

Categorical Probability Theory

Università di Pisa, May 5, 2026

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

Outline

Introduction

Multisets

Probability distributions

Channels / Kleisli maps

Draw distributions

Probabilistic updating

Updating in graphical form

Conclusions



Categorical Probability

Where we are, so far

Introduction

Multisets

Probability distributions

Channels / Kleisli maps

Draw distributions

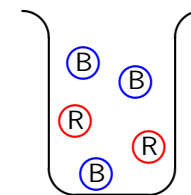
Probabilistic updating

Updating in graphical form

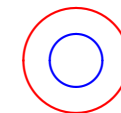
Conclusions



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 $X \rightarrow \mathcal{D}(Y)$ or $X \rightarrow \mathcal{G}(Y)$
 - for finite/discrete distribution monad \mathcal{D} , and continuous distribution “Giry” monad \mathcal{G}
 - the Kleisli categories $\mathcal{Kl}(\mathcal{D})$ and $\mathcal{Kl}(\mathcal{G})$ 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: $\mathcal{Kl}(\mathcal{D})$ and $\mathcal{Kl}(\mathcal{G})$ are symmetric monoidal
 - with a final object as tensor unit
- ▶ These Kleisli categories are also **Markov categories**
 - there are copy maps $X \rightarrow X \times X$, forming comonoids
 - Kleisli maps that commute with copiers are called **deterministic**
- ▶ Moreover, they are **effectuses**
 - they have coproducts $+$, which are suitably well-behaved
 - ‘predicates’ $X \rightarrow 1 + 1$ 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

Introduction

Multisets

Probability distributions

Channels / Kleisli maps

Draw distributions

Probabilistic updating

Updating in graphical form

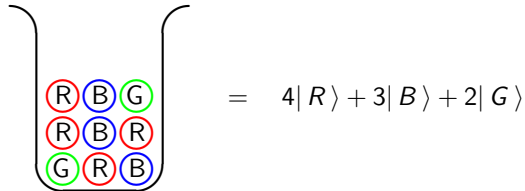
Conclusions



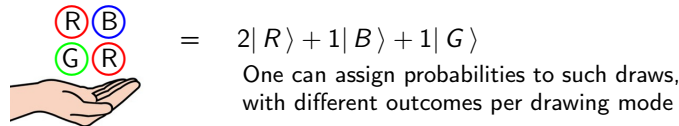
Drawing in terms of multisets

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:



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



From lists to multisets to subsets

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



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: $\sum_i n_i|x_i\rangle$ with $n_i \in \mathbb{N}$ and $x_i \in X$
- ▶ Alternatively, a multiset is a **function** $\varphi: X \rightarrow \mathbb{N}$ 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. $\|4|R\rangle + 3|B\rangle + 2|G\rangle\| = 4 + 3 + 2 = 9$.
- ▶ $\mathcal{M}[K](X) \hookrightarrow \mathcal{M}(X)$ is written for the subset of multisets of size $K \in \mathbb{N}$



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 $\binom{n+K-1}{K}$: if $|X| = n \geq 1$, then:

$$|\mathcal{M}[K](X)| = \binom{n+K-1}{K} = \frac{(n+K-1)!}{K! \cdot (n-1)!}$$

- ▶ For $\varphi \in \mathcal{M}(X)$, the number of lists $\ell \in \mathcal{L}(X)$ with $\text{acc}(\ell) = \varphi$ is given by the **multiset coefficient** $(\varphi) \in \mathbb{N}$, defined as:

$$(\varphi) := \frac{\|\varphi\|!}{\varphi_{\!|}!} \quad \text{where} \quad \varphi_{\!|} := \prod_{x \in X} \varphi(x)!$$

For $\varphi = \sum_i n_i |x_i\rangle$ with $n = \sum_i n_i$ this (φ) is $\binom{n}{n_1, \dots, n_k}$.

