From Multisets over Distributions to Distributions over Multisets

Bart Jacobs
Institute for Computing and Information Sciences, Radboud University Nijmegen, The Netherlands.
Web address: www.cs.ru.nl/B.Jacobs
Email: bart@cs.ru.nl

***Abstract*—A well-known challenge in the semantics of programming languages is how to combine non-determinism and probability. At a technical level, the problem arises from the fact that there is a no distributive law between the powerset monad and the distribution monad — as noticed some twenty years ago by Plotkin. More recently, it has become clear that there is a distributive law of the multiset monad over the distribution monad. This article elaborates the details of this distributivity and shows that there is a rich underlying theory relating multisets and probability distributions. It is shown that the new distributive law, called *parallel multinomial law*, can be defined in (at least) four equivalent ways. It involves putting multinomial distributions in parallel and commutes with hypergeometric distributions. Further, it is shown that this distributive law commutes with a new form of zipping for multisets. Abstractly, this can be described in terms of monoidal structure for a fixed-size multiset functor, when lifted to the Kleisli category of the distribution monad. Concretely, an application of the theory to sampling semantics is included.**

I. Introduction

Monads are used in the semantics of programming languages to classify different notions of computation, such non-deterministic, probabilistic, effectful, with exceptions, etc., see [1]. An obvious question is if such notions of computation can be combined via composition of monads. In general, monads do not compose, but they do compose in presence of a distributive law. This involves classic work in category theory, due to Beck [2]. This topic continues to receive attention today, especially because finding such distributive laws is a subtle and error-prone matter, see for instance [3] where it is shown that, somewhat surprisingly, the powerset monad does not distribute over itself. See also [4], [5] for wider investigations.

Sometimes it is also interesting whether a functor distributes over a monad, for instance in operational semantics, see the pioneering work of [6], which led to much follow-up work, especially in the theory of coalgebras [7]. There, distributive laws are also used for enhanced coinduction principles [8], [9].

Here we concentrate on the combination of non-deterministic computation. Some twenty years ago Gordon Plotkin noticed that the powerset monad $\mathcal{P}$ does not distribute over the probability distributions monad $\mathcal{D}$. He never published this important no-go result himself. Instead, it appeared in [10], [11], with full credits, where variations have been investigated (see also [12]). The lack of a distributive law between $\mathcal{P}$ and $\mathcal{D}$ means that there is no semantically clean way to combine non-deterministic and probabilistic computation.

In contrast, multisets do distribute over probability distributions. This seems to have been folklore knowledge for some time. It is therefore hard to give proper credit to this observation. In [13, p.82] it is described how the set of distributions $\mathcal{D}(M)$ on a commutative monoid M, can itself be turned into a commutative monoid. This is a somewhat isolated observation in [13], but by pushing this approach through one finds that it means that the distribution monad $\mathcal{D}$ on the category of sets lifts to the category of commutative monoids. Since the latter category is the category of Eilenberg-Moore algebras of the multiset monad $\mathcal{M}$, some abstract categorical result tells that this is equivalent to the existence of a distributive law of monads $\mathcal{M}\mathcal{D} \Rightarrow \mathcal{D}\mathcal{M}$. Hence one can say that this law is implicit in [13], but without an explicit description. The existence of this law is also discussed in [14], but also without an explicit definition. A further source is [15], where the existence of this law is taken for granted, since the resulting composite $\mathcal{D}\mathcal{M}$ is used as a monad. We simplify things, since [15] uses continuous instead of discrete distributions, but that is not essential. Also there, no explicit description of the distributive law is given.

This absence of an explicit description of the distributive law $\mathcal{M}\mathcal{D} \Rightarrow \mathcal{D}\mathcal{M}$ is understandable, since it is quite complex. The main contribution of this paper is that it describes the distributive law in full detail — actually in four different ways — and that it develops a rich theory surrounding this law.

The distributive law $\mathcal{M}\mathcal{D} \Rightarrow \mathcal{D}\mathcal{M}$ turns a multiset over distributions into a distribution over multisets. At the heart of this theory that we develop is the interaction of multisets and distributions. Recall that a multiset is like a set, except that elements may occur multiple times. A prime example of a multiset is an urn, containing multiple coloured balls. If the urn contains three red, five blue and two green balls, then we shall write it as a multiset $3|R\rangle + 5|B\rangle + 2|G\rangle$ over the set $\{R, B, G\}$ of colours. This multiset clearly has size 10. A basic property of the distributive law is that it preserves size: it turns a K-sized multiset of distributions into a distribution over multisets of size $K \in \mathbb{N}$.

When we draw a handful of balls from an urn, the draw itself may also be represented as a multiset. The classical multino-

In: Logic in Computer Science (LICS), 2021

mial and hypergeometric distributions assign probabilities to such draws — with and without replacement, respectively. We shall describe them as distributions on K-sized multisets. In fact, we shall describe the distributive law of this paper as a parallel composition of multinomial distributions. Therefore we propose to name this law the *parallel multinomial law*, written as *pml*.

We shall systematically describe multinomial and hypergeometric distributions via "channels", that is, via Kleisli maps in the Kleisli category $\mathcal{K}\ell(\mathcal{D})$ of the distribution monad $\mathcal{D}$. It is within this category that we formulate many of the key properties of multinomial and hypergeometric channels, using both sequential and parallel composition.

For the parallel behaviour of these channels and of the distributive law we introduce a new "zip" operation on multisets, called multizip. It turns out to be mathematically well-behaved, in its interaction with frequentist learning, with multinomial and hypergeometric channels, and also with the new parallel multinomial law. Technically, multizip makes the fixed-sized multiset functor, lifted to $\mathcal{K}\ell(\mathcal{D})$ via the distributive law *pml*, a monoidal functor. This multizip is a separate contribution of the paper.

The multinomial and hypergeometric distributions are classics in probability theory, but they are traditionally not studied as channels (Kleisli maps), and so their behaviour under sequential and parallel composition is typically ignored in the literature. This is a pity, since this behaviour involves beautiful structure, as will be demonstrated in this paper.

The paper is organised as follows. It first describes the relevant basics of multisets and probability distributions in Sections II and III, including the 'frequentist learning' map between them. Section IV introduces the operations of accumulation and arrangement for going back and forth between sequences and multisets. These two operations play a fundamental role, later on in the paper, in the description of multinomial channels, the parallel multinomial role, and also of multizip. Subsequent Sections V and VI describe multinomial and hypergeometric distributions as channels (Kleisli maps), together with a number of basic results, expressed via commuting diagrams, in the Kleisli category of the distribution monad $\mathcal{D}$. This is much like in [16], [17]. The real work starts in Section VII where the parallel multinomial law *pml* is introduced in four different, but equivalent, ways. Basic (sequential, via Kleisli composition) properties of this law are given in Section VIII, including commutation with multinomial and geometric channels, and with frequentist learning.

At this point the paper turns to putting things in parallel, again, in the Kleisli category of the distribution monad $\mathcal{D}$. For this purpose, Section IX introduces a new probabilistic "zip" for multisets. It combines two multisets of size K, on sets X, Y respectively, into a distribution on K-sized multisets on $X \times Y$. This multiset-zip (or simply: mulitzip, written as *mzip*) is thus a channel, and as such interacts well, not only with multinomial and geometric channels, but also with the parallel multinomial law *pml*, see Section X.

In the end, the newly developed theory of multisets and distributions is applied to sampling semantics. Section XI illustrates the correctness of the standard sampling approach for parallel and sequential composition (with a distribution obtained via Kleisli extension, *i.e.* state transformation) and for updating (conditioning of states). These correctness results rely on the commutation of frequentist learning and multizip with (parallel) multinomials and on the commutation of multinomials with updating.

II. Multisets

A multiset is a finite 'set' in which elements may occur multiple times. We shall write $3|a\rangle + 2|b\rangle$ for a multiset in which the element a occurs three times and the element b two times. The ket notation $|-\rangle$ is meaningless syntax that is used to keep elements (here: a, b) and their multiplicities (the numbers: $3, 2$) apart. We shall write $\mathcal{M}(X)$ for the set of finite multisets $n_1|x_1\rangle + \cdots + n_\ell|x_\ell\rangle$ over the set X, where $x_i \in X$ and $n_i \in \mathbb{N}$. Such a formal sum can also be written in functional form, as function $\varphi\colon X \to \mathbb{N}$, whose support $supp(\varphi) = \{x \in X \mid \varphi(x) \neq 0\}$ is finite. The *size* of a multiset is its total number of elements, written as $\|\varphi\| = \sum_x \varphi(x)$. Thus, $\big\| 3|a\rangle + 2|b\rangle \big\| = 5$.

For multisets $\varphi \in \mathcal{M}(X)$ and $\psi \in \mathcal{M}(Y)$ we can form their parallel product $\varphi \otimes \psi \in \mathcal{M}(X \times Y)$ via the product of multiplicities:

$$\big(\varphi \otimes \psi\big)(x, y) \coloneqq \varphi(x) \cdot \psi(y).$$

For instance, $\big(3|a\rangle + 2|b\rangle + 1|c\rangle\big) \otimes \big(2|0\rangle + 4|1\rangle\big) = 6|a,0\rangle + 12|a,1\rangle + 4|b,0\rangle + 8|b,1\rangle + 2|c,0\rangle + 4|c,1\rangle$.

Taking multisets is functorial: for a function $f\colon X \to Y$ one defines $\mathcal{M}(f)\colon \mathcal{M}(X) \to \mathcal{M}(Y)$ as $\mathcal{M}(f)(\sum_i n_i|x_i\rangle) \coloneqq \sum_i n_i|f(x_i)\rangle$. It is easy to see that $\|\mathcal{M}(f)(\varphi)\| = \|\varphi\|$.

The functor $\mathcal{M}\colon \mathbf{Sets} \to \mathbf{Sets}$ is actually a monad, with unit $\eta\colon X \to \mathcal{M}(X)$ given by $\eta(x) = 1|x\rangle$. The multiplication $\mu\colon \mathcal{M}^2(X) \to \mathcal{M}(X)$ is $\mu(\sum_i n_i|\varphi_i\rangle)(x) = \sum_i n_i \cdot \varphi_i(x)$. The Eilenberg-Moore algebras of this monad are precisely the commutative monoids: for a commutative monoid $(M, 0, +)$ the corresponding algebra $\mathcal{M}(M) \to M$ sends formal sums to actual sums: $\sum_i n_i|m_i\rangle \mapsto \sum_i n_i \cdot m_i$, for $m_i \in M$.

We often use multisets of a fixed size $K \in \mathbb{N}$, and so we write $\mathcal{M}[K](X) \subseteq \mathcal{M}(X)$ for the subset of $\varphi \in \mathcal{M}(X)$ with $\|\varphi\| = K$. This $\mathcal{M}[K]$ is also a functor, but not a monad. As an aside, when the set X has N elements, then $\mathcal{M}[K](X)$ has $\left(\!\binom{N}{K}\!\right) \coloneqq \binom{N+K-1}{K}$ elements, where $\left(\!\binom{N}{K}\!\right)$ is called the multiset number (also known as multichoose).

III. Probability distributions

In this paper, 'distribution' means 'finite discrete probability distribution'. Such a distribution is a formal convex finite sum $\sum_i r_i|x_i\rangle$ of elements x_i from some set X, and probabilities $r_i \in [0, 1]$ with $\sum_i r_i = 1$. We write $\mathcal{D}(X)$ for the set of such distributions over X. The mapping $X \mapsto \mathcal{D}(X)$ forms a monad on $\mathbf{Sets}$, just like $\mathcal{M}$. We shall write $\mathcal{K}\ell(\mathcal{D})$ for the Kleisli category of the monad $\mathcal{D}$.

