Connections Between Bernoulli Strings and Random Permutations

Jayaram Sethuraman[†] **and Sunder Sethuraman***

Dedicated to the memory of Professor Alladi Ramakrishnan.
We dedicate this paper to the memory of Professor Alladi Ramakrishnan, the founder of the world renown Mathematical Sciences Institute of Chennai. We fondly continue to cherish our memories of several contacts at a personal level with Professor Ramakrishnan.

Summary A sequence of random variables, each taking only two values "0" or "1," is called a Bernoulli sequence. Consider the counts of occurrences of strings of the form $\{11\}, \{101\}, \{1001\}, \ldots$ in Bernoulli sequences. Counts of such Bernoulli strings arise in the study of the cycle structure of random permutations, Bayesian nonparametrics, record values etc.

The joint distribution of such counts is a problem worked on by several researchers. In this paper, we summarize the recent technique of using conditional marked Poisson processes which allows to treat all cases studied previously. We also give some related open problems.

Mathematics Subject Classification (2000) Primary 60C05; Secondary 60K99

Key words and phrases Bernoulli · Cycles · Strings · Spacings · Nonhomogeneous · Poisson processes · Random permutations · Records

Research partially suported by ARO-W911NF-09-1-0338[†] and NSF-DMS 0906713*. Approved for public release, distribution unlimited.

J. Sethuraman
Department of Statistics, Florida State University, Tallahassee, FL 32306, USA
e-mail: sethu@stat.fsu.edu

S. Sethuraman
Department of Mathematics, 396 Carver Hall, Iowa State University, Ames, IA 50011, USA
e-mail: sethuram@iastate.edu

K. Alladi et al. (eds.), *The Legacy of Alladi Ramakrishnan in the Mathematical Sciences*, DOI 10.1007/978-1-4419-6263-8_24,

1 Introduction

Consider a sequence of independent Bernoulli random variables Y_k where $P(Y_k = 1) = 1 - P(Y_k = 0) = 1/k$ for $k \geq 1$. We will call such a sequence as a Bern$(1, 0)$ sequence. Such a sequence notes the outcomes, "1" for a success and "0" for a failure, in an experiment conducted over times $k = 1, 2, \ldots$. Notice that there is an infinite number of successes in the sequence, that is $\sum_{k\geq 1} Y_k = \infty$ a.s. since $\sum_{k\geq 1} E[Y_k] = \infty$. However, the number, Z_1, of consecutive pairs of successes, or strings $\{11\}$, is a.s. finite since

$$E(Z_1) = \sum_{k\geq} E[Y_k Y_{k+1}] \; = \; \sum_{k\geq 1} \frac{1}{k(k+1)} \; = \; 1.$$

As an illustration of counting strings of the form $\{11\}$, we see that there are five strings of the form $\{11\}$ in the truncated sequence $\{01011101111\}$. What can one say about the distribution of Z_1?

Persi Diaconis, around 1996, surprisingly recognized that Z_1 is distributed as a Poisson random variable with mean 1! Several studies of Z_1 and related counts of other strings followed from this observation, which became the subject of friendly mathematical conversation. In fact, we learned of the problem from Krishna Athreya, who heard it during the course of a dinner at a conference.

This topic can be generalized. For $m \geq 2$, let $Z_m = \sum_{k\geq 1} X_k[\prod_{l=1}^{m-1}(1 - X_{k+l})]X_{k+m}$, be the count of strings where a success is followed by exactly $m - 1$ failures before the next success, that is the number of strings of the form $\{1\underbrace{0\ldots0}_{m-1}1\}$. Analogous to Z_1, all the counts Z_m for $m \geq 2$ are finite a.s. Intriguingly, the counts $\mathbf{Z} = \{Z_m\}_{m\geq 1}$ turn out to be independent random variables, and the distribution of Z_m is Poisson with mean $1/m$ for $m \geq 1$.

How to explain this phenomena, and how robust and relevant is it? Consider the situation where the success probabilities are "perturbed" in certain ways, that is when $X_1, X_2, \ldots$ are independent with Bernoulli distributions satisfying $P(X_k = 1) = a/(a + b + k - 1)$ for $a > 0$, $b \geq 0$, and $k \geq 1$. We will call such a Bernoulli sequence as a Bern(a, b) sequence. In this case also, it turns out the joint distribution of the counts $\mathbf{Z}$ can be described in terms of a mixture of Poisson variables. Interestingly, Bern(a, b) sequences have been found to arise naturally in the study of random permutations, record values, Bayesian nonparametrics, and species allocation models.

