Drawing from an Urn is Isometric

Bart Jacobs(✉)

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

Abstract. Drawing (a multiset of) coloured balls from an urn is one of the most basic models in discrete probability theory. Three modes of drawing are commonly distinguished: multinomial (draw-replace), hypergeometric (draw-delete), and Pólya (draw-add). These drawing operations are represented as maps from urns to distributions over multisets of draws. The set of urns is a metric space via the Wasserstein distance. The set of distributions over draws is also a metric space, using Wasserstein-over-Wasserstein. The main result of this paper is that the three draw operations are all isometries, that is, they preserve the Wasserstein distances.

Keywords: probability · urn drawing · Wasserstein distance.

1 Introduction

We start with an illustration of the topic of this paper. We consider a situation with a set $C = \{R, G, B\}$ of three colours: red, green, blue. Assume that we have two urns υ_1, υ_2 with 10 coloured balls each. We describe these urns as multisets of the form:

$$\upsilon_1 = 8|G\rangle + 2|B\rangle \qquad \text{and} \qquad \upsilon_2 = 5|R\rangle + 4|G\rangle + 1|B\rangle.$$

Recall that a multiset is like a set, except that elements may occur multiple times. Here we describe urns as multisets using 'ket' notation $| - \rangle$. It separates multiplicities of elements (before the ket) from the elements in the multiset (inside the ket). Thus, urn υ_1 contains 8 green balls and 2 blue balls (and no red ones). Similarly, urn υ_2 contains 5 red, 4 green, and 1 blue ball(s).

Below, we shall describe the Wasserstein distance between multisets (of the same size). How this works does not matter for now; we simply posit that the Wasserstein distance $d(\upsilon_1, \upsilon_2)$ between these two urns is $\frac{1}{2}$ — where we assume the discrete distance on the set C of colours.

We turn to draws from these two urns, in this introductory example of size two. These draws are also described as multisets, with elements from the set $C = \{R, G, B\}$ of colours. There are six multisets (draws) of size 2, namely:

$$2|R\rangle \quad 1|R\rangle + 1|G\rangle \quad 2|G\rangle \quad 1|R\rangle + 1|B\rangle \quad 2|B\rangle \quad 1|G\rangle + 1|B\rangle. \tag{1}$$

As we see, there are three draws with 2 balls of the same colour, and three draws with balls of different colours.

© The Author(s) 2024

N. Kobayashi and J. Worrell (Eds.): FoSSaCS 2024, LNCS 14574, pp. 101–120, 2024.
https://doi.org/10.1007/978-3-031-57228-9_6

We consider the hypergeometric probabilities associated with these draws, from the two urns. Let's illustrate this for the draw $1|G\rangle + 1|B\rangle$ of one green ball and one blue ball from the urn v_1. The probability of drawing $1|G\rangle + 1|B\rangle$ is $\frac{16}{45}$; it is obtained as sum of:

- first drawing-and-deleting a green ball from $v_1 = 8|G\rangle + 2|B\rangle$, with probability $\frac{8}{10}$. It leaves an urn $7|G\rangle + 2|B\rangle$, from which we can draw a blue ball with probability $\frac{2}{9}$. Thus drawing "first green then blue" happens with probability $\frac{8}{10} \cdot \frac{2}{9} = \frac{8}{45}$.
- Similarly, the probability of drawing "first blue then green" is $\frac{2}{10} \cdot \frac{8}{9} = \frac{8}{45}$.

We can similarly compute the probabilities for each of the above six draws (1) from urn v_1. This gives the hypergeometric distribution, which we write using kets-over-kets as:

$$\mathit{hg}[2](v_1) \;=\; \tfrac{28}{45}\Big|\, 2|G\rangle \Big\rangle + \tfrac{16}{45}\Big|\, 1|G\rangle + 1|B\rangle \Big\rangle + \tfrac{1}{45}\Big|\, 2|B\rangle \Big\rangle .$$

The fraction written before a big ket is the probability of drawing the multiset (of size 2), written inside that big ket, from the urn v_1.

Drawing from the second urn v_2 gives a different distribution over these multisets (1). Since urn v_2 contains red balls, they additionally appear in the draws.

$$\begin{aligned}\mathit{hg}[2](v_2) \;=\; & \tfrac{2}{9}\Big|\, 2|R\rangle \Big\rangle + \tfrac{4}{9}\Big|\, 1|R\rangle + 1|G\rangle \Big\rangle + \tfrac{2}{15}\Big|\, 2|G\rangle \Big\rangle \\ & + \tfrac{1}{9}\Big|\, 1|R\rangle + 1|B\rangle \Big\rangle + \tfrac{4}{45}\Big|\, 1|G\rangle + 1|B\rangle \Big\rangle .\end{aligned}$$

We can also compute the distance between these two hypergeometric distributions over multisets. It involves a Wasserstein distance, over the space of multisets (of size 2) with their own Wasserstein distance. Again, details of the calculation are skipped at this stage. The distance between the above two hypergeometric draw-distributions is:

$$d\Big(\mathit{hg}[2](v_1),\, \mathit{hg}[2](v_2)\Big) \;=\; \tfrac{1}{2} \;=\; d\big(v_1,\, v_2\big).$$

This coincidence of distances is non-trivial. It holds, in general, for arbitrary urns (of the same size) over arbitrary metric spaces of colours, for draws of arbitrary sizes. Moreover, the same coincidence of distances holds for the multinomial and Pólya modes of drawing. These coincidences are the main result of this paper, see Theorems 1, 2, and 3 below.

In order to formulate and obtain these results, we describe multinomial, hypergeometric and Pólya distributions in the form of (Kleisli) maps:

$$\begin{tikzcd} \mathcal{D}(X) \arrow[r, "mn{[K]}"] & \mathcal{D}\big(\mathcal{M}[K](X)\big) & \mathcal{M}[L](X) \arrow[l, shift right, "hg{[K]}"'] \arrow[l, shift left, "pl{[K]}"] \end{tikzcd} \qquad (2)$$

They all produce distributions (indicated by $\mathcal{D}$), in the middle of this diagram, on multisets (draws) of size K, indicated by $\mathcal{M}[K]$, over a set X of colours. Details will be provided below. Using the maps in (2), the coincidence of distances that we saw above can be described as a preservation property, in terms of distance preserving maps — called isometries. At this stage we wish to emphasise that the representation of these different drawing operations as maps in (2) has a categorical background. It makes it possible to formulate and prove basic properties of drawing from an urn, such as naturality in the set X of colours. Also, as shown in [8] for the multinomial and hypergeometric case, drawing forms a monoidal transformation (with ‘zipping’ for multisets as coherence map). This paper demonstrates that the three draw maps (2) are even more well-behaved: they are all isometries, that is, they preserve Wasserstein distances. This is a new and amazing fact.

This paper concentrates on the mathematics behind these isometry results, and not on interpretations or applications. We do like to refer to interpretations in machine learning [14] where the distance that we consider on colours in an urn is called the *ground distance*. Actual distances between colours are used there, based on experiments in psychophysics, using perceived differences [16].

The Wasserstein — or Wasserstein-Kantorovich, or Monge-Kantorovich — distance is the standard distance on distributions and on multisets, going back to [12]. After some preliminaries on multisets and distributions, and on distances in general, Sections 4 and 5 of this paper recall the Wasserstein distance on distributions and on multisets, together with some basic results. The three subsequent Sections 6 – 8 demonstrate that multinomial, hypergeometric and Pólya drawing are all isometric. Distances occur on multiple levels: on colours, on urns (as multisets or distributions) and on draw-distributions. This may be confusing, but many illustrations are included.

2 Preliminaries on multisets and distributions

A *multiset* over a set X is a finite formal sum of the form $\sum_i n_i|x_i\rangle$, for elements $x_i \in X$ and natural numbers $n_i \in \mathbb{N}$ describing the multiplicities of these elements x_i. We shall write $\mathcal{M}(X)$ for the set of such multisets over X. A multiset $\varphi \in \mathcal{M}(X)$ may equivalently be described in functional form, as a function $\varphi\colon X \to \mathbb{N}$ with finite support: $\mathit{supp}(\varphi) := \{x \in X \mid \varphi(x) \neq 0\}$. Such a function $\varphi\colon X \to \mathbb{N}$ can be written in ket form as $\sum_{x\in X} \varphi(x)|x\rangle$. We switch back-and-forth between the ket and functional form and use the formulation that best suits a particular situation.

For a multiset $\varphi \in \mathcal{M}(X)$ we write $\|\varphi\| \in \mathbb{N}$ for the *size* of the multiset. It is the total number of elements, including multiplicities:

$$\|\varphi\| := \sum_{x\in X} \varphi(x).$$

For a number $K \in \mathbb{N}$ we write $\mathcal{M}[K](X) \subseteq \mathcal{M}(X)$ for the subset of multisets of size K. There are ‘accumulation’ maps $\mathit{acc}\colon X^K \to \mathcal{M}[K](X)$ turning lists into multisets via $\mathit{acc}\big(x_1, \ldots, x_K\big) := 1|x_1\rangle + \cdots + 1|x_K\rangle$. For instance $\mathit{acc}\big(c, b, a, c, a, c\big) = 2|a\rangle + 1|b\rangle + 3|c\rangle$. A standard result (see [10]) is that for a multiset $\varphi \in \mathcal{M}[K](X)$ there are $(\!(\varphi)\!) := \frac{K!}{\varphi\,!}$ many sequences $\boldsymbol{x} \in X^K$ with $\mathit{acc}(\boldsymbol{x}) = \varphi$, where $\varphi\,! = \prod_x \varphi(x)!$.