A Kleisli map $f\colon X \to \mathcal{D}(Y)$ will also be written as $f\colon X \rightsquigarrow Y$ using a special arrow $\rightsquigarrow$. For a distribution $\omega \in \mathcal{D}(X)$ we write $f \gg \omega \in \mathcal{D}(Y)$ for the distribution obtained by Kleisli extension (state transformation), where:

$$\big(f \gg \omega\big)(y) \,=\, \mu\Big(\mathcal{D}(f)(\omega)\Big)(y) \,=\, \sum_{x\in X} \omega(x)\cdot f(x)(y).$$

Composition in $\mathcal{K}\ell(\mathcal{D})$ will be written as $\circ$ and can be described as $(g \circ f)(x) = g \gg f(x)$. An arbitrary *function* $h\colon X \to Y$ can be promoted to a *channel* $\langle h\rangle \coloneqq \eta \circ h\colon X \rightsquigarrow Y$. We frequently promote implicitly, when needed, and omit the brackets $\langle - \rangle$.

There is a natural transformation $\mathit{Flrn}\colon \mathcal{M}_* \Rightarrow \mathcal{D}$ which we call frequentist learning, since it involves learning by counting. It amounts to normalisation:

$$\mathit{Flrn}\Big(\sum\nolimits_i n_i|x_i\rangle\Big) \coloneqq \sum\nolimits_i \frac{n_i}{n}|x_i\rangle \quad \text{where} \quad n = \sum\nolimits_i n_i.$$

This *Flrn* is defined on non-empty multisets, given by $\mathcal{M}_*$. It is *not* a map of monads.

For two distributions $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$ we write $\omega \otimes \rho \in \mathcal{D}(X \times Y)$ for the product distribution. It is given by $(\omega \otimes \rho)(x, y) = \omega(x)\cdot\rho(y)$, like for multisets. This tensor $\otimes$ can be extended to channels as $(f \otimes g)(x, y) = f(x) \otimes g(y)$. In this way the Kleisli category $\mathcal{K}\ell(\mathcal{D})$ becomes symmetric monoidal. The tensor comes with projections, since the monad $\mathcal{D}$ is affine: $\mathcal{D}(1) = 1$, see *e.g.* [18] for details.

For a fixed number $K \in \mathbb{N}$ we will use a 'big' tensor $\bigotimes\colon \mathcal{D}(X)^K \to \mathcal{D}(X^K)$ given by $\bigotimes(\omega_1, \ldots, \omega_K) = \omega_1 \otimes \cdots \otimes \omega_K$. This $\bigotimes$ is natural in X and commutes appropriately with η and μ, making it a distributive law of the functor $(-)^K$ over the monad $\mathcal{D}$. We shall write $\mathit{iid}\colon \mathcal{D}(X) \rightsquigarrow X^K$ for $\mathit{iid}(\omega) = \bigotimes(\omega, \ldots, \omega) = \omega \otimes \cdots \otimes \omega = \omega^K$. This *iid* gives the independent and identical distribution.

IV. Accumulation and arrangement

There is a simple way to turn a list/sequence of elements into a multiset, simply by counting occurrences. We call this *accumulation* and write it as *acc*. Thus, for instance $\mathit{acc}(a, a, b, a) = 3|a\rangle + 1|b\rangle$. More generally we can simply define:

$$\mathit{acc}(x_1, \ldots, x_n) \coloneqq 1|x_1\rangle + \cdots + 1|x_n\rangle.$$

This accumulation is mathematically well-behaved. It forms for instance a map of monad, from lists to multisets.

A natural question that arises is: how many lists accumulate to a given multiset φ? The (standard) answer is given by what we call the *multiset coefficient* of the multiset φ and write as (φ). It is:

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

The latter $\big(\ -\ \big)$ notation is common for a multinomial coefficient, where $\mathit{supp}(\varphi) = \{x_1, \ldots, x_n\}$. For instance, for $\varphi = 2|a\rangle + 3|b\rangle$ there are $(\varphi) = \frac{5!}{2!\cdot 3!} = 10$ sequences with length 5 of a's and b's that accumulate to φ.

We shall concentrate on accumulation for a fixed size, and then write it as $\mathit{acc}[K]\colon X^K \to \mathcal{M}[K](X)$. The parameter $K \in \mathbb{N}$ is omitted when it is clear for the context. This accumulation map is the coequaliser of all permutation maps $X^K \to X^K$. We shall describe it more concretely.

***Lemma** 1:* For a number $K \in \mathbb{N}$, let $f\colon X^K \to Y$ be a function which is *stable under permutation*: for each permutation $\pi\colon \{1, \ldots, K\} \xrightarrow{\cong} \{1, \ldots, K\}$ one has $f\big(x_1, \ldots, x_K\big) = f\big(x_{\pi(1)}, \ldots, x_{\pi(K)}\big)$, for all sequences $(x_1, \ldots, x_K) \in X^K$. Then there is a unique function $\overline{f}\colon \mathcal{M}[K](X) \to Y$ with $\overline{f} \circ \mathit{acc} = f$, as in:

$$\begin{tikzcd} X^K \arrow[r, two heads, "\mathit{acc}"] \arrow[dr, bend right, "f"'] & \mathcal{M}[K](X) \arrow[d, dotted, "\overline{f}"] \\ & Y \end{tikzcd} \qquad \square$$

In the other direction, for a multiset $\varphi \in \mathcal{M}[K](X)$ we define $\mathit{arr}[K](\varphi) \in \mathcal{D}(X^K)$ as the (uniform) distribution over all sequences that accumulate to φ. Thus:

$$\mathit{arr}[K](\varphi) \coloneqq \sum_{\vec{x}\in \mathit{acc}^{-1}(\varphi)} \frac{1}{(\varphi)}\big|\vec{x}\big\rangle.$$

This is what we call arrangement. It may also be defined using that *acc* is coequaliser.

***Lemma** 2:* Consider accumulation and arrangement.

i) They are natural transformations $\mathit{acc}\colon (-)^K \Rightarrow \mathcal{M}[K]$ and $\mathit{arr}\colon \mathcal{M}[K] \Rightarrow \mathcal{D}((-)^K)$.

ii) $\mathit{acc} \circ \mathit{arr}$ is the identity channel $\mathcal{M}[K](X) \rightsquigarrow \mathcal{M}[K](X)$.

iii) The composite $\mathit{arr} \circ \mathit{acc}\colon X^K \rightsquigarrow X^K$ yields the uniform distribution of all $K!$ permutations. It commutes with the big tensor $\bigotimes$, as in the following diagram of channels.

$$\begin{tikzcd} \mathcal{D}(X)^K \arrow[r, "\mathit{acc}"] \arrow[d, "\bigotimes"'] & \mathcal{M}[K]\big(\mathcal{D}(X)\big) \arrow[r, "\mathit{arr}"] & \mathcal{D}(X)^K \arrow[d, "\bigotimes"] \\ X^K \arrow[r, "\mathit{acc}"] & \mathcal{M}[K](X) \arrow[r, "\mathit{arr}"] & X^K \end{tikzcd} \qquad \square$$

V. Multinomial distributions

We can think of a distribution $\omega \in \mathcal{D}(X)$ as an abstract urn, with X as set of colours for the balls in the urn. The number $\omega(x) \in [0, 1]$ gives the probability of drawing one ball of colour x. We consider drawing with replacement, so that the urn does not change. When we draw a handful of balls, we wish to know the probability of the draw. Such a draw, say of size K, will be represented as a multiset $\varphi \in \mathcal{M}[K](X)$. Multinomial distributions assign probabilities to fixed-size draws. They will be organised as a channel, like in [16], [17], of the form:

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

For clarity, we write this channel here as a function, with the distribution monad $\mathcal{D}$ explicitly present in its codomain. We shall see that writing it as channel $\mathcal{D}(X) \rightsquigarrow \mathcal{M}[K](X)$, and composing it as such, allows us to smoothly express various properties of multinomials. This demonstrates the power of (categorical) abstraction.

On $\omega \in \mathcal{D}(X)$, as abstract urn, the multinomial channel involves the convex sum over multisets, as draws:

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

The multiset coefficient $\boldsymbol{(}\varphi\boldsymbol{)}$ of a draw/multiset φ, from Section IV appears because the order of drawn elements is irrelevant.

The next result captures some basic intuitions about multinomials: they are suitably additive and the draws match the original distribution ω, so learning from them yields ω itself.

***Proposition** 3:* Consider the above multinomial channel.

i) Draws can be combined: for $K, L \in \mathbb{N}$,

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[r, "mn[K+L]"] \arrow[d, "\Delta"'] & \mathcal{M}[K+L](X) \\
\mathcal{D}(X)\times\mathcal{D}(X) \arrow[r, "mn[K]\otimes mn[L]"] & \mathcal{M}[K](X)\times\mathcal{M}[L](X) \arrow[u, "+"']
\end{tikzcd}$$

ii) $\big(\mathit{Flrn}\circ mn[K]\big)(\omega) = \omega.$ $\square$

***Theorem** 4:* Multinomial channels are related to accumulation, arrangement, and independent & identical distributions:

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[r, "mn[K]"] \arrow[dr, bend right, "iid[K]"'] & \mathcal{M}[K](X) \arrow[d, "arr[K]"] \\
 & X^K
\end{tikzcd}$$

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[rr, "mn[K]"] \arrow[dr, bend right, "iid[K]"'] & & \mathcal{M}[K](X) \\
 & X^K \arrow[ur, bend right, "acc[K]"'] &
\end{tikzcd}$$

An immediate consequence of this last diagram is that binomial channels form a natural transformation $\mathcal{D} \Rightarrow \mathcal{D}\mathcal{M}[K]$.

Proof For $\omega \in \mathcal{D}(X)$ and $\vec{x} = (x_1, \ldots, x_K) \in X^K$,

$$\begin{aligned}
&\big(arr \circ mn[K]\big)(\omega)(\vec{x}) \\
&= \sum_{\varphi\in\mathcal{M}[K](X)} arr(\varphi)(\vec{x})\cdot mn[K](\omega)(\varphi) \\
&= \frac{1}{\boldsymbol{(}acc(\vec{x})\boldsymbol{)}}\cdot mn[K](\omega)(acc(\vec{x})) \\
&= \frac{1}{\boldsymbol{(}acc(\vec{x})\boldsymbol{)}}\cdot \boldsymbol{(}acc(\vec{x})\boldsymbol{)}\cdot\prod\nolimits_y \omega(y)^{acc(\vec{x})(y)} \\
&= \prod\nolimits_i \omega(x_i) \;=\; \omega^K(\vec{x}) \;=\; iid(\omega)(\vec{x}).
\end{aligned}$$

The second diagram then commutes since $acc \circ arr = \mathrm{id}$. $\square$

We conclude our description of multinomial channels by repeating a result from [16]. It tells that multinomials form a cone for an infinite chain of draw-and-delete channels $DD[K] \colon \mathcal{M}[K+1](X) \rightsquigarrow \mathcal{M}[K](X)$. In [16] this forms the basis for a re-description of de Finetti's theorem in terms of a limit in a category of channels. We refer the reader to *loc. cit.* for further information; here we just need the cone property in Lemma 5 (i) below. First we define withdrawing a single element from a multiset/urn as a distribution.

$$DD[K](\psi) \;:=\; \sum_{x\in supp(\psi)} \frac{\psi(x)}{K+1}\big|\,\psi - 1|x\rangle\big\rangle.$$

Notice that we sum over elements x in the support of the multiset/urn ψ of size $K+1$, so that $\psi(x) > 0$. Hence we may remove x from ψ, as indicated by the subtraction $\psi - 1|x\rangle$. It leaves us with a multiset of size K. The probability of drawing x depends on the number of occurrences $\psi(x)$. We may also write the associated probability $\frac{\psi(x)}{K+1}$ as $\mathit{Flrn}(\psi)(x)$. For instance:

$$DD\big(3|a\rangle + 2|b\rangle\big) \;=\; \tfrac{3}{5}\big|2|a\rangle + 2|b\rangle\big\rangle + \tfrac{2}{5}\big|2|a\rangle + 1|b\rangle\big\rangle.$$

It is not hard to see that these DD maps are natural in X. The key property that we are interested in is that draw-and-delete commutes with multinomial channels.

