Small parts in the Bernoulli sieve

Alexander Gnedin[1], Alex Iksanov[2] and Uwe Roesler[3†]

[1]*Department of Mathematics, Utrecht University, Postbus 80010, 3508 TA Utrecht, The Netherlands*
[2]*Faculty of Cybernetics, National T. Shevchenko University of Kiev, Kiev-01033, Ukraine*
[3] *Mathematisches Seminar, Christian-Albrechts-Universität zu Kiel, Ludewig-Meyn-Str.4, D-24098 Kiel, Germany*

Sampling from a random discrete distribution induced by a 'stick-breaking' process is considered. Under a moment condition, it is shown that the asymptotics of the sequence of occupancy numbers, and of the small-parts counts (singletons, doubletons, etc) can be read off from a limiting model involving a unit Poisson point process and a self-similar renewal process on the half-line.

Keywords: Poisson process; multiplicative renewal process; random occupancy scheme.

1   Introduction

A multiplicative renewal process (also known as residual allocation model or stick-breaking) is a random sequence $B = (P_j : j \in \mathbb{N}_0)$ of the form

$$P_j = \prod_{i=1}^{j} W_i, \tag{1}$$

(so $P_0 = 1$) where $(W_i : i \in \mathbb{N})$ are independent copies of a random variable W taking values in $]0, 1[$. We shall assume that the support of the distribution of W is not a geometric sequence or, equivalently, that the distribution of the variable $|\log W|$ is non-lattice, and also assume that

$$\mu := \mathbb{E}[-\log W] < \infty. \tag{2}$$

The 'stick-breaking' set B will be viewed as a simple point process, with 0 being the only accumulation point. The complement $B^c = [0, 1] \setminus B$ is an open set comprised of the component intervals $]P_{j+1}, P_j[$ for $j \in \mathbb{N}_0$.

Let $U_1, U_2, \ldots$ be independent uniform [0,1] random points, also independent of B, and for each n let $U_{n,1} < \ldots < U_{n,n}$ be the order statistics of $U_1, \ldots, U_n$. These data define a random occupancy

---

scheme, called the Bernoulli sieve, in which n 'balls' $1, \ldots, n$ are dropped into infinitely many 'boxes' $j = 1, 2, \ldots$ according to the rule: ball i falls in box j if the event $U_i \in]P_j, P_{j-1}[$ occurs. The allocation of balls can be constructed sequentially by first sampling a value W_1 and dropping every ball into box 1 independently with probability $1 - W_1$, then sampling W_2 and dropping each of the remaining balls independently into box 2 with probability $1 - W_2$, and so on until all n balls are distributed into boxes. In the most studied and analytically best tractable case the law of W is beta$(\theta, 1)$, $\theta > 0$, when the allocation of balls-in-boxes belongs to the circle of questions around the Ewens sampling formula [1, 3].

In [4, 5] the renewal theory was applied to explore the spectrum of possible limit laws for the number of occupied boxes K_n, including normal, stable and Mittag-Leffler distributions. In the present note we retain the condition (2) and focus on the variables $K_{n,r}$, which count the number of intervals of B^c occupied by some r out of n uniform points, $r \in \mathbb{N}$. We also define $K_{n,0}$ to be the number of unoccupied interval components of $B^c \cap [U_{n,1}, 1]$, so that $K_n = I_n - K_{n,0}$, with $I_n := \min\{i : P_i < U_{n,1}\}$ being the index of the last occupied interval. Note that the intervals are labelled right-to-left. We approach the $K_{n,r}$'s via the *occupancy counts*

$$Z_n^{(i)} := \#\{1 \leq j \leq n : U_j \in \ ]P_{I_n - i + 1}, P_{I_n - i}[\}, \ \ i \in \mathbb{N},$$

labelled in the left-to-right order of the intervals, where we set $Z_n^{(i)} = 0$ for $i > I_n$. Extending a result from [5] about $Z_n^{(1)}$, we will show that the $Z_n^{(i)}$'s jointly converge to the sequence of occupancy numbers in a limiting model that involves a Poisson process and another self-similar point process on the half-line.