Multisets $\varphi, \psi \in \mathcal{M}(X)$ can be added and compared elementwise, so that $\big(\varphi + \psi\big)(x) = \varphi(x) + \psi(x)$ and $\varphi \leq \psi$ means $\varphi(x) \leq \psi(x)$ for all $x \in X$. In the latter case, when $\varphi \leq \psi$, we can also subtract $\psi - \varphi$ elementwise.

The mapping $X \mapsto \mathcal{M}(X)$ is functorial: for a function $f\colon X \to Y$ we have $\mathcal{M}(f)\colon \mathcal{M}(X) \to \mathcal{M}(Y)$ given by $\mathcal{M}(f)(\varphi)(y) = \sum_{x \in f^{-1}(y)} \varphi(x)$. This map $\mathcal{M}(f)$ preserves sums and size.

For a multiset $\tau \in \mathcal{M}(X \times Y)$ on a product set we can take its two marginals $\mathcal{M}(\pi_1)(\tau) \in \mathcal{M}(X)$ and $\mathcal{M}(\pi_2)(\tau) \in \mathcal{M}(Y)$ via functoriality, using the two projection functions $\pi_1\colon X \times Y \to X$ and $\pi_2\colon X \times Y \to Y$. Starting from $\varphi \in \mathcal{M}(X)$ and $\psi \in \mathcal{M}(Y)$, we say that $\tau \in \mathcal{M}(X \times Y)$ is a *coupling* of φ, ψ if φ and ψ are the two marginals of τ. We define the *decoupling* map:

$$\mathcal{M}(X \times Y) \xrightarrow{\;\mathit{dcpl} \,:=\, \langle \mathcal{M}(\pi_1), \mathcal{M}(\pi_2) \rangle\;} \mathcal{M}(X) \times \mathcal{M}(Y) \tag{3}$$

The inverse image $\mathit{dcpl}^{-1}(\varphi, \psi) \subseteq \mathcal{M}(X \times Y)$ is thus the subset of couplings of φ, ψ.

A *distribution* is a finite formal sum of the form $\sum_i r_i |x_i\rangle$ with multiplicities $r_i \in [0,1]$ satisfying $\sum_i r_i = 1$. Such a distribution can equivalently be described as a function $\omega\colon X \to [0,1]$ with finite support, satisfying $\sum_x \omega(x) = 1$. We write $\mathcal{D}(X)$ for the set of distributions on X. This $\mathcal{D}$ is functorial, in the same way as $\mathcal{M}$. Both $\mathcal{D}$ and $\mathcal{M}$ are monads on the category $\mathbf{Sets}$ of sets and functions, but we only use this for $\mathcal{D}$. The unit and multiplication / flatten maps $\mathit{unit}\colon X \to \mathcal{D}(X)$ and $\mathit{flat}\colon \mathcal{D}^2(X) \to \mathcal{D}(X)$ are given by:

$$\mathit{unit}(x) \,:=\, 1|x\rangle \qquad\qquad \mathit{flat}(\Omega) \,:=\, \sum_{x \in X} \Bigg(\sum_{\omega \in \mathcal{D}(X)} \Omega(\omega) \cdot \omega(x) \Bigg) \, |x\rangle. \tag{4}$$

Kleisli maps $c\colon X \to \mathcal{D}(Y)$ are also called channels and written as $c\colon X \rightsquigarrow Y$. Kleisli extension $c \gg\!= (-)\colon \mathcal{D}(X) \to \mathcal{D}(Y)$ for such a channel, is defined on $\omega \in \mathcal{D}(X)$ as:

$$c \gg\!= \omega \,:=\, \mathit{flat}\big(\mathcal{D}(c)(\omega)\big) \,=\, \sum_{y \in Y} \Bigg(\sum_{x \in X} \omega(x) \cdot c(x)(y) \Bigg) \, |y\rangle.$$

Channels $c\colon X \rightsquigarrow Y$ and $d\colon Y \rightsquigarrow Z$ can be composed to $d \odot c\colon X \rightsquigarrow Z$ via $(d \odot c)(x) := d \gg\!= c(x)$. Each function $f\colon X \to Y$ gives rise to a deterministic channel $\langle f \rangle := \mathit{unit} \circ f\colon X \rightsquigarrow Y$, that is, via $\langle f \rangle(x) = 1\big| f(x) \big\rangle$.

An example of a channel is arrangement $\mathit{arr}\colon \mathcal{M}[K](X) \to \mathcal{D}(X^K)$. It maps a multiset $\varphi \in \mathcal{M}[K](X)$ to the uniform distribution of sequences that accumulate to φ.

$$\mathit{arr}(\varphi) \,:=\, \sum_{\boldsymbol{x} \in acc^{-1}(\varphi)} \frac{1}{\boldsymbol{(}\varphi\boldsymbol{)}} \big|\boldsymbol{x}\big\rangle \,=\, \sum_{\boldsymbol{x} \in acc^{-1}(\varphi)} \frac{\varphi!}{K!} \big|\boldsymbol{x}\big\rangle. \tag{5}$$

One can show that $\langle \mathit{acc} \rangle \odot \mathit{arr} = \mathcal{D}(\mathit{acc}) \circ \mathit{arr} = \mathit{unit}\colon \mathcal{M}[K](X) \to \mathcal{D}\big(\mathcal{M}[K](X)\big)$. The composite in the other direction produces the uniform distribution of all permutations of a sequence:

$$\mathit{arr} \odot \langle \mathit{acc} \rangle \,=\, \mathit{arr} \circ \mathit{acc} \,=\, \mathit{prm} \qquad \text{where} \qquad \mathit{prm}(\boldsymbol{x}) \,:=\, \sum_{t\colon K \overset{\cong}{\to} K} \frac{1}{K!} \big|\underline{t}(\boldsymbol{x})\big\rangle, \tag{6}$$

in which $\underline{t}(x_1, \ldots, x_K) := (x_{t(1)}, \ldots, x_{t(K)})$. In writing $t\colon K \overset{\cong}{\to} K$ we implicitly identify the number K with the set $\{1, \ldots, K\}$.

Each multiset $\varphi \in \mathcal{M}(X)$ of non-zero size can be turned into a distribution via normalisation. This operation is called frequentist learning, since it involves learning a distribution from a multiset of data, via counting. Explicitly:

$$\mathit{Flrn}(\varphi) := \sum_{x \in X} \frac{\varphi(x)}{\|\varphi\|} \, |x\rangle.$$

For instance, if we learn from an urn with three red, two green and five blue balls, we get the probability distribution for drawing a ball of a particular colour from the urn:

$$\mathit{Flrn}\Big(3|R\rangle + 2|G\rangle + 5|B\rangle\Big) = \tfrac{3}{10}|R\rangle + \tfrac{1}{5}|G\rangle + \tfrac{1}{2}|B\rangle.$$

This map *Flrn* is a natural transformation (but not a map of monads).

Given two distributions $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$, we can form their parallel product $\omega \otimes \rho \in \mathcal{D}(X \times Y)$, given in functional form as:

$$\big(\omega \otimes \rho\big)(x, y) := \omega(x) \cdot \rho(y).$$

Like for multisets, we call a joint distribution $\tau \in \mathcal{D}(X \times Y)$ a *coupling* of $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$ if ω, ρ are the two marginals of τ, that is if, $\mathcal{D}(\pi_1)(\tau) = \omega$ and $\mathcal{D}(\pi_2) = \rho$. We can express this also via a decouple map $\mathit{dcpl} = \langle \mathcal{D}(\pi_1), \mathcal{D}(\pi_2) \rangle$ as in (3).

An *observation* on a set X is a function of the form $p\colon X \to \mathbb{R}$. Such a map p, together with a distribution $\omega \in \mathcal{D}(X)$, is called a random variable — but confusingly, the distribution is often left implicit. The map $p\colon X \to \mathbb{R}$ will be called a *factor* if it restricts to non-negative reals $X \to \mathbb{R}_{\geq 0}$. Each element $x \in X$ gives rise to a point observation $\mathbf{1}_x\colon X \to \mathbb{R}$, with $\mathbf{1}_x(x') = 1$ if $x = x'$ and $\mathbf{1}_x(x') = 0$ if $x \neq x'$. For a distribution $\omega \in \mathcal{D}(X)$ and an observation $p\colon X \to \mathbb{R}$ on the same set X we write $\omega \models p$ for the validity (expected value) of p in ω, defined as (finite) sum: $\sum_{x \in X} \omega(x) \cdot p(x)$. We shall write $\mathit{Obs}(X) = \mathbb{R}^X$ and $\mathit{Fact}(X) = (\mathbb{R}_{\geq 0})^X$ for the sets of observations and factors on X.

3 Preliminaries on metric spaces

A metric space will be written as a pair (X, d_X), where X is a set and $d_X\colon X \times X \to \mathbb{R}_{\geq 0}$ is a distance function, also called metric. This metric satisfies:

- $d_X(x, x') = 0$ iff $x = x'$;
- symmetry: $d_X(x, x') = d_X(x', x)$;
- triangular inequality: $d_X(x, x'') \leq d_X(x, x') + d_X(x', x'')$.

Often, we drop the subscript X in d_X if it is clear from the context. We use the standard distance $d(x, y) = |x - y|$ on real and natural numbers.