***Lemma** 5:* Draw-and-delete of a single element via the channel $DD\colon \mathcal{M}[K+1](X) \rightsquigarrow \mathcal{M}[K](X)$ satisfies the following properties.

i) Commutation with multinomials:

$$\begin{tikzcd}
\mathcal{M}[K+1](X) \arrow[rr, "DD"] & & \mathcal{M}[K](X) \\
 & \mathcal{D}(X) \arrow[ul, bend left, "mn[K+1]"] \arrow[ur, bend right, "mn[K]"'] &
\end{tikzcd}$$

ii) Commutation with frequentist learning:

$$\begin{tikzcd}
\mathcal{M}[K+1](X) \arrow[rr, "DD"] \arrow[dr, bend right, "Flrn"'] & & \mathcal{M}[K](X) \arrow[dl, bend left, "Flrn"] \\
 & X &
\end{tikzcd}$$

Proof The first point is proven in [16], so we concentrate on the second one. For $\psi \in \mathcal{M}[K+1](X)$ and $y \in X$,

$$\begin{aligned}
&\big(\mathit{Flrn}\circ DD\big)(\psi)(y) \\
&= \sum_{\varphi\in\mathcal{M}[K](X)} DD(\psi)(\varphi)\cdot \mathit{Flrn}(\varphi)(y) \\
&= \sum_{x\in X} \frac{\psi(x)}{K+1}\cdot \mathit{Flrn}(\psi - 1|x\rangle)(y) \\
&= \frac{\psi(y)}{K+1}\cdot\frac{\psi(y)-1}{K} + \sum_{x\neq y}\frac{\psi(x)}{K+1}\cdot\frac{\psi(y)}{K} \\
&= \frac{\psi(y)}{K(K+1)}\cdot\left(\psi(y) - 1 + \sum_{x\neq y}\psi(x)\right) \\
&= \frac{\psi(y)}{K(K+1)}\cdot\left(\left(\sum_x \psi(x)\right) - 1\right) \\
&= \frac{\psi(y)}{K(K+1)}\cdot((K+1)-1) \\
&= \frac{\psi(y)}{K+1} \\
&= \mathit{Flrn}(\psi)(y).
\end{aligned}$$

$\square$

VI. Hypergeometric distributions

Where multinomial distributions describe draws with replacement, hypergeometric distributions involve actual withdrawals, without replacement. The urn itself is thus a multiset, say of size N, and withdrawals are possible of multisets of (fixed) size K, for $K \leq N$. The type of the channel is thus:

$$\mathcal{M}[N](X) \xrightarrow{\;hg[K]\;} \mathcal{D}\big(\mathcal{M}[K](X)\big). \tag{1}$$

In the definition below we use a pointwise ordering $\varphi \leq \psi$, for multisets φ, ψ on the same set X. It means that $\varphi(x) \leq \psi(x)$ for each $x \in X$. We write $\varphi \leq_K \psi$ if $\varphi \in \mathcal{M}[K](X)$ and $\varphi \leq \psi$.

$$hg[K]\big(\psi\big) \coloneqq \sum_{\varphi \leq_K \psi} \frac{\prod_x \binom{\psi(x)}{\varphi(x)}}{\binom{N}{K}} \,\big|\,\varphi\,\big\rangle,$$

where $N = \|\psi\|$. The probabilities involved add up to one by (a generalisation of) Vandermonde's formula $\binom{M+L}{K} = \sum_{m \leq M, \ell \leq L, m+\ell=K} \binom{M}{m} \cdot \binom{L}{\ell}$.

The main result about these hypergeometric distributions in this section is that they can be obtained by iterated draw-and-delete. In a slightly different way it is shown in [17] that both multinomial and hypergeometric channels can be obtained via iterated drawing, with or without replacement, where iteration is described in terms of a Kleisli composition.

Theorem 6: For $K, L \in \mathbb{N}$, the hypergeometric channel $\mathcal{M}[K+L](X) \rightsquigarrow \mathcal{M}[K](X)$ equals L consecutive draw-and-delete's:

$$\begin{tikzcd}
\mathcal{M}[K+L](X) \arrow[rrr, "hg{[K]}"] \arrow[dr, "DD"'] & & & \mathcal{M}[K](X) \\
& \mathcal{M}[K+L-1](X) \arrow[r, dotted, bend right, "\underbrace{DD \circ \cdots \circ DD}_{L-2 \text{ times}}"'] & \mathcal{M}[K+1](X) \arrow[ur, "DD"'] &
\end{tikzcd}$$

Proof By induction on L. □

From this result, and Lemma 5 about draw-and-delete, we can deduce many additional facts about hypergeometric distributions.

Corollary 7:

i) Hypergeometric channels (1) are natural in X.
ii) Frequentist learning from hypergeometric draws is like learning from the urn:

$$\begin{tikzcd}
\mathcal{M}[N](X) \arrow[rr, "hg{[K]}"] \arrow[dr, "Flrn"'] & & \mathcal{M}[K](X) \arrow[dl, "Flrn"] \\
& X &
\end{tikzcd}$$

iii) Hypergeometric channels compose, as in:

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

iv) Hypergeometric and multinomial channels commute:

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

□

VII. The parallel multinomial law: 4 definitions

We have already seen the close connection between multisets and distributions. This section focuses on a very special 'distributivity' connection that shows how a multiset of distributions can be transformed into a distribution over multisets. This is a rather complicated operation, but it is fundamental; it can be described via a tensor product $\otimes$ of multinomials, and will therefor be called the *parallel multinomial law*, abbreviated as *pml*.

This law *pml* turns out to be a distributive law, in a categorical sense. It has popped up in [13], [14] and is used in [15], for continuous probability, to describe 'point processes' as distributions over multisets. This law satisfies several elementary properties that combine basic ingredients of probability theory.

This parallel multinomial law *pml* that we are after has the following type. For a number $K \in \mathbb{N}$ and a set X it is a function:

$$\mathcal{M}[K]\Big(\mathcal{D}(X)\Big) \xrightarrow{\;pml\;} \mathcal{D}\Big(\mathcal{M}[K](X)\Big). \tag{2}$$

The dependence of *pml* on K and X is left implicit. Later on we develop a version of *pml* without the size parameter K.

This map *pml* turns a K-element multiset of distributions over X into a distribution over K-element multisets over X. It is not immediately clear how to do this. It turns out that there are several ways to describe *pml*. This section is devoted solely to defining this law, in four different manners — yielding each time the same result. The subsequent section collects basic properties of *pml*.

First definition: Since the law that we are after is rather complicated, we start with an example.

Example 8: Let $X = \{a, b\}$ be a set with two distributions $\omega, \rho \in \mathcal{D}(X)$, given by:

$$\omega = \tfrac{1}{3}|a\rangle + \tfrac{2}{3}|b\rangle \qquad \text{and} \qquad \rho = \tfrac{3}{4}|a\rangle + \tfrac{1}{4}|b\rangle.$$

We will define *pml* on the multiset of distributions $2|\omega\rangle + 1|\rho\rangle$ of size $K = 3$. The result should be a distribution on multisets of size $K = 3$ over X. There are four such multisets, namely:

$$3|a\rangle \qquad 2|a\rangle + 1|b\rangle \qquad 1|a\rangle + 2|b\rangle \qquad 3|b\rangle.$$

The goal is to assign a probability to each of them. The map *pml* does this in the following way:

$$
\begin{aligned}
& \mathit{pml}\big(2|\omega\rangle + 1|\rho\rangle\big) \\
&\quad = \omega(a)\cdot\omega(a)\cdot\rho(a)\big|3|a\rangle\big\rangle \\
&\qquad + \Big(\omega(a)\cdot\omega(a)\cdot\rho(b) + \omega(a)\cdot\omega(b)\cdot\rho(a) \\
&\qquad\qquad + \omega(b)\cdot\omega(a)\cdot\rho(a)\Big)\big|2|a\rangle + 1|b\rangle\big\rangle \\
&\qquad + \Big(\omega(a)\cdot\omega(b)\cdot\rho(b) + \omega(b)\cdot\omega(a)\cdot\rho(b) \\
&\qquad\qquad + \omega(b)\cdot\omega(b)\cdot\rho(a)\Big)\big|1|a\rangle + 2|b\rangle\big\rangle \\
&\qquad + \omega(b)\cdot\omega(b)\cdot\rho(b)\big|3|b\rangle\big\rangle \\
&\quad = \tfrac{1}{3}\cdot\tfrac{1}{3}\cdot\tfrac{3}{4}\big|3|a\rangle\big\rangle \\
&\qquad + \Big(\tfrac{1}{3}\cdot\tfrac{1}{3}\cdot\tfrac{1}{4} + \tfrac{1}{3}\cdot\tfrac{2}{3}\cdot\tfrac{3}{4} + \tfrac{2}{3}\cdot\tfrac{1}{3}\cdot\tfrac{3}{4}\Big)\big|2|a\rangle + 1|b\rangle\big\rangle \\
&\qquad + \Big(\tfrac{1}{3}\cdot\tfrac{2}{3}\cdot\tfrac{1}{4} + \tfrac{2}{3}\cdot\tfrac{1}{3}\cdot\tfrac{1}{4} + \tfrac{2}{3}\cdot\tfrac{2}{3}\cdot\tfrac{3}{4}\Big)\big|1|a\rangle + 2|b\rangle\big\rangle \\
&\qquad + \tfrac{2}{3}\cdot\tfrac{2}{3}\cdot\tfrac{1}{4}\big|3|b\rangle\big\rangle \\
&\quad = \tfrac{1}{12}\big|3|a\rangle\big\rangle + \tfrac{13}{36}\big|2|a\rangle + 1|b\rangle\big\rangle \\
&\qquad + \tfrac{4}{9}\big|1|a\rangle + 2|b\rangle\big\rangle + \tfrac{1}{9}\big|3|b\rangle\big\rangle.
\end{aligned}
$$

Notice that the larger outer brackets $\big|-\big\rangle$ involve a distribution over multisets, given by the smaller inner brackets $|-\rangle$.

We now formulate the function *pml* from (2) in general, for arbitrary K and X. It is defined on a multiset $\sum_i n_i|\omega_i\rangle$ with multiplicities $n_i \in \mathbb{N}$ satisfying $\sum_i n_i = K$, and with distributions $\omega_i \in \mathcal{D}(X)$. The number $\mathit{pml}\big(\sum_i n_i|\omega_i\rangle\big)(\varphi)$ describes the probability of the K-element multiset φ over X, by using for each element occurring in φ the probability of that element in the corresponding distribution in $\sum_i n_i|\omega_i\rangle$.

