PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 137, Number 6, June 2009, Pages 2125–2134
S 0002-9939(08)09793-1
Article electronically published on December 30, 2008

A STUDY OF COUNTS OF BERNOULLI STRINGS VIA CONDITIONAL POISSON PROCESSES

FRED W. HUFFER, JAYARAM SETHURAMAN, AND SUNDER SETHURAMAN

(Communicated by Edward C. Waymire)

Abstract. A sequence of random variables, each taking values 0 or 1, is called a Bernoulli sequence. We say that a string of length d occurs in a Bernoulli sequence if a success is followed by exactly $(d-1)$ failures before the next success. The counts of such d-strings are of interest, and in specific independent Bernoulli sequences are known to correspond to asymptotic d-cycle counts in random permutations.

In this paper, we give a new framework, in terms of conditional Poisson processes, which allows for a quick characterization of the joint distribution of the counts of all d-strings, in a general class of Bernoulli sequences, as certain mixtures of the product of Poisson measures. In particular, this general class includes all Bernoulli sequences considered in the literature, as well as a host of new sequences.

1. Introduction

In this paper, we study the joint distribution of the counts of certain d-strings of all orders $d > 1$ arising in a class of Bernoulli sequences. Previous work has used several different methods, including combinatorial, factorial moment, and Pólya and Hoppe urn model methods to identify the joint count distribution with respect to a class of independent Bernoulli sequences. In this context, our main contribution is to introduce a new framework, using conditional Poisson processes, which allows for a concise derivation of the joint count distribution as a mixture of the product of Poisson measures with respect to all Bernoulli sequences considered before, as well as many others in a general collection, including some dependent Bernoulli sequences.

A Bernoulli sequence $\mathbf{Y} = \{Y_n\}_{n\geq 1}$ is a sequence of $\{0,1\}$-valued random variables. For $d \geq 1$, we say that a d-string occurs if a 1 is followed by exactly $(d-1)$ 0's before the next 1 in the Bernoulli sequence. Specifically, a d-string occurs at time $n \geq 1$ if $Y_{n,d} = 1$, where

$$Y_{n,d} = \begin{cases} Y_n Y_{n+1} & \text{for } d = 1, \\ Y_n(1-Y_{n+1})\cdots(1-Y_{n+d-1})Y_{n+d} & \text{for } d \geq 2, \end{cases}$$

Received by the editors January 14, 2008, and, in revised form, September 25, 2008.
2000 *Mathematics Subject Classification.* Primary 60C05; Secondary 60K99.
Key words and phrases. Bernoulli, cycles, strings, spacings, nonhomogeneous, Poisson processes, random permutations.
This research was partially supported by ARO-W911NF-04-1-0333, NSA-H982300510041, and NSF-DMS-0504193.

that is, if $\langle Y_n, \ldots, Y_{n+d}\rangle = \langle 1, \underbrace{0, \ldots, 0}_{d-1}, 1\rangle$.

Let $Z_d = \sum_{n\geq 1} Y_{n,d}$ be the count of all d-strings, for $d \geq 1$, and $\mathbf{Z} = \langle Z_d : d \geq 1\rangle$ be the "count vector" of strings. [In general, $\mathbf{Z}$ may have divergent components, but for the Bernoulli sequences considered in this article it is easily shown (by taking expectations) that all components Z_k are finite with probability 1.]

In this notation, the general problem is to understand the distribution of $\mathbf{Z}$ and its connection to the underlying sequence $\mathbf{Y}$. Although, in this generality, one cannot expect a "closed form" solution (especially with respect to dependent sequences), our aim is to understand when sequences $\mathbf{Y}$ are associated to Poisson-type counts $\mathbf{Z}$, as is the case in many applications with respect to random permutations, record values, Bayesian nonparametrics, and species allocation models through the Ewens sampling formula, through a flexible embedding scheme.

We will use "$\stackrel{d}{=}$" to signify "equals in distribution" and $\mathcal{L}(X)$ to denote the law or distribution of the random variable X. We will also denote $\mathrm{Po}(\lambda)$ as the Poisson measure on $\mathbb{R}$ with intensity λ, and $I(B)$ as the indicator of a set B.

Example 1.1. We follow the exposition in Sethuraman-Sethuraman [13]. Let $\mathbb{S}_n = \{1, 2, \ldots, n\}$, and consider the Feller algorithm to generate a permutation $\pi : \mathbb{S}_n \to \mathbb{S}_n$ uniformly among the $n!$ choices (cf. Feller [5]):

1. Draw an element uniformly from $\mathbb{S}_n$, and call it $\pi(1)$. If $\pi(1) = 1$, a 1-cycle is completed. If $\pi(1) \neq 1$, make another drawing uniformly from $\mathbb{S}_n \setminus \{\pi(1)\}$, and call it $\pi(\pi(1))$. Continue drawing from $\mathbb{S}_n \setminus \{\pi(1), \pi(\pi(1))\}, \ldots$, naming them $\pi(\pi(\pi(1)))$, and so on, until a cycle (of some length) is finished.
2. From the elements left in $\mathbb{S}_n \setminus \{\pi(1), \pi(\pi(1)), \ldots, 1\}$ after the first cycle is completed, follow the process in step 1, with the smallest remaining number taking the role of "1" to finish a second cycle. Repeat until all elements of $\mathbb{S}_n$ are exhausted.

