Electron. Commun. Probab. **19** (2014), no. 8, 1–12.
DOI: 10.1214/ECP.v19-3152
ISSN: 1083-589X

ELECTRONIC COMMUNICATIONS in PROBABILITY

On runs, bivariate Poisson mixtures and distributions that arise in Bernoulli arrays

Djilali Ait Aoudia* Éric Marchand† François Perron‡
Latifa Ben Hadj Slimene§

Abstract

Distributional findings are obtained relative to various quantities arising in Bernoulli arrays $\{X_{k,j}, k \geq 1, j = 1, \ldots, r+1\}$, where the rows $(X_{k,1}, \ldots, X_{k,r+1})$ are independently distributed as Multinomial$(1, p_{k,1}, \ldots, p_{k,r+1})$ for $k \geq 1$ with the homogeneity across the first r columns assumption $p_{k,1} = \cdots = p_{k,r}$. The quantities of interest relate to the measure of the number of runs of length 2 and are $\underline{S}_n = (S_{n,1}, \ldots, S_{n,r})$, $\underline{S} = \lim_{n\to\infty} \underline{S}_n$, $T_n = \sum_{j=1}^{r} S_{n,j}$, and $T = \lim_{n\to\infty} T_n$, where $S_{n,j} = \sum_{k=1}^{n} X_{k,j} X_{k+1,j}$. With various known results applicable to the marginal distributions of the $S_{n,j}$'s and to their limiting quantities $S_j = \lim_{n\to\infty} S_{n,j}$, we investigate joint distributions in the bivariate ($r = 2$) case and the distributions of their totals T_n and T for $r \geq 2$. In the latter case, we derive a key relationship between multivariate problems and univariate ($r = 1$) problems opening up the path for several derivations and representations such as Poisson mixtures. In the former case, we obtain general expressions for the probability generating functions, the binomial moments and the probability mass functions through conditioning, an analysis of a resulting recursive system of equations, and again by exploiting connections with the univariate problem. More precisely, for cases where $p_{k,j} = \frac{1}{b+k}$ for $j = 1, 2$ with $b \geq 1$, we obtain explicit expressions for the probability generating function of $\underline{S}_n$, $n \geq 1$, and $\underline{S}$, as well as a Poisson mixture representation : $\underline{S}|(V_1 = v_1, V_2 = v_2) \sim^{ind.}$ Poisson(v_i) with $(V_1, V_2) \sim$ Dirichlet$(1, 1, b-1)$ which nicely captures both the marginal distributions and the dependence structure. From this, we derive the fact that $S_1|S_1 + S_2 = t$ is uniformly distributed on $\{0, 1, \ldots, t\}$ for all $b \geq 1$. We conclude with yet another mixture representation for $p_{k,j} = \frac{1}{b+k}$ for $j = 1, 2$ with $b \geq 1$, where we show that $\underline{S}|\alpha \sim p_\alpha$, $\alpha \sim$ Beta$(1, b)$ with p_α a bivariate mass function with Poisson(α) marginals given by $p_\alpha(s_1, s_2) = \frac{e^{-\alpha}\alpha^{s_1+s_2}}{(s_1+s_2+1)!}(s_1 + s_2 + 1 - \alpha)$.

Keywords: Arrays; Bernoulli; Binomial moments; Dirichlet; Multinomial; Poisson distribution; Poisson mixtures; Runs.
AMS MSC 2010: 60C05; 60E05; 62E15.
Submitted to ECP on November 21, 2013, final version accepted on February 13, 2014.

1 Introduction

Consider Bernoulli arrays (Ait Aoudia and Marchand, 2010) $\{X_{k,j}, k \geq 1, j = 1, \ldots, r+1\}$, where the rows $\underline{X}_k = (X_{k,1}, \ldots, X_{k,r+1})$ are independently distributed as Multinomial $(1, p_{k,1}, \ldots, p_{k,r+1})$ for $k \geq 1$ and which arise for instance in sampling with replacement

*Université du Québec à Montréal, CANADA . E-mail: djilali.ait.aoudia@usherbrooke.ca
†Université de Sherbrooke, CANADA . E-mail: eric.marchand@usherbrooke.ca
‡Université de Montréal, CANADA. E-mail: perronf@dms.umontreal.ca
§Université de Sherbrooke, CANADA . E-mail: latifa.ben.hadj.slimene@usherbrooke.ca

one object at a time from an urn with $r+1$ colours. Quantities of interest include the number of runs $S_{n,j} = \sum_{k=1}^{n} X_{k,j}X_{k+1,j}$ of length 2 in the j^{th} column, their total $T_n = \sum_{j=1}^{r} S_{n,j}$ among the first r columns, and the limits $S_j = \lim_{n\to\infty} S_{n,j}$ and $T = \lim_{n\to\infty} T_n$. A fascinating result is as follows.

Result A. For the case $r = 1$ with $p_{k,1} = \frac{1}{k}$, the distribution of S_1 is Poisson(1).

This was recognized in the mid 1990's by Persi Diaconis, as well as Hahlin (1995), and earlier versions are due to Arratia, Barbour, Tavaré (1992), Kolchin (1971), and Goncharov (1944). This elegant result has inspired much interest and lead to various findings relative to the distributions of $S_{n,j}$ and S_j for various other configurations of the $\{p_{k,j}\}$'s, relationships and implications for Pólya urns, records, matching problems, marked Poisson processes, etc, as witnessed by the work of Chern, Hwang and Yeh (2000), Csörgó and Wu (2000), Holst (2007,2008), Huffer, Sethuraman and Sethuraman (2009), Joffe et al. (2004, 2000), Mori (2001), Sethuraman and Sethuraman (2004), among others. As an exemplar, a lovely generalization of **Result A** is the Poisson mixture representation

$$S_1|U \sim Poisson(aU) \text{ with } U \sim Beta(a,b)\,, \text{ for } p_{k,1} = \frac{a}{a+b+k-1},\ a>0, b\geq 0\,, \quad (1.1)$$

as obtained by Mori (2001), as well as Holst (2008).

