The Bernoulli sieve: an overview

Alexander Gnedin[1] Alexander Iksanov[2] Alexander Marynych[2]

[1]*Department of Mathematics, Utrecht University, Postbus 80010, 3508 TA Utrecht, The Netherlands*
[2]*Faculty of Cybernetics, National T. Shevchenko University of Kiev, 01033 Kiev, Ukraine*

The Bernoulli sieve is a version of the classical balls-in-boxes occupancy scheme, in which random frequencies of infinitely many boxes are produced by a multiplicative random walk, also known as the residual allocation model or stick-breaking. We give an overview of the limit theorems concerning the number of boxes occupied by some balls out of the first n balls thrown, and present some new results concerning the number of empty boxes within the occupancy range.

Keywords: the occupancy problem, random partitions

1 Introduction

In a classical occupancy scheme n balls are thrown independently in an infinite array of boxes with probability p_k of hitting box $k = 1, 2, \ldots$, where $(p_k)_{k\in\mathbb{N}}$ is a fixed sequence of positive frequencies summing up to one. The quantities of traditional interest are

- K_n the number of boxes occupied by at least one of n balls,
- $K_{n,r}$ the number of boxes occupied by exactly r out of n balls,
- M_n the range of occupancy, equal to the maximal index of occupied box,
- $L_n := M_n - K_n$ the number of empty boxes within the occupancy range,
- Z_n the number of balls in the M_nth box.

In applications 'boxes' are clusters, species, types of data etc. The quantities in the list characterise the sample variability, which for large n is dominantly determined by the boxes occupied by a few balls, thus determined by the way the frequencies p_k approach zero as $k \to \infty$. The first two variables are functionals of the induced partition of n, defined as the unordered collection of positive occupancy counts.

The *Bernoulli sieve* is a version of the occupancy scheme with random frequencies

$$p_k := W_1 W_2 \cdots W_{k-1}(1 - W_k), \quad k \in \mathbb{N}, \tag{1}$$

where $(W_k)_{k\in\mathbb{N}}$ are independent copies of a random variable W taking values in $(0, 1)$. The name derives from the following recursive construction based on i.i.d. $q_k =_d 1 - W$: at round 1 a coin with probability

q_1 for heads is flipped for each of n balls and every time it turns heads the ball is put in box 1, then at round 2 a coin with probability q_2 for heads is flipped for each of the remaining balls and every time it turns heads the ball is sent to box 2, and so on until all balls are allocated in boxes.

It is useful to identify frequencies (1) with the lengths of component intervals induced by splitting $[0,1]$ at points visited by a multiplicative random walk $(Q_k)_{k\in\mathbb{N}_0}$, where

$$Q_0 := 1, \ \ Q_j := \prod_{i=1}^{j} W_i, \ \ j \in \mathbb{N}.$$

In the spirit of Kingman's 'paintbox representation' of exchangeable partitions [18], we may identify the boxes with open intervals (Q_k, Q_{k-1}), and mark the balls by independent points $U_1, \dots, U_n$ sampled from the uniform $[0,1]$ distribution, independently of (Q_k). The event $U_i \in (Q_{k-1}, Q_k)$ then means that ball i falls in box k. Keep in mind that in the natural order the intervals are indexed from the right to the left, thus the occupancy range is determined by the interval containing the leftmost mark $\min(U_1, \dots, U_n)$.

The Bernoulli sieve has nonrandom frequencies only when the law of W is a Dirac mass δ_p located at some $p \in (0,1)$, the frequencies p_k comprise then a geometric distribution. Results for this case can be readily recast from the numerous studies on sampling from the geometric distribution [5, 6, 19, 25] and related models like the leader election algorithms [3, 11, 22, 28], absorption sampling [7, 24] etc. It is known that asymptotic expansions of the moments of K_n, M_n and many other quantities have a component that oscillates periodically on the $\log n$-scale with a small amplitude [11, 27]. The same applies to distributions of the L_n's [19, 26]. There are some peculiarities in the symmetric case $p = 1/2$ [11, 28].

The best analytically tractable case involves random factors having beta$(\theta, 1)$ density $\mathbb{P}\{W \in \mathrm{d}x\} = \theta x^{\theta-1}\mathrm{d}x$ on $(0,1)$ with parameter $\theta > 0$. In this case the Bernoulli sieve may be viewed as a way to generate a random partition of n which follows the multivariate distribution known as the Ewens sampling formula [1]. This model has been widely studied in connection with problems of combinatorics, statistics and biology. In particular, the case $\theta = 1$ of uniform factors is related to records and cycle patterns of random permutations under the uniform distribution on the symmetric group. It is well known [1] that $(K_n - \theta \log n)/(\theta \log n)^{1/2}$ is asymptotically normal, and that the $K_{n,r}$'s converge jointly to independent Poisson(θ/r) random variables. These classical results are complemented by the observation that M_n exhibits the same asymptotics of moments and distribution as K_n, and the number of empty boxes has the following surprising limit law:

Theorem 1.1 [16] *If W has* $\mathrm{beta}(\theta, 1)$ *distribution then $L_n \to_d L_\infty$, where L_∞ has probability generating function*

$$\mathbb{E}s^{L_\infty} = \frac{\Gamma(1+\theta)\Gamma(1+\theta-\theta s)}{\Gamma(1+2\theta-\theta s)}, \quad s \in [0,1],$$

which corresponds to a mixed Poisson distribution with the parameter distributed like $\theta\,|\log(1-W)|$.