- ▶ If $|X| = n$, then:

$$\sum_{\varphi \in \mathcal{M}[K](X)} (\varphi) = n^K.$$



Distributions (finite, discrete)

- ▶ In a distribution the multiplicities **add up to one**, as in:

$$\text{coin} = \frac{49}{100} |H\rangle + \frac{51}{100} |T\rangle$$

$$\text{dice} = \frac{1}{6} |1\rangle + \frac{1}{6} |2\rangle + \frac{1}{6} |3\rangle + \frac{1}{6} |4\rangle + \frac{1}{6} |5\rangle + \frac{1}{6} |6\rangle$$

- ▶ 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 $\omega: X \rightarrow [0, 1]$ 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**:

$$\text{Flrn}(4|R\rangle + 3|B\rangle + 2|G\rangle) = \frac{4}{9}|R\rangle + \frac{3}{9}|B\rangle + \frac{2}{9}|G\rangle.$$

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



Where we are, so far

Introduction

Multisets

Probability distributions

Channels / Kleisli maps

Draw distributions

Probabilistic updating

Updating in graphical form

Conclusions



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:

$$\omega \otimes \rho := \sum_{x \in X, y \in Y} \omega(x) \cdot \rho(y) |x, y\rangle$$

- ▶ For instance, for

$$\text{coin} = \frac{1}{2} |H\rangle + \frac{1}{2} |T\rangle \quad \text{dice} = \sum_{1 \leq i \leq 6} \frac{1}{6} |i\rangle$$

tossing them together is captured by the tensor product $\text{coin} \otimes \text{dice}$:

$$\begin{aligned} & \frac{1}{12} |H, 1\rangle + \frac{1}{12} |H, 2\rangle + \frac{1}{12} |H, 3\rangle + \frac{1}{12} |H, 4\rangle + \frac{1}{12} |H, 5\rangle + \frac{1}{12} |H, 6\rangle + \\ & + \frac{1}{12} |T, 1\rangle + \frac{1}{12} |T, 2\rangle + \frac{1}{12} |T, 3\rangle + \frac{1}{12} |T, 4\rangle + \frac{1}{12} |T, 5\rangle + \frac{1}{12} |T, 6\rangle \end{aligned}$$



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

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

▶ $\mathcal{D}(f): \mathcal{D}(X) \rightarrow \mathcal{D}(Y)$ and $\mathcal{M}(f): \mathcal{M}(X) \rightarrow \mathcal{M}(Y)$

▶ Explicitly:

$$\mathcal{D}(f)\left(\sum_i r_i |x_i\rangle\right) := \sum_i r_i |f(x_i)\rangle \quad \text{and similarly for } \mathcal{M}$$

▶ 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:

$$\mathcal{D}(\pi_1)(\tau) \in \mathcal{D}(X) \quad \text{and} \quad \mathcal{D}(\pi_2)(\tau) \in \mathcal{D}(Y)$$

▶ 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:

$$\text{shift}(s, \omega) := \mathcal{D}(s + (-))(\omega) \quad \text{where} \quad s + (-): \mathbb{R} \rightarrow \mathbb{R}$$

▶ Similarly, one can **scale** ω via:

$$\text{scale}(s, \omega) := \mathcal{D}(s \cdot (-))(\omega) \quad \text{using} \quad s \cdot (-): \mathbb{R} \rightarrow \mathbb{R}$$

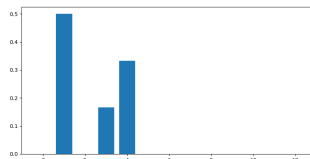
Shifting and scaling form **monoid actions** on distributions

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

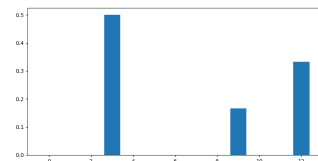
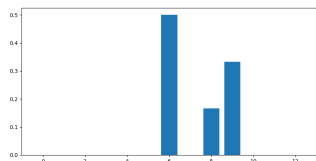


Shift & scale illustrations

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 Act_M .