In order to make this description precise we assume that the indices i are somehow ordered, say as $i_1, \ldots, i_m$ and use this ordering to form a product distribution $\bigotimes_i \omega_i^{n_i} \in \mathcal{D}\big(X^K\big)$.

$$\bigotimes_i \omega_i^{n_i} = \underbrace{\omega_{i_1}\otimes\cdots\otimes\omega_{i_1}}_{n_{i_1}\text{ times}} \otimes\ \cdots\ \otimes \underbrace{\omega_{i_m}\otimes\cdots\otimes\omega_{i_m}}_{n_{i_m}\text{ times}}.$$

Now we formulate the first definition:

$$
\begin{aligned}
& \mathit{pml}\big(\textstyle\sum_i n_i|\omega_i\rangle\big) \\
&\quad := \sum_{\vec{x}\in X^K} \Big(\bigotimes \omega_i^{n_i}\Big)(\vec{x})\,\big|\mathit{acc}(\vec{x})\big\rangle \\
&\quad = \sum_{\varphi\in\mathcal{M}[K](X)} \left(\sum_{\vec{x}\in \mathit{acc}^{-1}(\varphi)} \Big(\bigotimes \omega_i^{n_i}\Big)(\vec{x})\right)\big|\varphi\big\rangle.
\end{aligned}
\tag{3}
$$

This formulation has been used in Example 8.

Second definition: There is an alternative formulation of the parallel multinomial law, using multiple multinomial distributions, put in parallel via a tensor product $\otimes$. This formulation is the basis for the phrase 'parallel multinomial'.

$$
\begin{aligned}
& \mathit{pml}\big(\textstyle\sum_i n_i|\omega_i\rangle\big) \\
&\quad := \mathcal{D}(+)\Big(\bigotimes_i \mathit{mn}[n_i](\omega_i)\Big) \\
&\quad = \sum_{i,\,\varphi_i\in\mathcal{M}[n_i](X)} \Big(\textstyle\prod_i \mathit{mn}[n_i](\omega_i)(\varphi_i)\Big)\,\big|\textstyle\sum_i \varphi_i\big\rangle
\end{aligned}
\tag{4}
$$

The sum that we use here as type:

$$\mathcal{M}[n_{i_1}](X)\times\cdots\times\mathcal{M}[n_{i_m}](X) \xrightarrow{+} \mathcal{M}[\underbrace{n_{i_1}+\cdots+n_{i_m}}_{K}](X).$$

Third definition: Our third definition of *pml* is more abstract. It uses the coequaliser property of Lemma 1. It determines *pml* as the unique (dashed) map in:

$$
\begin{tikzcd}
\mathcal{D}(X)^K \arrow[rr, two heads, "\mathit{acc}"] \arrow[dr, "\otimes"'] & & \mathcal{M}[K]\big(\mathcal{D}(X)\big) \arrow[dd, dashed, "\mathit{pml}"] \\
& \mathcal{D}(X^K) \arrow[dr, "\mathcal{D}(\mathit{acc})"'] & \\
& & \mathcal{D}\big(\mathcal{M}[K](X)\big)
\end{tikzcd}
\tag{5}
$$

There is an important side-condition in Lemma 1, namely that the composite $\mathcal{D}(\mathit{acc}) \circ \bigotimes\colon \mathcal{D}(X)^K \to \mathcal{D}\big(\mathcal{M}[K](X)\big)$ is stable under permutations. This is easy to check.

This third formulation of the parallel multinomial law is not very useful for actual calculations, like in Example 8. But it is useful for proving properties about *pml*, via the uniqueness part of the third definition.

Fourth definition: For our fourth and last definition we have to piece together some basic observations.

i) (From [13, p.82]) If M is a commutative monoid, then so is the set $\mathcal{D}(M)$ of distributions on M, with sum:

$$
\begin{aligned}
\omega + \rho &:= \mathcal{D}(+)(\omega\otimes\rho) \\
&= \sum_{x,y\in M} (\omega\otimes\rho)(x,y)\big|x{+}y\big\rangle.
\end{aligned}
\tag{6}
$$

ii) Such commutative monoid structure corresponds to an Eilenberg-Moore algebra $\alpha\colon \mathcal{M}(\mathcal{D}(M)) \to \mathcal{D}(M)$ of the multiset monad $\mathcal{M}$, given by:

$$\alpha\big(\textstyle\sum_i n_i|\omega_i\rangle\big) = \displaystyle\sum_{\vec{x}\in M^K} \big(\textstyle\bigotimes_i \omega_i^{n_i}\big)(\vec{x})\,\big|\textstyle\sum\vec{x}\,\big\rangle$$

where $K = \sum_i n_i$.

iii) Applying the previous two points with commutative monoid $M = \mathcal{M}(X)$ of multisets yields an Eilenberg-Moore algebra:

$$\mathcal{M}\Big(\mathcal{D}\big(\mathcal{M}(X)\big)\Big) \xrightarrow{\alpha} \mathcal{D}\big(\mathcal{M}(X)\big) \tag{7}$$

We now define:

$$\mathit{pml} := \Big(\mathcal{M}\mathcal{D}(X) \xrightarrow{\mathcal{M}\mathcal{D}(\eta)} \mathcal{M}\mathcal{D}\mathcal{M}(X) \xrightarrow{\alpha} \mathcal{D}\mathcal{M}(X)\Big). \tag{8}$$

Proposition 9: The definition of *pml* in (8) restricts to $\mathcal{M}[K](\mathcal{D}(X)) \to \mathcal{D}\big(\mathcal{M}[K](X)\big)$, for each $K \in \mathbb{N}$. This restriction is the same *pml* as defined in (3), (4) and (5).

Proof For the equivalence of the first two formulations (3) and (4) we use Theorem 4. This definition of *pml* commutes with *acc* and is thus the same as the third one in (5). Finally, when we unravel the fourth formulation (8) we get the second one (4). □

VIII. The parallel multinomial law: properties

The next result enriches what we already know: the rectangle on the right below is added to the (known) one on the left.

***Proposition** 10:* The parallel multinomial law *pml* is the unique channel making both rectangles below commute.

$$\begin{tikzcd} \mathcal{D}(X)^K \arrow[r, "acc"] \arrow[d, "\bigotimes"'] & \mathcal{M}[K]\big(\mathcal{D}(X)\big) \arrow[r, "arr"] \arrow[d, dotted, "pml"'] & \mathcal{D}(X)^K \arrow[d, "\bigotimes"] \\ X^K \arrow[r, "acc"] & \mathcal{M}[K](X) \arrow[r, "arr"] & X^K \end{tikzcd}$$

Proof The rectangle on the left is the third formulation of *pml* in (5) and thus provides uniqueness. Commutation of the rectangle of the right follows from a uniqueness argument, using that the outer rectangle commutes by Lemma 2 (iii):

$$\begin{array}{rcll} \bigotimes \circ arr \circ acc & = & arr \circ acc \circ \bigotimes & \text{by Lemma 2 (iii)} \\ & = & arr \circ pml \circ acc & \text{by (5).} \end{array} \qquad \square$$

This result shows that *pml* is squeezed between $\bigotimes$, both on the left and on the right. This $\bigotimes$ is a distributive law. We shall prove the same about *pml* below.

But first we look at interaction with frequentist learning.

***Theorem** 11:* The distributive law *pml* commutes with frequentist learning, in the sense that for $\Psi \in \mathcal{M}[K]\big(\mathcal{D}(X)\big)$,

$$Flrn \gg pml(\Psi) \,=\, \mu\big(Flrn(\Psi)\big).$$

Equivalently, in diagrammatic form:

$$\begin{tikzcd} \mathcal{M}[K]\big(\mathcal{D}(X)\big) \arrow[r, "pml"] \arrow[d, "Flrn"'] & \mathcal{M}[K](X) \arrow[d, "Flrn"] \\ \mathcal{D}(X) \arrow[r, "\mathrm{id}"] & X \end{tikzcd}$$

The channel $\mathcal{D}(X) \rightsquigarrow X$ at the bottom is the identity function $\mathcal{D}(X) \to \mathcal{D}(X)$.

Proof Let $\Psi = \sum_i n_i|\omega_i\rangle \in \mathcal{M}[K]\big(\mathcal{D}(X)\big)$.

$$\begin{array}{ll} & Flrn \gg pml(\Psi) \\ \overset{(4)}{=} & Flrn \gg \mathcal{D}(+)\Big(\bigotimes_i mn[n_i](\omega_i)\Big) \\ \overset{(*)}{=} & \sum_i \frac{n_i}{K} \cdot \big(Flrn \gg mn[n_i](\omega_i)\big) \\ = & \sum_i \frac{n_i}{K} \cdot \omega_i \qquad \text{by Proposition 3 (ii)} \\ = & \mu\big(\sum_i \frac{n_i}{K}|\omega_i\rangle\big) \\ = & \mu\big(Flrn(\Psi)\big). \end{array}$$

The marked equation $\overset{(*)}{=}$ follows from (a generalisation of) the following fact, for $\Omega \in \mathcal{D}(\mathcal{M}[K](X))$, $\Theta \in \mathcal{D}(\mathcal{M}[L](X)$.

$$\begin{array}{ll} & Flrn \gg \mathcal{D}(+)(\Omega \otimes \Theta) \\ = & \frac{K}{K+L} \cdot \big(Flrn \gg \Omega\big) + \frac{L}{K+L} \cdot \big(Flrn \gg \Theta\big). \end{array} \qquad \square$$

Our next result shows that the parallel multinomial law commutes with hypergeometric channels.

***Theorem** 12:* The parallel multinomial *pml* commutes with draw-and-delete:

$$\begin{tikzcd} \mathcal{M}[K+1]\big(\mathcal{D}(X)\big) \arrow[r, "DD"] \arrow[d, "pml"'] & \mathcal{M}[K]\big(\mathcal{D}(X)\big) \arrow[d, "pml"] \\ \mathcal{M}[K+1](X) \arrow[r, "DD"] & \mathcal{M}[K](X) \end{tikzcd}$$

and then also with hypergeometrics: for $N \geq K$ one has,

$$\begin{tikzcd} \mathcal{M}[N]\big(\mathcal{D}(X)\big) \arrow[r, "hg{[K]}"] \arrow[d, "pml"'] & \mathcal{M}[K]\big(\mathcal{D}(X)\big) \arrow[d, "pml"] \\ \mathcal{M}[N](X) \arrow[r, "hg{[K]}"] & \mathcal{M}[K](X) \end{tikzcd}$$

Proof Showing commutation of *pml* with *DD* suffices, by Theorem 6. We use an auxiliary probabilistic projection channel $ppr\colon X^{K+1} \rightsquigarrow X^K$. It is a uniform distribution over projecting away at each position.

$$\begin{array}{l} ppr(x_1, \ldots, x_{K+1}) \\ \quad := \displaystyle\sum_{1 \leq k \leq K+1} \frac{1}{K+1}\big|x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_{K+1}\big\rangle. \end{array}$$

It is not hard to see that the following diagram commutes.

$$\begin{tikzcd} \mathcal{D}(X)^{K+1} \arrow[r, "\bigotimes"] \arrow[d, "ppr"'] & X^{K+1} \arrow[r, "acc"] \arrow[d, "ppr"'] & \mathcal{M}[K+1](X) \arrow[r, "arr"] \arrow[d, "DD"] & X^{K+1} \arrow[d, "ppr"] \\ \mathcal{D}(X)^K \arrow[r, "\bigotimes"] & X^K \arrow[r, "acc"] & \mathcal{M}[K](X) \arrow[r, "arr"] & X^K \end{tikzcd}$$

In the end we use that *acc* is coequaliser (so surjective/epic):

$$\begin{array}{rcl} DD \circ pml \circ acc & \overset{(5)}{=} & DD \circ acc \circ \bigotimes \\ & = & acc \circ ppr \circ \bigotimes \\ & = & acc \circ \bigotimes \circ ppr \\ & \overset{(5)}{=} & pml \circ acc \circ ppr \\ & = & pml \circ DD \circ acc. \end{array} \qquad \square$$