Throughout we shall use the following notation for the moments

$$\mu := \mathbb{E}|\log W|, \quad \sigma^2 := \mathrm{Var}\,(\log W), \quad \nu := \mathbb{E}|\log(1-W)|,$$

which may be finite or infinite. The *standing assumption* for what follows is that the distribution of $|\log W|$ is non-lattice. In particular, the case of sampling from the geometric distribution will be excluded.

2 Markov chains and distributional recursions

A random combinatorial structure which captures the occupancy of boxes by n indistinguishable balls is the *weak composition* C_n^* comprised of nonnegative integer parts summing up to n. The term *weak* composition means that zero parts are allowed, for instance, the sequence $(2,3,0,1,0,0,1,0,0,0,\ldots)$ (padded by infinitely many 0's) is a possible value of C_7^*. A related structure which contains less information is a composition C_n obtained by discarding zero parts of C_n^*. Discarding further the order of parts in C_n yields a random partition of n. The parts of C_n^* can be represented (see [18, p. 452]) as the magnitudes of jumps of a time-homogeneous nonincreasing Markov chain $Q_n^* = (Q_n^*(k))_{k\in\mathbb{N}_0}$ on integers, which starts at n and moves from n to m with transition probabilities

$$q^*(n,m) = \binom{n}{m}\mathbb{E}(1-W)^{n-m}W^m, \quad m = 0,\ldots,n.$$

In the same direction, parts of the composition C_n are the magnitudes of jumps of a Markov chain $Q_n = (Q_n(k))_{k\in\mathbb{N}_0}$ with transition probabilities

$$q(n,m) = \binom{n}{m}\frac{\mathbb{E}(1-W)^{n-m}W^m}{1-\mathbb{E}W^n}, \quad m = 0,\ldots,n-1.$$

This Markovian realisation implies the following distributional recursions (see [16, Section 3]):

$$M_0 = 0, \quad M_n =_d M_{Q_n^*(1)} + 1, \quad n \in \mathbb{N},$$

$$K_0 = 0, \quad K_n =_d K_{Q_n(1)} + 1, \quad n \in \mathbb{N}, \tag{2}$$

$$L_0 = 0, \quad L_n =_d L_{Q_n^*(1)} + 1_{\{Q_n^*(1)=n\}}, n \in \mathbb{N}, \tag{3}$$

where in the right-hand side $Q_n^*(1)$ is assumed independent of $\{M_n : n \in \mathbb{N}\}$ and $\{L_n : n \in \mathbb{N}\}$, and $Q_n(1)$ independent of $\{K_n : n \in \mathbb{N}\}$. Analysis of the recursions by known direct methods is difficult, as these impose restrictive conditions on the moments of $Q_n(1)$ or $Q_n^*(1)$. Nevertheless, coupling with the multiplicative random walk allows to gain a lot of information about the compositions. For instance, let $g(n,m)$ be the potential function, equal to the probability that Q_n ever visits state m,

$$g(n,m) = \sum_{j=0}^{\infty}\mathbb{P}\{Q_n(j) = m\}.$$

The coupling implies that ([12, Proposition 5])

$$\lim_{n\to\infty} g(n,m) = \frac{1-\mathbb{E}W^m}{\mu m}, \tag{4}$$

which is 0 if $\mu = \infty$.

The coupling readily implies stochastic subadditivity $M_{n+m} <_d M_n + M_m'$ where the terms in the right-hand side are independent. Indeed, note first that M_n is nondecreasing. Now, when n balls have been allocated within the range M_n, adding m new balls leads to (stochastically) maximal increase of the occupancy range when all m fall outside the old range M_n, in which event the new range of occupancy is distributed like $M_n + M_m'$. With analogous notation, $L_{n+m} <_d L_n + L_m'$ for exactly the same reason (although L_n is not monotone).

3 Asymptotics of M_n

Passing from the multiplicative to conventional (additive) random walk we introduce

$$S_0 := 0, \;\; S_k := |\log W_1| + \ldots + |\log W_k|, \;\; k \in \mathbb{N}. \tag{5}$$

In this scenario the Bernoulli sieve can be defined as allocation of balls with exponentially distributed marks $E_j = -\log U_j, \; 1 \le j \le n$, in boxes $(S_k, S_{k+1}), \; k \in \mathbb{N}_0$. Define

$$N_t := \inf\{k \ge 1 : S_k > t\}, \;\; t \ge 0, \tag{6}$$

which is the first time (S_k) enters (t, ∞). From the extreme-value theory we know that the maximum statistic $T_n := \max(E_1, \ldots, E_n))$ satisfies $T_n - \log n \to_d T$, where T has the standard Gumbel distribution $\mathbb{P}\{T \le x\} = \exp(-e^{-x})$, $x \in \mathbb{R}$. A key observation is that

$$M_n = N_{T_n},$$

thus the asymptotic behaviour of M_n is very much the same as that of $N_{\log n}$, and the latter can be concluded by means of the renewal theory. A complete description of possible limit laws and scaling/centering constants for the number of renewals N_t [16, Proposition A.1] leads to the following classification of possible limit laws for M_n.

