in $\mathcal{M}(\mathrm{Pred}(X))$, as used before, comes in batches: the multisets $\phi \in \mathrm{Pred}(\mathcal{M}(X))$ occurring as elements of Ψ. For a state $\omega \in \mathcal{D}(X)$ one can now define external-internal likelihood $\models_{\mathrm{EI}}$ as:

$$\omega \models_{\mathrm{EI}} \Psi \;:=\; \sum\nolimits_{\phi} \big(\omega \models_{\mathrm{I}} \phi\big)^{\Psi(\phi)} \;=\; \sum\nolimits_{\phi} \big(\omega \models \&_p\, p^{\phi(p)}\big)^{\Psi(\phi)}.$$

This uses external likelihood on the outside and internal likelihood on the inside. A different order does not make sense. One can then develop an associated form of external-internal learning. This actually occurs in the literature, namely in the leading example used in [18] to describe Expectation-Maximisation. Elaborating the details goes beyond the current setting, but we can briefly sketch the essentials (of a single E-step).

Two coins are given in [18] via a channel $c\colon \{1,2\} \mathrel{\circ\!\!\rightarrow} \{H,T\}$ with different biases: $c(1) = \frac{3}{5}|H\rangle + \frac{2}{5}|T\rangle$ and $c(2) = \frac{1}{2}|H\rangle + \frac{1}{2}|T\rangle$. There are 5 batches of point data with 10 coin flips each: $\phi_1 = 5|H\rangle + 5|T\rangle$, $\phi_2 = 9|H\rangle + 1|T\rangle$, $\phi_3 = 8|H\rangle + 2|T\rangle$, $\phi_4 = 4|H\rangle + 6|T\rangle$, $\phi_5 = 7|H\rangle + 3|T\rangle$. Starting from the uniform state $\upsilon \in \mathcal{D}(\{1,2\})$ one learns a better fit via 'external-internal' learning: take externally the weighted average (convex sum) over all batches, of the internally learned states per batch, giving:

$$\sum\nolimits_i \tfrac{1}{5} \cdot \mathit{Ilrn}(\upsilon, c, \phi_i) \;=\; 0.597|1\rangle + 0.403|2\rangle.$$

Thus, these data indicate that the first coin is a bit more likely.

7. The previous point, with combined likelihood $\models_{\mathrm{EI}}$, can offer a perspective on the question when to use external or internal likelihood. When we have data $\psi \in \mathcal{M}(\mathrm{Pred}(X))$ we can view it in two ways:

   – as multiset of batches of *separate* single data items $\psi_E := \sum_p \psi(p)\big|\, 1|\,p\,\rangle\,\big\rangle$ in $\mathcal{M}(\mathcal{M}(\mathrm{Pred}(X)))$, whose external-internal likelihood equals external likelihood: $\omega \models_{\mathrm{EI}} \psi_E \;=\; \omega \models_{\mathrm{E}} \psi$;

   – as a single batch of *joint* data items $\psi_I := 1|\,\psi\,\rangle$ in $\mathcal{M}(\mathcal{M}(\mathrm{Pred}(X)))$, with external-internal likelihood equal to internal likelihood: $\omega \models_{\mathrm{EI}} \psi_I \;=\; \omega \models_{\mathrm{I}} \psi$.

   This distinction is consistent with the one we made earlier in point (4) in terms of observers.
   Concretely, in the earlier setting of coin bias learning, suppose we have two separate batches of data, in the form of two multisets of predicates:

$$\phi_1 \;=\; 1|\,\mathit{flip} \ll \mathbf{1}_H\,\rangle + 1|\,\mathit{flip} \ll \mathbf{1}_T\,\rangle \qquad \phi_2 \;=\; 2|\,\mathit{flip} \ll \mathbf{1}_H\,\rangle + 1|\,\mathit{flip} \ll \mathbf{1}_T\,\rangle.$$

   Then $\varPsi = 1|\,\phi_1\,\rangle + 1|\,\phi_2\,\rangle$ has likelihood (in the uniform state υ):

$$\upsilon \models_{\mathrm{EI}} \varPsi \;=\; (\upsilon \models_{\mathrm{I}} \phi_1) \cdot (\upsilon \models_{\mathrm{I}} \phi_2) \;=\; \tfrac{19}{120} \cdot \tfrac{19}{240} \;=\; \tfrac{361}{28800}.$$

8. Along the way we have noticed several times — *e.g.* in (8) and (20) — that frequentist and external learning satisfy the more-is-the-same property: repeating the same data as input has no effect. In contrast, internal learning behaves like an action — see Lemma 10 (1) and Propositions 19 and 22 — where repeated (and multiple) data inputs do have effect and are processed sequentially via multiple Bayesian updates. This is a significant difference between 'internal' and 'external'.