From a viewpoint, B is the range of $(\exp(-S_t) : t \geq 0)$ for some compound Poisson process $(S_t : t \geq 0)$. Asymptotics of $K_n, K_{n,r}$'s have been studied in a similar situation with $(S_t : t \geq 0)$ being an increasing Lévy process whose Lévy measure is infinite [2, 7, 8]. In the infinite measure case neither the counts $Z_n^{(i)}$ nor I_n can be defined, because B has then the topology of a Cantor set, thus the interval components of B^c cannot be simply enumerated by integers consistently with their natural order.

2   Occupancy counts

For $0 \leq m \leq n$ the probability that the interval $]P_1, P_0[$ contains m out of n uniform points is

$$p(n : m) = \binom{n}{m} \mathbb{E}[W^{n-m}(1 - W)^m].$$

The sequence of cluster sizes of points within intervals can be represented as a *weak composition* of n, meaning that $n_1 > 0, n_2 \geq 0, \ldots, n_\ell \geq 0$ and $n_1 + \ldots + n_\ell = n$. The structure (1) and elementary properties of the uniform distribution imply the product formula

$$\binom{n}{n_1, \ldots, n_\ell} p(n_1 + \ldots + n_\ell : n_\ell) p(n_1 + \ldots + n_{\ell-1} : n_{\ell-1}) \ldots p(n_1 : n_1) \tag{3}$$

(where the multinomial coefficient can be factored as $\prod_{j=1}^{\ell} \binom{n_1 + \ldots + n_j}{n_j}$) for the probability that the intervals $]P_j, P_{j-1}[$ contain n_j uniform points, $j = 1, \ldots, \ell$. While this formula specifies the joint distribution of the occupancy counts read right-to-left, there is no simple formula for the joint distribution of the counts read left-to-right. We shall see, nevertheless, that in the $n \to \infty$ limit there is a considerable simplification, as in [6].

Observe that $Z_n := (Z_n^{(i)} : i \in \mathbb{N})$ can be defined in the same 'balls-in-boxes' fashion in terms of the inflated sets nB and $\mathcal{U}_n := \{nU_{n,j} : 1 \leq j \leq n\}$. From the extreme-value theory we know that, as $n \to \infty$, the point process $\mathcal{U}_n$ converges vaguely to a unit Poisson process $\mathcal{U}$ on $\mathbb{R}_+$. Here and henceforth, the vague convergence means weak convergence on every finite interval bounded away from 0. On the other hand, nB also converges vaguely to some point process $\mathcal{B}$ on $\mathbb{R}_+$ which is self-similar, i.e. $c\mathcal{B} =_d \mathcal{B}$ for every $c > 0$. The intensity measure of the process $\mathcal{B}$ is $(\mu x)^{-1} \mathrm{d}x$. The convergence of nB is a consequence of the classical renewal theorem applied to the finite-mean random walk $(-\log P_j : j \in \mathbb{N}_0)$. The self-similarity in this context is analogous to the stationarity in the (additive) renewal theory.

The set $\mathbb{R}_+ \setminus \mathcal{B}$ is itself a collection of open intervals ('boxes') which accumulate in some way the points of $\mathcal{U}$ ('balls'), hence we can define a nonnegative sequence of counts of 'balls-in-boxes' $Z := (Z^{(i)} : i \in \mathbb{N})$ which starts with some positive number $Z^{(1)}$ of Poisson points falling in the leftmost nonempty interval. In view of the convergence of the point processes, one can expect that the convergence of the counting sequences also holds. This is stated in the following theorem ($\to_d$ denotes the convergence in distribution).

