On the number of runs for Bernoulli arrays

Djilali Ait Aoudia

Éric Marchand [1]

Université de Sherbrooke, Département de mathématiques, Sherbrooke, QC, CANADA, J1K 2R1

Abstract

We introduce and motivate the study of $(n+1)\times r$ arrays X with Bernoulli entries $X_{k,j}$ and independently distributed rows. We study the distribution of $S_n = \sum_{j=1}^{r}\sum_{k=1}^{n} X_{k,j}X_{k+1,j}$ representing the number of consecutive pairs of successes (or runs of length 2) when reading the array down the columns and across the rows. With the case $r = 1$ having been studied by several authors, and permitting some initial inferences for the general case $r > 1$, we pursue by obtaining various distributional properties and representations of S_n for the case $r = 2$, and with a more explicit analysis for the case of multinomial and identically distributed rows. Applications are also given in cases where the array the array X arises from a Pólya sampling scheme.

AMS 2000 subject classifications: 60C05, 60E05, 60K99.

Keywords and phrases: Bernoulli; multinomial; Pólya urn; probability generating function, runs.

1. Introduction

For an array X of Bernoulli random variables $X_{k,j}; k \geqslant 1$ and $j = 1,\ldots,r$; such that the random vectors $\underline{X}_k = (X_{k,1},\ldots,X_{k,r})'$ are independent, we study the distributional properties of $S_n = \sum_{k=1}^{n} \underline{X}_k'\underline{X}_{k+1}$, $n \geq 1$, which counts the number of pairs of consecutive Bernoulli successes (or runs of length 2) in the array X reading down the lines and across the columns. Alternatively, we may

[1]Corresponding author: eric.marchand@usherbrooke.ca

write

$$S_n = \sum_{j=1}^{r} \sum_{k=1}^{n} X_{k,j} X_{k+1,j} = \sum_{j=1}^{r} Z_j \tag{1}$$

with

$$Z_j = \sum_{k=1}^{n} X_{k,j} X_{k+1,j}. \tag{2}$$

As an illustration, the array

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \end{pmatrix}$$

where $n = 3$, $r = 5$, yields $Z_1 = 1, Z_2 = 0, Z_3 = 0, Z_4 = 2, Z_5 = 1$, and $S_3 = \sum_{j=1}^{5} Z_j = 4$.

Example 1 *A class of applications where S_n arises consists of $n+1$ draws with replacement from an urn with r columns of cardinalities $\alpha_1, \cdots, \alpha_r$ $(\alpha_i > 0)$, where we define*

$$X_{k,j} = \mathbf{I}_{\{kth \ \ draw \ \ is \ \ of \ \ type \ \ j\}}, \qquad 1 \leq k \leq n, \ \ 1 \leq j \leq r. \tag{3}$$

We thus have $\underline{X}_k \sim Multinomial(1, \theta_{k,1}, \ldots, \theta_{k,r})$, with $\theta_{k,j} = \alpha_j / \sum_{j=1}^{r} \alpha_j$, and S_n counting the number of pairs of consecutive draws with matching colours. Observe that the independence of the vectors $\underline{X}_k$ is a consequence of the independent draws with replacement. In this illustration, the random vectors $\underline{X}_k$ are also identically distributed but this need not be the case as we can envisage non-identical draws such as single draws from several distinct non-homogeneous urns.

Remark 1 *Our setup does not directly encompass Pólya urn sampling schemes where the rows of X are dependent. However, by de Finetti's representation theorem, the study of Example 1's class of applications, along with the results of this paper, do indeed lead to implications for Pólya urns (for $r = 2$), as expanded upon in Section 3.*

More generally, the spectrum of distribution of S_n is governed, of course, by the possible distributions for the random matrix X, subject to the rows $\underline{X}_1, \ldots, \underline{X}_{n+1}$, being independent, but not necessarily identically distributed. The univariate column case (i.e., $r = 1$ or as pertaining to the marginal distribution of Z_j) has generated much recent interest, analysis, and interpretation, as witnessed by the contributions of Holst (2008, 2007), Joffe et al., (2004), Mori (2001), Huffer, Sethuraman and Sethuraman (2009), Sethuraman and Sethuraman (2004), among others. In particular, for $r = 1$ and $P(X_{k,1} = 1) = a/(a + b + k - 1)$ with $a > 0, b \geq 0$, the limiting distribution of S_n admits the representation:

$$S|L \sim Poisson(aL) \text{ with } L \sim Beta(a, b), \tag{4}$$

(Holst, 2008; Joffe et al., 2004, $a = 1$; Mori 2001, $b = 0$), while the constant case $P(X_{k,1} = 1) = p$ relates to type **II** Binomial distributions (see Joffe et al., 2004 and the references therein). Even though the marginal distributions of $Z_1, \ldots, Z_r$ are known in the above cases, the difficulty of studying the distribution of S_n, as given in (1), is of course that the columns of X, and consequently $Z_1, \ldots, Z_r$, are not assumed to be independent. Nevertheless, it is worthwhile first summarizing results that may be inferred in cases where the columns are either independent, or in perfect dependence.

