Categorical Probability Theory

Macquarie, Sydney, Nov. 5, 2025

Bart Jacobs, Radboud University Nijmegen
https://www.cs.ru.nl/B.Jacobs/   bart@cs.ru.nl

Outline

Where we are, so far

Probability is about counting and measuring

probability-from-counting: draw a ball   **probability-from-measuring: throw a dart**

- there is a $\frac{3}{5}$ chance of drawing a blue ball
- there is $\frac{3}{4}$ chance of throwing a dart in the red circle but not in the blue one — with radiuses 1 and 2

Categorical Probability Theory

- involves the application of categorical techniques to probability theory — uncovering new structure & results
- first steps in 1980s by Lawvere and Giry — involving the Giry monad of continuous distributions on measurable spaces
- new impetus in last 5-10 years through:
  - usage of string diagrams for graphical modeling
  - work on (semantics of) probabilistic programming languages — including higher order
- see e.g. work of Tobias Fritz, Sam Staton (with their teams) & others
- own work resulting in a book "Structured Probabilistic Reasoning"
  - to be published by CUP, with introductory "teaser"
  - see: www.cs.ru.nl/B.Jacobs/PAPERS/ProbabilisticReasoning.pdf
- Today's topic: gentle introduction/overview to its topic & results
  - no "categorical air guitar playing", but connecting to what happens

Why using categorical language in probability?

- **Primary reason**
  - conditional probabilities $p(y|x)$ are Kleisli maps
  - they map an element x to a probability distribution, on y's
  - they form maps $X \to \mathcal{D}(Y)$ or $X \to \mathcal{G}(Y)$
  - for finite/discrete distribution monad $\mathcal{D}$, and continuous distribution "Giry" monad $\mathcal{G}$
  - the Kleisli categories $\mathcal{K}\ell(\mathcal{D})$ and $\mathcal{K}\ell(\mathcal{G})$ are symmetric monoidal, supporting string diagrams
- **Secondary reason**
  - the traditional language of probability theory is horrible
  - everything is called 'p', in many confusing forms
  - (too) much is left implicit; calculation rules are often missing
  - the language is so bad, that basic results have been missed
  - Wittgenstein: "the limits of our language determine the limits of our thinking". A language update is badly needed.

Monads, Kleisli categories and beyond

- Both $\mathcal{D}$ and $\mathcal{G}$ are commutative, affine monads
  - as a result: $\mathcal{K}\ell(\mathcal{D})$ and $\mathcal{K}\ell(\mathcal{G})$ are symmetric monoidal
  - with a final object as tensor unit
- These Kleisli categories are also Markov categories
  - there are copy maps $X \to X \times X$, forming comonoids
  - Kleisli maps that commute with copiers are called deterministic
- Moreover, they are effectuses
  - they have coproducts $+$, which are suitably well-behaved
  - 'predicates' $X \to 1 + 1$ form effect modules (probabilistic analogues of Boolean algebras)

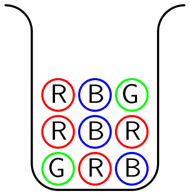
This talk will mostly use discrete probability distributions (via $\mathcal{D}$), to avoid technicalities in the continuous case. Formally, the differences are limited.

Where we are, so far

Drawing in terms of multisets

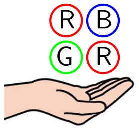
Informally, a multiset is a 'set' in which elements may occur multiple times. Multisets occur frequently in probability theory

- An urn with coloured balls is a multiset, over the colours:



$$= \quad 4|\,R\,\rangle + 3|\,B\,\rangle + 2|\,G\,\rangle$$

- A draw of multiple balls from such an urn is also a multiset



$$= \quad 2|\,R\,\rangle + 1|\,B\,\rangle + 1|\,G\,\rangle$$

One can assign probabilities to such draws, with different outcomes per drawing mode

Multisets, more formally

- We use 'ket' notation to separate multiplicities from elements, as:
$$4|\,R\,\rangle + 3|\,B\,\rangle + 2|\,G\,\rangle$$
- For a set X we write $\mathcal{M}(X)$ for the multisets over X, written as finite formal sums:
$$\textstyle\sum_i n_i|\,x_i\,\rangle \qquad \text{with } n_i \in \mathbb{N} \text{ and } x_i \in X$$
- Alternatively, a multiset is a function $\varphi\colon X \to \mathbb{N}$ with finite support set $supp(\varphi) := \{x \in X \mid \varphi(x) > 0\}$
  - we switch freely between ket & function notation
- The set of multisets $\mathcal{M}(X)$ is the free commutative monoid on X
  - addition of multisets works element-wise
- Write $\|\varphi\| \in \mathbb{N}$ for the size of a multiset, e.g.
$$\Big\|\, 4|\,R\,\rangle + 3|\,B\,\rangle + 2|\,G\,\rangle \,\Big\| = 4 + 3 + 2 = 9.$$
- $\mathcal{M}[K](X) \hookrightarrow \mathcal{M}(X)$ is written for the subset of multisets of size $K \in \mathbb{N}$

From lists to multisets to subsets

- **From lists to multisets**, via "accumulation"
  - Idea: count elements, but forget their order
  - like in: $acc\big(c, b, a, a, b, c\big) = 3|\,a\,\rangle + 2|\,b\,\rangle + 2|\,c\,\rangle$
  - In general: $acc\big(x_1, \ldots, x_n\big) = 1|\,x_1\,\rangle + \cdots + 1|\,x_n\,\rangle$
  - Accumulation preserves size and restricts to $acc\colon X^K \to \mathcal{M}[K](X)$
- **From multisets to subsets**, via "support"
  - Idea: forget multiplicities
  - as in: $supp\big(3|\,a\,\rangle + 2|\,b\,\rangle + 2|\,c\,\rangle\big) = \{a, b, c\}$
  - In general: $supp\Big(\sum_i n_i|\,x_i\,\rangle\Big) = \{x_1, \ldots, x_n\}$, assuming $n_i > 0$

Multisets are 'inbetween' lists and subsets

Comparison of datatypes:

	lists	multisets	subsets
order of elements matters	+	-	-
multiplicity of elements matters	+	+	-

More categorically:

$$\mathcal{L}(X) \xrightarrow[\text{forget order}]{acc} \mathcal{M}(X) \xrightarrow[\text{forget multiplicity}]{supp} \mathcal{P}(X)$$

