Drawing from an Urn is Isometric

FoSSaCS 2024, Luxemburg, April 9, 2024

Bart Jacobs
Radboud University Nijmegen, bart@cs.ru.nl

Outline

Where we are, so far

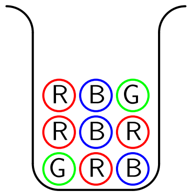
General remarks about drawing from an urn

- 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

Drawing in terms of multisets

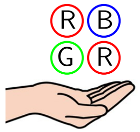
Informally, a multiset is a 'set' in which elements may occur multiple times. Multisets occur frequently in probability theory

- 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 for the different modes

Multisets and distributions — first steps

- For a set X, write:
  - $\mathcal{M}[K](X)$ for the set of multisets of size K with elements from X
  - $\mathcal{D}(X)$ for the set of probability distributions over X
- Hypergeometric K-sized drawing from L-sized urns forms a map:

$$\mathcal{M}[L](X) \xrightarrow{\;hg[K]\;} \mathcal{D}(\mathcal{M}[K](X))$$

  (with restriction: $K \leq L$)
- Pólya drawing has the same form:

$$\mathcal{M}[L](X) \xrightarrow{\;pol[K]\;} \mathcal{D}(\mathcal{M}[K](X))$$

- For multinomial (draw-replace) drawing one may describe the urn as a distribution, giving:

$$\mathcal{D}(X) \xrightarrow{\;mn[K]\;} \mathcal{D}(\mathcal{M}[K](X))$$

Adding metric structure

- If X is a metric space, then so are $\mathcal{M}[K](X)$ and $\mathcal{D}(X)$
  - this involves the Wasserstein metric, see later for details
- A function $f: X \to Y$ is an isometry if it preserves the metric on-the-nose, i.e. for all $x, x' \in X$,

$$d_Y\big(f(x), f(x')\big) \;=\; d_X\big(x, x'\big).$$

- The main result is that all drawing maps are isometries in:

$$\mathcal{D}(X) \xrightarrow{\;mn[K]\;} \mathcal{D}(\mathcal{M}[K](X)) \underset{pol[K]}{\overset{hg[K]}{\leftleftarrows}} \mathcal{M}[L](X)$$

  In the middle this involves a complicated "Wasserstein over Wasserstein" distance
- Drawing from an urn is thus spectacularly well-behaved

A categorical perspective

- Earlier (own) results (LICS'21):
  - draw maps are natural transformations — in the set of colours
  - even monoidal transformations
- These result appear in a categorical perspective on probability theory
  - they have not emerged earlier in the probability literature
- Also the present isometry results benefit/arise from this categorical perspective
- The new, general approach of categorical probability theory (Fritz, Staton, ...) also makes use of string diagrams for clarification
  - boxes are channels (Kleisli maps)
  - they are not used here — but could be

Where we are, so far

Multisets

- 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

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 $x_i \in X$.
  - alternative, 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:

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

From lists to multisets, and back

- Write $\|\varphi\|$ 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 the subset of multisets of size $K \in \mathbb{N}$
- There is an accumulation function

$$X^K \xrightarrow{acc} \mathcal{M}[K](X) \quad \text{e.g.} \quad acc\big(a,b,a,c,c\big) = 2|a\rangle + 1|b\rangle + 2|c\rangle$$

- In the other direction there is a probabilistic function (Kleisli map, channel)

$$\mathcal{M}[K](X) \xrightarrow{arr} \mathcal{D}(X^K) \qquad \text{or} \qquad \mathcal{M}[K](X) \xrightarrow{arr}\!\!\!\!\circ\;\; X^K$$

It assigns to a multiset φ a uniform distribution over all lists that accumulate to φ.