Example 2 *(independent columns and $P(X_{k,j} = 1) = a_j/(a_j + b_j + k - 1)$)*

Whenever the columns are independent, representation (1) implies that an efficient analysis of S_n passes through the convolution of r independent one dimensional problems. For instance, whenever the $X_{k,j}$'s, $k \geq 1, j = 1, \ldots, r$, are independent, and $P(X_{k,j} = 1) = a_j/(a_j + b_j + k - 1)$, the distribution of $S = \lim_{n\to\infty} S_n$ admits the following representation:

$$S|Y \sim Poisson(Y) \qquad with \qquad Y = \sum_{j=1}^{r} a_j L_j,$$

and $L_1, \cdots, L_r$ independent random variables such that $L_j \sim Beta(a_j, b_j)$. This may be established by exploiting representations (1) and (4), the independence of the Z_j's, so that for all $t \in \Re$

$$\begin{aligned} E[t^S] &= E[t^{\sum_{j=1}^n Z_j}] = \prod_{j=1}^r E[t^{Z_j}] \\ &= \prod_{j=1}^r E\left[E[t^{Z_j}]|L_j\right] = \prod_{j=1}^r E[e^{a_j L_j (t-1)}] \\ &= E[e^{\sum_{j=1}^r a_j L_j (t-1)}] = E[e^{Y(t-1)}]\,. \end{aligned}$$

Example 3 *Whenever the $X_{k,j}$'s, $j = 1, \cdots, r$, are equal with probability one (i.e., $P(\underline{X}_k = (1, \cdots, 1)) + P(\underline{X}_k = (0, \cdots, 0)) = 1$) for all $k \in \{1, \cdots, n+1\}$), it follows that $P(Z_1 = \cdots = Z_r) = 1$ and hence $\mathcal{L}(S_n) = \mathcal{L}(rZ_1)$. In such cases, univariate results, such as (4) apply directly.*

The findings that follow in this paper relate to the bidimensional case ($r = 2$). In section 2, we give a general recurrence (Lemma 1) for the probability generating function (pgf) of S_n. We then obtain more explicit representations (Corollary 1) for the particular case of multinomial rows (as in Example 1), which will lead to an explicit expression (Theorem 1) for the probability mass function of S_n in the identically distributed case. The family of distributions thus obtained is quite interesting on its own and may be viewed as one that contains (for fixed n) the Binomial$(n, 1/2)$ distribution. These distributions also admit a large n normal approximation (Remark 5). Via de Finetti's representation theorem, the last results of Section 2 permit us to obtain useful representations for the distribution of S_n when arising from a Pólya urn sampling scheme. This development is presented in Section 3.

2. The bidimensional case ($r = 2$)

We begin by analyzing the general bidimensional case ($r = 2$) where

$$f_{i,j}^{(k)} = P\left(\underline{X}_k = (i, j)\right), \qquad 1 \le k \le n+1, \qquad i, j = 0, 1\,.$$

Define S_n as in (1) above with $S_0 = 0$ in other words

$$S_n = S_{n-1} + X_{n,1}X_{n+1,1} + X_{n,2}X_{n+1,2}\,;\; n \geq 1. \tag{5}$$

To derive a useful expression for the probability generating function (pgf) $\varphi_{S_n}(t)(= E[t^{S_n}])$ of S_n, we introduce the auxiliary random variables $W_{l,n}; l = 1, 2, 3;$ defined for $n \geq 0$ as

$$W_{1,n} := S_n + X_{n+1,1} + X_{n+1,2}\,, \qquad W_{2,n} := S_n + X_{n+1,1}\,, \qquad W_{3,n} := S_n + X_{n+1,2}\,. \tag{6}$$

We denote their pgf's as $\varphi_{W_{l,n}}$, and we also write

$$\underline{\varphi}_n(\cdot) = \left(\varphi_{S_n}(\cdot), \varphi_{W_{1,n}}(\cdot), \varphi_{W_{2,n}}(\cdot), \varphi_{W_{3,n}}(\cdot)\right)'\,;\; n \geq 0.$$

The next result establishes an explicit recurrence and expression for $\underline{\varphi}_n(\cdot)\,,\; n \geq 1\,.$