- $\mathcal{L}$, $\mathcal{M}$, $\mathcal{P}$ are all monads — $\mathcal{M}$, $\mathcal{P}$ commutative, but $\mathcal{L}$ not
- Accumulation and support are maps of monads
- They also preserve the monoid structures

General point: multisets are undervalued and often overlooked / ignored

Basic facts about multisets

- Multisets are not counted via binomial coefficients $\binom{n}{K}$, but via multichoose coefficients $\left(\!\binom{n}{K}\!\right)$: if $|X| = n \geq 1$, then:
$$\Big|\mathcal{M}[K](X)\Big| = \left(\!\!\binom{n}{K}\!\!\right) = \binom{n+K-1}{K} = \frac{(n+K-1)!}{K! \cdot (n-1)!}$$
- For $\varphi \in \mathcal{M}(X)$, the number of lists $\ell \in \mathcal{L}(X)$ with $acc(\ell) = \varphi$ is given by the multiset coefficient $(\varphi) \in \mathbb{N}$, defined as:
$$(\varphi) := \frac{\|\varphi\|!}{\varphi!} \qquad \text{where} \qquad \varphi! := \prod_{x \in X} \varphi(x)!$$
For $\varphi = \sum_i n_i | x_i \rangle$ with $n = \sum_i n_i$ this (φ) is $\binom{n}{n_1, \ldots, n_k}$.
- If $|X| = n$, then:
$$\sum_{\varphi \in \mathcal{M}[K](X)} (\varphi) = n^K.$$

Where we are, so far

Distributions (finite, discrete)

- In a distribution the multiplicities add up to one, as in:
$$coin = \tfrac{49}{100}| H \rangle + \tfrac{51}{100}| T \rangle$$
$$dice = \tfrac{1}{6}| 1 \rangle + \tfrac{1}{6}| 2 \rangle + \tfrac{1}{6}| 3 \rangle + \tfrac{1}{6}| 4 \rangle + \tfrac{1}{6}| 5 \rangle + \tfrac{1}{6}| 6 \rangle$$
- In general, the set $\mathcal{D}(X)$ contains distributions as formal sums $\sum_i r_i | x_i \rangle$ with $r_i \in [0,1]$ satisfying $\sum_i r_i = 1$ and $x_i \in X$.
  - alternatively, a distribution is a function $\omega \colon X \to [0,1]$ with finite support and $\sum_x \omega(x) = 1$
- There is frequentist learning map $Flrn$ turning a (non-empty) multiset into a distribution via normalisation:
$$Flrn\Big(4| R \rangle + 3| B \rangle + 2| G \rangle\Big) = \tfrac{4}{9}| R \rangle + \tfrac{3}{9}| B \rangle + \tfrac{2}{9}| G \rangle.$$

(this $Flrn$ is *not* a map of monads, from non-empty multisets to distributions)

Tensors of distributions

- For two distributions $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$ define their parallel product as tensor $\omega \otimes \rho \in \mathcal{D}(X \times Y)$
  - in functional form as:
$$(\omega \otimes \rho)(x,y) := \omega(x) \cdot \rho(y)$$
  - or, equivalently, in ket form as:
$$\omega \otimes \rho := \sum_{x \in X,\, y \in Y} \omega(x) \cdot \rho(y) \, \big| x, y \big\rangle$$
- For instance, for
$$coin = \tfrac{1}{2}| H \rangle + \tfrac{1}{2}| T \rangle \qquad dice = \sum_{1 \leq i \leq 6} \tfrac{1}{6}| i \rangle$$
tossing them together is captured by the tensor product $coin \otimes dice$:
$$\tfrac{1}{12}\big|H,1\big\rangle + \tfrac{1}{12}\big|H,2\big\rangle + \tfrac{1}{12}\big|H,3\big\rangle + \tfrac{1}{12}\big|H,4\big\rangle + \tfrac{1}{12}\big|H,5\big\rangle + \tfrac{1}{12}\big|H,6\big\rangle +$$
$$+ \tfrac{1}{12}\big|T,1\big\rangle + \tfrac{1}{12}\big|T,2\big\rangle + \tfrac{1}{12}\big|T,3\big\rangle + \tfrac{1}{12}\big|T,4\big\rangle + \tfrac{1}{12}\big|T,5\big\rangle + \tfrac{1}{12}\big|T,6\big\rangle$$

Functoriality of $\mathcal{D}$ (and $\mathcal{M}$)

Each function $f\colon X \to Y$ gives rise to:

- $\mathcal{D}(f)\colon \mathcal{D}(X) \to \mathcal{D}(Y)$ and $\mathcal{M}(f)\colon \mathcal{M}(X) \to \mathcal{M}(Y)$
- Explicitly:
$$\mathcal{D}(f)\Big(\sum_i r_i|x_i\rangle\Big) := \sum_i r_i\big|f(x_i)\big\rangle \qquad \text{and similarly for } \mathcal{M}$$

- Functoriality is used e.g. for marginalisation of a 'joint' distribution $\tau \in \mathcal{D}(X \times Y)$
- Via projections $X \xleftarrow{\pi_1} X \times Y \xrightarrow{\pi_2} Y$ one gets:
$$\mathcal{D}(\pi_1)(\tau) \in \mathcal{D}(X) \qquad \text{and} \qquad \mathcal{D}(\pi_2)(\tau) \in \mathcal{D}(Y)$$
- In general, $\tau \neq \mathcal{D}(\pi_1)(\tau) \otimes \mathcal{D}(\pi_2)(\tau)$
  - In case of equality, τ is called "non-entwined" or "non-entangled" or its parts are "indendent"

Functoriality for shifting and scaling

Let a distribution $\omega \in \mathcal{D}(\mathbb{R})$ on the reals be given, with $s \in \mathbb{R}$.