***Proposition** 13:* The parallel multinomial law commutes with sums of multisets:

$$\begin{tikzcd} \mathcal{M}[K]\big(\mathcal{D}(X)\big) \times \mathcal{M}[L]\big(\mathcal{D}(X)\big) \arrow[r, "pml \otimes pml"] \arrow[d, "+"'] & \mathcal{M}[K](X) \times \mathcal{M}[L](X) \arrow[d, "+"] \\ \mathcal{M}[K+L]\big(\mathcal{D}(X)\big) \arrow[r, "pml"] & \mathcal{M}[K+L](X) \end{tikzcd}$$

Proof Via a suitable generalisation of Proposition 3 (i). $\square$

***Theorem** 14:* The parallel multinomial law *pml* is a *distributive law* in several ways.

i) It is a distributive law of the multiset *functor* $\mathcal{M}[K]$ over the distribution monad $\mathcal{D}$. This means that *pml* commutes with the unit and multiplication operations of $\mathcal{D}$. Equivalently, $\mathcal{M}[K]$ can be lifted to a functor $\mathcal{K}\ell(\mathcal{D}) \to \mathcal{K}\ell(\mathcal{D})$.
ii) It is thereby also a law $\mathcal{M}\mathcal{D} \Rightarrow \mathcal{D}\mathcal{M}$ of *functor* $\mathcal{M}$ over the monad $\mathcal{D}$, as in the fourth definition (8), so that $\mathcal{M}$ also lifts to $\mathcal{K}\ell(\mathcal{D})$.
iii) In addition, $pml\colon \mathcal{M}\mathcal{D} \Rightarrow \mathcal{D}\mathcal{M}$ is a distributive law of *monads*, of $\mathcal{M}$ over $\mathcal{D}$. Equivalently, $\mathcal{D}$ can be

lifted to a monad $\mathcal{D}\colon \mathbf{CMon} \to \mathbf{CMon}$, on the category $\mathbf{CMon} = \mathcal{EM}(\mathcal{M})$ of commutative monoids (Eilenberg-Moore algebras of $\mathcal{M}$).

See *e.g.* [7] for a detailed description of the correspondence between laws and lifting. The last point means that the composite $\mathcal{DM}\colon \mathbf{Sets} \to \mathbf{Sets}$ is a monad. A continuous version of this monad is used in [15], as the monad of point processes.

Proof An easy way to prove the first point is to use that $\bigotimes\colon \mathcal{D}(X)^K \to \mathcal{D}(X^K)$ is a distributive law and that $acc\colon X^K \to \mathcal{M}[K](X)$ is surjective. The second point then follows easily. For the third point it is easiest to check that $\mathcal{D}(M)$ is a commutative monoid if M is, using the formula (6) and that this gives a monad on $\mathbf{CMon}$. □

IX. Zipping multisets

So far we have seen (parallel) multinomial and hypergeometric channels and looked at some basic properties, involving their sequential behaviour, in the Kleisli category $\mathcal{K}\ell(\mathcal{D})$. We now move to their parallel properties, in terms of tensors $\otimes$. This requires a new (probabilistic) operation to combine multisets (of the same size) which we call multizip, and write as *mzip*. This section concentrates on this *mzip*, and the next section will look into its interaction with the main probabilistic channels of this paper.

In (functional) programming zipping two lists together is a common operation. Mathematically it takes the form of function $zip\colon X^K \times Y^K \xrightarrow{\cong} (X \times Y)^K$. This section shows how to do a form of zipping for multisets. This is a new and very useful operation, that, as we shall see later, commutes with basic probabilistic mappings.

But first we need to collect two auxiliary properties of ordinary zip.

***Lemma** 15:* Zipping commutes with tensors of distributions in the following ways.

$$\begin{tikzcd}
\mathcal{D}(X)\times\mathcal{D}(Y) \arrow[r, "iid\otimes iid"] \arrow[d, "\otimes"'] & X^K\times Y^K \arrow[d, "zip"] \\
\mathcal{D}(X\times Y) \arrow[r, "iid"] & (X\times Y)^K
\end{tikzcd}$$

$$\begin{tikzcd}
\mathcal{D}(X)^K \times \mathcal{D}(Y)^K \arrow[rr, "\bigotimes\otimes\bigotimes"] \arrow[d, "zip"'] & & X^K \times Y^K \arrow[d, "zip"] \\
\big(\mathcal{D}(X) \times \mathcal{D}(Y)\big)^K \arrow[r, "\otimes^K"] & \mathcal{D}(X \times Y)^K \arrow[r, "\bigotimes"] & (X \times Y)^K
\end{tikzcd}$$

Proof By straightforward calculation. □

When we take sizes into account, we see that the sum of two multisets is a function $\mathcal{M}[K](X) \times \mathcal{M}[L](X) \to \mathcal{M}[K+L](X)$. The product (tensor) has type $\mathcal{M}[K](X) \times \mathcal{M}[L](Y) \to \mathcal{M}[K\cdot L](X \times Y)$. The multizip function *mzip* that we are about to introduce takes inputs of the same size and does not change this size. Its type is thus given by the top row of the next diagram. The rest of the diagram forms its definition:

$$\begin{tikzcd}
\mathcal{M}[K](X) \times \mathcal{M}[K](Y) \arrow[r, "mzip{[K]}"] \arrow[d, "arr\otimes arr"'] \arrow[dr, phantom, "\text{definition of } mzip"] & \mathcal{M}[K](X \times Y) \\
X^K \times Y^K \arrow[r, "zip"] & (X \times Y)^K \arrow[u, "acc"']
\end{tikzcd}$$

Explicitly, for multisets $\varphi \in \mathcal{M}[K](X)$ and $\psi \in \mathcal{M}[K](Y)$,

$$mzip(\varphi, \psi) \coloneqq \sum_{\vec{x}\in acc^{-1}(\varphi)} \; \sum_{\vec{y}\in acc^{-1}(\psi)} \frac{1}{(\!(\varphi)\!)\cdot(\!(\psi)\!)}\big|\, acc\big(zip(\vec{x},\vec{y})\big)\big\rangle.$$

Thus, in zipping two multisets φ and ψ we first look at all their arrangements $\vec{x}$ and $\vec{y}$, zip these together in the ordinary way, and accumulate the resulting list of pairs into a multiset again. We elaborate how this works concretely.

Example 16: Let's use two set $X = \{a, b\}$ and $Y = \{0, 1\}$ with two multisets of size three:

$$\varphi = 1|a\rangle + 2|b\rangle \qquad \text{and} \qquad \psi = 2|0\rangle + 1|1\rangle.$$

Then:

$$(\!(\varphi)\!) = \binom{3}{1,2} = 3 \qquad (\!(\psi)\!) = \binom{3}{2,1} = 3$$

The sequences in X^3 and Y^3 that accumulate to φ and ψ are:

$$\begin{cases} a, b, b \\ b, a, b \\ b, b, a \end{cases} \qquad \text{and} \qquad \begin{cases} 0, 0, 1 \\ 0, 1, 0 \\ 1, 0, 0 \end{cases}$$

Zipping them together gives the following nine sequences in $(X \times Y)^3$.

$$\begin{array}{lll}
(a,0),(b,0),(b,1) & (b,0),(a,0),(b,1) & (b,0),(b,0),(a,1) \\
(a,0),(b,1),(b,0) & (b,0),(a,1),(b,0) & (b,0),(b,1),(a,0) \\
(a,1),(b,0),(b,0) & (b,1),(a,0),(b,0) & (b,1),(b,0),(a,0)
\end{array}$$

By applying the accumulation function *acc* to each of these only two multisets remain:

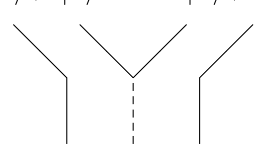
$$\text{three of } 1|a,1\rangle+2|b,0\rangle \qquad \text{six of } 1|a,0\rangle+1|b,0\rangle+1|b,1\rangle.$$

Finally, multiplication with $\frac{1}{(\!(\varphi)\!)\cdot(\!(\psi)\!)} = \frac{1}{9}$ gives the outcome:

$$\begin{array}{l}
mzip[3]\Big(1|a\rangle+2|b\rangle,\; 2|0\rangle+1|1\rangle\Big) \\
\quad = \frac{1}{3}\big|1|a,1\rangle+2|b,0\rangle\big\rangle \;+\; \frac{2}{3}\big|1|a,0\rangle+1|b,0\rangle+1|b,1\rangle\big\rangle.
\end{array}$$

This shows that calculating *mzip* is laborious. But it is quite mechanical and easy to implement. The picture below suggests to look at *mzip* as a funnel with two input pipes in which multiple elements from both sides can be combined into a probabilistic mixture.

$1|a\rangle+2|b\rangle \qquad 2|0\rangle+1|1\rangle$



$\frac{1}{3}\big|1|a,1\rangle+2|b,0\rangle\big\rangle \;+\; \frac{2}{3}\big|1|a,0\rangle+1|b,0\rangle+1|b,1\rangle\big\rangle$

***Lemma** 17:* Consider multizip *mzip* as channel $\mathcal{M}[K](X)\times\mathcal{M}[K](Y) \rightsquigarrow \mathcal{M}[K](X\times Y)$.

i) Zipping of multisets is a natural transformation $\mathcal{M}[K](-)\times\mathcal{M}[K](-) \Rightarrow \mathcal{D}\mathcal{M}[K](-\times-)$.
ii) $mzip(\varphi, K|y\rangle) = 1\big|\varphi\otimes 1|y\rangle\big\rangle$.
iii) *mzip* is associative, in $\mathcal{K}\ell(\mathcal{D})$.
iv) *mzip* commutes with projections, as in $\mathcal{M}[K](\pi_i) \circ mzip = \pi_i$, but not with diagonals.
v) Arrangement *arr* relates *zip* and *mzip* as in:

$$\begin{tikzcd} \mathcal{M}[K](X)\times\mathcal{M}[K](Y) \arrow[r, "arr\otimes arr"] \arrow[d, "mzip"'] & X^X\times Y^K \arrow[d, "zip"] \\ \mathcal{M}[K](X\times Y) \arrow[r, "arr"] & (X\times Y)^K \end{tikzcd}$$

But *mzip* and *zip* do not commute with *acc* instead of *arr*.
vi) Draw-and-delete commutes with *mzip*, as in:

$$\begin{tikzcd} \mathcal{M}[K+1](X)\times\mathcal{M}[K+1](Y) \arrow[r, "DD\otimes DD"] \arrow[d, "mzip"'] & \mathcal{M}[K](X)\times\mathcal{M}[K](Y) \arrow[d, "mzip"] \\ \mathcal{M}[K+1](X\times Y) \arrow[r, "DD"] & \mathcal{M}[K](X\times Y) \end{tikzcd}$$

Proof Most of these points are relatively straightforward, except the last point. Recall the probabilistic projection channel $ppr\colon X^{K+1} \rightsquigarrow X^K$ from the proof of Theorem 12. It commutes with *acc* and *arr*, but not with (ordinary) *zip*. The actual projection function $\pi\colon X^{K+1}\to X^K$, which removes the last element, does commute with *zip*. One can change appropriately from *ppr* to π since $ppr \circ arr = \pi \circ arr$. □

The following result deserves a separate status. It tells that what we learn from a zip of multisets is the same as what we learn from a parallel product (of these multisets). On a related note we illustrate in Section XI that multizip can be used for compositional parallel sampling.