Theorem 2.1 *As $n \to \infty$,*

$$(Z_n^{(1)}, Z_n^{(2)}, \ldots) \to_d (Z^{(1)}, Z^{(2)}, \ldots). \tag{4}$$

The distribution of the limit sequence is given by the formula

$$\mathbb{P}(Z^{(1)} = n_1, \ldots, Z^{(\ell)} = n_\ell) =$$

$$\frac{1}{\mu(n_1 + \ldots + n_\ell)} \binom{n_1 + \ldots + n_\ell}{n_1, \ldots, n_\ell} p(n_1 + \ldots + n_\ell : n_\ell) p(n_1 + \ldots + n_{\ell-1} : n_{\ell-1}) \ldots p(n_1 : n_1) \tag{5}$$

for any $\ell > 0$ and $n_1 > 0, n_2 \geq 0, \ldots, n_\ell \geq 0$.

Proof: Fix $\epsilon > 0$ and restrict all point processes to $[\epsilon, \epsilon^{-1}]$. By Skorohod's theorem we can select probability space in such a way that the convergence of $(nB, \mathcal{U}_n)$ to $(\mathcal{B}, \mathcal{U})$ holds almost surely, then for the continuity reasons the occupancy numbers of the intervals within $[\epsilon, \epsilon^{-1}]$ converge. The weak convergence (4) follows by sending $\epsilon \to 0$ and noting that the probability that any m leftmost points of $\mathcal{U}$ fit in $[\epsilon, \epsilon^{-1}]$ goes to one.

Let $n = n_1 + \ldots + n_\ell$ and denote by X the $(n + 1)$st leftmost point of $\mathcal{U}$. The generic sequence of occupancy numbers which gives rise to the event in (5) is of the form $(n_1, \ldots, n_\ell, 0, \ldots, 0, m)$ where m is some positive number and the number of 0's is arbitrary. Let $G = \max(\mathcal{B} \cap [0, X])$ be the largest point of $\mathcal{B}$ smaller than X; from selfsimilarity and [6] we know that the distribution of G/X has density $(\mu x)^{-1} \mathbb{P}(W < x)$ on $[0, 1]$, and from the order statistics property of the Poisson process we know that given X the first n points of $\mathcal{U}$ are distributed as a uniform sample from $[0, X]$. The pattern $(n_1, \ldots, n_\ell, 0, \ldots, 0, m)$ occurs when the uniform n-sample does not hit $[G, X]$ (event E_1) and within $[0, G]$ the occupancy numbers are $(n_1, \ldots, n_\ell, 0, \ldots, 0)$ (event E_2). Integrating by parts, the probability of E_1 is

$$\int_0^1 \frac{x^n \mathbb{P}(W < x)}{\mu x} \, \mathrm{d}x = \frac{1}{\mu n} \left(1 - \mathbb{E}[W^n]\right).$$

For $i \in \mathbb{N}_0$ let $E_{2,i}$ be the event that the pattern $(n_1, \ldots, n_l, 0, \ldots, 0)$ with exactly i zeros occurs. In view of the equality $\dfrac{\mathcal{B} \bigcap [0, G]}{G} =_d B \setminus \{1\}$, the conditional probability $\mathbb{P}(E_{2,i} | X = x, E_1)$ equals the

probability (3) with $k = l + i$ and $n_{l+1} = \ldots = n_k = 0$. Since $E_2 = \bigcup_{i=0}^{\infty} E_{2,i}$ then summing the last probability over i, we have

$$\mathbb{P}(E_2|X = x, E_1) =$$

$$\frac{1}{1 - \mathbb{E}[W^n]} \binom{n}{n_1, \ldots, n_k} p(n_1 + \ldots + n_k : n_k) p(n_1 + \ldots + n_{k-1} : n_{k-1}) \ldots p(n_1 : n_1) =$$

$$\mathbb{P}(E_2|E_1).$$

Since the probability in (5) equals $\mathbb{P}(E_1 \bigcap E_2)$, the proof is complete. $\qquad\qquad \square$

3   Small parts

We wish to connect the asymptotics of r-counts to Theorem 2.1. Let Y be the leftmost atom of $\mathcal{U}$. For $r \geq 0$ let K_r^* be the number of intervals of $]Y, \infty[\setminus \mathcal{B}$ that contain exactly r points of $\mathcal{U}$. For $r > 0$ we can take $]0, \infty[$ instead of $]Y, \infty[$ in this definition.