Theorem 3.1 [16] *The following assertions are equivalent:*

(i) *There exist sequences $\{a_n, b_n : n \in \mathbb{N}\}$ with $a_n > 0$ and $b_n \in \mathbb{R}$ such that, as $n \to \infty$, the variable $(M_n - b_n)/a_n$ converges weakly to some non-degenerate and proper distribution.*

(ii) *The distribution of $|\log W|$ either belongs to the domain of attraction of a stable law, or the function $\mathbb{P}\{|\log W| > x\}$ slowly varies at ∞.*

Accordingly, there are five possible modes of convergence:

(a) *If $\sigma^2 < \infty$ then, with constants $b_n = \mu^{-1} \log n$ and $a_n = (\mu^{-3}\sigma^2 \log n)^{1/2}$, the limiting distribution of $(M_n - b_n)/a_n$ is standard normal.*

(b) *If $\sigma^2 = \infty$, and*

$$\int_0^x y^2 \, \mathbb{P}\{|\log W| \in \mathrm{d}y\} \;\sim\; L(x) \;\; x \to \infty,$$

for some function L slowly varying at ∞, then, with $b_n = \mu^{-1} \log n$ and $a_n = \mu^{-3/2} c_{[\log n]}$, where $c(x)$ is any positive function satisfying $\lim_{x\to\infty} xL(c(x))/c^2(x) = 1$, the limiting distribution of $(M_n - b_n)/a_n$ is standard normal.

(c) *If*

$$\mathbb{P}\{|\log W| > x\} \;\sim\; x^{-\alpha} L(x), \;\; x \to \infty, \tag{7}$$

for some L slowly varying at ∞ and $\alpha \in (1, 2)$ then, with $b_n = \mu^{-1} \log n$ and $a_n = \mu^{-\frac{\alpha+1}{\alpha}} c_{\log n}$, where $c(x)$ is any positive function satisfying $\lim_{x\to\infty} xL(c(x))/c^{\alpha}(x) = 1$, the limiting distribution of $(M_n - b_n)/a_n$ is α-stable with characteristic function

$$t \mapsto \exp\{-|t|^{\alpha}\Gamma(1-\alpha)(\cos(\pi\alpha/2) + i \sin(\pi\alpha/2)\,\mathrm{sgn}(t))\}, \; t \in \mathbb{R}.$$

(d) *Assume that the relation* (7) *holds with* $\alpha = 1$. *Let* $r : \mathbb{R} \to \mathbb{R}$ *be any nondecreasing function such that* $\lim_{x\to\infty} x\mathbb{P}\{|\log W| > r(x)\} = 1$ *and set*

$$m(x) := \int_0^x \mathbb{P}\{|\log W| > y\}\mathrm{d}y, \;\; x > 0.$$

Then, with $b_n = \log n/(m(\log n/r(m(\log n))))$ *and*

$$a_n := \frac{r(\log n/m(\log n))}{m(\log n)},$$

the limiting distribution of $(M_n - b_n)/a_n$ *is* 1*-stable with characteristic function*

$$t \mapsto \exp\{-|t|(\pi/2 - i\log|t|\,\mathrm{sgn}(t))\}, \; t \in \mathbb{R}.$$

(e) *If the relation* (7) *holds for* $\alpha \in [0,1)$ *then, with* $b_n \equiv 0$ *and* $a_n := \log^\alpha n/L(\log n)$, *the limiting distribution of* M_n/a_n *is the Mittag-Leffler law* θ_α *with moments*

$$\int_0^\infty x^k\, \theta_\alpha(\mathrm{d}x) \;=\; \frac{k!}{\Gamma^k(1-\alpha)\Gamma(1+\alpha k)}, \;\; k \in \mathbb{N}.$$

4 Asymptotics of K_n

Loosely speaking, ν controls the mean number of empty boxes, so that $\nu < \infty$ implies $\lim_{n\to\infty} \mathbb{E}L_n = \nu/\mu < \infty$ (Theorem 7.1 to follow). Thus when $\nu < \infty$ the identity $K_n = M_n - L_n$ suggests that K_n does not differ much from M_n. A first result of this kind was obtained in [12]: assuming $\nu < \infty$ and $\sigma^2 < \infty$ it was shown that

$$(K_n - \mu^{-1}\log n)/\sqrt{\sigma^2\mu^{-3}\log n} \to_d \text{normal}\,(0,1), \;\; n \to \infty.$$

The proof was based on a careful analysis of the recursion (2) to conclude on the asymptotics of $\mathrm{Var}\, K_n$ and to eventually prove the normal limit.

The similarity between M_n and K_n was justified in full generality in [16], where is was shown that under the assumption $\nu < \infty$ Theorem 3.1 remains valid if M_n is replaced by K_n.

Another approach which allows one to treat the cases of finite and infinite ν in a unified way was proposed in [15]. It was suggested to approximate K_n by $N^*(\log n)$, where

$$\begin{aligned} N^*(x) \;&:=\; \#\{k \in \mathbb{N} : p_k \ge e^{-x}\} \\ &=\; \#\{k \in \mathbb{N} : W_1 \cdots W_{k-1}(1 - W_k) \ge e^{-x}\}, \quad x > 0. \end{aligned}$$

The connection exemplifies the general idea that the variability of K_n stems from randomness in frequencies (p_k) superposed with randomness in sampling, and the first often plays a dominating role through the conditional law of large numbers $K_n \sim \mathbb{E}(K_n \mid (p_k))$ a.s. (see [23]). Thus we believe that the approach based on $N^*(x)$ offers a natural and the most adequate way to study the asymptotics of K_n. The following result was proved in [15].