Definition 1. *Let (X, d_X), (Y, d_Y) be two metric spaces.*

1. *A function $f\colon X \to Y$ is called* short *(or also* non-expansive*) if:*
$$d_Y\big(f(x), f(x')\big) \leq d_X\big(x, x'\big), \qquad \textit{for all } x, x' \in X.$$
Such a map is called an isometry *or an* isometric embedding *if the above inequality $\leq$ is an actual equality $=$. This implies that the function f is injective, and thus an 'embedding'.*
We write $\mathbf{Met}_S$ for the category of metric spaces with short maps between them.
2. *A function $f\colon X \to Y$ is* Lipschitz *or M-*Lipschitz*, if there is a number $M \in \mathbb{R}_{>0}$ such that:*
$$d_Y\big(f(x), f(x')\big) \leq M \cdot d_X\big(x, x'\big), \qquad \textit{for all } x, x' \in X.$$
The number M is sometimes called the Lipschitz constant. Thus, a short function is Lipschitz, with constant 1. We write $\mathbf{Met}_L$ for the category of metric spaces with Lipschitz maps between them (with arbitrary Lipschitz constants).

Lemma 1. *For two metric spaces (X_1, d_1) and (X_2, d_2) we equip the cartesian product $X_1 \times X_2$ of sets with the sum of the two metrics:*
$$d\Big((x_1, x_2), (x_1', x_2')\Big) := d_{X_1}(x_1, x_1') + d_{X_2}(x_2, x_2'). \tag{7}$$
With the usual projections and tuples this forms a product in the category $\mathbf{Met}_L$. ❑

The product $\times$ also exists in the category $\mathbf{Met}_S$ of metric spaces with short maps. There, it forms a *monoidal* product (a tensor $\otimes$) since there are no diagonals. In the setting of $[0,1]$-bounded metrics (with short maps) one uses the maximum instead of the sum (7) in order to form products (possibly infinite). In the category $\mathbf{Met}_L$ the products $X_1 \times X_2$ with maximum and with sum of distances are isomorphic, via the identity maps. This works since for $r, s \in \mathbb{R}_{\geq 0}$ one as $\max(r, s) \leq r + s$ and $r + s \leq 2 \cdot \max(r, s)$.

4 The Wasserstein distance between distributions

This section introduces the Wasserstein distance between probability distributions and recalls some basic results. There are several equivalent formulations for this distance. We express it in terms of validity and couplings, see also *e.g.* [1,3,6,4].

Definition 2. *Let (X, d_X) be a metric space. The* Wasserstein *metric $d\colon \mathcal{D}(X) \times \mathcal{D}(X) \to \mathbb{R}_{\geq 0}$ is defined by any of the three equivalent formulas:*
$$\begin{aligned} d\big(\omega, \omega'\big) &:= \bigwedge_{\tau \in \mathit{dcpl}^{-1}(\omega, \omega')} \tau \models d_X \\ &= \bigvee_{p,\, p' \in \mathit{Obs}(X),\, p \oplus p' \leq d_X} \omega \models p + \omega' \models p' \\ &= \bigvee_{q \in \mathit{Fact}_S(X)} \big|\, \omega \models q - \omega' \models q \,\big|. \end{aligned} \tag{8}$$

This turns $\mathcal{D}(X)$ into a metric space. The operation $\oplus$ in the second formulation is defined as $(p \oplus p')(x, x') = p(x) + p'(x')$. The set Fact$_S(X) \subseteq$ *Fact(X) in the third formulation is the subset of short factors $X \to \mathbb{R}_{\geq 0}$. To be precise, we should write* Fact$_S(X, d_X)$ *since the distance d_X on X is a parameter, but we leave it implicit for convenience. The meet $\bigwedge$ and joins $\bigvee$ in* (8) *are actually reached, by what are called the* optimal *coupling and the* optimal *observations / factor.*

In this definition it is assumed that X is a metric space. This includes the case where X is simply a set, with the discrete metric (where different elements have distance 1). The above Wasserstein distance can then be formulated as what is often called the *total variation distance*. For distributions $\omega, \omega' \in \mathcal{D}(X)$ it is:

$$d(\omega, \omega') = \tfrac{1}{2} \sum_{x \in X} \big| \omega(x) - \omega'(x) \big|.$$

This discrete case is quite common, see *e.g.* [11] and the references given there.

The equivalence of the first and second formulation in (8) is an instance of strong duality in linear programming, which can be obtained via Farkas' Lemma, see *e.g.* [13]. The second formulation is commonly associated with Monge. The single factor q in the third formulation can be obtained from the two observations p, p' in the second formulation, and vice-versa. What we call the Wasserstein distance is also called the Monge-Kantorovich distance.

We do not prove the equivalence of the three formulations for the Wasserstein distance $d(\omega, \omega')$ between two distributions ω, ω' in (8), one with a meet $\bigwedge$ and two with a join $\bigvee$. This is standard and can be found in the literature, see *e.g.* [15]. These three formulations do not immediately suggest how to calculate distances. What helps is that the minimum and maxima are actually reached and can be computed. This is done via linear programming, originally introduced by Kantorovich, see [13,15,3]. In the sequel, we shall see several examples of distances between distributions. They are obtained via our own Python implementation of the linear optimisation, which also produces the optimal coupling, observations or factor. This implementation is used only for illustrations.

Example 1. Consider the set X containing the first eight natural numbers, so $X = \{0, 1, \ldots, 7\} \subseteq \mathbb{N}$, with the usual distance, written as d_X, between natural numbers: $d_X(n, m) = |n - m|$. We look at the following two distributions on X.

$$\omega = \tfrac{1}{2}|0\rangle + \tfrac{1}{2}|4\rangle \qquad \omega' = \tfrac{1}{8}|2\rangle + \tfrac{1}{8}|3\rangle + \tfrac{1}{8}|6\rangle + \tfrac{5}{8}|7\rangle.$$

We claim that the Wasserstein distance $d(\omega, \omega')$ is $\frac{15}{4}$. This will be illustrated for each of the three formulations in Definition 2.

- The optimal coupling $\tau \in \mathcal{D}(X \times X)$ of ω, ω' is:

$$\tau = \tfrac{1}{8}\big|0, 2\big\rangle + \tfrac{1}{8}\big|0, 3\big\rangle + \tfrac{1}{8}\big|0, 6\big\rangle + \tfrac{1}{8}\big|0, 7\big\rangle + \tfrac{1}{2}\big|4, 7\big\rangle.$$

It is not hard to see that τ's first marginal is ω, and its second marginal is ω'. We compute the distances as:

$$\begin{aligned} d(\omega,\omega') &= \tau \models d_X \\ &= \tfrac{1}{8}\cdot d_X(0,2) + \tfrac{1}{8}\cdot d_X(0,3) + \tfrac{1}{8}\cdot d_X(0,6) + \tfrac{1}{8}\cdot d_X(0,7) + \tfrac{1}{2}\cdot d_X(4,7) \\ &= \tfrac{2}{8} + \tfrac{3}{8} + \tfrac{6}{8} + \tfrac{7}{8} + \tfrac{3}{2} = \tfrac{18}{8} + \tfrac{3}{2} = \tfrac{9}{4} + \tfrac{6}{4} = \tfrac{15}{4}. \end{aligned}$$

- There are the following two optimal observations $p, p' \colon X \to \mathbb{R}$, described as sums of weighted point predicates:

$$\begin{aligned} p &= -1\cdot \mathbf{1}_1 - 2\cdot \mathbf{1}_2 - 3\cdot \mathbf{1}_3 - 4\cdot \mathbf{1}_4 - 5\cdot \mathbf{1}_5 - 6\cdot \mathbf{1}_6 - 7\cdot \mathbf{1}_7 \\ p' &= 1\cdot \mathbf{1}_1 + 2\cdot \mathbf{1}_2 + 3\cdot \mathbf{1}_3 + 4\cdot \mathbf{1}_4 + 5\cdot \mathbf{1}_5 + 6\cdot \mathbf{1}_6 + 7\cdot \mathbf{1}_7. \end{aligned}$$

It is not hard to see that $(p \oplus p')(i,j) \coloneqq p(i) + p'(j) \le d_X(i,j)$ holds for all $i, j \in X$. Using the second formulation in (8) we get:

$$\begin{aligned} &\big(\omega \models p\big) + \big(\omega' \models p'\big) \\ &= \tfrac{1}{2}\cdot p(0) + \tfrac{1}{2}\cdot p(4) + \tfrac{1}{8}\cdot p'(2) + \tfrac{1}{8}\cdot p'(3) + \tfrac{1}{8}\cdot p'(6) + \tfrac{5}{8}\cdot p'(7) \\ &= \tfrac{-4}{2} + \tfrac{2}{8} + \tfrac{3}{8} + \tfrac{6}{8} + \tfrac{35}{8} = -2 + \tfrac{46}{8} = \tfrac{30}{8} = \tfrac{15}{4}. \end{aligned}$$

- Finally, there is a (single) short factor $q \colon X \to \mathbb{R}_{\ge 0}$ given by:

$$q = 7\cdot \mathbf{1}_0 + 6\cdot \mathbf{1}_1 + 5\cdot \mathbf{1}_2 + 4\cdot \mathbf{1}_3 + 3\cdot \mathbf{1}_4 + 2\cdot \mathbf{1}_5 + 1\cdot \mathbf{1}_6.$$

Then:

$$\begin{aligned} &\big(\omega \models q\big) - \big(\omega' \models q\big) \\ &= \tfrac{1}{2}\cdot q(0) + \tfrac{1}{2}\cdot q(4) - \Big(\tfrac{1}{8}\cdot q(2) + \tfrac{1}{8}\cdot q(3) + \tfrac{1}{8}\cdot q(6) + \tfrac{5}{8}\cdot q(7)\Big) \\ &= \tfrac{7}{2} + \tfrac{3}{2} - \Big(\tfrac{5}{8} + \tfrac{4}{8} + \tfrac{1}{8}\Big) = \tfrac{10}{2} - \tfrac{10}{8} = \tfrac{20}{4} - \tfrac{5}{4} = \tfrac{15}{4}. \end{aligned}$$

From the fact that the coupling τ, the two observations p, p', and the single factor q produce the same distance one can deduce that they are optimal, using the formula (8).