Let $I_k^{(n)}$ be the indicator that a cycle is completed at the kth Feller drawing from $\mathbb{S}_n$. A moment's thought convinces one that $\{I_k^{(n)}\}_{k=1}^n$ are independent Bernoulli random variables with $P(I_k^{(n)} = 1) = 1/(n-k+1)$ because, at time k, independent of the past, exactly one choice from the remaining $n-k+1$ members left in $\mathbb{S}_n$ completes the cycle. Denote $C_k^{(n)}$ as the number of k-cycles in π,

$$C_k^{(n)} = \begin{cases} I_1^{(n)} + \sum_{i=1}^{n-1} I_i^{(n)} I_{i+1}^{(n)} & \text{for } k = 1, \\ \prod_{l=1}^{k-1}(1 - I_l^{(n)}) I_k^{(n)} + \sum_{i=1}^{n-k} I_i^{(n)} \prod_{l=i+1}^{i+k-1}(1 - I_l^{(n)}) I_{i+k}^{(n)} & \text{for } 2 \leq k \leq n. \end{cases}$$

Now let $\mathbf{Y}$ be the independent sequence where $P(Y_k = 1) = 1/k$ for $k \geq 1$, so that $Y_k \stackrel{d}{=} I_{n-k+1}^{(n)}$ for $1 \leq k \leq n$. Then, as Y_n, and $Y_{n-k+1}\prod_{l=n-k+2}^{n}(1 - Y_l)$ for $2 \leq k \leq n$ all vanish in probability as $n \uparrow \infty$, we conclude for each $k \geq 1$ that $\lim_{n\to\infty} C_k^{(n)} \stackrel{d}{=} Z_k$.

Finally, as is well known, the asymptotic cycle counts $\{\lim_n C_k^{(n)}\}_{k\geq 1}$ are distributed as independent Poisson random variables with respective means $1/k$ for $k \geq 1$ (cf. Kolchin [9]). Hence, $\mathbf{Z} \stackrel{d}{=} \prod_{k\geq 1} \mathrm{Po}(1/k)$. [Example 2.1, in section 2, gives a derivation in our Poisson process framework. See also Arratia-Barbour-Tavaré ([1, 2]) for more discussion of the Ewens sampling formula.]

Example 1.2. Consider the standard nonparametric problem of estimating the unknown distribution function F from independent and identically distributed observations $\{X_i\}_{i\geq 1}$. A Bayesian may place on F a Dirichlet prior with parameters $a\mu$ where $a > 0$ and μ is a nonatomic probability measure.

Let $Y_1 = 1$, and for $n \geq 2$ define $Y_n = 1$ if X_n is a new observation, that is, if $X_n \notin \{X_1, \ldots, X_{n-1}\}$, and $Y_n = 0$ otherwise. Then, it can be shown that $\mathbf{Y}$ is an independent Bernoulli sequence with $P(Y_n = 1) = a/(a+n-1)$ for $n \geq 1$ and that $(\log n)^{-1} \sum_{i=1}^{n} Y_i \to a$ a.s. The latter result can be interpreted in terms of counts of strings in this Bernoulli sequence. See Korwar-Hollander [10] for more details, and also Ghosh-Ramamoorthi [6].

In the literature, to our knowledge, only the count vectors of the following class of underlying independent Bernoulli sequences have been investigated. Denote the independent Bernoulli sequence $\mathbf{Y}$ where $P(Y_n = 1) = a/(a+b+n-1)$ for $n \geq 1$ as $\mathbf{Y} = \text{Bern}(a, b)$. The case $a = 1$, $b = 0$ is Example 1.1 (see also Arratia-Tavaré [3]). The case $a > 0$, $b = 0$ is Example 1.2. For this case, Arratia-Barbour-Tavaré [1] observe that the associated $\mathbf{Z} \stackrel{d}{=} \prod_{k\geq 1} \text{Po}(a/k)$ through connections with Ewens sampling formula. When $a = 1$, $b > 0$, Sethuraman-Sethuraman [13], employing factorial moments, show that, given the value x_0 of a $\text{Beta}(b, 1)$ random variable, $\mathbf{Z} \stackrel{d}{=} \prod_{k\geq 1} \text{Po}((1 - x_0^k)/k)$. Such a distribution will be called a "mixture of independent Poisson factors." When $a > 0$ and $b > 0$, Holst [7] goes further, using Pólya and Hoppe urns, and establishes that, given the value x_0 of a $\text{Beta}(b, a)$ random variable, $\mathbf{Z} \stackrel{d}{=} \prod_{k\geq 1} \text{Po}(a(1 - x_0^k)/k)$, again a mixture of independent Poisson factors.

We note also that several interesting studies of 1-strings, and other strings, preceded some of the above work, e.g. an unpublished manuscript of Diaconis, Chern-Hwang-Yeh [4] (which derives approximations via several probability distances), Móri [11] (which uses generating functions), Joffe-Marchand-Perron-Popadiuk [8] (which gives a formula for the 1-string count in a general finite independent Bernoulli sequence in terms of a nonhomogeneous Markov chain and which uses generating functions), and references therein in these and the above papers.