Theorem 4.1 *If there exist functions $f : \mathbb{R}_+ \to \mathbb{R}_+$ and $g : \mathbb{R}_+ \to \mathbb{R}$ such that $(N_t - g(t))/f(t)$ converges weakly (as $t \to \infty$) to some non-degenerate and proper distribution, then also $(K_n - b_n)/a_n$ converge weakly (as $n \to \infty$) to the same distribution, where the constants are given by*

$$b_n = \int_0^{\log n} g(\log n - y)\,\mathbb{P}\{|\log(1-W)| \in \mathrm{d}y\}, \quad a_n = f(\log n).$$

As in [16], the convergence criterion for N_t leads to a complete characterisation of possible normalisations and limiting laws for K_n, see Corollary 1.1 in [15]. But Theorem 4.1 says more: if $\nu = \infty$ the behaviour of L_n may affect the asymptotics of $K_n = M_n - L_n$. The following example illustrates the phenomenon.

Example 4.1 Assume that, for some $\gamma \in (0, 1/2)$,

$$\mathbb{P}\{W > x\} = \frac{1}{1 + |\log(1-x)|^\gamma}, \;\; x \in [0,1).$$

Then

$$\mathbb{E}\log^2 W < \infty \;\text{ and }\; \mathbb{P}\{|\log(1-W)| > x\} \sim x^{-\gamma} \quad \text{as } x \to \infty,$$

and in this case,

$$a_n = \text{const}\log^{1/2} n \;\text{ and }\; d_n = \mu^{-1}(\log n - (1-\gamma)^{-1}\log^{1-\gamma} n + o(\log^{1-\gamma} n)).$$

Thus we see that the second term $d_n - \mu^{-1}\log n$ of centering cannot be ignored. Moreover, one can check that

$$\mathbb{E}L_n \sim \frac{1}{\mu}\sum_{k=1}^n \frac{\mathbb{E}W^k}{k} \;\sim\; b_n - \mu^{-1}\log n \;\sim\; \frac{1}{\mu(1-\gamma)}\log^{1-\gamma} n,$$

which reveals the indispensable contribution of L_n.

5 Weak convergence of $K_{n,r}$

Assume $\mu < \infty$. For $B := \{\prod_{i=1}^k W_i : k \in \mathbb{N}_0\}$ the set of sites visited by the multiplicative random walk, consider a point process with unit atoms located at points of $-\log B$ (which are the sites visited by $S_k,\ k \in \mathbb{N}_0$). By the renewal theorem the point process $-\log B - \log n$ vaguely converges to a shift-invariant renewal process $\mathcal{P}$ on the whole line. Therefore, the point process nB converges vaguely to a point process $\mathcal{B} := \exp(-\mathcal{P})$ on $\mathbb{R}_+$. Think of intervals between consequitive points of $\mathcal{B}$ as a series of boxes. Note that the process is self-similar, meaning that $c\mathcal{B} =_d \mathcal{B}$ for every $c > 0$, and has the intensity measure $(\mu x)^{-1}\mathrm{d}x$, so the atoms accumulate at 0 and ∞. In the role of balls assume the points of a unit Poisson process $\mathcal{U}$ independent of $\mathcal{B}$. A well-known fact of extreme value theory is that $\mathcal{U}$ is the vague limit of the point process with unit atoms located at $nU_j,\ 1 \le j \le n$. The location of the leftmost atom of $\mathcal{U}$, say Y, has exponential distribution. For $r \ge 0$ define $\hat{K}_r$ to be the number of component intervals of $(Y, \infty) \setminus \mathcal{B}$ that contain exactly r atoms of $\mathcal{U}$. The existence of weak limits for the occupancy counts is read off from the convergence of point processes:

Theorem 5.1 [17] *As $n \to \infty$ we have the joint convergence in distribution*

$$(L_n, K_{n,1}, K_{n,2}, \ldots) \to_d (\hat{K}_0, \hat{K}_1, \hat{K}_2, \ldots)$$

along with

$$\mathbb{E}K_{n,r} \to \mathbb{E}\hat{K}_r = \frac{1}{r\mu},\ r > 0.$$

When $W =_d \mathrm{beta}(\theta, 1)$ the process $\mathcal{B}$ is Poisson with intensity $\theta x^{-1}\mathrm{d}x$. By self-similarity, the partition induced by allocation of n leftmost atoms of $\mathcal{U}$ is the Ewens partition. The theorem allows to re-prove the results on asymptotics of the Ewens partition mentioned in Introduction, along with Theorem 1.1. Except the $\mathrm{beta}(\theta, 1)$ case no explicit formulas for the distribution of occupancy counts are known; in general the $\hat{K}_r$'s are neither independent, nor Poisson. See more on self-similar partitions in [13, Section 5].

6 Asymptotics of Z_n

The variable Z_n is analogous to the number of winners in the leader election algorithm [4, 5, 6, 25].