We proceed with several standard properties of the Wasserstein distance on distributions.

Lemma 2. *In the context of Definition 2, the following properties hold.*

1. *For an M-Lipschitz function $f \colon X \to Y$, the pushforward map $\mathcal{D}(f) \colon \mathcal{D}(X) \to \mathcal{D}(Y)$ is also M-Lipschitz; as a result, $\mathcal{D}$ lifts to a functor $\mathcal{D} \colon \mathbf{Met}_L \to \mathbf{Met}_L$, and also to $\mathcal{D} \colon \mathbf{Met}_S \to \mathbf{Met}_S$.*
2. *If $f \colon X \to Y$ is an isometry, then so is $\mathcal{D}(f) \colon \mathcal{D}(X) \to \mathcal{D}(Y)$.*
3. *For an M-Lipschitz factor $q \colon X \to \mathbb{R}_{\ge 0}$, the validity-of-$q$ factor $(-) \models q \colon \mathcal{D}(X) \to \mathbb{R}_{\ge 0}$ is also M-Lipschitz.*
4. *For each element $x \in X$ and distribution $\omega \in \mathcal{D}(X)$ one has: $d\big(1|x\rangle, \omega\big) = \omega \models d_X(x, -)$; especially, $d\big(1|x\rangle, 1|x'\rangle\big) = d_X(x, x')$, making the map* unit*: $X \to \mathcal{D}(X)$ an isometry.*

5. *The monad multiplication flat*: $\mathcal{D}^2(X) \to \mathcal{D}(X)$ *is short, so that* $\mathcal{D}$ *lifts from a monad on* **Sets** *to a monad on* $\mathbf{Met}_S$ *and on* $\mathbf{Met}_L$.
6. *If a channel* $c\colon X \to \mathcal{D}(Y)$ *is* M*-Lipschitz, then so is its Kleisli extension* $c \gg= (-) := \mathit{flat} \circ \mathcal{D}(c)\colon \mathcal{D}(X) \to \mathcal{D}(Y)$.
7. *If channel* $c\colon X \rightsquigarrow Y$ *is* M*-Lipschitz and channel* $d\colon Y \rightsquigarrow Z$ *is* K*-Lipschitz, then their (channel) composite* $d \odot c\colon X \rightsquigarrow Z$ *is* $(M \cdot K)$*-Lipschitz.*
8. *For distributions* $\omega_i, \omega_i' \in \mathcal{D}(X)$ *and numbers* $r_i \in [0,1]$ *with* $\sum_i r_i = 1$ *one has:*

$$d\Big(\textstyle\sum_i r_i \cdot \omega_i,\, \sum_i r_i \cdot \omega_i'\Big) \;\leq\; \textstyle\sum_i r_i \cdot d\big(\omega_i, \omega_i'\big).$$

9. *The permutation channel prm*: $X^K \to \mathcal{D}(X^K)$ *from* (6) *is short.*

Proof. We skip the first two points since they are standard.

3. Let $q\colon X \to \mathbb{R}_{\geq 0}$ be M-Lipschitz, then $\frac{1}{M} \cdot q\colon X \to \mathbb{R}_{\geq 0}$ is short. The function $(-) \models q\colon \mathcal{D}(X) \to \mathbb{R}_{\geq 0}$ is then also M-Lipschitz, since for $\omega, \omega' \in \mathcal{D}(X)$,

$$\begin{aligned}\big|\, \omega \models q - \omega' \models q \,\big| \;&=\; M \cdot \big|\, \omega \models \tfrac{1}{M} \cdot q - \omega' \models \tfrac{1}{M} \cdot q \,\big| \\ &\leq\; M \cdot \bigvee_{p \in \mathit{Fact}_S(X)} \big|\, \omega \models p - \omega' \models p \,\big| \\ &=\; M \cdot d\big(\omega, \omega'\big).\end{aligned}$$

4. The only coupling of $1|x\rangle, \omega \in \mathcal{D}(X)$ is $1|x\rangle \otimes \omega \in \mathcal{D}(X \times X)$. Hence:

$$d\big(1|x\rangle, \omega\big) \;=\; 1|x\rangle \otimes \omega \models d_X \;=\; \sum_{x' \in X} \omega(x') \cdot d_X(x, x') \;=\; \omega \models d_X(x, -).$$

5. We first note that for a distribution of distributions $\Omega \in \mathcal{D}^2(X)$ and a short factor $p\colon X \to \mathbb{R}_{\geq 0}$ the validity in Ω of the short validity factor $(-) \models p\colon \mathcal{D}(X) \to \mathbb{R}_{\geq 0}$ from item 3 satisfies:

$$\begin{aligned}\Omega \models \big((-) \models p\big) \;&=\; \sum_{\omega \in \mathcal{D}(X)} \Omega(\omega) \cdot \big(\omega \models p\big) \\ &=\; \sum_{\omega \in \mathcal{D}(X)} \sum_{x \in X} \Omega(\omega) \cdot \omega(x) \cdot p(x) \\ &\overset{(4)}{=}\; \sum_{x \in X} \mathit{flat}(\Omega)(x) \cdot p(x) \\ &=\; \mathit{flat}(\Omega) \models p.\end{aligned}$$

Thus for $\Omega, \Omega' \in \mathcal{D}^2(X)$,

$$\begin{aligned}& d_X\Big(\mathit{flat}(\Omega), \mathit{flat}(\Omega')\Big) \\ &= \bigvee_{p \in \mathit{Fact}_S(X)} \big|\, \mathit{flat}(\Omega) \models p \,-\, \mathit{flat}(\Omega') \models p \,\big| \\ &= \bigvee_{p \in \mathit{Fact}_S(X)} \big|\, \Omega \models \big((-) \models p\big) \,-\, \Omega' \models \big((-) \models p\big) \big| && \text{as just shown} \\ &\leq \bigvee_{Q \in \mathit{Fact}_S(\mathcal{D}(X))} \big|\, \Omega \models Q \,-\, \Omega' \models Q \,\big| && \text{by item 3} \\ &= d_{\mathcal{D}(X)}\big(\Omega, \Omega'\big).\end{aligned}$$

6. Directly by points (1) and (5).
7. The channel composite $d \circledcirc c = \mathit{flat} \circ \mathcal{D}(d) \circ c$ consists of a functional composite of M-Lipschitz, K-Lipschitz, and 1-Lipschitz maps, and is thus $(M \cdot K \cdot 1)$-Lipschitz. This uses items 1 and (5).
8. If we have couplings τ_i for ω_i, ω_i', then $\sum_i r_i \cdot \tau_i$ is a coupling of $\sum_i r_i \cdot \omega_i$ and $\sum_i r_i \cdot \omega_i'$. Moreover:

$$d\Big(\textstyle\sum_i r_i \cdot \omega_i,\, \sum_i r_i \cdot \omega_i' \Big) \;\leq\; \Big(\textstyle\sum_i r_i \cdot \tau_i \Big) \models d_X \;=\; \textstyle\sum_i r_i \cdot \Big(\tau_i \models d_X \Big).$$

   Since this holds for all τ_i, we get: $d\big(\sum_i r_i \cdot \omega_i, \sum_i r_i \cdot \omega_i' \big) \leq \sum_i r_i \cdot d\big(\omega_i, \omega_i'\big)$.
9. We unfold the definition of the *prm* map from (6) and use the previous item in the first step below. We also use that the distance between two sequences is invariant under permutation (of both).

$$\begin{aligned} d_{\mathcal{D}(X^K)}\big(\mathit{prm}(\boldsymbol{x}), \mathit{prm}(\boldsymbol{y})\big) \;&\leq\; \sum_{t\colon K \overset{\cong}{\to} K} \frac{1}{K!} \cdot d_{\mathcal{D}(X^K)}\big(1\big|\underline{t}(\boldsymbol{x})\big\rangle, 1\big|\underline{t}(\boldsymbol{y})\big\rangle\big) \\ &=\; \sum_{t\colon K \overset{\cong}{\to} K} \frac{1}{K!} \cdot d_{X^K}\big(\underline{t}(\boldsymbol{x}), \underline{t}(\boldsymbol{y})\big) \qquad \text{by item 4} \\ &=\; \sum_{t\colon K \overset{\cong}{\to} K} \frac{1}{K!} \cdot d_{X^K}\big(\boldsymbol{x}, \boldsymbol{y}\big) \;=\; d_{X^K}\big(\boldsymbol{x}, \boldsymbol{y}\big). \end{aligned} \qquad ❑$$

Later on we need the following facts about tensors of distributions.

