Electronic Notes in Theoretical Informatics and Computer Science

HTTPS://ENTICS.EPISCIENCES.ORG

Volume 5
Proceedings of
MFPS 2025

A Fresh Look at Bivariate Binomial Distributions

Bart Jacobs[1]

iHub
Radboud University
Nijmegen, The Netherlands

Abstract

Binomial distributions capture the probabilities of 'heads' outcomes when a (biased) coin is tossed multiple times. The coin may be identified with a distribution on the two-element set $\{0, 1\}$, where the 1 outcome corresponds to 'head'. One can also toss two separate coins, with different biases, in parallel and record the outcomes. This paper investigates a slightly different 'bivariate' binomial distribution, where the two coins are dependent (also called: entangled, or entwined): the two-coin is a distribution on the product $\{0, 1\} \times \{0, 1\}$. This bivariate binomial exists in the literature, with complicated formulations. Here we use the language of category theory to give a new succint formulation. This paper investigates, also in categorically inspired form, basic properties of these bivariate distributions, including their mean, variance and covariance, and their behaviour under convolution and under updating, in Laplace's rule of succession. Furthermore, it is shown how Expectation Maximisation works for these bivariate binomials, so that mixtures of bivariate binomials can be recognised in data. This paper concentrates on the bivariate case, but the binomial distributions may be generalised to the multivariate case, with multiple dimensions, in a straightforward manner.

Keywords: Probability theory, bivariate binomial distribution, convolution Laplace's rule of Succession, Expectation Maximisation

1 Introduction

A very basic physical model of a random phenomenon is the tossing (or 'flipping' or 'throwing') of a coin. The possible outcomes are 'head' and 'tail', or '1' and '0' in this paper. The coin may have a bias $r \in [0, 1]$, which we incorporate in a 'flip' distribution on the set of outcomes $\{0, 1\}$. Using 'ket' notation $|-\rangle$ it is written as:

$$flip(r) = r\big|1\big\rangle + (1-r)\big|0\big\rangle. \tag{1}$$

This says that the probability of outcome 1 is r and the probability of outcome 0 is $1 - r$.

Repeating such coin flipping $K \in \mathbb{N}$ times, still with a bias $r \in [0, 1]$, yields the binomial distribution on the set $\{0, 1, \ldots, K\}$ where a number n in the set gets the probability that n out of these K flips are head. Again using ket notation this binomial distribution is written as:

$$bn[K](r) = \sum_{0 \le n \le K} \binom{K}{n} \cdot r^n \cdot (1-r)^{K-n} \big|n\big\rangle. \tag{2}$$

[1] Email: bart@cs.ru.nl

Rec'd Apr 3,2025; Pub'd Dec 15,2025
10.46298/entics.17042

Proceedings Available Online at
ENTICS MFPS 41 Proceedings

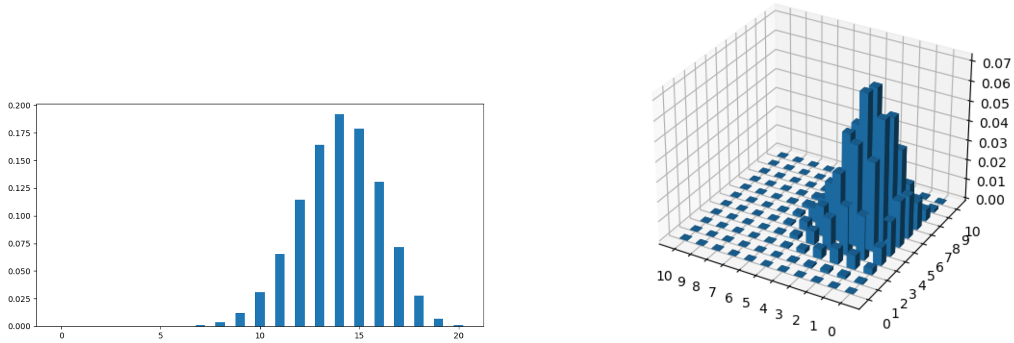
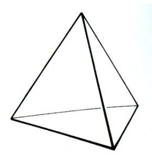


Fig. 1. Plots of an ordinary binomial distribution (20 tosses with bias $\frac{7}{10}$) on the left, and a bivariate binomial distribution on the right, with 10 tosses of a two-coin $\frac{3}{8}\big|0,0\big\rangle + \frac{5}{12}\big|0,1\big\rangle + \frac{1}{12}\big|1,0\big\rangle + \frac{1}{8}\big|1,1\big\rangle$.

The probabilities add up to one by the Binomial Theorem. On the left in Figure 1 one sees the plot of such a binomial distribution with $K = 20$ tosses of a coin with bias $r = \frac{7}{10}$. Via the bias this typical (discrete) bell shape can be shifted to the left or right.

On the right in Figure 1 we see a plot of a bivariate binomial distribution. It has a bell shape in two dimensions. It does not arise from a single coin distribution *flip*(r) on $\{0,1\}$ but from what we call a two-coin distribution on $\{0,1\} \times \{0,1\}$. Such a two-coin may arise as a tensor product *flip*$(r) \otimes$ *flip*(s) of two separate coins, but in general it is more than such a product. It may incorporate the important feature of 'joint' distributions on product spaces, namely that the whole distribution does not coincide with the product of its marginals. In that case the joint distribution is called dependent, entangled, or entwined.



A possible physical model of such a two-coin distribution on $\{0,1\} \times \{0,1\} = \{(0,0),(0,1),(1,0),(1,1)\}$ is given by a tetrahedron, as on the side, also known as a triangular pyramid. Its four faces can be labeled 00, 01, 10 and 11. Such a tetrahedron can be tossed, where the outcome is the label on the face on which the tetrahedron lands. One could imagine that the weight distribution inside the tetrahedron is such that the outcomes are determined by a general two-coin distribution:

$$r_1\big|0,0\big\rangle + r_2\big|0,1\big\rangle + r_3\big|1,0\big\rangle + r_4\big|1,1\big\rangle \qquad \text{where} \qquad r_1 + r_2 + r_3 + r_4 = 1.$$

This paper formulates this bivariate binomial distribution, for such a two-coin, using some basic category theory, namely the functoriality of sending a set X to the set $\mathcal{D}(X)$ of distributions on X. For a number $K \in \mathbb{N}$ of tosses, it produces a distribution on the product set $\{0,1,\ldots,K\} \times \{0,1,\ldots,K\}$ — as on the right in Figure 1 where $K = 10$. This bivariate binomial is introduced as a (functorial) pushforward of a multinomial distribution along a 'marginal heads' function, see Section 3 for the mathematical details. Intuitively, one throws the tetrahedron K-many times and one separately counts the number of 1's in the first a and in the second b position in the label ab written on the faces that the tetrahedron lands on (in K-many tosses). We focus on a 2-dimensional extension of the binomial distribution. One can also extend to N-dimensional form, via a 'hedron' object with 2^N sides and labels from $\{0,1\}^N$. It gives a distribution on $\{0,1,\ldots,K\}^N$ when one counts the 1's in each of the N positions $1,\ldots,N$ of the (down-facing) labels, after multiple tosses.

The (ordinary) binomial distribution (2) can be seen as discrete version of the widely used Gaussian (normal) distribution. Gaussians often occur in N-dimensional form. The N-dimensional discrete version, however, is less familiar. In Section 6 we show how to recognise (mixtures of) bivariate binomials in (multisets of) data items. This technique may have practical value and could increase the popularity of bivariate binomial distributions.

This paper starts by describing the relevant background information, on multisets and distributions, tensor products and convolutions of distributions, channels and their inversions, and on binomial and multinomial distributions. Section 3 then introduces bivariate binomial distributions, via functoriality, and establishes some basic properties with respect to marginalisation and also closure under convolution. Subsequently, Section 4 determines the mean, variance and covariance of bivariate binomials. It turns out that they can be expressed, respectively, as K times the mean, variance and covariance of the underlying two-coin distribution, where K is the number of tosses. Section 5 briefly discusses Laplace's rule of succession, which is in generally used to describe the expected outcome after an update. Here it is formulated in general terms, as a mean of a dagger (channel), and more specifically for bivariate distributions, with Dirichlet and Poisson distributions as priors. Section 6 gives an illustration of how bivariate binomials can be recognised in suitable data multisets.

2 Preliminaries

2.1 Multisets, distributions and random variables

A *multiset* is like a subset, except that its elements may occur multiple times. Alternatively, a multiset is like a list in which the order of the elements does not matter. For instance, an urn filled with three red, two green and five blue balls forms a multiset. We write it using ket notation as $3|R\rangle + 2|G\rangle + 5|B\rangle$, where the ket brackets $|-\rangle$ serve to separate the elements R, G, B for the three colours, from their multiplicities $3, 2, 5$. Also a draw of multiple balls from such an urn forms a multiset. In general, multiple data items, as used for learning (see Section 6) are naturally described as a multiset.

For an arbitrary set X we shall write $\mathcal{M}(X)$ for the set of (finite) multisets with elements from X. A multiset $\varphi \in \mathcal{M}(X)$ may be described in two equivalent ways, namely via kets as a formal sum $\varphi = n_1|x_1\rangle + \cdots + n_K|x_K\rangle = \sum_i n_i|x_i\rangle$, with elements $x_i \in X$ and associated multiplicities $n_i \in \mathbb{N}$. Alternatively, we may describe a multiset $\varphi \in \mathcal{M}(X)$ as a function $\varphi\colon X \to \mathbb{N}$ with finite support, where $\mathrm{supp}(\varphi) = \{x \in X \mid \varphi(x) \neq 0\}$. We can combine these notations and write $\varphi = \sum_x \varphi(x)|x\rangle$, where the summands with multiplicity zero are omitted.