Theorem

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

$$\begin{array}{ccc} \text{Act}_M & \xrightarrow{\mathcal{D}} & \text{Act}_M \\ \downarrow & & \downarrow \\ \text{Sets} & \xrightarrow{\mathcal{D}} & \text{Sets} \end{array}$$

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:

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

- ▶ Similarly one may compute: $\mathcal{D}(sum)(dice \otimes dice \otimes dice)$
- ▶ The **types** involved are:

$$dice \in \mathcal{D}(\mathbb{N}) \quad dice^K = dice \otimes \dots \otimes dice \in \mathcal{D}(\mathbb{N}^K) \quad sum: \mathbb{N}^K \rightarrow \mathbb{N}$$



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:

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

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:

$$\begin{array}{ccc} \underline{\text{CMon}} & \xrightarrow{\mathcal{D}} & \underline{\text{CMon}} \\ \downarrow & & \downarrow \\ \underline{\text{Sets}} & \xrightarrow{\mathcal{D}} & \underline{\text{Sets}} \end{array}$$

Aside: this lifting does not extend to vector spaces



Sum of dices as convolution

- ▶ Using convolution we can describe the two-dice-sum distribution as:

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

- ▶ Similarly, the three-dice-sum distribution is:

$$\begin{aligned} dice + dice + dice &= \frac{1}{216}|3\rangle + \frac{1}{72}|4\rangle + \frac{1}{36}|5\rangle + \frac{5}{108}|6\rangle + \frac{5}{72}|7\rangle + \\ &\quad + \frac{7}{72}|8\rangle + \frac{25}{216}|9\rangle + \frac{1}{8}|10\rangle + \frac{1}{8}|11\rangle + \frac{25}{216}|12\rangle + \\ &\quad + \frac{7}{72}|13\rangle + \frac{5}{72}|14\rangle + \frac{5}{108}|15\rangle + \frac{1}{36}|16\rangle + \frac{1}{72}|17\rangle + \frac{1}{216}|18\rangle \end{aligned}$$



Multiple coin flips

- ▶ For a **bias** $r \in [0, 1]$ we have $flip(r) \in \mathcal{D}(\{0, 1\}) \leftrightarrow \mathcal{D}(\mathbb{N})$ via:

$$flip(r) := r|1\rangle + (1-r)|0\rangle \quad \text{where } 1 = \text{head}, 0 = \text{tail}.$$

- ▶ We can now look at the **sum of two coin flips**:

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

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

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



Result: map of monoids

We now have equations:

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

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

Introduction

Multisets

Probability distributions

Channels / Kleisli maps

Draw distributions

Probabilistic updating

Updating in graphical form

Conclusions



Basics of channels

- ▶ A Kleisli map $X \rightarrow \mathcal{D}(Y)$ will be called a **channel**
 - the name 'channel' is borrowed from information theory
 - there, transmission errors arise probabilistically
- ▶ A channel $X \rightarrow \mathcal{D}(Y)$ is a map $X \rightarrow Y$ in $\mathcal{Kl}(\mathcal{D})$
 - we write it as $X \rightarrowtail Y$, with a circle on its shaft
- ▶ Channels, as maps in the SMC $\mathcal{Kl}(\mathcal{D})$, may be composed, both **sequentially** and **in parallel**
- ▶ Ordinary functions $f: X \rightarrow Y$ form **deterministic** channels $x \mapsto 1|f(x)\rangle$
 - inclusion via functor $\text{Sets} \rightarrow \mathcal{Kl}(\mathcal{D})$
- ▶ Traditional probabilistic view of a channel $X \rightarrowtail Y$ 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 $c: X \rightarrowtail Y$ and distribution $\omega \in \mathcal{D}(X)$ one gets $c_*(\omega) \in \mathcal{D}(Y)$ via:

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

- ▶ For channels $X \rightarrowtail Y \xrightarrow{d} Z$ defined **sequential composition** $d \circ c: X \rightarrowtail Z$ as:

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

- ▶ For channels $X \rightarrowtail Y$ and $A \rightarrowtail B$ one gets $c \otimes e: X \times A \rightarrowtail Y \times B$ via pointwise tensors:

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



Example channels

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

$$arr(\varphi) := acc^{\sim 1}(\varphi) = \sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{1}{(\varphi)} |\vec{x}\rangle.$$