With this background, our main idea is that it is easier to study $\mathbf{Z}$ starting from an extrinsic "conditional marked Poisson process model" (CMPP) rather than directly from the Bernoulli sequence. Namely, we prove that when the underlying Bernoulli sequence $\mathbf{Y}$ is generated through a CMPP model, the count vector $\mathbf{Z}$ is distributed as a mixture of independent Poisson factors in terms of model parameters (Theorem 2.2). As remarked earlier, the Poisson process techniques used here are quite different from previous methods and allow quick derivations. Perhaps interestingly, the sequences $\mathbf{Y}$ found in our model include many dependent Bernoulli sequences (some explicit examples are in section 5). However, the most general sequence studied until now, the independent sequence $\text{Bern}(a, b)$ with $a > 0$ and $b \geq 0$, can also be realized in our framework (Proposition 3.1), yielding a new proof of its count vector distribution.

Our conditional marked Poisson process model also yields a new class of independent Bernoulli sequences, which we call $\text{Bern}_1(a, b)$. Denote the independent Bernoulli sequence $\mathbf{Y}$ where $P(Y_1 = 1) = 1$, and $P(Y_n = 1) = a/(a + b + n - 2)$ for $n \geq 2$ as $\mathbf{Y} = \text{Bern}_1(a, b)$. The $\text{Bern}_1(a, b)$ sequence appends a 1 to the $\text{Bern}(a, b)$ sequence and picks up one more d-string contributed by any leading 0's in $\text{Bern}(a, b)$.

We show that the distribution of the count vector $\mathbf{Z}$ for $\text{Bern}_1(a,b)$ for $a>0, b\geq 1$ is a mixture of independent Poisson factors (Proposition 4.1). This result fails for $0\leq b<1$, and in this case even the distribution of Z_1, the count of 1-strings in $\text{Bern}_1(a,b)$, is not a mixture of Poisson distributions (Proposition 4.5). However, the distribution of $\mathbf{Z}$ in $\text{Bern}_1(a,b)$ can be expressed through a recurrence relation for all values of b including $0\leq b<1$ (Proposition 4.3).

The plan of the article is to discuss the CMPP model and prove the main theorem in section 2. In sections 3 and 4, the main theorem is applied to the independent sequences $\text{Bern}(a,b)$ and $\text{Bern}_1(a,b)$ respectively. Last, in section 5, two explicit dependent Bernoulli sequences, arising from the CMPP model, are given.

2. CMPP models

The following "Poisson process" derivation of the distribution of $\mathbf{Z}$ with respect to $\text{Bern}(1,0)$ (cf. Example 1.1) motivates subsequent developments.

Example 2.1. Consider the following standard way to generate a $\text{Bern}(1,0)$ sequence. Let $\{\beta_i\}_{i\geq 1}$ be independent, identically distributed (iid) Uniform$[0,1]$ random variables, and define $Y_n = I(\beta_n \text{ is a record}),\ n\geq 1$. Rényi's theorem shows that $\{Y_n\}_{n\geq 1}$ are independent and $P(Y_n=1)=1/n$ for $n\geq 1$, that is, $\mathbf{Y}=\text{Bern}(1,0)$. Let $\{X_i\}_{i\geq 1}$ be the record values among $\{\beta_i\}_{i\geq 1}$. Notice that the point process N on $[0,1]$ defined by $N(A)=\sum_{i\geq 1}\delta_{X_i}(A)$ is a nonhomogeneous Poisson process on $[0,1]$ with intensity $1/(1-x)$ (cf. Resnick [12]). For each point X_i, we can associate a Geometric$(1-X_i)$ variable L_i (a "mark") corresponding to the number of uniform random variables in $\{\beta_i\}_{i\geq 1}$ to the next record. Then, by thinning decompositions, $Z_k=\sum_{i\geq 1} I(L_i=k)=\sum_{i\geq 1}\delta_{X_i}([0,1])I(L_i=k)$ for $k\geq 1$ are independent Poisson variables with respective means $\int_0^1 (1-x)^{-1}x^{k-1}(1-x)dx = 1/k$ for $k\geq 1$.

In a sense, the thrust of the following CMPP model and our main result (Theorem 2.2) below is to reverse the procedure in Example 2.1. By beginning with a given Poisson process and spacing variables, which themselves determine the count vector $\mathbf{Z}$, we then see which associated Bernoulli sequence $\mathbf{Y}$ arises.

Consider a sequence of random variables $(\mathbf{X},\mathbf{L})=\{(X_i,L_i)\}_{i\geq 0}$ on $\mathbb{R}\times\mathbb{N}$, where $\mathbb{N}=\{1,2,\ldots\}$, and the point process N on $\mathbb{R}$ is given by $N(A)=\sum_{i\geq 1}\delta_{X_i}(A)$. Also let $g:\mathbb{R}\to[0,\infty)$ be a probability density function (pdf), and for each $x\in\mathbb{R}$ let $r(x,\cdot), q(x,\cdot):\mathbb{N}\to[0,1]$ be probability mass functions, and let $\lambda_x:\mathbb{R}\to[0,\infty)$ be an intensity function.

Then, we say $(\mathbf{X},\mathbf{L})$ is the conditional marked Poisson process $\mathcal{M}(g,r,\lambda,q)$ if the following hold:

1. X_0 has pdf g,
2. conditional on $X_0=x_0$, N is a nonhomogeneous Poisson process with intensity function $\lambda_{x_0}(\cdot)$,
3. $P(L_0=k|\mathbf{X})=r(X_0,k)$ for $k\geq 1$, and
4. $P(L_n=k|\mathbf{X},L_0,L_1,\ldots,L_{n-1})=q(X_n,k)$ for $k,n\geq 1$.