The size $\|\varphi\| \in \mathbb{N}$ of a multiset $\varphi \in \mathcal{M}(X)$ is the total number of its elements, including multiplicities. Thus, $\|\varphi\| = \sum_x \varphi(x)$. The abovementioned urn $\upsilon = |R\rangle + 2|G\rangle + 5|B\rangle$ has size $\|\upsilon\| = 10$. There is one multiset in $\mathcal{M}(X)$ of size zero, namely the empty multiset, written as $\mathbf{0} \in \mathcal{M}(X)$. As a function, it satisfies $\mathbf{0}(x) = 0$, for each $x \in X$, giving x multiplicity 0. We write $\mathcal{M}[K](X) \subseteq \mathcal{M}(X)$ for the subset of multisets of size $K \in \mathbb{N}$.

The mapping $X \mapsto \mathcal{M}(X)$ forms a monad on the category of sets. We only use functoriality: a function $f\colon X \to Y$ yields another function $\mathcal{M}(f)\colon \mathcal{M}(X) \to \mathcal{M}(Y)$ via the following two equivalent definitions, in ket form on the left, and in function form on the right.

$$\mathcal{M}(f)\Big(\sum_i n_i\big|x_i\big\rangle\Big) = \sum_i n_i\big|f(x_i)\big\rangle \qquad\qquad \mathcal{M}(f)(\varphi)(y) = \sum_{x \in f^{-1}(y)} \varphi(x). \tag{3}$$

We shall especially use functoriality for a projection function $\pi_1\colon X \times Y \to X$. Then $\mathcal{M}(\pi_1)\colon \mathcal{M}(X \times Y) \to \mathcal{M}(X)$ performs marginalisation.

One more observation is that $\mathcal{M}(X)$ is the free commutative monoid on the set X. The monoid structure is given by pointwise addition of multisets: $\big(\varphi + \psi\big)(x) = \varphi(x) + \psi(x)$, with the everywhere zero multset $\mathbf{0} \in \mathcal{M}(X)$ as neutral element. Functions $\mathcal{M}(f)$ preserve this monoid structure.

We turn to distributions: a (finite discrete probability) distribution $r_1|x_1\rangle + \cdots + r_K|x_K\rangle$ is written as a multiset, except that the multiplicities r_i are now probabilities from the unit interval $[0,1]$ that add up to one: $\sum_i r_i = 1$. We shall write $\mathcal{D}(X)$ for the set of distributions with elements from X. A distribution $\omega \in \mathcal{D}(X)$ may be written in ket form, as we just did, but also as a function $\omega\colon X \to [0,1]$ with finite support $\mathrm{supp}(\omega) = \{x \in X \mid \omega(x) \neq 0\}$ and with $\sum_x \omega(x) = 1$. When X is a finite set, we say that $\omega \in \mathcal{D}(X)$ has full support when $\mathrm{supp}(\omega) = X$, that is, when $\omega(x) > 0$ for each $x \in X$. The operation $\mathcal{D}$ is also a monad on the category of sets. Functoriality works as for $\mathcal{M}$, in (3), in particular with projections $\mathcal{D}(\pi_i)$ for marginalisation.

Each non-empty multiset $\varphi \in \mathcal{M}(X)$ can be turned into a distribution $\mathit{Flrn}(\varphi) \in \mathcal{D}(X)$, where *Flrn* stands for frequentist learning. This involves learning by counting:

$$Flrn(\varphi) \coloneqq \sum_{x\in \mathrm{supp}(\varphi)} \frac{\varphi(x)}{\|\varphi\|} \,\big|\, x\,\big\rangle. \tag{4}$$

For instance, for the above urn $\upsilon = 3|R\rangle + 2|G\rangle + 5|B\rangle$ we get $Flrn(\upsilon) = \frac{3}{10}|R\rangle + \frac{1}{5}|G\rangle + \frac{1}{2}|B\rangle$. It captures the probabilities of drawing a ball of a particular colour from the urn.

A random variable is a pair (ω, p) where $\omega \in \mathcal{D}(X)$ is a distribution and $p\colon X \to \mathbb{R}$ an 'observable'. The validity (or expected value) is:

$$\omega \models p \;\coloneqq\; \sum_{x\in X} \omega(x)\cdot p(x). \tag{5}$$

2.2 Products and convolutions of distributions

For two distributions $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$ one can form the (parallel) product distribution $\omega\otimes\rho \in \mathcal{D}(X\times Y)$, via $\big(\omega\otimes\rho\big)(x,y) = \omega(x)\cdot\rho(y)$. Marginalisation of such a product yields the components:

$$\mathcal{D}(\pi_1)\big(\omega\otimes\rho\big) = \omega \qquad\text{and}\qquad \mathcal{D}(\pi_2)\big(\omega\otimes\rho\big) = \rho.$$

In general, for a 'joint' distribution $\tau \in \mathcal{D}(X\times Y)$, on a product space, one does *not* have $\tau = \mathcal{D}(\pi_1)(\tau)\otimes \mathcal{D}(\pi_2)(\tau)$. If this equation does hold, τ is called independent, non-entangled, or non-entwined.

We write $\mathbf{2} = \{0,1\}$. A distribution $\tau \in \mathcal{D}(\mathbf{2}\times\mathbf{2})$, say $\tau = r_1|0,0\rangle + r_2|0,1\rangle + r_3|1,0\rangle + r_4|1,1\rangle$ with $r_1+r_2+r_3+r_4 = 1$, is non-entwined if and only if $r_1\cdot r_4 = r_2\cdot r_3$.

Tensors of distributions are used for a special sum of distributions, called convolution. Let $M = (M,+,0)$ be a commutative monoid and let $\omega,\rho \in \mathcal{D}(M)$ be two distributions on the underlying set M. Then we can form their convolution $\omega+\rho \in \mathcal{D}(M)$ as, via functoriality of $\mathcal{D}$, applied to the addition $+\colon M\times M \to M$, as in:

$$\omega+\rho \coloneqq \mathcal{D}(+)\Big(\omega\otimes\rho\Big) = \sum_{x,y\in M} \omega(x)\cdot\rho(y)\,\big|\, x+y\,\big\rangle. \tag{6}$$

This turns $\mathcal{D}(M)$ into a commutative monoid, with $1|0\rangle \in \mathcal{D}(M)$ as unit, for $0 \in M$. It is not hard to see that if $f\colon M \to M'$ is a homomorphism of monoids, then so is $\mathcal{D}(f)\colon \mathcal{D}(M) \to \mathcal{D}(M')$.

In the sequel we shall encounter distributions on $\{0,1\}$ or more generally on $\{0,\ldots,K\}$. By considering these sets as subsets of $\mathbb{N}$, with its additive commutative monoid structure, we can apply the convolution definition (6) to such distributions. Commutative monoids are closed under products $\times$, so that $\mathbb{N}^2 = \mathbb{N}\times\mathbb{N}$ is also a commutative monoid, via component-wise sums and zeros. Also, as we have seen, the set $\mathcal{M}(X)$ of multisets over a set X is a commutative monoid, so that convolution can be used for distributions on multisets. This applies to multinomial distributions, see Subsection 2.4.

2.3 Channels and their inversions / daggers

A channel is a function of the form $c\colon X \to \mathcal{D}(Y)$. It maps an element $x \in X$ to a distribution $c(x)$ on Y. As such it occurs often as conditional probability, written as $\Pr(y|x)$. Channels are well-behaved probabilistic computations that can be composed sequentially and in parallel. They form the morphisms in the symmetric monoidal Kleisli category $\mathcal{K}\ell(\mathcal{D})$ of the distribution monad $\mathcal{D}$. In fact we simplify the notation and use a special arrow $\rightsquigarrow$ for channels. Thus we simply write $c\colon X \rightsquigarrow Y$ for the more cumbersome $c\colon X \to \mathcal{D}(Y)$.

A basic operation is pushforward along a channel. Given a channel $c\colon X \rightsquigarrow Y$ and a distribution $\omega \in \mathcal{D}(X)$ on its domain X, we can form a new distribution $c_*(\omega) \in \mathcal{D}(Y)$ on its codomain Y, namely:

$$c_*(\omega) \coloneqq \sum_{x\in X} \omega(x)\cdot c(x) = \sum_{y\in Y}\left(\sum_{x\in X}\omega(x)\cdot c(x)(y)\right)\big|\, y\,\big\rangle. \tag{7}$$

Under suitable circumstances a channel can be turned around, via an operation called Bayesian inversion, see *e.g.* [4,3,5]. This involves turning a conditional probability $\Pr(y|x)$ into $\Pr(x|y)$. We present it in channel form: so let $c\colon X \rightsquigarrow Y$ be a channel from X to Y, where the set Y is finite. We will turn it into a channel $Y \rightsquigarrow X$, in the opposite direction. This inversion requires a 'prior' distribution $\omega \in \mathcal{D}(X)$ such that the pushforward $c_*(\omega) \in \mathcal{D}(Y)$ has full support. Then we can define a new channel, written as $c^\dagger_\omega\colon Y \rightsquigarrow X$ via:

$$c^\dagger_\omega(y) \coloneqq \sum_{x\in X} \frac{\omega(x)\cdot c(x)(y)}{c_*(\omega)(y)}\,\big|\, x\,\big\rangle. \tag{8}$$

The full support requirement is needed to prevent division-by-zero errors.

2.4 Binomial and multinomial distributions

We have seen the binomial distribution $bn[K](r) \in \mathcal{D}\big(\{0,\ldots,K\}\big)$ in Equation (2), for K-many tosses of a coin with bias $r \in [0,1]$. A single coin toss is given by the flip distribution $\mathit{flip}(r)$ in (1).

Lemma 2.1 (i) *The binomial distribution is an iterated convolution of the flip distribution:*

$$bn[K](r) = K\cdot \mathit{flip}(r) = \underbrace{\mathit{flip}(r)+\cdots+\mathit{flip}(r)}_{K\ \textit{times}}.$$

Hence, binomial distributions are closed under convolution: $bn[K](r) + bn[L](r) = bn[K+L](r)$.