9. Finally, we like to point at an analogy with the distinction between Pearl's updating and Jeffrey's adaptation along a channel, as described in [24]. The description of the externally learned state $e_\omega^\dagger \gg \mathit{Flrn}(\phi)$ via a dagger channel in Proposition 23 is an instance of Jeffrey's adaptation rule. Internally learning along a channel is an instance of Pearl's updating rule. In follow-up work it will be shown that Jeffrey's rule is about decreasing divergence and Pearl's rule is about increasing validity. Proposition 20 shows that external learning can also be expressed in terms of decreasing divergence. Hence it seems that external learning and Jeffrey's rule belong to the same "decreasing divergence" school, whereas internal learning and Pearl's rule are in the "increasing likelihood" school.

References

1. S. Abramsky. No-cloning in categorical quantum mechanics. In S. Gay and I. Mackie, editors, *Semantical Techniques in Quantum Computation*, pages 1–28. Cambridge Univ. Press, 2010.

2. S. Abramsky. Coalgebras, Chu spaces, and representations of physical systems. *Journ. Phil. Logic*, 42(3):551–574, 2013.
3. S. Abramsky. Contextual semantics: From quantum mechanics to logic, databases, constraints, and complexity. *EATCS Bulletin*, 113, 2014.
4. S. Abramsky. Arrow's theorem by Arrow theory. In Å. Hirvonen, J. Kontinen, R. Kossak, and A. Villaveces, editors, *Logic Without Borders – Essays on Set Theory, Model Theory, Philosophical Logic and Philosophy of Mathematics*, pages 15–30. De Gruyter, 2015.
5. S. Abramsky, R. Barbosa, M. Karvonen, and S. Mansfield. A comonadic view of simulation and quantum resources. In *Logic in Computer Science*, pages 1–12. IEEE, 2019.
6. S. Abramsky, R. Blute, and P. Panangaden. Nuclear and trace ideals in tensored *-categories. *Journ. of Pure & Appl. Algebra*, 143:3–47, 2000.
7. S. Abramsky and A. Brandenburger. The sheaf-theoretic structure of non-locality and contextuality. *New Journ. of Physics*, 13:113036, 2011.
8. S. Abramsky and B. Coecke. A categorical semantics of quantum protocols. In K. Engesser, Dov M. Gabbay, and D. Lehmann, editors, *Handbook of Quantum Logic and Quantum Structures: Quantum Logic*, pages 261–323. North-Holland, Elsevier, Computer Science Press, 2009.
9. S. Abramsky and C. Heunen. H*-algebras and nonunital Frobenius algebras: first steps in infinite-dimensional categorical quantum mechanics. *Clifford Lectures, AMS Proceedings of Symposia in Applied Mathematics*, 71:1–24, 2012.
10. L. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statistics*, 41:164–171, 1970.
11. C. Bishop. *Pattern Recognition and Machine Learning.* Information Science and Statistics. Springer, 2006.
12. P. Bruza and S. Abramsky. Probabilistic programs: Contextuality and relational database theory. In *Quantum Interaction*, pages 163–174, 2016.
13. H. Chan and A. Darwiche. On the revision of probabilistic beliefs using uncertain evidence. *Artif. Intelligence*, 163:67–90, 2005.
14. K. Cho and B. Jacobs. Disintegration and Bayesian inversion via string diagrams. *Math. Struct. in Comp. Sci.*, 29(7):938–971, 2019.
15. F. Clerc, F. Dahlqvist, V. Danos, and I. Garnier. Pointless learning. In J. Esparza and A. Murawski, editors, *Foundations of Software Science and Computation Structures*, number 10203 in Lect. Notes Comp. Sci., pages 355–369. Springer, Berlin, 2017.
16. A. Darwiche. *Modeling and Reasoning with Bayesian Networks.* Cambridge Univ. Press, 2009.
17. A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journ. Royal Statistical Soc.*, 39(1):1–38, 1977.
18. C. Do and S. Batzoglou. What is the expectation maximization algorithm? *Nature Biotechnology*, 26:897–899, 2008.
19. T. Fritz. A synthetic approach to Markov kernels, conditional independence, and theorems on sufficient statistics. *Advances in Math.*, 370:107239, 2020.
20. M. Giry. A categorical approach to probability theory. In B. Banaschewski, editor, *Categorical Aspects of Topology and Analysis*, number 915 in Lect. Notes Math., pages 68–85. Springer, Berlin, 1982.
21. B. Jacobs. New directions in categorical logic, for classical, probabilistic and quantum logic. *Logical Methods in Comp. Sci.*, 11(3), 2015.