Lemma 1 *We have for $n \geq 1, t \geq 0$:*

$$\underline{\varphi}_n(t) = M_{n+1}\ \underline{\varphi}_{n-1}(t) \tag{7}$$

and

$$\underline{\varphi}_n(t) = M_{n+1}\cdots M_2\underline{\varphi}_0(t) = M_{n+1}\cdots M_1\mathbf{1}, \tag{8}$$

with

$$M_n = \begin{pmatrix} f_{0,0}^{(n)} & f_{1,1}^{(n)} & f_{1,0}^{(n)} & f_{0,1}^{(n)} \\ f_{0,0}^{(n)} & t^2 f_{1,1}^{(n)} & t f_{1,0}^{(n)} & t f_{0,1}^{(n)} \\ f_{0,0}^{(n)} & t f_{1,1}^{(n)} & t f_{1,0}^{(n)} & f_{0,1}^{(n)} \\ f_{0,0}^{(n)} & t f_{1,1}^{(n)} & f_{1,0}^{(n)} & t f_{0,1}^{(n)} \end{pmatrix}$$

and $\mathbf{1} = (1, 1, 1, 1)'$.

Proof. We condition on $\underline{X}_{n+1}$. For $S_n, n \geq 1$, we obtain from (5) and (6) and the independence of the $\underline{X}_k$'s:

$$\mathcal{L}\left(S_n \mid \underline{X}_{n+1} = (1,1)\right) = \mathcal{L}\left(S_{n-1} + X_{n,1} + X_{n,2}\right) = \mathcal{L}\left(W_{1,n-1}\right),$$

$$\mathcal{L}\left(S_n \mid \underline{X}_{n+1} = (1,0)\right) = \mathcal{L}\left(S_{n-1} + X_{n,1}\right) = \mathcal{L}\left(W_{2,n-1}\right),$$

$$\mathcal{L}\left(S_n \mid \underline{X}_{n+1} = (0,1)\right) = \mathcal{L}\left(S_{n-1} + X_{n,2}\right) = \mathcal{L}\left(W_{3,n-1}\right),$$

$$\text{and } \mathcal{L}\left(S_n \mid \underline{X}_{n+1} = (0,0)\right) = \mathcal{L}\left(S_{n-1}\right).$$

Since

$$\varphi_{S_n}(t) = E\left[E\left[t^{S_n} | \underline{X}_{n+1}\right]\right],$$

the above translates to

$$\varphi_{S_n}(t) = \left(f_{0,0}^{(n+1)}, f_{1,1}^{(n+1)}, f_{1,0}^{(n+1)}, f_{0,1}^{(n+1)}\right) \underline{\varphi}_{n-1}(t),$$

(i.e., the scalar product of the first row of M_{n+1} and $\underline{\varphi}_{n-1}(t)$). The rest of the system is obtained along the same lines. Finally (8) follows from (7), with $\underline{\varphi}_0(t)$ derived directly as $M_1\mathbf{1}$ from the definitions of $W_{l,0}$ in (6), and since $S_0 = 0$. □

We now pursue with further analysis for multinomial distributed rows $\underline{X}_k$, where, for all $k \in \{1, \ldots, n+1\}$,

$$f_{1,0}^{(k)} = 1 - f_{0,1}^{(k)} = p_k \text{ (say)}. \tag{9}$$

This corresponds to non-homogeneous (or non-identically distributed) draws in Example 1. We do not assume for the time being that the $\underline{X}_k$'s are identically distributed, but the more explicit results that follow later (e.g., part (b) of Corollary 1) do apply to cases where

$$f_{1,0}^{(k)} = 1 - f_{0,1}^{(k)} = p, \tag{10}$$

for all $k \in \{1, \ldots, n+1\}$. We require the following result.

Lemma 2 *If A is a 2×2 matrix with distinct eigenvalues λ_1 and λ_2 and I_2 is the 2×2 identity matrix, then for all $n \geq 2$*

$$A^n = \left(\frac{\lambda_2^n - \lambda_1^n}{\lambda_2 - \lambda_1}\right) A - \lambda_1\lambda_2 \left(\frac{\lambda_2^{n-1} - \lambda_1^{n-1}}{\lambda_2 - \lambda_1}\right) I_2\,. \tag{11}$$

Proof. See for instance Ardakov (1997).

The next result consists of specializations of Lemma 1 in cases where (9) or (10) hold. In such cases with the two first columns of M_n becoming 0 vectors, we obtain a useful and more explicit representation for the pgf of S_n.