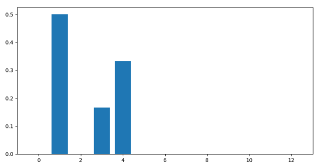
- One can shift ω via:
$$shift(s, \omega) := \mathcal{D}\Big(s + (-)\Big)(\omega) \quad \text{where} \quad s + (-)\colon \mathbb{R} \to \mathbb{R}$$
- Similarly, one can scale ω via:
$$scale(s, \omega) := \mathcal{D}\Big(s \cdot (-)\Big)(\omega) \quad \text{using} \quad s \cdot (-)\colon \mathbb{R} \to \mathbb{R}$$

Shifting and scaling form monoid actions on distributions

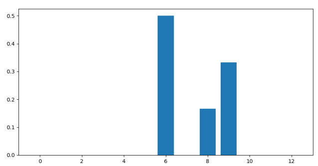
- w.r.t. the additive and multiplicative monoids on the reals

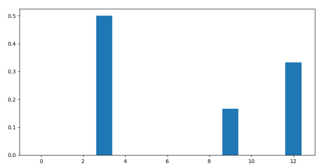
Shift & scale illustrations

Consider as original distribution:



It is shifted by a factor 3 on the left, and scaled by 3 on the right:





A general result about actions

Let M be a monoid, with category of actions $\underline{\mathrm{Act}}_M$.

Theorem

The distribution functor $\mathcal{D}\colon \underline{\mathrm{Sets}} \to \underline{\mathrm{Sets}}$ can be lifted to a functor $\underline{\mathrm{Act}}_M \to \underline{\mathrm{Act}}_M$, also written $\mathcal{D}$, in a commuting diagram:

$$\begin{tikzcd}
\underline{\mathrm{Act}}_M \arrow[r, "\mathcal{D}"] \arrow[d] & \underline{\mathrm{Act}}_M \arrow[d] \\
\underline{\mathrm{Sets}} \arrow[r, "\mathcal{D}"] & \underline{\mathrm{Sets}}
\end{tikzcd}$$

Actually,

- this is a lifting of monads
- this holds much more generally, for a monoid in a symmetric monoidal category and a strong monad on that category
- but it is nice to recognise the relevant structure in shifting/scaling

Sum of dices

- Suppose I have two dices, throw them both, and I want to know the distribution of the sum of the pips.
- We can do this systematically, using categorical notation:
  - throwing both involves the tensor $dice \otimes dice$
  - their sum is obtained via functoriality: $\mathcal{D}(sum)(dice \otimes dice)$
  - The outcome can be calculated easily:

$$\begin{aligned}&\mathcal{D}(sum)\Big(dice \otimes dice\Big)\\ &= \tfrac{1}{36}\big|2\big\rangle + \tfrac{1}{18}\big|3\big\rangle + \tfrac{1}{12}\big|4\big\rangle + \tfrac{1}{9}\big|5\big\rangle + \tfrac{5}{36}\big|6\big\rangle + \tfrac{1}{6}\big|7\big\rangle\\ &\quad + \tfrac{5}{36}\big|8\big\rangle + \tfrac{1}{9}\big|9\big\rangle + \tfrac{1}{12}\big|10\big\rangle + \tfrac{1}{18}\big|11\big\rangle + \tfrac{1}{36}\big|12\big\rangle\end{aligned}$$

- Similarly one may compute: $\mathcal{D}(sum)\Big(dice \otimes dice \otimes dice\Big)$
- The types involved are:
  $dice \in \mathcal{D}(\mathbb{N}) \quad dice^K = dice \otimes \cdots \otimes dice \in \mathcal{D}(\mathbb{N}^K) \quad sum\colon \mathbb{N}^K \longrightarrow \mathbb{N}$

The general construction: convolution

Definition

Let $M = (M, +, 0)$ be a commutative monoid, with two distributions $\omega, \rho \in \mathcal{D}(M)$. Their convolution sum and unit are defined as:

$$\omega + \rho := \mathcal{D}(+)(\omega \otimes \rho) \in \mathcal{D}(M) \qquad \text{with} \qquad 1\big|0\big\rangle \in \mathcal{D}(M).$$

This turns $\mathcal{D}(M)$ into a commutative monoid. In fact there is another lifting.

Theorem

The distribution monad $\mathcal{D}$ on Sets can be lifted to a monad on commutative monoids, as in:

$$\begin{tikzcd} \underline{\mathrm{CMon}} \arrow[r, "\mathcal{D}"] \arrow[d] & \underline{\mathrm{CMon}} \arrow[d] \\ \underline{\mathrm{Sets}} \arrow[r, "\mathcal{D}"] & \underline{\mathrm{Sets}} \end{tikzcd}$$

Aside: this lifing does not extend to vector spaces

Multiple coin flips

- For a bias $r \in [0,1]$ we have $flip(r) \in \mathcal{D}(\{0,1\}) \hookrightarrow \mathcal{D}(\mathbb{N})$ via:
  $flip(r) := r\big|1\big\rangle + (1-r)\big|0\big\rangle \qquad$ where $1 =$ head, $0 =$ tail.
- We can now look at the sum of two coin flips:

$$\begin{aligned} flip(r) + flip(r) &= \mathcal{D}(+)\Big(flip(r) \otimes flip(r)\Big)\\ &= r^2\big|2\big\rangle + 2r(1-r)\big|1\big\rangle + (1-r)^2\big|0\big\rangle\end{aligned}$$

- The sum of K-many coin flips gives the binomial distribution

$$\begin{aligned} K \cdot flip(r) &= flip(r) + \cdots + flip(r)\\ &= \mathcal{D}(sum)\Big(flip(r) \otimes \cdots \otimes flip(r)\Big)\\ &= \sum_{0 \le i \le K} \binom{K}{i} \cdot r^i \cdot (1-r)^{K-i}\big|i\big\rangle\\ &= bn[K](r) \in \mathcal{D}\big(\{0,1,\ldots,K\}\big).\end{aligned}$$

Result: map of monoids

We now have equations:

$$bn[K](r) + bn[L](r) = bn[K+L](r) \qquad \text{and} \qquad bn[0](r) = 1\big|0\big\rangle$$

Equivalent, binomials form a map of monoids:

$$(\mathbb{N}, +, 0) \xrightarrow{\;bn[-](r)\;} (\mathcal{D}(\mathbb{N}), +, 0)$$

This happens more often, for instance for the rate of Poisson distributions, with infinite support:

$$(\mathbb{R}_{\geq 0}, +, 0) \xrightarrow{\;pois[-]\;} (\mathcal{D}_\infty(\mathbb{N}), +, 0)$$

