

# Gaps and interleaving of point processes in sampling from a residual allocation model

JIM PITMAN<sup>1</sup> and YURI YAKUBOVICH<sup>2</sup>

<sup>1</sup>Statistics Department, University of California, 367 Evans Hall # 3860, Berkeley, CA 94720-3860, USA.  
E-mail: [pitman@stat.berkeley.edu](mailto:pitman@stat.berkeley.edu)

<sup>2</sup>St. Petersburg State University, Universitetskaya nab. 7-9, St. Petersburg 199034, Russia.  
E-mail: [y.yakubovich@spbu.ru](mailto:y.yakubovich@spbu.ru)

This article presents a limit theorem for the gaps  $\widehat{G}_{i:n} := X_{n-i+1:n} - X_{n-i:n}$  between order statistics  $X_{1:n} \leq \dots \leq X_{n:n}$  of a sample of size  $n$  from a random discrete distribution on the positive integers  $(P_1, P_2, \dots)$  governed by a residual allocation model (also called a Bernoulli sieve)  $P_j := H_j \prod_{i=1}^{j-1} (1 - H_i)$  for a sequence of independent random hazard variables  $H_i$  which are identically distributed according to some distribution of  $H \in (0, 1)$  such that  $-\log(1 - H)$  has a non-lattice distribution with finite mean  $\mu_{\log}$ . As  $n \rightarrow \infty$  the finite dimensional distributions of the gaps  $\widehat{G}_{i:n}$  converge to those of limiting gaps  $G_i$  which are the numbers of points in a stationary renewal process with i.i.d. spacings  $-\log(1 - H_j)$  between times  $T_{i-1}$  and  $T_i$  of births in a Yule process, that is  $T_i := \sum_{k=1}^i \varepsilon_k / k$  for a sequence of i.i.d. exponential variables  $\varepsilon_k$  with mean 1. A consequence is that the mean of  $\widehat{G}_{i:n}$  converges to the mean of  $G_i$ , which is  $1/(i\mu_{\log})$ . This limit theorem simplifies and extends a result of Gnedin, Iksanov and Roesler for the Bernoulli sieve.

*Keywords:* GEM distribution; interleaving of simple point processes; residual allocation model; stars and bars duality; stationary renewal process; Yule process

## 1. Introduction

Let  $P_\bullet := (P_1, P_2, \dots)$  be a random discrete distribution on the positive integers described by a *residual allocation* or *stick-breaking* model (RAM), [11,23,43], also known as a *Bernoulli sieve* [16,19] in which

$$P_j := H_j \prod_{i=1}^{j-1} (1 - H_i), \tag{1.1}$$

where the  $H_i \in (0, 1)$ , called *residual fractions*, *random discrete hazards*, or *factors*, are independent and identically distributed (i.i.d.) according to some distribution of  $H \in (0, 1)$ . Given  $P_\bullet$ , let  $X_1, \dots, X_n$  be an i.i.d. sample of size  $n$  from  $P_\bullet$ , and define counting variables

$$N_{b:n} := \sum_{i=1}^n \mathbf{1}(X_i = b) \quad (b = 1, 2, \dots). \tag{1.2}$$

Study of the *reversed counts*, coming back from the maximum of the sample,

$$\widehat{N}_{\bullet,n} := (N_{M_n-i:n}, i = 0, 1, \dots), \quad \text{where } M_n := \max\{b : N_b > 0\} = \max_{1 \leq i \leq n} X_i \quad (1.3)$$

has been motivated by a number applications in ecology, population biology, and computer science [9,11,16,39,43]. A model of particular interest in these applications is the GEM(0,  $\theta$ ) model, when it is assumed that  $H$  has the beta(1,  $\theta$ ) distribution with density  $\theta(1 - u)^{\theta-1}$  at  $u \in (0, 1)$ . To provide some notation, let

$$\mu_{i,j} := \mathbb{E}H^i(1 - H)^j \stackrel{\theta}{=} \frac{(1)_i(\theta)_j}{(1 + \theta)_{i+j}}, \quad (1.4)$$

where  $\stackrel{\theta}{=}$  indicates an evaluation for the GEM(0,  $\theta$ ) model with  $\mathbb{P}(H > u) \stackrel{\theta}{=} (1 - u)^\theta$ , and

$$(x)_i := x(x + 1) \cdots (x + i - 1) = \frac{\Gamma(x + i)}{\Gamma(x)}$$

is Pochhammer’s rising factorial function. Let

$$\mu_{\log} := \mathbb{E} - \log(1 - H) = \sum_{m=1}^{\infty} \frac{\mu_{m,0}}{m} = \int_0^1 \frac{\mathbb{P}(H > u)}{1 - u} du \stackrel{\theta}{=} \frac{1}{\theta}. \quad (1.5)$$

The starting point of this article is the following limit theorem.

**Theorem 1.1 (Gnedin, Iksanov and Roesler [17], Theorem 2.1).** *If*

$$- \log(1 - H) \text{ has a non-lattice distribution with finite mean } \mu_{\log} \quad (1.6)$$

*then*

$$\widehat{N}_{\bullet,n} \xrightarrow{(d)} N_{\bullet} \quad \text{as } n \rightarrow \infty, \quad (1.7)$$

*meaning that the finite dimensional distributions of the sequence  $\widehat{N}_{\bullet,n}$  converge to those of a limit sequence  $N_{\bullet}$ , which are determined by the formula*

$$\mathbb{P}(N_i = n_i, 0 \leq i \leq k) = \frac{(n_0 + \cdots + n_k - 1)!}{n_0! \cdots n_k!} \frac{\prod_{i=0}^k \mu_{n_i, n_0 + \cdots + n_{i-1}}}{\mu_{\log}} \quad (1.8)$$

*for every finite sequence of non-negative integers  $(n_0, \dots, n_k)$  with  $n_0 > 0$ .*

Formula (1.8) corrects a confusing mistake in the corresponding formula of [17], Theorem 2.1. The proof of this result in [17] involved a construction of  $N_{\bullet}$  from a limiting point process, recalled later in (3.19). This article offers a simpler approach to an extended version of Theorem 1.1, with a simpler point process representation of the limiting sequence  $N_{\bullet}$ , and some corollaries which clarify special properties of the limit for the GEM(0,  $\theta$ ) model.

Let

$$X_{1:n} := \min_{1 \leq i \leq n} X_i \leq X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n} = M_n := \max_{1 \leq i \leq n} X_i, \tag{1.9}$$

denote the order statistics of the sample  $X_1, \dots, X_n$  from the RAM  $P_\bullet$ . It was recently shown in [39] that most known properties of  $\text{GEM}(0, \theta)$  samples are consequences of the following simple description of the *sequence of gaps*  $\widehat{G}_{\bullet:n}$  between order statistics, in reversed order

$$\widehat{G}_{i:n} := X_{n+1-i:n} - X_{n-i:n} \quad (1 \leq i \leq n) \tag{1.10}$$

with the convention  $X_{0:n} := 1$ , so that  $\widehat{G}_{n:n} := X_{1:n} - 1$ . According to the main result of [39]

*in sampling from  $\text{GEM}(0, \theta)$  the  $\widehat{G}_{i:n}$  are independent, with geometric distributions whose parameters depend only on  $i$  and  $\theta$ , and not on the sample size  $n \geq i$ , according to the equality in distribution*

$$(\widehat{G}_{i:n}, 1 \leq i \leq n) \stackrel{(d)}{=} (G_i, 1 \leq i \leq n), \tag{1.11}$$

where  $G_\bullet := (G_i, i = 1, 2, \dots)$  is an infinite sequence of independent random variables such that  $G_i$  has the geometric( $i/(i + \theta)$ ) distribution on  $\{0, 1, \dots\}$  with  $\mathbb{E}G_i = \theta/i$  and

$$\mathbb{P}(G_i \geq k) = \left(\frac{\theta}{i + \theta}\right)^k \quad (k = 0, 1, 2, \dots). \tag{1.12}$$

It is an obvious consequence of (1.11) that  $\widehat{G}_{\bullet:n} \xrightarrow{(d)} G_\bullet$  in the  $\text{GEM}(0, \theta)$  model. The aim of the present study is to relate this limit sequence of reversed gaps  $G_\bullet$  to the limiting reversed counts  $N_\bullet$  in (1.7), first of all for the  $\text{GEM}(0, \theta)$  model, then to describe a corresponding limit sequence of reversed gaps  $G_\bullet$  for a more general RAM.

In sampling from any random discrete distribution  $P_\bullet$ , the two sequences  $\widehat{G}_{\bullet:n}$  and  $\widehat{N}_{\bullet:n}$  are related via the *time-reversed tail count sequence*  $\widehat{Q}_{\bullet:n}$  defined by the partial sums of  $\widehat{N}_{\bullet:n}$ . So, using the well-known balls in the box description of the sampling procedure, recalled in Section 2.1 below,

$$\widehat{Q}_{k:n} := \sum_{i=0}^k \widehat{N}_{i:n} := \sum_{i=0}^k N_{M_n-i:n} \quad (k = 0, 1, 2, \dots) \tag{1.13}$$

is the number of balls in the last  $k + 1$  boxes when  $n$  balls are distributed, counting back from the rightmost occupied box. The count of balls in that box is the initial reversed tail count, indexed by  $k = 0$ :

$$\widehat{Q}_{0:n} := \widehat{N}_{0:n} := N_{M_n:n} := \sum_{i=1}^n \mathbf{1}(X_i = M_n) > 0. \tag{1.14}$$

For each  $g \geq 0$ , a value  $j < n$  appears exactly  $g$  times in the sequence  $\widehat{Q}_{\bullet:n}$  iff there is a corresponding gap between order statistics  $\widehat{G}_{j:n} = g$ . Thus,

$$\widehat{G}_{j:n} = \sum_{k=0}^n \mathbf{1}(\widehat{Q}_{k:n} = j) \quad (1 \leq j < n). \tag{1.15}$$

For a RAM subject to (1.6), the connection (1.13)–(1.15) between the reversed gaps  $\widehat{G}_{\bullet;n}$  and reversed counts  $\widehat{N}_{\bullet;n}$  shows that convergence in distribution of  $\widehat{G}_{\bullet;n}$  to a limiting sequence of gaps  $G_{\bullet}$  holds jointly with the convergence in distribution (1.7) of  $\widehat{N}_{\bullet;n}$  to  $N_{\bullet}$ , with  $G_{\bullet}$  the sequence of occupation times of states by the partial sums  $Q_{\bullet}$  of  $N_{\bullet}$ . The conclusion of this argument is expressed by the following extension of Theorem 1.1.

**Theorem 1.2.** *Suppose the common distribution of factors  $H$  in the RAM, with moments (1.4), is such that  $-\log(1 - H)$  is non-lattice with finite mean  $\mu_{\log}$ . Then as  $n \rightarrow \infty$  there is the joint convergence of finite-dimensional distributions*

$$(\widehat{Q}_{\bullet;n}, \widehat{N}_{\bullet;n}, \widehat{G}_{\bullet;n}) \xrightarrow{(d)} (Q_{\bullet}, N_{\bullet}, G_{\bullet}) \tag{1.16}$$

with limit sequences  $Q_{\bullet}$  and  $N_{\bullet}$  indexed by  $k = 0, 1, 2, \dots$ , and  $G_{\bullet}$  indexed by  $j = 1, 2, \dots$ . The joint distribution of these three limit sequences is defined as follows:

- The sequence  $Q_{\bullet}$  is a Markov chain with stationary transition probability matrix

$$p_{m,n} := \binom{n-1}{m-1} \mu_{n-m,m} \quad (1 \leq m \leq n), \tag{1.17}$$

where  $\mu_{n-m,m} := \mathbb{E}H^{n-m}(1 - H)^m$ , and initial distribution

$$\mathbb{P}(Q_0 = m) = \frac{\mu_{m,0}}{m\mu_{\log}} \quad (m = 1, 2, \dots) \tag{1.18}$$

which is the limit distribution of  $N_{M_n;n}$ , the number of balls in the last occupied box, as  $n \rightarrow \infty$ .