- Then: $acc \circ arr = id$

Accumulation and arrangement as (co)equalisers

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

$$\mathcal{M}[K](X) \xrightarrow{acc^{\sim 1}} X^K \begin{array}{c} \xrightarrow{\quad \circ \quad} \\ \vdots \text{ swaps} \\ \xrightarrow{\quad \circ \quad} \end{array} X^K \xrightarrow{acc} \mathcal{M}[K](X)$$

where acc is (also) deterministic

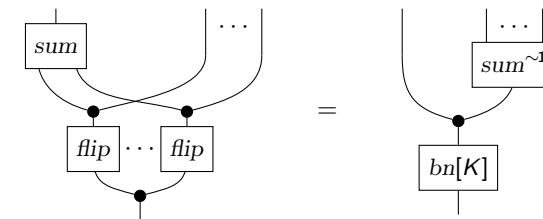


String diagrams & Markov category

- ▶ $\mathcal{Kl}(\mathcal{D})$ is a **Markov category**:
 - it is symmetric monoidal
 - each object X carries a **copier** $\Delta: X \rightarrow X \times X$
 - the tensor unit $1 \in \mathcal{Kl}(\mathcal{D})$ 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 Υ and discard's $\bar{\Upsilon}$

Basic properties expressed via string diagrams

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



The partial inverse of the sum is defined as:

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



Where we are, so far

Introduction

Multisets

Probability distributions

Channels / Kleisli maps

Draw distributions

Probabilistic updating

Updating in graphical form

Conclusions



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
$$\tau := \frac{1}{2}|c\rangle + \frac{1}{3}|f\rangle + \frac{1}{6}|p\rangle \in \mathcal{D}(\{c, f, p\})$$
- ▶ 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")



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



Proceeding systematically, with "kets over kets"

- ▶ Suppose there are $K = 2$ treatments per day
 - The probabilities of **tuples** of treatments arise as:
$$\tau \otimes \tau = \frac{1}{4}|c, c\rangle + \frac{1}{6}|c, f\rangle + \frac{1}{12}|c, p\rangle + \frac{1}{6}|f, c\rangle + \frac{1}{9}|f, f\rangle + \frac{1}{18}|f, p\rangle + \frac{1}{6}|p, c\rangle + \frac{1}{9}|p, f\rangle + \frac{1}{18}|p, p\rangle$$
 - The probabilities of **multisets** is obtained via **accumulation** acc
$$\mathcal{D}(acc)(\tau \otimes \tau) = \frac{1}{4}|2|c\rangle\rangle + \frac{1}{3}|1|c\rangle + 1|f\rangle\rangle + \frac{1}{6}|1|c\rangle + 1|p\rangle\rangle + \frac{1}{9}|2|f\rangle\rangle + \frac{1}{9}|1|f\rangle + 1|p\rangle\rangle + \frac{1}{36}|2|p\rangle\rangle$$
- ▶ Similarly, the probabilities for $K = 3$ treatments are:
$$\mathcal{D}(acc)(\tau \otimes \tau \otimes \tau) = \frac{1}{8}|3|c\rangle\rangle + \frac{1}{4}|2|c\rangle + 1|f\rangle\rangle + \frac{1}{6}|1|c\rangle + 2|f\rangle\rangle + \frac{1}{27}|3|f\rangle\rangle + \frac{1}{8}|2|c\rangle + 1|p\rangle\rangle + \frac{1}{6}|1|c\rangle + 1|f\rangle + 1|p\rangle\rangle + \frac{1}{18}|2|f\rangle + 1|p\rangle\rangle + \frac{1}{24}|1|c\rangle + 2|p\rangle\rangle + \frac{1}{36}|1|f\rangle + 2|p\rangle\rangle + \frac{1}{216}|3|p\rangle\rangle$$



The multinomial distribution

Definition

Fix an “urn” $\omega \in \mathcal{D}(X)$ and a draw-size $K \in \mathbb{N}$. The **multinomial** distribution $mn[K](\omega) \in \mathcal{D}(\mathcal{M}[K](X))$ is defined as:

$$\begin{aligned} mn[K](\omega) &:= \mathcal{D}(acc)(\omega^K) = \mathcal{D}(acc)(\omega \otimes \cdots \otimes \omega) \\ &= \sum_{\varphi \in \mathcal{M}[K](X)} (\varphi) \cdot \prod_{x \in X} \omega(x)^{\varphi(x)} | \varphi \rangle. \end{aligned}$$