Corollary 1 **(a)** *Under assumption (9), we have for all* $n \geq 1, t \geq 0$*:*

$$\varphi_{S_n}(t) = p_{n+1}\varphi_{W_{2,n-1}}(t) + (1 - p_{n+1})\varphi_{W_{3,n-1}}(t) \tag{12}$$

with

$$\begin{bmatrix} \varphi_{W_{2,n}}(t) \\ \varphi_{W_{3,n}}(t) \end{bmatrix} = \begin{bmatrix} tp_{n+1} & 1 - p_{n+1} \\ p_{n+1} & t(1 - p_{n+1}) \end{bmatrix} \begin{bmatrix} \varphi_{W_{2,n-1}}(t) \\ \varphi_{W_{3,n-1}}(t) \end{bmatrix}, \tag{13}$$

and $(\varphi_{W_{2,0}}(t), \varphi_{W_{3,0}}(t))' = (p_1 t + (1 - p_1), p_1 + (1 - p_1)t)'$.

(b) *For all* $p \in [0, 1], n \geq 1, t \geq 0$*, set*

$$\lambda_1 = (t + \sqrt{t^2 - 4p(1-p)(t^2-1)})/2, \quad \lambda_2 = t - \lambda_1, \quad \alpha_n = \frac{\lambda_1^n - \lambda_2^n}{\lambda_1 - \lambda_2}.$$

Under assumption (10)*, we have for all* $n \geq 1, t \geq 0$*:*

$$\varphi_{S_n}(t) = \alpha_n \left[2p(1-p) + t(1 - 2p(1-p))\right] + \alpha_{n-1} p(1-p)(1-t^2). \tag{14}$$

Proof. **(a)** The result follows directly from Lemma 1, and since $W_{2,0} = X_{1,1}, W_{3,0} = X_{1,2}$.

(b) From part (a), we have for all $n \geq 1$ and under assumption (10):

$$\begin{bmatrix} \varphi_{W_{2,n-1}}(t) \\ \varphi_{W_{3,n-1}}(t) \end{bmatrix} = B^{n-1} \begin{bmatrix} \varphi_{W_{2,0}}(t) \\ \varphi_{W_{3,0}}(t) \end{bmatrix} = B^n \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

with

$$B = \begin{bmatrix} tp & 1-p \\ p & t(1-p) \end{bmatrix}.$$

Observe that λ_1 and λ_2 are the eigenvalues of B, so that

$$B^n = \alpha_n B - (t^2-1)p(1-p)\alpha_{n-1}\mathbf{I}_2,$$

by virtue of Lemma 2. From this, we obtain

$$\begin{bmatrix} \varphi_{W_{2,n-1}}(t) \\ \varphi_{W_{3,n-1}}(t) \end{bmatrix} = \alpha_n \begin{bmatrix} tp+1-p \\ p+t(1-p) \end{bmatrix} + \alpha_{n-1} \begin{bmatrix} (1-t^2)\,p(1-p) \\ (1-t^2)\,p(1-p) \end{bmatrix}.$$

and the result follows by applying (12). □

Before proceeding with an evaluating of the probability mass function of S_n, a couple of observations merit mention.

Remark 2 *(The probabilities $P(S_n = 0)$ and $P(S_n = n)$)*

S_n is supported on $\{0, 1, \ldots, n\}$ and it is easy to evaluate $P(S_n = n)$ directly, obtaining $P(S_n = n) = P\left(\bigcap_{k=1}^{n+1}\{X_{k,1} = 1\}\right) + P\left(\bigcap_{k=1}^{n+1}\{X_{k,1} = 0\}\right) = \prod_{k=1}^{n+1} p_k + \prod_{k=1}^{n+1}(1-p_k)$, under (9); and $P(S_n = n) = p^{n+1} + (1-p)^{n+1}$ under (10).

Consider now $\varphi_{S_n}(0) = P(S_n = 0)$. Under assumption (10), relationship (14) applies with $t = 0, \lambda_1 = -\lambda_2 = \sqrt{p(1-p)}$, and $\alpha_n = ((p(1-p))^{(n-1)/2}$ for odd n, and $\alpha_n = 0$ for even n. We therefore obtain from (14)

$$P(S_n = 0) = 2\,(p(1-p))^{(n+1)/2}\,\mathbf{I}_{\{n\ odd\}} + (p(1-p))^{n/2}\,\mathbf{I}_{\{n\ even\}}\,. \tag{15}$$