- $acc \odot arr = \mathrm{id}$, where $\odot$ is Kleisli composition, in $\mathcal{K}\ell(\mathcal{D})$

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

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

- $\mathcal{D}(f): \mathcal{D}(X) \to \mathcal{D}(Y)$ and $\mathcal{M}(f): \mathcal{M}(X) \to \mathcal{M}(Y)$
- Explicitly:

$$\mathcal{D}(f)\Big(\textstyle\sum_i r_i|\, x_i\,\rangle\Big) \coloneqq \textstyle\sum_i r_i\big|\, f(x_i)\big\rangle \qquad \text{and similarly for } \mathcal{M}$$

- Functoriality is used for marginalisation of 'joint' distribution $\tau \in \mathcal{D}(X \times Y)$
- Via projections $X \xleftarrow{\pi_1} X \times Y \xrightarrow{\pi_2} Y$ we get:

$$\begin{cases} \mathcal{D}(\pi_1)(\tau) \in \mathcal{D}(X) \\ \mathcal{D}(\pi_2)(\tau) \in \mathcal{D}(Y) \end{cases}$$

- Given $\omega, \omega' \in \mathcal{D}(X)$, one calls $\tau \in \mathcal{D}(X \times X)$ a coupling of ω, ω' if τ has ω, ω' as marginals

Tensors and pushforward of distributions

Parallel product / tensor of $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$

- It forms a new distributions $\omega \otimes \rho \in \mathcal{D}(X \times Y)$
- Defined pointwise as: $(\omega \otimes \rho)(x, y) \coloneqq \omega(x) \cdot \rho(y)$
- This $\omega \otimes \rho$ is a coupling of ω, ρ

Pushforward along a channel $c: X \to \mathcal{D}(Y)$

- A distribution $\omega \in \mathcal{D}(X)$ is pushed along the channel to $c \gg\!= \omega \in \mathcal{D}(Y)$
- Explicitly, $(c \gg\!= \omega)(y) \coloneqq \sum_x \omega(x) \cdot c(x)(y)$
- This pushforward is Kleisli extension

Predicates and their validity

For a distribution $\omega \in \mathcal{D}(X)$ and a 'factor' $p: X \to \mathbb{R}_{\geq 0}$ we write:

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

This is validity or expected value of p in ω.

Where we are, so far

Metric spaces and their maps

- A metric space (X, d) is a set with a distance function $d\colon X \times X \to \mathbb{R}_{\geq 0}$
  - Examples: numbers $\mathbb{N}, \mathbb{R}$ with Euclidean distance $d(r,s) = |r - s|$
  - Discrete metrics space $d(x, x') = 1$ when $x \neq x'$
- For product space $X_1 \times X_2$ we use the sum metric:

$$d_{X_1 \times X_2}\Big((x_1, x_2), (x_1', x_2')\Big) := d_{X_1}(x_1, x_1') + d_{X_2}(x_2, x_2')$$

Maps of metric spaces $f\colon X \to Y$

(1) f is called M-Lipschitz, for $M \in \mathbb{R}_{>0}$, if for all $x, x' \in X$,

$$d_Y\big(f(x), f(x')\big) \leq M \cdot d_X(x, x').$$

(2) When $M = 1$, the map f is called short or non-expansive

(3) When $\leq$ in (1) is $=$, this f is called isometric, or an isometry

The Wasserstein metric between distributions

For distributions $\omega, \omega' \in \mathcal{D}(X)$ on a metric space X there are three equivalent ways to define the Wasserstein / Kantorovic / Monge distance between them:

$$\begin{aligned} d(\omega, \omega') &:= \bigwedge_{\tau \text{ is coupling of } \omega, \omega'} \tau \models d_X \\ &= \bigvee_{p, p'\colon X \to \mathbb{R},\, p \oplus p' \leq d_X} \omega \models p + \omega' \models p' \\ &= \bigvee_{q\colon X \to \mathbb{R}_{\geq 0} \text{ short}} \big|\, \omega \models q - \omega' \models q \,\big|. \end{aligned}$$

where $(p \oplus p')(x, x') = p(x) + p'(x')$.

This forms a metric that is widely used in e.g. program semantics and machine learning

The Wasserstein distance between multisets

For multisets $\varphi, \varphi' \in \mathcal{M}[K](X)$ of the same size on a metric space X there is a similar Wasserstein distance:

$$\begin{aligned} d(\varphi, \varphi') &:= \bigwedge_{\tau \text{ is coupling of } \varphi, \varphi'} \mathit{Flrn}(\tau) \models d_X \\ &= \bigwedge_{\vec{x} \in acc^{-1}(\varphi),\, \vec{y} \in acc^{-1}(\varphi')} \tfrac{1}{K} \cdot d_{X^K}(\vec{x}, \vec{y}) \\ &= \bigwedge_{\vec{x} \in acc^{-1}(\varphi),\, \vec{y} \in acc^{-1}(\varphi')} \sum_{1 \leq i \leq K} \tfrac{1}{K} \cdot d_X(x_i, y_i). \end{aligned}$$