With known marginal distributions for the $S_{n,j}$'s and the S_j's for various configurations of the $p_{k,j}$'s and a clearly negative pairwise dependence, further questions of interest concern the joint distribution of the vectors $\underline{S}_n = (S_{n,1}, \ldots, S_{n,r})$ and $\underline{S} = \lim_{n\to\infty} \underline{S}_n$. As well, the distribution of the totals T_n and T are also of related interest. With such Poisson distributions and Poisson mixtures arising naturally in these univariate ($r = 1$) situations, it seems natural to investigate multivariate versions of such results. Said otherwise, in what sense and for which configurations of the $\{p_{k,j}\}$'s, can analytical extensions of **Result A** and (1.1) be obtained?

In this paper, we obtain multivariate generalizations for the homogeneous along column case (i.e., first r row components identically distributed) where $p_{k,1} = \cdots = p_{k,r}$. As in Joffe et al. (2004) and Ait Aoudia and Marchand (2010), we first obtain by conditioning a recursive system of equations involving the probability generating functions of the $S_{n,j}$'s in Section 2.1. This permits us, in Section 3, to obtain key result (Theorem 3.1) linking the multivariate $r \geq 2$ distributions of T_n and T to univariate ($r = 1$) analogs. This is especially useful given that results like (1.1) are available and, hence, corollaries are derived. As an illustration, for $p_{k,j} = \frac{\lambda}{\lambda r+k-1}$, we show that the distribution of T is Poisson(λ) and, for $p_{k,j} = \frac{a}{k-1+rb}$, we obtain a Poisson mixture representation for the distribution of T with a Beta mixing variable.

In Section 4, we obtain (Theorems 4.1 and 4.3) for the bivariate case with $p_{k,j} = \frac{1}{b+k}$, $b \geq 1$, explicit expressions and representations for the distributions of $\underline{S}_n$, $n \geq 1$, and $\underline{S}$. For instance, we show (Theorem 4.3 **(b)**) that the distribution of $\underline{S}$ is the mixture of two independent Poisson(V_i), $i = 1, 2$ with (V_1, V_2) having a Dirichlet distribution, with some definitions and preliminary results on such mixtures given earlier in Section 2.2. This represents a natural extension of (1.1) for $a = 1$ as one recovers the univariate result with the Beta marginals of the Dirichlet and the sought-after dependence structure as reflected by the dependence of the Dirichlet components V_1 and V_2. Yet another mixture representation is given in part **(c)** of Theorem 4.3. But it is quite different as the mixing parameter is univariate and the dependence is reflected otherwise through a bivariate distribution with Poisson and non-independent marginals.

2 Preliminary results, definitions and notations

2.1 Definitions, recurrences for pgf's and binomial moments

We work with the quantities $\underline{X}_k$, $S_{n,j}$, S_j, T_n, T, $\underline{S}_n$ and $\underline{S}$ as defined in the Introduction. Ait Aoudia and Marchand (2010) studied the distribution of T_n for the bivariate case ($r = 2$) and the homogeneous (in k) case with $p_{k,1} = p, p_{k,2} = 1 - p$ (and $p_{k,3} = 0$) for all k. We obtain here representations and relationships for the distributions of the vectors $\underline{S}_n$ and $\underline{S}$ as well as those of the totals T_n and T for various non-homogeneous in k configurations of the $p_{k,j}$'s but with identically distributed components of $\underline{X}_k$, in other words

$$p_{k,1} = \cdots = p_{k,r} = p_k \text{ (say)}. \tag{2.1}$$

We pursue by setting $S_{0,j} = 0$ for all j and by introducing the auxiliary random variables $W_{n,1}, \ldots, W_{n,r}$ where

$$W_{n,j} := S_{n-1,j} + X_{n,j}\,, n \geq 1, j \in \{1, \ldots, r\}\,. \tag{2.2}$$

By conditioning, we obtain the following recurrence for the probability generating functions (pgf) $G_0, G_1, G_2, \ldots$ of the random vectors $\underline{S}_0, \underline{S}_1, \underline{S}_2, \ldots$, which also involves the array of pgf's

$$H_{n,j}(t_1, \ldots, t_r) = E[t_j^{W_{n,j}} \prod_{i \neq j} t_i^{S_{n-1,i}}]\,; n \geq 1, j \in \{1, \ldots, r\}\,.$$

Lemma 2.1. *We have for all $\underline{t} = (t_1, \ldots, t_r)$, $n \geq 1, j \in \{1, \ldots, r\}$,*

$$(1) \qquad G_n(\underline{t}) = (1 - \sum_{i=1}^{r} p_{n+1,i})\, G_{n-1}(\underline{t}) + \sum_{i=1}^{r} p_{n+1,i}\, H_{n,i}(\underline{t})$$

$$(2) \qquad H_{n+1,j}(\underline{t}) = (1 - \sum_{i=1}^{r} p_{n+1,i})\, G_{n-1}(\underline{t}) + \sum_{i=1}^{r} p_{n+1,i}\, t_i^{1_{\{i=j\}}} H_{n,i}(\underline{t})\,,$$

with $G_1(\underline{t}) = 1 + \sum_{j=1}^{r} p_{1,j}\, p_{2,j}\, (t_j - 1)$ and $H_{1,j}(\underline{t}) = 1 + p_{1,j}(t_j - 1)$.

Proof. We condition on $\underline{X}_{n+1}$. We obtain with the independence of the $\underline{X}_k$'s and the definitions of the sequences S_n and $W_{n,j}$

$$\begin{aligned}
&\mathcal{L}\left(\underline{S}_n \mid \underline{X}_{n+1} = (0, 0, \ldots, 0)\right) = \mathcal{L}\left(\underline{S}_{n-1}\right), \\
&\mathcal{L}\left(\underline{S}_n \mid \underline{X}_{n+1} = (1, 0, \ldots, 0)\right) = \mathcal{L}\left(W_{n,1}, S_{n-1,2}, \ldots, S_{n-1,r}\right), \\
&\mathcal{L}\left(\underline{S}_n \mid \underline{X}_{n+1} = (0, 1, 0, \cdots, 0)\right) = \mathcal{L}\left(S_{n-1,1}, W_{n,2}, \ldots, S_{n-1,r}\right), \\
&\qquad\qquad \vdots \\
&\mathcal{L}\left(\underline{S}_n \mid \underline{X}_{n+1} = (0, \ldots, 0, 1)\right) = \mathcal{L}\left(S_{n-1,1}, \ldots, S_{n-1,r-1}, W_{n,r}\right).
\end{aligned}$$

Result (1) follows since

$$G_n(\underline{t}) = E(\prod_{i=1}^{r} t_i^{S_{n,i}}) = E\left(E\left(\prod_{i=1}^{r} t_i^{S_{n,i}} | \underline{X}_{n+1}\right)\right).$$

Equations (2) follow along the same lines by conditioning again on $\underline{X}_{n+1}$. Finally, the initial values for $n = 1$ follow from definitions. □

A rearrangement of the above system of equations is as follows.