- The limit sequence of reversed counts  $N_{\bullet}$  is the difference sequence of  $Q_{\bullet}$ , with  $N_0 := Q_0$  and  $N_i := Q_i - Q_{i-1}$  for  $i = 1, 2, \dots$
- The limit sequence of reversed gaps  $G_{\bullet}$  is the sequence of occupation counts of the Markov chain  $Q_{\bullet}$ :

$$G_j := \sum_{k=0}^{\infty} \mathbf{1}(Q_k = j) \quad (j = 1, 2, \dots). \tag{1.19}$$

It should be clear from the above discussion that Theorem 1.2 can easily be deduced from Theorem 1.1, and vice versa. In particular, the formula (1.18) for the initial distribution of  $Q_0 = N_0$  is the instance  $k = 0$  of the formula (1.8) for the limiting distribution of  $N_{M_n;n}$ , which was first found in [20]. See also [16], Theorem 6.1. The equivalence of formula (1.8) for general  $k \geq 0$  with the Markov property of  $Q_{\bullet}$  expressed in Theorem 1.2 is also easily checked. Further study of the chain  $Q_{\bullet}$  in the GEM(0,  $\theta$ ) case leads to a remarkable connection to the Riemann zeta function [10]. Theorem 1.2 will be proved in Section 3, along with the following corollaries.

**Corollary 1.3.** *In the setting of Theorem 1.2:*

- The distribution of  $G_j$ , the limit in distribution of  $\widehat{G}_{j;n}$  as  $n \rightarrow \infty$ , is the zero-modified geometric distribution with parameters  $(h_j, 1 - \mu_{0,j})$ , meaning that

$$\mathbb{P}(G_j \geq k) = h_j \mu_{0,j}^{k-1} \quad (j, k = 1, 2, \dots), \tag{1.20}$$

where

$$h_j := \mathbb{P}(G_j \geq 1) = \frac{1 - \mu_{0,j}}{j \mu_{\log}} \quad (j = 1, 2, \dots). \tag{1.21}$$

- The conditional distribution of  $G_j$  given  $G_j \geq 1$  is the  $\text{geometric}(1 - \mu_{0,j})$  distribution on  $\{1, 2, \dots\}$ .
- All moments of  $G_j$  are finite, in particular

$$\mathbb{E}G_j = \frac{h_j}{1 - \mu_{0,j}} = \frac{1}{j \mu_{\log}} \tag{1.22}$$

and there is convergence of moments  $\lim_{n \rightarrow \infty} \mathbb{E}\widehat{G}_{j;n}^p = \mathbb{E}G_j^p$  for every  $p \geq 0$ .

**Corollary 1.4.** *In the setting of Theorem 1.2, for a RAM with i.i.d. factors, the following four conditions are equivalent:*

- the model is  $\text{GEM}(0, \theta)$ , meaning that  $H$  is  $\text{beta}(1, \theta)$ , for some  $\theta > 0$ ;
- distribution of  $G_j$  is  $\text{geometric}(p_j)$  on  $\{0, 1, \dots\}$  for some  $p_j$ , for all  $j = 1, 2, \dots$ ;
- the limiting gaps  $G_j$  are independent random variables;
- for the prelimit gaps the probability  $\mathbb{P}(\widehat{G}_{1;n} \geq 1)$  does not depend on  $n$ .

Then  $p_j = j/(j + \theta)$  and the formulas of Corollary 1.3 hold with

$$\mu_{\log} = \frac{\theta}{\theta}, \quad h_j = \frac{\theta}{j + \theta} \mu_{0,j} = \frac{\theta}{j + \theta}. \tag{1.23}$$

As discussed in [17] for a general RAM, and in [39] for  $\text{GEM}(0, \theta)$ , the limit sequences  $N_\bullet$  and  $G_\bullet$  may be encoded in various ways as counts in suitable point processes. But those studies overlooked the simple basic structure of the Markov chain  $Q_\bullet$  exposed by the next two corollaries.

**Corollary 1.5.** *For any distribution of  $H$  on  $(0, 1)$ , formula (1.17) defines the transition mechanism of the Markov chain  $Q_\bullet$  defined by a Galton–Watson branching process in a random environment, in which at each generation  $k$  the offspring distribution of all individuals present is  $\text{geometric}(1 - H_k)$  on  $\{1, 2, \dots\}$ , with the  $H_k$  picked i.i.d. from one generation to the next according to the distribution of  $H$ . Thus for each  $k \geq 0$ , by conditioning on  $H_k$ ,*

$$\mathbb{E}(z^{Q_{k+1} - Q_k} \mid Q_k = m) = \mathbb{E}\left(\frac{1 - H}{1 - Hz}\right)^m \tag{1.24}$$

which when expanded in powers of  $z$  is equivalent to (1.17).

This is a mixed negative binomial( $m, 1 - H$ ) distribution for the increment  $Q_{k+1} - Q_k$  given  $Q_k = m$ . With the shift by  $m$ , the conditional distribution of  $Q_{k+1}$  given  $Q_k = m$  is the mixed Pascal( $m, 1 - H$ ) distribution of the waiting time until the  $m$ th success in a series of trials, which given  $H$  are i.i.d. with success probability  $1 - H$  per trial. In particular, if  $H$  is assigned a beta( $a, b$ ) distribution on  $(0, 1)$ , the distribution of  $Q_{k+1} - Q_k$  given  $Q_k = m$  is known as a *beta mixed negative binomial* or *inverse Polya–Eggenberger* distribution [30], Section 8.4. Some basic properties of the chain  $Q_\bullet$ , such as the formula (1.31) below for  $\mathbb{E}(Q_k)$  and asymptotic growth of  $Q_k$  for large  $k$ , can be read from general theory of branching processes in a random environment. See, for instance, [3,24,44]. But it seems easiest to obtain results about  $Q_k$  for large  $k$  from the following simple construction:

**Corollary 1.6.** *For  $k = 0, 1, 2, \dots$  let  $Y_k := \sum_{i=1}^k \varepsilon_i / i$  where the  $\varepsilon_i$  are independent standard exponential variables, so that  $0 = Y_0 < Y_1 < Y_2 < \dots$  are the times of births in a standard Yule process  $N_Y(t) := \sum_{k=0}^\infty \mathbf{1}(Y_k \leq t)$ ,  $t \geq 0$ . Independent of this Yule process, let  $0 < S_0^* < S_1^* < \dots$  be the times of arrivals in a stationary renewal counting process  $N_S^*(t) := \sum_{i=0}^\infty \mathbf{1}(S_i^* \leq t)$ ,  $t \geq 0$ , with  $S_i^* - S_{i-1}^*$  for  $i \geq 1$  a sequence of i.i.d. copies of  $-\log(1 - H)$ , and  $S_0^*$  independent of the  $S_i^* - S_{i-1}^*$  for  $i \geq 1$ , with the stationary delay distribution*

$$\mathbb{P}(S_0^* \in ds) = \frac{\mathbb{P}(-\log(1 - H) > s)}{\mu_{\log}} ds = \frac{\mathbb{P}(H > 1 - e^{-s})}{\mu_{\log}} ds \quad (s > 0), \tag{1.25}$$

where  $\mu_{\log} := \mathbb{E} -\log(1 - H)$  as in (1.5). Then the limiting reversed tail count Markov chain  $Q_\bullet$  may be constructed as

$$Q_k := N_Y(S_k^*) \quad (k = 0, 1, 2, \dots) \tag{1.26}$$

along with a corresponding representation of the cumulative sums of gaps  $G_\bullet$ :

$$G_1 + \dots + G_j = N_S^*(Y_j) \quad (j = 1, 2, \dots). \tag{1.27}$$

These results may be summarized less formally as follows. Regard the sample from the RAM as a pattern of  $n$  balls in an infinite row of boxes labeled by positive integers, from left to right. Read the pattern from right to left, starting with the rightmost occupied box, as a list of  $n$  stars and some number of bars, with a star for each ball, and each bar representing a barrier between boxes. The pattern starts with  $N_{0:n} \geq 1$  stars followed by a bar, then  $N_{1:n} \geq 0$  stars, followed by a bar, and so on. Then  $G_{j:n}$  for  $1 \leq j < n$  is the number of bars between the  $j$ th and  $(j + 1)$ th star.

As  $n \rightarrow \infty$ , the distribution of each initial finite segment of this pattern of stars and bars converges to that of an initial segment of an infinite pattern of stars and bars. The limit pattern starts with a star and contains both infinitely many stars and infinitely many bars almost surely. This limit pattern is generated by the superposition of two independent simple point processes on  $[0, \infty)$ , a Yule birth process of stars at times  $0 = Y_0 < Y_1 < \dots$ , and a stationary renewal process of bars at times  $0 < S_0^* < S_1^* < \dots$ , by listing the stars and bars in their order of appearance, always starting with a star from initial birth of the Yule process at time 0.

Thus the two limit sequences  $N_\bullet$  and  $G_\bullet$  provide dual encodings of the interleaving of points of the two simple point processes  $S^*$  and  $Y$  on  $[0, \infty)$ :

- $G_\bullet$  counts  $S^*$  renewals between consecutive Yule birth times;
- $N_\bullet$  counts Yule birth times between consecutive  $S^*$  renewal points.

For instance, if the interleaving of birth times and renewal points in increasing order is

$$(0 = Y_0, Y_1, S_0^*, Y_2, S_1^*, S_2^*, S_3^*, Y_3, Y_4, Y_5, Y_6, S_4^*, Y_7, Y_8, Y_9, \dots)$$

which information is fully encoded by the string  $YYSYSSSYYYYSYYY \dots$ , the two sequences describing this interleaving are

$$(G_1, G_2, \dots) = (0, 1, 3, 0, 0, 0, 1, 0, 0, \dots) \quad \text{and} \quad (N_0, N_1, \dots) = (2, 1, 0, 0, 4, 3+, \dots),$$

where the  $3+$  indicates a value which is at least 3, but which cannot be determined exactly without examining more points of the interleaving. This is a limiting form of the classical “stars and bars” duality of combinatorics for enumerating various kinds of integer compositions.

In the usual representation of sampling from a RAM, by the locations of  $n$  uniform sample points in the unit interval, relative to a sequence of bins of lengths  $P_1, P_2, \dots$ , the two point processes  $N_Y$  and  $N_S^*$  arise from an asymptotic analysis on a logarithmic scale. The structure of the sample points on that scale, coming down from the maximum, is that of the birth times of the Yule process, with  $Y_0 = 0$  corresponding to the maximum value in the sample of  $n$  independent variables. The two point processes which play dual roles in this asymptotic description are

$$N_S^*(t) := \sum_{i=0}^{\infty} \mathbf{1}(S_i^* \leq t) \text{ has stationary increments with } \mathbb{E}N_S^*(t) = t/\mu_{\log}; \tag{1.28}$$

$$N_Y(t) := \sum_{i=0}^{\infty} \mathbf{1}(Y_i \leq t) \text{ is a Markovian branching process with } \mathbb{E}N_Y(t) = e^t. \tag{1.29}$$

The GEM(0,  $\theta$ ) model is by far the simplest, because of the result of Ignatov [26]:

$$\text{in the GEM}(0, \theta) \text{ model } (N_S^*(t), t \geq 0) \text{ is a Poisson process with constant rate } \theta. \tag{1.30}$$

See [39] for further analysis in this case.

**Corollary 1.7.** *In the setting of Theorem 1.2 for all  $j \geq 0$*

$$\lim_{n \rightarrow \infty} \mathbb{E} \widehat{Q}_{j:n} = \mathbb{E}Q_j = \frac{(\mu_{0,-1} - 1)\mu_{0,-1}^j}{\mu_{\log}}, \tag{1.31}$$

where  $\mu_{0,-1} := \mathbb{E}(1 - H)^{-1}$ , along with a corresponding result for  $\mathbb{E}N_j$  by differencing, with  $\mathbb{E}Q_j = \mathbb{E}N_j = \infty$  for all  $j$  if  $\mu_{0,-1} := \mathbb{E}(1 - H)^{-1} = \infty$ . For the GEM(0,  $\theta$ ) model,  $\mu_{0,-1} \stackrel{\theta}{=} 1/(\theta - 1)_+$ , with value  $\infty$  iff  $0 < \theta \leq 1$ . Also, in the setting of Theorem 1.2, the Markov chain  $Q_\bullet$  is subject to the exponential growth

$$\lim_{k \rightarrow \infty} Q_k^{1/k} = \exp(\mu_{\log}) \text{ almost surely.} \tag{1.32}$$

The formula (1.31) for  $\mathbb{E}Q_j$  is read from the branching process description of  $Q_\bullet$ , and the convergence of expectations is easily justified, much as in [17] for the limiting counts  $K_j = \sum_{i=0}^\infty \mathbf{1}(N_i = j)$ . The limit formula (1.32) also follows easily from Corollary 1.6. The particular case of (1.32) for the Markov chain  $Q_\bullet$  associated with the GEM(0, 1) model with  $\mu_{\log} = 1$  was first encountered by Borel [5] and Lévy [34] in connection with the representation of a number between 0 and 1 in the so-called Engel series, see also [12] for detailed proofs.