Proposition 1. *Let X, Y be metric spaces, and K be a positive natural number.*

1. *The tensor map $\otimes\colon \mathcal{D}(X) \times \mathcal{D}(Y) \to \mathcal{D}(X \times Y)$ is an isometry.*
2. *The K-fold tensor map $\mathit{iid}[K]\colon \mathcal{D}(X) \to \mathcal{D}(X^K)$, given by $\mathit{iid}[K](\omega) := \omega^K = \omega \otimes \cdots \otimes \omega$, is K-Lipschitz. Actually, there is an equality: $d(\omega^K, \rho^K) = K \cdot d(\omega, \rho)$.*

Proof. 1. Let distributions $\omega, \omega' \in \mathcal{D}(X)$ and $\rho, \rho' \in \mathcal{D}(Y)$ be given. For the inequality $d_{\mathcal{D}(X) \times \mathcal{D}(Y)}\big((\omega, \rho), (\omega', \rho')\big) \leq d_{\mathcal{D}(X \times Y)}\big(\omega \otimes \rho, \omega' \otimes \rho'\big)$ one uses that a coupling $\tau \in \mathcal{D}\big((X \times Y) \times (X \times Y)\big)$ of $\omega \otimes \rho, \omega' \otimes \rho' \in \mathcal{D}(X \times Y)$ can be turned into two couplings τ_1, τ_2 of ω, ω' and of ρ, ρ', namely as $\tau_i := \mathcal{D}\big(\pi_i \times \pi_i\big)(\tau)$. For the reverse inequality one turns two couplings τ_1, τ_2 of ω, ω' and ρ, ρ' into a coupling τ of $\omega \otimes \rho, \omega' \otimes \rho'$ via $\tau := \mathcal{D}\big(\langle \pi_1 \times \pi_1, \pi_2 \times \pi_2 \rangle\big)\big(\tau_1 \otimes \tau_2\big)$.

2. For $\omega, \rho \in \mathcal{D}(X)$ and $K \in \mathbb{N}$, using the previous item, we get:

$$d_{\mathcal{D}(X^K)}\big(\omega^K, \rho^K\big) \overset{1}{=} d_{\mathcal{D}(X)^K}\Big((\omega, \ldots, \omega), (\rho, \ldots, \rho)\Big) \overset{(7)}{=} K \cdot d_{\mathcal{D}(X)}\big(\omega, \rho\big). \quad ❑$$

5 The Wasserstein distance between multisets

There is also a Wasserstein distance between multisets of the same size. This section recalls the definition and the main results.

Definition 3. *Let (X, d_X) be a metric space and $K \in \mathbb{N}$ a natural number. We can turn the metric $d_X \colon X \times X \to \mathbb{R}_{\geq 0}$ into the* Wasserstein *metric $d \colon \mathcal{M}[K](X) \times \mathcal{M}[K](X) \to \mathbb{R}_{\geq 0}$ on multisets (of the same size), via:*

$$\begin{aligned} d\big(\varphi, \varphi'\big) &\coloneqq \bigwedge_{\tau \in dcpl^{-1}(\varphi,\varphi')} Flrn(\tau) \models d_X \\ &= \frac{1}{K} \cdot \bigwedge_{\boldsymbol{x} \in acc^{-1}(\varphi),\, \boldsymbol{x}' \in acc^{-1}(\varphi')} d_{X^K}\big(\boldsymbol{x}, \boldsymbol{x}'\big) \\ &\overset{(7)}{=} \frac{1}{K} \cdot \bigwedge_{\boldsymbol{x} \in acc^{-1}(\varphi),\, \boldsymbol{x}' \in acc^{-1}(\varphi')} \sum_{0 \leq i < K} d_X(x_i, x'_i). \end{aligned} \tag{9}$$

All meets in (9) are finite and can be computed via enumeration. Alternatively, one can use linear optimisation. We give an illustration below. The equality of the first two formulations is standard, like in Definition 2, and is used here without proof. There is an alternative formulation of the above distance between multisets that uses bistochastic matrices, see *e.g.* [2,6], but we do not need it here.

Example 2. Consider the following two multisets of size 4 on the set $X = \{1, 2, 3\} \subseteq \mathbb{N}$, with standard distance between natural numbers.

$$\varphi = 3|1\rangle + 1|2\rangle \qquad\qquad \varphi' = 2|1\rangle + 1|2\rangle + 1|3\rangle.$$

The optimal coupling $\tau \in \mathcal{M}[4](X \times X)$ is:

$$\tau = 2\big|1, 1\big\rangle + 1\big|1, 2\big\rangle + 1\big|2, 3\big\rangle.$$

The resulting Wasserstein distance $d(\varphi, \varphi')$ is:

$$Flrn(\tau) \models d_X = \tfrac{1}{2} \cdot d_X(1, 1) + \tfrac{1}{4} \cdot d_X(1, 2) + \tfrac{1}{4} \cdot d_X(2, 3) = \tfrac{1}{4} \cdot 1 + \tfrac{1}{4} \cdot 1 = \tfrac{1}{2}.$$

Alternatively, we may proceed as follows. There are $\big(\varphi\big) = \frac{4!}{3! \cdot 1!} = 4$ lists that accumulate to φ, and $\big(\varphi'\big) = \frac{4!}{2! \cdot 1! \cdot 1!} = 12$ lists that accumulate to φ'. We can align them all and compute the minimal distance. It is achieved for instance at:

$$\tfrac{1}{4} \cdot d_{X^4}\Big((1, 1, 1, 2), (1, 1, 2, 3)\Big) \overset{(7)}{=} \tfrac{1}{4} \cdot \big(0 + 0 + 1 + 1\big) = \tfrac{2}{4} = \tfrac{1}{2}.$$

Lemma 3. *We consider the situation in Definition 3.*

1. *Frequentist learning $Flrn \colon \mathcal{M}[K](X) \to \mathcal{D}(X)$ is an isometry, for $K > 0$.*
2. *For numbers $K, n \geq 1$ the scalar multiplication function $n \cdot (-) \colon \mathcal{M}[K](X) \to \mathcal{M}[n \cdot K](X)$ is an isometry.*
3. *The sum of distributions $+ \colon \mathcal{M}[K](X) \times \mathcal{M}[L](X) \to \mathcal{M}[K + L](X)$ is short.*
4. *If $f \colon X \to Y$ is M-Lipschitz, then $\mathcal{M}[K](f) \colon \mathcal{M}[K](X) \to \mathcal{M}[K](Y)$ is M-Lipschitz too. Thus, the fixed size multiset functor $\mathcal{M}[K]$ lifts to categories of metric spaces $\mathbf{Met}_S$ and $\mathbf{Met}_L$.*
5. *For $K > 0$ the accumulation map $acc \colon X^K \to \mathcal{M}[K](X)$ is $\frac{1}{K}$-Lipschitz, and thus short.*

6. *The arrangement channel* $\mathit{arr}\colon \mathcal{M}[K](X) \rightsquigarrow X^K$ *is* K*-Lipschitz; in fact there is an equality* $d\big(\mathit{arr}(\varphi), \mathit{arr}(\varphi')\big) = K \cdot d(\varphi, \varphi')$.

Proof. 1. Via naturality of frequentist learning: if $\tau \in \mathcal{M}[K](X \times X)$ is a coupling of $\varphi, \varphi' \in \mathcal{M}[K](X)$, then $\mathit{Flrn}(\tau) \in \mathcal{D}(X \times X)$ is a coupling of $\mathit{Flrn}(\varphi), \mathit{Flrn}(\varphi') \in \mathcal{D}(X)$. This gives $d\big(\varphi, \varphi'\big) \leq d\big(\mathit{Flrn}(\varphi), \mathit{Flrn}(\varphi')\big)$. The reverse inequality is a bit more subtle. Let $\sigma \in \mathcal{D}(X \times X)$ be an optimal coupling of $\mathit{Flrn}(\varphi), \mathit{Flrn}(\varphi')$. Then, since any coupling $\tau \in \mathcal{M}[K](X \times X)$ of φ, φ' gives, as we have just seen, a coupling $\mathit{Flrn}(\tau) \in \mathcal{D}(X \times X)$ of $\mathit{Flrn}(\varphi), \mathit{Flrn}(\varphi')$, we obtain, by optimality:

$$d\big(\mathit{Flrn}(\varphi), \mathit{Flrn}(\varphi')\big) \;=\; \sigma \models d_X \;\leq\; \mathit{Flrn}(\tau) \models d_X.$$

Since this holds for any coupling τ, we get $d\big(\mathit{Flrn}(\varphi), \mathit{Flrn}(\varphi')\big) \leq d\big(\varphi, \varphi'\big)$.