Where we are, so far

Basics of channels

- A Kleisli map $X \to \mathcal{D}(Y)$ will be called a channel
  - the name 'channel' is borrowed from information theory
  - there, transmission errors arise probabilistically
- A channel $X \to \mathcal{D}(Y)$ is a map $X \to Y$ in $\mathcal{K}\ell(\mathcal{D})$
  - we write it as $X \mathrel{\circ\!\!\!\to} Y$, with a circle on its shaft
- Channels, as maps in the SMC $\mathcal{K}\ell(\mathcal{D})$, may be composed, both sequentially and in paralell
- Ordinary functions $f\colon X \to Y$ form deterministic channels $x \mapsto 1\big|f(x)\big\rangle$
  - inclusion via functor $\underline{\mathrm{Sets}} \to \mathcal{K}\ell(\mathcal{D})$
- Tranditional probabilistic view of a channel $X \mathrel{\circ\!\!\!\to} Y$ is as conditional probability $p(y|x)$
  - sequential & parallel composition is not used for such conditional probabilities

Formulas for sequential & parallel composition

- First, Kleisli extension yields pushforward: for channel $c\colon X \mathrel{\circ\!\!\!\to} Y$ and distribution $\omega \in \mathcal{D}(X)$ one gets $c_*(\omega) \in \mathcal{D}(Y)$ via:

$$c_*(\omega) \coloneqq \sum_{y\in Y}\left(\sum_{x\in X}\omega(x)\cdot c(x)(y)\right)\big|\,y\,\big\rangle$$

- For channels $X \overset{c}{\mathrel{\circ\!\!\!\to}} Y \overset{d}{\mathrel{\circ\!\!\!\to}} Z$ defined sequential composition $d \mathbin{\circ\!\!\cdot} c\colon X \mathrel{\circ\!\!\!\to} Z$ as:

$$\big(d \mathbin{\circ\!\!\cdot} c\big)(x) \coloneqq d_*(c(x)) = \sum_{z\in Z}\left(\sum_{y\in Y} c(x)(y)\cdot d(y)(z)\right)\big|\,z\,\big\rangle$$

- For channels $X \overset{c}{\mathrel{\circ\!\!\!\to}} Y$ and $A \overset{e}{\mathrel{\circ\!\!\!\to}} B$ one gets $c \otimes e : X \times A \mathrel{\circ\!\!\!\to} Y \times B$ via pointwise tensors:

$$\big(c \otimes e\big)(x, a) \coloneqq c(x) \otimes e(a) = \sum_{y\in Y,\, b\in B} c(x)(y)\cdot d(a)(b)\,\big|\,y, b\big\rangle.$$

Example channels

- $\mathit{flip}\colon [0,1] \mathrel{\circ\!\!\!\to} \{0,1\}$, where, recall $\mathit{flip}(r) = r\big|\,1\,\big\rangle + (1-r)\big|\,0\,\big\rangle$
  - in fact, this flip is an isomorphism
- Binomial $\mathit{bn}[K]\colon [0,1] \mathrel{\circ\!\!\!\to} \{0,1,\ldots,K\}$
- A probabilistic inverse of $\mathit{acc}\colon X^K \to \mathcal{M}[K](X)$
  - it takes a multiset $\varphi \in \mathcal{M}[K](X)$ to the uniform distribution over all sequences that accumulate to φ
  - recall, there are $(\!(\varphi)\!)$ many such sequences
  - we call this probabilistic inverse arrangement, written as $\mathit{arr} = \mathit{acc}^{\sim 1}\colon \mathcal{M}[K](X) \mathrel{\circ\!\!\!\to} X^K$
  - Excplicitly,

$$\mathit{arr}(\varphi) \coloneqq \mathit{acc}^{\sim 1}(\varphi) = \sum_{\vec{x}\in \mathit{acc}^{-1}(\varphi)} \frac{1}{(\!(\varphi)\!)}\,\big|\,\vec{x}\,\big\rangle.$$

  - Then: $\mathit{acc} \mathbin{\circ\!\!\cdot} \mathit{arr} = \mathrm{id}$

Accumulation and arrangement as (co)equalisers

- Consider all permutations / transpositions $X^K \xrightarrow{\cong} X^K$ as deterministic swap maps in $\mathcal{K}\ell(\mathcal{D})$.
- There is an equaliser & coequaliser diagram in $\mathcal{K}\ell(\mathcal{D})$ of the form:

$$\mathcal{M}[K](X) \xrightarrow{acc^{\sim 1}} X^K \underset{\text{swaps}}{\overset{\vdots}{\rightrightarrows}} X^K \xrightarrow{acc} \mathcal{M}[K](X)$$

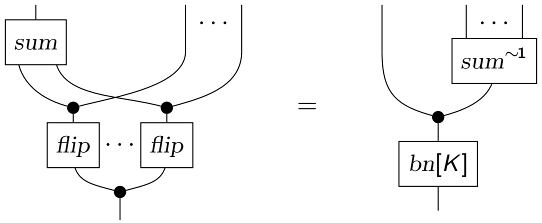
where acc is (also) deterministic

String diagrams & Markov category

- $\mathcal{K}\ell(\mathcal{D})$ is a Markov category:
  - it is symmetric monoidal
  - each object X carries a copier $\Delta \colon X \to X \times X$
  - the tensor unit $1 \in \mathcal{K}\ell(\mathcal{D})$ is final
- For such Markov categories there are convenient string diagrams
  - channels (Kleisli maps) form boxes
  - they can be combined sequentially and in parallel
  - there are copy's $\curlyvee$ and discard's $\overline{\top}$

Basic properties expressed via string diagrams

The addition function $sum \colon \mathbb{N}^K \to \mathbb{N}$ is a sufficient statistic for the K-fold parallel product of flip's, as expressed by the following equality between channels $[0,1] \rightsquigarrow \mathbb{N} \times \{0,1\}^K$.