(ii) *The mean and variance of the binomial distribution are given by:*

$$\mathrm{mean}\Big(bn[K](r)\Big) = K\cdot r \qquad \textit{and} \qquad \mathrm{var}\Big(bn[K](r)\Big) = K\cdot(K-1)\cdot r. \qquad \square$$

We turn to the multinomial distribution. We can view a distribution $\omega \in \mathcal{D}(X)$ as an abstract urn, where X is the set of colours and $\omega(x)\in[0,1]$ gives the probability of drawing a ball of colour $x \in X$. We are interested in drawing multiple balls, in 'multinomial mode', with replacement. A draw of K-many balls from the urn ω can be identified with a K-sized multiset $\varphi \in \mathcal{M}[K](X)$. The multinomial distribution $mn[K](\omega) \in \mathcal{D}\big(\mathcal{M}[K](X)\big)$ assigns probabilites to such draws of size K. It is defined as:

$$mn[K](\omega) \coloneqq \sum_{\varphi\in\mathcal{M}[K](X)} (\varphi)\cdot \prod_{x\in\mathrm{supp}(\omega)} \omega(x)^{\varphi(x)}\,\big|\,\varphi\,\big\rangle \quad \text{where} \quad (\varphi) \coloneqq \frac{\|\varphi\|!}{\prod_x \varphi(x)!}. \tag{9}$$

The precise form of this multinomial distribution does not matter so much at this stage; interested readers may consult [9,11]. We make the following facts explicit. The first two items are standard. For details about the third item, see *e.g.* [13, Prop. 7.2] or [14].

Lemma 2.2 *Consider the multinomial distribution* (9) *for* $\omega \in \mathcal{D}(X)$ *and* $K \in \mathbb{N}$.

(i) *The binomial and multinomial distribution can be identified when* $X = \mathbf{2} = \{0,1\}$*, via the isomorphisms flip*: $[0,1] \overset{\cong}{\Rightarrow} \mathcal{D}(\mathbf{2})$ *and* $\{0,\ldots,K\} \overset{\cong}{\Rightarrow} \mathcal{M}[K](\mathbf{2})$*, via* $n \mapsto n|1\rangle + (K-n)|0\rangle$*, in:*

$$\begin{array}{ccc} [0,1] & \xrightarrow{bn[K]} & \mathcal{D}\big(\{0,\ldots,K\}\big) \\ \wr\| & & \wr\| \\ \mathcal{D}(\mathbf{2}) & \xrightarrow{mn[K]} & \mathcal{D}\big(\mathcal{M}[K](\mathbf{2})\big) \end{array}$$

(ii) *Multinomials are closed under convolution:* $mn[K](\omega) + mn[L](\omega) = mn[K+L](\omega)$.

(iii) *For elements* $y, z \in X$ *with* $y \neq z$,

$$\begin{aligned} \sum_{\varphi\in\mathcal{M}[K](X)} mn[K](\omega)(\varphi)\cdot\varphi(y) &= K\cdot\omega(y) \\ \sum_{\varphi\in\mathcal{M}[K](X)} mn[K](\omega)(\varphi)\cdot\varphi(y)\cdot\varphi(z) &= K\cdot(K-1)\cdot\omega(y)\cdot\omega(z). \\ \sum_{\varphi\in\mathcal{M}[K](X)} mn[K](\omega)(\varphi)\cdot\varphi(y)^2 &= K\cdot(K-1)\cdot\omega(y)^2 \;+\; K\cdot\omega(y). \end{aligned}$$

□

3 Introducing bivariate distributions

This section introduces the main distribution of this paper, namely the bivariate binomial distribution. It also describes some of its most basic properties. Our definition exploits functoriality of distributions $\mathcal{D}$, via a marginal heads function that is defined first. This definition is compared to explicit formulations in the literature in Remark 3.5 below.

Definition 3.1 Fix a number $K \in \mathbb{N}$ and consider the following 'marginal heads' function:

$$\mathcal{M}[K]\big(\mathbf{2}\times\mathbf{2}\big) \xrightarrow{heads} \{0,1,\ldots,K\}\times\{0,1,\ldots,K\},$$

defined as pair of marginals evaluated at $1 \in \mathbf{2}$, in:

$$heads(\varphi) \coloneqq \big(\,\varphi(1,0)+\varphi(1,1),\, \varphi(0,1)+\varphi(1,1)\,\big) = \big(\,\mathcal{M}(\pi_1)(\varphi)(1),\, \mathcal{M}(\pi_2)(\varphi)(1)\,\big).$$

The last line makes clear that the multiplicities of 1, that is of 'head', in the marginal mulitsets are counted by this function *heads*.

More generally, for a 'dimension' $N \geq 1$ one can define:

$$\begin{array}{ccc} \mathcal{M}[K]\big(\mathbf{2}^N\big) & \xrightarrow{\qquad heads \qquad} & \{0,\ldots,K\}^N \\ \varphi & \longmapsto & \big(\,\mathcal{M}(\pi_1)(\varphi)(1),\ldots,\mathcal{M}(\pi_N)(\varphi)(1)\,\big) \end{array} \tag{10}$$

We shall need the following combinatorial result about the marginal heads function.

Lemma 3.2 *For* $K \geq 0$ *and* $N \geq 1$ *and numbers* $\boldsymbol{n} = n_1,\ldots,n_N \in \{0,\ldots,K\}$ *one has:*

$$\sum_{\varphi\in\mathcal{M}[K](\mathbf{2}^N),\, heads(\varphi)=\boldsymbol{n}} (\varphi) \;=\; \prod_{1\leq i\leq N} \binom{K}{n_i}.$$

Proof. We do the proof for $N = 2$ and use Vandermonde's formula, see *e.g.* [6]: for numbers $B, G \in \mathbb{N}$ and $K \leq B+G$,

$$\binom{B+G}{K} = \sum_{b\leq B,\, g\leq G,\, b+g=K} \binom{B}{b}\cdot\binom{G}{g}. \tag{11}$$

Now, for a multiset $\varphi \in \mathcal{M}[K](\mathbf{2}\times\mathbf{2})$ we use new variables i, j below, where $i = \varphi(1,1)$ and $j = \varphi(0,1)$.

$$
\begin{aligned}
&\sum_{\varphi\in\mathcal{M}[K](\mathbf{2}\times\mathbf{2}),\, heads(\varphi)=(n_1,n_2)} (\varphi) \\
&= \sum_{\varphi\in\mathcal{M}[K](\mathbf{2}\times\mathbf{2}),\, \varphi(1,0)+\varphi(1,1)=n_1,\, \varphi(0,1)+\varphi(1,1)=n_2} \frac{K!}{\varphi(0,0)!\cdot\varphi(1,0)!\cdot\varphi(0,1)!\cdot\varphi(1,1)!} \\
&= \sum_{i\leq n_1,\, j\leq K-n_1,\, i+j=n_2} \frac{K!}{(K-n_1-j)!\cdot(n_1-i)!\cdot j!\cdot i!} \\
&= \sum_{i\leq n_1,\, j\leq K-n_1,\, i+j=n_2} \frac{K!}{n_1!\cdot(K-n_1)!}\cdot\frac{n_1!}{i!\cdot(n_1-i)!}\cdot\frac{(K-n_1)!}{j!\cdot(K-n_1-j)!} \\
&= \binom{K}{n_1}\cdot\sum_{i\leq n_1,\, j\leq K-n_1,\, i+j=n_2}\binom{n_1}{i}\cdot\binom{K-i}{j} \overset{(11)}{=} \binom{K}{n_1}\cdot\binom{K}{n_2}.
\end{aligned}
$$
□

Definition 3.3 For a 'two-coin' distribution $\gamma \in \mathcal{D}(\mathbf{2}\times\mathbf{2})$ and a 'toss number' K the *bivariate* binomial distribution is defined as pushforward along the marginal heads function, of a multinomial distribution:

$$
\begin{aligned}
bvbn[K](\gamma) &:= \mathcal{D}(heads)\Big(mn[K](\gamma)\Big) \in \mathcal{D}\Big(\{0,1,\ldots,K\}\times\{0,1,\ldots,K\}\Big) \\
&= \sum_{0\leq n_1,n_2\leq K}\ \sum_{\varphi\in\mathcal{M}[K](\mathbf{2}\times\mathbf{2}),\, heads(\varphi)=(n_1,n_2)} mn[K](\gamma)(\varphi)\,\big|\, n_1,n_2\big\rangle.
\end{aligned}
\tag{12}
$$

The N-dimensional version of this binomial distribution is obtained by first forming the multinomial distribution $mn[K](\gamma) \in \mathcal{D}(\mathcal{M}[K](\mathbf{2}^N))$, for an N-coin $\gamma \in \mathcal{D}(\mathbf{2}^N)$, and then pushing along the N-dimensional marginal heads function (10).