Let $L_0^*=L_0$, and $L_r^*=L_{r-1}^*+L_r$ for $r\geq 1$. We now define a Bernoulli sequence $\mathbf{Y}$ based on $(\mathbf{X},\mathbf{L})$ as follows: $Y_n=1$ if n is of the form L_r^* for some $r\geq 0$, and $Y_n=0$ otherwise. Another way to say this is

$$Y_n = \begin{cases} 0 & \text{when } n<L_0^*, \text{ or } L_r^*<n<L_{r+1}^* \text{ for } r\geq 0,\\ 1 & \text{when } n=L_r^* \text{ for } r\geq 0.\end{cases} \tag{2.1}$$

Then, the count vector $\mathbf{Z}$ is given by

$$Z_k \;=\; \sum_{n\geq 1} I(L_n = k), \qquad \text{for } k \geq 1. \tag{2.2}$$

We note the zeroth mark L_0 is not included in the above summation, since any Y_i with $i < L_0$ is part of an initial segment of zeros of the sequence not preceded by a 1 and so does not contribute to any d-string, for $d \geq 1$.

Theorem 2.2. *Suppose $\int \lambda_w(x)q(x,k)dx < \infty$ for all $w \in \mathbb{R}$ and $k \geq 1$. Then, the count vector $\mathbf{Z}$ associated with the sequence $\mathbf{Y}$, defined through CMPP $(\mathbf{X}, \mathbf{L}) = \mathcal{M}(g, r, \lambda, q)$, is distributed as follows. Given the value $X_0 = x_0$,*

$$\mathbf{Z} \;\stackrel{d}{=}\; \prod_{k\geq 1} \mathrm{Po}\Big(\int \lambda_{x_0}(x)q(x,k)dx \Big).$$

Remark 2.3. The distribution of $\mathbf{Z}$ does not depend on the transition function r, consistent with the discussion of L_0 before the theorem.

Also, for a given $k \geq 1$, Z_k is infinite with positive probability exactly when there is a set B such that $P(X_0 \in B) > 0$ and $\int \lambda_w(x)q(x,k)dx = \infty$ for $w \in B$.

Proof of Theorem 2.2. Recall the count vector representation (2.2). Conditional on $X_0 = x_0$, the point process M on $\mathbb{R}\times\mathbb{N}$ given by $M(A\times\{k\}) = \sum_{i\geq 1} \delta_{X_i}(A) I(L_i = k)$ is a Poisson process on $\mathbb{R}\times\mathbb{N}$ with intensity function $\lambda_{x_0}(x)q(x,k)$ (cf. Proposition 4.10.1 (b) in Resnick [12]). Hence, it follows that, given $X_0 = x_0$, the variables $M(\mathbb{R}\times\{k\}) = \sum_{n\geq 1} I(L_n = k) = Z_k$ are independent Poisson variables with respective means $\int \lambda_{x_0}(x)q(x,k)dx$, for $k \geq 1$. □

3. The sequence Bern(a, b)

We now derive the count vector distribution for the sequence Bern(a, b) using a CMPP model. Denote, as usual, for $\alpha, \beta > 0$, the Beta function

$$B(\alpha,\beta) \;=\; \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{3.1}$$

and let

1. $\bar{g}(x) = x^{b-1}(1-x)^{a-1}/B(b,a)$ on $0 < x < 1$, the Beta(b, a) pdf,
2. $\bar{r}(x,k) = x^{k-1}(1-x)$ for $k \geq 1$,
3. $\bar{\lambda}_w(x) = [a/(1-x)]I(w < x < 1)$, and
4. $\bar{q}(x,k) = x^{k-1}(1-x)$ for $k \geq 1$.

Proposition 3.1. *The model $(\mathbf{X}, \mathbf{L}) = \mathcal{M}(\bar{g}, \bar{r}, \bar{\lambda}, \bar{q})$ produces an independent Bernoulli sequence $\mathbf{Y} \stackrel{d}{=}$ Bern(a, b) for $a > 0$ and $b > 0$ whose count vector $\mathbf{Z}$, conditional on the value x_0 of a* Beta(b, a) *random variable is distributed as $\prod_{k\geq 1} \mathrm{Po}(a(1-x_0^k)/k)$.*

Remark 3.2. As a corollary, by taking $b \downarrow 0$, we recover the count vector distribution for Bern$(a, 0)$ already considered in the literature as simply $\mathbf{Z} \stackrel{d}{=} \prod_{k\geq 1} \mathrm{Po}(a/k)$. Note that $(X_0, L_0) \to (0, 1)$ in distribution as $b \downarrow 0$.

The Poisson process in the above CMPP model with intensity $\bar{\lambda}_w(\cdot)$ can be generated in the following way. First, the point process formed by the record values from an iid sequence of Beta$(1, a)$ random variables is a Poisson process

with intensity $a/(1-x)$, the Beta$(1,a)$ failure rate (cf. Resnick [12], Proposition 4.11.1 (b)). Next, we thin this process as follows. Let $X_0 \stackrel{d}{=} \text{Beta}(b,a)$ and $\{X_i\}_{i\geq 1}$ be the record values from an iid sequence of Beta$(1,a)$ random variables, subject to $X_i > X_0$ for $i \geq 1$. Then, conditional on $X_0 = x_0$, the point process $\bar{N}$ defined by $\bar{N}(A) = \sum_{i\geq 1} \delta_{X_i}(A)$ is the desired Poisson process with intensity function $\bar{\lambda}_{x_0}(x) = [a/(1-x)]I(x_0 < x < 1)$.