Theorem 6.1 [16] *The number of balls in the last occupied box satisfies:*

(1) *If $\mu < \infty$ then $Z_n \to_d Z$, $n \to \infty$, where the variable Z has distribution*

$$\mathbb{P}\{Z = k\} = \frac{\mathbb{E}(1 - W)^k}{\mu k},\ \ k \in \mathbb{N}.$$

(2) *If* (7) *holds with $\alpha \in [0, 1)$ then*

$$\frac{\log Z_n}{\log n} \to_d Z^{(\alpha)},\ \ n \to \infty,$$

where the law of $Z^{(0)}$ is δ_1, while for $\alpha \in (0, 1)$ we have $Z^{(\alpha)} =_d \mathrm{beta}(1 - \alpha, \alpha)$.

(3) *If* (7) *holds with $\alpha = 1$ and $\mu = \infty$, then*

$$\frac{m(\log Z_n)}{m(\log n)} \to_d Z^{(1)},\ \ n \to \infty,$$

where $m(x) = \int_0^x \mathbb{P}\{|\log W| > y\}\mathrm{d}y$, and $Z^{(1)} =_d \mathrm{uniform}[0, 1]$.

The case $\mu < \infty$ is quite elementary, as is seen from

$$\mathbb{P}\{Z_n = m\} = g(n, m)\mathbb{P}\{Q_m(1) = 0\} = g(n, m)\frac{\mathbb{E}(1 - W)^m}{1 - \mathbb{E}W^m} \tag{8}$$

and (4). In the case $\mu = \infty$ the result follows from the known limit distribution of the undershoot $U(z) = z - S_{N(z)-1}$ (see [8, 10]) and the representation

$$\mathbb{P}\{Z_n > k\} = \mathbb{P}\{U(E_{n,n}) > E_{n,n} - E_{n-k,n}\},\ \ k \in \mathbb{N},$$

where $E_{1,n} \le \ldots \le E_{n,n} = T_n$ are the order statistics of the exponential variables $E_j,\ 1 \le j \le n$.

7 Asymptotics of L_n

Although there is an explicit formula

$$\mathbb{E}L_n = \sum_{k=1}^{n}(-1)^{k+1}\binom{n}{k}\frac{1-\mathbb{E}(1-W)^k}{1-\mathbb{E}W^k}, \tag{9}$$

it does not seem possible to employ it in order to conclude on the asymptotic behaviour of $\mathbb{E}L_n$ without restrictive additional assumptions.

Using a different approach we arrived at

Theorem 7.1 *The expectation $\mathbb{E}L_n$ exhibits the following asymptotic behaviour:*

(i) *If $\mu = \infty$ and $\nu = \infty$ then*

$$\liminf_{n\to\infty}\frac{\mathbb{E}W^n}{\mathbb{E}(1-W)^n} \le \liminf_{n\to\infty}\mathbb{E}L_n \le \limsup_{n\to\infty}\mathbb{E}L_n \le \limsup_{n\to\infty}\frac{\mathbb{E}W^n}{\mathbb{E}(1-W)^n}.$$

In particular,

$$\lim_{n\to\infty}\frac{\mathbb{E}W^n}{\mathbb{E}(1-W)^n} = \gamma_0 \in [0,\infty]$$

implies $\lim_{n\to\infty}\mathbb{E}L_n = \gamma_0$.

(ii) *If $\nu < \infty$ and $\mu \le \infty$ then*

$$\lim_{n\to\infty}\mathbb{E}L_n = \nu/\mu.$$

(iii) *If $\mu < \infty$ and $\nu = \infty$ then, as $n \to \infty$,*

$$\mathbb{E}L_n \sim \frac{1}{\mu}\int_1^n \frac{\mathbb{E}e^{-y(1-W)}}{y}\mathrm{d}y.$$

Proof: Part (i). Set $s_m = \frac{\mathbb{E}W^m}{\mathbb{E}(1-W)^m}$. We will use the representation

$$\mathbb{E}L_n = \mathbb{E}s_{Z_n}, \tag{10}$$

which follows from (8). The array $c_{n,m} := \mathbb{P}\{Z_n = m\}$ verifies the conditions of Lemma 8.1 in Appendix, in particular by the assumption $\mu = \infty$. Hence the lemma can be applied to $t_n = \mathbb{E}L_n$, whence the assertion. When γ_0 is well defined the proof is simpler, as in this case the statement follows from (10), divergence of Z_n, and by using dominated convergence in the case $\gamma_0 < \infty$, respectively using Fatou's lemma in the case $\gamma_0 = \infty$.

See [16] and [17] for (ii).

For part (iii) we use the poissonised version of the Bernoulli sieve, in which balls are thrown one-by-one at the epochs of a unit Poisson process $(\Pi_t)_{t\ge 0}$, independent of W_k's. One can check that

$$\mathbb{E}(L_{\Pi_t}|(W_k)_{k\in\mathbb{N}}) = \sum_{k=1}^{\infty}\left(e^{-tW_1\cdot\ldots\cdot W_{k-1}(1-W_k)} - e^{-tW_1\cdot\ldots\cdot W_{k-1}}\right).$$

Recalling definitions (5),(6) and setting $\varphi(t) := \mathbb{E}e^{-t(1-W)}$, $U(x) := \mathbb{E}N_x = \sum_{k=1}^{\infty} \mathbb{P}\{S_{k-1} \le x\}$, we have

$$\begin{aligned}
\mathbb{E}L_{\Pi_t} &= \mathbb{E}\sum_{k=1}^{\infty}\Big(\varphi(te^{-S_{k-1}}) - \exp(-te^{-S_{k-1}})\Big) \\
&= \int_0^{\infty}\Big(\varphi(te^{-x}) - \exp(-te^{-x})\Big)U(\mathrm{d}x) \qquad (11)\\
&= \sum_{k=1}^{\infty}(-1)^{k+1}\frac{t^k}{k!}\frac{(1-\mathbb{E}(1-W)^k)}{(1-\mathbb{E}W^k)}, \qquad (12)
\end{aligned}$$

where the familiar formula for Laplace transform of the potential measure,

$$\int_0^{\infty} e^{-sx}U(\mathrm{d}x) = \frac{1}{1-\mathbb{E}W^s}, \quad s > 0,$$

has been utilised. Note that (12) is an obvious counterpart of (9).