Lemma 2.2. *We have for all* $\underline{t} = (t_1, \ldots, t_r)$, $n \geq 2, j \in \{1, \ldots, r\}$:

$$(1') \qquad G_n(\underline{t}) = G_{n-1}(\underline{t}) + \sum_{j=1}^{r} p_{n,j}\, p_{n+1,j}\, (t_j - 1)\, H_{n-1,j}(\underline{t})$$

$$(2') \qquad H_{n,j}(\underline{t}) = G_{n-1}(\underline{t}) + p_{n,j}\, (t_j - 1)\, H_{n-1,j}(\underline{t})\,.$$

Proof. Equation ($2'$) follows at once from the difference (2) - (1) in Lemma 2.1, while ($1'$) follows by rewriting (1) in Lemma 2.1 as $G_n(\underline{t}) = G_{n-1}(\underline{t}) + \sum_{j=1}^{r} p_{n+1,j}(H_{n,j}(\underline{t}) - G_{n-1}(\underline{t}))$ and making use of ($2'$). □

As in Holst (2008), we will make use of probability generating function and probability mass function representations of a non-negative integer valued random variable Z which involve the binomial moments $E\binom{Z}{k}; k \in \{0, 1, 2, \ldots\}$; where $\binom{z}{r}$ is taken to be equal to 0 for $r > z$. Indeed, the Taylor series expansion about 1 of the probability generating function ψ_Z, of an non-negative integer-valued random variable Z, having radius of convergence greater than 1 can be expressed as

$$\psi_Z(t) = E(t^Z) = \sum_{k \geq 0} E\binom{Z}{k} (t-1)^k\,, t \in [0, 1], \tag{2.3}$$

and the probability function of Z can be written as

$$P(Z = i) = \sum_{k=i}^{\infty} (-1)^{k-i} \binom{k}{i} E\binom{Z}{k}\,. \tag{2.4}$$

For the bivariate case with non-negative integer valued components Z_1 and Z_2, analogous relationships are

$$E(t_1^{Z_1}\, t_2^{Z_2}) = \sum_{k \geq 0, l \geq 0} \mathbb{E}\binom{Z_1}{k}\binom{Z_2}{l} (t_1 - 1)^k (t_2 - 1)^l\,, (t_1, t_2) \in [0, 1]^2\,, \tag{2.5}$$

$$\text{and } \mathbb{P}\,(Z_1 = x_1, Z_2 = x_2) = \sum_{k \geq x_1, l \geq x_2} (-1)^{k+l-x_1-x_2}\, \mathbb{E}\left[\binom{Z_1}{k}\binom{Z_2}{l}\right] \binom{k}{x_1}\binom{l}{x_2}, \tag{2.6}$$

as long as the Taylor series expansion at $(t_1, t_2) = (1, 1)$ of the probability generating function converges on an open set containing the origin.

2.2 Bivariate Poisson mixtures

We elaborate here on bivariate Poisson mixtures which will arise in the limiting distribution of $\underline{S}_n$ in Section 4. We denote $(\gamma)_k$ as an ascending factorial with $(\gamma)_0 = 1$ and $(\gamma)_k = \prod_{j=0}^{k-1}(\gamma + j)$ for $k = 1, 2, \ldots$. As well, we denote ${}_1F_1(\gamma_1, \gamma_2; z)$ as the Gauss hypergeometric function given by $\sum_{k \geq 0} \frac{(\gamma_1)_k}{(\gamma_2)_k} \frac{z^k}{k!}; z \in \mathbb{R}$.

Definition 2.3. *We will say that the distribution* $U = (U_1, U_2)$ *is a bivariate Poisson mixture with mixing parameter* F *whenever there exists a bivariate random vector* $V = (V_1, V_2)$ *on* $[0, \infty) \times [0, \infty)$ *with cdf* F *such that* $U_i | V \sim^{indep.} Poisson(V_i)$ *.*

The next lemma brings into play bivariate Dirichlet(a_1, a_2, a_3); $a_i > 0$; densities supported on the simplex $S = \{(v_1, v_2) : v_1 \geq 0, v_2 \geq 0, v_1 + v_2 \leq 1\}$ and given by

$$\frac{\Gamma(\sum_i a_i)}{\prod_i \Gamma(a_i)} v_1^{a_1 - 1} v_2^{a_2 - 1} (1 - v_1 - v_2)^{a_3 - 1}\, \mathbb{I}_S(v_1, v_2)\,,$$

as well as Dirichlet$(a_1, a_2, 0)$ distributions defined as $(V_1, V_2, V_3) = (V_1, 1 - V_1, 0)$ with $V_1 \sim$ Beta(a_1, a_2), and the bivariate hypergeometric or Humbert Φ_2 function given by

$$\Phi_2(a, b, c, x, y) = \sum_{j,k\geq 0} \frac{(a)_j (b)_k}{(c)_{j+k}} \frac{x^j y^k}{j!k!} .$$

The connection between these two entities, which we exploit in the following lemma, is that the probability generating function of a Dirichlet(a_1, a_2, a_3) random vector V is given by (Lee, 1971)

$$\psi_V(t_1, t_2) = E(t_1^{V_1} t_2^{V_2}) = \Phi_2(a_1, a_2, a_1 + a_2 + a_3, \log(t_1), \log(t_2)) \,, \text{ for } t_1, t_2 > 0 \,. \quad (2.7)$$

This is obtained in a straightforward manner by expanding the exponential terms in the evaluation of the moment generating function and is also valid for cases where $a_3 = 0$ by a direct evaluation of $E(t_1^{V_1} t_2^{1-V_1})$.