Basic results about Wasserstein

Theorem

(1) The tensor $\otimes\colon \mathcal{D}(X) \times \mathcal{D}(Y) \to \mathcal{D}(X \times Y)$ is isometric

(2) The K-fold tensor $\omega \mapsto \omega^K$ as map $\mathcal{D}(X) \to \mathcal{D}\big(X^K\big)$ is K-Lipschitz

(3) Frequentist learning $\mathit{Flrn}\colon \mathcal{M}[K](X) \to \mathcal{D}(X)$ is isometric

(4) Accumulation $acc\colon X^K \to \mathcal{M}[K](X)$ if $\frac{1}{K}$-Lipschitz

(5) Arrangement $arr\colon \mathcal{M}[K](X) \to \mathcal{D}\big(X^K\big)$ is K-Lipschitz

(6) if $f\colon X \to Y$ is M-Lipschitz, then so is $\mathcal{D}(f)\colon \mathcal{D}(X) \to \mathcal{D}(Y)$

(7) if $c\colon X \to \mathcal{D}(Y)$ is M-Lipschitz, then so is $c \gg\!\!= (-)\colon \mathcal{D}(X) \to \mathcal{D}(Y)$

Where we are, so far

Drawing from an urn

- Recall the types of multinomial, hypergeometric and Pólya drawing:

$$\mathcal{D}(X) \xrightarrow{mn[K]} \mathcal{D}\big(\mathcal{M}[K](X)\big) \underset{pol[K]}{\overset{hg[K]}{\leftleftarrows}} \mathcal{M}[L](X)$$

- They all interact nicely with frequentist learning $Flrn$, as in:

$$\begin{aligned} Flrn \gg\!= mn[K](\omega) &= \omega \\ Flrn \gg\!= hg[K](\upsilon) &= Flrn(\upsilon) \\ Flrn \gg\!= pol[K](\upsilon) &= Flrn(\upsilon). \end{aligned}$$

- This gives one inequality-part of the isometry:

$$\begin{aligned} d\big(\omega, \omega'\big) &= d\big(Flrn \gg\!= mn[K](\omega), Flrn \gg\!= mn[K](\omega')\big) \\ &\leq d\big(mn[K](\omega), mn[K](\omega')\big) \end{aligned}$$

And similarly for hypergeomtric and Pólya

Main result

Theorem

Multinomial, hypergeometric and Pólya drawing are isometric, as maps:

$$\mathcal{D}(X) \xrightarrow{mn[K]} \mathcal{D}\big(\mathcal{M}[K](X)\big) \underset{pol[K]}{\overset{hg[K]}{\leftleftarrows}} \mathcal{M}[L](X)$$

Proof for multinomial $mn[K](\omega) \coloneqq \mathcal{D}(acc)\big(\omega^K\big)$

Only shortness is needed. $\omega \mapsto \omega^K$ is K-Lipschitz and acc is $\frac{1}{K}$-Lipschitz.
The composition is then $K \cdot \frac{1}{K} = 1$-Lipschitz. QED

The proof for hypergeometric is more work, and for even more for Pólya.

Isometry illustration, for multinomial, part I

- Consider the distributions $\omega, \omega' \in \mathcal{D}(\mathbb{N})$.

$$\omega = \tfrac{1}{3}|0\rangle + \tfrac{2}{3}|2\rangle \quad \text{and} \quad \omega' = \tfrac{1}{2}|1\rangle + \tfrac{1}{2}|2\rangle \quad \text{with} \quad d(\omega, \omega') = \tfrac{1}{2}$$

- There are 10 multisets of size 3 over $\{0, 1, 2\}$:

$$\begin{array}{llll} \varphi_1 = 3|0\rangle & \varphi_2 = 2|0\rangle + 1|1\rangle & \varphi_3 = 1|0\rangle + 2|1\rangle & \varphi_4 = 3|1\rangle \\ \varphi_5 = 2|0\rangle + 1|2\rangle & \varphi_6 = 1|0\rangle + 1|1\rangle + 1|2\rangle & \varphi_7 = 2|1\rangle + 1|2\rangle & \\ \varphi_8 = 1|0\rangle + 2|2\rangle & \varphi_9 = 1|1\rangle + 2|2\rangle & \varphi_{10} = 3|2\rangle. & \end{array}$$