***Theorem** 18:* Multizip and frequentist learning interact well, namely as:

$$Flrn \gg mzip(\varphi,\psi) \;=\; Flrn(\varphi\otimes\psi).$$

Equivalently, in diagrammatic form:

$$\begin{tikzcd} \mathcal{M}[K](X)\times\mathcal{M}[K](Y) \arrow[r, "mzip"] \arrow[d, "\otimes"'] & \mathcal{M}[K](X\times Y) \arrow[d, "Flrn"] \\ \mathcal{M}[K^2](X\times Y) \arrow[r, "Flrn"] & X\times Y \end{tikzcd}$$

Proof Let multisets $\varphi\in\mathcal{M}[K](X)$ and $\psi\in\mathcal{M}[K](Y)$ be given and let $a\in X$ and $b\in Y$ be arbitrary elements. We need to show that the probability:

$$\begin{aligned} &\Big(Flrn \gg mzip(\varphi,\psi)\Big)(a,b) \\ &= \sum_{\vec{x}\in acc^{-1}(\varphi)}\;\sum_{\vec{y}\in acc^{-1}(\psi)} \frac{acc\big(zip(\vec{x},\vec{y})\big)(a,b)}{K\cdot (\varphi)\cdot(\psi)} \end{aligned}$$

is the same as the probability:

$$Flrn(\varphi\otimes\psi)(a,b) \;=\; \frac{\varphi(a)\cdot\psi(a)}{K\cdot K}.$$

We reason informally, as follows. For arbitrary $\vec{x}\in acc^{-1}(\varphi)$ and $\vec{y}\in acc^{-1}(\psi)$ we need to find the fraction of occurrences (a,b) in $zip(\vec{x},\vec{y})$. The fraction of occurrences of a in $\vec{x}$ is $Flrn(\varphi)(a) = \frac{\varphi(a)}{K}$, and the fraction of occurrences of b in $\vec{y}$ is $Flrn(\psi)(b) = \frac{\psi(b)}{K}$. Hence the fraction of occurrences of (a,b) in $zip(\vec{x},\vec{y})$ is $Flrn(\varphi)(a)\cdot Flrn(\psi)(b) = Flrn(\varphi\otimes\psi)(a,b)$. □

Once we have seen the definition of *mzip*, via 'deconstruction' of multisets into lists, a *zip* operation on lists, and 'reconstruction' to a multiset result, we can try to apply this approach more widely. For instance, instead of using a *zip* on lists we can simply concatenate ($+\!\!+$) the lists — assuming they contain elements from the same set. This yields, like in the definition of *mzip*, a composite channel:

$$\begin{tikzcd} \mathcal{M}[K](X)\times\mathcal{M}[L](X) \arrow[d, "arr\otimes arr"'] & & \mathcal{D}\big(\mathcal{M}[K+L](X)\big) \\ \mathcal{D}\big(X^K\times X^L\big) \arrow[rr, "\mathcal{D}(+\!\!+)", "\cong"'] & & \mathcal{D}\big(X^{K+L}\big) \arrow[u, "\mathcal{D}(acc)"'] \end{tikzcd}$$

It is easy to see that this yields addition of multisets, as a deterministic channel.

We don't get the tensor $\otimes$ of multisets in this way, because there is no tensor of lists, because lists form a non-commutative theory.

X. Multizip and probabilistic operations

Having seen the essentials of the multizip operation *mzip* we proceed to demonstrate how it interacts with the main probabilistic operations of this paper, namely (parallel) multinomials and hypergeometric channels. It turns out that multizip is a mathematically well-behaved operation.

***Theorem** 19:* Multizip commutes with multinomial and hypergeometric channels, as in:

$$\begin{tikzcd} \mathcal{D}(X)\times\mathcal{D}(Y) \arrow[r, "mn{[K]}\otimes mn{[K]}"] \arrow[d, "\otimes"'] & \mathcal{M}[K](X)\times\mathcal{M}[K](Y) \arrow[d, "mzip"] \\ \mathcal{D}(X\times Y) \arrow[r, "mn{[K]}"] & \mathcal{M}[K](X\times Y) \end{tikzcd}$$

$$\begin{tikzcd} \mathcal{M}[N](X)\times\mathcal{M}[N](Y) \arrow[r, "hg{[K]}\otimes hg{[K]}"] \arrow[d, "mzip"'] & \mathcal{M}[K](X)\times\mathcal{M}[K](Y) \arrow[d, "mzip"] \\ \mathcal{M}[N](X\times Y) \arrow[r, "hg{[K]}"] & \mathcal{M}[K](X\times Y) \end{tikzcd}$$

Proof Commutation with the hypergeometric channels follows from Lemma 17 (vi), using Theorem 6. To prove commutation of the multinomial rectangle, we expand the definition of *mzip*, on the right below.

$$\begin{tikzcd} \mathcal{D}(X)\times\mathcal{D}(Y) \arrow[r, "mn{[K]}\otimes mn{[K]}"] \arrow[dr, "iid\otimes iid"'] \arrow[ddd, "\otimes"'] & \mathcal{M}[K](X)\times\mathcal{M}[K](Y) \arrow[d, "arr\otimes arr"] \arrow[ddd, bend left=60, "mzip"] \\ & X^K\times Y^K \arrow[d, "zip"] \\ & (X\times Y)^K \arrow[d, "acc"] \\ \mathcal{D}(X\times Y) \arrow[ur, "iid"] \arrow[r, "mn{[K]}"] & \mathcal{M}[K](X\times Y) \end{tikzcd}$$

The (triangle) subdiagrams at the top and bottom commute by Theorem 4. The middle subdiagram is the first rectangle in Lemma 15. $\square$

Remark 20: In the first diagram in Theorem 19 we see that multinomial channels commute with $\otimes$ and *mzip*. One may wonder about the analogous diagram for tensors only, see below. This diagram does *not* commute.

$$\begin{tikzcd}
\mathcal{D}(X) \times \mathcal{D}(Y) \arrow[r, "mn[K]\otimes mn[L]"] \arrow[d, "\otimes"'] \arrow[dr, phantom, "\neq"] & \mathcal{M}[K](X) \times \mathcal{M}[L](Y) \arrow[d, "\otimes"] \\
\mathcal{D}(X \times Y) \arrow[r, "mn[K\cdot L]"'] & \mathcal{M}[K\cdot L](X \times Y)
\end{tikzcd}$$

Take for instance $X = Y = \{a, b\}$, $K = 1$, $L = 2$ with uniform distribution $\omega = \frac{1}{2}|a\rangle + \frac{1}{2}|b\rangle$. The two legs of the above diagram differ when applied to (ω, ω).

The *pièce de résistance* of this paper is the commutation of multizip *mzip* with the parallel multinomial law *pml*, as expressed by the next two results.

***Lemma** 21:* The following diagram commutes.

$$\begin{tikzcd}
\mathcal{M}[K]\big(\mathcal{D}(X)\big)\times\mathcal{M}[K]\big(\mathcal{D}(Y)\big) \arrow[r, "pml\otimes pml"] \arrow[d, "mzip"'] & \mathcal{M}[K](X)\times\mathcal{M}[K](Y) \arrow[dd, "mzip"] \\
\mathcal{M}[K]\big(\mathcal{D}(X)\times\mathcal{D}(Y)\big) \arrow[d, "\mathcal{M}[K](\otimes)"'] & \\
\mathcal{M}[K]\big(\mathcal{D}(X\times Y)\big) \arrow[r, "pml"] & \mathcal{M}[K](X\times Y)
\end{tikzcd}$$

Proof The result follows from a big diagram chase in which the *mzip* operations on the left and on the right are expanded.

$$\begin{tikzcd}
\mathcal{M}[K]\mathcal{D}(X)\times\mathcal{M}[K]\mathcal{D}(Y) \arrow[rr, "pml\otimes pml"] \arrow[d, "arr\otimes arr"'] \arrow[ddd, bend right=60, "mzip"'] & & \mathcal{M}[K](X)\times\mathcal{M}[K](Y) \arrow[d, "arr\otimes arr"] \arrow[dddd, bend left=60, "mzip"] \\
\mathcal{D}(X)^K\times\mathcal{D}(Y)^K \arrow[rr, "\bigotimes\otimes\bigotimes"] \arrow[d, "zip"'] & & X^K\times Y^K \arrow[d, "zip"] \\
\big(\mathcal{D}(X)\times\mathcal{D}(Y)\big)^K \arrow[r, "\otimes^K"] \arrow[d, "acc"'] & \mathcal{D}(X\times Y)^K \arrow[r, "\bigotimes"] \arrow[ddl, bend left, "acc"] & (X\times Y)^K \arrow[dd, "acc"] \\
\mathcal{M}[K]\big(\mathcal{D}(X)\times\mathcal{D}(Y)\big) \arrow[d, "\mathcal{M}[K](\otimes)"'] & & \\
\mathcal{M}[K]\mathcal{D}(X\times Y) \arrow[rr, "pml"] & & \mathcal{M}[K](X\times Y)
\end{tikzcd}$$

The upper rectangle commutes by Proposition 10 and the middle one by Lemma 15. The lower-left subdiagram commutes by naturality of *acc* and the lower-right one via the third definition of *pml* in (5). $\square$

***Theorem** 22:* The lifted functor $\mathcal{M}[K]\colon \mathcal{K}\ell(\mathcal{D}) \to \mathcal{K}\ell(\mathcal{D})$ from Theorem 14 (i) commutes with multizipping: for channels $f\colon X \rightsquigarrow U$ and $g\colon Y \rightsquigarrow V$ one has:

$$\begin{tikzcd}
\mathcal{M}[K](X) \times \mathcal{M}[K](Y) \arrow[r, "mzip"] \arrow[d, "\mathcal{M}[K](f)\otimes\mathcal{M}[K](g)"'] & \mathcal{M}[K](X \times Y) \arrow[d, "\mathcal{M}[K](f\otimes g)"] \\
\mathcal{M}[K](U) \times \mathcal{M}[K](V) \arrow[r, "mzip"] & \mathcal{M}[K](U \times V)
\end{tikzcd}$$

This channel-naturality of *mzip*, in combination with the unit and associativity of Lemma 17 (ii) and (iii), means that the lifted functor $\mathcal{M}[K]\colon \mathcal{K}\ell(\mathcal{D}) \to \mathcal{K}\ell(\mathcal{D})$ is monoidal via *mzip*.

This result is rather subtle, since f, g are used as channels. So when we write $\mathcal{M}[K](f)$ we mean the application of the lifted functor $\mathcal{M}[K]\colon \mathcal{K}\ell(\mathcal{D}) \to \mathcal{K}\ell(\mathcal{D})$ to f, which is:

$$\mathcal{M}[K](X) \xrightarrow{\mathcal{M}[K](f)} \mathcal{M}[K]\big(\mathcal{D}(Y)\big) \xrightarrow{pml} \mathcal{D}\big(\mathcal{M}[K](Y)\big).$$

Proof We use:

$$\begin{aligned}
& mzip \circ \Big(\mathcal{M}[K](f) \otimes \mathcal{M}[K](g)\Big) \\
&= mzip \circ (pml \otimes pml) \circ \big(\mathcal{M}[K](f) \times \mathcal{M}[K](g)\big) \\
&= pml \circ \mathcal{M}[K](\otimes) \circ mzip \circ \big(\mathcal{M}[K](f) \times \mathcal{M}[K](g)\big) \\
&\qquad \text{by Lemma 21} \\
&= pml \circ \mathcal{M}[K](\otimes) \circ \mathcal{M}[K](f \times g) \circ mzip \\
&\qquad \text{by Lemma 17 (i)} \\
&= pml \circ \mathcal{M}[K](f \otimes g) \circ mzip \\
&= \mathcal{M}[K]\big(f \otimes g\big) \circ mzip.
\end{aligned}$$

$\square$

Remark 23: One may wonder: why bother about these K-sized multisets $\mathcal{M}[K]$ and not work with $\mathcal{M}$ instead? After all, $\mathcal{M}$ also lifts to a monad on $\mathcal{K}\ell(\mathcal{D})$, as we have seen in Theorem 14 (ii). We can use as tensor for $\mathcal{DM}$ the combination of the tensor for $\mathcal{M}$ and for $\mathcal{D}$.