Example 3.4 Consider the two-coin distribution $\gamma = \frac{3}{8}|0,0\rangle + \frac{5}{12}|0,1\rangle + \frac{1}{12}|1,0\rangle + \frac{1}{8}|1,1\rangle$. It is not hard to see that it is entwined, *i.e.* that it is different from the product $\otimes$ of its marginals. We use $K = 2$ and first consider the multinomial distribution on $\mathcal{M}[2](\mathbf{2}\times\mathbf{2})$. It is written via 'kets of kets' in:

$$
\begin{aligned}
mn[2](\gamma) = {} & \tfrac{9}{64}\big|\, 2|0,0\rangle\big\rangle + \tfrac{5}{16}\big|\, 1|0,0\rangle + 1|0,1\rangle\big\rangle + \tfrac{25}{144}\big|\, 2|0,1\rangle\big\rangle + \tfrac{1}{16}\big|\, 1|0,0\rangle + 1|1,0\rangle\big\rangle \\
& + \tfrac{5}{72}\big|\, 1|0,1\rangle + 1|1,0\rangle\big\rangle + \tfrac{1}{144}\big|\, 2|1,0\rangle\big\rangle + \tfrac{3}{32}\big|\, 1|0,0\rangle + 1|1,1\rangle\big\rangle \\
& + \tfrac{5}{48}\big|\, 1|0,1\rangle + 1|1,1\rangle\big\rangle + \tfrac{1}{48}\big|\, 1|1,0\rangle + 1|1,1\rangle\big\rangle + \tfrac{1}{64}\big|\, 2|1,1\rangle\big\rangle.
\end{aligned}
$$

We now obtain the associated bivariate binomial distribution by applying the marginals heads function *heads* to the multisets inside the above big kets $\big|\, - \big\rangle$. By counting the total number of 1's on the left and on the right in these multisets one obtains:

$$
\begin{aligned}
bvbn[2](\gamma) = \mathcal{D}(heads)\Big(mn[2](\gamma)\Big) = {} & \tfrac{9}{64}|0,0\rangle + \tfrac{5}{16}|0,1\rangle + \tfrac{25}{144}|0,2\rangle + \tfrac{1}{16}|1,0\rangle + \tfrac{47}{288}|1,1\rangle \\
& + \tfrac{5}{48}|1,2\rangle + \tfrac{1}{144}|2,0\rangle + \tfrac{1}{48}|2,1\rangle + \tfrac{1}{64}|2,2\rangle.
\end{aligned}
$$

The bivariate binomial for draw size $K = 10$, of the same γ, is plotted on the right in Figure 1.

Remark 3.5 (i) Our formulation of the bivariate binomial in Definition 3.3 uses the functoriality of $\mathcal{D}$. Common definitions in the literature do not exploit this functoriality and involve more complicated explicit formulations. For instance, in [1, Eqn. (1)] one finds the following expression, translated into ket form:

$$
\sum_{0\leq k,l\leq K}\ \sum_{\max(0,k+l-K)\leq i\leq\min(k,l)} \frac{K!}{i!\cdot(k-i)!\cdot(l-i)!\cdot(K-k-l+i)!} \cdot\gamma(0,0)^i\cdot\gamma(0,0)^{k-i}\cdot\gamma(0,0)^{l-i}\cdot\gamma(0,0)^{K-k-l+i}\,\big|\, k,l\big\rangle
\tag{13}
$$

One recognises elements of the multinomial distribution (9) in this formulation, but it is hard to follow what is going on. In fact, this formulation (13) is subtly different, from the one that we use in (12), since it does not count *heads* / *ones* but it counts *tails* / *zeros*, in each coordinate separately.
Multi-dimensional (multivariate) versions of the formula (13) become even harder, see *e.g.* [1,17]. Here, the multi-dimensional form is obtained in a straightforward manner: the definition (12) still works for an N-coin $\gamma \in \mathcal{D}\big(\mathbf{2}^N\big)$ with the marginal heads function (10) used in N-ary form.

(ii) We can also give a more explicit formulation of our functorial definition (12) of the bivariate binomial, in the style of (13), via the following characterisation of the inverse image.

$$\begin{array}{l}\big\{\varphi \in \mathcal{M}[K]\big(\mathbf{2}\times\mathbf{2}\big) \;\big|\; heads(\varphi) = (n_1, n_2)\big\} \\ = \begin{cases} \big\{(K-n_2-i)\big|0,0\big\rangle + (n_2-n_1+i)\big|0,1\big\rangle + i\big|1,0\big\rangle + (n_1-i)\big|1,1\big\rangle \\ \qquad\qquad \big|\; 0 \le i \le \min(n_1, K-n_2)\big\} & \text{if } n_1 \le n_2 \\ \big\{(K-n_1-i)\big|0,0\big\rangle + i\big|0,1\big\rangle + (n_1-n_2+i)\big|1,0\big\rangle + (n_2-i)\big|1,1\big\rangle \\ \qquad\qquad \big|\; 0 \le i \le \min(n_2, K-n_1)\big\} & \text{if } n_2 < n_1. \end{cases}\end{array} \tag{14}$$

We first collect some basic, standard properties about the bivariate binomial distribution, in particular in relation to the ordinary binomial distribution. The definition (12) of the bivariate binomial distribution via functoriality allows equational reasoning in proofs. The details are relegated to the appendix.

Lemma 3.6 *Let a number $K \in \mathbb{N}$ and a two-coin distribution $\gamma \in \mathcal{D}\big(\mathbf{2}\times\mathbf{2}\big)$ be given. For clarity we use different notation for the various projections:*

$$\mathbf{2}\times\mathbf{2} \underset{\pi_2}{\overset{\pi_1}{\rightrightarrows}} \mathbf{2} \qquad\qquad \{0,\ldots,K\}\times\{0,\ldots,K\} \underset{\Pi_2}{\overset{\Pi_1}{\rightrightarrows}} \{0,\ldots K\} \tag{15}$$

We then write $\gamma_1 \coloneqq \mathcal{D}(\pi_1)(\gamma)$ and $\gamma_2 \coloneqq \mathcal{D}(\pi_2)(\gamma)$ in $\mathcal{D}(\mathbf{2})$ for the two marginal coins. They are determined by the 'heads' probabilities $\gamma_1(1), \gamma_2(1) \in [0,1]$, via the isomorphism $\mathcal{D}(\mathbf{2}) \cong [0,1]$.

(i) *The marginals of the bivariate binomial distribution are the (ordinary) binomials of the marginals of the two-coin distribution:*

$$\mathcal{D}\big(\Pi_1\big)\Big(bvbn[K](\gamma)\Big) \;=\; bn[K]\Big(\gamma_1(1)\Big) \qquad\qquad \mathcal{D}\big(\Pi_2\big)\Big(bvbn[K](\gamma)\Big) \;=\; bn[K]\Big(\gamma_2(1)\Big).$$

(ii) *When γ is non-entwined, that is, when $\gamma = \gamma_1 \otimes \gamma_2$, then $bvbn[K](\gamma)$ is also non-entwined, with:*

$$bvbn[K]\Big(\gamma_1 \otimes \gamma_2\Big) \;=\; bn[K]\Big(\gamma_1(1)\Big) \otimes bn[K]\Big(\gamma_2(1)\Big).$$

This second result confirms that the notion of bivariate binomial is a well-defined extension of the ordinary binomial: if the two-coin distribution γ is a tensor product of two coins, the bivariate binomial is a tensor product of the two (marginal, ordinary) binomial distributions.

There is one more well-behavedness property that we add, namely closure under convolution, as we have seen for binomial and multinomial distributions, in Lemmas 2.1 and 2.2. Again, the proofs are in the appendix.

Proposition 3.7 *Let $\gamma \in \mathcal{D}\big(\mathbf{2}\times\mathbf{2}\big)$ be given, with numbers $K, L \in \mathbb{N}$. Bivariate binomials are closed under convolution, as expressed on the left below.*

$$bvbn[K](\gamma) \;+\; bvbn[L](\gamma) \;=\; bvbn[K+L](\gamma) \qquad\qquad bvbn[K](\gamma) \;=\; K\cdot\gamma \;=\; \gamma+\cdots+\gamma.$$

Thus, the bivariate binomial is a K-fold convolution of the two-coin distribution γ, as on the right. □

4 Mean, variance and covariance

This section describes common statistical characteristics of the bivariate binomial distribution, namely the mean, variance and covariance. It turns out that these values are K-times the values for the underlying two-coin distributions. This gives nice formulas. The proofs require some work.

First, for a distribution $\omega \in \mathcal{D}\big(\mathbb{R}^N\big)$, for some number $N \geq 1$, the mean is described as the tuple in $\mathbb{R}^N$ given by the validities of the projections $\pi_i \colon \mathbb{R}^N \to \mathbb{R}$, viewed as observables. This gives the formula:

$$\text{mean}(\omega) \coloneqq \big(\, \omega \models \pi_1, \, \ldots, \, \omega \models \pi_N \,\big). \tag{16}$$

Lemma 4.1 *For $K \in \mathbb{N}$ and $\gamma \in \mathcal{D}\big(\mathbf{2} \times \mathbf{2}\big)$,*

$$\text{mean}\Big(bvbn[K](\gamma)\Big) \;=\; K \cdot \text{mean}\big(\gamma\big) \;\in\; \mathbb{R} \times \mathbb{R}.$$

Proof. An abstract argument uses that the mean ommutes with convolutions, in combination with Proposition 3.7. But one can also proceed in a more concrete manner. First, the mean of $\gamma \in \mathcal{D}\big(\mathbf{2} \times \mathbf{2}\big)$ can be expressed in terms of γ's marginals $\gamma_1 \coloneqq \mathcal{D}(\pi_1)(\gamma), \gamma_2 \coloneqq \mathcal{D}(\pi_2)(\gamma) \in \mathcal{D}(\mathbf{2})$, namely as:

$$\begin{aligned}\text{mean}\big(\gamma\big) \overset{(16)}{=} \big(\, \omega \models \pi_1, \, \omega \models \pi_2 \,\big) &= \left(\sum_{b_1, b_2 \in \mathbf{2}} \gamma(b_1, b_2) \cdot b_1, \; \sum_{b_1, b_2 \in \mathbf{2}} \gamma(b_1, b_2) \cdot b_2\right) \\ &= \left(\sum_{b_2 \in \mathbf{2}} \gamma(1, b_2), \; \sum_{b_1 \in \mathbf{2}} \gamma(b_1, 1)\right) \;=\; \Big(\gamma_1(1), \, \gamma_2)(1)\Big).\end{aligned} \tag{$*$}$$