Lemma 2.4. *Consider a bivariate Poisson mixture distribution U with mixing variable $V \sim$ Dirichlet(a_1, a_2, a_3), and $a_1 > 0, a_2 > 0, a_3 \geq 0$. Then,*

(a) *U has probability generating function $\Phi_2(a_1, a_2, a_1 + a_2 + a_3, t_1 - 1, t_2 - 1)$, and probability mass function given by*

$$P(U_1 = i, U_2 = j) = \frac{1}{i!j!} \sum_{k\geq 0, l\geq 0} \frac{(-1)^{k+l}}{k!\, l!} \frac{(a_1)_{k+i} (a_2)_{l+j}}{(a_1 + a_2 + a_3)_{k+l+i+j}} . \quad (2.8)$$

(b) *For cases where $a_1 = a_2 = 1$, the distribution of $U_1 + U_2$ is:* **(i)** *a Poisson mixture with $U_1 + U_2 | W \sim Poisson(W)$, $W \sim$ Beta$(2, a_3)$ for $a_3 > 0$; and* **(ii)** *Poisson(1) for $a_3 = 0$.*

Proof. **(a)** With the conditional Poisson representation of U and (2.7), we have $E[t_1^{U_1} t_2^{U_2}] = E^V \left(E[t_1^{U_1} t_2^{U_2} | V] \right) = E^V [e^{V_1(t_1-1)+V_2(t_2-1)}] = \psi_V(e^{t_1-1}, e^{t_2-1}) = \Phi_2(a_1, a_2, a_1+a_2+a_3, t_1 - 1, t_2 - 1)$.[1] The probability function is obtained with the help of (2.5) and (2.6).
(b) With the conditional Poisson representation, we have $U_1 + U_2 | V \sim$ Poisson$(V_1 + V_2)$ so that the result is immediate when $a_3 = 0$. For $a_3 > 0$, the result follows by verifying directly that $V_1 + V_2 \sim$ Beta$(2, a_3)$ whenever $(V_1, V_2) \sim$ Dirichlet$(1, 1, a_3)$. □

3 Distribution of the totals T_n and T

For studying the distribution of T_n, it suffices to consider the probability generating function $G_n(t_1, \ldots, t_r)$ of $\underline{S}_n$ evaluated at $t_1 = \cdots = t_r$. Simplifications will thus follow when applying Lemma 2.2. Moreover, in the particular cases where we have assumption (2.1), the components $S_{n,1}, \ldots, S_{n,r}$ are equidistributed for a given n, and the same is true for the $W_{n,j}$'s, $j = 1, \ldots, r$. As a consequence, the quantities $H_{n,j}(t, \ldots, t)$ will be, for fixed t and $n \geq 1$ constant in j, $j = 1, \ldots, r$. And this also leads to further simplifications when applying Lemma 2.2.

Our key finding, which we now proceed to describe, establishes a **link** between a multivariate problem with $r \geq 2$ and an univariate problem where $r = 1$. This will be especially useful given the known results in the literature applicable to the distribution of T_n and T for $r = 1$.

[1] We note here the general relationship between the probability generating function of the Poisson mixture U with the moment generating function of the mixing variable V, which also illustrates that the Poisson mixtures in Definition 2.3 are identifiable.

Theorem 3.1. *Let $\psi^r_{n,p_1,p_2,\dots,p_{n+1}}(\cdot)$ be the probability generating function of T_n with assumption (2.1). Let $\psi^1_{n,rp_1,rp_2,\dots,rp_{n+1}}(\cdot)$ be the probability generating function of ${S_n}' = \sum_{k=1}^n X_k X_{k+1}$ where $X_k \sim^{ind.} Ber(rp_k), k \geq 1$.[2] Then, we have for all $n \geq 1$, $r \geq 2$:*

$$\psi^r_{n,p_1,p_2,\dots,p_{n+1}}(u) = \psi^1_{n,rp_1,rp_2,\dots,rp_{n+1}}(1+\frac{u-1}{r})\,;u\in[0,1]\,; \tag{3.1}$$

$$E\binom{T_n}{k} = E\binom{S'_n}{k}\frac{1}{r^k}\,;k\in\{0,1,2,\dots\}\,; \tag{3.2}$$

$$P(T_n=j) = \sum_{k=j}^{n}(-1)^{k-j}\binom{k}{j}E\binom{S'_n}{k}\frac{1}{r^k}\,;j\in\{0,1,\dots,n\}\,. \tag{3.3}$$

Furthermore, the above applies to $T=\lim_{n\to\infty}T_n$ by replacing T_n by T, ${S_n}'$ by $S'=\lim_{n\to\infty}{S'}_n$, and taking $n\to\infty$; for $u\in[-1,1]$ in (3.1) and with (3.2) and (3.3) applicable as long as the probability generating function of S' has radius of convergence about 1 greater than $1/r$.

Proof. (3.2) and (3.3) follow from (3.1), (2.3) and (2.4). There remains (3.1). As remarked upon above, we have for fixed t: $G_n(t,\dots,t)=\psi^r_{n,p_1,p_2,\dots,p_{n+1}}(t)=\psi_n(t)$ (say, for short), and $H_{n,j}(t,\dots,t)=h_n(t)$ (say) by virtue of assumption (2.1). Rewrite Lemma 2.2's system of equations as

$$\begin{aligned}\psi_n(t) &= \psi_{n-1}(t)+(rp_n)\,(rp_{n+1})\,\frac{t-1}{r}\,h_{n-1}(t)\,,\\ h_n(t) &= \psi_{n-1}(t)+rp_n\,\frac{t-1}{r}\,h_{n-1}(t)\,,\end{aligned}$$

for $n\geq 2$, and with the initial values $\psi_1(t)=1+rp_1p_2(t-1)$ and $h_1(t)=1+p_1(t-1)$. Now, we set $\psi_n(t)=a_n(\frac{t-1}{r})$, $h_n(t)=b_n(\frac{t-1}{r})$, and ${p_n}'=rp_n$, so that the above system of equations becomes

$$\begin{aligned}a_n(u) &= a_{n-1}(u)+{p_n}'\,{p_{n+1}}'\,u\,b_{n-1}(u)\\ b_n(u) &= a_{n-1}(u)+{p_n}'\,u\,b_{n-1}(u)\,,\end{aligned}$$

for $n\geq 2$, $u\in\mathbb{R}$. The last two systems of equations tell us that $\psi^r_{n,p_1,p_2,\dots,p_{n+1}}(ru+1)=a_n(u)=\psi^1_{n,rp_1,rp_2,\dots,rp_{n+1}}(u+1)$, for all $n\geq 2$ and the result follows. Finally, results carry over to T in the same manner by making use of (2.3) and (2.4), and with a radius of convergence greater than $1/r$ for the pgf of S' implies a radius of convergence greater than 1 for the pgf of T. □

Theorem 3.1 describes a powerful relationship between our $r\geq 2$ problem of identifying the distribution of T_n, or of T, and a corresponding univariate or $r=1$ problem for which there exists already a certain number of results in the literature. We conclude this section with applications of Theorem 3.1.