The partial inverse of the sum is defined as:

$$sum^{\sim 1}(n) = \sum_{\vec{b} \in sum^{-1}(n)} \frac{1}{\binom{K}{n}} \big| \vec{b} \big\rangle.$$

Where we are, so far

General remarks about drawing from an urn

- Drawing coloured balls from an urn is a basic probabilistic model
- The urn contains multiple balls of multiple colours: 5 red, 3 blue, ...
- A draw may consist of a single ball or of multiple balls
  - the proportions of colours in the urn determines the probabilities
- Commonly, three modes of drawing are distinguished
  - **draw-delete**: "hypergeometric"
    - each drawn ball is deleted from the urn
    - the urn shrinks — and drawing stops when the urn is empty
  - **draw-replace**: "multinomial"
    - each drawn ball is returned to the urn before the next draw
    - the urn remains the same
  - **draw-add**: "Pólya"
    - each drawn ball is returned to the urn together with an extra ball of the same colour
    - the urn grows — and displays clustering behaviour
- Multinomial and hypergeometric draws will be discussed here

A 'draw' model for a dentist

- Suppose I run a very basic dental clinic, with only three treatments: cleaning (c) of teeth, filling (f) of holes (cavities) in teeth, and pulling (p) of teeth
- Suppose the proportions of treatments is given by the distribution
$$\tau \coloneqq \tfrac{1}{2}|\,c\,\rangle + \tfrac{1}{3}|\,f\,\rangle + \tfrac{1}{6}|\,p\,\rangle \in \mathcal{D}\big(\{c, f, p\}\big)$$
- Now suppose that $K \geq 1$ patients arrive for treatment, say on a single day, and I wish to compute what are the probabilities for the various combinations of K treatments that I have to perform.
  - this is relevant for scheduling and preparation of resources
  - I like to know the probabilities of multisets of treatments
  - this will be described via "kets over kets"
- The answer corresponds to "drawing" a multiset of size K from the "urn" of treatments τ.

(There is no connection here between "drawing" and "pulling")

Proceeding systematically, with "kets over kets"

- Suppose there are $K = 2$ treatments per day
  - The probabilities of tuples of treatments arise as:
$$\begin{aligned}\tau \otimes \tau = {} & \tfrac{1}{4}\big|\,c, c\,\big\rangle + \tfrac{1}{6}\big|\,c, f\,\big\rangle + \tfrac{1}{12}\big|\,c, p\,\big\rangle + \tfrac{1}{6}\big|\,f, c\,\big\rangle + \tfrac{1}{9}\big|\,f, f\,\big\rangle + \tfrac{1}{18}\big|\,f, p\,\big\rangle \\ & + \tfrac{1}{6}\big|\,p, c\,\big\rangle + \tfrac{1}{9}\big|\,p, f\,\big\rangle + \tfrac{1}{18}\big|\,p, p\,\big\rangle\end{aligned}$$
  - The probabilities of multisets is obtained via accumulation acc
$$\begin{aligned}\mathcal{D}(acc)\big(\tau \otimes \tau\big) = {} & \tfrac{1}{4}\Big|\,2|\,c\,\rangle\Big\rangle + \tfrac{1}{3}\Big|\,1|\,c\,\rangle + 1|\,f\,\rangle\Big\rangle + \tfrac{1}{6}\Big|\,1|\,c\,\rangle + 1|\,p\,\rangle\Big\rangle \\ & + \tfrac{1}{9}\Big|\,2|\,f\,\rangle\Big\rangle + \tfrac{1}{9}\Big|\,1|\,f\,\rangle + 1|\,p\,\rangle\Big\rangle + \tfrac{1}{36}\Big|\,2|\,p\,\rangle\Big\rangle\end{aligned}$$
- Similarly, the probabilities for $K = 3$ treatments are:

$$\begin{aligned}\mathcal{D}(acc)\big(\tau \otimes \tau \otimes \tau\big) = {} & \tfrac{1}{8}\Big|\,3|\,c\,\rangle\Big\rangle + \tfrac{1}{4}\Big|\,2|\,c\,\rangle + 1|\,f\,\rangle\Big\rangle + \tfrac{1}{6}\Big|\,1|\,c\,\rangle + 2|\,f\,\rangle\Big\rangle \\ + \tfrac{1}{27}\Big|\,3|\,f\,\rangle\Big\rangle & + \tfrac{1}{8}\Big|\,2|\,c\,\rangle + 1|\,p\,\rangle\Big\rangle + \tfrac{1}{6}\Big|\,1|\,c\,\rangle + 1|\,f\,\rangle + 1|\,p\,\rangle\Big\rangle \\ + \tfrac{1}{18}\Big|\,2|\,f\,\rangle + 1|\,p\,\rangle\Big\rangle & + \tfrac{1}{24}\Big|\,1|\,c\,\rangle + 2|\,p\,\rangle\Big\rangle + \tfrac{1}{36}\Big|\,1|\,f\,\rangle + 2|\,p\,\rangle\Big\rangle + \tfrac{1}{216}\Big|\,3|\,p\,\rangle\Big\rangle\end{aligned}$$

The multinomial distribution

Definition

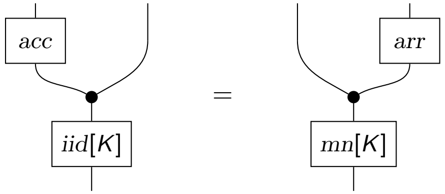
Fix an "urn" $\omega \in \mathcal{D}(X)$ and a draw-size $K \in \mathbb{N}$. The multinomial distribution $mn[K](\omega) \in \mathcal{D}\big(\mathcal{M}[K](X)\big)$ is defined as:
$$\begin{aligned}mn[K](\omega) \coloneqq \mathcal{D}(acc)\big(\omega^K\big) & = \mathcal{D}(acc)\big(\omega \otimes \cdots \otimes \omega\big) \\ & = \sum_{\varphi \in \mathcal{M}[K](X)} (\varphi) \cdot \prod_{x \in X} \omega(x)^{\varphi(x)} \,\big|\,\varphi\,\big\rangle.\end{aligned}$$

In the (traditional) literature you do not find:

- the snappy, conceptually clear formulation $mn[K](\omega) \coloneqq \mathcal{D}(acc)\big(\omega^K\big)$
- The fundamental interaction with frequentist learning $Flrn\colon \mathcal{M}[K](X) \to \mathcal{D}(X)$, via Kleisli extension namely:
$$Flrn_*\Big(mn[K](\omega)\Big) = \omega$$

Other fundamental properties of multinomial

- Closure under convolution, using that multisets form a commutative monoid:
$$mn[K](\omega) + mn[L](\omega) = mn[K+L](\omega)$$
This gives a map of monoids $mn[-](\omega)\colon \mathbb{N} \to \mathcal{D}(\mathcal{M}(X))$
- Accumulation is a sufficient statistic, via an equality of channels $\mathcal{D}(X) \rightsquigarrow \mathcal{M}[K](X) \times X^K$ in:



where:
$$iid[K](\omega) = \omega^K$$
$$arr = acc^{\sim 1}$$

- Law of large number, see later ...
- Also, multinomials form a monoidal natural transformation $\mathcal{D} \Rightarrow \mathcal{M}$, in a lifted setting