Set $K(t) = \varphi(e^t) - \exp(-e^t)$, $t \in \mathbb{R}$. Since $\nu = \infty$ and

$$\int_0^{\infty}\frac{e^{-z(1-W)} - e^{-z}}{z}\mathrm{d}z = |\log(1-W)|,$$

we conclude that

$$\lim_{t\to\infty}\int_{-\infty}^{t} K(z)\mathrm{d}z = \infty. \qquad (13)$$

Applying a minor extension of [29, Theorem 5] to the equality

$$\mathbb{E}L_{\Pi_{e^t}} = \int_0^{\infty} K(t-x)U(\mathrm{d}x), \qquad (14)$$

which is equivalent to (11), yields

$$\mathbb{E}L_{\Pi_{e^t}} \sim \frac{1}{\mu}\int_0^t \varphi(e^x)\mathrm{d}x \sim \frac{1}{\mu}\int_1^{e^t}\frac{\varphi(x)}{x}\mathrm{d}x.$$

The asymptotics of $\mathbb{E}L_n$ is now obtained by the depoissonisation Lemma 8.2 in Appendix. The lemma is applicable because $\mathbb{E}L_{\Pi(t)}$ is slowly varying. Indeed, slow variation of $\int_1^t \varphi(u)\mathrm{d}u/u$ is checked straightforwardly from $\varphi(t) \downarrow 0$ and the divergence of the integral for $t = \infty$. $\square$

Similarly to the above, the proof of the next theorem is based on the poissonisation technique.

Theorem 7.2 [16] *If $\mu < \infty$ and $\nu < \infty$ then $L_n \to_d L_\infty$ as $n \to \infty$ for some random variable L_∞ whose distribution satisfies*

$$\mathbb{P}\{L_\infty \ge i\} = \frac{1}{\mu}\sum_{j=1}^{\infty}\frac{\mathbb{E}W^j}{j}\mathbb{P}\{L_j = i\}, \;\; i \in \mathbb{N}.$$

Moreover, the convergence of all moments holds, i.e. $\mathbb{E}L_n^k \to \mathbb{E}L_\infty^k < \infty$ for $k \in \mathbb{N}$.

It is also known that if $\mu < \infty$ and $\nu = \infty$ then $L_n \to_d \infty$ (see [17]), and that $L_n \to_P 0$ if $\nu < \infty$ and $\mu = \infty$. In the cases not covered by these results the question about the weak convergence of L_n is open.

Note that Theorem 7.2 only gives implicit specification of the limit law through distributions of L_n's, which are not easy to determine, with one remarkable exception. Obviously from the recursive construction of the Bernoulli sieve, the distribution of L_1 is geometric with parameter $\mathbb{E}W$. Curiously, the same is true for all n provided the law of W is symmetric about the midpoint $1/2$.

Proposition 7.1 *If $W =_d 1 - W$ then L_n is geometrically distributed with parameter $1/2$ for all $n \in \mathbb{N}$.*

Proof: The argument is based on the recursion (3) for marginal distributions of the L_n's. The symmetry $W =_d 1 - W$ yields $\mathbb{E}W^k = \mathbb{E}(1 - W)^k$ for all $k \in \mathbb{N}$ and

$$\mathbb{P}\{Q_n^*(1) = n\} = \mathbb{P}\{Q_n^*(1) = 0\} \tag{15}$$

for all $n \in \mathbb{N}$. We will show by induction on n that $\mathbb{P}\{L_n = k\} = 2^{-k-1}$ for all $k \in \mathbb{N}_0$. Using (3) and (15) we obtain

$$\begin{aligned}
\mathbb{P}\{L_n = 0\} &= \mathbb{P}\{Q_n^*(1) = 0\} + \sum_{k=1}^{n-1} \mathbb{P}\{L_k = 0\}\mathbb{P}\{Q_n^*(1) = k\} \\
&= \mathbb{P}\{Q_n^*(1) = 0\} + \frac{1}{2}\Big(1 - 2\mathbb{P}\{Q_n^*(1) = 0\}\Big) = \frac{1}{2},
\end{aligned}$$

by the induction hypothesis. Assuming now that $\mathbb{P}\{L_n = i\} = 2^{-i-1}$ for all $i < k$ we have

$$\begin{aligned}
\mathbb{P}\{L_n = k\} &= \sum_{j=1}^{n-1} \mathbb{P}\{Q_n^*(1) = j\}\mathbb{P}\{L_j = k\} + \mathbb{P}\{Q_n^*(1) = n\}\mathbb{P}\{L_n = k - 1\} \\
&= 2^{-k-1}\Big(1 - 2\mathbb{P}\{Q_n^*(1) = 0\}\Big) + \mathbb{P}\{Q_n^*(1) = 0\}2^{-k} = 2^{-k-1},
\end{aligned}$$

and the proof is complete. □

Alternatively, one can use a representation of L_n through the sojourns of the Markov chain Q_n^* in positive states. Indeed, recall that L_1 has geometric distribution with parameter $\mathbb{E}W$. Then using (15) and induction it can be checked that the distribution of L_n does not depend on $n \geq 1$.