However, for strings which are not of the form $\{10\ldots01\}$, it seems that "nice" distributional expressions for their counts may not be available even with respect to sequence Bern$(1, 0)$. For instance, although the generating function for the count $W_3 = \sum_{k\geq 1} X_k X_{k+1} X_{k+2}$, of three consecutive successes, i.e., of the string $\{111\}$, can be found, its distribution is not known in a "closed form." See [15] for more details.

As another independent Bernoulli sequence, consider $Y_1, Y_2, \ldots$ where $Y_1 \equiv 1$, $P(Y_k = 1) = a/(a + b + k - 2)$ for $k \geq 2$ for $a > 0$, $b \geq 0$, which we call $\text{Bern}_1(a, b)$. This sequence appends a 1 to a $\text{Bern}(a, b)$ sequence, thereby picking up an additional k-string corresponding to any leading 0's in the $\text{Bern}(a, b)$ sequence. Another interpretation of $\text{Bern}_1(a, b)$ arises from the following observation. The conditional distribution of the tail segment $(Y_n, Y_{n+1}, \ldots)$ in a $\text{Bern}(a, b)$ sequence given $Y_n = 1$ is $\text{Bern}_1(a, b + n + 2)$.

It can be proved that the joint distribution of $\mathbf{Z}$ is sensitive to the value of b in a $\text{Bern}_1(a, b)$ sequence. Namely, when $b \geq 1$, the joint distribution is again a mixture of Poisson variables, but is not when $0 \leq b < 1$.

By now there are several different ways to find the joint distribution of $\mathbf{Z} = \{Z_m\}_{m \geq 1}$, for instance by using combinatorial techniques [1–4], generating functions of moments [12, 17], Polya and Hoppe urns [7], and Poisson process embedding [8–10]. The purpose of this note is to summarize existing results, and to describe the last method in [10], the technique of using conditional marked Poisson process models, through which the joint distribution of $\mathbf{Z}$ can be found for a large class of Bernoulli sequences including all sequences studied before, in particular $\text{Bern}(a, b)$, $\text{Bern}_1(a, b)$, and dependent sequences.

The plan of the article is to give motivating examples in Sect. 2, and to detail the technique of conditional marked Poisson processes in Sect. 3. In Sects. 4, 5, and 6, this method is applied to find the joint distribution of $\mathbf{Z}$ when $\mathbf{Y} = \text{Bern}(a, b)$, when $\mathbf{Y} = \text{Bern}_1(a, b)$, and also when $\mathbf{Y}$ are some types of dependent Bernoulli sequences. In the following, we rely on the exposition in [10, 17].

2 Examples

Bernoulli sequences arise naturally in several situations. We give four examples below with respect to random permutations, Bayesian nonparametric statistics, production failures, and record values.

Example 2.1. This example will show that the Bernoulli sequence $\text{Bern}(1, 0)$ arises in the limit in the study of cycles in random permutations. 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. [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 draw uniformly from $\mathbb{S}_n \setminus \{\pi(1)\}$, and call it $\pi(\pi(1))$. If $\pi(\pi(1)) = 1$, a 2-cycle is completed. If $\pi(\pi(1)) \neq 1$, 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 draw from $\mathbb{S}_n$. A moment's thought convinces us that $\{I_k^{(n)}\}_{k=1}^n$ are independent Bernoulli random variables with $P(I_k^{(n)} = 1) = 1/(n-k+1)$ since, at time k and 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,

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

Now let $\mathbf{Y}$ be the sequence Bern$(1, 0)$ where $P(Y_k = 1) = 1/k$ for $k \ge 1$ so that $Y_k \stackrel{d}{=} I_{n-k+1}^{(n)}$ in distribution, for $1 \le k \le n$. Since Y_n, and $Y_{n-k+1} \prod_{l=n-k+2}^{n} (1 - Y_l)$ for $2 \le k \le n$ all vanish in probability as $n \uparrow \infty$, we can conclude, for each $k \ge 1$, that $\lim_{n\to\infty} C_k^{(n)} \stackrel{d}{=} Z_k$ in distribution.

Finally, as is well-known, the asymptotic cycle counts $\{\lim_n C_k^{(n)}\}_{k\ge1}$ are distributed as independent Poisson random variables with respective means $1/k$ for $k \ge 1$ (cf. [13]). Hence, $\mathbf{Z} \stackrel{d}{=} \prod_{k\ge1} \text{Po}(1/k)$. See also [1, 2] for more discussion with Ewens sampling formula.