## 2. Preliminaries

### 2.1. Stars and bars

The notion of a random sample from a random discrete distribution admits a variety of possible interpretations. See, for instance, [39] for recent review. But the balls-in-boxes metaphor from recent studies of the Bernoulli sieve [16] seems to provide the most intuitive language for the present analysis. Regard the sample  $X_1, \dots, X_n$  as an allocation of  $n$  balls labeled by  $i = 1, 2, \dots, n$  into an unlimited number of boxes labeled by  $b \in \{1, 2, \dots\}$ . So  $X_i$  is the label of the box into which ball  $i$  is thrown. Given  $P_\bullet$  the  $X_i$  are independent allocations with  $\mathbb{P}(X_i = b \mid P_\bullet) = P_b$ . The count  $N_{b:n}$  is the number of balls thrown into box  $b$ , the sample maximum  $X_{n:n} = \max\{b : N_b > 0\}$  is the label of the rightmost occupied box, and so on.

Represent a configuration of balls in boxes by its classical combinatorial encoding as a list of *stars and bars* [45], page 15. For instance, the configurations of values and their multiplicities in two possible samples of size 8 from positive integers may be indicated as

$$N_{\bullet:8} = (2, 0, 1, 2, 0, 3) \quad \text{or} \quad (0, 0, 0, 2, 0, 3, 0, 2, 1) \tag{2.1}$$

which correspond to the stars and bars sequences

$$**||**||***||| \dots \quad \text{or} \quad |||**||***||**|| \dots \tag{2.2}$$

Here  $N_{b:8}$  is the number of stars between the  $(b - 1)$ th and  $b$ th bar, reading from left to right, and the sequence is terminated for convenience at the rightmost occupied box  $b = M_8$ , beyond which all the counts are 0. Each star represents a ball, and each bar a *barrier* between boxes. Box 1 is the container to the left of the first bar, box 2 is between the first bar and the second bar, and so on. The  $\dots$  represent an unlimited number of additional boxes, all of which are empty in these configurations with only 8 balls. These configurations are just as well encoded by their order statistics, or by the gaps between their order statistics. The sequence of gaps between order statistics  $\widehat{G}_{\bullet:8}$ , defined from the top down as in (1.10), counts numbers of bars between stars (barriers between balls), from right to left, starting with the rightmost ball:

$$\widehat{G}_{\bullet:8} = (0, 0, 2, 0, 1, 2, 0, 0) \quad \text{or} \quad (1, 0, 2, 0, 0, 2, 0, 3). \tag{2.3}$$

These two configurations illustrate for  $n = 8$  the cases

$$(N_{1:n} > 0) \Leftrightarrow (\widehat{G}_{n:n} = 0) \quad \text{or} \quad (N_{1:n} = 0) \Leftrightarrow (\widehat{G}_{n:n} > 0). \tag{2.4}$$



These relations hold no matter what the sample, by the identity of events  $(N_{1:n} > 0) = (X_{1:n} = 1)$  and the definition of  $\widehat{G}_{n:n} := X_{1:n} - 1$ .

For sampling from any random discrete distribution, the distribution of  $N_{\bullet:n}$  is just a mixture of infinitinomial( $P_{\bullet}$ ) distributions, as treated in [4,15,32] for a fixed sequence of probability parameters  $P_{\bullet}$  on the positive integers, with the parameter sequence  $P_{\bullet}$  assigned some probability distribution. Thus for every sequence of integers  $(n_1, \dots, n_k)$  with  $n_i \geq 0$  for  $1 \leq i < k$  and  $n_k > 0$  and  $\sum_{i=1}^n n_i = n$ ,

$$\mathbb{P}[N_{\bullet:n} = (n_1, \dots, n_k, 0, 0, \dots)] = \binom{n}{n_1, \dots, n_k} \mathbb{E} \prod_{i=1}^k P_i^{n_i} \tag{2.5}$$

$$= \binom{n}{n_1, \dots, n_k} \mathbb{E} \prod_{i=1}^k H_i^{n_i} (1 - H_i)^{n_{i+1} + \dots + n_k} \tag{2.6}$$

$$= \binom{n}{n_1, \dots, n_k} \prod_{i=1}^k \mu_{n_i, n_{i+1} + \dots + n_k}, \tag{2.7}$$

where in the last expression it is assumed that  $P_{\bullet}$  follows a RAM with i.i.d. hazards  $H_k \stackrel{(d)}{=} H$  with moments  $\mu_{i,j} := \mathbb{E}[H^i (1 - H)^j]$ . This expression for a RAM corrects a formula of Gnedin, Iksanov and Roesler [17], (3), in which the multinomial coefficient should be omitted, and the order of indices reversed. Formula (2.5) also gives

$$\mathbb{P}[\widehat{G}_{\bullet:n} = (g_1, \dots, g_n)] \text{ for } (g_1, \dots, g_n) \leftrightarrow (n_1, \dots, n_k, 0, 0, \dots) \tag{2.8}$$

via the bijection (1.13)–(1.15) mediated by the stars and bars representation between the  $n$ -tuples of non-negative integers  $(g_1, \dots, g_n)$  which are the possible values of the gap sequence  $\widehat{G}_{\bullet:n}$  and the weak compositions  $(n_1, \dots, n_k, 0, 0, \dots)$  of  $n$  with  $n_j \geq 0$  and  $\sum_j n_j = n$  which are possible values of the count sequence  $N_{\bullet:n}$ . In principle, formula (2.5) specifies the distribution of the gaps  $\widehat{G}_{\bullet:n}$  for a sample of size  $n$  from any random discrete distribution  $P_{\bullet}$ , with some simplification for a RAM. The distribution of the gap sequence  $\widehat{G}_{\bullet:n}$  derived from GEM(0,  $\theta$ ) is especially simple, as indicated in (1.11), and far simpler than its logically equivalent description in terms of the counts sequence  $N_{\bullet:n}$ , via the stars and bars bijection.

### 2.2. Point processes

For a random discrete distribution  $P_{\bullet}$  on the positive integers, assumed to be of the stick-breaking form (1.1) for some (possibly dependent) random factors  $H_i \in [0, 1]$ , let

$$F_0 := 0 \quad \text{and} \quad F_j := \sum_{i=1}^j P_i = 1 - \prod_{i=1}^j (1 - H_i) \quad \text{for } j = 1, 2, \dots \tag{2.9}$$

In the stick-breaking interpretation, the  $F_k \in [0, 1]$  are called the *break points*. For a random sample  $X_1, X_2, \dots$  from  $P_\bullet$ ,

$$\mathbb{P}(X_i \leq j \mid P_\bullet) = F_j, \tag{2.10}$$

so the sequence  $F_\bullet = (F_0, F_1, F_2, \dots)$  gives the evaluations at  $j = 0, 1, 2, \dots$  of the random discrete cumulative distribution function derived from  $P_\bullet$ . Suppose that almost surely

$$P_j > 0 \quad \text{and} \quad \sum_j P_j = 1, \quad \text{or, equivalently} \quad 0 < F_1 < F_2 < \dots \uparrow 1. \tag{2.11}$$

Following the method used by Ignatov [26] for  $\text{GEM}(0, \theta)$ , and further developed by Gnedin and coauthors [17] for various other models of random discrete distributions, the counts of break points  $F_j$ , not including either end of the interval  $[0, 1]$ , define a simple point process  $N_F$ , with

$$N_F(a, b) := \sum_{k=1}^{\infty} \mathbf{1}(a < F_k \leq b) \quad (0 \leq a < b < 1)$$

the number of break points in  $(a, b)$ . It is convenient to make the change of variable from  $[0, 1]$  to  $[0, \infty)$  by the map  $u \mapsto x = -\log(1 - u)$ . This is the inverse of the cumulative distribution function  $x \mapsto 1 - e^{-x}$  of a standard exponential variable  $\varepsilon := -\log(1 - U)$  for  $U$  uniform on  $(0, 1)$ . Let  $S_0 := 0$  and

$$S_j := -\log(1 - F_j) = \sum_{i=1}^j -\log(1 - H_i) \quad (j = 1, 2, \dots).$$

Regard these images  $S_j$  of the break points  $F_j$  as the points of an associated point process  $N_S$  on  $(0, \infty)$ ,

$$N_S(s, t] := \sum_{j=1}^{\infty} \mathbf{1}(s < S_j \leq t) = N_F(1 - e^{-s}, 1 - e^{-t}] \quad (0 \leq s < t < \infty)$$

and consider the *spacings*  $S_j - S_{j-1}$  between these points. Note that the assumption (2.11) translates into  $0 = S_0 < S_1 < S_2 < \dots \uparrow \infty$ . The following lemma is obvious from these definitions.

**Lemma 2.1.**

- (i) *In the representation of  $P_\bullet$  as a residual allocation model (1.1) or (2.9), the factors  $H_i$  are independent iff the spacings  $S_j - S_{j-1}$ ,  $j = 1, 2, \dots$  are independent.*
- (ii) *The  $H_i$  are i.i.d. iff the  $S_j - S_{j-1}$  are i.i.d., in which case the  $S_j$  are the points of a renewal process on  $(0, \infty)$ , with spacings distributed according to  $S_1 \stackrel{(d)}{=} -\log(1 - H_1)$ .*

The key simplifying property of the  $\text{GEM}(0, \theta)$  model is Ignatov’s result [26] that its associated renewal process is a homogeneous Poisson process with rate  $\theta$ . To analyse the process of

sampling from a random discrete distribution  $P_\bullet$ , the sample  $X_1, X_2, \dots$  may be constructed in the usual way from the discrete random cumulative values  $F_\bullet$  as

$$X_i := 1 + N_F(0, U_i] := 1 + \sum_{j=1}^{\infty} \mathbf{1}(F_j \leq U_i) \quad (i = 1, 2, \dots), \tag{2.12}$$

where  $U_1, U_2, \dots$  is a sequence of i.i.d. uniform  $(0, 1)$  variables, independent of the break points  $F_\bullet$  counted by  $N_F$ . Here, the parts of the stick between break points are labeled from left to right by  $1, 2, \dots$ . Then  $X_i$  is the label of the part of the stick containing  $U_i$ . That label is conveniently evaluated in (2.12) as 1 plus the number of break points to the left of  $U_i$ . The next lemma too follows immediately from these definitions.

**Lemma 2.2.** *Whatever the random discrete distribution  $P_\bullet$  subject to (2.11), let the sample  $(X_1, \dots, X_n)$  be constructed by (2.12) from the break points  $F_\bullet$  and an i.i.d. uniform sample  $(U_1, \dots, U_n)$ , with  $S_j := -\log(1 - F_j)$  the transformed break points, and  $\varepsilon_i := -\log(1 - U_i)$  the transformed uniforms, which are an i.i.d. standard exponential sample. Then the order statistics of the  $X$ -sample, the  $U$ -sample and the  $\varepsilon$ -sample are related by*

$$X_{i:n} = 1 + N_F(0, U_{i:n}] = 1 + N_S(0, \varepsilon_{i:n}] \quad (1 \leq i \leq n), \tag{2.13}$$

while the gaps of the  $X$ -sample in descending order  $\widehat{G}_{i:n} := X_{n-i+1:n} - X_{n-i,n}$  may be recovered from counts in either of the point processes  $N_F$  or  $N_S$  as

$$\widehat{G}_{i:n} = N_F(U_{n-i:n}, U_{n-i+1:n}] = N_S(\varepsilon_{n-i:n}, \varepsilon_{n-i+1:n}] \quad (1 \leq i \leq n). \tag{2.14}$$

### 2.3. The Yule process

The following proposition presents a number of known characterizations of the Yule process.