Proof of Proposition 3.1. The second part on the count vector distribution follows from Theorem 2.2, noting for $k \geq 1$ that

$$\int_0^1 \bar{\lambda}_{x_0}(x)\bar{q}(x,k)dx \;=\; \int_{x_0}^1 ax^{k-1}dx \;=\; \frac{a(1-x_0^k)}{k}. \tag{3.2}$$

For the first part, we observe that the distribution of $\{Y_i\}_{i\geq 1}$ given through (2.1) is uniquely determined by the probabilities of cylinder sets of the form

$$\begin{aligned} E(k_0,\ldots,k_n) &= (L_0 = k_0, L_1 = k_1, \ldots, L_n = k_n) \\ &= \Big(Y_t = 1 \text{ for } t \in \{K_0, K_1, \ldots, K_n\}, \text{and } Y_t = 0 \text{ otherwise for } 1 \leq t \leq K_n\Big) \end{aligned} \tag{3.3}$$

where $k_0, k_1, \ldots, k_n$ are positive integers and $K_0 = k_0, K_1 = K_0 + k_1, \ldots, K_n = K_{n-1} + k_n$ are their partial sums. If the probability of sets of the form $E \stackrel{def}{=} E(k_0,\ldots,k_n)$ is a product of appropriate marginal probabilities, then $\{Y_n, n \geq 1\}$ will be the Bernoulli sequence Bern(a,b). We will proceed to establish this.

Let $A_n = \{0 < x_0 < x_1 < \cdots < x_n < 1\}$. Using the Beta variables representation in Remark 3.2, write

$$P(E) \;=\; \int_{A_n} \bar{g}(x_0)\bar{r}(x_0,k_0)\prod_{i=1}^n \Big[P(X_i \in dx_i | X_i > x_{i-1})\bar{q}(x_i,k_i)\Big]dx_0.$$

Since $P(X_i \in dx_i | X_i > x_{i-1}) = a(1-x_i)^{a-1}/(1-x_{i-1})^a\, dx_i$ for $1 \leq i \leq n$, we have further that the last line equals

$$\begin{aligned} &\frac{a^n}{B(b,a)}\int_{A_n} x_0^{b+k_0-2}\prod_{i=1}^n x_i^{k_i-1}(1-x_n)^a dx_0\ldots dx_n \\ &\qquad = \;\frac{B(b+K_n-1,a+1)}{B(b,a)}\cdot\frac{a^n}{\prod_{s=0}^{n-1}(b+K_s-1)} \end{aligned} \tag{3.4}$$

and, noting (3.1) and $\alpha\Gamma(\alpha) = \Gamma(\alpha+1)$, that (3.4) becomes

$$\frac{a\prod_{r=0}^{K_n-2}(b+r)}{\prod_{r=0}^{K_n-1}(a+b+r)}\cdot\frac{a^n}{\prod_{s=0}^{n-1}(b+K_s-1)} \;=\; \prod_{i=1}^{K_n}\frac{b+i-1}{a+b+i-1}\prod_{r=0}^{n}\frac{a}{b+K_r-1},$$

which is exactly $\prod_{i=1}^{K_n} P(Y_i = 0)\prod_{r=0}^n [P(Y_{K_r} = 1)/P(Y_{K_r} = 0)]$ with $\mathbf{Y}$ specified as Bern(a,b). □

4. The sequence $\text{Bern}_1(a,b)$

We will derive the count vector distribution for the sequence $\text{Bern}_1(a,b)$, and show a dichotomy depending on whether $b \geq 1$ or $b < 1$. We first consider the case where $a > 0$ and $b > 1$. Define

1. $g^*(x) = x^{b-2}(1-x)^a/B(b-1,a+1)$ on $0 < x < 1$, the Beta$(b-1,a+1)$ pdf,
2. $r^*(x,1) = 1$,

3. $\lambda_w^*(x) = [a/(1-x)]I(w < x < 1)$, and
4. $q^*(x,k) = x^{k-1}(1-x)$ for $k \geq 1$.

Proposition 4.1. *The CMPP model* $(\mathbf{X}, \mathbf{L}) = \mathcal{M}(g^*, r^*, \lambda^*, q^*)$ *produces an independent Bernoulli sequence* $\mathbf{Y} \stackrel{d}{=} \text{Bern}_1(a,b)$ *for* $a > 0$ *and* $b > 1$*, and, conditional on a* $\text{Beta}(b-1, a+1)$ *variable* $X_0 = x_0$*, the distribution of its count vector* $\mathbf{Z}$ *is* $\prod_{k\geq 1} \text{Po}(a(1-x_0^k)/k)$.

Remark 4.2. As a corollary, by taking $b \downarrow 1$, we find the count vector distribution for $\text{Bern}_1(a,1)$ to be simply $\mathbf{Z} \stackrel{d}{=} \prod_{k\geq 1} \text{Po}(a/k)$. (In fact, $\text{Bern}_1(a,1)$ coincides with the sequence $\text{Bern}(a,0)$ mentioned earlier in Remark 3.2.)

Also, we note that the Poisson process in the above CMPP model with intensity λ^* can be generated, as in Proposition 3.1, by taking $X_0 \stackrel{d}{=} \text{Beta}(b-1, a+1)$ and $\{X_i\}_{i\geq 1}$ as the sequence of records from an iid sequence of $\text{Beta}(1,a)$ random variables, subject to the condition $X_1 > X_0$.