Lemma 3.1 *Let A, B be two simple (i.e. without multiple points) point processes defined and a.s. finite in some interval $[s, t]$, and such that $A \cap B = \varnothing$ a.s. Suppose we have weak convergence $(A_n, B_n) \to_d (A, B)$ for a sequence of bivariate point processes. Define a gap to be a subinterval of $[s, t]$ whose endpoints are consecutive atoms of B. Let L_k be the number of gaps in B that contain exactly k points of A (with the convention that L_0 counts the gaps to the right of the leftmost A-point in $[s, t]$), and let $L_{n,k}$ be defined similarly in terms of (A_n, B_n). Then $(L_{n,0}, L_{n,1}, \ldots) \to_d (L_0, L_1, \ldots)$ as $m \to \infty$.*

Proof: By Skorohod's theorem a version of the processes can be defined on some probability space in such a way that with probability one the convergence is pointwise. That is to say, for large enough n, $\#B_n$ and $\#B$ are equal and the points of B_n (labelled, e.g. in the increasing order) are ϵ-close to the points of B. Same for A_n, A. Thus for large n, there is a bijection between the gaps in B and in B_n and between the points of A and A_n that fall in each particular gap. $\qquad \square$

A variation of the lemma allows accumulation of atoms of the gaps-generating process at the left endpoint of the underlying interval. In our situation both $\mathcal{B}$ and $\mathcal{U}$ live on the half-line and accumulate at infinity, hence to pass from the occupancy counts to $K_{n,r}$'s we need to take further care by showing that the contribution of the counts within $[s, \infty]$ is negligible for large s. To this end, it is enough to work with expected values. Now, the mean contribution of $[0, s]$ to $\mathbb{E}[K_r^*]$ can be estimated by the expected number of points in $\mathcal{B} \cap [\min(Y, s), s]$,

$$\mathbb{E}\left[\int_{\min(Y,s)}^{s} \frac{\mathrm{d}x}{\mu x}\right] = \int_0^{\infty} e^{-z} \mathrm{d}z \int_{\min(z,s)}^{s} \frac{\mathrm{d}x}{\mu x} = \int_0^s e^{-z} \mathrm{d}z \int_z^s \frac{\mathrm{d}x}{\mu x} < \infty. \qquad (6)$$

Lemma 3.2 *For $r > 0$ we have $\mathbb{E}[K_r^*] = (\mu r)^{-1}$. Furthermore, $\mathbb{E}[K_0^*] = \nu/\mu$, where*

$$\nu := \mathbb{E}[-\log(1 - W)]$$

which may be finite or infinite.

Proof: Understanding a $\mathcal{B}$-atom in $\mathrm{d}x$ as the *right* endpoint of a gap we obtain for $r > 0$

$$\mathbb{E}[K_r^*] = \mathbb{E}\left[\int_0^\infty e^{-x(1-W)} \frac{x^r(1-W)^r}{r!} \frac{\mathrm{d}x}{\mu x}\right] = \frac{1}{\mu\,r!}\,\mathbb{E}\left[\int_0^\infty e^{-y}y^{r-1}\mathrm{d}y\right] = \frac{1}{\mu r}\,.$$

For $r = 0$ we have

$$\mathbb{E}[K_0^*] = \mathbb{E}\left[\int_0^\infty e^{-(1-W)x}(1-e^{-Wx})\frac{\mathrm{d}x}{\mu x}\right] = \frac{\mathbb{E}[-\log(1-W)]}{\mu} = \frac{\nu}{\mu}\,,$$

where the second factor in the integrand stands for the event that X is smaller than the left endpoint of the gap. □

In the case $\nu = \infty$ the source of divergence of K_0^* is ∞ and not 0, as is seen from (6).

Theorem 3.3 *As $n \to \infty$ we have*

$$(K_{n,0}, K_{n,1}, \ldots) \to_d (K_0^*, K_1^*, \ldots),$$

along with

$$\mathbb{E}[K_{n,r}] \to \mathbb{E}[K_r^*].$$

The three conditions $\nu = \infty$, $\mathbb{E}K_0^* = \infty$ and $K_0^* = \infty$ a.s. are equivalent.