Alternatively, since the event $\{S_n = 0\}$ is equivalent to the event of alternating 0's and 1's in the first column of X, we may infer directly for odd n: $P(S_n = 0) = P(X_{1,1} = 1, X_{1,2} = 0, \ldots, X_{1,n} = 1, X_{1,n+1} = 0) + P(X_{1,1} = 0, X_{1,2} = 1, \ldots, X_{1,n} = 0, X_{1,n+1} = 1) = 2[p(1-p)]^{\frac{n+1}{2}}$; and similarly $P(S_n = 0) = [p(1-p)]^{\frac{n}{2}}$ as above in (15) for even n. For the non-homogeneous case in (9), we may proceed in an analogous manner to evaluate $P(S_n = 0)$ either directly as in (15), or by making use

of (13) with $t = 0$, obtaining

$$P(S_n = 0) = (1 - p_{n+1})^{[n \text{ even}]} \prod_{k=1}^{\lceil n/2 \rceil} p_{2k}(1 - p_{2k-1}) + (p_{n+1})^{[n \text{ even}]} \prod_{k=1}^{\lceil n/2 \rceil} p_{2k-1}(1 - p_{2k}),$$

where $\lceil x \rceil$ is the ceiling function given by $\lceil x \rceil = \min \{y \in \mathbf{Z} : y \geq x\}$.

Remark 3 *(The case $p = 1/2$). For the homogeneous case in* (10) *with $p = 1/2$, we can show that $S_n \sim Bin(n, 1/2)$. Indeed, evaluating* (14) *with $p = 1/2$, we obtain*

$$\alpha_n = \left(\frac{t+1}{2}\right)^n - \left(\frac{t-1}{2}\right)^n,$$

and

$$\varphi_{S_n}(t) = \alpha_n \left(\frac{t+1}{2}\right) - \alpha_{n-1} \left(\frac{t-1}{2}\right)\left(\frac{t+1}{2}\right) = \left(\frac{t+1}{2}\right)^n,$$

which implies $S_n \sim Bin(n, 1/2)$. Alternatively, setting $Y_k = \mathbf{I}_{\{X_{k+1} = X_k\}}$, we have

$$S_n =^d \sum_{k=1}^{n} Y_k \tag{16}$$

with $Y_k \sim Ber(p^2 + (1-p)^2)$, for all $k \geq 1$. Observe that $\forall k \geq 1, \forall i_l \in \{0, 1\}, l = 1, \cdots, k$,

$$P\left(Y_{k+1} = 1 | X_k = i_k, \cdots, X_1 = i_1\right) = P\left(Y_{k+1} = 1 | X_k = i_k\right) = p^{i_k}(1-p)^{1-i_k}.$$

We thus infer for $p = 1/2$ that the Y_k's are independent $Ber(1/2)$, and that consequently $S_n \sim Bin(n, 1/2)$, by virtue of (16). *Finally, notice that, for $p \in (0, 1), E(S_n) = \sum_{k=1}^{n} E(Y_k) = n(p^2 + (1-p)^2)$, which equals np iff $p = 1/2$, and implies $S_n \sim Bin(n, 1/2)$ if and only if $p = 1/2$.*

We pursue with an expansion of φ_{S_n} as given in (14) in order to derive an explicit form for the probability function of S_n.

Theorem 1 *Let $\rho = 4p(1-p)$, n, k be non negative integers, and*

$$a_{n,k}(\rho) = \left(\frac{\rho}{1-\rho}\right)^{\lceil \frac{n}{2} \rceil - k - 1} \sum_{j = \lceil \frac{n}{2} \rceil - k - 1}^{\lceil \frac{n}{2} \rceil - 1} \binom{j}{\lceil \frac{n}{2} \rceil - k - 1} \binom{n}{2j+1} (1-\rho)^j \; 1_{\{0 \leq k \leq \lceil \frac{n}{2} \rceil - 1\}}.$$

Under assumption (10), we have for all $p \in (0, 1)$, $p \neq \frac{1}{2}$,

(a) $P(S_n = 2k) = \frac{\rho a_{n,k}(\rho)}{2^n}$, *and* $P(S_n = 2k+1) = \frac{1}{2^n}\left[(2-\rho)a_{n,k}(\rho) + \rho a_{n-1,k}(\rho) - \rho a_{n-1,k-1}(\rho)\right]$, *for* n *odd; and*

(b) $P(S_n = 2k) = \frac{1}{2^n}\left[(2-\rho)a_{n,k-1}(\rho) + \rho a_{n-1,k}(\rho) - \rho a_{n-1,k-1}(\rho)\right]$, *and* $P(S_n = 2k+1) = \frac{\rho a_{n,k}(\rho)}{2^n}$, *for* n *even.*