**Proposition 2.3.** *Let  $Y_0 := 0 < Y_1 < Y_2 < \dots$  be the points of a simple point process on  $[0, \infty)$ , with an initial point at 0, and associated counting process*

$$N_Y(t) := \sum_{k=0}^{\infty} \mathbf{1}(Y_k \leq t) \quad (t \geq 0).$$

The following conditions are equivalent:

- The counting process  $N_Y$  is a standard Yule process, that is a pure birth Markov process, with state space  $\{1, 2, \dots\}$ , initial state  $N_Y(0) = 1$ , right continuous step function paths, constant transition rate  $k$  for transitions from  $k$  to  $k + 1$ , and all other transition rates 0.
- The counting process  $N_Y$  is a Markov process with stationary transition probabilities, such that  $N_Y(0) = 1$  and for  $s, t \geq 0$ , for each  $m = 1, 2, \dots$  the conditional distribution of  $N_Y(s + t)$  given  $N_Y(s) = m$  is the same as the distribution of  $N_Y(t)$  given  $N_Y(0) = m$ .

$t) - N_Y(s)$  given  $N_Y(s) = m$  is the negative binomial( $m, e^{-t}$ ) distribution with generating function

$$\mathbb{E}(z^{N_Y(s+t) - N_Y(s)} \mid N_Y(s) = m) = \left( \frac{e^{-t}}{1 - (1 - e^{-t})z} \right)^m. \tag{2.15}$$

- There is the representation

$$Y_k = \sum_{i=1}^k \frac{\varepsilon_i}{i} \quad (k \geq 0) \tag{2.16}$$

for a sequence of i.i.d. standard exponential variables  $\varepsilon_i$ , with  $Y_0 = 0$ .

- For each fixed  $k$  there is the equality of joint laws

$$(Y_j, 0 \leq j \leq k) \stackrel{(d)}{=} (\varepsilon_{n:n} - \varepsilon_{n-j:n}, 0 \leq j \leq k) \quad \text{for every } n \geq k, \tag{2.17}$$

where  $\varepsilon_{0:n} := 0 < \varepsilon_{1:n} < \dots < \varepsilon_{n:n}$  is the sequence of order statistics of i.i.d. standard exponential variables  $(\varepsilon_i, 1 \leq i \leq n)$ .

- There is the representation

$$N_Y(t) = 1 + N_Y((e^t - 1)\widehat{\varepsilon}_1), \tag{2.18}$$

where  $N_Y(v) := \sum_{i=1}^{\infty} \mathbf{1}(\gamma_i \leq v)$  is a rate 1 Poisson process with arrival times  $\gamma_k = \sum_{i=1}^k \tilde{\varepsilon}_i$  for a sequence of i.i.d. standard exponential variables  $\tilde{\varepsilon}_i$ , and  $\widehat{\varepsilon}_1$  a further standard exponential variable, independent of all the  $\tilde{\varepsilon}_i$ , which may be identified as

$$\widehat{\varepsilon}_1 = \lim_{t \rightarrow \infty} e^{-t} N_Y(t) = \lim_{k \rightarrow \infty} k e^{-Y_k} \quad \text{almost surely} \tag{2.19}$$

while

$$\gamma_k = (e^{Y_k} - 1)\widehat{\varepsilon}_1 \quad (k = 1, 2, \dots). \tag{2.20}$$

- There is the representation

$$Y_k = \log(\widehat{\gamma}_{k+1}/\widehat{\gamma}_1) \quad (k = 0, 1, 2, \dots), \tag{2.21}$$

where  $\widehat{\gamma}_{k+1} = \sum_{i=1}^{k+1} \widehat{\varepsilon}_i$  for a sequence of i.i.d. standard exponential variables  $\widehat{\varepsilon}_i$ , which may be identified as  $\widehat{\gamma}_1 = \widehat{\varepsilon}_1$  as in (2.19) for  $k = 0$ , and  $\widehat{\gamma}_{k+1} = \widehat{\varepsilon}_1 + \gamma_k$  as in (2.20) for  $k \geq 1$ .

The equivalence of the first two descriptions can be found in Feller [14], Section XVII.3. The representation (2.16) of the pure birth process in terms of independent exponential holding times is well known, as is the equivalence of (2.16) and the representation (2.17) in terms of the order statistics of a sequence of i.i.d. exponential variables, which is due to Sukhatme [46] and Rényi [42]. The representation (2.18) is due to Kendall [33], Theorem 1. See also [2], page 127, Theorem 2, [31], Theorem 3.12, [36], and [21], Theorem 4.3, for a more general characterization due to Lundberg [35] of when a mixed Poisson process can be time-changed to a Markovian birth process with stationary birth rates. See also [22], [40], Lemma 6, [8], page 532, [6,13] for

variations and applications of the representation (2.18) of the Yule process. The identification (2.19) follows immediately from (2.18) and the strong law of large numbers for the Poisson process  $N_\gamma$ . Then (2.20) follows, because the random time  $t = Y_k$  when  $N_Y$  first reaches  $k + 1$ , corresponds to the random time  $v = (e^{Y_k} - 1)\widehat{\varepsilon}_1$  when  $N_\gamma$  first reaches  $k$ , that is  $v = \gamma_k$ . The final representation (2.21) is just a rearrangement of (2.20), with the indicated change of variables, as given in [48], (5.3). The equivalence of the two very different looking representations (2.16) and (2.21) of  $(Y_1, Y_2, \dots)$  can be quickly checked as follows. By well-known beta-gamma algebra (see, e.g., [7]), in the construction (2.21) the random variables  $\widehat{\gamma}_1/\widehat{\gamma}_2, \widehat{\gamma}_2/\widehat{\gamma}_3, \dots, \widehat{\gamma}_i/\widehat{\gamma}_{i+1}$  are independent, with

$$\widehat{\gamma}_i/\widehat{\gamma}_{i+1} \stackrel{(d)}{=} 1 - \beta_{1,i} \quad \text{and} \quad -\log(1 - \beta_{1,i}) \stackrel{(d)}{=} \varepsilon_i/i$$

as required. Tracing back through the definitions, this calculation also implies Kendall’s representation (2.18).

### 3. Proofs

In a sample  $X_1, \dots, X_n$  from  $P_\bullet$  defined by the RAM (1.1), the number of values  $X_i$  strictly greater than 1 has the mixed binomial( $n, p$ ) distribution with  $p$  assigned the distribution of  $1 - H$ . That is

$$\mathbb{P}\left(\sum_{i=1}^n \mathbf{1}(X_i > 1) = m\right) = q^*(n, m) := \binom{n}{m} \mu_{n-m,m}, \tag{3.1}$$

where  $\mu_{n-m,m} := \mathbb{E}H^{n-m}(1 - H)^m$ . This  $q^*(\ell, m)$  is the transition probability function of each of the *tail count Markov chains*  $Q_{\bullet:n}^* := (Q_{k:n}^*, k = 0, 1, \dots)$  which describe the sequences of tail counts obtained when  $n$  balls are distributed according to the RAM with i.i.d. factors distributed like  $H$ :

$$Q_{k:n}^* := \sum_{i=1}^n \mathbf{1}(X_i > k) \quad (k = 0, 1, \dots). \tag{3.2}$$

So  $Q_{\bullet:n}^*$  is the Markov chain with transition matrix (3.1) and initial state  $Q_{0:n}^* = n$ . The  $*$  is used here to match notation with the overview [16], Section 2, where  $W := 1 - H$  so that  $\mu_{n-m,m} := \mathbb{E}(1 - W)^{n-m}W^m$ , and to distinguish  $q^*$  from the decrement matrix  $q$  that appears in Gnedin [19] and Gnedin and Pitman [18]. Those articles are concerned with the *composition of  $n$*  generated by the configuration of  $n$  balls in boxes, that is the list of counts of non-empty boxes. For that purpose, the chain with transition matrix  $q^*$  is watched only when it moves. By renormalizing the off-diagonal elements of  $q^*$ , the resulting transition matrix  $q$  with zero diagonal entries is given by

$$q(n, m) := \frac{q^*(n, m)\mathbf{1}(n > m)}{1 - q^*(n, n)} = \binom{n}{m} \frac{\mu_{n-m,m}\mathbf{1}(n > m)}{1 - \mu_{0,n}} \tag{3.3}$$

which explains the factor  $1 - \mu_{0,n}$  in the denominator of formulas of [19] and [18], Example 2. Beware that the *decrement matrix* of [19] and [18], (11), is the matrix with entries  $q(n, n - m)$  rather than  $q(n, m)$  as in (3.3).

By definition,

$$M_n := \max_{1 \leq i \leq n} X_i = \min\{k : Q_{k:n}^* = 0\}, \tag{3.4}$$

is the almost surely finite random time at which the transient Markov chain  $Q_{\bullet:n}^*$  is absorbed in state 0. Theorem 1.2 is concerned with the time reversal of this Markov chain, with time counted back from the step before the absorption time, that is

$$\widehat{Q}_{k:n} := Q_{M_n-1-k:n}^* \quad (k \geq 0) \tag{3.5}$$

with the convention that  $\widehat{Q}_{k:n} := \infty$  for  $k \geq M_n$ , meaning that  $\widehat{Q}_{\bullet:n}$  reaches an absorbing state  $\infty$  at time  $M_n$ , just as  $Q_{\bullet:n}^*$  reaches its absorbing state 0 at time  $M_n$ . The initial state of  $\widehat{Q}_{\bullet:n}$  is  $\widehat{Q}_{0:n} := Q_{M_n-1:n}^* = N_{M_n:n}$ , the number of balls in the rightmost occupied box after  $n$  balls have been thrown. Also, the last state of  $\widehat{Q}_{\bullet:n}$  before absorption at  $\infty$  is  $\widehat{Q}_{M_n-1:n} = Q_{0:n}^* = n$ . Hunt [25] showed that the time reversal of a Markov chain with stationary transition probabilities such as (3.5) is another Markov chain with stationary transition probabilities, say  $\widehat{q}_n(\ell, m)$ , which may be described as follows. For  $n, m \geq 1$  let

$$g_{m:n} := \sum_{k=0}^{\infty} \mathbb{P}(Q_{k:n}^* = m) = \sum_{k=0}^{\infty} \mathbb{P}(\widehat{Q}_{k:n} = m) \tag{3.6}$$

be the potential function giving the expected number of times that either  $Q_{\bullet:n}^*$  or its time reversal  $\widehat{Q}_{\bullet:n}$  hits state  $m$ . Then for positive integers  $m$  and  $\ell$  the expected number of  $m$  to  $\ell$  transitions of the original tail count chain  $Q_{\bullet:n}^*$  can be computed in two different ways as

$$g_{m:n} q^*(m, \ell) = g_{\ell:n} \widehat{q}_n(\ell, m) \tag{3.7}$$

which rearranges to give

$$\widehat{q}_n(\ell, m) = \frac{g_{m:n}}{g_{\ell:n}} q^*(m, \ell). \tag{3.8}$$

A proof of Theorem 1.2 will combine the following two lemmas.

**Lemma 3.1.** *Let  $Q_{\bullet:n}^*$  be a sequence of Markov chains, each with the same transition probability function  $q^*(m, \ell)$  on the set  $\{0, 1, 2, \dots\}$ , such that  $\mathbb{P}(M_n < \infty) = 1$  for every  $n$ , where  $M_n$  is the time  $Q_{\bullet:n}^*$  first hits state 0, assumed to be an absorbing state. Suppose further that for every  $m = 1, 2, \dots$*

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}(Q_{0:n}^* = m) &= 0 \quad \text{and} \\ \lim_{n \rightarrow \infty} g_{m:n} &= g_m \quad \text{with } 0 < g_m < \infty \text{ and } \sum_{m=1}^{\infty} g_m q^*(m, 0) = 1. \end{aligned} \tag{3.9}$$

Let  $\widehat{Q}_{\bullet:n}$  be the chain  $Q_{\bullet:n}^*$  reversed back from time  $M_n - 1$  as in (3.5). Then, as  $n \rightarrow \infty$ , the finite-dimensional distributions of  $\widehat{Q}_{\bullet:n}$  converge to those of a limit process  $\widehat{Q}_{\bullet:\infty}$ , which is a transient Markov chain with state space  $\{1, 2, \dots\}$ , transition probability matrix  $\widehat{q}(\ell, m) := g_m q^*(m, \ell)/g_\ell$ , initial probability distribution

$$\mathbb{P}(\widehat{Q}_{0:\infty} = m) = g_m q^*(m, 0) \quad (m = 1, 2, \dots) \tag{3.10}$$

and potential function  $\sum_{k=0}^\infty \mathbb{P}(\widehat{Q}_{k:\infty} = m) = g_m$  for all  $m = 1, 2, \dots$ .