In the (traditional) literature you do **not** find:

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

$$Flrn_* (mn[K](\omega)) = \omega$$

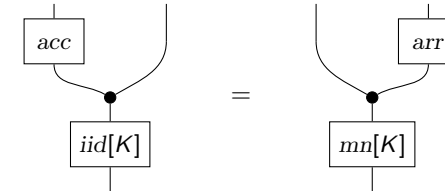
Other fundamental properties of multinomial

- ▶ 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** $mn[-](\omega): \mathbb{N} \rightarrow \mathcal{D}(\mathcal{M}(X))$

- ▶ Accumulation is a **sufficient statistic**, via an equality of channels $\mathcal{D}(X) \rightarrow \mathcal{M}[K](X) \times X^K$ in:



where:

$$\begin{aligned} iid[K](\omega) &= \omega^K \\ arr &= acc^{\sim 1} \end{aligned}$$

- ▶ 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 $DD: \mathcal{M}[K+1](X) \rightarrow \mathcal{M}[K](X)$
 - and formula for random draw from urn / multiset $v \in \mathcal{M}[K+1](X)$,

$$DD(v) := \sum_{x \in \text{supp}(v)} \frac{v(x)}{K+1} |v - 1|x\rangle$$

- ▶ Now define **hypergeometric** drawing as channel
 $hg[K]: \mathcal{M}[L](X) \rightarrow \mathcal{M}[K](X)$, for $L \geq K$, via Kleisli iteration:

$$hg[K] := DD^K = DD \circ \cdots \circ DD, \quad K \text{ times}$$

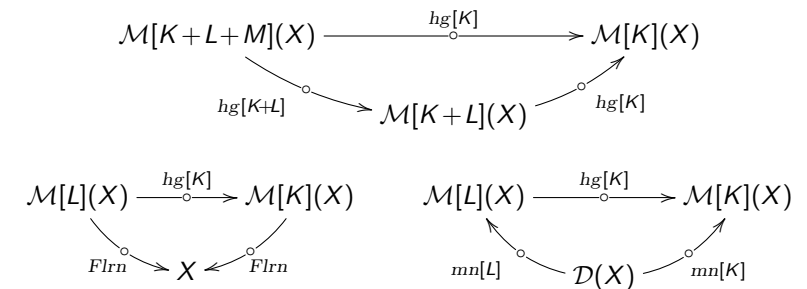
- ▶ **Explicit formula**, for urn / multiset $v \in \mathcal{M}[L](X)$,

$$hg[K](v) = \sum_{\varphi \leq_K v} \frac{\binom{v}{\varphi}}{\binom{L}{K}} | \varphi \rangle \quad \text{where} \quad \binom{v}{\varphi} := \prod_{x \in X} \binom{v(x)}{\varphi(x)}$$

and where $\varphi \leq_K v$ means $\|\varphi\| = K$ and $\varphi \leq v$ pointwise

Basic hypergeometric channel properties

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



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



Where we are, so far

Introduction

Multisets

Probability distributions

Channels / Kleisli maps

Draw distributions

Probabilistic updating

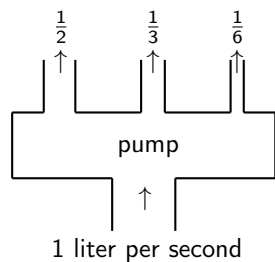
Updating in graphical form

Conclusions



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

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

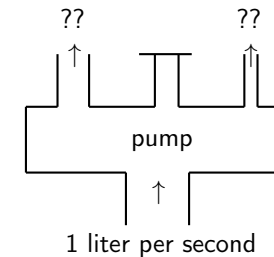
Basic questions

- ▶ 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 neuroscientists 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 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?