Proof. Begin with standard operations to express α_n as a polynomial in t. Write Corollary 1's λ_1 and λ_2 as $\lambda_1 = (t+\Delta)/2$, $\lambda_2 = (t-\Delta)/2$, with $\Delta = \sqrt{\rho + t^2(1-\rho)}$, so that $\lambda_1 - \lambda_2 = \Delta$, and

$$\begin{aligned}
\alpha_n = \frac{\lambda_1^n - \lambda_2^n}{\lambda_1 - \lambda_2} &= \frac{1}{\Delta 2^n}\left\{(t+\Delta)^n - (t-\Delta)^n\right\} = \frac{1}{\Delta 2^n}\sum_{k=0}^{n}\binom{n}{k}t^{n-k}\Delta^k(1^k - (-1)^k) \\
&= \frac{1}{2^{n-1}}\sum_{k=0,k\,\text{odd}}^{n}\binom{n}{k}t^{n-k}\Delta^{k-1} = \frac{1}{2^{n-1}}\sum_{j=0}^{\lceil\frac{n}{2}\rceil-1}\binom{n}{2j+1}t^{n-2j-1}\left(\rho + t^2(1-\rho)\right)^j \\
&= \frac{1}{2^{n-1}}\sum_{j=0}^{\lceil\frac{n}{2}\rceil-1}\binom{n}{2j+1}t^{n-2j-1}\sum_{k=0}^{j}\binom{j}{k}\left(\frac{\rho}{1-\rho}\right)^k(1-\rho)^j\,t^{2j-2k} \\
&= \frac{1}{2^{n-1}}\sum_{k=0}^{\lceil\frac{n}{2}\rceil-1}t^{n-2k-1}\left(\frac{\rho}{1-\rho}\right)^k\sum_{j=k}^{\lceil\frac{n}{2}\rceil-1}\binom{j}{k}\binom{n}{2j+1}(1-\rho)^j \\
&= \frac{t^{[\text{n even}]}}{2^{n-1}}\sum_{k=0}^{\lceil\frac{n}{2}\rceil-1}t^{2k}\left(\frac{\rho}{1-\rho}\right)^{\lceil\frac{n}{2}\rceil-k-1}\sum_{j=\lceil\frac{n}{2}\rceil-k-1}^{\lceil\frac{n}{2}\rceil-1}\binom{j}{\lceil\frac{n}{2}\rceil-k-1}\binom{n}{2j+1}(1-\rho)^j \\
&= \frac{t^{[\text{n even}]}}{2^{n-1}}\sum_{k}a_{n,k}(\rho)\,t^{2k}. \qquad (17)
\end{aligned}$$

Making use of (14) and (17), we obtain for n odd

$$\begin{aligned}
\varphi_{S_n}(t) &= \left\{\frac{1}{2^{n-1}}\Big(\sum_k a_{n,k}(\rho)t^{2k}\Big)\left(\frac{\rho}{2} + t(1-\frac{\rho}{2})\right)\right\} + \left\{\frac{1}{2^{n-2}}\Big(\sum_k a_{n-1,k}(\rho)t^{2k+1}\Big)\left(\frac{\rho(1-t^2)}{4}\right)\right\} \\
&= \frac{1}{2^n}\Big\{\sum_k \rho\, a_{n,k}(\rho)t^{2k} + \sum_k (2-\rho)a_{n,k}(\rho)t^{2k+1} + \sum_k \rho\, a_{n-1,k}(\rho)t^{2k+1} - \sum_k \rho\, a_{n-1,k-1}(\rho)t^{2k+1}\Big\},
\end{aligned}$$

and the result follows by collecting terms in the representation $\varphi_{S_n}(t) = \sum_k t^{2k}P(S_n = 2k) + \sum_k t^{2k+1}P(S_n = 2k+1)$. Finally, a similar development leads to the stated result for n even. □

Remark 4 *The probability functions given in Theorem 1 and Remark 3 form an interesting family on their own with parameter* ρ, $\rho \in (0,1]$, *with the case* $\rho = 1$ *corresponding to a* $Bin(n, 1/2)$

distribution. As seen below in Remark 5, all of these distributions may be described by a large n normal approximation. It is also instructive to consider some values of the probability function of S_n as prescribed by Theorem 1, such as the previously studied (Remark 2) probabilities of the events $\{S_n = 0\}$ and $\{S_n = n\}$. As an illustration, consider $P(S_n = n)$ for n even in which case (part (a) with $k = n/2$)