Example 3.2. *(Constant case $p_k=p$) For the constant case $p_k=p$, $p\leq\frac{1}{r}$, we have ${p'}_k=rp$ and (e.g., Hirano et al, 1991; Holst, 2008)*

$$E\binom{{S'}_n}{k} = (rp)^k\sum_{i=1}^{k}\binom{k-1}{k-i}\binom{n+1-k}{i}(rp)^i\,,$$

[2] Note that we must have $rp_k\leq 1$ for all $k\in\{1,\dots,n+1\}$ given (2.1).

for $k \geq 1$, with (3.2) and (3.3) yielding the binomial moments and probability mass function of T_n. For instance, we obtain for $j = 0, 1, \ldots, n$,

$$P(T_n = j) = \sum_{k=j}^{n} (-1)^{k-j} \binom{k}{j} p^k \{\mathbb{I}_{\{k=0\}} + (\sum_{i=1}^{k} \binom{k-1}{k-i} \binom{n+1-k}{i} (rp)^i) \mathbb{I}_{\{k \geq 1\}}\} .$$

To conclude, observe that for $p = \frac{1}{r}$, we have $p'_k = 1$ so that $P(S'_n = n) = 1$ and $E(t^{S'_n}) = t^n$. Theorem 3.1 still applies and (3.1) yielding $E(u^{T_n}) = (1 + \frac{u-1}{r})^n$, i.e., $T_n \sim$ Bin$(n, \frac{1}{r})$. This serves more as an illustration as the result here for the distribution of T_n follows at once from the representation $T_n = \sum_{k=1}^{n} I_{\{\underline{X}_k = \underline{X}_{k+1}\}}$ with the indicator variables $I_{\{\underline{X}_k = \underline{X}_{k+1}\}}$ independently distributed as Bernoulli$(\frac{1}{r})$.

Example 3.3. *(Case where $p_k = \frac{a}{r(a+b+k-1)}$)*
In the setup for T_n with assumption (2.1), consider cases where $p_k = \frac{a}{r(a+b+k-1)}$ with $a > 0, b \geq 0$. The analysis for general a, b will cover many interesting particular cases which we will point out below. First, following Theorem 3.1, we consider the univariate sequence S'_n with $p'_k = rp_k = \frac{a}{a+b+k-1}$. From Holst (2008), we have for $k \in \{1, \ldots, n\}$:

$$E\binom{S'_n}{k} = \frac{a^k}{(a+b+n)^{\underline{k}}} \sum_{j=1}^{k} \binom{k-1}{k-j} \binom{n+1-k}{j} \frac{(a)_j}{(a+b)_j} , \tag{3.4}$$

$$and \;\; E\binom{S'}{k} = \frac{a^k}{k!} \frac{(a)_k}{(a+b)_k} . \tag{3.5}$$

Theorem 3.1 along with expressions (2.3) and (2.4) yield immediate expressions for the binomial moments, the probability generating function and the probability mass function of T_n through the binomial moment identity (3.2). Similarly, for the distribution of T, we obtain

$$\begin{aligned} E\binom{T}{k} &= \frac{a^k}{r^k\, k!} \frac{(a)_k}{(a+b)_k} , \;\; k = 0, 1, \ldots ; \\ E(u^T) &= {}_1F_1(a, a+b; \frac{a(u-1)}{r}) ; \\ P(T = j) &= \sum_{k \geq j} (-1)^{k-j} \binom{k}{j} \frac{a^k}{r^k\, k!} \frac{(a)_k}{(a+b)_k} . \end{aligned}$$

Moreover, it is straightforward to verify the following representation from the above, which constitutes a multivariate ($r > 1$) generalization of (1.1).

Corollary 3.4. *For cases where $p_k = \frac{a}{r(a+b+k-1)}$ with $a > 0, b \geq 0$, the distribution of T admits the following Poisson mixture representation:*

$$T|L = l \sim Poisson(\frac{a\, l}{r}) , \;\; L \sim \text{Beta}(a, b) . \tag{3.6}$$

We signal the following further applications.

- *(I) With $b = 0$ and $a = r\lambda$; $\lambda > 0$; , i.e., $p_k = \frac{\lambda}{\lambda\, r+k-1}$, we obtain that T has a Poisson distribution with mean equal to λ. When $r = 1$, this corresponds to result (1.1) with $b = 0$.*
- *(II) For the distributions of T_n and T with the configuration $p_k = \frac{a'}{k-1+rb'}$ with $b' \geq a'$, the results above also apply by taking $a = ra'$ and $b = r(b' - a')$. Corollary 3.4 hence yields the representation $T|L = l \sim$ Poisson$(a'\, l)$, $L \sim$ Beta$(ra', r(b' - a'))$, for such p_k's.*

4 Distributions of $\underline{S}_n$ and $\underline{S}$: bivariate case with $p_k = \frac{1}{b+k}$

In this section, we obtain the probability generating functions of $\underline{S}_n\,, n \geq 1$, and $\underline{S}$ in the bivariate case ($r = 2$) with $p_k = \frac{1}{b+k}$, $b \geq 1$. Moreover, by taking $n \to \infty$ and by making use of a representation (Lemma 4.2) for the pgf of $\underline{S}$ in terms of the marginal binomial moments, we arrive at explicit forms for the probability generating and mass functions of $\underline{S}$, as well as mixture representations (Theorem 4.3). This is achieved by first solving the recurrence given in Lemma 2.2 yielding explicit expressions for the probability generating functions $G_n, H_{n+1,1}$ and $H_{n+1,2}$ for $n \geq 1$ (Theorem 4.1).