22. B. Jacobs. From probability monads to commutative effectuses. *Journ. of Logical and Algebraic Methods in Programming*, 94:200–237, 2018.
23. B. Jacobs. Learning along a channel: the Expectation part of Expectation-Maximisation. In B. König, editor, *Math. Found. of Programming Semantics*, number 347 in Elect. Notes in Theor. Comp. Sci., pages 143–160. Elsevier, Amsterdam, 2019.
24. B. Jacobs. The mathematics of changing one's mind, via Jeffrey's or via Pearl's update rule. *Journ. of Artif. Intelligence Research*, 65:783–806, 2019.
25. B. Jacobs. A channel-based perspective on conjugate priors. *Math. Struct. in Comp. Sci.*, 30(1):44–61, 2020.
26. B. Jacobs and F. Zanasi. The logical essentials of Bayesian reasoning. In G. Barthe, J.-P. Katoen, and A. Silva, editors, *Foundations of Probabilistic Programming*, pages 295–331. Cambridge Univ. Press, 2021.
27. F. Jensen and T. Nielsen. *Bayesian Networks and Decision Graphs.* Statistics for Engineering and Information Science. Springer, 2^{nd} rev. edition, 2007.
28. D. Koller and N. Friedman. *Probabilistic Graphical Models. Principles and Techniques.* MIT Press, Cambridge, MA, 2009.
29. D. Kozen. Semantics of probabilistic programs. *Journ. Comp. Syst. Sci*, 22(3):328–350, 1981.
30. D. Kozen. A probabilistic PDL. *Journ. Comp. Syst. Sci*, 30(2):162–178, 1985.
31. G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Wiley, New York, 1997.
32. A. Mrad, V. Delcroix, S. Piechowiak, P. Leicester, and M. Abid. An explication of uncertain evidence in Bayesian networks: likelihood evidence and probabilistic evidence. *Applied Intelligence*, 23(4):802–824, 2015.
33. H. Pishro-Nik. *Introduction to probability, statistics, and random processes.* Kappa Research LLC, 2014. Available at `https://www.probabilitycourse.com`.
34. S. Ross. *A first course in probability.* Pearson Education, tenth edition edition, 2018.
35. S. Russell and P. Norvig. *Artificial Intelligence. A Modern Approach.* Prentice Hall, Englewood Cliffs, NJ, 2003.
36. M. Valtorta, Y.-G. Kim, and J. Vomlel. Soft evidential update for probabilistic multiagent systems. *Int. Journ. of Approximate Reasoning*, 29(1):71–106, 2002.

A Appendix

We provide the missing proofs of Theorem 18 (1) and of Proposition 16. The proof of the latter proposition is standard, but is included because it forms a proper preparation for the proofs of the two theorems — which are new results. All proofs rely on some basic real analysis for finding the maximum of functions with constraints on their inputs. This is done via the Lagrange multiplier method, see *e.g.* [11, §2.2]. This will be illustrated first. Subsequently we make use of a 'sum-increase' lemma to prove the other results.

Proof. (of Proposition 16) Let $\phi \in \mathcal{M}(X)$ be a fixed non-empty multiset. We need to prove that the external likelihood function $(-) \models_{\mathrm{E}} \phi \colon \mathcal{D}(X) \to [0,1]$ takes its maximum at $\mathit{Flrn}(\phi)$. We will thus seek the maximum of the function

$\omega \mapsto \omega \models_{\mathrm{E}} \phi$ by taking the derivative with respect to $\omega \in \mathcal{D}(X)$. We will work with the 'log-validity', that is, with the function $\omega \mapsto \ln(\omega \models_{\mathrm{E}} \phi)$, where ln is the monotone (natural) logarithm function. It reduces the product $\prod$ of powers in the definition of $\models_{\mathrm{E}}$ to a sum $\sum$ of multiplications.