Hypergeometric drawing

- First define a single-draw with deletion
  - with channel type $DD\colon \mathcal{M}[K+1](X) \rightsquigarrow \mathcal{M}[K](X)$
  - and formula for random draw from urn / multiset $\upsilon \in \mathcal{M}[K+1](X)$,
$$DD(\upsilon) := \sum_{x \in supp(\upsilon)} \frac{\upsilon(x)}{K+1} \Big| \upsilon - 1 | x \rangle \Big\rangle$$
- Now define hypergeometric drawing as channel $hg[K]\colon \mathcal{M}[L](X) \rightsquigarrow \mathcal{M}[K](X)$, for $L \geq K$, via Kleisli iteration:
$$hg[K] := DD^K = DD \circ \cdots \circ DD, \qquad K \text{ times}$$
- Explicit formula, for urn / multiset $\upsilon \in \mathcal{M}[L](X)$,
$$hg[K](\upsilon) = \sum_{\varphi \leq_K \upsilon} \frac{\binom{\upsilon}{\varphi}}{\binom{L}{K}} |\varphi\rangle \quad \text{where} \quad \binom{\upsilon}{\varphi} := \textstyle\prod_{x \in X} \binom{\upsilon(x)}{\varphi(x)}$$
and where $\varphi \leq_K \upsilon$ means $\|\varphi\| = K$ and $\varphi \leq \upsilon$ pointwise

Basic hypergeometric channel properties

Hypergeometric channels (Kleisli) compose, and they commute with frequentist learning and with multinomials:

$$\begin{tikzcd}
\mathcal{M}[K+L+M](X) \arrow[rr, "hg[K]"] \arrow[dr, "hg[K+L]"'] & & \mathcal{M}[K](X) \\
& \mathcal{M}[K+L](X) \arrow[ur, "hg[K]"'] &
\end{tikzcd}$$

$$\begin{tikzcd}
\mathcal{M}[L](X) \arrow[rr, "hg[K]"] \arrow[dr, "Flrn"'] & & \mathcal{M}[K](X) \arrow[dl, "Flrn"] \\
& X &
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathcal{M}[L](X) \arrow[rr, "hg[K]"] & & \mathcal{M}[K](X) \\
& \mathcal{D}(X) \arrow[ul, "mn[L]"] \arrow[ur, "mn[K]"'] &
\end{tikzcd}$$

Again, you don't find these in the traditional literature, since the monad structure (and thus Kleisli composition) is not recognised.

The law of large urns

Using the total variation or Kantorovic distance d between distributions one gets:
$$\lim_{\upsilon \to \infty} d\Big(hg[K](\upsilon),\, mn[K]\big(Flrn(\upsilon)\big)\Big) = 0.$$

Informally: for a fixed draw-size K, there is for large urns no difference between hypergeometric draws (with deletion) and multinomial draws (without deletion).

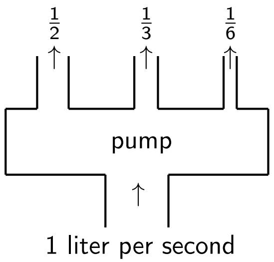
Where we are, so far

Basic questions

- A friend of mine has three children, but I don't know their sexes. The boy/girl probability is 50%.
  - What is the probability that there are three girls?
- A mutual acquaintance now tells me that there is at least one daughter in the family
  - Given this extra information ("evidence"), what is now the probability that there are three girls?
- Many say $\frac{1}{4}$, but it is $\frac{1}{7}$. Why is this so bloody difficult?
  - we have no logic and no good mental models for such reasoning
  - (despite the fact that according to neurscientists we have a "Bayesian brain", at the neuronal level)
- Small variation of the question, that will be elaborated later:
  - there are four children in the family
  - I've been told there are at least two girls
  - what is now the four-girl probability?

A physical model for updating I

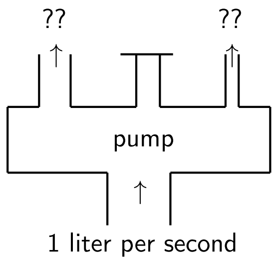
Consider a pump with one input pipe at the bottom and three output pipes at the top. The outgoing pipes have relative diameters, as indicated at the top.



representing the distribution $\omega = \frac{1}{2}|L\rangle + \frac{1}{3}|M\rangle + \frac{1}{6}|R\rangle$.

A physical model for updating II

There is now evidence that the middle pipe is blocked. The pump keeps on operating and still realises the throughput of one liter per second (with increased pressure). What are the new outgoing flows?

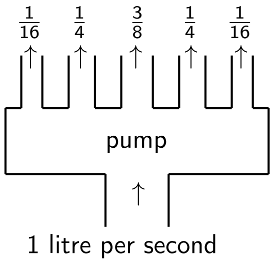


Answer

- Recall the left and right pipes have diameter $\frac{1}{2}$ and $\frac{1}{6}$, with ratio 3 : 1
- The new, updated distribution is thus $\frac{3}{4}|L\rangle + \frac{1}{4}|R\rangle$.

The family with four children example

The uniform girl-boy distribution is $\upsilon = \frac{1}{2}|\, g\,\rangle + \frac{1}{2}|\, b\,\rangle$. The four-children options are given by the multinomial distribution $mn[4](\upsilon)$.



representing the distribution:

$$\begin{aligned} mn[4](\upsilon) \;=\; & \tfrac{1}{16}\Big|\, 4|\, g\,\rangle \Big\rangle \\ & + \tfrac{1}{4}\Big|\, 3|\, g\,\rangle + 1|\, b\,\rangle \Big\rangle \\ & + \tfrac{3}{8}\Big|\, 2|\, g\,\rangle + 2|\, b\,\rangle \Big\rangle \\ & + \tfrac{1}{4}\Big|\, 1|\, g\,\rangle + 3|\, b\,\rangle \Big\rangle \\ & + \tfrac{1}{16}\Big|\, 4|\, b\,\rangle \Big\rangle . \end{aligned}$$