8 Appendix.

For ease of reference we include a result due to Toeplitz and Schur (see [20], Theorem 2 on p. 43 and Theorem 9 on p. 52). We rewrite it in a form suitable for our purposes.

Lemma 8.1 *Let $\{s_n, n \in \mathbb{N}\}$ be any sequence of real numbers and let $\{c_{nm},\ n, m \in \mathbb{N}\}$ be a nonnegative array. Define another sequence $\{t_n, n \in \mathbb{N}\}$ by $t_n = \sum_{m=1}^{n} c_{nm} s_m$. If*

(i) $\lim_{n\to\infty} c_{nm} = 0$ *for all m,*

(ii) $\lim_{n\to\infty} \sum_{m=1}^{n} c_{nm} = 1$, *then*

$$\liminf_{n\to\infty} s_n \le \liminf_{n\to\infty} t_n \le \limsup_{n\to\infty} t_n \le \limsup_{n\to\infty} s_n \le +\infty.$$

Now we address the issue of depoissonisation. Recall that the function $\mathbb{E}L_{\Pi(t)}$ is slowly varying.

Lemma 8.2 *If* $\lim_{t\to\infty}\mathbb{E}L_{\Pi_t} = +\infty$ *then* $\mathbb{E}L_n \sim \mathbb{E}L_{\Pi_n}$, as $n\to\infty$.

Proof: For any fixed $\varepsilon\in(0,1)$,

$$\mathbb{E}L_{\Pi_t} \quad=\quad \mathbb{E}L_{\Pi_t}1_{\{|\Pi_t - t|>\epsilon t\}} + \mathbb{E}L_{\Pi_t}1_{\{|\Pi_t-t|\le\epsilon t\}} =: A(t)+B(t).$$

Sublinearity of $\mathbb{E}L_{\Pi_t}$ and the elementary large deviation bound for the Poisson distribution [2],

$$\mathbb{P}\{|\Pi_t - t|>\epsilon t\} < c_1 e^{-c_2 t},\ t>0 \tag{16}$$

with some $c_1, c_2>0$, yield $A(t)\to 0$.

It remains to evaluate $B(t)$. Since both M_n and K_n are non-decreasing, we have

$$B(t) = \mathbb{E}(M_{\Pi_t} - K_{\Pi_t})1_{\{|\Pi_t-t|\le\varepsilon t\}} \le \mathbb{E}L_{[(1-\varepsilon)t]} + \mathbb{E}(M_{[(1+\varepsilon)t]} - M_{[(1-\varepsilon)t]}).$$

Similarly, $B(t)\ge \mathbb{E}L_{[(1+\varepsilon)t]} - \mathbb{E}(M_{[(1+\varepsilon)t]} - M_{[(1-\varepsilon)t]})$. First step is to prove that

$$\lim_{\varepsilon\downarrow 0}\lim_{n\to\infty}\mathbb{E}(M_{[(1+\varepsilon)n]} - M_n) = 0. \tag{17}$$

Recalling the notation $T_n = \max(E_1,\dots,E_n)$ and using the subadditivity of $U(\cdot)$ we obtain

$$\begin{aligned} D(n) \quad &:= \quad \mathbb{E}\Big(M_{[(1+\varepsilon)n]} - M_n\Big) = \mathbb{E}\Big(U(T_{[(1+\varepsilon)n]}) - U(T_n)\Big) \\ &\le \quad \mathbb{E}U(T_{[(1+\varepsilon)n]} - T_n)1_{\{T_{[(1+\varepsilon)n]}>T_n\}}. \end{aligned}$$

An appeal to estimate $U(x)<ax+b$ (with some $a,b>0$) allows us to conclude that

$$D(n)\le \mathbb{E}\Big(a(T_{[(1+\varepsilon)n]} - T_n)+b\Big)1_{\{T_{[(1+\varepsilon)n]}>T_n\}} \le a(H_{[(1+\varepsilon)n]} - H_n) + b\mathbb{P}\{T_{[(1+\varepsilon)n]}>T_n\},$$

where the equality $\mathbb{E}T_n = H_n := \sum_{k=1}^n \frac{1}{k}$ has been utilised. Since

$$\lim_{\varepsilon\downarrow 0}\lim_{n\to\infty}(H_{[(1+\varepsilon)n]} - H_n) = 0$$

and by exchangeability

$$\mathbb{P}\{T_{[(1+\varepsilon)n]}>T_n\} = 1 - \frac{n}{[(1+\varepsilon)n]} \to \varepsilon,$$

as $n\to\infty$, we arrive at (17).