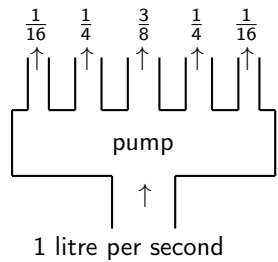
Answer

- ▶ 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 $v = \frac{1}{2}|g\rangle + \frac{1}{2}|b\rangle$. The four-children options are given by the **multinomial** distribution $mn[4](v)$.



representing the distribution:

$$mn[4](v) = \frac{1}{16} |4|g\rangle\rangle + \frac{1}{4} |3|g\rangle + 1|b\rangle\rangle + \frac{3}{8} |2|g\rangle + 2|b\rangle\rangle + \frac{1}{4} |1|g\rangle + 3|b\rangle\rangle + \frac{1}{16} |4|b\rangle\rangle.$$

- ▶ 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: $\frac{1}{11} |4|g\rangle\rangle + \frac{4}{11} |3|g\rangle + 1|b\rangle\rangle + \frac{6}{11} |2|g\rangle + 2|b\rangle\rangle$

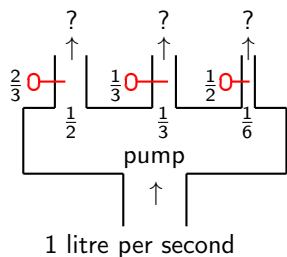
Now in terms of “cross-out and renormalise”

$$\left(\begin{array}{l} \frac{1}{16} |4|g\rangle\rangle + \\ \frac{1}{4} |3|g\rangle + 1|b\rangle\rangle + \\ \frac{3}{8} |2|g\rangle + 2|b\rangle\rangle + \\ \frac{1}{4} |1|g\rangle + 3|b\rangle\rangle + \\ \frac{1}{16} |4|b\rangle\rangle \end{array} \right) \xrightarrow{\text{renormalise}} \left(\begin{array}{l} \frac{1}{11} |4|g\rangle\rangle + \\ \frac{4}{11} |3|g\rangle + 1|b\rangle\rangle + \\ \frac{6}{11} |2|g\rangle + 2|b\rangle\rangle \end{array} \right)$$



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: $(\frac{1}{2} \cdot \frac{2}{3}) : (\frac{1}{3} \cdot \frac{1}{3}) : (\frac{1}{6} \cdot \frac{1}{2})$, 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 $p: X \rightarrow \mathbb{R}$
 - a **random variable** is a pair of $\omega \in \mathcal{D}(X)$, $p: X \rightarrow \mathbb{R}$
 - (this is a fundamental concept, but hardly ever defined explicitly)
- ▶ **Fuzzy** and **sharp** predicates are special cases of observables, with $p: X \rightarrow [0, 1]$ and $p: X \rightarrow \{0, 1\}$
- ▶ The **validity** $\omega \models p$ of observable $p: X \rightarrow \mathbb{R}$ in distribution $\omega \in \mathcal{D}(X)$ is:

$$\omega \models p := \sum_{x \in X} \omega(x) \cdot p(x)$$

This is commonly written as **expected value** $E(p)$, leaving ω implicitly



Pulling back

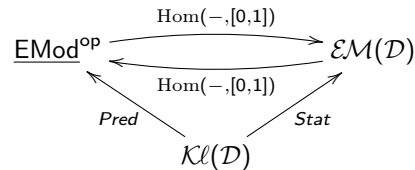
Observables can be **pulled back** along channels: for $c: X \rightarrow Y$ and $q: Y \rightarrow \mathbb{R}$ we get $c^*(q): X \rightarrow \mathbb{R}$ via:

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

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

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

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



Bayesian update

Definition

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

$$\omega|_p := \sum_{x \in X} \frac{\omega(x) \cdot p(x)}{\omega \models p} |x\rangle$$

- ▶ $\omega|_1 = \omega$ and $\omega|_{p \& q} = \omega|_p|_q$

- ▶ Bayes' (product) laws:

$$\omega|_p \models q = \frac{\omega \models p \& q}{\omega \models p} \quad \text{and} \quad \omega|_p \models q = \frac{(\omega|_q \models p) \cdot (\omega \models q)}{\omega \models p}$$

- ▶ **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 $\text{dice} = \sum_{1 \leq i \leq 6} \frac{1}{6} |i\rangle \in \mathcal{D}(X)$