Proof: The limit set satisfies $\mathcal{B} \cap [0,1] =_d W_0 B$ where B and W_0 are independent, and W_0 has the density

$$(\mu x)^{-1}\mathbb{P}(W < x)\mathrm{d}x, \quad x \in [0,1]. \tag{7}$$

We shall speak of $[r]$-counts meaning the intervals within $[U_{n,1}, 1]$ (or $[Y, \infty]$, depending on the context) that contain at most r sampling points, and denote $K_{[r]}^* = \sum_{i=0}^r K_i^*$, $K_{n,[r]} = \sum_{i=0}^r K_{n,i}$. Replacing in the proof of Lemma 3.2 the lower limit of integration 0 by sW_0 we see that choosing s large enough we can achieve that the contribution to $\mathbb{E}[K_{[r]}^*]$ of the intervals with right endpoint in $[sW_0, \infty]$ is arbitrarily small. It remains to show that the contribution to $\mathbb{E}[K_{n,[r]}]$ of $[s/n, 1]$ is small for large enough s uniformly in n.

Observe that the number of components of $B^c \cap [\epsilon, 1]$ that contain no more than r uniform points is nonincreasing with n, because the number of 'balls' in a 'box' can only grow as more 'balls' are thrown. Furthermore, observe that, for the purpose of estimate, the fixed-n uniform sample can be replaced by the Poisson sample of rate n on $[0,1]$. Indeed, the probability that a gap of size x is hit by $r \geq 0$ uniform points is $\binom{n}{r}x^r(1-x)^{n-r}$, and in the poissonised model it is $e^{-nx}(nx)^r/r!$. Using the elementary inequality $1-x < e^{-x}$ and some algebra it is easy to estimate from the above the mean number of $[r]$-counts coming from the gaps of size $0 < x < 1/2$ within $[s/n, 1]$ for the fixed-n sample through the analogous quantity in the Poisson model. The intervals larger $1/2$ can be ignored, since the probability that they accomodate at most r sample points decays exponentially with n.

Arguing within the framework of Poisson sample $\mathcal{U} \cap [0,1]$, we compare occupancy of 'boxes' generated by B with that for $W_0 B$. The 'meander interval' $[W_0, 1]$ gives negligible contribution to $[r]$-counts hence will be ignored. Because $\mathcal{B} \cap [W_0 s/n, W_0]$ is a zoomed-in copy of $[s/n, 1]$, the sequence of occupancy counts for $\mathcal{B} \cap [W_0 s/n, W_0]$ has the same distribution as if we had $B \cap [s/n, 1]$ in the role of 'boxes' and a mixed Poisson sample with rate nW_0 in the role of 'balls'. By monotonicity and because $W_0 < 1$,

the number of $[r]$-counts derived from $\mathcal{B} \cap [W_0 s/n, W_0]$ is larger than the number of r-counts from $B \cap [s/n, W_0]$, which implies that the mean number of such counts can be kept small by the choice of s. This implies the desired estimate of the contribution of $[s/n, W_0]$ to $\mathbb{E}[K_{n,[r]}]$.

If $K_0^* = \infty$ a.s. then $\mathbb{E}[K_0^*] = \infty$ and by Lemma 3.2 we conclude that $\nu = \infty$. The equivalence of the conditions $\nu < \infty$ and $K_0^* < \infty$ a.s. will follow from the next lemma. $\qquad\square$

Lemma 3.4 *If $\nu = \infty$ then $K_0^* = \infty$ a.s.*

Proof: Let $(T_i : i \in \mathbb{N})$ be the points of $\mathcal{U}$ in increasing order. Consider the random points defined as follows:

$$\xi_1 := \frac{T_1}{W_0} \quad \text{and} \quad \xi_k := \frac{T_1}{W_0} \prod_{i=1}^{k-1} \frac{1}{W_i}, \quad i = 2, 3, \ldots,$$

where W_0 is independent of $(W_i : i \in \mathbb{N})$ and has distribution as in (7). For $i \in \mathbb{N}$ denote by E_i the event that the interval $[\xi_i, \xi_{i+1}]$ contains none of T_k for $k \geq 2$. Then $K_0^* = \sum_{i=1}^{\infty} 1_{E_i}$ and we wish to show that the series diverges with probability one.