We are ready to finish the proof. Divide the inequality

$$\mathbb{E}L_{\Pi_{n/(1-\varepsilon)}} \le A(n/(1-\varepsilon)) + \mathbb{E}L_n + \mathbb{E}(M_{[\frac{1+\varepsilon}{1-\varepsilon}n]} - M_n), \tag{18}$$

by $\mathbb{E}L_{\Pi_n}$. Letting $n \to \infty$ then $\varepsilon \to 0$ and using the slow variation of $\mathbb{E}L_{\Pi_n}$ we obtain

$$\lim\inf_{n\to\infty} \frac{\mathbb{E}L_n}{\mathbb{E}L_{\Pi_n}} \geq 1.$$

The upper bound follows in the same way from the inequality

$$\mathbb{E}L_{\Pi_{n/(1+\varepsilon)}} \geq \mathbb{E}L_n - \mathbb{E}(M_n - M_{[\frac{1-\varepsilon}{1+\varepsilon}n]}). \tag{19}$$

□

References

[1] R. Arratia, A. Barbour, and S. Tavaré. *Logarithmic combinatorial structures*. European Mathematical Society, 2003.

[2] R. R. Bahadur. Some limit theorems in statistics. *CBMS Regional conference series in applied mathematics. Philadelphia: SIAM*, 4, 1971.

[3] Y. Baryshnikov, B. Eisenberg, and G. Stengle. A necessary and sufficient condition for the existence of the limiting probability of a tie for first place. *Statistics and Probability Letters*, 23(3):203–209, 1995.

[4] J. Brands, F. Steutel, and R. Wilms. On the number of maxima in a discrete sample. *Statistics and Probability Letters*, 20:209–217, 1994.

[5] F. T. Bruss and R. Grübel. On the multiplicity of the maximum in a discrete random sample. *Ann. Appl. Prob.*, 13(4):1252–1263, 2003.

[6] F. T. Bruss and C. A. O'Cinneide. On the maximum and its uniqueness for geometric random samples. *J. Appl. Probab.*, 27:598–610, 1990.

[7] C. F. Dunkl. The absorption distribution and the q-binomial theorem. *Communications in Statistics - Theory and Methods*, 10(19):1915–1920, 1981.

[8] E. B. Dynkin. Some limit theorems for sums of independent random variables with infinite mathematical expectations. *Selected Transl. in Math. Statist. and Probability*, 1:171–189, 1961.

[9] B. Eisenberg, G. Stengle, and G. Strang. The asymptotic probability of a tie for first place. *Ann. Appl. Prob.*, 3:731–745, 1993.

[10] K. B. Erickson. Strong renewal theorems with infinite mean. *Trans. Amer. Math. Soc.*, 151:263–291, 1970.

[11] J. Fill, H. Mahmoud, and W. Szpankowski. On the distribution for the duration of a randomized leader election algorithm. *Ann. Appl. Prob.*, 6:1260–1283, 1996.

[12] A. Gnedin. The Bernoulli sieve. *Bernoulli*, 10:79–96, 2004.

[13] A. Gnedin. Regeneration in random combinatorial structures. *Probability Surveys*, 7:105–156, 2010.

[14] A. Gnedin, A. Hansen, and J. Pitman. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability Surveys*, 4:146–171, 2007.

[15] A. Gnedin, A. Iksanov, and A. Marynych. Limit theorems fot the number of occupied boxes in the Bernoulli sieve. *Submitted to Theory of Stochastic Processes*, 2010.

[16] A. Gnedin, A. Iksanov, P. Negadajlov, and U. Rösler. The Bernoulli sieve revisited. *Ann. Appl. Prob.*, 19:1634–1655, 2009.

[17] A. Gnedin, A. Iksanov, and U. Rösler. Small parts in the Bernoulli sieve. *Discrete Mathematics and Theoretical Computer Science*, Proceedings Series Volume AI:239–246, 2008.

[18] A. Gnedin and J. Pitman. Regenerative composition structures. *Ann. Probab.*, 33:445–479, 2005.

[19] W. M. Y. Goh and P. Hitczenko. Gaps in samples of geometric random variables. *Discrete Mathematics*, 22:2871–2890, 2007.

[20] G. H. Hardy. *Divergent Series*. AMS Bookstore, 2000.

[21] P. Hitczenko and A. Knopfmacher. Gap-free compositions and gap-free samples of geometric random variables. *Discrete Mathematics*, 294(3):225–239, 2005.

[22] S. Janson and W. Szpankowski. Analysis of an asymmetric leader election algorithm. *Electron. J. Combin.*, 4(1), Art. #R17, 1997.

[23] S. Karlin. Central limit theorems for certain infinite urn schemes. *J. Math. Mech.*, 17:373–401, 1967.

[24] A. W. Kemp. Absorption sampling and the absorption distribution. *J. Appl. Prob.*, 35:489–494, 1998.

[25] P. Kirschenhofer and H. Prodinger. The number of winners in a discrete geometrically distributed sample. *Ann. Appl. Probab.*, 6:687–694, 1996.

[26] G. Louchard and H. Prodinger. On gaps and unoccupied urns in sequence of geometrically distributed random variables. *Discrete Mathematics*, 308(9):1538–1562, 2008.

[27] G. Louchard and H. Prodinger. The asymmetric leader election algorithm: another approach. *Annals of Combinatorics*, 12(4):449–478, 2009.

[28] H. Prodinger. How to select a loser. *Discrete Math.*, 120:149–159, 1993.

[29] M. S. Sgibnev. Renewal theorem in the case of an infinite variance. *Siberian Math. J.*, 22:787–796, 1981.