2. For multisets $\varphi, \varphi' \in \mathcal{M}[K](X)$, by the previous item:

$$\begin{aligned} d_{\mathcal{M}[K](X)}\big(\varphi, \varphi'\big) &= d_{\mathcal{D}(X)}\big(\mathit{Flrn}(\varphi), \mathit{Flrn}(\varphi')\big) \\ &= d_{\mathcal{D}(X)}\big(\mathit{Flrn}(n \cdot \varphi), \mathit{Flrn}(n \cdot \varphi')\big) \\ &= d_{\mathcal{M}[n \cdot K](X)}\big(n \cdot \varphi, n \cdot \varphi'\big). \end{aligned}$$

3. For multisets $\varphi, \varphi' \in \mathcal{M}[K](X)$ and $\psi, \psi' \in \mathcal{M}[L](X)$, using Lemma 2 (8),

$$\begin{aligned} & d\Big(\varphi + \psi, \varphi' + \psi'\Big) \\ & \overset{1}{=} d\Big(\mathit{Flrn}(\varphi + \psi), \mathit{Flrn}(\varphi' + \psi')\Big) \\ & = d\Big(\tfrac{K}{K+L} \cdot \mathit{Flrn}(\varphi) + \tfrac{L}{K+L} \cdot \mathit{Flrn}(\psi), \tfrac{K}{K+L} \cdot \mathit{Flrn}(\varphi') + \tfrac{L}{K+L} \cdot \mathit{Flrn}(\psi'), \Big) \\ & \leq \tfrac{K}{K+L} \cdot d\big(\mathit{Flrn}(\varphi), \mathit{Flrn}(\varphi')\big) + \tfrac{L}{K+L} \cdot d\big(\mathit{Flrn}(\psi), \mathit{Flrn}(\psi')\big) \\ & \overset{1}{=} \tfrac{K}{K+L} \cdot d\big(\varphi, \varphi'\big) + \tfrac{L}{K+L} \cdot d\big(\psi, \psi'\big) \\ & \leq d\big(\varphi, \varphi'\big) + d\big(\psi, \psi'\big) \\ & \overset{(7)}{=} d\Big((\varphi, \psi), (\varphi', \psi')\Big). \end{aligned}$$

4. Let $f\colon X \to Y$ be M-Lipschitz. We use that frequentist learning Flrn is an isometry and a natural transformation $\mathcal{M}[K] \Rightarrow \mathcal{D}$. For multisets $\varphi, \varphi' \in \mathcal{M}[K](X)$,

$$\begin{aligned} & d_{\mathcal{M}[K](Y)}\Big(\mathcal{M}(f)(\varphi), \mathcal{M}(f)(\varphi')\Big) \\ & \overset{1}{=} d_{\mathcal{D}(Y)}\Big(\mathit{Flrn}\big(\mathcal{M}(f)(\varphi)\big), \mathit{Flrn}\big(\mathcal{M}(f)(\varphi')\big)\Big) \\ & = d_{\mathcal{D}(Y)}\Big(\mathcal{D}(f)\big(\mathit{Flrn}(\varphi)\big), \mathcal{D}(f)\big(\mathit{Flrn}(\varphi')\big)\Big) && \text{by naturality of } \mathit{Flrn} \\ & \leq M \cdot d_{\mathcal{D}(X)}\big(\mathit{Flrn}(\varphi), \mathit{Flrn}(\varphi')\big) && \text{by Lemma 2 (1)} \\ & \overset{1}{=} d_{\mathcal{M}[K](X)}\big(\varphi, \varphi'\big). \end{aligned}$$

5. The map $\mathit{acc}\colon X^K \to \mathcal{M}[K](X)$ is $\frac{1}{K}$-Lipschitz since for $\boldsymbol{y}, \boldsymbol{y'} \in X^K$,

$$\begin{aligned} d\Big(\mathit{acc}(\boldsymbol{y}), \mathit{acc}(\boldsymbol{y'})\Big) &= \frac{1}{K} \cdot \bigwedge_{\boldsymbol{x} \in \mathit{acc}^{-1}(\mathit{acc}(\boldsymbol{y})),\, \boldsymbol{x'} \in \mathit{acc}^{-1}(\mathit{acc}(\boldsymbol{y'}))} d_{X^K}\big(\boldsymbol{x}, \boldsymbol{x'}\big) \\ &\leq \frac{1}{K} \cdot d_{X^K}\big(\boldsymbol{y}, \boldsymbol{y'}\big). \end{aligned}$$

6. For fixed $\varphi, \varphi' \in \mathcal{M}[K](X)$, take arbitrary $\boldsymbol{x} \in acc^{-1}(\varphi)$ and $\boldsymbol{x'} \in acc^{-1}(\varphi')$. Then:

$$\begin{aligned} d_{\mathcal{D}(X^K)}\big(arr(\varphi), arr(\varphi')\big) &= d_{\mathcal{D}(X^K)}\Big(arr(acc(\boldsymbol{x})), arr(acc(\boldsymbol{x'}))\Big) \\ &\overset{(6)}{=} d_{\mathcal{D}(X^K)}\Big(prm(\boldsymbol{x}), prm(\boldsymbol{x'})\Big) \\ &\leq d_{X^K}\big(\boldsymbol{x}, \boldsymbol{x'}\big) \qquad \text{by Lemma 2 (9).} \end{aligned}$$

Since this holds for all $\boldsymbol{x} \in acc^{-1}(\varphi)$, $\boldsymbol{x'} \in acc^{-1}(\varphi')$ we get an inequaltiy $d_{\mathcal{D}(X^K)}\big(arr(\varphi), arr(\varphi')\big) \leq K \cdot d_{\mathcal{M}[K](X)}\big(\varphi, \varphi'\big)$, see Definition 3. This inequality is an actual equality since acc, and thus $\mathcal{D}(acc)$, is $\frac{1}{K}$-Lipschitz:

$$\begin{aligned} d_{\mathcal{M}[K](X)}\big(\varphi, \varphi'\big) &= d_{\mathcal{D}(\mathcal{M}[K](X))}\big(1|\varphi\rangle, 1|\varphi'\rangle\big) \\ &= d_{\mathcal{D}(\mathcal{M}[K](X))}\Big(\mathcal{D}(acc)\big(arr(\varphi)\big), \mathcal{D}(acc)\big(arr(\varphi')\big)\Big) \\ &\leq \tfrac{1}{K} \cdot d_{\mathcal{D}(X^K)}\Big(arr(\varphi), arr(\varphi')\Big) \end{aligned}$$

❑

6 Multinomial drawing is isometric

Multinomial draws are of the draw-and-replace kind. This means that a drawn ball is returned to the urn, so that the urn remains unchanged. Thus we may use a distribution $\omega \in \mathcal{D}(X)$ as urn. For a draw size number $K \in \mathbb{N}$, the multinomial distribution $mn[K](\omega) \in \mathcal{D}\big(\mathcal{M}[K](X)\big)$ on multisets / draws of size K can be defined via accumulated sequences of draws:

$$\begin{aligned} mn[K](\omega) &:= \mathcal{D}(acc)\big(\omega^K\big) \\ &= \mathcal{D}(acc)\big(iid[K](\omega)\big) \\ &= \sum_{\varphi \in \mathcal{M}[K](X)} (\!(\varphi)\!) \cdot \prod_{x \in X} \omega(x)^{\varphi(x)} \, \big|\varphi\big\rangle. \end{aligned} \tag{10}$$

We recall that $(\!(\varphi)\!) = \frac{K!}{\prod_x \varphi(x)!}$ is the number of sequences that accumulate to a multiset / draw $\varphi \in \mathcal{M}[K](X)$. A basic result from [8, Prop. 3] is that applying frequentist learning to the draws yields the original urn:

$$Flrn \gg\!= mn[K](\omega) = \omega. \tag{11}$$

We can now formulate and prove our first isometry result.

Theorem 1. *Let X be an arbitrary metric space (of colours), and $K > 0$ be a positive natural (draw size) number. The multinomial channel*

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

is an isometry. This involves the Wasserstein metric (8) *for distributions over X on the domain $\mathcal{D}(X)$, and the Wasserstein metric for distributions over multisets of size K, with their Wasserstein metric* (9)*, on the codomain $\mathcal{D}\big(\mathcal{M}[K](X)\big)$.*

Proof. Let distributions $\omega, \omega' \in \mathcal{D}(X)$ be given. The map $mn[K]$ is short since:

$$\begin{array}{ll} d_{\mathcal{D}(\mathcal{M}[K](X))}\Big(mn[K](\omega),\, mn[K](\omega')\Big) & \\ \overset{(10)}{=}\; d_{\mathcal{D}(\mathcal{M}[K](X))}\Big(\mathcal{D}(acc)(iid[K](\omega)),\, \mathcal{D}(acc)(iid[K](\omega'))\Big) & \\ \leq\; \frac{1}{K}\cdot d_{\mathcal{D}(X^K)}\Big(iid[K](\omega),\, iid[K](\omega')\Big) & \text{by Lemma 3 (5)} \\ =\; \frac{1}{K}\cdot K \cdot d_{\mathcal{D}(X)}\big(\omega,\, \omega'\big) & \text{by Proposition 1 (2)} \\ =\; d_{\mathcal{D}(X)}\big(\omega,\, \omega'\big). & \end{array}$$

There is also an inequality in the other direction, via:

$$\begin{aligned} d_{\mathcal{D}(X)}\big(\omega,\, \omega'\big) &\overset{(11)}{=} d_{\mathcal{D}(X)}\Big(Flrn \gg\!= mn[K](\omega),\, Flrn \gg\!= mn[K](\omega')\Big) \\ &\leq\; d_{\mathcal{D}(\mathcal{M}[K](X))}\Big(mn[K](\omega),\, mn[K](\omega')\Big). \end{aligned}$$

The latter inequality follows from the fact that frequentist learning *Flrn* is short, see Lemma 3 (1), and that Kleisli extension $Flrn \gg\!= (-)$ is thus short too, see Lemma 2 (6). ❏

Example 3. Consider the following two distributions $\omega, \omega' \in \mathcal{D}(\mathbb{N})$.

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

This distance $d(\omega, \omega')$ involves the standard distance on $\mathbb{N}$, using the optimal coupling $\frac{1}{3}|0,1\rangle + \frac{1}{6}|2,1\rangle + \frac{1}{2}|2,2\rangle \in \mathcal{D}\big(\mathbb{N}\times\mathbb{N}\big)$.

We take draws of size $K = 3$. There are 10 multisets of size 3 over $\{0, 1, 2\}$:

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

These multisets occur in the following multinomial distributions of draws of size 3.