Proof of Proposition 4.1. We need only establish the distribution of $\mathbf{Y}$, as the last statement follows from Theorem 2.2 and the computation (3.2). The calculations are similar to the proof of Proposition 3.1. Let $k_0 = 1, k_1, k_2, \ldots, k_n$ be positive integers, and $K_0 = k_0 = 1, K_1 = K_0 + k_1, \ldots, K_n = K_{n-1} + k_n$ be their partial sums. Recall the cylinder set defined in (3.3) and let

$$E_1 \stackrel{def}{=} E(1, k_1, \ldots, k_n) = (L_0 = 1, L_1 = k_1, \ldots, L_n = k_n),$$

and set $A_n = \{0 < x_0 < x_1 < \cdots < x_n < 1\}$. Write, using the construction in Remark 4.2, that

$$\begin{aligned} P(E_1) &= \frac{1}{B(b-1,a+1)} \int_{A_n} \left[x_0^{b-2}(1-x_0)^a\right] \cdot 1 \\ &\qquad \times \prod_{i=1}^{n} \left[a(1-x_i)^{a-1}/(1-x_{i-1})^a\right] \left[x_i^{k_i-1}(1-x_i)\right] dx_0 \ldots dx_n \\ &= \frac{a^n}{B(b-1,a+1)} \int_{A_n} x_0^{b-2} \prod_{i=1}^{n} x_i^{k_i-1}(1-x_n)^a dx_0 \ldots dx_n. \end{aligned}$$

Then, with (3.1) and $\alpha\Gamma(\alpha) = \Gamma(\alpha+1)$, the last line equals

$$\begin{aligned} &\frac{B(b+K_n-2,a+1)}{B(b-1,a+1)} \cdot \frac{a^n}{(b-1)\prod_{s=1}^{n-1}(b+K_s-2)} \\ &\qquad = \frac{\prod_{r=0}^{K_n-2}(b-1+r)}{\prod_{r=0}^{K_n-2}(a+b+r)} \cdot \frac{a^n}{(b-1)\prod_{s=1}^{n-1}(b+K_s-2)} \\ &\qquad = \prod_{i=1}^{K_n-1} \frac{b+i-1}{a+b+i-1} \prod_{r=1}^{n} \frac{a}{b+K_r-2}, \end{aligned}$$

which is exactly $P(Y_1 = 1)\prod_{i=2}^{K_n} P(Y_i = 0)\prod_{r=1}^{n}[P(Y_{K_r} = 1)/P(Y_{K_r} = 0)]$ with $\mathbf{Y}$ specified as $\text{Bern}_1(a,b)$. □

We now give the distribution of the count vector under $\text{Bern}_1(a,b)$ for all $a > 0$ and $b \geq 0$ by conditioning on the location of the second 1 in the sequence $\mathbf{Y}$. Denote $\mathbf{Z}(a,b)$ as the count vector with respect to $\text{Bern}_1(a,b)$ for $a > 0$ and $b \geq 0$. Let $\mathbf{W_n}$

be the sequence whose nth coordinate is 1 and all the other coordinates are zero, for $n \geq 1$. Also let

$$p_n \;=\; \begin{cases} \frac{a}{a+b} & \text{for } n = 2, \\ \frac{a}{a+b+n-2} \prod_{r=0}^{n-3} \frac{b+r}{a+b+r} & \text{for } n \geq 3 \end{cases}$$

be the probability that the second 1 in $\text{Bern}_1(a,b)$ occurs at time $n \geq 2$.

Proposition 4.3. *For $a > 0$ and $b \geq 0$, we have*

$$\mathcal{L}\left(\mathbf{Z}(a,b)\right) \;=\; \sum_{n\geq 2} p_n \, \mathcal{L}\Big(\mathbf{Z}(a, b+n-1) + \mathbf{W}_{n-1}\Big), \tag{4.1}$$

and $\mathbf{Z}(a, b+n-1)$, conditional on the value x_0 of a $\text{Beta}(b+n-2, a+1)$ random variable, is distributed as $\prod_{k\geq 1} \text{Po}(a(1-x_0^k)/k)$, for $b > 0$ and $n \geq 2$.

Remark 4.4. The special case $b = 0$ is interesting. The sequence $\text{Bern}_1(a, 0)$ is the independent sequence where $Y_1 = Y_2 = 1$ and $P(Y_n = 1) = a/(a+n-2)$ for $n \geq 3$. That is, starting from time $n = 2$, the sequence is $\text{Bern}_1(a, 1) = \text{Bern}(a, 0)$. Hence, by Proposition 3.1 (see Remark 3.2), $\mathbf{Z}(a, 0)$ is distributed as $\hat{\mathbf{Z}} + \mathbf{W}_1$, where $\hat{\mathbf{Z}} \stackrel{d}{=} \prod_{k\geq 1} \text{Po}(a/k)$ is the count vector for $\text{Bern}(a, 0)$. This agrees with (4.1), since $p_2 = 1$ (when $b = 0$) and $\mathbf{Z}(a, 1) = \hat{\mathbf{Z}}$.

Proof of Proposition 4.3. The distribution of $\mathbf{Z}(a, b)$ follows by conditioning on the first time that $Y_n = 1$ for $n \geq 2$. The distributions of $\mathbf{Z}(a, b+n-1)$ are completely specified by Proposition 4.1 and Remark 4.2, since $b + n - 1 \geq 1$ for $n \geq 2$. □

From (4.1), it is not clear whether or not the distribution of $\mathbf{Z}(a, b)$ is a mixture of product Poisson factors for $0 \leq b < 1$. We show now that even the first component $Z_1(a, b)$ is not a mixture of Poissons when $0 \leq b < 1$.