**Proof.** By decomposing over values of  $M_n$ , formula (3.10) holds with  $\widehat{Q}_{0:n}$  instead of  $\widehat{Q}_{0:\infty}$  and  $g_{m:n}$  instead of  $g_m$  for each  $n$ . The assumptions (3.9) ensure that the initial state  $\widehat{Q}_{0:n}$  of  $\widehat{Q}_{\bullet:n}$  converges in distribution to  $\widehat{Q}_{0:\infty}$  with distribution (3.10). Together with the well-known equation (A.5) recalled in the Appendix, which is satisfied by the potential function

$$g_{\ell:n} = \mathbb{P}(Q_{0:n}^* = \ell) + \sum_{m=1}^\infty g_{m:n} q^*(m, \ell) \tag{3.11}$$

for every  $n$ , these assumptions imply that a stochastic matrix  $\widehat{q}(\ell, m)$  is obtained as the  $n \rightarrow \infty$  limit of  $\widehat{q}_n(\ell, m)$  in (3.8). The probability of any particular finite sequence of values of  $(\widehat{Q}_{i:n}, 0 \leq i \leq k)$ , say  $(m_0, \dots, m_k)$  is

$$g_{m_0:n} q^*(m_0, 0) \prod_{i=0}^{k-1} \widehat{q}_n(m_i, m_{i+1}) \rightarrow g_{m_0} q^*(m_0, 0) \prod_{i=0}^{k-1} \widehat{q}(m_i, m_{i+1}) \quad \text{as } n \rightarrow \infty$$

and the conclusion follows. □

Lemma 3.1 also applies to the theory of random walks conditioned to stay positive. For instance, for  $Q_{\bullet:n}^*$  a simple symmetric random walk with increments of  $\pm 1$  stopped when it first reaches 0, the time reversed limit chain  $\widehat{Q}_{\bullet:\infty}$  is the random walk started at 1 and conditioned never to hit 0, as discussed in [41].

That the assumptions of Lemma 3.1 are satisfied by the particular sequence of tail count Markov chains  $Q_{\bullet:n}^*$  derived from a RAM is a consequence of the next lemma.

**Lemma 3.2.** *In the setting of Theorem 1.2, for a RAM such that  $-\log(1 - H)$  has a non-lattice distribution with finite mean  $\mu_{\log}$ , the limit of the potential function (3.6) of the chain  $Q_{\bullet:n}^*$  as  $n \rightarrow \infty$  is*

$$\lim_{n \rightarrow \infty} g_{m:n} = \frac{1}{m \mu_{\log}}. \tag{3.12}$$

**Proof.** From the transition matrix (3.1), given that the chain  $Q_{\bullet:n}^*$  ever hits state  $m$ , the expected number of times it stays there is  $1/(1 - q^*(m, m)) = (1 - \mu_{0,m})^{-1}$ . So

$$g_{m:n} = \mathbb{P}(Q_{k:n}^* = m \text{ for some } k \geq 0) (1 - \mu_{0,m})^{-1}$$

and (3.12) is equivalent to the result of Gnedin [19], Proposition 5, that under the above assumptions

$$\lim_{n \rightarrow \infty} \mathbb{P}(Q_{k:n}^* = m \text{ for some } k \geq 0) = \frac{1 - \mu_{0,m}}{m\mu_{\log}}. \tag{3.13}$$

Gnedin’s proof of (3.13) can be simplified to obtain (3.12) directly as follows. Let  $S_k := \sum_{i=1}^k -\log(1 - H_i)$  so the  $S_k$  with  $S_0 = 0$  are the times of renewals in a renewal point process on  $[0, \infty)$ , with associated counts  $N_S(s, t] := \sum_{k=1}^{\infty} \mathbf{1}(S_k \in (s, t])$  for  $0 \leq s < t < \infty$ . Note that this renewal point process has zero delay as in Section 2.2, rather than the stationary delay as in Corollary 1.6. Let  $\varepsilon_{1:n} < \dots < \varepsilon_{n:n}$  be the order statistics of  $n$  i.i.d. exponential variables  $\varepsilon_1, \dots, \varepsilon_n$ , so that the tail count chains may be constructed according to Lemma 2.2 as

$$Q_{k:n}^* := \sum_{i=1}^n \mathbf{1}(\varepsilon_i > S_k) = \sum_{i=1}^n \mathbf{1}(\varepsilon_{i:n} > S_k). \tag{3.14}$$

Then for  $1 \leq m < n$  and  $k \geq 0$ , there is the almost sure identity of events

$$(Q_{k:n}^* = m) = (\varepsilon_{n-m:n} < S_k \leq \varepsilon_{n-m+1:n}) \tag{3.15}$$

and hence as  $n \rightarrow \infty$

$$g_{m:n} := \mathbb{E} \sum_{k=0}^{\infty} \mathbf{1}(Q_{k:n}^* = m) = \mathbb{E} N_S(\varepsilon_{n-m:n}, \varepsilon_{n-m+1:n}] \rightarrow \frac{1}{m\mu_{\log}} \tag{3.16}$$

by application of Blackwell’s renewal theorem and the representation (2.17) of exponential order statistics, which together with (2.16) gives  $\mathbb{E}(\varepsilon_{n-m+1:n} - \varepsilon_{n-m:n}) = 1/m$ . □

**Proof of Theorem 1.2.** Lemma 3.2 gives the convergence of potential functions required in (3.9) for application of the time reversal Lemma 3.1 to conclude that  $\widehat{Q}_{\bullet:N}$  converges in distribution to  $Q_{\bullet} := \widehat{Q}_{\bullet:\infty}$  as in the lemma. Since  $Q_{0:n}^* = n$  the first condition in (3.9) is satisfied, and so is the last condition, because switching the order of summation and integration shows that the limit distribution (1.18) of  $Q_0$  sums to 1 for any distribution of  $H$  with  $\mu_{\log} < \infty$ . Formula (1.17) comes from combination of (3.1) and (3.12). The joint convergence in distribution of  $Q$ ,  $N$  and  $G$  sequences as in (1.16) follows easily from the result for the  $Q$  sequences, using the explicit formulas for  $N$  and  $G$  in terms of  $Q$ , both before and after passage to the limit, which are also summarized in the claim. □

**Proof of Corollary 1.5.** This is clear by inspection of formula (1.17) for the transition mechanism of  $Q_{\bullet}$ . □

**Proof of Corollary 1.6.** This follows by comparison of Corollary 1.5 with the transition mechanism of the Yule process described in Proposition 2.3. A more insightful argument can be given as follows, by further development of the argument around (3.14). Blackwell’s renewal theorem



gives  $\lim_{t \rightarrow \infty} \mathbb{E}N_S(t, t + h) = h/\mu_{\log}$ , as well as convergence in distribution as  $t \rightarrow \infty$  of the shifted counting process  $(N_S(t, t + h], h \geq 0)$  to that of the stationary renewal process

$$N_S^*(0, h) := \sum_{i=0}^{\infty} \mathbf{1}(S_i^* \leq h) \quad (h \geq 0), \tag{3.17}$$

where  $(S_z^*, z \in \mathbb{Z})$  is the collection of points of the two-sided stationary renewal point process with spacing distribution that of  $-\log(1 - H)$  and the stationary start  $S_0^*$  assigned the distribution (1.25). Together with the well known reversibility property of  $N_S^*$  and the Yule representation (2.17) of the order statistics of exponential variables, this gives the description of Corollary 1.6 for the limit in distribution of the  $G$ -sequence in Theorem 1.2. The corresponding results for the  $N$ - and  $Q$ -sequences follow easily using the duality between the  $N$ - and  $G$ -sequences described in the introduction, in terms of interleaving the points of the two point processes.  $\square$

The convergence part of the above argument glosses over a few details, but they are technically easier than those dealt with in [17] in a quite similar proof. The present method of looking down from the maximum of exponential order statistics means it is only necessary to deal with convergence in distribution of counts of a one-sided renewal process, looking back from a large random time, which is easily done. Moreover, the limit comes out simply expressed in terms of just one side of the two-sided stationary renewal process, rather than involving both sides of the two-sided process, as in [17].

**Alternative proof of Corollary 1.6.** As indicated already in the introduction, it is quite easy to pass algebraically between the alternate formulations of Theorems 1.1 and 1.2. Gnedin, Iksanov and Roesler already gave an interpretation of their Theorem 1.1 in terms of an interleaving of points of two point processes. However, their point process interpretation of the limiting  $N_\bullet$  sequence is not obviously equivalent to the much simpler point process description of Corollary 1.6. So it seems worth indicating how Corollary 1.6 can nonetheless be derived from the construction of  $N_\bullet$  given in [17]. In [17], Theorem 2.1, and the proof of [17], Lemma 3.4, the limit sequence  $N_\bullet$  was constructed by first exponentiating the two-sided stationary renewal process  $S_z^*, z \in \mathbb{Z}$  on the logarithmic scale to create a scale-invariant (self-similar) point process  $\xi_z^* := \exp(S_z^*), z \in \mathbb{Z}$  with

$$0 < \dots < \xi_{-2}^* < \xi_{-1}^* < 1 < \xi_0^* < \xi_1^* \tag{3.18}$$

together with an independent homogeneous Poisson rate 1 sampling process, with points at say  $0 < \gamma_1 < \gamma_2 < \dots$  with independent standard exponential spacings. The informal description from [17] is to regard the  $\gamma_k$  as the locations of balls thrown into a doubly infinite array of boxes  $(\xi_z^*, \xi_{z+1}^*)$  indexed by  $z \in \mathbb{Z}$ , then let  $(N_0, N_1, \dots)$  be the counts of balls in consecutive boxes, starting with  $N_0$  the count of balls in the first occupied box, that is the box  $(\xi_Z^*, \xi_{Z+1}^*)$  with  $\xi_Z^* < \gamma_1 < \xi_{Z+1}^*$  for the random integer index  $Z$  determined by this condition, ignoring the event of probability zero that there might be any ties between the  $\xi_z$  and  $\gamma_k$ . That is to say

$$N_i := \sum_{k=1}^{\infty} \mathbf{1}(\xi_i < \gamma_k < \xi_{i+1}), \quad \text{where } \xi_i := \xi_{Z+i}^* \quad (i = 0, 1, 2, \dots). \tag{3.19}$$

As argued in the proof of [17], Lemma 3.4, the  $\xi_k, k \geq 1$  may be constructed directly as

$$\xi_k = \frac{\gamma_1}{W_0} \prod_{i=1}^{k-1} (1 - H_i)^{-1}, \tag{3.20}$$

where the  $H_i$  are i.i.d. copies of the basic hazard variable,  $1/W_0$  with  $1/W_0 \stackrel{(d)}{=} \exp(S_0^*)$  is a further independent multiplicative version of the stationary delay variable, and the Poisson points  $\gamma_k$  are assumed to be independent of all these variables. Note that the precise value of  $\xi_0$  is of no importance in (3.19), except for the fact that  $\xi_0 < \gamma_1 < \xi_1$ , so that (3.19) for  $i = 0$  reduces to

$$N_0 := 1 + \sum_{k=2}^{\infty} \mathbf{1}(\gamma_k < \xi_1). \tag{3.21}$$

This formula states that count of balls in the leftmost occupied box is 1 for  $\gamma_1$  plus however many following  $\gamma_k$  are to the left of the first split  $\xi_1$  between boxes. This construction is made difficult by the way the definition of the break points between boxes involves one of the Poisson sample points  $\gamma_1$ , which then affects the distribution of the other Poisson sample points which are all conditioned to be greater than  $\gamma_1$ . However, the construction can be simplified by remarking that the dual sequence  $G_\bullet$  counting box dividers between balls is

$$\begin{aligned} G_i &:= \sum_{k=1}^{\infty} \mathbf{1}(\gamma_i < \xi_k < \gamma_{i+1}) = N_S^*(\log \gamma_i, \log \gamma_{i+1}) \\ &\stackrel{(d)}{=} N_S^*(\log(\gamma_i/\gamma_1), \log(\gamma_{i+1}/\gamma_1)) \quad \text{by stationarity of } N_S^* \\ &\stackrel{(d)}{=} N_S^*(Y_{i-1}, Y_i) \end{aligned} \tag{3.22}$$

according to the representation (2.21) of the birth times  $Y_i$  of a Yule process. Moreover, it is clear that this identity in distribution holds jointly as  $i$  varies. Thus the point process description of  $N_\bullet$  in [17], implies the characterization of  $G_\bullet$  provided in Corollary 1.6, hence also the dual description of  $N_\bullet$  in terms of the Yule process.  $\square$

The proof of Corollaries 1.4 and 1.3 is based on theory of increasing Markov chains recalled in the Appendix.