However, an important fact is that the parallel multinomial law *pml* does *not* commute with these two tensors (of multisets and distributions), as in the following diagram.

$$\begin{tikzcd}
\mathcal{M}[K]\big(\mathcal{D}(X)\big)\times\mathcal{M}[L]\big(\mathcal{D}(Y)\big) \arrow[r, "pml\otimes pml"] \arrow[d, "\otimes"'] & \mathcal{M}[K](X)\times\mathcal{M}[K](Y) \arrow[dd, "\otimes"] \\
\mathcal{M}[K\cdot L]\big(\mathcal{D}(X)\times\mathcal{D}(Y)\big) \arrow[d, "\mathcal{M}[K\cdot L](\otimes)"'] \arrow[r, phantom, "\neq"] & \\
\mathcal{M}[K\cdot L]\big(\mathcal{D}(X\times Y)\big) \arrow[r, "pml"] & \mathcal{M}[K\cdot L](X\times Y)
\end{tikzcd}$$

Take for instance $X = \{a, b\}$, $Y = \{0, 1\}$ with $K = 2$, $L = 1$ and with multisets $\varphi = 2\big|\frac{3}{4}|a\rangle + \frac{1}{4}|b\rangle\big\rangle \in \mathcal{M}[K](\mathcal{D}(X))$ and $\psi = 1\big|\frac{2}{3}|0\rangle + \frac{1}{3}|1\rangle\big\rangle \in \mathcal{M}[L](\mathcal{D}(Y))$. With some perseverance one sees that the two legs of the above diagrams give different outcomes on (φ, ψ).

This non-commutation generalises what we have seen in Remark 20 for multinomial channels.

***Theorem** 24:* The lifted functors $\mathcal{M}[K]\colon \mathcal{K}\ell(\mathcal{D}) \to \mathcal{K}\ell(\mathcal{D})$ commute with sums of multisets: for a channel $f\colon X \rightsquigarrow Y$,

$$\begin{tikzcd}
\mathcal{M}[K](X) \times \mathcal{M}[L](Y) \arrow[r, "+"] \arrow[d, "\mathcal{M}[K](f)\otimes\mathcal{M}[L](f)"'] & \mathcal{M}[K+L](X) \arrow[d, "\mathcal{M}[K+L](f)"] \\
\mathcal{M}[K](Y) \times \mathcal{M}[L](Y) \arrow[r, "+"] & \mathcal{M}[K+L](Y)
\end{tikzcd}$$

Proof By Proposition 13, in combination with the standard fact that the multiset functor $\mathcal{M}\colon \mathbf{Sets} \to \mathbf{Sets}$ commutes with sums of multisets. $\square$

Now that we know, via Theorem 14 (i), that K-sized multisets form a (lifted) functor $\mathcal{M}[K]\colon \mathcal{K}\ell(\mathcal{D}) \to \mathcal{K}\ell(\mathcal{D})$, we can ask whether the natural transformations that we have used before are also natural with respect to this lifted functor. This involves naturality with respect to channels instead of with respect to ordinary functions. The proofs of the following result are left to the reader.

***Lemma** 25:* Arrangement and accumulation, and draw-delete are natural transformation in the situations:

$$\begin{tikzcd} \mathcal{K}\ell(\mathcal{D}) \arrow[r, bend left=50, "\mathcal{M}[K]"] \arrow[r, "(-)^K"] \arrow[r, bend right=50, "\mathcal{M}[K]"'] \arrow[r, phantom, shift left=3, "\Downarrow arr"] \arrow[r, phantom, shift right=3, "\Downarrow acc"] & \mathcal{K}\ell(\mathcal{D}) \end{tikzcd} \qquad \begin{tikzcd} \mathcal{K}\ell(\mathcal{D}) \arrow[r, bend left, "\mathcal{M}[K+1]"] \arrow[r, bend right, "\mathcal{M}[K]"'] \arrow[r, phantom, "\Downarrow DD"] & \mathcal{K}\ell(\mathcal{D}) \end{tikzcd}$$

The power functor $(-)^K$ on $\mathbf{Sets}$ lifts to $\mathcal{K}\ell(\mathcal{D})$ via the big tensor $\bigotimes\colon \mathcal{D}(X)^K \to \mathcal{D}(X^K)$. $\square$

The distribution functor $\mathcal{D}$ on $\mathbf{Sets}$ lifts in a standard way to a (monoidal) functor on its Kleisli category $\mathcal{K}\ell(\mathcal{D})$, also written as $\mathcal{D}$. This is used in the following channel-naturality result for multinomial and hypergeometric distributions.

***Theorem** 26:* The multinomial and hypergeometric channels are natural with respect to lifted functors:

$$\begin{tikzcd} \mathcal{K}\ell(\mathcal{D}) \arrow[r, bend left, "\mathcal{D}"] \arrow[r, bend right, "\mathcal{M}[K]"'] \arrow[r, phantom, "\Downarrow mn[K]"] & \mathcal{K}\ell(\mathcal{D}) \end{tikzcd} \qquad \begin{tikzcd} \mathcal{K}\ell(\mathcal{D}) \arrow[r, bend left, "\mathcal{M}[L]"] \arrow[r, bend right, "\mathcal{M}[K]"'] \arrow[r, phantom, "\Downarrow hg[K]"] & \mathcal{K}\ell(\mathcal{D}) \end{tikzcd}$$

where $L \geq K$. These natural transformations are monoidal since they commute with *mzip*, see Theorem 19. $\square$

This result shows that the basic probabilistic operations of drawing from an urn, with or without replacement, come with rich categorical structure.

XI. Application in sampling semantics

The above theory about the distributive law of multisets over distributions can be used to demonstrate the correctness of (basic aspects of) sampling semantics in probabilistic programming languages (see *e.g.* [19], [20], [21] for an overview). We concentrate on parallel and sequential composition and on conditioning.

We formalise sampling via the multinomial channel $mn[K]\colon \mathcal{D}(X) \rightsquigarrow \mathcal{M}[K](X)$. It assigns probabilities to samples of size K, from a distribution $\omega \in \mathcal{D}(X)$. The fact that $Flrn \gg mn[K](\omega) = \omega$, as we have seen in Proposition 3 (ii), shows that learning from these samples yields the original distribution ω — a key correctness property for sampling.

The first diagram in Theorem 19 shows that sampling a product distribution:

$$mn[K](\omega \otimes \rho)$$

is the same as the mzip of parallel (separate) sampling:

$$mzip \gg \Big(mn[K](\omega) \otimes mn[K](\rho)\Big).$$

This gives a basic compositionality result in sampling semantics. At the same time it illustrates the usefulness of *mzip*. Applying frequentist learning to these expressions yields the original product distribution $\omega \otimes \rho$, see also Theorem 18.

We continue with sequential sampling.

> Suppose we have a distribution $\omega \in \mathcal{D}(X)$ and a channel $c\colon X \rightsquigarrow Y$. We wish to sample the distribution $c \gg \omega \in \mathcal{D}(Y)$ obtained by state transformation. A common way to do so is to first sample elements x from ω and then elements y from $c(x)$. These y's are the samples that capture $c \gg \omega$.

We can first sample from ω via $mn[K](\omega) \in \mathcal{M}[K](X)$. Subsequently we can apply the channel c to each of the elements in the sample, via $\mathcal{M}[K](c)\colon \mathcal{M}[K](X) \to \mathcal{M}[K](\mathcal{D}(Y))$. Our next step is to sample from all of these outcomes $c(x)$. But this is precisely what the parallel multinomial law *pml* does. This gives the samples (multisets) of Y that we are after, via the composite of channels:

$$\mathcal{D}(X) \xrightarrow{mn[K]} \mathcal{M}[K](X) \xrightarrow{\mathcal{M}[K](c)} \mathcal{M}[K](\mathcal{D}(Y)) \overset{pml}{\rightsquigarrow} \mathcal{M}[K](Y)$$

The last two arrows describe the lifted functor $\mathcal{M}[K]$ applied to the channel $c\colon X \rightsquigarrow Y$. We claim that the next result is a correctness result for the above sampling method for $c \gg \omega$.

***Proposition** 27:* For a distribution $\omega \in \mathcal{D}(X)$ and channel $c\colon X \rightsquigarrow Y$,

$$Flrn \gg \Big(pml \odot \mathcal{M}[K](c) \odot mn[K]\Big)(\omega) \;=\; c \gg \omega.$$

Proof Because:

$$\begin{aligned}
&\big(Flrn \odot pml \odot \mathcal{M}[K](c) \odot mn[K]\big)(\omega) \\
&= \big(\mathrm{id} \odot Flrn \odot \mathcal{M}[K](c) \odot mn[K]\big)(\omega) \text{ by Theorem 11} \\
&= \big(\mathrm{id} \odot \mathcal{D}(c) \odot Flrn \odot mn[K]\big)(\omega) \quad \text{by naturality} \\
&= \big(\mathrm{id} \odot \mathcal{D}(c)\big)(\omega) \quad \text{by Proposition 3 (ii)} \\
&= \big(\mu \circ \mathcal{D}(c)\big)(\omega) \\
&= c \gg \omega.
\end{aligned}$$

$\square$

Updating (or conditioning) of a distribution is a basic operation in probabilistic programming. We briefly sketch how it fits in the current (sampling) setting. Let $\omega \in \mathcal{D}(X)$ be distribution and $p\colon X \to [0,1]$ be a (fuzzy) predicate. We write $\omega \models p := \sum_x \omega(x) \cdot p(x)$ for the validity (expected value) of p in ω. If this validity is non-zero, we can define the updated distribution $\omega|_p \in \mathcal{D}(X)$ as the normalised product:

$$\omega|_p(x) := \frac{\omega(x) \cdot p(x)}{\omega \models p}.$$

See *e.g.* [22], [23], [17] for more information.

The question arises: how to sample from an updated state $\omega|_p$? This involves the interaction of updating with (parallel) multinomials. It is addressed below. There we use the free extension $\overline{p}\colon \mathcal{M}(X) \to [0,1]$ of a predicate $p\colon X \to [0,1]$ to multisets, via conjunction &, that is, via pointwise multiplication:

$$\overline{p}(\varphi) := \prod_{x \in X} \underbrace{p(x) \cdot \ldots \cdot p(x)}_{\varphi(x) \text{ times}}.$$

Point (ii) below shows that sampling of an updated distribution $\omega|_p$ can be done compositionally, by first sampling ω and then updating the samples with the free extension $\overline{p}$. If p is a sharp (non-fuzzy) predicate, this means throwing out the sampled multisets where p does not hold for all elements, followed by re-normalisation.

***Theorem** 28:* For a distribution $\omega \in \mathcal{D}(X)$, a predicate $p\colon X \to [0,1]$, and a number $K \in \mathbb{N}$, one has:

i) $\mathit{mn}[K](\omega) \models \overline{p} \;=\; \big(\omega \models p\big)^{K}$;
ii) $\mathit{mn}[K](\omega)\big|_{\overline{p}} = \mathit{mn}[K](\omega|_p)$;
iii) Similarly, for the parallel multinomial law *pml*,

$$\mathit{pml}\big(\textstyle\sum_i n_i|\omega_i\rangle\big) \models \overline{p} \;=\; \prod_i \big(\omega_i \models p\big)^{n_i}.$$

iv) And:

$$\mathit{pml}\big(\textstyle\sum_i n_i|\omega_i\rangle\big)\big|_{\overline{p}} \;=\; \mathit{pml}\big(\textstyle\sum_i n_i|\,\omega_i|_p\rangle\big).$$

An example that combines multinomials and updating occurs in [24, §6.4], but without a general formulation, as given above, in point (ii).