Proposition 4.5. *The distribution of $Z_1 \equiv Z_1(a, b)$, the count of 1-strings in the $\text{Bern}_1(a, b)$ sequence, is not a mixture of Poissons when $0 \leq b < 1$; that is, there is no measure μ on $[0, \infty)$ such that*

$$E\Big[\exp\{tZ_1\}\Big] \;=\; \int_{[0,\infty)} e^{v(e^t-1)} d\mu(v). \tag{4.2}$$

Proof. It is well known that when (4.2) holds, the variable Z_1 is *over*-dispersed, that is, $O(Z_1) \stackrel{def}{=} \text{Var}(Z_1) - E(Z_1) \geq 0$. The proof now follows by showing that $O(Z_1) < 0$ in (4.4).

Let $\mathbf{Y} = \text{Bern}_1(a, b)$. Then,

$$Z_1 \;=\; Y_2 + \hat{Z}_1 = Y_2 + Y_2 Y_3 + Z_1^+, \tag{4.3}$$

where $\hat{Z}_1 = \sum_{i\geq 2} Y_i Y_{i+1}$ and $Z_1^+ = \sum_{i\geq 3} Y_i Y_{i+1}$, and the latter is independent of Y_2. Furthermore $\hat{Z}_1$, Z_1^+ are the counts of strings of order 1 from $\text{Bern}(a, b)$, $\text{Bern}(a, b+1)$, respectively, and their distributions are known from Proposition 3.1. Hence, by easy calculations,

$$E(\hat{Z}_1) = \frac{a^2}{(a+b)},\; E(Z_1^+) = \frac{a^2}{(a+b+1)},\; E(\hat{Z}_1^{\,2}) = \frac{a^3(a+1)}{(a+b)(a+b+1)} + \frac{a^2}{(a+b)}.$$

From the identities in (4.3), we have

$$E(Z_1) = \frac{a(a+1)}{(a+b)},\; E(Z_1^2) = \frac{a(a+1)}{(a+b)} + \frac{a^2(a+1)(a+2)}{(a+b)(a+b+1)}.$$

This leads to

$$O(Z_1) \;=\; \frac{a^2(a+1)(b-1)}{(a+b)^2(a+b+1)}, \tag{4.4}$$

which is negative for $b<1$ and positive for $b>1$. □

5. Some dependent Bernoulli sequences

Two examples of dependent Bernoulli sequences, arising in CMPP models with simple structures, whose count vector distributions are mixtures of independent Poisson factors are given.

First sequence. For $a>0$ and $b>0$, denote $P_{a,b}$ as the probability distribution of the CMPP $\mathcal{M}(\bar{g},\bar{r},\bar{\lambda},\bar{q})$ described in Proposition 3.1 which gives rise to the Bernoulli sequence Bern(a,b). Now let $r^+(x,k)=kx^{k-1}(1-x)^2$ for $k\geq 1$. Consider the associated CMPP model $\mathcal{M}(\bar{g},r^+,\bar{\lambda},\bar{q})$ with $\bar{g},\bar{\lambda},\bar{q}$ the same as in Proposition 3.1. Denote the probability measure under this model as $P^+=P^+_{a,b}$. Note that $r^+(x,k)=k[\bar{r}(x,k)-\bar{r}(x,k+1)]$, where $\bar{r}(x,k)=x^{k-1}(1-x)$. Recall the cylinder set $E \stackrel{def}{=} E(k_0,\ldots,k_n)$ from (3.3), where $k_0,k_1,\ldots,k_n$ are positive integers, and $K_0,K_1,\ldots,K_n$ their partial sums. It is easy to see that

$$P^+(E) \;=\; k_0\Big[P_{a,b}\Big(E(k_0,\ldots,k_n)\Big)-P_{a,b}\Big(E(k_0+1,k_1,\ldots,k_n)\Big)\Big].$$

From this expression, the distribution of $\mathbf{Y}$ can be recovered and shown not to be that of independent Bernoulli variables. For instance,

$$\begin{aligned} P^+(Y_1=1) \;&=\; P_{a,b}(Y_1=1)-P_{a,b}(Y_1=0,Y_2=1) \;=\; \frac{a(a+1)}{(a+b)(a+b+1)},\\ P^+(Y_2=1) \;&=\; \frac{a^2(a+2)+2ba(a+1)}{(a+b)(a+b+1)(a+b+2)}. \end{aligned}$$

Thus

$$P^+(Y_1=1)P^+(Y_2=1) \;=\; \frac{a^2(a+1)(a^2+2a+2ba+2b)}{(a+b)^2(a+b+1)^2(a+b+2)},$$

which does not match, for $a,b>0$,

$$P^+(Y_1=1,Y_2=1) \;=\; \frac{a^2(a+2)}{(a+b)(a+b+1)(a+b+2)}.$$

Finally, by Remark 2.3, we note that the count vectors under $P_{a,b}$ and P^+ have the same distribution, and by Proposition 3.1 conditional on the value of x_0 of a Beta(b,a) variable, the count vectors are distributed as $\prod_{k\geq 1}\text{Po}(a(1-x_0^k)/k)$.