- The multinomial distributions are:

$$\begin{aligned} mn[3](\omega) &= \tfrac{1}{27}\big|\varphi_1\big\rangle + \tfrac{2}{9}\big|\varphi_5\big\rangle + \tfrac{4}{9}\big|\varphi_8\big\rangle + \tfrac{8}{27}\big|\varphi_{10}\big\rangle \\ mn[3](\omega') &= \tfrac{1}{8}\big|\varphi_4\big\rangle + \tfrac{3}{8}\big|\varphi_7\big\rangle + \tfrac{3}{8}\big|\varphi_9\big\rangle + \tfrac{1}{8}\big|\varphi_{10}\big\rangle. \end{aligned}$$

Isometry illustration, for multinomial, part II

- The 'optimal' coupling $\tau \in \mathcal{D}\big(\mathcal{M}[3](\mathbb{N}) \times \mathcal{M}[3](\mathbb{N})\big)$ between the multinomial distributions is:
$$\tau = \tfrac{1}{27}\Big|\varphi_1, \varphi_4\Big\rangle + \tfrac{19}{216}\Big|\varphi_5, \varphi_4\Big\rangle + \tfrac{1}{8}\Big|\varphi_{10}, \varphi_{10}\Big\rangle + \tfrac{29}{216}\Big|\varphi_5, \varphi_7\Big\rangle + \tfrac{5}{72}\Big|\varphi_8, \varphi_7\Big\rangle + \tfrac{3}{8}\Big|\varphi_8, \varphi_9\Big\rangle + \tfrac{37}{216}\Big|\varphi_{10}, \varphi_7\Big\rangle.$$
- The distance between the multinomial distributions, using $d_{\mathcal{M}} = d_{\mathcal{M}[3](\mathbb{N})}$, is:

$$\begin{aligned} d\big(mn[3](\omega), mn[3](\omega')\big) &= \tau \models d_{\mathcal{M}} \\ &= \tfrac{1}{27}\cdot d_{\mathcal{M}}\big(\varphi_1, \varphi_4\big) + \tfrac{19}{216}\cdot d_{\mathcal{M}}\big(\varphi_5, \varphi_4\big) + \tfrac{1}{8}\cdot d_{\mathcal{M}}\big(\varphi_{10}, \varphi_{10}\big) + \tfrac{29}{216}\cdot d_{\mathcal{M}}\big(\varphi_5, \varphi_7\big) \\ &\quad + \tfrac{5}{72}\cdot d_{\mathcal{M}}\big(\varphi_8, \varphi_7\big) + \tfrac{3}{8}\cdot d_{\mathcal{M}}\big(\varphi_8, \varphi_9\big) + \tfrac{37}{216}\cdot d_{\mathcal{M}}\big(\varphi_{10}, \varphi_7\big) \\ &= \tfrac{1}{27}\cdot 1 + \tfrac{19}{216}\cdot 1 + \tfrac{1}{8}\cdot 0 + \tfrac{29}{216}\cdot\tfrac{2}{3} + \tfrac{5}{72}\cdot\tfrac{2}{3} + \tfrac{3}{8}\cdot\tfrac{1}{3} + \tfrac{37}{216}\cdot\tfrac{2}{3} = \tfrac{1}{2} \text{ !!} \end{aligned}$$

- This Wasserstein-over-Wasserstein computation is much more complex, but still gives the same outcome

Concluding remarks

- Drawing from an urn is mathematically incredibly well-behaved
  - the isometry results give a glimpse of "Plato's heaven"
- Are the isometry results usefull, in applications?
  - Do they need to be?
  - In machine learning one sometimes uses a "ground distance" between colours in experiments in psychophysics
  - Possible applications in sensitivity analysis
- Extensions to infinite discrete distributions exist and give similar results, e.g.
$$d\Big(pois[\lambda_1], pois[\lambda_2]\Big) = \big|\,\lambda_1 - \lambda_2\,\big|$$
- Extensions to continuous probability theory are less clear