The events $(E_i : i \in \mathbb{N})$ are adapted to the filtration $(\mathcal{F}_i : i \in \mathbb{N}_0)$, where $\mathcal{F}_i$ is the σ-field generated by $T_1, W_0, (W_k : k \in \mathbb{N})$ and $(E_k : k \leq i)$. By the conditional Borel-Cantelli lemma [9, Theorem 2.8.5], it is enough to show that

$$\sum_{i=1}^{\infty} \mathbb{P}(E_i | \mathcal{F}_{i-1}) = \infty \quad \text{a.s.} \tag{8}$$

In view of the conditional independence of the E_i's given T_1, W_0 and $(W_i : i \in \mathbb{N})$, we have $\mathbb{P}(E_i | \mathcal{F}_{i-1}) = \mathbb{P}(E_i | \mathcal{F}_0)$. Therefore,

$$\alpha_i := \mathbb{P}(E_i | \mathcal{F}_{i-1}) = \exp\left(-\frac{T_1}{W_0} \frac{1 - W_i}{W_1 \cdots W_i}\right).$$

By the strong law of large numbers, there exists $c > 1$ such that

$$\alpha_i \geq \exp\left(-\frac{T_1}{W_0} c^i (1 - W_i)\right) =: \beta_i$$

for large enough i. Conditioning on the value of T_1/W_0 we need to show divergence of the series $\sum_{i=1}^{\infty} \beta_i$. By the virtue of the ordinary Borel-Cantelli lemma, $\nu = \infty$ implies

$$\limsup_{i \to \infty} \frac{-\log(1 - W_i)}{i} = \infty \quad \text{a.s.}$$

Consequently, there exists an increasing sequence (i_k) such that $\beta_{i_k} \to 1$ as $k \to \infty$, which means that the series $\sum_{i=1}^{\infty} \beta_i$ cannot converge with probability one. Since, given $T_1/W_0, \beta_1, \beta_2, \ldots$ are independent and positive, the sum of the series is ∞ with probability one, whence (8). $\qquad\square$

References

[1] ARRATIA, R., BARBOUR, A.D. AND TAVARÉ, S. (2003) *Logarithmic combinatorial structures: a probabilistic approach*, European Math. Soc. Publ. House, Zürich.

[2] BARBOUR, A.D. AND GNEDIN, A.V. (2006). Regenerative compositions in the case of slow variation. *Stoch. Process. Appl.* **116,** 1012–1047.

[3] GNEDIN, A.V. (2004) Three Sampling Formulas, *Combinatorics, Probability and Computing* **13**, 185-193.

[4] GNEDIN, A. V. (2004). The Bernoulli sieve. *Bernoulli* **10,** 79–96.

[5] GNEDIN, A., IKSANOV, A., NEGADAJLOV, P. AND ROESLER, U. (2008). The Bernoulli sieve revisited, submitted.

[6] GNEDIN, A. AND PITMAN, J. (2005) Self-similar and Markov composition structures. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **326**, Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 13, 59–84, translation in *J. Math. Sci. (N. Y.)* **140** (2007) 376–390.

[7] GNEDIN, A., PITMAN, J., AND YOR, M. (2006). Asymptotic laws for regenerative compositions: gamma subordinators and the like. *Probab. Theory Relat. Fields* **135,** 576–602.

[8] GNEDIN, A., PITMAN, J., AND YOR, M. (2006). Asymptotic laws for compositions derived from transformed subordinators. *Ann. Probab.* **34,** 468–492.

[9] STOUT, W.F. (1974) *Almost sure convergence*, Academic Press, New York - San Francisco - London.