Second sequence. Consider $P_{1,0}$, the measure for the CMPP model discussed in Example 2.1 and Remark 3.2, with respect to the Bernoulli sequence Bern$(1,0)$, where $(X_0,L_0)\equiv(0,1)$, $\{X_i\}_{i\geq 1}$ are the records from an iid Uniform$[0,1]$ sequence, and L_i are Geometric$(1-X_i)$ for $i\geq 1$.

Let P' stand for the measure under the "switched" CMPP model where (X_1,L_1) and (X_2,L_2) are interchanged. The probabilities of $\mathbf{Y}$ on cylinder sets (cf. (3.3), under P', are given by

$$P'\Big(E(1,k_1,\ldots,k_n)\Big) \;=\; P_{1,0}(L_2=k_1,L_1=k_2,\text{ and } L_i=k_i \text{ for } 3\leq i\leq n)$$

for positive integers $k_0 = 1, k_1, \ldots, k_n$, with $K_0 = 1, K_1 = K_0 + k_1, \ldots, K_n = K_{n-1} + k_n$ as their partial sums. Under both models $P_{1,0}$ and P', as only two terms (L_1, L_2) exchange places, the associated count vectors are the same and by Proposition 3.1, are distributed as $\prod_{k\geq 1} \mathrm{Po}(1/k)$.

We now show that $\{Y_i\}_{i\geq 1}$ is not an independent sequence under P'. From the calculation in (3.4) with $(X_0, L_0) \equiv (0, 1)$, $Y_1 \equiv 1$ and $\bar{r}(x, 1) = 1$ (take $b \downarrow 0$), and $a = 1$, we can write $P'(Y_2 = 1) = P_{1,0}(L_2 = 1) = \sum_{k\geq 1} P_{1,0}(L_1 = k, L_2 = 1) = \sum_{k\geq 1} \int_{0<x_1<x_2<1} x_1^{k-1}(1 - x_2)dx_1dx_2 = 1/4$.

Also, $P'(Y_2 = 1, Y_3 = 1) = P_{1,0}(L_1 = 1, L_2 = 1) = P_{1,0}(Y_2 = 1, Y_3 = 1) = 1/6$, and $P'(Y_2 = 0, Y_3 = 1) = P_{1,0}(L_2 = 2) = \sum_{k\geq 1} \int_{0<x_1<x_2<1} x_1^{k-1}$. $x_2(1 - x_2)dx_1dx_2 = 5/36$, which give $P'(Y_3 = 1) = 11/36$.

However, $P'(Y_2 = 1)P'(Y_3 = 1) = 11/144 \neq 1/6 = P'(Y_2 = 1, Y_3 = 1)$.

References

[1] Arratia, R., Barbour, A.D., and Tavaré, S. (1992), Poisson process approximations for the Ewens sampling formula. *Ann. Appl. Probab.* **2** 519-535. MR1177897 (94a:60003)

[2] Arratia, R., Barbour, A.D., and Tavaré, S. (2003), *Logarithmic Combinatorial Structures: A Probabilistic Approach*. European Mathematical Society, Zürich. MR2032426 (2004m:60004)

[3] Arratia, R., and Tavaré, S. (1992), The cycle structure of random permutations. *Ann. Probab.* **20** 1567-1591. MR1175278 (93g:60013)

[4] Chern, H.-H., Hwang, H.-K., and Yeh, Y.-N. (2000), Distribution of the number of consecutive records. *Random Structures and Algorithms* **17** 169-196. MR1801131 (2002c:60006)

[5] Feller, W. (1945), The fundamental limit theorems in probability. *Bull. Amer. Math. Soc.* **51** 800-832. MR0013252 (7:128i)

[6] Ghosh, J.K., and Ramamoorthi, R.V. (2003), *Bayesian Nonparametrics*, Springer-Verlag, New York. MR1992245 (2004g:62004)

[7] Holst, Lars (2007), Counts of failure strings in certain Bernoulli sequences. *J. Appl. Probab.* **44** 824-830. MR2355594 (2008i:60014)

[8] Joffe, A., Marchand, E., Perron, F., and Popadiuk, P. (2004), On sums of products of Bernoulli variables and random permutations. *Journal of Theoretical Probability* **17** 285-292. MR2054589 (2005e:60023)

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

[10] Korwar, R.M., and Hollander, M. (1973), Contributions to the theory of Dirichlet processes. *Ann. Probab.* **1** 705-711. MR0350950 (50:3442)

[11] Móri, T.F. (2001), On the distribution of sums of overlapping products. *Acta Scientarium Mathematica (Szeged)* **67** 833-841. MR1876470 (2002h:60024)

[12] Resnick, S.I. (1994), *Adventures in Stochastic Processes*. Second Ed., Birkhäuser, Boston. MR1181423 (93m:60004)

[13] Sethuraman, Jayaram, and Sethuraman, Sunder (2004), On counts of Bernoulli strings and connections to rank orders and random permutations. In *A festschrift for Herman Rubin. IMS Lecture Notes Monograph Series* **45** 140-152. MR2126893 (2006d:60020)

Department of Statistics, Florida State University, Tallahassee, Florida 32306
E-mail address: huffer@stat.fsu.edu

Department of Statistics, Florida State University, Tallahassee, Florida 32306
E-mail address: sethu@stat.fsu.edu

Department of Mathematics, 396 Carver Hall, Iowa State University, Ames, Iowa 50011
E-mail address: sethuram@iastate.edu