Assume that the support of $\phi = \sum_i n_i|x_i\rangle$ is $\{x_1, \ldots, x_n\} \subseteq X$. We look at distributions $\omega \in \mathcal{D}(\{x_1, \ldots, x_n\})$; they may be identified with numbers $\vec{v} = v_1, \ldots, v_n \in \big(\mathbb{R}_{\geq 0}\big)^n$ with $\sum_i v_i = 1$. We thus seek the maximum of the log-validity function:

$$k(\vec{v}) \coloneqq \ln\Big(\sum_i v_i|x_i\rangle \models_{\mathrm{E}} \sum_i n_i|x_i\rangle\Big) = \ln\Big(\prod_i v_i^{n_i}\Big) = \sum_i n_i \cdot \ln(v_i).$$

Since we have a constraint $(\sum_i v_i) - 1 = 0$ on the inputs, we can use the Lagrange multiplier method for finding the maximum. We thus take another parameter λ in a new function:

$$K(\vec{v}, \lambda) \coloneqq k(\vec{v}) - \lambda \cdot \big((\sum_i v_i) - 1\big) = \big(\sum_i n_i \ln(v_i)\big) - \lambda \cdot \big((\sum_i v_i) - 1\big).$$

The partial derivatives of K are:

$$\frac{\partial K}{\partial v_i}(\vec{v}, \lambda) = \frac{n_i}{v_i} - \lambda \qquad\qquad \frac{\partial K}{\partial \lambda}(\vec{v}, \lambda) = 1 - \sum_i v_i.$$

Setting all of these to 0 and solving gives the required maximum. First, we have:

$$1 = \sum_i v_i = \sum_i \frac{n_i}{\lambda} = \frac{\sum_i n_i}{\lambda}.$$

Hence $\lambda = \sum_i n_i$ and thus:

$$v_i = \frac{n_i}{\lambda} = \frac{n_i}{\sum_i n_i} \overset{(7)}{=} \mathit{Flrn}(\phi)(x_i). \qquad \square$$

We now come to an auxiliary result which we shall call the sum-increase lemma. It is a special (discrete) case of a more general result [10, Thm. 2.1]. It describes how to find increases for sum expressions in general.

Lemma 27. *Let X, Y be finite sets, and let $F\colon X \times Y \to \mathbb{R}_{\geq 0}$ be a given function. For each $x \in X$, write $F_1(x) \coloneqq \sum_{y \in Y} F(x, y)$ for the sum that we wish to increase. Assume that there is an $x' \in X$ with:*

$$x' = \operatorname*{argmax}_z G(x, z) \qquad \textit{where} \qquad G(x, z) \coloneqq \sum_{y \in Y} F(x, y) \cdot \ln\big(F(z, y)\big).$$

Then $F_1(x') \geq F_1(x)$.

The proof uses Jensen's inequality: for $a_1, \ldots, a_n \in \mathbb{R}_{>0}$ and $r_1, \ldots, r_n \in [0, 1]$ with $\sum_i r_i = 1$ one has $\ln(\sum_i r_i a_i) \geq \sum_i r_i \ln(a_i)$. This gives a strict increase, except in 'corner' cases. The same holds for the above sum-increase lemma. The actual maximum x' in that lemma can in many situation be determined analytically — using the Lagrange multiplier method — but it need not be unique.