Example 2.2. Consider the standard nonparametric inference problem of estimating the unknown distribution function F from data $X_1, X_2, \ldots$ which are independently and identically distributed as F. In Bayesian inference, one would place a Dirichlet prior $\mathcal{D}(\alpha)$ on F. Here α is a finite measure on $\mathbb{R}_1$ with $a = \alpha(\mathbb{R}_1) > 0$. Under these circumstances, one can show that there will be repetitions among $X_1, X_2, \ldots$. Let $\beta_1 = 1, \beta_n = I(X_n \notin \{X_1, \ldots, X_{(n-1)}\})$ for $n = 2, 3, \ldots$. Thus, $\beta = 1$ if X_n is different from $X_1, \ldots, X_{(n-1)}$ and zero other wise. It is well-known that $\beta_1, \beta_2, \ldots$ are independent and $P(\beta_n = 1) = a/(a + n - 1)$ for $n = 1, 2, \ldots$ and thus form a Bern$(a, 0)$ sequence. For details, see [6, 14]. This example is also relevant in counting species among animals that are captured, and is part of the definition of species allocation models.

Example 2.3. Suppose items are produced and examined routinely over time. Alternatively, the item can be a long "chip" with successive spatial components. The data consist of a Bernoulii sequence $\{Y_1, Y_2, \ldots\}$, where $Y_n = 1$ means that there is a flaw (and $Y_n = 0$ means that there is no flaw) at time n or at the nth spatial component. In practice, given improvements in production scheme or other attributes, $P(Y_n = 1)$ will go to 0 as n gets large. Isolated flaws do not signify failures. However, successive flaws like $\{11\}$, $\{101\}, \ldots$ signify failures of say of type $1, 2, \ldots$. One would like to know the distribution of the number of failures of type $1, 2, \ldots$, e.g., the distribution of the joint distribution $\mathbf{Z}$.

Example 2.4. The following is another way to generate a Bern$(1, 0)$ sequence from record values. Let $\{\beta_i\}_{i\ge1}$ be independent, identically distributed (iid) Uniform$[0, 1]$ random variables, and define $Y_1 = 1$ and $Y_n = I(\beta_n \text{ is a record}) = I(\beta_n > \max(\beta_1, \ldots, \beta_{n-1}))$, $n \ge 2$. Rènyi's theorem shows that $\{Y_n\}_{n\ge1}$ are independent and $P(Y_n = 1) = 1/n$ for $n \ge 1$, that is $\mathbf{Y} = \text{Bern}(1, 0)$.

3 Conditional Marked Poisson Process (CMPP)

To introduce the technique of conditional marked Poisson processes, let us further examine Example 2.4 of Sect. 2, and derive in its context the joint distribution of the count vector $\mathbf{Z}$ associated with sequence $\mathbf{Y} = \text{Bern}(1, 0)$. With the same notations, define $\tau_1 = 1, X_1 = \beta_1$, $\tau_n = \inf\{m : m > \tau_{n-1}, \beta_m > X_{\tau_{n-1}}\}$, $X_n = \beta_{\tau_n}$ for $n \geq 2$. Then, $\{X_i\}_{i\geq 1}$ are the record values among $\{\beta_i\}_{i\geq 1}$ and $\{\tau_n\}_{i\geq 1}$ are the record times. 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. [16]).

For each record value 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 $\int_0^1 (1 - x)^{-1}x^{k-1}(1 - x)\mathrm{d}x = 1/k$ for $k \geq 1$.

The idea now is to reverse the discussion above, and starting from what we call a conditional marked Poisson process (CMPP), which is slightly more general than a marked Poisson process, we determine a Bernoulli sequence $\mathbf{Y}$ and compute the corresponding joint distribution of $\mathbf{Z}$ through Poisson thinning decompositions.

Conditional Marked Poisson Process 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 $\mathbb{R}$ given by $N(A) = \sum_{i\geq 1} \delta_{X_i}(A)$. Let also $g : \mathbb{R} \to [0, \infty)$ be a probability density function (pdf), and for each $x \in \mathbb{R}$ $r(x, \cdot), q(x, \cdot) : \mathbb{N} \to [0, 1]$ be probability mass functions, and $\lambda_x(\cdot) : \mathbb{R} \to [0, \infty)$ be an intensity function.