Theorem 4.1. *Under assumption (2.1) with $r = 2$ and $p_k = \frac{1}{b+k}$, the probability generating functions $G_n(\underline{t}) = E[t_1^{S_{n,1}}\, t_2^{S_{n,2}}]$, $H_{n+1,1}(\underline{t}) = E[t_1^{W_{n+1,1}}\, t_2^{S_{n,2}}]$, and $H_{n+1,2}(\underline{t}) = E[t_1^{S_{n,1}}\, t_2^{W_{n+1,2}}]$ are given by*

$$\begin{aligned} G_n(\underline{t}) &= \frac{b}{s_1 - s_2} \sum_{k\geq 1} [\mathbb{I}(k \leq n+1) - \frac{s_1 + s_2}{n+1+b} \mathbb{I}(k \leq n) \\ &+ \frac{s_1 s_2}{(b+n)(b+n+1)} \mathbb{I}(k \leq n-1)] \frac{s_1^k - s_2^k}{(b)_k} \\ H_{n+1\,j}(\underline{t}) &= \frac{s_j^{n+1}}{(1+b)_{n+1}} + \frac{b}{s_1 - s_2} \sum_{k\geq 1} [\mathbb{I}(k \leq n+1) - \frac{s_{3-j}}{n+1+b} \mathbb{I}(k \leq n)] \frac{s_1^k - s_2^k}{(b)_k}\,, \end{aligned}$$

for $s_1 \neq s_2$, $j = 1, 2$, $n \geq 1$, and $s_i = t_i - 1$.

Proof. We proceed by induction. A direct evaluation yields $G_1(t_1, t_2) = 1 + \frac{1}{(1+b)_2}(s_1 + s_2)$ and $H_{2,j}(t_1, t_2) = 1 + \frac{1}{(1+b)} s_j + \frac{1}{(1+b)_2} s_j^2 + \frac{1}{(1+b)_2} s_{3-j}, \quad j = 1, 2$, which matches the given formulas for $n = 1$. Now suppose the above formulas hold for $n = 1, \ldots, m$. A slight reorganization of Lemma 2.2 tells us that

$$\begin{aligned} G_{n+1}(\underline{t}) &= p_{n+2,3}\, G_n(t) + \sum_{j=1}^{2} p_{n+2,j}\, H_{n+1,j}(\underline{t}) \\ H_{n+1,\,j}(\underline{t}) &= G_n(\underline{t}) + p_{n+1,j}\, s_j H_{n,j}(\underline{t}), \quad j = 1, 2, \end{aligned}$$

for $n \geq 1$, recalling that by definition $p_{n+2,3} = 1 - p_{n+2,1} - p_{n+2,2} = 1 - \frac{2}{b+n+2}$. With the following decomposition, which is verified directly,

$$\frac{s_j^{m+1}}{(b+1)_{m+2}} = \frac{b}{s_j - s_{3-j}} \Big[\frac{s_j^{m+2}}{(b)_{m+2}} - \frac{s_1 + s_2}{m+b+2} \frac{s_j^{m+1}}{(b)_{m+1}} + \frac{s_1 s_2}{(m+b+1)(m+b+2)} \frac{s_j^m}{(b)_m} \Big]\,,$$

a calculation of $p_{m+2,3} G_m + \sum_{j=1}^{2} p_{m+2,j} H_{m+1,j}$ yields the desired expression for G_{m+1}. Similarly, an evaluation of $G_{m+1} + p_{m+2,j} s_j H_{m+1,j}$ leads to the desired expression for H_{m+2}. □

Lemma 4.2. *Suppose ψ is the probability generating function of a random vector (Z_1, Z_2) on $\mathbb{N}^2$ satisfying the equation*

$$\psi(t_1, t_2)(t_1 - t_2) \;=\; (t_1 - 1)\, \psi(t_1, 1) \;-\; (t_2 - 1)\, \psi(1, t_2)\,, \tag{4.1}$$

$(t_1, t_2) \in [-1, 1]^2$, and that the Taylor series development at $(1, 1)$ of ψ converges on an open set containing the origin. Then, ψ is given by

$$\psi(t_1, t_2) \;=\; \sum_{k,l \geq 0} E\binom{Z_1}{k+l} (t_1 - 1)^k\, (t_2 - 1)^l\,. \tag{4.2}$$

Proof. With the series representation (2.5) and the uniqueness of the coefficients, it suffices to show that

$$E\binom{Z_1}{k}\binom{Z_2}{l} = E\binom{Z_1}{k+l}, \text{ for all } k,l \geq 0\,. \tag{4.3}$$

Now, equation (4.1) implies, for all (t_1,t_2), with $c_{k,l} = E\binom{Z_1}{k}\binom{Z_2}{l}$ and $s_i = t_i - 1$,

$$\begin{aligned}\sum_{k,l\geq 0} c_{k,l}\, s_1^{k+1}\, s_2^{l} - \sum_{k,l\geq 0} c_{k,l}\, s_1^{k}\, s_2^{l+1} &= \sum_{k\geq 0} c_{k,0} s_1^{k+1} - \sum_{l\geq 0} c_{0,l} s_2^{l+1} \\ \Longrightarrow \sum_{k,l\geq 1} c_{k-1,l}\, s_1^{k}\, s_2^{l} &= \sum_{k,l\geq 1} c_{k,l-1}\, s_1^{k}\, s_2^{l}\,.\end{aligned}$$

But the above is equivalent to $c_{k-1,l} = c_{k,l-1}$ for all $k,l \geq 1$, which implies $c_{0,k+l-1} = \cdots = c_{k-1,l} = c_{k,l-1} = \cdots = c_{k+l-1,0}$ for all $k,l \geq 1$, which is (4.3). □

The key result that follows concerns the limiting distribution of $\underline{S}_n$ for the homogeneous bivariate case $p_{k,1} = p_{k,2} = \frac{1}{b+k}$. A first explicit form (equation 4.8) for the probability generating function is easily derived from Theorem 4.1. A second explicit form is obtained via Lemma 4.2 by verifying directly that the probability generating function of $\underline{S}$ verifies (4.1). This permits to write down the probability generating function of $\underline{S}$ in terms of the binomial moments of S_1, that either can be derived from our expressions or taken from known results in the univariate case.