$$\begin{aligned} P(S_n = n) &= \frac{1}{2^n}[\,(2-\rho)\,a_{n,\frac{n}{2}-1}(\rho) + \rho\,a_{n-1,\frac{n}{2}-1}(\rho) - \rho\,a_{n-1,\frac{n}{2}-1}(\rho)\,] \\ &= \frac{1}{2^n}[(2-\rho)\sum_{j=0}^{\frac{n}{2}-1}\binom{n}{2j+1}(1-\rho)^j + 0 - \rho\sum_{j=0}^{\frac{n}{2}-1}\binom{n-1}{2j+1}(1-\rho)^j] \\ &= \frac{1}{2^{n+1}}\,[\,(\sqrt{1-\rho}+\frac{1}{\sqrt{1-\rho}})\{(1+\sqrt{1-\rho})^n - (1-\sqrt{1-\rho})^n\} \\ &+ (\sqrt{1-\rho}-\frac{1}{\sqrt{1-\rho}})\{(1+\sqrt{1-\rho})^{n-1} - (1-\sqrt{1-\rho})^{n-1}\}\,], \end{aligned}$$

by virtue of the identity $\sum_{j=0}^{\frac{s}{2}-1}\binom{s}{2j+1}w^j = \frac{1}{2\sqrt{w}}\,[(1+\sqrt{w})^s - (1-\sqrt{w})^s]$; s an even positive integer, $w > 0$. Finally, with a little algebra, one obtains the equivalent form

$$P(S_n = n) = \frac{1}{2^{n+1}}\,[\,(1+\sqrt{1-\rho})^{n+1} + (1-\sqrt{1-\rho})^{n+1}\,] = p^{n+1} + (1-p)^{n+1},$$

as in Remark 2.

We conclude this section with a large sample normal approximation for the distribution of S_n.

Remark 5 *(Large sample normal approximation for S_n) In view of representation (16), the results of Stein (1972, Corollary 3.1) can be applied here to derive a large n normal approximation for the distribution of S_n. Indeed, the Bernoulli sequence $Y_1, Y_2, \ldots$ in (16) is stationary and 1-dependent (i.e., Y_i and Y_j are independent for all $|i-j| > 1$), and the results of Stein imply that*

$$\frac{\frac{S_n}{n} - E(\frac{S_n}{n})}{\sqrt{Var(\frac{S_n}{n})}} \to^d N(0,1). \tag{18}$$

Setting $\rho = 4p(1-p)$, *we have* $Y_k \sim$ *Bernoulli*$(1-\rho/2)$, $E(\frac{S_n}{n}) = 1-\rho/2$ *(as in Remark 3);* $Var(Y_1) = \frac{\rho}{2}(1-\frac{\rho}{2})$; $Cov(Y_1, Y_2) = P(X_1 = X_2 = X_3) - (p^2 + (1-p)^2)^2 = (p^3 + (1-p)^3) - (p^2 + (1-p)^2)^2 = \rho(1-2\rho)/4$;

$$\begin{aligned} Var(\frac{S_n}{n}) = \frac{1}{n^2} Var(\sum_{k=1}^{n} Y_k) &= \frac{1}{n}\{Var(Y_1) + 2(1-1/n)Cov(Y_1, Y_2)\} \\ &= \frac{1}{n}\{\frac{\rho}{2}(1-\frac{\rho}{2}) + (1-1/n)\frac{\rho(1-2\rho)}{2}\} \\ &= \frac{\rho - 3\rho^2/4}{n} + O(n^{-2}), \end{aligned}$$

and finally from (18)

$$\frac{\frac{S_n}{n} - (1-\rho/2)}{\sqrt{\frac{\rho - 3\rho^2/4}{n}}} \rightarrow^d N(0,1). \tag{19}$$

Alternatively, the result is also available with an asymptotic expansion of (14).

3. Applications for Pólya urns

By virtue of de Finetti's representation theorem for sequences of $0-1$ exchangeable random variables, the results above under assumption (10) (Corollary 1 part b; Theorem 1) permit us to describe the distribution of S_n in (1) for $r = 2$ where the rows $(X_{k,1}, X_{k,2}), k \geq 1$ are no longer independent, but arise in the context of a Pólya urn sampling scheme. In such schemes, an urn initially contains b black balls and w white balls. At step k, a ball is drawn randomly and uniformly from urn, and returned to the urn with s ($s > 0$) balls of the same colour. This generates the sequence of rows $(X_{k,1}, X_{k,2}), k \geq 1$, where $X_{k,1} = 1 - X_{k,2}$ is 1 or 0 according to whether the colour of the ball selected on the k^{th} draw is black or white (respectively). Hence, S_n given in (1), and as described in Example 1, counts the number of consecutive pairs of draws with matching colours, among the first $n+1$ draws. Now, in such a case, de Finetti's representation theorem (see for instance Feller,

1971) implies the representation

$$X_{1,1},\ldots,X_{n+1,1} \text{ i.i.d. Bernouilli}(\theta), \text{ with } \theta \sim \text{ Beta}(\frac{b}{s},\frac{w}{s}). \tag{20}$$

We thus obtain the following.