Then, we say that $(\mathbf{X}, \mathbf{L})$ forms a CMPP $\mathcal{M}(g, r, \lambda, q)$ 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. A different way to say this is

$$Y_n = \begin{cases} 0, & \text{when } n < L_0^*, \text{ or } L_r^* < n < L_{r+1}^* \text{ for } r \geq 0 \\ 1, & \text{when } n = L_r^* \text{ for some } r \geq 0. \end{cases} \tag{3.1}$$

In the Bernoulli sequence $\mathbf{Y}$, there is a $1 : 1$ correspondence between k-strings and marks $L_n = k$, which signify a "1" followed by $(k - 1)$ "0"s and then succeeded by a "1." Thus, the count vector $\mathbf{Z}$ associated with $\mathbf{Y}$ is given by

$$Z_k = \sum_{n \geq 1} I(L_n = k), \quad \text{for } k \geq 1. \tag{3.2}$$

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 k-string, for $k \geq 1$.

Theorem 3.1. *Suppose* $\int \lambda_w(x)q(x,k)\mathrm{d}x < \infty$ *for all* $w \in \mathbb{R}$ *and* $k \geq 1$. *Then, the count vector* $\mathbf{Z}$ *associated with sequence* $\mathbf{Y}$*, defined through CMPP* $(\mathbf{X}, \mathbf{L}) = \mathcal{M}(g, r, \lambda, q)$*, is distributed as follows. Given the value* $X_0 = x_0$*,*

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

Remark 3.2. The distribution of $\mathbf{Z}$ does not depend on the transition function r, consistent with the discussion of L_0 before the theorem.

Proof of Theorem 3.1. Recall the count vector representation (3.2). 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) [16]). 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 $\int \lambda_{x_0}(x)q(x,k)\mathrm{d}x$, for $k \geq 1$. □

4 The Sequence Bern(a, b)

We now give a CMPP model which produces a Bern(a, b) sequence. Recall that a sequence $\mathbf{Y}$ is a Bern(a, b) sequence if $Y_1, Y_2, \ldots$ are independent and $P(Y_k = 1) = a/(a + b + k - 1)$ for $k = 1, 2, \ldots$. Denote, as usual, for $\alpha, \beta > 0$, the Beta function

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}. \tag{4.1}$$

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 $k \geq 1$.

We note 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. [16] 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 $X_0 = w$, 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 $\bar{\lambda}_w(x) = [a/(1-x)]I(w < x < 1)$.

Proposition 4.1. *The model* $(\mathbf{X}, \mathbf{L}) = \mathcal{M}(\bar{g}, \bar{r}, \bar{\lambda}, \bar{q})$ *produces an independent Bernoulli sequence* $\mathbf{Y} \stackrel{d}{=} \text{Bern}(a,b)$ *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* $\prod_{k\geq 1} \text{Po}(a(1-x_0^k)/k)$.

Remark 4.2. As a corollary, by taking $b \downarrow 0$, we recover the count vector distribution for Bern$(a,0)$ as simply $\mathbf{Z} \stackrel{d}{=} \prod_{k\geq 1} \text{Po}(a/k)$. Note that $(X_0, L_0) \to (0,1)$ in distribution as $b \downarrow 0$.

Proof of Proposition 4.1. The second part on the count vector distribution follows from Theorem 3.1, noting for $k \geq 1$, that

$$\int_0^1 \bar{\lambda}_{x_0}(x)\bar{q}(x,k)\mathrm{d}x = \int_{x_0}^1 a x^{k-1}\mathrm{d}x = \frac{a(1-x_0^k)}{k}. \tag{4.2}$$

The first part is proved by showing that the finite dimensional distributions of the Bernoulli sequence $\mathbf{Y}$ agree with those of a Bern(a,b). Observe that the distribution of $\{Y_i\}_{i\geq 1}$ given through (3.1) is uniquely determined by the probabilities of cylinder sets of the form $E = E(k_0, \ldots, k_n)$,

$$\begin{aligned} E &= (L_0 = k_0, L_1 = k_1, \ldots, L_n = k_n) \\ &= \Big(Y_t = 1 \text{ for } t \in \{K_0, K_1, \ldots, K_n\}, \text{and } Y_t = 0 \text{ otherwise for } 1 \leq t \leq K_n\Big), \end{aligned} \tag{4.3}$$

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. The random variables $\{Y_n\}$ will form a Bern(a,b) sequence if

$$P(E(k_0, \ldots, k_n)) = \prod_{i=1}^{K_n} \frac{b+i-1}{a+b+i-1} \prod_{r=0}^{n} \frac{a}{b+K_r-1}. \tag{4.4}$$

Let $A_n = \{0 < x_0 < x_1 < \cdots < x_n < 1\}$. We now use the Beta variables representation given just above Proposition 4.1. Observe

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

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 $1 \le i \le n$, we have further that the last line equals

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

which is equal to the probability in (4.4). □

We note following ideas based on Theorem 2.2 in [9] (Theorem 3.1 in this note). Holst [8] shows that an alternate CMPP model based on iid exponential random variables can also give rise to a Bern(a,b) and yield the same results for $\mathbf{Z}$.