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

The optimal coupling $\tau \in \mathcal{D}\big(\mathcal{M}[3](\mathbb{N}) \times \mathcal{M}[3](\mathbb{N})\big)$ between these two multinomial distributions is:

$$\begin{aligned} \tau \;=\; &\tfrac{1}{27}\Big|\varphi_1, \varphi_4\Big\rangle + \tfrac{19}{216}\Big|\varphi_5, \varphi_4\Big\rangle + \tfrac{1}{8}\Big|\varphi_{10}, \varphi_{10}\Big\rangle + \tfrac{29}{216}\Big|\varphi_5, \varphi_7\Big\rangle \\ &+ \tfrac{5}{72}\Big|\varphi_8, \varphi_7\Big\rangle + \tfrac{3}{8}\Big|\varphi_8, \varphi_9\Big\rangle + \tfrac{37}{216}\Big|\varphi_{10}, \varphi_7\Big\rangle. \end{aligned}$$

We compute the distance between the multinomial distributions, using $d_{\mathcal{M}} = d_{\mathcal{M}[3](\mathbb{N})}$.

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

As predicted in Theorem 1, this distance coincides with the distance $d(\omega, \omega') = \frac{1}{2}$ between the original urn distributions. One sees that the computation of the distance between the draw distributions is more complex, involving ‘Wasserstein over Wasserstein’.

7 Hypergeometric drawing is isometric

We start with some preparatory observations on probabilistic projection and drawing.

Lemma 4. *For a metric space X and a number K, consider the probabilistic projection-delete PD and probabilistic draw-delete DD channels.*

$$X^{K+1} \xrightarrow{PD} \mathcal{D}\big(X^K\big) \qquad\qquad \mathcal{M}[K+1](X) \xrightarrow{DD} \mathcal{D}\big(\mathcal{M}[K](X)\big)$$

They are defined via deletion of elements from sequences and from multisets:

$$\begin{aligned} PD(x_1, \ldots, x_{K+1}) &\coloneqq \sum_{1 \le i \le K+1} \frac{1}{K+1} \big| x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{K+1} \big\rangle \\ DD(\psi) &\coloneqq \sum_{x \in supp(\psi)} \frac{\psi(x)}{K+1} \big| \psi - 1|x\rangle \big\rangle \\ &= \sum_{x \in supp(\psi)} Flrn(\psi)(x) \big| \psi - 1|x\rangle \big\rangle. \end{aligned}$$

Then:

1. *$\langle acc \rangle \circledcirc PD = DD \circledcirc \langle acc \rangle$;*
2. *$Flrn \ggg DD(\psi) = Flrn(\psi)$;*
3. *PD is $\frac{K}{K+1}$-Lipschitz, and thus short;*
4. *DD is an isometry.*

Proof. The first point is easy and the second one is [8, Lem. 5 (ii)].

3. For $\boldsymbol{x}, \boldsymbol{y} \in X^{K+1}$, via Lemma 2 (8) and (4),

$$\begin{aligned} d\Big(PD(\boldsymbol{x}), PD(\boldsymbol{y})\Big) &= d\left(\sum_{1 \le i \le K+1} \frac{1}{K+1} \big| x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{K+1} \big\rangle, \right. \\ &\qquad \left. \sum_{1 \le i \le K+1} \frac{1}{K+1} \big| y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_{K+1} \big\rangle \right) \\ &\le \sum_{1 \le i \le K+1} \frac{1}{K+1} \cdot d\Big(1 \big| x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{K+1} \big\rangle, \\ &\qquad 1 \big| y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_{K+1} \big\rangle \Big) \\ &= \sum_{1 \le i \le K+1} \frac{1}{K+1} \cdot d_{X^K}\Big((x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{K+1}), \\ &\qquad (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_{K+1}) \Big) \\ &= \sum_{1 \le i \le K+1} \frac{1}{K+1} \cdot K \cdot d_X(x_i, y_i) \\ &\overset{(7)}{=} \frac{K}{K+1} \cdot d_{X^{K+1}}\big(\boldsymbol{x}, \boldsymbol{y}\big). \end{aligned}$$

4. Via item 1 we get:

$$\langle acc\rangle \circledcirc PD \circledcirc arr \;=\; DD \circledcirc \langle acc\rangle \circledcirc arr \;=\; DD \circledcirc unit \;=\; DD. \qquad (*)$$

Now we can show that DD is short: for $\psi, \psi' \in \mathcal{M}[K+1](X)$

$$\begin{array}{l} d_{\mathcal{D}(\mathcal{M}[K](X))}\big(DD(\psi),\, DD(\psi')\big) \\ \overset{(*)}{=}\; d_{\mathcal{D}(\mathcal{M}[K](X))}\Big(\mathcal{D}(acc)\big(PD \gg\!= arr(\psi)\big),\, \mathcal{D}(acc)\big(PD \gg\!= arr(\psi')\big)\Big) \\ \leq\; \frac{1}{K} \cdot d_{\mathcal{D}(X^K)}\Big(PD \gg\!= arr(\psi),\, PD \gg\!= arr(\psi')\Big) \\ \leq\; \frac{1}{K} \cdot \frac{K}{K+1} \cdot d_{\mathcal{D}(X^{K+1})}\big(arr(\psi),\, arr(\psi')\big) \\ =\; \frac{1}{K+1} \cdot (K+1) \cdot d_{\mathcal{M}[K+1](X))}\big(\psi,\, \psi'\big) \\ =\; d_{\mathcal{M}[K+1](X))}\big(\psi,\, \psi'\big). \end{array}$$

For the reverse inequality we use item 2 and the fact that $Flrn$ is a short:

$$\begin{array}{l} d_{\mathcal{D}(\mathcal{M}[K](X))}\Big(DD(\psi), DD(\psi')\Big) \\ \geq\; d_{\mathcal{D}(\mathcal{M}[K](X))}\Big(Flrn \gg\!= DD(\psi), Flrn \gg\!= DD(\psi')\Big) \\ =\; d_{\mathcal{D}(X)}\big(Flrn(\psi), Flrn(\psi')\big) \\ =\; d_{\mathcal{M}[K+1](X)}\big(\psi, \psi'\big). \end{array} \qquad ❑$$

The hypergeometric channel $hg[K]\colon \mathcal{M}[L](X) \to \mathcal{D}\big(\mathcal{M}[K](X)\big)$, for urn size $L \geq K$, where K is the draw size, is an iteration of draw-delete's, see [8, Thm. 6]:

$$hg[K](\upsilon) \;:=\; \underbrace{DD \circledcirc \cdots \circledcirc DD}_{L-K \text{ times}} \;=\; \sum_{\varphi \in \mathcal{M}[K](X),\, \varphi \leq \upsilon} \frac{\binom{\upsilon}{\varphi}}{\binom{L}{K}} \,\big|\varphi\big\rangle, \qquad (12)$$

where $\binom{\upsilon}{\varphi} := \prod_{x\in X} \binom{\upsilon(x)}{\varphi(x)}$.

Theorem 2. *The hypergeometric channel* $hg[K]\colon \mathcal{M}[L](X) \to \mathcal{D}\big(\mathcal{M}[K](X)\big)$ *defined in* (12)*, for* $L \geq K$*, is an isometry.*

Proof. We see in (12) that $hg[K]$ is a (channel) iteration of isometries DD, and thus of short maps; hence it it short itself. Via iterated use of Lemma 4 (2) we get $Flrn \gg\!= hg[K](\psi) = Flrn(\psi)$. This gives the inequality in the other direction, like in the proof of Lemma 4 (2):

$$\begin{array}{rl} d_{\mathcal{M}[K+1](X)}\big(\psi, \psi'\big) & =\; d_{\mathcal{D}(X)}\big(Flrn(\psi), Flrn(\psi')\big) \\ & =\; d_{\mathcal{D}(\mathcal{M}[K](X))}\Big(Flrn \gg\!= hg[K](\psi), Flrn \gg\!= hg[K](\psi')\Big) \\ & \leq\; d_{\mathcal{D}(\mathcal{M}[K](X))}\big(hg[K](\psi), hg[K](\psi')\big). \end{array} \qquad ❑$$

The very beginning of this paper contains an illustration of this result, for urns over the set of colours $C = \{R, G, B\}$, considered as a discrete metric space.

8 Pólya drawing is isometric

Hypergeometric distributions use the draw-delete mode: a drawn ball is removed from the urn. The less well-known Pólya draws [7] use the draw-add mode. This means that a drawn ball is returned to the urn, together with another ball of the same colour (as the drawn ball). Thus, with hypergeometric draws the urn decreases in size, so that only finitely many draws are possible, whereas with Pólya draws the urn grows in size, and the drawing may be repeated arbitrarily many times. As a result, for Pólya distributions we do not need to impose restrictions on the size K of draws. We do have to restrict draws from urn υ to multisets $\varphi \in \mathcal{M}[K](X)$ with $\mathit{supp}(\varphi) \subseteq \mathit{supp}(\upsilon)$ since we can only draw balls of colours that are in the urn. Pólya distributions are formulated in terms of multi-choose binomials $\left(\!\binom{n}{m}\!\right) := \binom{n+m-1}{m} = \frac{(n+m-1)!}{m!\cdot(n-1)!}$, for $n > 0$. This multi-choose number $\left(\!\binom{n}{m}\!\right)$ is the number of multisets of size m over a set with n elements, see [9,10] for details.

$$\mathit{pl}[K](\upsilon) := \sum_{\varphi\in\mathcal{M}[K](X),\, \mathit{supp}(\varphi)\subseteq \mathit{supp}(\upsilon)} \frac{\left(\!\binom{\upsilon}{\varphi}\!\right)}{\left(\!\binom{L}{K}\!\right)} \big|\varphi\big\rangle, \tag{13}$$

where $\left(\!\binom{\upsilon}{\varphi}\!\right) := \prod_{x\in \mathit{supp}(\upsilon)} \left(\!\binom{\upsilon(x)}{\varphi(x)}\!\right)$.