**Proof of Corollary 1.3.** This follows easily from Proposition A.1.  $\square$

**Proof of Corollary 1.4.** That the limiting gaps  $G_\bullet$  for the GEM(0,  $\theta$ ) model are independent with the indicated geometric distributions can be read from the exact result (1.11) for a sample of size  $n$ , as explained in the introduction. As for the converse assertions, suppose first that the limiting gaps  $G_\bullet$  are geometrically distributed. Then

$$h_j := \mathbb{P}(G_j \geq 1) = 1 - p_{j,j} = 1 - \mu_{0,j}.$$

On the other hand, (3.13) gives  $h_j = (1 - \mu_{0,j})/(j\mu_{\log})$ . Equating these two expressions for  $h_j$  gives for  $j = 1, 2, \dots$

$$\mathbb{E}(1 - H)^j = \mu_{0,j} = \frac{\theta}{\theta + j} = \mathbb{E}(1 - \beta_{1,\theta})^j,$$

where  $\theta := 1/\mu_{\log}$  and  $\beta_{1,\theta}$  has the beta(1,  $\theta$ ) distribution with density  $\theta(1 - u)^{\theta-1}$  at  $0 < u < 1$ . Hence  $H \stackrel{(d)}{=} \beta_{1,\theta}$ .

Suppose next that  $\mathbb{P}(\widehat{G}_{1:n} \geq 1)$  does not depend on  $n$ . Given  $P_\bullet$ , the conditional probability

$$\mathbb{P}(\widehat{G}_{1:n} \geq 1 \mid P_\bullet) = n \sum_{m=1}^{\infty} P_m (P_1 + \dots + P_{m-1})^{n-1}$$

by the decomposition over the maximal value  $m$  which appears only once iff  $\widehat{G}_{1:n} \geq 1$ . This gives for a RAM with i.i.d. factors  $H_1, H_2, \dots$

$$\begin{aligned} \mathbb{P}(\widehat{G}_{1:n} \geq 1) &= n \sum_{m=1}^{\infty} \mathbb{E} \left[ H_m \prod_{i=1}^{m-1} (1 - H_i) \left( 1 - \prod_{i=1}^{m-1} (1 - H_i) \right)^{n-1} \right] \\ &= n\mu_{1,0} \sum_{m=1}^{\infty} \sum_{k=0}^{n-1} \binom{n-1}{k} (-1)^k (\mu_{0,k+1})^{m-1} \\ &= n\mu_{1,0} \sum_{k=0}^{n-1} \binom{n-1}{k} (-1)^k \frac{1}{1 - \mu_{0,k+1}}. \end{aligned} \tag{3.23}$$

By a simple algebra, the equality  $\mathbb{P}(\widehat{G}_{1:n} \geq 1) = \mathbb{P}(\widehat{G}_{1:n+1} \geq 1)$  translates into the following recursion

$$\frac{n+1}{1 - \mu_{0,n+1}} = \sum_{k=0}^{n-1} \binom{n}{k} (-1)^{n-k-1} \frac{k+1}{1 - \mu_{0,k+1}} \tag{3.24}$$

which allows to define  $\mu_{0,n}$  one by one starting from arbitrary  $\mu_{0,1} \in (0, 1)$ . Since the moments  $\mu_{0,n}$  determine the distribution of  $H$ , and in GEM(0,  $\theta$ ) model satisfy (3.24) as follows from (1.11) or can be easily verified directly using (1.23), the claim follows.

Finally, suppose that the limiting gaps  $G_j$  are independent as  $j$  varies. According to Proposition A.2, the jumping chain derived from  $Q_\bullet$  is the strict record chain derived from the initial distribution  $p_{0,j} := \mu_{j,0}/(j\mu_{\log})$ . This implies that the transition matrix  $p$  of  $Q_\bullet$  must satisfy the identity

$$\frac{p_{m,n}}{1 - p_{m,m}} = \frac{p_{0,n}}{1 - p_{0,1} - \dots - p_{0,m}} \quad \text{for } 1 \leq m \leq n. \tag{3.25}$$

Plugging in the formula (1.17) for  $p$ , this becomes

$$\frac{\binom{n-1}{m-1} \mu_{n-m,m}}{1 - \mu_{0,m}} = \frac{\mu_{n,0}/n}{\mu_{\log} - \mu_{1,0}/1 - \dots - \mu_{m,0}/m}. \tag{3.26}$$

Consider now a distribution of  $H$  on  $(0, 1)$  with  $\mathbb{E}(H) = \mu_1$  and  $\mathbb{E} - \log(1 - H) = \mu_{\log}$  for some specified values of  $\mu_1$  and  $\mu_{\log}$  subject to  $0 < \mu_1 < \mu_{\log} < \infty$ . Let  $\theta := \mu_1/(\mu_{\log} - \mu_1)$  and  $\mu_n := \mu_{n,0} = \mathbb{E}H^n$ . Then (3.26) for  $m = 1$  yields easily

$$\mu_n = \frac{n\mu_{n-1}}{\theta + n} \quad (n \geq 2)$$

and hence by induction on  $n$

$$\mu_n = (1 + \theta)\mu_1 \frac{(1)_n}{(1 + \theta)_n} = (1 + \theta)\mu_1 \mathbb{E}\beta_{1,\theta}^n \quad (n \geq 1).$$

Thus  $\mathbb{E}H^n = c\mathbb{E}\beta_{1,\theta}^n$  for every  $n \geq 1$  for a constant factor  $c := (1 + \theta)\mu_1$ . To complete the argument, it only remains to show that  $c = 1$ , which achieved by the following lemma.  $\square$

**Lemma 3.3.** *Let  $\tilde{H}$  and  $H$  be two random variables with*

$$\mathbb{P}(0 < \tilde{H} < 1) = \mathbb{P}(0 < H < 1) = 1 \quad \text{and} \tag{3.27}$$

$$\mathbb{E}\tilde{H}^n = c\mathbb{E}H^n \quad \text{for some constant } c \text{ and all } n = 1, 2, \dots \tag{3.28}$$

Then  $c = 1$  and  $\tilde{H} \stackrel{(d)}{=} H$ .

**Proof.** By switching the roles of  $H$  and  $\tilde{H}$  if  $c > 1$ , it is enough to consider the case when  $0 < c < 1$ . Let  $B_c$  denote a Bernoulli( $c$ ) variable independent of  $H$ . Then

$$\mathbb{E}(B_c H)^n = \mathbb{E}(B_c^n H^n) = \mathbb{E}(B_c H^n) = \mathbb{E}(B_c)\mathbb{E}H^n = c\mathbb{E}H^n \quad (n \geq 1).$$

Because a distribution on  $[0, 1]$  is uniquely determined by its moments, it follows from (3.28) that  $\tilde{H} \stackrel{(d)}{=} B_c H$ , the distribution of which has an atom of magnitude  $1 - c > 0$  at 0. Thus (3.28) for  $0 < c < 1$  violates (3.27), and the conclusion follows.  $\square$

**Remark.** An alternative proof that the limiting gaps  $G_\bullet$  derived from GEM(0,  $\theta$ ) are both independent and geometrically distributed is obtained via Proposition A.2. According to that proposition, it is enough to check that the transition matrix  $p$  of  $Q_\bullet$  satisfies (3.25) for all  $1 \leq m \leq n$ , that is to check (3.26) for the various moments derived from  $H$  with beta(1,  $\theta$ ) distribution. But using the evaluations (1.4) of beta(1,  $\theta$ ) moments, after cancelling common factors, the identity (3.26) reduces to the summation formula

$$\sum_{i=1}^m \frac{(1)_{i-1}\theta}{(1 + \theta)_i} = 1 - \frac{(1)_m}{(1 + \theta)_m} \quad (m = 1, 2, \dots). \tag{3.29}$$

This is easily checked by induction on  $m$ , along with its probabilistic interpretation

$$\mathbb{P}(Q_0 > m) = \frac{\theta}{(1 + \theta)_m} \quad (m = 1, 2, \dots). \tag{3.30}$$

The limit case

$$\sum_{i=1}^{\infty} \frac{(1)_{i-1}\theta}{(1 + \theta)_i} = 1 \quad (\theta > 0) \tag{3.31}$$

is the evaluation of  ${}_2F_1(1, 1; 2 + \theta; z)\theta/(1 + \theta)$  at  $z = 1$  using Gauss’s hypergeometric summation formula  ${}_2F_1(a, b; c; 1) = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}$  ( $\text{Re}(c - b - a) > 0$ ), for  $a = b = 1$  and  $c = 2 + \theta$ .

## 4. Further remarks

### 4.1. Checking the entrance law

Implicit in Theorem 1.2 is the fact that for any distribution of  $H$  on  $(0, 1)$  such that  $\mu_{\log} := \mathbb{E} - \log(1 - H) < \infty$  and  $-\log(1 - H)$  is non-lattice, and for the Markov chain  $Q_\bullet$  with mixed negative binomial transition matrix (1.17) derived from moments of  $H$ , with initial distribution

$$p_{0,n} := \frac{\mu_{n,0}}{n\mu_{\log}}$$

as in (1.18), the hitting probabilities  $h_n := \mathbb{P}(Q_k = n \text{ for some } k \geq 0)$  are given by the formula

$$h_n = \frac{1 - \mu_{0,n}}{n\mu_{\log}} \quad (n \geq 1).$$

In fact, this is the case whenever  $\mu_{\log} < \infty$ , even if  $-\log(1 - H)$  has a lattice distribution. For, by the theory of increasing Markov chains recalled in Proposition A.1, the sequence  $h_n$  is the unique solution of the system of last exit equations

$$h_n = p_{0,n} + \sum_{m=1}^{n-1} \frac{h_m p_{m,n}}{1 - p_{m,m}} \quad (n = 1, 2, \dots)$$

which for the particular mixed negative binomial transition matrix (1.17) expands to

$$\frac{1 - \mu_{0,n}}{n\mu_{\log}} = \frac{\mu_{n,0}}{n\mu_{\log}} + \sum_{m=1}^{n-1} \frac{1}{m\mu_{\log}} \binom{n-1}{m-1} \mu_{n-m,m}.$$

After writing  $\binom{n-1}{m-1} = \frac{m}{n} \binom{n}{m}$ , cancelling common factors, and some rearrangement of terms, this reduces to

$$1 = \sum_{m=0}^n \binom{n}{m} \mu_{n-m,m} = \mathbb{E}[H + (1 - H)]^n$$

which is true for any distribution of  $H$  on  $[0, 1]$  by the binomial theorem.

**4.2. Some checks for  $N_0$  and  $N_1$**

According to Corollary 1.6, the distribution of  $N_0 = Q_0$ , representing the limit distribution of the number of balls  $N_{M_n, n}$  in the last box as  $n \rightarrow \infty$ , and the probability of corresponding events in terms of the limiting gaps  $G_i$ , can be evaluated as

$$\mathbb{P}(N_0 > k) = \mathbb{P}(G_i = 0, 1 \leq i \leq k) = \mathbb{P}[N_S^*(Y_k) = 0] = \mathbb{P}(Y_k < S_0^*) \tag{4.1}$$

for  $k = 1, 2, \dots$ . This probability can be computed by conditioning on  $S_0^*$ , and using the representation  $Y_k \stackrel{(d)}{=} \max_{1 \leq i \leq k} \varepsilon_i$  of (2.17), which makes  $\mathbb{P}(Y_k \leq s) = (1 - e^{-s})^k$ . Thus,

$$\begin{aligned} \mathbb{P}(N_0 > k) &= \frac{1}{\mu_{\log}} \int_0^\infty \mathbb{P}(-\log(1 - H) > s) (1 - e^{-s})^k ds \\ &= \frac{1}{\mu_{\log}} \int_0^1 \mathbb{P}(H > u) \frac{u^k}{1 - u} du, \end{aligned} \tag{4.2}$$

by the change of variables  $u = 1 - e^{-s}$ ,  $ds = du/(1 - u)$ . In the special case of  $\text{GEM}(0, \theta)$ , with  $\mu_{\log} \stackrel{\theta}{=} \theta^{-1}$  and  $\mathbb{P}(H > u) \stackrel{\theta}{=} (1 - u)^\theta$ , this gives easily

$$\mathbb{P}(N_0 > k) \stackrel{\theta}{=} \theta \int_0^1 u^k (1 - u)^{\theta-1} du \stackrel{\theta}{=} \frac{(1)_k}{(1 + \theta)_k}. \tag{4.3}$$

But for the general case it is easier to work with the complementary event. The same change of variables gives

$$\begin{aligned} \mathbb{P}(N_0 \leq k) &= \frac{1}{\mu_{\log}} \int_0^\infty \mathbb{P}(-\log(1 - H) > s) (1 - (1 - e^{-s})^k) ds \\ &= \frac{1}{\mu_{\log}} \int_0^1 \mathbb{P}(H > u) \frac{1 - u^k}{1 - u} du \\ &= \frac{1}{\mu_{\log}} \int_0^1 \mathbb{P}(H > u) (1 + u + \dots + u^{k-1}) du \\ &= \frac{1}{\mu_{\log}} \left( \frac{\mathbb{E}H}{1} + \frac{\mathbb{E}H^2}{2} + \dots + \frac{\mathbb{E}H^k}{k} \right) \end{aligned} \tag{4.4}$$

in agreement with formula (1.18) for  $\mathbb{P}(N_0 = k) = \mathbb{P}(Q_0 = k)$ .