Theorem 4.3. *Under assumption (2.1) with $r = 2$ and with $p_k = \frac{1}{b+k}$, $b \geq 1$,*

(a) *The probability generating and probability mass functions of $\underline{S}$ are given by*

$$G(t_1,t_2) = \Phi_2(1,1,b+1,t_1-1,t_2-1)\,, \tag{4.4}$$

$$\mathbb{P}\left(S_1 = x_1, S_2 = x_2\right) = \frac{1}{x_1!\,x_2!}\sum_{k,l\geq 0}\frac{(-1)^{k+l}}{(b+1)_{k+l+x_1+x_2}}\,(k+1)_{x_1}\,(l+1)_{x_2}\,. \tag{4.5}$$

(b) *The distribution of $\underline{S}$ is a bivariate Poisson mixture, as in Definition 2.3 and Lemma 2.4, with $V \sim$ Dirichlet$\,(a_1 = 1, a_2 = 1, a_3 = b-1)$.*

(c) *The distribution of $\underline{S}$ admits the representation*

$$\underline{S}|\alpha \sim p_\alpha\,,\ \alpha \sim \text{ Beta}(1,b)\,, \tag{4.6}$$

with p_α the bivariate probability mass function on $\mathbb{N}^2$ given by

$$p_\alpha(s_1,s_2) = \frac{e^{-\alpha}\alpha^{s_1+s_2}}{(s_1+s_2+1)!}\,(s_1+s_2+1-\alpha)\,;\ \alpha \in (0,1]\,. \tag{4.7}$$

(d) *For $t \in \mathbb{N}$, the distribution of $S_1|S_1+S_2 = t$ is uniform on $\{0,1,\ldots,t\}$.*

Proof. **(a)** The probability function in (4.5) follows from (4.4) and part (a) of Lemma 2.4. For establishing (4.4), start with Theorem 4.1, where we obtain (for $t_1 \neq t_2$):

$$G(t_1,t_2) = E(t_1^{S_1}\,t_2^{S_2}) = \lim_{n\to\infty} G_n(t_1,t_2) = \frac{b}{t_1-t_2}\left({}_1F_1(1,b,t_1-1) - {}_1F_1(1,b,t_2-1)\right)\,. \tag{4.8}$$

From this expression, or alternatively from Holst (2008) and (2.3), we obtain

$$(t_1-1)G(t_1,1) = b\,\left[{}_1F_1(1,b,t_1-1)\,-\,1\right] \text{ and } (t_2-1)G(1,t_2) = b\,\left[{}_1F_1(1,b,t_2-1)\,-\,1\right].$$

Now, observe that $\psi \equiv G$ satisfies (4.1), so that (4.2) applies with $E\binom{S_1}{k+l} = \frac{1}{(b+1)_{k+l}}$; again derived from Holst (2008) or (4.8); yielding $G(t_1,t_2) = \sum_{k,l\geq 0} \frac{1}{(b+1)_{k+l}} (t_1-1)^k (t_2-1)^l = \Phi_2(1,1,b+1,t_1-1,t_2-1)$.
(b) Part **(a)** paired with Lemma 2.4 imply the given representation.
(c) Given that the probability generating function of $\underline{S}$ is necessarily expressible as in (2.5), it will suffice given part **(a)** to show that

$$\mathbb{E}\binom{S_1}{k}\binom{S_2}{l} = \frac{1}{(b+1)_{k+l}}, \tag{4.9}$$

under representation (4.6)-(4.7). In turn, it will suffice to show that the mixed binomial moments of (S_1,S_2) under probability function p_α are given by

$$\mathbb{E}\left[\binom{S_1}{k}\binom{S_2}{l}|\alpha\right] = \frac{\alpha^{k+l}}{(k+l)!}\,, \text{ for non negative integers } k,l, \tag{4.10}$$

since this would imply along with representation (4.6) that $\mathbb{E}\left[\binom{S_1}{k}\binom{S_2}{l}\right] = \mathbb{E}(\frac{\alpha^{k+l}}{(k+l)!}) = \frac{1}{(b+1)_{k+l}}$, which is (4.9). Finally, manipulations lead to (4.10) as follows:

$$\begin{aligned}
\mathbb{E}\left[\binom{S_1}{k}\binom{S_2}{l}|\alpha\right] &= \sum_{s_1\geq 0, s_2\geq 0}\binom{s_1+k}{k}\binom{s_2+l}{l} p_\alpha(s_1+k,s_2+l) \\
&= \sum_{y\geq 0}\frac{e^{-\alpha}\,\alpha^{y+k+l}}{(y+k+l+1)!}(y+l+k+1-\alpha)\sum_{x=0}^{y}\binom{x+k}{k}\binom{y-x+l}{l} \\
&= \sum_{y\geq 0}\frac{e^{-\alpha}\,\alpha^{y+k+l}}{(y+k+l+1)!}(y+l+k+1-\alpha)\binom{y+l+k+1}{y} \\
&= \frac{e^{-\alpha}\,\alpha^{k+l}}{(k+l+1)!}\sum_{y\geq 0}\frac{\alpha^y}{y!}\,(y+k+l+1-\alpha) = \frac{\alpha^{k+l}}{(k+l)!}\,.
\end{aligned}$$

(d) It suffices to show that $P(S_1=s_1,S_2=s_2)$ is a function of s_1+s_2, $s_1,s_2\geq 0$. From part (c), we have $P(S_1=s_1,S_2=s_2) = \int_0^1 p_\alpha(s_1,s_2)\,b\,(1-\alpha)^{b-1}\,d\alpha$, with p_α given in (4.7), and which indeed depends on (s_1,s_2) only through the sum s_1+s_2. □

The bivariate Poisson mixture representation of $\underline{S}$ extends in a most interesting way the known marginal distribution representation for S_1 and S_2 (i.e., (1.1)) expressible as a Beta mixture of Poisson distributions. Here S_1 and S_2 are clearly dependent but the representation tells us that they are conditionally independent and the dependence is reflected through the dependence of the mixing components of the Dirichlet. Similarly, we obtain from part **(b)** of the Theorem and Lemma 2.4 the mixture representation

$$T|W=w \sim \text{ Poisson}(w)\,,\; W\sim \text{Beta}(2,b-1)\,,$$

with this, alternatively, following also from Corollary 3.4. In contrast to the Dirichlet mixing (when $b>1$), the dependence in representation **(d)** is reflected through the conditional distributions of (S_1,S_2), and the mixing variable α is univariate. Furthermore, it is readily verified that the conditional marginal distributions of $S_i|\alpha$ are Poisson(α), which is consistent with the univariate result in (1.1). We conclude with some observations on the probability functions p_α in (4.7).