5 The Sequence Bern$_1(a,b)$

Recall Bern$_1(a,b)$ is the independent Bernoulli sequence $\mathbf{Y}$ where $P(Y_1 = 1) = 1$ and $P(Y_k = 1) = a/(a+b+k-2), k = 2, 3, \ldots$. We now construct a CMPP model corresponding to Bern$_1(a,b)$ sequence when $a > 0, b > 1$. Thus, the joint distribution of strings $\mathbf{Z}$ in a Bern$_1(a,b)$ sequence when $a > 0, b > 1$ can be written as a certain mixture of Poissons.

Let $a > 0$ 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 $k \ge 1$.

Note that the Poisson process in the above CMPP model with intensity λ^* can be generated, as in Proposition 4.1, by taking $X_0 \stackrel{d}{=}$ Beta$(b-1, a+1)$, and $\{X_i\}_{i \ge 1}$ as the sequence of records from an iid sequence of Beta$(1,a)$ random variables, subject to the condition $X_1 > X_0$.

Proposition 5.1. *The CMPP model* $(\mathbf{X}, \mathbf{L}) = \mathcal{M}(g^*, r^*, \lambda^*, q^*)$ *produces an independent Bernoulli sequence* $\mathbf{Y} \stackrel{d}{=}$ Bern$_1(a,b)$ *for* $a > 0$ *and* $b > 1$, *and, conditional on a Beta*$(b-1, a+1)$ *variable* $X_0 = x_0$, *the distribution of its count vector* $\mathbf{Z}$ *is* $\prod_{k \ge 1}$ Po$(a(1-x_0^k)/k)$.

Remark 5.2. As a corollary, by taking $b \downarrow 1$, we find the count vector distribution for Bern$_1(a,1)$ to be simply $\mathbf{Z} \stackrel{d}{=} \prod_{k \ge 1}$ Po(a/k). [In fact, Bern$_1(a,1)$ coincides with the sequence Bern$(a,0)$ mentioned earlier in Remark 4.2.]

Proof of Proposition 5.1. That the Bernoulli sequence $\mathbf{Y}$ defined from $\mathbf{X}, \mathbf{L}$ is $\text{Bern}_1(a, b)$, and the associated counts $\mathbf{Z}$ are the desired mixture of Poissons follows from the same method as in Proposition 4.1. See [10] for details. □

Although a CMPP model does not lead to a $\text{Bern}_1(a, b)$ sequence for $a > 0$, $0 \le b < 1$, the distributions of the associated count vector $\mathbf{Z}$ can still be described with direct calculations in terms of a recurrence relation. However, it can be shown the distribution of $\mathbf{Z}$ is not a mixture of Poissons. For more specifics, see [10].

6 Dependent Bernoulli Sequences

The CMPP model given in Sect. 3 can also produce dependent Bernoulli sequences. In all these cases, as a consequence of Theorem 3.1, the joint distribution of the count vector $\mathbf{Z}$ are fully described as a mixture of Poisson variables.

We describe briefly two such examples.

Example 6.1. 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 4.1 which gives rise to the Bernoulli sequence $\text{Bern}(a, b)$. Let now $r^+(x, k) = kx^{k-1}(1-x)^2$ for $k \ge 1$. 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 4.1. 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 (4.3) 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 with a few calculations not to be an independent sequence, e.g., $P^+(Y_1 = Y_2 = 1) \ne P^+(Y_1 = 1)P^+(Y_2 = 1)$. For details see [10].

However, by noting Remark 3.2, the count vectors under $P_{a,b}$ and P^+ have the same distribution $\prod_{k \ge 1} \text{Po}(a(1 - x_0^k)/k)$.