Next, consider the joint limit law of  $N_0$  and  $N_1$  for a general RAM, given  $N_0 := N_Y(S_0^*) = m \geq 1$  say, meaning that the event  $Y_{m-1} < S_0^* < Y_m$  has occurred. The case  $m = 1$  is easiest. Then, as in the previous computation, except that the increment  $N_Y(S_1^*) - N_Y(S_0^*)$  involves  $S_1^* - S_0^* \stackrel{(d)}{=} -\log(1 - H)$  instead of the stationary delay, for  $k \geq 1$

$$\mathbb{P}(N_1 \geq k \mid N_0 = 1) = \int_0^\infty \mathbb{P}(-\log(1 - H) \in ds) (1 - e^{-s})^k = \mathbb{E}H^k \tag{4.5}$$

by the now familiar change of variable. That is to say, the distribution of  $N_1$  given  $N_0 = 1$  is mixed geometric( $1 - H$ ). Given  $N_0 = m$  instead of  $N_0 = 1$ , the situation is similar, except that there are  $m$  independent lines of descent in the Yule process, and the total number of births in the Yule process in all  $m$  lines of descent must be counted before the next renewal time. Now the mixing variable  $1 - H = e^{-(S_1^* - S_0^*)}$  is the same for all lines of descent, so conditional on  $H$  and  $N_0 = m$  the distribution of  $N_1$  is the sum of  $m$  independent geometric( $1 - H$ ) variables. That is to say,  $N_1$  given  $N_0 = m$  has the mixed negative binomial( $m, 1 - H$ ) distribution, as in Corollary 1.5.

### 4.3. Results when $\mu_{\log} = \infty$

It is known that

$$\text{if } \mu_{\log} = \infty \text{ then } \widehat{G}_{i:n} \text{ converges in probability to 0 for every } i, \tag{4.6}$$

corresponding to a piling up of values at the sample maximum  $M_n$ , so the number  $\widehat{N}_{0:n} := N_{M_n:n}$  of ties with the maximum converges in probability to  $\infty$ . See [20,27–29] for further treatment of limit theorems in this case.

### 4.4. Limits for small counts

The simple formula (1.22) for the limiting mean  $\mathbb{E}G_j$  is a companion of similar limit results for counts  $K_{j:n} := \sum_{b=1}^{M_n} \mathbf{1}(N_b = j) = \sum_{i=0}^{M_n-1} \mathbf{1}(\widehat{N}_i = j)$ , which were derived in [17], Lemma 3.2 and Theorem 3.4, and are checked again here:

$$\lim_{n \rightarrow \infty} \mathbb{E}K_{j:n} = \mathbb{E}K_j = \frac{1}{j\mu_{\log}} \quad \text{for } j = 1, 2, \dots, \tag{4.7}$$

$$\lim_{n \rightarrow \infty} \mathbb{E}K_{0:n} = \mathbb{E}K_0 = \frac{\mathbb{E}(-\log H)}{\mu_{\log}}, \tag{4.8}$$

where the limit variables  $K_j$  are encoded in  $N_\bullet$  as

$$K_j := \sum_{i=0}^{\infty} \mathbf{1}(N_i = j) \quad (j = 0, 1, \dots). \tag{4.9}$$

It is known [1] that in the GEM( $0, \theta$ ) model the  $K_j$  for  $j \geq 1$  are independent Poisson variables with means  $\theta/j$  for  $j = 1, 2, \dots$ . It would be interesting to have converses of these properties, similar to Corollary 1.4. It is also known [28] that for a general RAM subject to (1.6) the distribution of  $K_0$  is always mixed Poisson, that is  $K_0 \stackrel{(d)}{=} N(\tau)$  for  $N$  a rate one Poisson process independent of some non-negative random variable  $\tau$ , with  $\tau \stackrel{\theta}{=} \theta |\log \beta_{1,\theta}|$  in the GEM( $0, \theta$ ) case with  $H = \beta_{1,\theta}$ .

### 4.5. Examples

Even without detailing the bijection between the limiting  $N$  and  $G$  sequences, it is easy to express the limits in distribution of various statistics of interest in terms of either of these sequences. Most obviously, the number  $L_n$  of ties with the maximum and the number  $K_{0:n}$  of missing values are encoded in the counts and gaps as

$$L_n = \min\{i : G_{i:n} > 0\} = N_{M_n:n} \quad \text{and} \tag{4.10}$$

$$K_{0:n} = \sum_{i=1}^n (G_{i:n} - 1)_+ = \sum_{r=0}^{M_n-1} \mathbf{1}(N_{M_n-r:n} = 0)$$

and it follows from Theorem 1.2 that

$$(L_n, K_{0:n}) \xrightarrow{(d)} \left( \min\{i : G_i > 0\}, \sum_{i=1}^{\infty} (G_i - 1)_+ \right) = \left( N_0, \sum_{r=0}^{\infty} \mathbf{1}(N_r = 0) \right). \tag{4.11}$$

Here the convergence in distribution of  $L_n$  follows immediately from either of the sequence limit theorems. The convergence in distribution of  $L_n$  to  $N_0$ , and the description (1.18) of its limit law, was given in [17], Theorem 2.1, as well as in [20], Theorem 2.4. The result for  $K_{0:n}$  takes more work, because it is not immediately obvious that its limit can be read as indicated just from convergence of finite dimensional distributions in (1.16). Also, according to [17], Proposition 3.3, unlike the case for  $L_n$ , the limit of  $K_{0:n}$  can be infinite almost surely, and is so iff  $\mathbb{E}[-\log H] = \infty$ . These issues were taken care of in [17] for the representation in terms of the  $N_i$ , and a similar discussion can be provided for the representation above in terms of the  $G_i$ . This  $G$ -representation of  $K_{0:n}$  and its limit in distribution can also be seen in the special case of GEM(0,  $\theta$ ) in [20], (19).

### Appendix: Increasing Markov chains

Consider a Markov chain  $Q_\bullet$  with state space  $\{1, 2, \dots\}$  and stationary transition matrix  $(p_{i,j})$ . Call  $Q_\bullet$  *weakly increasing* if  $p_{i,j} = 0$  for  $j < i$ , so that

$$\mathbb{P}(Q_0 \leq Q_1 \leq \dots) = 1$$

and call  $Q_\bullet$  *strictly increasing* if  $p_{i,j} = 0$  for  $j \leq i$ , so that

$$\mathbb{P}(Q_0 < Q_1 < \dots) = 1.$$

The following proposition recalls some standard facts about such Markov chains [38].

**Proposition A.1.** *Let  $Q_\bullet$  be a weakly increasing Markov chain with state space  $\{1, 2, \dots\}$ , initial distribution*

$$p_{0,j} := \mathbb{P}(Q_0 = j) \tag{A.1}$$



and stationary transition matrix  $(p_{i,j})$ . Suppose that  $Q_\bullet$  has no absorbing states, that is  $p_{j,j} < 1$  for all  $j$ . Let  $G_j := \sum_{k=0}^{\infty} \mathbf{1}(Q_k = j)$  be the number of visits to state  $j$ . Then

- (i)  $G_j$  has the zero-modified geometric distribution with parameters  $(h_j, 1 - p_{j,j})$ , meaning that

$$\mathbb{P}(G_j \geq k) = h_j p_{j,j}^{k-1} \quad (k = 1, 2, \dots), \quad (\text{A.2})$$

where

$$h_j := \mathbb{P}(G_j \geq 1) = \mathbb{P}(Q_k = j \text{ for some } k \geq 0); \quad (\text{A.3})$$

- (ii) the conditional distribution of  $G_j$  given  $G_j \geq 1$  is the geometric  $(1 - p_{j,j})$  distribution on  $\{1, 2, \dots\}$ , so

$$g_j := \mathbb{E}G_j = \sum_{k=1}^{\infty} \mathbb{P}(G_j \geq k) = \frac{h_j}{1 - p_{j,j}}; \quad (\text{A.4})$$

- (iii) the potential sequence  $(g_j, j \geq 1)$  is the unique non-negative solution of the system of equations

$$g_j = p_{0,j} + \sum_{i=1}^{\infty} g_i p_{i,j} \quad (j \geq 1); \quad (\text{A.5})$$

- (iv) equivalently, the sequence of hitting probabilities  $(h_j, j \geq 1)$  is the unique non-negative solution of the system of equations

$$h_j = p_{0,j} + \sum_{i \neq j}^{\infty} \frac{h_i p_{i,j}}{1 - p_{i,i}} \quad (j \geq 1). \quad (\text{A.6})$$

**Proof.** For transient Markov chains, it is well known (e.g., [38], Lemma 1.5.2) that claim (ii) holds with the probability to ever return back from state  $j$  instead of  $p_{j,j}$ ; that this probability is  $p_{j,j}$  for weakly increasing Markov chains is obvious. Then (A.2) follows directly from the definitions and implies (A.4) by summation. The mean number of visits to state  $j$  satisfies (A.5) because one can either start with  $Q_0 = j$ , which happens with probability  $p_{0,j}$ , or come to  $j$  from some state  $i$ , and the mean number of such transitions is  $g_i p_{i,j}$ , so equation (A.5) holds for arbitrary Markov chains with countable state space. The uniqueness of its solution is evident for weakly increasing Markov chains with  $p_{i,j} = 0$  for  $j < i$ . Finally, the change of variables (A.4) implies (A.6) by rearranging terms.  $\square$

Let  $p_{0,\bullet} := (p_{0,j}, j = 1, 2, \dots)$  be a fixed probability distribution with  $p_{0,j} > 0$  for infinitely many  $j$ . Then there is the following well-known construction of two Markov chains derived from  $p_{0,\bullet}$ . The *weak record chain* generated by  $p_{0,\bullet}$  is the sequence of weak upper record values generated by an i.i.d. sequence with distribution  $p_{0,\bullet}$ . The *strict record chain* generated by  $p_{0,\bullet}$  is similarly generated sequence of strict upper record values, which is the sequence of values of the weak record chain watched only when there is a change of value. It is known [47] and easily

verified that the weak record chain is Markov with initial distribution  $p_{0,\bullet}$  and transition matrix

$$p_{i,j}^{\leq} = \frac{p_{0,j}\mathbf{1}(i \leq j)}{p_{0,i} + p_{0,i+1} + \dots} \tag{A.7}$$

while the strict record chain is Markov with initial distribution  $p_{0,\bullet}$  and transition matrix

$$p_{i,j}^{\lt} = \frac{p_{0,j}\mathbf{1}(i < j)}{p_{0,i+1} + p_{0,i+2} + \dots}. \tag{A.8}$$

It is also well known [37], Theorems 16.1, 16.8, that each of these Markov chains derived from  $p_{0,\bullet}$  has the special property that

$$\text{the counts } G_j := \sum_{k=0}^{\infty} \mathbf{1}(Q_k = j) \text{ are independent as } j \text{ varies,} \tag{A.9}$$

where the distribution of  $G_j$  is

- geometric( $1 - p_{j,j}^{\leq}$ ) for the weak record chain, and
- Bernoulli( $p_{j,j}^{\lt}$ ) for the strict record chain.