Remark 4.4. *The bivariate probability function in (4.7) has a simple enough form so that it possibly has arisen in previous work, but we cannot identify such a source. Anyhow, it is most interesting that it arises here in a natural way from the Bernoulli array in*

the representation of $\underline{S}$ for $r = 2$ and the configuration $p_k = \frac{1}{b+k}$, $b \geq 1$. The probability generating function, using the binomial moments at the end of the proof of Theorem 4.3 and (2.5), may be written as

$$\psi_\alpha(t_1, t_2) = E_\alpha(t_1^{S_1}\, t_2^{S_2}) \;=\; \sum_{k,l \geq 0} \alpha^{k+l} \frac{(t_1 - 1)^k\, (t_2 - 1)^l}{(k + l)!}\,.$$

As seen above, the marginals are Poisson(α) distributed. These distributions, as expanded on by Ait Aoudia and Marchand (2013), possess at least two other interesting properties:

(i) the distribution of $S_1 + S_2$ (conditional on α) is given by the convolution of a Poisson(α) with a Bernoulli(α);

(ii) the correlation coefficient between S_1 and S_2 (conditional on α) is equal to $-\frac{\alpha}{2}$.

Concluding Remarks

The main findings in this paper concern the numbers of runs of length 2 in Bernoulli arrays with independently distributed multinomial distributed rows and identically distributed row components. Exploiting the structure of the problem through a recurrence involving probability generating functions and building on known marginal distribution results in the literature, we have explored the distributions of totals across columns and joint distributions of column sums. Elegant representations have been obtained: **(i)** through Section 3's correspondence between multivariate and univariate problems to describe the distribution of a total, and **(ii)** with Section 4's bivariate distributions, where we have for instance obtained in a specific situation a bivariate Poisson mixture with a Dirichlet mixing parameter. Many other open and interesting problems can be envisioned. These include an analysis of the distributions of $\underline{S}_n$ and $\underline{S}$ for $r > 2$, closed form distributional results in the absence of assumption (2.1), and an extended framework for probability models other than multinomial.

Acknowledgments

We are indebted and thankful to an anonymous reviewer for constructive comments and for identifying a mistake in an earlier version of the manuscript. Éric Marchand and François Perron gratefully acknowledge the support for research provided by NSERC of Canada.

References

[1] Ait Aoudia, D. and Marchand É. (2014). On a simple construction of bivariate probability functions with fixed marginals. Technical report 136 `http://www.usherbrooke.ca/mathematiques/recherche/publications/rapports-recherche/`

[2] Ait Aoudia, D. and Marchand É. (2010). On the number of runs for Bernoulli arrays. *Journal of Applied Probability*, **47**, 367-377. MR-2668494

[3] Arratia, R., Barbour, A.D., and Tavaré, S. (1992). Poisson process approximations for the Ewens sampling formula. *Annals of Applied Probability*, **2**, 519-535. MR-1177897

[4] Chern, H.-H., Hwang, H.-K., and Yeh, Y.-N. (2000). Distribution of the number of consecutive records. *Random Structures Algorithms*, **17**, 169-196. MR-1801131

[5] Csörgó, M. and Wu, B.W. (2000). On sums of overlapping products of independent Bernoulli random variables. *Ukrainian Matematical Journal*, **52**, 1304-1309. MR-1816943

[6] Goncharov, V. (1944). On the field of combinatory analysis. *Soviet Math. Izv., Ser. Math.*, **8**, 3-48. In Russian.

[7] Hahlin, L.O. (1995). Double Records. *Research Report # 12*, Department of Mathematics, Uppsala University.

[8] Hirano, K., Aki, S., Kashiwagi, N., and Kuboki, H. (1991). On Ling's binomial and negative binomial distributions of order k. *Statistics & Probability Letters*, **11**, 503-509. MR-1116744

[9] Holst, L. (2008). The number of two-consecutive successes in a Hoppe-Polyá urn. *Journal of Applied Probability*, **45**, 901-906. MR-2455191

[10] Holst, L. (2007). Counts of failure strings in certain Bernoulli sequences. *Journal of Applied Probability*, **44**, 824-830. MR-2355594

[11] Huffer, W. F., Sethuraman, J. and Sethuraman, S. (2009). A study of counts of Bernoulli strings via conditional Poisson processes. *Proceedings of the American Mathematical Society*, **137**, 2125-2134. MR-2480294

[12] Joffe, A., Marchand É., Perron, F., and Popadiuk, P. (2004). On sums of products of Bernoulli variables and random permutations. *Journal of Theoretical Probability*, **17**, 285-292. MR-2054589

[13] Joffe, A., Marchand É., Perron, F., and Popadiuk, P. (2000). On sums of products of Bernoulli variables and random permutations. *Research Report # 2686*, Centre de Recherches Mathématiques, Montréal, Canada.

[14] Kolchin, V.F. (1971). A problem of the allocation of particles in cells and cycles of random permutations. *Theory of Probability and its Applications*, **16**, 74-90.

[15] Lee, P.A. (1971). A diagonal expansion for the 2-variate Dirichlet probability density function. *SIAM Journal on Applied Mathematics*, **21**, 155-165. MR-0288805

[16] Mori, T.F. (2001). On the distribution of sums of overlapping products. *Acta Scientiarum Mathematica (Szeged)*, **67**, 833-841. MR-1876470

[17] Sethuraman, J. and Sethuraman, S. (2004). On counts of Bernoulli strings and connections to rank orders and random permutations. *A Festschrift for Herman Rubin. IMS Lecture Note Monograph Series*, **45**, Institute of Mathematical Statistics, pp. 140-152. MR-2126893