Proof The first point follows from the Multinomial Theorem. Then, for $\varphi \in \mathcal{M}[K](X)$,

$$\begin{aligned}
\mathit{mn}[K](\omega)\big|_{\overline{p}}(\varphi) &= \frac{\mathit{mn}[K](\omega)(\varphi)\cdot\overline{p}(\varphi)}{\mathit{mn}[K](\omega)\models\overline{p}} \\
&\overset{\text{(i)}}{=} (\varphi)\cdot\frac{\prod_x \omega(x)^{\varphi(x)}\cdot\prod_x p(x)^{\varphi(x)}}{(\omega\models p)^K} \\
&= (\varphi)\cdot\frac{\prod_x \big(\omega(x)\cdot p(x)\big)^{\varphi(x)}}{(\omega\models p)^K} \\
&= (\varphi)\cdot\prod_x\left(\frac{\omega(x)\cdot p(x)}{\omega\models p}\right)^{\varphi(x)} \\
&= (\varphi)\cdot\prod_x \omega|_p(x)^{\varphi(x)} \\
&= \mathit{mn}[K](\omega|_p)(\varphi).
\end{aligned}$$

For point (iii) we use the second formulation (4) of *pml* in:

$$\begin{aligned}
&\mathit{pml}\big(\sum_i n_i|\omega_i\rangle\big)\models\overline{p} \\
&= \sum_{i,\varphi_i\in\mathcal{M}[n_i](X)} \Big(\textstyle\prod_i \mathit{mn}[n_i](\omega_i)(\varphi_i)\Big)\cdot\overline{p}\big(\textstyle\sum_i\varphi_i\big) \\
&= \sum_{i,\varphi_i\in\mathcal{M}[n_i](X)} \Big(\textstyle\prod_i \mathit{mn}[n_i](\omega_i)(\varphi_i)\Big)\cdot\Big(\textstyle\prod_i\overline{p}(\varphi_i)\Big) \\
&\qquad\text{since } \overline{p} \text{ is by definition a map of monoids} \\
&= \sum_{i,\varphi_i\in\mathcal{M}[n_i](X)} \textstyle\prod_i \mathit{mn}[n_i](\omega_i)(\varphi_i)\cdot\overline{p}(\varphi_i) \\
&= \prod_i \sum_{\varphi_i\in\mathcal{M}[n_i](X)} \mathit{mn}[n_i](\omega_i)(\varphi_i)\cdot\overline{p}(\varphi_i) \\
&= \prod_i \mathit{mn}[n_i](\omega_i)\models\overline{p} \\
&\overset{\text{(i)}}{=} \prod_i \big(\omega_i\models p\big)^{n_i}.
\end{aligned}$$

This formula for validity is used to prove the final equation about updating. Let $K = \sum_i n_i$ in:

$$\begin{aligned}
&\mathit{pml}\big(\textstyle\sum_i n_i|\omega_i\rangle\big)\big|_{\overline{p}} \\
&= \sum_{\varphi\in\mathcal{M}[K](X)} \frac{\mathit{pml}\big(\sum_i n_i|\omega_i\rangle\big)(\varphi)\cdot\overline{p}(\varphi)}{\mathit{pml}\big(\sum_i n_i|\omega_i\rangle\big)\models\overline{p}}\,\big|\varphi\big\rangle \\
&= \sum_{i,\varphi_i\in\mathcal{M}[n_i](X)} \frac{\prod_i \mathit{mn}[n_i](\omega_i)(\varphi_i)\cdot\overline{p}(\varphi_i)}{\prod_i(\omega_i\models p)^{n_i}}\,\big|\textstyle\sum_i\varphi_i\big\rangle \\
&= \sum_{i,\varphi_i\in\mathcal{M}[n_i](X)} \prod_i\,(\varphi_i)\cdot\frac{\prod_x(\omega(x)\cdot p(x))^{\varphi_i(x)}}{(\omega_i\models p)^{n_i}}\,\big|\textstyle\sum_i\varphi_i\big\rangle \\
&= \sum_{i,\varphi_i\in\mathcal{M}[n_i](X)} \prod_i(\varphi_i)\cdot\prod_x\left(\frac{\omega(x)\cdot p(x)}{\omega_i\models p}\right)^{\varphi_i(x)}\big|\textstyle\sum_i\varphi_i\big\rangle \\
&= \sum_{i,\varphi_i\in\mathcal{M}[n_i](X)} \prod_i\,(\varphi_i)\cdot\prod_x \omega_i|_p(x)^{\varphi_i(x)}\,\big|\textstyle\sum_i\varphi_i\big\rangle \\
&= \sum_{i,\varphi_i\in\mathcal{M}[n_i](X)} \prod_i\,\mathit{mn}[n_i](\omega_i|_p)(\varphi_i)\,\big|\textstyle\sum_i\varphi_i\big\rangle \\
&= \mathit{pml}\big(\textstyle\sum_i n_i|\,\omega_i|_p\rangle\big). \qquad\square
\end{aligned}$$

XII. Final remarks

This paper has given an exposition on the distributivity of multisets over probability distributions, via what has been named the parallel multinomial law. It demonstrates (once more) that basic results in probability theory find a natural formulation in the categorical framework of channels / Kleisli-maps. Highlights of this paper are the commutation of the parallel multinomial law with hypergeometric distributions (Theorem 12), and the multizip operation that interacts smoothly with frequentist learning, (parallel) multinomials and hypergeometrics, see Theorems 18 and 19 and Lemma 21. The results make the K-sized multiset functor on $\mathcal{K}\ell(\mathcal{D})$ monoidal and also make various natural transformations monoidal. This may enrich axiomatic approaches to probability like in [25]. The channel-based approach could also be of more practical interest for popular probabilistic computation tools like SPSS or R.

The probabilistic channels in this paper have been implemented in Python (via the EfProb library [26]). This greatly helped in grasping what's going on, and for testing results (with random inputs). The commuting diagrams (of channels) in this paper may look simple, but the computations involved are quite complicated and quickly consume many resources (and then run out of memory or take a long time).

This work offers ample room for extension, for instance with Pólya channels (like in [16]) or to continuous probability (like in [15]).

References

[1] E. Moggi, "Notions of computation and monads," *Information & Computation*, vol. 93(1), pp. 55–92, 1991.

[2] J. Beck, "Distributive laws," in *Seminar on Triples and Categorical Homology Theory*, ser. Lect. Notes Math., B. Eckman, Ed., no. 80. Springer, Berlin, 1969, pp. 119–140.

[3] B. Klin and J. Salamanca, "Iterated covariant powerset is not a monad," in *Math. Found. of Programming Semantics*, ser. Elect. Notes in Theor. Comp. Sci., S. Staton, Ed., no. 341. Elsevier, Amsterdam, 2018, pp. 261–276.

[4] M. Zwart, "On the non-compositionality of monads via distributive laws," Ph.D. dissertation, Dep. of Comp. Science, Univ. of Oxford, 2020.

[5] M. Zwart and D. Marsden, "No-go theorems for distributive laws," in *Logic in Computer Science*, IEEE. Computer Science Press, 2019, pp. 1–13.

[6] D. Turi and G. Plotkin, "Towards a mathematical operational semantics," in *Logic in Computer Science*, IEEE. Computer Science Press, 1997, pp. 280–291.

[7] B. Jacobs, *Introduction to Coalgebra. Towards Mathematics of States and Observations*, ser. Tracts in Theor. Comp. Sci. Cambridge Univ. Press, 2016, no. 59.

[8] F. Bartels, "Generalised coinduction," *Math. Struct. in Comp. Sci.*, vol. 13(2), pp. 321–348, 2003.

[9] F. Bonchi, D. Petrişan, D. Pous, and J. Rot, "A general account of coinduction up-to," *Acta Informatica*, vol. 54(2), pp. 127–190, 2017.

[10] D. Varacca, "Probability, nondeterminism and concurrency: Two denotational models for probabilistic computation," Ph.D. dissertation, Univ. Aarhus, 2003, bRICS Dissertation Series, DS-03-14.

[11] D. Varacca and G. Winskel, "Distributing probability over nondeterminism," *Math. Struct. in Comp. Sci.*, vol. 16, pp. 87–113, 2006.

[12] A. Goy and D. Petrişan, "Combining probabilistic and non-deterministic choice via weak distributive laws," in *Logic in Computer Science*, H. Hermanns, L. Zhang, N. Kobayashi, and D. Miller, Eds., ACM/IEEE. Computer Science Press, 2020, pp. 454–464.

[13] K. Keimel and G. Plotkin, "Mixed powerdomains for probability and nondeterminism," *Logical Methods in Comp. Sci.*, vol. 13(1), 2017, see https://lmcs.episciences.org/2665.

[14] F. Dahlqvist, L. Parlant, and A. Silva, "Layer by layer – combining monads," in *Theoretical Aspects of Computing*, ser. Lect. Notes Comp. Sci., B. Fischer and T. Uustalu, Eds., no. 11187. Springer, Berlin, 2018, pp. 153–172.

[15] S. Dash and S. Staton, "A monad for probabilistic point processes," in *Applied Category Theory Conference*, ser. Elect. Proc. in Theor. Comp. Sci., D. Spivak and J. Vicary, Eds., 2020, https://cgi.cse.unsw.edu.au/~eptcs/paper.cgi?ACT2020:61.

[16] B. Jacobs and S. Staton, "De Finetti's construction as a categorical limit," in *Coalgebraic Methods in Computer Science (CMCS 2020)*, ser. Lect. Notes Comp. Sci., D. Petrişan and J. Rot, Eds., no. 12094. Springer, Berlin, 2020, pp. 90–111.

[17] B. Jacobs, "Multisets and distributions, in drawing and learning," in *Samson Abramsky on Logic and Structure in Computer Science and Beyond*, A. Palmigiano and M. Sadrzadeh, Eds. Springer, 2021, to appear.

[18] ——, "From probability monads to commutative effectuses," *Journ. of Logical and Algebraic Methods in Programming*, vol. 94, pp. 200–237, 2018.

[19] A. Gordon, T. Henzinger, A. Nori, and S. Rajamani, "Probabilistic programming," in *Int. Conf. on Software Engineering*, 2014.

[20] A. Ścibior, Z. Ghahramani, and A. Gordon, "Practical probabilistic programming with monads," in *Proc. 2015 ACM SIGPLAN Symp. on Haskell*. ACM, 2015, pp. 165–176.

[21] G. Barthe, J.-P. Katoen, and A. Silva, Eds., *Foundations of Probabilistic Programming*. Cambridge Univ. Press, 2020.

[22] B. Jacobs, "Learning along a channel: the Expectation part of Expectation-Maximisation," in *Math. Found. of Programming Semantics*, ser. Elect. Notes in Theor. Comp. Sci., B. König, Ed., no. 347. Elsevier, Amsterdam, 2019, pp. 143–160.

[23] ——, "The mathematics of changing one's mind, via Jeffrey's or via Pearl's update rule," *Journ. of Artif. Intelligence Research*, vol. 65, pp. 783–806, 2019.

[24] S. Ross, *A first course in probability*, 10^{th} ed. Pearson Education, 2018.

[25] T. Fritz, "A synthetic approach to Markov kernels, conditional independence, and theorems on sufficient statistics," *Advances in Math.*, vol. 370, p. 107239, 2020.

[26] K. Cho and B. Jacobs, "The EfProb library for probabilistic calculations," in *Conference on Algebra and Coalgebra in Computer Science (CALCO 2017)*, ser. LIPIcs, F. Bonchi and B. König, Eds., vol. 72. Schloss Dagstuhl, 2017.