The following proposition offers a converse:

**Proposition A.2.** *Let  $Q_{\bullet}$  be a weakly increasing Markov chain with transition matrix  $(p_{i,j})$  on the set of positive integers, with initial distribution  $p_{0,\bullet}$  such that  $p_{0,j} > 0$  for infinitely many  $j$ . Suppose that the occupation times  $G_j$  defined by (A.9) satisfy  $\mathbb{P}(G_j < \infty) = 1$  and  $\mathbb{P}(G_j = 0) < 1$  for every  $j$ , and are independent. Then*

$$h_j := \mathbb{P}(G_j \geq 1) = p_{j,j}^{\leq} \tag{A.10}$$

is derived from  $p_{0,\bullet}$  as in (A.7), which implies

$$\prod_{j=1}^{\infty} (1 - h_j) = 0. \tag{A.11}$$

Moreover, the transition matrix  $(p_{i,j})$  of  $Q_{\bullet}$  is of the form

$$p_{i,j} = \mathbf{1}(j = i)p_{i,i} + \mathbf{1}(j > i)(1 - p_{i,i})h_j \prod_{k=i+1}^{j-1} (1 - h_k) \tag{A.12}$$

$$= \mathbf{1}(j = i)p_{i,i} + \mathbf{1}(j > i)(1 - p_{i,i})p_{i,j}^{\lt} \tag{A.13}$$

for some arbitrary sequence of self-transition probabilities  $p_{i,i}$  with  $p_{i,i} < 1$ , where  $p_{i,j}^{\lt}$  is defined by (A.8). In particular, continuing to assume that the  $G_j$  are independent, the following three conditions are equivalent:

- $p_{i,i} = h_i$  for all  $i$ ;
- each of the  $G_j$ 's is geometrically distributed on  $\{0, 1, 2, \dots\}$ ;
- $Q_\bullet$  is the weak record chain derived from  $p_{0,\bullet}$ ;

and so too are the following three conditions

- $p_{i,i} = 0$  for all  $i$ ;
- each of the  $G_j$ 's has a Bernoulli distribution on  $\{0, 1\}$ ;
- $Q_\bullet$  is the strict record chain derived from  $p_{0,\bullet}$ .

In any case, whatever the self-transition probabilities  $p_{i,i}$ , the chain  $Q_\bullet$  watched only when it changes state is a copy of the strict record chain derived from  $p_{0,\bullet}$ .

**Proof.** By the general theory of transient Markov chains reviewed in Proposition A.1, the distribution of  $G_\bullet$  is that of a sequence of random variables with zero-modified  $\text{geometric}(h_j, 1 - p_{j,j})$  distributions, as displayed in (A.2). The assumed independence of the counts  $G_\bullet$  implies that for  $j \geq 1$

$$\mathbb{P}(Q_0 > j) = \mathbb{P}(G_1 = 0, \dots, G_{j-1} = 0, G_j = 0) \quad (\text{A.14})$$

$$= \prod_{i=1}^j (1 - h_i). \quad (\text{A.15})$$

The assumption that  $\mathbb{P}(Q_0 < \infty)$  then gives (A.11). On the other hand, it is clear that the entire path of the weakly increasing Markov chain  $Q_\bullet$  can be recovered with probability one from its sequence of occupation times  $G_\bullet$ . So the transition matrix of  $Q_\bullet$  is a function of the two sequences  $(h_j)$  and  $(p_{j,j})$  which should be chosen consistently with the initial distribution to make counts  $G_\bullet$  independent. If the chain jumps from  $i$  to  $j > i$  it means that it does not stay at  $i$ , which happens with probability  $1 - p_{i,i}$ , and given this event that  $G_{i+1} = \dots = G_{j-1} = 0$  and  $G_j > 0$ , so (A.12) follows from independence of the counts  $G_\bullet$ . Then a tedious but straightforward calculation shows by induction on  $j$  that  $h_j = p_{j,j}^{\leq}$  provides a solution for (A.6), so (A.10) holds and gives (A.13) after some algebra. Finally, both assertions about equivalence of three conditions and the last assertion of the proposition now follow easily from comparison of (A.2), (A.7) and (A.8).  $\square$

## 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](#)
- [2] Athreya, K.B. and Ney, P.E. (1972). *Branching Processes. Die Grundlehren der Mathematischen Wissenschaften* **176**. New York: Springer. [MR0373040](#)
- [3] Birkner, M., Geiger, J. and Kersting, G. (2005). Branching processes in random environment—A view on critical and subcritical cases. In *Interacting Stochastic Systems* 269–291. Berlin: Springer. [MR2118578](#)
- [4] Bogachev, L.V., Gnedin, A.V. and Yakubovich, Y.V. (2008). On the variance of the number of occupied boxes. *Adv. in Appl. Math.* **40** 401–432. [MR2412154](#)

- [5] Borel, É. (1947). Sur les développements unitaires normaux. *C. R. Acad. Sci. Paris* **225** 51. [MR0023007](#)
- [6] Bruss, F.T. and Rogers, L.C.G. (1991). Pascal processes and their characterization. *Stochastic Process. Appl.* **37** 331–338. [MR1102879](#)
- [7] Carmona, P., Petit, F. and Yor, M. (1998). Beta–gamma random variables and intertwining relations between certain Markov processes. *Rev. Mat. Iberoam.* **14** 311–367. [MR1654531](#)
- [8] Daley, D.J. and Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes. Springer Series in Statistics*. New York: Springer. [MR0950166](#)
- [9] Donnelly, P. and Tavaré, S. (1986). The ages of alleles and a coalescent. *Adv. in Appl. Probab.* **18** 1–19. [MR0827330](#)
- [10] Duchamps, J.-J., Pitman, J. and Tang, W. Renewal sequences and record chains related to multiple zeta sums. *Trans. Amer. Math. Soc.* Published online: September 18, 2018.
- [11] Engen, S. (1975). A note on the geometric series as a species frequency model. *Biometrika* **62** 697–699. [MR0411097](#)
- [12] Erdős, P., Rényi, A. and Szűsz, P. (1958). On Engel’s and Sylvester’s series. *Ann. Univ. Sci. Budapest. Eötvös Sect. Math.* **1** 7–32. [MR0102496](#)
- [13] Feigin, P.D. (1979). On the characterization of point processes with the order statistic property. *J. Appl. Probab.* **16** 297–304. [MR0531764](#)
- [14] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, Vol. I*, 3rd ed. New York: Wiley. [MR0228020](#)
- [15] Gnedin, A., Hansen, B. and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws. *Probab. Surv.* **4** 146–171. [MR2318403](#)
- [16] Gnedin, A., Iksanov, A. and Marynych, A. (2010). The Bernoulli sieve: An overview. In *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA’10)*. *Discrete Math. Theor. Comput. Sci. Proc.*, AM 329–341. Nancy: Assoc. Discrete Math. Theor. Comput. Sci. [MR2735350](#)
- [17] Gnedin, A., Iksanov, A. and Roesler, U. (2008). Small parts in the Bernoulli sieve. In *Fifth Colloquium on Mathematics and Computer Science*. *Discrete Math. Theor. Comput. Sci. Proc.*, AI 235–242. Nancy: Assoc. Discrete Math. Theor. Comput. Sci. [MR2508790](#)
- [18] Gnedin, A. and Pitman, J. (2005). Regenerative composition structures. *Ann. Probab.* **33** 445–479. [MR2122798](#)
- [19] Gnedin, A.V. (2004). The Bernoulli sieve. *Bernoulli* **10** 79–96. [MR2044594](#)
- [20] Gnedin, A.V., Iksanov, A.M., Negadajlov, P. and Rösler, U. (2009). The Bernoulli sieve revisited. *Ann. Appl. Probab.* **19** 1634–1655. [MR2538083](#)
- [21] Grandell, J. (1997). *Mixed Poisson Processes. Monographs on Statistics and Applied Probability* **77**. London: CRC Press. [MR1463943](#)
- [22] Greenwood, P. and Pitman, J. (1980). Construction of local time and Poisson point processes from nested arrays. *J. Lond. Math. Soc. (2)* **22** 182–192. [MR0579823](#)
- [23] Halmos, P.R. (1944). Random alms. *Ann. Math. Stat.* **15** 182–189. [MR0010342](#)
- [24] Hambly, B. (1992). On the limiting distribution of a supercritical branching process in a random environment. *J. Appl. Probab.* **29** 499–518. [MR1174427](#)
- [25] Hunt, G.A. (1960). Markoff chains and Martin boundaries. *Illinois J. Math.* **4** 313–340. [MR0123364](#)
- [26] Ignatov, T. (1982). A constant arising in the asymptotic theory of symmetric groups, and Poisson–Dirichlet measures. *Teor. Veroyatn. Primen.* **27** 129–140. [MR0645134](#)
- [27] Iksanov, A. (2012). On the number of empty boxes in the Bernoulli sieve II. *Stochastic Process. Appl.* **122** 2701–2729. [MR2926172](#)
- [28] Iksanov, A. (2013). On the number of empty boxes in the Bernoulli sieve I. *Stochastics* **85** 946–959. [MR3176494](#)

- [29] Iksanov, A.M., Marynych, A.V. and Vatutin, V.A. (2015). Weak convergence of finite-dimensional distributions of the number of empty boxes in the Bernoulli sieve. *Theory Probab. Appl.* **59** 87–113. [MR3416065](#)
- [30] Johnson, N.L., Kemp, A.W. and Kotz, S. (2005). *Univariate Discrete Distributions*, 3rd ed. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley Interscience. [MR2163227](#)
- [31] Kallenberg, O. (2017). *Random Measures, Theory and Applications. Probability Theory and Stochastic Modelling* **77**. Cham: Springer. [MR3642325](#)
- [32] Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *J. Math. Mech.* **17** 373–401. [MR0216548](#)
- [33] Kendall, D.G. (1966). Branching processes since 1873. *J. Lond. Math. Soc.* **41** 385–406 (1 plate). [MR0198551](#)
- [34] Lévy, P. (1947). Remarques sur un théorème de M. Émile Borel. *C. R. Acad. Sci. Paris* **225** 918–919. [MR0023008](#)
- [35] Lundberg, O. (1940). On random processes and their application to sickness and accident statistics. Ph.D. thesis, Univ. Stockholm, Uppsala. [MR0003509](#)
- [36] Neuts, M.F. and Resnick, S.I. (1971). On the times of births in a linear birthprocess. *J. Aust. Math. Soc.* **12** 473–475. [MR0293732](#)
- [37] Nevzorov, V.B. (2001). *Records: Mathematical Theory. Translations of Mathematical Monographs* **194**. Providence, RI: Amer. Math. Soc. Translated from the Russian manuscript by D.M. Chibisov. [MR1791071](#)
- [38] Norris, J.R. (1998). *Markov Chains. Cambridge Series in Statistical and Probabilistic Mathematics* **2**. Cambridge: Cambridge Univ. Press. Reprint of 1997 original. [MR1600720](#)
- [39] Pitman, J. and Yakubovich, Y. (2017). Extremes and gaps in sampling from a GEM random discrete distribution. *Electron. J. Probab.* **22** Paper No. 44, 26. [MR3646070](#)
- [40] Pitman, J. and Yor, M. (2001). On the distribution of ranked heights of excursions of a Brownian bridge. *Ann. Probab.* **29** 361–384. [MR1825154](#)
- [41] Pitman, J.W. (1975). One-dimensional Brownian motion and the three-dimensional Bessel process. *Adv. in Appl. Probab.* **7** 511–526. [MR0375485](#)
- [42] Rényi, A. (1953). On the theory of order statistics. *Acta Math. Acad. Sci. Hung.* **4** 191–231. [MR0061792](#)
- [43] Sawyer, S. and Hartl, D. (1985). A sampling theory for local selection. *J. Genet.* **64** 21–29.
- [44] Smith, W.L. and Wilkinson, W.E. (1969). On branching processes in random environments. *Ann. Math. Stat.* **40** 814–827. [MR0246380](#)
- [45] Stanley, R.P. (2012). *Enumerative Combinatorics, Vol. 1*, 2nd ed. *Cambridge Studies in Advanced Mathematics* **49**. Cambridge: Cambridge Univ. Press. [MR2868112](#)
- [46] Sukhatme, P.V. (1937). Tests of significance for samples of the  $\chi^2$ -population with two degrees of freedom. *Ann. Eugen.* **8** 52–56.
- [47] Vervaat, W. (1973). Limit theorems for records from discrete distributions. *Stochastic Process. Appl.* **1** 317–334. [MR0362457](#)
- [48] Waugh, W.A.O’N. (1970). Transformation of a birth process into a Poisson process. *J. Roy. Statist. Soc. Ser. B* **32** 418–431. [MR0287620](#)

Received April 2018