Corollary 2 *Let* $c_{n,j,k} = \binom{j}{\lceil \frac{n}{2} \rceil - k - 1} \binom{n}{2j+1} 1_{\{\lceil \frac{n}{2} \rceil - k - 1 \le j \le \lceil \frac{n}{2} \rceil - 1\}} 1_{\{0 \le k \le \lceil \frac{n}{2} \rceil - 1\}}$, *and*

$$B_{n,k,m} = \sum_j c_{n,j,k} \sum_{i=0}^{j-(\lceil \frac{n}{2} \rceil - k - 1)} \binom{j-(\lceil \frac{n}{2} \rceil - k - 1)}{i} (-1)^i \, 4^{j+m-i} \frac{(\frac{b}{s})_{j+m-i} (\frac{w}{s})_{j+m-i}}{(\frac{b+w}{s})_{2(j+m-i)}}, \tag{21}$$

for positive integers n, k, $m = 0$ *or* 1, *and where* $(a)_x$ *is the usual Pochhammer function defined as* $(a)_0 = 1$, *and* $(a)_x = \prod_{i=0}^{x-1}(a+i)$ *for* $x = 1, 2, \ldots$ *Then, for a Pólya urn as described above with parameters* b, w, s, *we have*

(a) $P(S_n = 2k) = \frac{B_{n,k,1}}{2^n}$, *and* $P(S_n = 2k+1) = \frac{1}{2^n}\left[(2B_{n,k,0} - B_{n,k,1} + B_{n-1,k,1} - B_{n-1,k-1,1}\right]$, *for* n *odd; and*

(b) $P(S_n = 2k) = \frac{1}{2^n}\left[(2B_{n,k-1,0} - B_{n,k-1,1} + B_{n-1,k,1} - B_{n-1,k-1,1}\right]$, *and* $P(S_n = 2k+1) = \frac{B_{n,k,1}}{2^n}$, *for* n *even.*

Proof. It follows directly from representation (20) and Theorem 1 that the probability function of S_n in the context here of a Pólya urn is given by the above equations with $B_{n,k,m} = E[Z^m a_{n,k}(Z)]$, with $Z =^d 4\theta(1-\theta)$, $\theta \sim Beta(\frac{b}{s},\frac{w}{s})$. There remains to show that (21) is a valid expression for $B_{n,k,m}$. Indeed, we have

$$\begin{aligned} E[Z^m a_{n,k}(Z)] &= \sum_j c_{n,j,k} \, E[Z^{m+\lceil \frac{n}{2} \rceil - k - 1}(1-Z)^{j-(\lceil \frac{n}{2} \rceil - k - 1)}] \\ &= \sum_j c_{n,j,k} \sum_{i=0}^{j-(\lceil \frac{n}{2} \rceil - k - 1)} \binom{j-(\lceil \frac{n}{2} \rceil - k - 1)}{i} (-1)^i \, E[(4\theta(1-\theta))^{j+m-i}] \\ &= B_{n,k,m}, \end{aligned} \tag{22}$$

by the evaluation

$$E(Z^u(1-Z)^v)] = \frac{(a_1)_u(a_2)_v}{(a_1+a_2)_{u+v}}, \text{ for } Z \sim \text{ Beta}(a_1, a_2);\ u, v \geq 0. \tag{23}$$

Remark 6 *(Case $b = w$) For cases where $b = w$ (equal number of black and white balls initially in the Pólya urn), the expression of $B_{n,k,m}$ in Corollary 2 is also equivalent to the simpler expression*

$$B_{n,k,m} = (\frac{2b}{s} - 1)_{m+\lceil \frac{n}{2} \rceil - k - 1} \sum_j c_{n,j,k} \frac{(\frac{1}{2})_{j-(\lceil \frac{n}{2} \rceil - k - 1)}}{(\frac{2b}{s} - \frac{1}{2})_{j+m}}.$$

This is verified by establishing that: **(i)** $Z = 4\theta(1-\theta) \sim$ *Beta*$(2b^{'}-1, \frac{1}{2})$ *whenever* $\theta \sim$ *Beta*$(b^{'}, b^{'})$, *and* **(ii)** *by evaluating directly (22) via (23).*

Acknowledgements

The research work of Éric Marchand is partially supported by NSERC of Canada.

References

[1] Ardakov, K. (1997). Powers and other functions of 2×2 matrices. *The Mathematical Gazette*, **4**, 434-431.

[2] Feller, W. (1971). *An Introduction to Probability theory and its Applications*, volume II, Wiley, New York.

[3] 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.

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

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

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

[7] 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.

[8] 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.

[9] Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. II, pp. 583-602.