Next, using the projection notation from (15),

$$\begin{aligned}\text{mean}\Big(bvbn[K](\gamma)\Big) &\overset{(16)}{=} \big(\, bvbn[K](\gamma) \models \Pi_1, \, bvbn[K](\gamma) \models \Pi_2 \,\big) \\ &= \left(\sum_{n_1, n_2 \in \mathbb{N}} bvbn[K](\gamma)(n_1, n_2) \cdot n_1, \; \sum_{n_1, n_2 \in \mathbb{N}} bvbn[K](\gamma)(n_1, n_2) \cdot n_2\right) \\ &= \left(\sum_{n_1 \in \mathbb{N}} \mathcal{D}(\Pi_1)\Big(bvbn[K](\gamma)\Big)(n_1) \cdot n_1, \; \sum_{n_2 \in \mathbb{N}} \mathcal{D}(\Pi_2)\Big(bvbn[K](\gamma)\Big)(n_2) \cdot n_2\right) \\ &= \left(\sum_{n_1 \in \mathbb{N}} bn[K]\big(\gamma_1(1)\big)(n_1) \cdot n_1, \; \sum_{n_2 \in \mathbb{N}} bn[K]\big(\gamma_2(1)\big)(n_2) \cdot n_2\right) \qquad \text{by Lemma 3.6 (i)} \\ &= \Big(\text{mean}\big(bn[K]\big(\gamma_1(1)\big)\big), \, \text{mean}\big(bn[K]\big(\gamma_2(1)\big)\big)\Big) \\ &= \Big(K \cdot \gamma_1(1), \, K \cdot \gamma_2(1)\Big) \qquad \text{by Lemma 2.1 (ii)} \\ &= K \cdot \Big(\gamma_1(1), \, \gamma_2)(1)\Big) \overset{(*)}{=} \text{mean}\big(\gamma\big).\end{aligned}$$

□

Analogous results exist for the variance and covariance of the bivariate binomial. We first recall the definitions, for a distribution $\omega \in \mathcal{D}(X)$ and two observables $p, q \colon X \to \mathbb{R}$ on the same set:

$$\begin{aligned}\text{var}\big(\omega, p\big) &= \Big(\omega \models p \mathbin{\&} p\Big) - \Big(\omega \models p\Big)^2 \\ \text{cov}\big(\omega, p, q\big) &= \Big(\omega \models p \mathbin{\&} q\Big) - \Big(\omega \models p\Big) \cdot \Big(\omega \models q\Big).\end{aligned} \tag{17}$$

The conjunction & of observables is given by point-wise multiplication. When the underlying set X is (a subset of) $\mathbb{R}^N$ one typically computes these variances and covariances for the projections. That will be done below.

Theorem 4.2 *For $K \in \mathbb{N}$ and $\gamma \in \mathcal{D}\big(\mathbf{2} \times \mathbf{2}\big)$ one has, for $i, j \in \{1, 2\}$,*

$$\mathrm{var}\Big(bvbn[K](\gamma), \Pi_i\Big) \;=\; K \cdot \mathrm{var}(\gamma, \pi_i) \qquad \textit{and} \qquad \mathrm{cov}\Big(bvbn[K](\gamma), \Pi_i, \Pi_j\Big) \;=\; K \cdot \mathrm{cov}(\gamma, \pi_i, \pi_j).$$

Proof. We do the case for $i = 1$. First,

$$\begin{aligned}
\mathrm{var}\Big(\gamma, \pi_1\Big) \overset{(17)}{=} \Big(\gamma \models \pi_1 \,\&\, \pi_1\Big) \;-\; \Big(\gamma \models \pi_1\Big)^2 &= \Bigg(\sum_{b_1, b_2 \in \mathbf{2}} \gamma(b_1, b_2) \cdot b_1 \cdot b_1\Bigg) \;-\; \Bigg(\sum_{b_1, b_2 \in \mathbf{2}} \gamma(b_1, b_2) \cdot b_1\Bigg)^2 \\
&= \gamma(1,0) + \gamma(1,1) \;-\; \Big(\gamma(1,0) + \gamma(1,1)\Big)^2.
\end{aligned}$$

We turn to the bivariate binomial and make use of Lemma 2.2 (iii).

$$\begin{aligned}
\mathrm{var}\Big(bvbn[K](\gamma), \Pi_i\Big) &= \Big(bvbn[K](\gamma) \models \Pi_1 \,\&\, \Pi_1\Big) \;-\; \Big(bvbn[K](\gamma) \models \Pi_1\Big)^2 \\
&= \Bigg(\sum_{n_1, n_2 \in \mathbb{N}} bvbn[K](\gamma)(n_1, n_2) \cdot n_1 \cdot n_1\Bigg) - \Bigg(\sum_{n_1, n_2 \in \mathbb{N}} bvbn[K](\gamma)(n_1, n_2) \cdot n_1\Bigg)^2 \\
&= \Bigg(\sum_{\varphi \in \mathcal{M}[K](\mathbf{2}\times\mathbf{2})} mn[K](\gamma)(\varphi) \cdot \big(\varphi(1,0) + \varphi(1,1)\big)^2\Bigg) \\
&\qquad - \Bigg(\sum_{\varphi \in \mathcal{M}[K](\mathbf{2}\times\mathbf{2})} mn[K](\gamma)(\varphi) \cdot \big(\varphi(0,1) + \varphi(1,1)\big)\Bigg)^2 \\
&= K \cdot (K-1) \cdot \gamma(1,0)^2 + K \cdot \gamma(1,0) \;+\; 2 \cdot K \cdot (K-1) \cdot \gamma(1,0) \cdot \gamma(1,1) \\
&\qquad + K \cdot (K-1) \cdot \gamma(1,1)^2 + K \cdot \gamma(1,1) - \Big(K \cdot \gamma(1,0) + K \cdot \gamma(1,1)\Big)^2 \\
&= K \cdot \Big(- \gamma(1,0)^2 + \gamma(1,0) - 2 \cdot \gamma(1,0) \cdot \gamma(1,1) - \gamma(1,1)^2 + \gamma(1,1)\Big) \\
&= K \cdot \Big(\gamma(1,0) + \gamma(1,1) - \big(\gamma(1,0) + \gamma(1,1)\big)^2\Big) \\
&= K \cdot \mathrm{var}\Big(\gamma, \pi_1\Big) \quad \text{as shown above.}
\end{aligned}$$

For covariance, the cases where $i = j$ are covered by the previous point, since covariance with equal observables is variance. We do the case $i = 1, j = 2$, which gives the same outcome as for $i = 2, j = 1$. We first look at the covariance of the two-coin distribution γ.

$$\begin{aligned}
\mathrm{cov}\Big(\gamma, \pi_1, \pi_2\Big) &\overset{(17)}{=} \Big(\gamma \models \pi_1 \,\&\, \pi_2\Big) \;-\; \Big(\gamma \models \pi_1\Big) \cdot \Big(\gamma \models \pi_2\Big) \\
&= \Bigg(\sum_{b_1, b_2 \in \mathbf{2}} \gamma(b_1, b_2) \cdot b_1 \cdot b_2\Bigg) \;-\; \Bigg(\sum_{b_1, b_2 \in \mathbf{2}} \gamma(b_1, b_2) \cdot b_1\Bigg) \cdot \Bigg(\sum_{b_1, b_2 \in \mathbf{2}} \gamma(b_1, b_2) \cdot b_2\Bigg) \\
&= \gamma(1,1) \;-\; \Big(\gamma(1,0) + \gamma(1,1)\Big) \cdot \Big(\gamma(0,1) + \gamma(1,1)\Big).
\end{aligned}$$

Next, again by Lemma 2.2 (iii):

$$\begin{aligned}
&\mathrm{cov}\Big(bvbn[K](\gamma), \Pi_1, \Pi_2\Big) \\
&= \Big(bvbn[K](\gamma) \models \Pi_1 \,\&\, \Pi_2\Big) \;-\; \Big(bvbn[K](\gamma) \models \Pi_1\Big) \cdot \Big(bvbn[K](\gamma) \models \Pi_2\Big) \\
&= \left(\sum_{n_1, n_2 \in \mathbb{N}} bvbn[K](\gamma)(n_1, n_2) \cdot n_1 \cdot n_2\right) \\
&\qquad - \left(\sum_{n_1, n_2 \in \mathbb{N}} bvbn[K](\gamma)(n_1, n_2) \cdot n_1\right) \cdot \left(\sum_{n_1, n_2 \in \mathbb{N}} bvbn[K](\gamma)(n_1, n_2) \cdot n_2\right) \\
&= \left(\sum_{\varphi \in \mathcal{M}[K](\mathbf{2} \times \mathbf{2})} mn[K](\gamma)(\varphi) \cdot \big(\varphi(1,0) + \varphi(1,1)\big) \cdot \big(\varphi(0,1) + \varphi(1,1)\big)\right) \\
&\qquad - \Big(K \cdot \gamma(1,0) + K \cdot \gamma(1,1)\Big) \cdot \Big(K \cdot \gamma(0,1) + K \cdot \gamma(1,1)\Big) \\
&= K \cdot (K-1) \cdot \Big(\gamma(1,0) \cdot \gamma(0,1) + \gamma(1,0) \cdot \gamma(1,1) + \gamma(1,1) \cdot \gamma(0,1) \\
&\qquad + \gamma(1,1)^2\Big) + K \cdot \gamma(1,1) \;-\; K^2 \cdot \Big(\gamma(1,0) + \gamma(1,1)\Big) \cdot \Big(\gamma(0,1) + \gamma(1,1)\Big) \\
&= K \cdot \Big(\gamma(1,1) \;-\; \big(\gamma(1,0) + \gamma(1,1)\big) \cdot \big(\gamma(0,1) + \gamma(1,1)\big)\Big) \\
&= \mathrm{cov}\Big(\gamma, \pi_1, \pi_2\Big), \quad \text{see before.}
\end{aligned}$$

□

We conclude with an observation which will be useful for Expectation Maximisation in Section 6.