Example 6.2. Let $\mathbf{Y}$ be the Bernoulli sequence $\text{Bern}(1, 0)$ generated by the CMPP model based on $(\mathbf{X}, \mathbf{L})$ discussed in Example 2.4 and Remark 4.2. Note that the count vector $\mathbf{Z}$ does not change if one interchanges (X_1, L_1) and (X_2, L_2). More precisely, let $X_0^* = X_0, L_0^* = L_0, X_1^* = X_2, L_1^* = L_2, X_2^* = X_1, L_2^* = L_1$, and $X_n^* = X_n, L_n^* = L_n$ for $n = 3, 4, \ldots$. Then, as the counts are invariant under such a switch, $\mathbf{Z}^* = \mathbf{Z}$ still has distribution $\prod_1^\infty \text{Po}(1/k)$. However, the underlying Bernoulli sequence $\mathbf{Y}^*$ generated by $(\mathbf{X}^*, \mathbf{L}^*)$ is no longer independent. Again, one can show $P(Y_1^* = Y_2^* = 1) \ne P(Y_1^* = 1)P(Y_2^* = 1)$. Details can be found in [10].

7 Some Open Problems

We indicate two intriguing questions, although certainly many more can be envisioned.

1. As indicated in the introduction, the generating function of W_3, the count of strings of the form $\{111\}$ in Bern(a, b) has been identified in the nice paper [15]. However, we do not have a good specification of the exact distribution. We know even less about counts of strings of the form $\{1111\}, \{11111\}$, etc. although some recursions are given in [15]. Can one say something more about these counts?
2. In an interesting paper [11], the following question is raised. Consider the sequence Bern$(a, 0)$. We know that the count Z_1 of strings of the form $\{11\}$ is finite. Let N_1 be the last n such that $Y_{n-1}Y_n = 1$. Is there a stopping time τ on $\mathbf{Y}$ such that $P(\tau = N_1)$ is maximized among all stopping times? Hsiau [11] constructs such a τ and shows that it is of a threshold type, that is there is a $t \in \mathbb{N}$ such that $\tau = \min\{n : n \geq t, Y_{n-1}Y_n = 1\}$. It will be interesting to answer this question for Bernoulli sequences Bern(a, b) and for other counts $Z_2, Z_3, \ldots$.

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.
[2] Arratia, R., Barbour, A.D. and Tavaré, S. (2003) *Logarithmic Combinatorial Structures: A Probabilistic Approach*. European Mathematical Society, Zürich.
[3] Arratia, R., and Tavaré, S. (1992) The cycle structure of random permutations. *Ann. Probab.* **20** 1567–1591.
[4] Chern, H.-H., Hwang, H.-K. and Yeh, Y.-N. (2000) Distribution of the number of consecutive records. *Random Struct. Algor.* **17** 169–196.
[5] Feller, W. (1945) The fundamental limit theorems in probability. *Bull. Amer. Math. Soc.* **51** 800–832.
[6] Ghosh, J.K., and Ramamoorthi, R.V. (2003) *Bayesian Nonparametrics*. Springer, New York.
[7] Holst, L. (2007) Counts of failure strings in certain Bernoulli sequences. *J. Appl. Probab.* **44** 824–830.
[8] Holst, L. (2008) A note on embedding certain Bernoulli sequences in marked Poisson processes *J. Appl. Probab.* **45** 1181–1185.
[9] Huffer, F., Sethuraman, J. and Sethuraman, S. (2008) A study of counts of Bernoulli strings via conditional Poisson processes. Available at arXiv 0801.2115v1.pdf.
[10] Huffer, F., Sethuraman, J. and Sethuraman, S. (2009) A study of counts of Bernoulli strings via conditional Poisson processes. *Proc. Amer. Math. Soc.* **137** 2125–2134.
[11] Hsiau S. (2009) Selecting the last consecutive records in a record process. *Technical Report.*
[12] Joffe, A., Marchand, E., Perron, F. and Popadiuk, P. (2004) On sums of products of Bernoulli variables and random permutations. *J. Theoret. Probab.* **17** 285–292.
[13] Kolchin, V.F. (1971) A problem of the allocation of particles in cells and cycles of random permutations. *Theory Probab. Appl.* **16** 74–90.
[14] Korwar, R.M., and Hollander, M. (1973) Contributions to the theory of Dirichlet processes. *Ann. Probab.* **1** 705–711.

[15] Móri, T.F. (2001) On the distribution of sums of overlapping products. *Acta Scientarium Mathematica (Szeged)* **67** 833–841.
[16] Resnick, S.I. (1994) *Adventures in Stochastic Processes.* Second Ed. Birkhäuser, Boston.
[17] Sethuraman, J., and Sethuraman, S. (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.