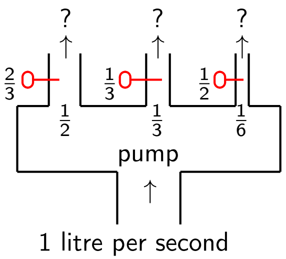
- Recall, there are at least two girls, so the last two pipes are blocked
- the remaining ratios $\frac{1}{16} : \frac{1}{4} : \frac{3}{8}$ are $1 : 4 : 6$, adding up to 11
- the update is: $\frac{1}{11}\Big|\, 4|\, g\,\rangle \Big\rangle + \frac{4}{11}\Big|\, 3|\, g\,\rangle + 1|\, b\,\rangle \Big\rangle + \frac{6}{11}\Big|\, 2|\, g\,\rangle + 2|\, b\,\rangle \Big\rangle$

Now in terms of "cross-out and renormalise"

$$\left(\begin{array}{c} \frac{1}{16}\Big|\, 4|\, g\,\rangle \Big\rangle + \\ \frac{1}{4}\Big|\, 3|\, g\,\rangle + 1|\, b\,\rangle \Big\rangle + \\ \frac{3}{8}\Big|\, 2|\, g\,\rangle + 2|\, b\,\rangle \Big\rangle + \\ \cancel{\frac{1}{4}\Big|\, 1|\, g\,\rangle + 3|\, b\,\rangle \Big\rangle} + \\ \cancel{\frac{1}{16}\Big|\, 4|\, b\,\rangle \Big\rangle} \end{array}\right) \xmapsto{\text{renormalise}} \left(\begin{array}{c} \frac{1}{11}\Big|\, 4|\, g\,\rangle \Big\rangle + \\ \frac{4}{11}\Big|\, 3|\, g\,\rangle + 1|\, b\,\rangle \Big\rangle + \\ \frac{6}{11}\Big|\, 2|\, g\,\rangle + 2|\, b\,\rangle \Big\rangle \end{array}\right)$$

A physical model for updating III

Instead of "sharp" blocking (yes/no) we can add taps to the pipes, for "fuzzy" evidence, as in:



the fractions left of the taps describe their openness

- The ratios are now: $\left(\frac{1}{2}\cdot\frac{2}{3}\right) : \left(\frac{1}{3}\cdot\frac{1}{3}\right) : \left(\frac{1}{6}\cdot\frac{1}{2}\right)$, that is $12 : 4 : 3$, adding up to 19
- the update is then: $\frac{12}{19}|\, L\,\rangle + \frac{4}{19}|\, M\,\rangle + \frac{3}{19}|\, R\,\rangle$

Predicates, observables, and validity

- An observable is a function $p\colon X \to \mathbb{R}$
  - a random variable is a pair of $\omega \in \mathcal{D}(X)$, $p\colon X \to \mathbb{R}$
  - (this is a fundamental concept, but hardly ever defined explicitly)
- Fuzzy and sharp predicates are special cases of observables, with $p\colon X \to [0,1]$ and $p\colon X \to \{0,1\}$
- The validity $\omega \models p$ of observable $p\colon X \to \mathbb{R}$ in distribution $\omega \in \mathcal{D}(X)$ is:
  $$\omega \models p \;\coloneqq\; \sum_{x\in X} \omega(x)\cdot p(x)$$
  This is commonly written as expected value $E(p)$, leaving ω implicity

The law of large numbers, in terms of validity

For a distribution ω,

$$\lim_{K\to\infty} mn[K](\omega) \models d\Big(\omega, Flrn(-)\Big) \;=\; 0.$$

Informally, for very large draws / samples φ from ω, the distance between ω and $Flrn(\varphi)$ is zero, in probability.

Pulling back

Observables can be pulled back along channels: for $c\colon X \to Y$ and $q\colon Y \to \mathbb{R}$ we get $c^*(q)\colon X \to \mathbb{R}$ via:

$$c^*(q)(x) \;\coloneqq\; \sum_{y\in X} c(x)(y)\cdot q(y)$$

Then: $c_*(\omega) \models q \;=\; \omega \models c^*(q)$

- "the law of total expectation" or "conditional expectation formula"
- confusingly written as $E(Y) = E\big(E(Y|X)\big)$, leaving relevant distribution, channel and observable implicit

Bigger picture: "state-and-effect" or "Heisenberg/Schrödinger" triangles:

$$\begin{tikzcd}
\underline{\mathrm{EMod}}^{\mathrm{op}} \arrow[rr, bend left=10, "\mathrm{Hom}(-,[0,1])"] & & \mathcal{EM}(\mathcal{D}) \arrow[ll, bend left=10, "\mathrm{Hom}(-,[0,1])"] \\
& \mathcal{K}\ell(\mathcal{D}) \arrow[ul, "Pred"] \arrow[ur, "Stat"'] &
\end{tikzcd}$$

Bayesian update

Definition

Consider a distribution $\omega \in \mathcal{D}(X)$ and predicate $p\colon X \to [0,1]$ with non-zero validity $\omega \models p$. The Bayesian update $\omega|_p \in \mathcal{D}(X)$ is the normalised product:

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

- $\omega|_1 = \omega$ and $\omega|_{p\&q} = \omega|_p|_q$
- Bayes' (product) laws:

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

- Validity increase through updating — absent in the literature!!

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

Validity and conditioning example

- Take $X = \{1,2,3,4,5,6\}$ with $dice = \sum_{1\leq i\leq 6} \frac{1}{6}|\,i\,\rangle \in \mathcal{D}(X)$
- Take the fuzzy predicate $evenish\colon X \to [0,1]$

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

- The validity of $evenish$ for our fair dice is:

$$\begin{aligned}
dice \models evenish \;&=\; \textstyle\sum_i dice(i)\cdot evenish(i) \\
&=\; \tfrac{1}{6}\cdot\tfrac{1}{5} + \tfrac{1}{6}\cdot\tfrac{9}{10} + \tfrac{1}{6}\cdot\tfrac{1}{10} + \tfrac{1}{6}\cdot\tfrac{9}{10} + \tfrac{1}{6}\cdot\tfrac{1}{10} + \tfrac{1}{6}\cdot\tfrac{4}{5} \;=\; \tfrac{1}{2}
\end{aligned}$$