Fact 4.3 *Suppose we have a distribution $\sigma \in \mathcal{D}\Big(\{0, \ldots, K\} \times \{0, \ldots, K\}\Big)$, for some given $K \in \mathbb{N}$. We know that σ is a bivariate binomial distribution, of the form $\sigma = \mathrm{bvbn}[K](\gamma)$ but we do not know $\gamma \in \mathcal{D}\big(\mathbf{2} \times \mathbf{2}\big)$. We can compute the mean and the covariance of σ. Via Proposition 4.1 and Theorem 4.2 we can then obtain the probabilities of γ, and thus γ itself, in the following way.*

- $\gamma(1,1) = \frac{1}{K} \cdot \Big(\mathrm{cov}\big(\sigma, \Pi_1, \Pi_2\big) - \big(\sigma \models \Pi_1\big) \cdot \big(\sigma \models \Pi_2\big)\Big)$;
- $\gamma(1,0) = \frac{1}{K} \cdot \big(\sigma \models \Pi_1\big) - \gamma(1,1)$;
- $\gamma(0,1) = \frac{1}{K} \cdot \big(\sigma \models \Pi_2\big) - \gamma(1,1)$;
- $\gamma(0,0) = 1 - \gamma(1,0) - \gamma(0,1) - \gamma(1,1)$.

Thus, when the draw size K is given, a bivariate binomial distribution is entirely determined by its mean and covariance, since its mean is given by the pair of validities $(\, \mathrm{bvbn}[K](\gamma) \models \Pi_1,\, \mathrm{bvbn}[K](\gamma) \models \Pi_2 \,)$, see (16).

5 Laplace's rule of succession

Laplace's rule of succession provides an answer to what was originally formulated as the sunrise problem: suppose I have seen the sun rise in n successive days, what is the probability that it will rise tomorrow? The answer according to Laplace's rule is $\frac{n+1}{n+2}$. It will go to one as n goes to infinity.

The binomial distribution plays a role in a systematic answer formulation of the rule of succession. Hence the question arises: is there a role for the bivariate binomial too? In order to answer this question we first provide a general, modern interpretation of the rule of succession. It is generally understood as providing the expected outcome after an update, see *e.g.* [16] or [2]. We reformulate this as: the rule of succession calculates the *mean of a dagger*.

The relevant calculations involve continuous distributions as priors, such as the Beta and Dirichlet distributions. We have to assume that the reader is reasonably familiar with these distributions and with the associated results. Therefore, this section only gives a sketch of the situation at hand, from a structural angle. We start by recalling, without proof, basic facts about the Beta distribution on $[0,1]$ and about the Dirichlet distribution on discrete distributions $\mathcal{D}(X)$. These continuous distributions are typically used for the bias parameter $r \in [0,1]$ of a flip distribution *flip*(r) and for the distribution parameter $\omega \in \mathcal{D}(X)$ of a multinomial distribution $mn[K](\omega)$. The results below are standard. The formulation of the conjugate prior property in terms of daggers is in line with [8].

Lemma 5.1 *We write Beta(α,β) for the Beta distribution on the unit interval $[0,1]$, where $\alpha,\beta \in \mathbb{N}_{>0}$.*

(i) $\text{mean}\big(Beta(\alpha,\beta)\big) = \frac{\alpha}{\alpha+\beta}$.

(ii) *The Beta distribution on $[0,1]$ is conjugate prior to the binomial channel $\text{bn}[K]\colon [0,1] \rightsquigarrow \{0,\dots,K\}$. The dagger channel $\text{bn}[K]^{\dagger}_{\text{Beta}(\alpha,\beta)}\colon \{0,\dots,K\} \rightsquigarrow [0,1]$, with Beta$(\alpha,\beta)$ as prior distribution, gives a parameter update of the form:*

$$bn[K]^{\dagger}_{Beta(\alpha,\beta)}(n) \;=\; Beta\big(\alpha+n,\, \beta+K-n\big).$$

(iii) *Laplace's rule of succession gives in this case as mean of the dagger:*

$$\text{mean}\Big(bn[K]^{\dagger}_{Beta(\alpha,\beta)}(n)\Big) \;=\; \text{mean}\Big(Beta\big(\alpha+n,\, \beta+K-n\big)\Big) \;=\; \frac{\alpha+n}{\alpha+\beta+K}. \qquad \square$$

Starting from a uniform prior on $[0,1]$, for $\alpha=\beta=1$, the probability of seeing another sunrise after seeing K out of K sunrises is thus:

$$\text{mean}\Big(bn[K]^{\dagger}_{Beta(1,1)}(K)\Big) \;=\; \frac{1+K}{1+1+K} \;=\; \frac{K+1}{K+2}.$$

This is what Laplace calculated.

There is a similar situation for multinomial distributions (9), where the appropriate distribution is the Dirichlet distribution $Dir(\psi)$ on $\mathcal{D}(X)$, for a finite set X. This distribution involves a multiset $\psi \in \mathcal{M}(X)$ with full support as parameter.

Lemma 5.2 (i) $\text{mean}\big(Dir(\psi)\big) = Flrn(\psi)$.

(ii) *The Dirichlet distribution on $\mathcal{D}(X)$ is conjugate prior to the multinomial channel $\text{mn}[K]\colon \mathcal{D}(X) \rightsquigarrow \mathcal{M}[K](X)$ with parameter update induced by observation $\varphi \in \mathcal{M}[K](X)$ is:*

$$mn[K]^{\dagger}_{Dir(\psi)}(\varphi) \;=\; Dir\big(\psi+\varphi\big).$$

(iii) *Laplace's rule of succession thus gives:*

$$\text{mean}\Big(mn[K]^{\dagger}_{Dir(\psi)}(\varphi)\Big) \;=\; \text{mean}\Big(Dir\big(\psi+\varphi\big)\Big) \;=\; Flrn\big(\psi+\varphi\big). \qquad \square$$

We now come to the bivariate binomial situation.

Proposition 5.3 *For $K \in \mathbb{N}$ consider the bivariate binomial as a channel $\text{bvbn}[K]\colon \mathcal{D}\big(\mathbf{2}\times\mathbf{2}\big) \to \{0,\dots,K\}\times\{0,\dots,K\}$, with a Dirichlet distribution $\text{Dir}(\psi)$ on $\mathcal{D}\big(\mathbf{2}\times\mathbf{2}\big)$, for a multiset parameter $\psi \in \mathcal{M}\big(\mathbf{2}\times\mathbf{2}\big)$ with full support. Laplace's rule of succession gives for observations $n_1,n_2 \in \{0,\dots,K\}$ as outcome:*

$$\text{mean}\Big(bvbn[K]^{\dagger}_{Dir(\psi)}(n_1,n_2)\Big) \;=\; Flrn\left(\sum_{\varphi\in\mathcal{M}[K](\mathbf{2}\times\mathbf{2}),\, heads(\varphi)=(n_1,n_2)} \psi+\varphi\right).$$

Proof. (Sketch) We write Z for a suitable normalisation constant in:

$$\begin{aligned}
&\mathrm{mean}\Big(bvbn[K]^{\dagger}_{Dir(\psi)}(n_1,n_2)\Big) \\
&= \frac{1}{Z}\cdot\int_{\gamma\in\mathcal{D}(\mathbf{2}\times\mathbf{2})} bvbn[K](\gamma)(n_1,n_2)\cdot Dir(\psi)(\gamma)\cdot\gamma\,\mathrm{d}\gamma \\
&= \frac{1}{Z}\cdot\sum_{\varphi\in\mathcal{M}[K](\mathbf{2}\times\mathbf{2}),\, heads(\varphi)=(n_1,n_2)}\int_{\gamma\in\mathcal{D}(\mathbf{2}\times\mathbf{2})} mn[K](\gamma)(\varphi)\cdot Dir(\psi)(\gamma)\cdot\gamma\,\mathrm{d}\gamma \\
&= Flrn\left(\sum_{\varphi\in\mathcal{M}[K](\mathbf{2}\times\mathbf{2}),\, heads(\varphi)=(n_1,n_2)} \mathrm{mean}\Big(mn[K]^{\dagger}_{Dir(\psi)}(\varphi)\Big)\right) \\
&= Flrn\left(\sum_{\varphi\in\mathcal{M}[K](\mathbf{2}\times\mathbf{2}),\, heads(\varphi)=(n_1,n_2)} Flrn(\psi+\varphi)\right), \quad \text{by Lemma 5.2 (iii)} \\
&= Flrn\left(\sum_{\varphi\in\mathcal{M}[K](\mathbf{2}\times\mathbf{2}),\, heads(\varphi)=(n_1,n_2)} \psi+\varphi\right).
\end{aligned}$$

□

In Lemma 5.1 and 5.2 there are Laplace succession rules arising from conjugate prior situations, but this not a necessary ingredient, as Equation (18) below shows. It involves the combination of the binomial and Poisson distributions, see *e.g.* [15, §6.11]. This applies for instance when one has an imperfect particle counter/detector in physics, where particles are emitted by some source according to a Poisson distribution (with rate λ) and the imperfect detection is captured by a subsequent binomial channel (with chance r of being detected). When n particles are detected, the question is: what is the expected number of emitted particles? The distribution of these total numbers is obtained by updating the prior Poisson distribution. Interestingly, this yields a new Poisson distribution, but with a shift. The Laplace succession rule says in this case:

$$\mathrm{mean}\Big(bn[-](r)^{\dagger}_{pois[\lambda]}(n)\Big) = n + (1-r)\cdot\lambda. \tag{18}$$

Here one views binomial as a channel $bn[-](r)\colon \mathbb{N} \rightsquigarrow \{0,\ldots,K\}$ that is reversed via the dagger, with the Poisson distribution on $\mathbb{N}$ as prior. The formula (18) expresses that after detecting n particles the expected number of emitted particles is $n+(1-r)\cdot\lambda$. The factor $(1-r)\cdot\lambda$ captures the expected number of emitted particles that were not detected. This makes sense.

Without proof we mention the following analogue for bivariate binomials.