Proof. Let x' be the element where $G(x,-)\colon Y \to \mathbb{R}_{\geq 0}$ takes its maximum. This x' satisfies $F_1(x') \geq F_1(x)$, since:

$$\begin{aligned}
\ln\left(\frac{F_1(x')}{F_1(x)}\right) &= \ln\left(\sum\nolimits_y \frac{F(x',y)}{F_1(x)}\right) \\
&= \ln\left(\sum\nolimits_y \frac{F(x,y)}{F_1(x)} \cdot \frac{F(x',y)}{F(x,y)}\right) \\
&\geq \sum\nolimits_y \frac{F(x,y)}{F_1(x)} \cdot \ln\left(\frac{F(x',y)}{F(x,y)}\right) \qquad \text{by Jensen's inequality} \\
&= \frac{1}{F_1(x)} \cdot \sum\nolimits_y F(x,y) \cdot \Big(\ln\big(F(x',y)\big) - \ln\big(F(x,y)\big)\Big) \\
&= \frac{1}{F_1(x)} \cdot \Big(G(x,x') - G(x,x)\Big) \;\geq\; 0.
\end{aligned}$$

□

Proof. (Theorem 18 (1)) Let $\omega \in \mathcal{D}(X)$ be state on a finite set X and let $p_1, \ldots, p_n$ be predicates on X, all with non-zero validity $\omega \models p_i$. We claim that the state $\omega' = \sum_i \frac{1}{n} \cdot \omega|_{p_i}$ then satisfies:

$$\prod\nolimits_i (\omega' \models p_i) \;\geq\; \prod\nolimits_i (\omega \models p_i). \tag{25}$$

The inequality in Theorem 18 (1) is a direct consequence of (25). We shall prove (25) for $n = 2$. The generalisation to arbitrary n should then be obvious, but involves much more book-keeping of additional variables.

We use Lemma 27 with function $F\colon \mathcal{D}(X) \times X \times X \to \mathbb{R}_{\geq 0}$ given by:

$$F(\omega,x,y) \coloneqq \omega(x) \cdot p_1(x) \cdot \omega(y) \cdot p_2(y).$$

Then by distributivity of multiplication over addition:

$$\sum\nolimits_{x,y} F(\omega,x,y) \;=\; \big(\sum\nolimits_x \omega(x) \cdot p_1(x)\big) \cdot \big(\sum\nolimits_y \omega(y) \cdot p_2(y)\big) \;=\; (\omega \models p_1) \cdot (\omega \models p_2).$$

Let $X = \{x_1, \ldots, x_n\}$ and let the function H be given by:

$$H(\vec{v},\lambda) \coloneqq \sum\nolimits_{i,j} F(\omega,x_i,x_j) \cdot \ln\big(v_i \cdot p_1(x_i) \cdot v_j \cdot p_2(x_j)\big) \;-\; \lambda \cdot \big((\sum\nolimits_i v_i) - 1\big).$$

Then:

$$\frac{\partial H}{\partial v_k}(\vec{v},\lambda) \;=\; \sum\nolimits_i \frac{F(\omega,x_k,x_i) + F(\omega,x_i,x_k)}{v_k} \;-\; \lambda \qquad \frac{\partial H}{\partial \lambda}(\vec{v},\lambda) \;=\; 1 - \sum\nolimits_i v_i.$$

Setting these to zero gives:

$$1 \;=\; \sum\nolimits_k v_k \;=\; \frac{\sum_{k,i} F(\omega,x_k,x_i) + F(\omega,x_i,x_k)}{\lambda} \;=\; \frac{2 \cdot (\omega \models p_1) \cdot (\omega \models p_2)}{\lambda}.$$

Hence $\lambda = 2 \cdot (\omega \models p_1) \cdot (\omega \models p_2)$ so that:

$$\begin{aligned} v_k &= \frac{\sum_i F(\omega, x_k, x_i) + F(\omega, x_i, x_k)}{\lambda} \\ &= \tfrac{1}{2} \cdot \frac{\omega(x_k) \cdot p_1(x_k) \cdot (\omega \models p_2)}{(\omega \models p_1) \cdot (\omega \models p_2)} + \tfrac{1}{2} \cdot \frac{(\omega \models p_1) \cdot \omega(x_k) \cdot p_2(x_k)}{(\omega \models p_1) \cdot (\omega \models p_2)} \\ &= \tfrac{1}{2} \cdot \frac{\omega(x_k) \cdot p_1(x_k)}{\omega \models p_1} + \tfrac{1}{2} \cdot \frac{\omega(x_k) \cdot p_2(x_k)}{\omega \models p_2} \\ &= \tfrac{1}{2} \cdot \omega|_{p_1}(x_k) + \tfrac{1}{2} \cdot \omega|_{p_2}(x_k). \end{aligned}$$

□