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

$$\begin{aligned}
dice\big|_{evenish} \;&=\; \textstyle\sum_i \frac{dice(i)\cdot evenish(i)}{dice \models evenish}\,\big|\,x\,\big\rangle \\
&=\; \tfrac{1/6\cdot 1/5}{1/2}\big|1\big\rangle + \tfrac{1/6\cdot 9/10}{1/2}\big|2\big\rangle + \tfrac{1/6\cdot 1/10}{1/2}\big|3\big\rangle + \tfrac{1/6\cdot 9/10}{1/2}\big|4\big\rangle + \tfrac{1/6\cdot 1/10}{1/2}\big|5\big\rangle + \tfrac{1/6\cdot 4/5}{1/2}\big|6\big\rangle \\
&=\; \tfrac{1}{15}\big|1\big\rangle + \tfrac{3}{10}\big|2\big\rangle + \tfrac{1}{30}\big|3\big\rangle + \tfrac{3}{10}\big|4\big\rangle + \tfrac{1}{30}\big|5\big\rangle + \tfrac{4}{15}\big|6\big\rangle.
\end{aligned}$$

Backward inference: updating along a channel, part I

- Consider a disease with *a priori* probability (or 'prevalence') of 10%
  - we thus have a prior distribution $\omega = \frac{1}{10}|\,d\,\rangle + \frac{9}{10}|\,d^{\perp}\,\rangle$
- There is a test for the disease with:
  - ('sensitivity') If someone has the disease, then the test is positive with probability of 90%
  - ('specificity') If someone does not have the disease, there is a 95% chance that the test is negative.
- The test gives a channel $t\colon \{d, d^{\perp}\} \to \mathcal{D}(\{p, n\})$

$$t(d) = \tfrac{9}{10}|\,p\,\rangle + \tfrac{1}{10}|\,n\,\rangle \quad \text{and} \quad t(d^{\perp}) = \tfrac{1}{20}|\,p\,\rangle + \tfrac{19}{20}|\,n\,\rangle$$

Backward inference: updating along a channel, part II

- Suppose you have a positive test. What is the probability that you have the disease?
  - the positive test is a point predicate 1_p on $\{p, n\}$
  - we pull it back along the channel t to a predicate $t^*(1_p)$ on $\{d, d^{\perp}\}$
  - now we can perform the update $\omega|_{t^*(1_p)}$, giving as posterior distribution: $$\omega|_{t^*(1_p)} = \tfrac{2}{3}|\,d\,\rangle + \tfrac{1}{3}|\,d^{\perp}\,\rangle.$$
- This approach also works with:
  - fuzzy test evidence: "I'm 80% sure the test is positive"
  - multiple tests, although then different update methods of Pearl and of Jeffrey can be used

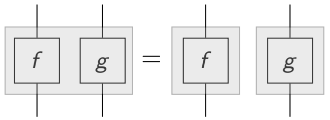
Where we are, so far

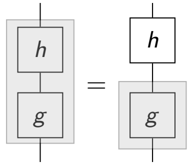
Normalisation boxes

- Probabilistic updating involves normalisation
  - this turns subdistributions, with sum ≤ 1, into proper distributions
- This normalisation can be expressed graphically via shaded (or dashed) boxes.
- It became clear recently that normalisation (boxes) behave reasonably well
  - there are several compositional rules
  - there is also a removal rule for shaded boxes
- This allows graphical rewriting for conditioning in Bayesian networks and also for causality
  - see own MFPS'25 paper and work of Sean Tull and others

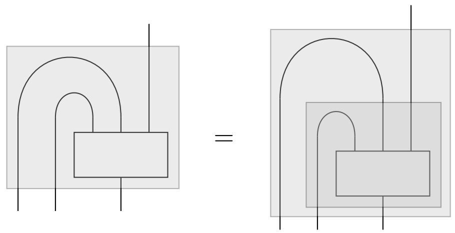
Some shaded box rules

- Parallel and sequential rules (when h is a proper channel)



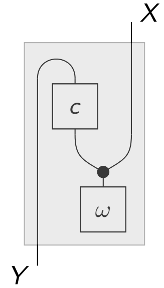


- Multiple normalisation:

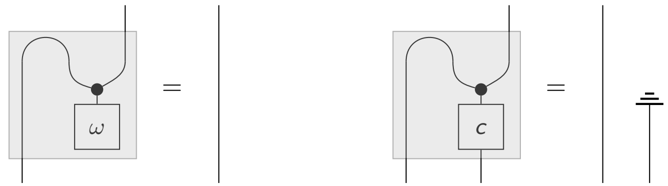


Dagger and box removal

$c^\dagger_\omega \quad :=$

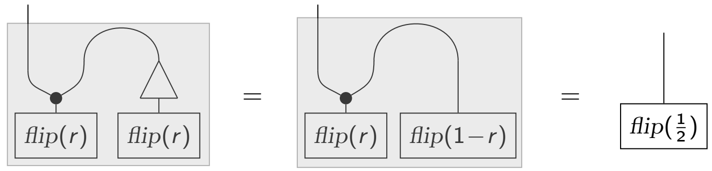


for $c\colon X \rightsquigarrow Y$, $\omega \in \mathcal{D}(X)$

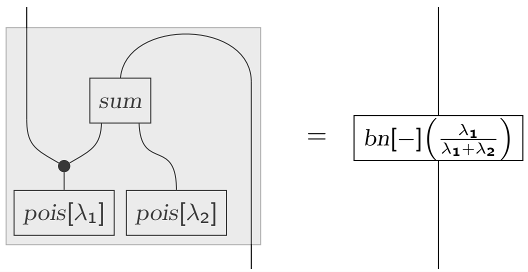


Two illustrations

Von Neumann showed how to get a fair coin from a biased one:



A basic result about the sum of two Poisson distributions:



Where we are, so far

Concluding remarks

- **Message**: there is so much beautiful (unexpolored, categorical) structure in probability
  - categorical notation trumps traditional notation
  - better expresses what's going on, uncovering overlooked properties
- Multisets are an essential but (largely) ignored part of the story
- There is much more to be said, e.g. about
  - a distributive law $\mathcal{MD} \Rightarrow \mathcal{DM}$
  - formalisation of updating, including rules of Jeffrey and Pearl
  - causality, a hot topic
  - continuous probability, etc.
- This precise approach may be useful for understanding AI, as XAI
  - **big goal**: a symbolic, formal logic for probability, with updating
- Current version of my "fat" book (now ± 800 pages):
  - www.cs.ru.nl/B.Jacobs/PAPERS/ProbabilisticReasoning.pdf
  - Feedback is welcome!

iCIS | Digital Security
Radboud University