Proposition 5.4 *Fix $\lambda\in\mathbb{R}_{\geq 0}$, $K\in\mathbb{N}$, $\gamma\in\mathcal{D}(\mathbf{2}\times\mathbf{2})$. For observations $n_1,n_2\in\{0,\ldots,K\}$, say with $n_1\leq n_2$, one has:*

$$\mathrm{mean}\Big(bvbn[-](\gamma)^{\dagger}_{pois[\lambda]}(n_1,n_2)\Big) = n_2 + \gamma(0,0)\cdot\lambda + \frac{\sum_{0\leq i\leq n_1} ppp(i)\cdot i}{\sum_{0\leq i\leq n_1} ppp(i)},$$

where $ppp(i) \coloneqq pois[\gamma(0,1)\cdot\lambda](n_2-n_1+i)\cdot pois[\gamma(1,0)\cdot\lambda](i)\cdot pois[\gamma(1,1)\cdot\lambda](n_1-i)$. □

6 Expectation Maximisation for bivariate binomials

Expectation Maximisation is an unsupervised machine learning technique for classifying data in a finite number of classes/features. The classification takes place by finding a 'mixture' distribution that assigns propertions (in the form of probabilities) to the different classification features. This section gives an impression how bivariate distributions — in the formulation of this paper — may be useful in concrete applications in machine learning. Below we use the channel-based analysis of Expectation Maximisation developed in [12,14] and demonstrate how it works in a concrete example. It avoids explicit updates of

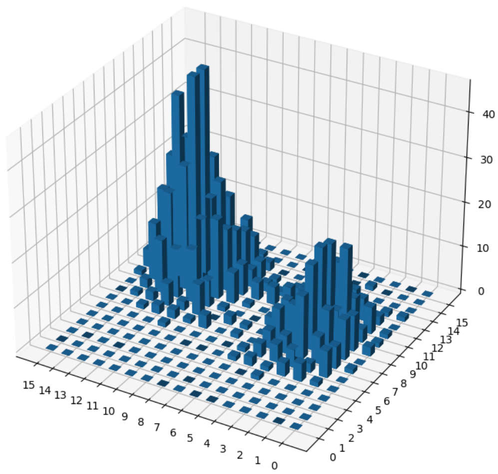


Fig. 2. Plots of 1000 samples from the distribution $\sigma \in \mathcal{D}\big(\{0,\ldots,15\} \times \{0,\ldots,15\}\big)$ from (19).

```
def BivariateBinomialEM (dist, chan, data_dist):
    dagger = chan†_dist
    # E-step, as Jeffrey update
    new_dist = dagger_*(data_dist)
    # M-step, via double dagger
    double_dagger = dagger†_data_dist
    def new_chan(x):  return bvbn[K]( recover(double_dagger(x)) )
    return (new_dist, new_chan)
```

Fig. 3. Expectation Maximisation for finding a mixture of bivariate binomial distributions.

parameters, as happens in common formulations in Expectation Maximisation. However, implicitly, a two-coin distribution parameter is recomputed in each iteration via the steps described in Fact 4.3.

The approach is as follows. We start from a known mixture of two bivariate binomial distributions, with $K = 15$, namely:

$$\sigma := \tfrac{1}{3} \cdot bvbn[K](\gamma_0) + \tfrac{2}{3} \cdot bvbn[K](\gamma_1) \quad \text{where} \quad \begin{cases} \gamma_0 = \frac{3}{8}\big|0,0\big\rangle + \frac{5}{12}\big|0,1\big\rangle + \frac{1}{12}\big|1,0\big\rangle + \frac{1}{8}\big|1,1\big\rangle \\ \gamma_1 = \frac{1}{10}\big|0,0\big\rangle + \frac{1}{10}\big|0,1\big\rangle + \frac{1}{5}\big|1,0\big\rangle + \frac{3}{5}\big|1,1\big\rangle \end{cases} \tag{19}$$

We form a multiset $\psi \in \mathcal{M}[1000]\big(\mathbf{2} \times \mathbf{2}\big)$ via 1000 random samples from this distribution. This multiset is plotted in Figure 2. One recognises two humps that indicate that there may be an underlying mixture of bivariate binomial distributions. Our aim is now to reconstruct not only the $\frac{1}{3} - \frac{2}{3}$ mixture but also the two distributions γ_1, γ_2 from these data alone — where we only assume that two numbers are known, namely the number $K = 15$ of tosses and the number 2 of binomials in the mixture.

Expectation Maximisation is the technique for doing this. It seeks a mixture distribution ω, in this case on $\mathbf{2} = \{0,1\}$ since we seek a mixture of 2 binomials, and a channel $c \colon \mathbf{2} \rightsquigarrow \{0,\ldots,15\}^2$, where both $c(0)$ and $c(1)$ are bivariate distributions. The aim is to minimise the KL-divergence $D_{KL}\big(\mathit{Flrn}(\psi), c_*(\omega)\big)$, so that the divergence between the data distribution and the prediction is minimal, see [12] for details.

This reduction happens via several iterations. A single iteration, in Python-like pseudo code, is described in Figure 3. It produces a new mixture distribution via Jeffrey's update rule [7,10], using the dagger of the original channel, with the mixture distribution as prior. The new channel is obtained by first forming the double dagger, with the data distribution $\mathit{Flrn}(\psi)$ as prior. This double dagger may not be of the right bivariate binomial form, but it is forced into this form via a `recover` function. It uses the steps described in Fact 4.3, to find a distribution on $\mathbf{2} \times \mathbf{2}$ by computing the mean and covariance.

We run the algorithm in Figure 3 five times on an arbitrary mixture distribution $\omega \in \mathcal{D}(\mathbf{2})$ and with a channel $c \colon \mathbf{2} \rightsquigarrow \{0, \ldots, 15\}^2$ consisting of arbitrary bivariate distributions $c(0)$ and $c(1)$. One such run gives as successive divergences:

$$1.991 \qquad 0.272 \qquad 0.093 \qquad 0.087 \qquad 0.087.$$

Starting from a divergence resulting from random choices, we see that these divergence quickly stabilise (in three decimals). The mixture distribution and associated (channel of bivariate binomials) given by two-coin distributions that emerge after these five runs are:

$$0.326\big|\,0\,\big\rangle + 0.674\big|\,1\,\big\rangle \qquad \begin{array}{c} 0.378\big|\,0,0\,\big\rangle + 0.416\big|\,0,1\,\big\rangle + 0.0795\big|\,1,0\,\big\rangle + 0.127\big|\,1,1\,\big\rangle \\ 0.094\big|\,0,0\,\big\rangle + 0.104\big|\,0,1\,\big\rangle + 0.205\big|\,1,0\,\big\rangle + 0.597\big|\,1,1\,\big\rangle \end{array}$$

We see that they are pretty close to values in the mixture (19), from which the 1000 samples in Figure 2 were taken. This example illustrates how Expectation Maximisation can be performed for bivariate binomial distributions.

7 Concluding remarks

This paper reformulates bivariate binomial distributions in a concise manner, by exploiting the functoriality of the mapping $X \mapsto \mathcal{D}(X)$, sending a set X to the set of probability distributions on X. It demonstrates the clarifying power of the categorical approach to probability theory by (re)describing some of the basic properties of bivariate binomials, for instance via (dagger) channels. It also shows that this categorical formulation makes it possible to prove basic properties via equational reasoning. The paper concentrates on two dimensions, for simplicity, but the extension to multiple dimensions is straightforward.

The multivariate binomial distribution be understood as a discrete version of the widely used multi-dimensional Gaussian distribution. By the famous theorem of De Moivre-Laplace every binomial distributions approximate a Gaussian distribution. This is a special case of the Central Limit Theorem. The approximation also works for multivariate binomials, see [18].

References

[1] I. Bayramoglu and O. Elmastas. Discrete distributions connected with the bivariate binomial. *Hacettepe Journ. of Mathematics and Statistics*, 39(1):109–120, 2010. https://dergipark.org.tr/en/pub/hujms/issue/46741/531103.

[2] J. Blitzstein and J. Hwang. *Introduction to Probability.* CRC Press, 2^{nd} edition, 2019, ISBN: 9780429766732, 0429766734

[3] K. Cho and B. Jacobs. Disintegration and Bayesian inversion via string diagrams. *Math. Struct. in Comp. Sci.*, 29(7):938–971, 2019. doi:10.1017/s0960129518000488.

[4] F. Clerc, F. Dahlqvist, V. Danos, and I. Garnier. Pointless learning. In J. Esparza and A. Murawski, editors, *Foundations of Software Science and Computation Structures*, number 10203 in Lect. Notes Comp. Sci., pages 355–369. Springer, Berlin, 2017. doi:10.1007/978-3-662-54458-7_21.

[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] R. Graham, D. Knuth, and O. Patashnik. *Concrete Mathematics.* Addison-Wesley, 2^{nd} edition, 1994, ISBN: 0-201-55802-5

[7] B. Jacobs. The mathematics of changing one's mind, via Jeffrey's or via Pearl's update rule. *Journ. of Artif. Intelligence Research*, 65:783–806, 2019. doi:10.1613/jair.1.11349.

[8] B. Jacobs. A channel-based perspective on conjugate priors. *Math. Struct. in Comp. Sci.*, 30(1):44–61, 2020. doi:10.1017/S0960129519000082.

[9] B. Jacobs. From multisets over distributions to distributions over multisets. In *Logic in Computer Science*. Computer Science Press, 2021.
doi:10.1109/lics52264.2021.9470678.

[10] B. Jacobs. Learning from what's right and learning from what's wrong. In A. Sokolova, editor, *Math. Found. of Programming Semantics*, number 351 in Elect. Proc. in Theor. Comp. Sci., pages 116–133, 2021.
doi:10.4204/EPTCS.351.8.

[11] B. Jacobs. Urns & tubes. *Compositionality*, 4(4), 2022.
doi:10.32408/compositionality-4-4.

[12] B. Jacobs. A principled approach to Expectation Maximisation and Latent Dirichlet Allocation using Jeffrey's update rule. In H. Hansen, A. Scedrov, and R. de Queiroz, editors, *Worksh. on Logic, Language, Inf. & Comp.*, number 13923 in Lect. Notes Comp. Sci., pages 256–273, 2023.
doi:10.1007/978-3-031-39784-4_16.