- ▶ Take the **fuzzy predicate** $\text{evenish}: X \rightarrow [0, 1]$

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

- ▶ The **validity** of evenish for our fair dice is:

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

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

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

New validity: $\frac{58}{75}$



Backward inference: updating along a channel, part I

- ▶ Consider a **disease** with a *prior* 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 $t: \{d, d^\perp\} \rightarrow \mathcal{D}(\{p, n\})$

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



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:

$$\omega|_{t^*(1_p)} = \frac{2}{3}|d\rangle + \frac{1}{3}|d^\perp\rangle.$$
- ▶ 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



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



Where we are, so far

Introduction

Multisets

Probability distributions

Channels / Kleisli maps

Draw distributions

Probabilistic updating

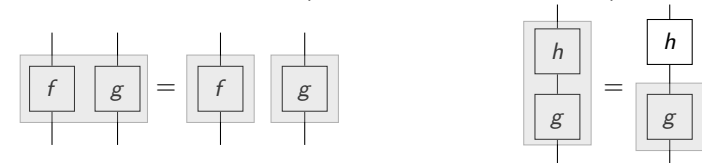
Updating in graphical form

Conclusions

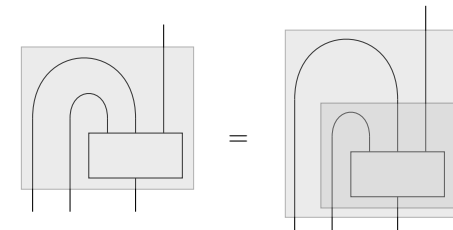


Some shaded box rules

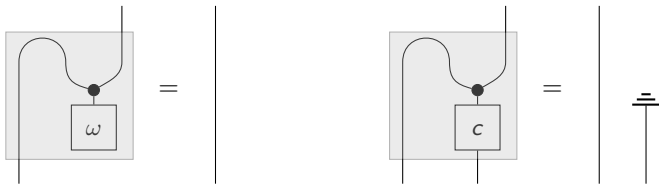
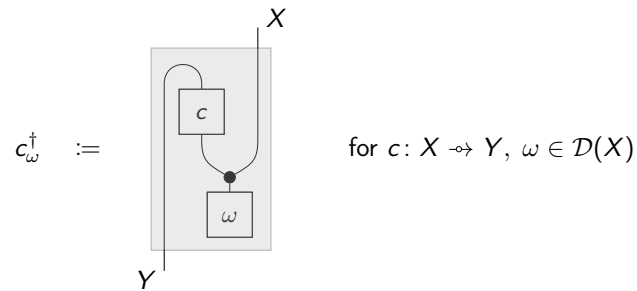
- ▶ Parallel and sequential rules (when h is a proper channel)



- ▶ Multiple normalisation:

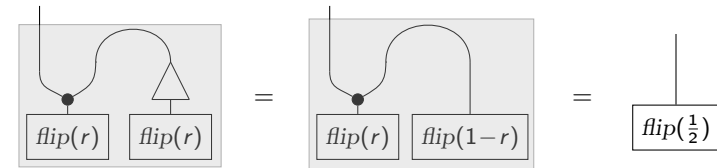


Dagger and box removal

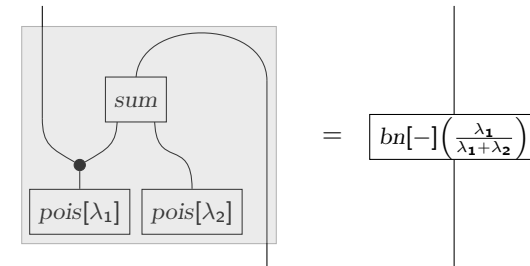


Two illustrations

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

Introduction

Multisets

Probability distributions

Channels / Kleisli maps

Draw distributions

Probabilistic updating

Updating in graphical form

Conclusions

Concluding remarks

- ▶ **Message:** there is so much beautiful (unexplored, 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 ± 1000 pages):
 - www.cs.ru.nl/B.Jacobs/PAPERS/ProbabilisticReasoning.pdf
 - To appear with CUP, after this summer.