Theorem 3. *Each Pólya channel* $\mathit{pl}[K]\colon \mathcal{M}[L](X) \to \mathcal{D}\big(\mathcal{M}[K](X)\big)$*, for urn and draw sizes* $L > 0, K > 0$*, is an isometry.*

Proof. One inequality follows by exploiting the equation $\mathit{Flrn} \gg\!= \mathit{pl}[K](\psi) = \mathit{Flrn}(\psi)$ like in previous sections. The reverse inequality, for shortness, involves a draw-store-add channel of the form:

$$\mathcal{M}[L](X) \times \mathcal{M}[N](X) \xrightarrow{\quad \mathit{DSA} \quad} \mathcal{D}\big(\mathcal{M}[L](X) \times \mathcal{M}[N+1](X)\big)$$

defined as:

$$\begin{aligned} \mathit{DSA}(\upsilon,\varphi) &:= \sum_{x\in \mathit{supp}(\upsilon+\varphi)} \mathit{Flrn}(\upsilon+\varphi)(x)\,\Big|\upsilon, \varphi + 1|x\rangle\Big\rangle \\ &= 1\big|\upsilon\big\rangle \otimes \left(\sum_{x\in \mathit{supp}(\upsilon+\varphi)} \mathit{Flrn}(\upsilon+\varphi)(x)\big|\varphi+1|x\rangle\big\rangle\right). \end{aligned}$$

With some effort one shows that this channel DSA is short and that the Pólya channel can be expressed via iterated draw-store-add's, namely as:

$$\mathit{pl}[K](\upsilon) = \mathcal{D}(\pi_2)\Big(\big(\underbrace{\mathit{DSA} \circ \cdots \circ \mathit{DSA}}_{K \text{ times}}\big)(\upsilon, \mathbf{0})\Big),$$

where $\mathbf{0} \in \mathcal{M}[0](X)$ is the empty multiset. This makes the Pólya channel $\mathit{pl}[K]$ short, and thus an isometry. ❑

We illustrate that the Pólya channel is an isometry.

Example 4. We take as space of colours $X = \{0, 10, 50\} \subseteq \mathbb{N}$ with two urns:

$$\upsilon_1 = 3|0\rangle + 1|10\rangle \qquad \upsilon_2 = 1|0\rangle + 2|10\rangle + 1|50\rangle.$$

The distance between these urns is 15, via the optimal coupling $1|0,0\rangle + 2|0,10\rangle + 1|10,50\rangle$, yielding $\frac{1}{4}\cdot(0-0) + \frac{1}{2}\cdot(10-0) + \frac{1}{4}\cdot(50-10) = 5 + 10 = 15$.

We look at Pólya draws of size $K = 2$. This gives distributions:

$$\begin{aligned}
\mathit{pl}[2](\upsilon_1) &= \tfrac{3}{5}\Big|2|0\rangle\Big\rangle + \tfrac{3}{10}\Big|1|0\rangle + 1|10\rangle\Big\rangle + \tfrac{1}{10}\Big|2|10\rangle\Big\rangle \\
\mathit{pl}[2](\upsilon_2) &= \tfrac{1}{10}\Big|2|0\rangle\Big\rangle + \tfrac{1}{5}\Big|1|0\rangle + 1|10\rangle\Big\rangle + \tfrac{3}{10}\Big|2|10\rangle\Big\rangle + \tfrac{1}{10}\Big|1|0\rangle + 1|50\rangle\Big\rangle \\
&\qquad + \tfrac{1}{5}\Big|1|10\rangle + 1|50\rangle\Big\rangle + \tfrac{1}{10}\Big|2|50\rangle\Big\rangle
\end{aligned}$$

We compute the distance between these two distributions via the last formulation in (8), using the optimal short factor $p \colon \mathcal{M}[2](X) \to \mathbb{R}_{\geq 0}$ given by:

$$\begin{array}{lll}
p\big(2|0\rangle\big) = 0 & p\big(1|0\rangle + 1|10\rangle\big) = 5 & p\big(2|10\rangle\big) = 10 \\
p\big(1|0\rangle + 1|50\rangle\big) = 25 & p\big(1|10\rangle + 1|50\rangle\big) = 30 & p\big(2|50\rangle\big) = 50.
\end{array}$$

Then:

$$\begin{aligned}
\mathit{pl}[2](\upsilon_1) \models p &= \tfrac{3}{5}\cdot 0 + \tfrac{3}{10}\cdot 5 + \tfrac{1}{10}\cdot 10 = \tfrac{5}{2} \\
\mathit{pl}[2](\upsilon_2) \models p &= \tfrac{1}{10}\cdot 0 + \tfrac{1}{5}\cdot 5 + \tfrac{3}{10}\cdot 10 + \tfrac{1}{10}\cdot 25 + \tfrac{1}{5}\cdot 30 + \tfrac{1}{10}\cdot 50 = \tfrac{35}{2}.
\end{aligned}$$

As predicted by Theorem 3, the distance between the Pólya distributions then coincides with the distance between the urns:

$$\begin{aligned}
d\Big(\mathit{pl}[2](\upsilon_1), \mathit{pl}[2](\upsilon_2)\Big) &= \Big|\mathit{pl}[2](\upsilon_1) \models p - \mathit{pl}[2](\upsilon_2) \models p\Big| \\
&= \tfrac{35}{2} - \tfrac{5}{2} = 15 = d\big(\upsilon_1, \upsilon_2\big).
\end{aligned}$$

9 Conclusions

Category theory provides a fresh look at the area of probability theory, see *e.g.* [5] or [10] for an overview. Its perspective allows one to formulate and prove new results. This paper demonstrates that draw operations, viewed as (Kleisli) maps, are incredibly well-behaved: they preserve Wasserstein distances. Such distances on urns filled with coloured balls are relatively simple, starting from a 'ground' metric on the set of colours. But on draw distributions, the distances involve Wasserstein-over-Wasserstein. This paper concentrates on drawing from an urn. A natural question is whether other probabilistic operations, as Kleisli maps, preserve distance. This is a topic for further investigation.

Acknowledgments

Thanks are due to the anonymous reviewers for their detailed comments that improved an earlier version of this work.

References

1. F. van Breugel. An introduction to metric semantics: operational and denotational models for programming and specification languages. *Theor. Comp. Sci.*, 258(1-2):1–98, 2001. doi:10.1016/S0304-3975(00)00403-5.
2. H. Brezis. Remarks on the Monge-Kantorovich problem in the discrete setting. *Comptes Rendus Mathematique*, 356(2):207–213, 2018. doi:10.1016/j.crma.2017.12.008.
3. Y. Deng and W. Du. The Kantorovich metric in computer science: A brief survey. In C. Baier and A. di Pierro, editors, *Quantitative Aspects of Programming Languages*, number 253(3) in Elect. Notes in Theor. Comp. Sci., pages 73–82. Elsevier, Amsterdam, 2009. doi:10.1016/j.entcs.2009.10.006.
4. J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden. Metrics for labelled Markov processes. *Theor. Comp. Sci.*, 318:232–354, 2004.
5. T. Fritz. A synthetic approach to Markov kernels, conditional independence, and theorems on sufficient statistics. *Advances in Math.*, 370:107239, 2020. doi:10.1016/J.AIM.2020.107239.
6. T. Fritz and P. Perrone. A probability monad as the colimit of spaces of finite samples. *Theory and Appl. of Categories*, 34(7):170–220, 2019. doi:10.48550/arXiv.1712.05363.
7. F. Hoppe. Pólya-like urns and the Ewens' sampling formula. *Journ. Math. Biology*, 20:91–94, 1984. doi:10.1007/BF00275863.
8. B. Jacobs. From multisets over distributions to distributions over multisets. In *Logic in Computer Science*. IEEE, Computer Science Press, 2021. doi:10.1109/lics52264.2021.9470678.
9. B. Jacobs. Urns & tubes. *Compositionality*, 4(4), 2022. doi:10.32408/compositionality-4-4.
10. B. Jacobs. Structured probabilitistic reasoning. Book, in preparation, see http://www.cs.ru.nl/B.Jacobs/PAPERS/ProbabilisticReasoning.pdf, 2023.
11. B. Jacobs and A. Westerbaan. Distances between states and between predicates. *Logical Methods in Comp. Sci.*, 16(1), 2020. See https://lmcs.episciences.org/6154.
12. L. Kantorovich and G. Rubinshtein. On a space of totally additive functions. *Vestnik Leningrad Univ.*, 13:52–59, 1958.
13. J. Matouěk and B. Gärtner. *Understanding and Using Linear Programming*. Springer Verlag, Berlin, 2006. doi:10.1007/978-3-540-30717-4.
14. Y. Rubner, C. Tomasi, and L. Guibas. The Earth Mover's Distance as a metric for image retrieval. *Int. Journ. of Computer Vision*, 40:99–121, 2000. doi:10.1023/A:1026543900054.
15. C. Villani. *Optimal Transport — Old and New*. Springer, Berlin Heidelberg, 2009. doi:10.1007/978-3-540-71050-9.
16. G. Wyszecki and W. Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae*. Wiley, 1982.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.