[13] B. Jacobs. Drawing with distance. *Logical Methods in Comp. Sci.*, 21(2), 2025.
doi:10.46298/lmcs-21(2:11)2025.

[14] B. Jacobs. Structured probabilitistic reasoning. Book, in preparation, see http://www.cs.ru.nl/B.Jacobs/PAPERS/ProbabilisticReasoning.pdf, 2025.

[15] E. Jaynes. *Probability Theory: the Logic of Science*. Cambridge Univ. Press, 2003, ISBN: 9780521592710, 0521592712

[16] S. Ross. *A first course in probability*. Pearson Education, 10th edition, 2018, ISBN: 9780024038722, 0024038725

[17] J. Teugels. Some representations of the multivariate Bernoulli and binomial distributions. *Journ. of Multivariate Analysis*, 32(2):256–268, 1990.
doi:10.1016/0047-259X(90)90084-U.

[18] J. Veeh. The multivariate laplace-de moivre theorem. *Journ. of Multivariate Analysis*, 8(1):46–51, 1986.
doi:10.1016/0047-259X(86)90057-6.

Appendix

We add proofs that are skipped in the main text.

Proof. (of Lemma 3.6)

(i) We do only the first equation, using twice the Binomial Theorem. For $0 \leq n_1 \leq K$ we have:

$$\begin{aligned}
\mathcal{D}\big(\Pi_1\big)\Big(bvbn[K](\gamma)\Big)(n_1) &= \sum_{0\leq n_2\leq K} \mathcal{D}\big(heads\big)\Big(mn[K](\gamma)\Big)(n_1,n_2) \\
&= \sum_{0\leq n_2\leq K} \;\sum_{\varphi\in\mathcal{M}[K](\mathbf{2}\times\mathbf{2}),\, heads(\varphi)=(n_1,n_2)} mn[K](\gamma)(\varphi) \\
&= \sum_{\varphi\in\mathcal{M}[K](\mathbf{2}\times\mathbf{2}),\, \varphi(1,0)+\varphi(1,1)=n_1} (\varphi)\cdot \prod_{i,j\in\mathbf{2}} \gamma(i,j)^{\varphi(i,j)} \\
&= \sum_{0\leq i\leq n_1}\;\sum_{0\leq j\leq K-n_1} \frac{K!}{(K-n_1-j)!\cdot j!\cdot(n_1-i)!\cdot i!} \\
&\qquad \cdot\gamma(0,0)^{K-n_1-j}\cdot\gamma(0,1)^j\cdot\gamma(1,0)^{n_1-i}\cdot\gamma(1,1)^i \\
&= \sum_{0\leq i\leq n_1}\;\sum_{0\leq j\leq K-n_1} \frac{K!}{n_1!\cdot(K-n_1)!}\cdot\frac{n_1!}{i!\cdot(n_1-i)!}\cdot\frac{(K-n_1)!}{j!\cdot(K-n_1-j)!} \\
&\qquad \cdot\gamma(0,0)^{K-n_1-j}\cdot\gamma(0,1)^j\cdot\gamma(1,0)^{n_1-i}\cdot\gamma(1,1)^i \\
&= \binom{K}{n_1}\cdot\big(\gamma(1,0)+\gamma(1,1)\big)^{n_1}\cdot\big(\gamma(0,0)+\gamma(0,1)\big)^{K-n_1} \\
&= \binom{K}{n_1}\cdot\gamma_1(1)^{n_1}\cdot\big(1-\gamma_1(1)\big)^{K-n_1} \;=\; bn[K]\Big(\gamma_1(1)\Big)(n_1).
\end{aligned}$$

(ii) We now get for $0 \leq n_1, n_2 \leq K$, via Lemma 3.2,

$$\begin{aligned}
bvbn[K]\Big(\gamma_1 \otimes \gamma_2\Big)(n_1, n_2) &= \sum_{\varphi \in \mathcal{M}[K](\mathbf{2}\times\mathbf{2}),\, heads(\varphi)=(n_1,n_2)} mn[K]\Big(\gamma_1 \otimes \gamma_2\Big)(\varphi) \\
&= \sum_{\varphi \in heads^{-1}(n_1,n_2)} (\varphi) \cdot \gamma_1(1)^{\varphi(1,0)+\varphi(1,1)} \cdot \gamma_1(0)^{\varphi(0,0)+\varphi(0,1)}. \\
&\qquad \cdot \gamma_2(1)^{\varphi(0,1)+\varphi(1,1)} \cdot \gamma_2(0)^{\varphi(0,0)+\varphi(1,0)} \\
&= \sum_{\varphi \in heads^{-1}(n_1,n_2)} (\varphi) \cdot \gamma_1(1)^{n_1} \cdot \gamma_1(0)^{K-n_1} \cdot \gamma_2(1)^{n_2} \cdot \gamma_2(0)^{K-n_2} \\
&= \binom{K}{n_1} \cdot \binom{K}{n_2} \cdot \gamma_1(1)^{n_1} \cdot \big(1-\gamma_1(1)\big)^{K-n_1} \cdot \gamma_2(1)^{n_2} \cdot \big(1-\gamma_2(1)\big)^{K-n_2} \\
&= bn[K]\big(\gamma_1(1)\big)(n_1) \cdot bn[K]\big(\gamma_2(1)\big)(n_2) \\
&= \Big(bn[K]\big(\gamma_1(1)\big) \otimes bn[K]\big(\gamma_2(1)\big)\Big)(n_1, n_2).
\end{aligned}$$

□

Proof. (of Proposition 3.7) For the first equation we use the closure of multinomial distributions under convolution, from Lemma 2.2 (ii) and note that the commutative monoid involved is $\mathbb{N}^2$ with the component-wise sum $+ \cdot \mathbb{N}^2 \times \mathbb{N}^2 \to \mathbb{N}^2$.

$$\begin{aligned}
& bvbn[K](\gamma) + bvbn[L](\gamma) \\
&= \mathcal{D}(+)\Big(bvbn[K](\gamma) + bvbn[L](\gamma)\Big) \\
&= \sum_{n_1,n_2 \in \mathbb{N}} \sum_{m_1,m_2 \in \mathbb{N}} bvbn[K](\gamma)(n_1,n_2) \cdot bvbn[L](\gamma)\Big)(m_1,m_2)\,\big|\,(n_1,n_2)+(m_1,m_2)\big\rangle \\
&= \sum_{n_1,n_2 \in \mathbb{N}} \sum_{m_1,m_2 \in \mathbb{N}} \left(\sum_{\varphi \in heads^{-1}(n_1,n_2)} mn[K](\gamma)(\varphi)\right) \cdot \left(\sum_{\psi \in heads^{-1}(m_1,m_2)} mn[L](\gamma)(\psi)\right) \\
&\qquad \big|\,(n_1+m_1, n_2+m_2)\big\rangle \\
&= \sum_{\varphi \in \mathcal{M}[K](\mathbf{2}\times\mathbf{2})} \sum_{\psi \in \mathcal{M}[L](\mathbf{2}\times\mathbf{2})} mn[K](\gamma)(\varphi) \cdot mn[L](\gamma)(\psi) \\
&\qquad \big|\,\big(\mathcal{M}(\pi_1)(\varphi)(1) + \mathcal{M}(\pi_1)(\psi)(1), \mathcal{M}(\pi_2)(\varphi)(1) + \mathcal{M}(\pi_2)(\psi)(1)\big)\big\rangle \\
&= \sum_{\varphi \in \mathcal{M}[K](\mathbf{2}\times\mathbf{2})} \sum_{\psi \in \mathcal{M}[L](\mathbf{2}\times\mathbf{2})} mn[K](\gamma)(\varphi) \cdot mn[L](\gamma)(\psi) \\
&\qquad \big|\,\big(\mathcal{M}(\pi_1)(\varphi + \pi_1)(\psi)(1), \mathcal{M}(\pi_2)(\varphi + \psi)(1)\big)\big\rangle \\
&= \mathcal{D}\big(heads\big)\left(\sum_{\varphi \in \mathcal{M}[K](\mathbf{2}\times\mathbf{2})} \sum_{\psi \in \mathcal{M}[L](\mathbf{2}\times\mathbf{2})} mn[K](\gamma)(\varphi) \cdot mn[L](\gamma)(\psi)\,\big|\,\varphi + \psi\big\rangle\right) \\
&= \mathcal{D}\big(heads\big)\Big(mn[K](\gamma) + mn[L](\gamma)\Big) \;=\; \mathcal{D}\big(heads\big)\Big(mn[K+L](\gamma)\Big) \;=\; bvbn[K+L](\gamma).
\end{aligned}$$

The second equation $bvbn[K](\gamma) = K \cdot \gamma$ is obtained via this closure under convolution, using that $bvbn[0](\gamma) = 1\big|0,0\big\rangle = 0 \cdot \gamma$ and:

$$\begin{aligned} bvbn[1](\gamma) &= bvbn[1](\gamma)(0,0)\big|0,0\big\rangle + bvbn[1](\gamma)(0,1)\big|0,1\big\rangle \\ &\quad + bvbn[1](\gamma)(1,0)\big|1,0\big\rangle + bvbn[1](\gamma)(1,1)\big|1,1\big\rangle \\ &= mn[1](\gamma)\big(1|0,0\rangle\big)\big|0,0\big\rangle + mn[1](\gamma)\big(1|0,1\rangle\big)\big|0,1\big\rangle \\ &\quad + mn[1](\gamma)\big(1|1,0\rangle\big)\big|1,0\big\rangle + mn[1](\gamma)\big(1|1,1\rangle\big)\big|1,1\big\rangle \\ &= \gamma(0,0)\big|0,0\big\rangle + \gamma(1,0)\big|1,0\big\rangle + \gamma(0,1)\big|0,1\big\rangle + \gamma(1,1)\big|1,1\big\rangle \;=\; \gamma. \end{aligned}$$

□