Bernoulli **10**(1), 2004, 79–96

The Bernoulli sieve

ALEXANDER V. GNEDIN

Department of Mathematics, Utrecht University, Postbus 80010, 3508 TA Utrecht, The Netherlands. E-mail: gnedin@math.uu.nl

The Bernoulli sieve is a recursive construction of a random composition (ordered partition) of an integer n. This composition can be induced by sampling from a random discrete distribution which has frequencies equal to the sizes of components of a stick-breaking interval partition of [0, 1]. We exploit the Markov property of the composition and its renewal representation to study the number of its parts. We derive asymptotics of the moments and prove a central limit theorem.

Keywords: composition; renewal; sampling; stick-breaking

1. Introduction

The Bernoulli sieve can be seen as a generalization of the 'game' found in Bruss and O'Cinneide (1990). The first round of the game starts with n players and amounts to tossing a coin with probability X_1 for tails. Each of the players tosses once and those getting tails must drop out. If all n get heads the trial is disregarded and must be repeated with all n players, as many times as necessary until some players are eliminated. If at least one player remains after the first round, the second round continues with the remaining players, who must toss another coin with probability X_2 for tails. The game continues with probabilities $X_3, X_4, \ldots$ for tails until all players have been sorted. It is assumed that the probabilities $X_1, X_2, \ldots$ are independent random variables with a given distribution ω on $]0, 1[$, and that, given X_j, the individual outcomes at round j are conditionally independent. It readily follows that, as far as only the number of players is concerned, the outcome of a round depends on the past solely through the number of players which proceed that far.

A random composition C_n of an integer n arises, in which part j is the number of players dropping out at round j. In this paper we shall focus on some of the properties of C_n; in particular, we are interested in the distribution of the number of parts of the composition, which may be thought of as the duration of the game.

There is a natural way to settle all C_ns on the same probability space in a consistent fashion. Consider a random interval partition of [0, 1] by points $1 - (1 - X_1)(1 - X_2) \ldots (1 - X_j)$, $j = 1, 2, \ldots,$ and assign each player a random uniform tag, independent of the X_js. The tags group within the intervals, and recording the cluster sizes from left to right yields a composition (intervals containing no tags are ignored). To establish equivalence with the coin-tossing construction we only need to note that the probability of a particular player remaining in the game for at least j rounds is precisely $(1 - X_1)(1 - X_2) \ldots (1 - X_j)$.

The game considered by Bruss and O'Cinneide (1990) corresponds to ω supported by a

1350–7265 © 2004 ISI/BS

single point, in which case the X_js are all equal. The composition is induced then by sampling from a geometric distribution and, of course, had already appeared many times in the literature in different guises. Karlin (1967) distinguished this case in the context of a general occupation problem with infinitely many boxes and derived distributions for the number of parts, number of singletons, doubletons, etc. The feature studied in Bruss and O'Cinneide (1990) and Kirschenhofer and Prodinger (1996) was the probability that there is exactly one winner – a player remaining in the last round (that is to say, the last part of the composition is 1).

When ω is a Beta$(1, \theta)$ distribution the law of C_n is known as the ordered Ewens sampling formula (ESF). This structure is well understood; see Arratia *et al.* (2002) for a recent account and Pitman (2002), Gnedin (2003) and Gnedin and Pitman (2003) for generalizations.

Our interest in the construction arose in connection with the regenerative composition structures introduced by Gnedin and Pitman (2003). Within this more general setting the Bernoulli sieve composition may be seen as a discretization of a subordinator with finite Lévy measure and zero drift. In what follows we shall treat the general probability measures, with the sole constraint that ω is not supported by a geometric sequence like $1 - x^j$ (in particular, sampling from the geometric distribution is ruled out) and such that ω does not settle too much mass near the end-points of $[0, 1]$. Our method relies on renewal theory and the analysis of 'divide-and-conquer' recurrences, the techniques intended to replace the independence-based tools available in the ESF case; see Arratia *et al.* (2002).

By virtue of the exchangeability among the players the compositions C_n are *sampling consistent* for different values of n. That is to say, if a part of C_n is selected at random, in a size-biased fashion, and decremented by one unit, then the resulting composition of $n - 1$ (possibly with fewer parts) has the same distribution as C_{n-1}. The sequence (C_n) forms a composition structure (see Gnedin 1997; 1998) and determines a random exchangeable composition of a countable set.

There are two further constructions of C_n featuring renewal and Markov properties. The *renewal representation* is obtained from the stick-breaking construction by applying the transformation $\phi(x) = -\log(1 - x)$ which maps $[0, 1]$ onto $[0, \infty]$. Consider the range $\mathcal{R}$ of a renewal process with initial state 0 and step distribution $\Omega = \omega_\phi$, and let $E_1, \ldots, E_n$ be increasing order statistics from the standard exponential distribution (which correspond to exponentially distributed tags). The points of $\mathcal{R}$ induce a partition of $[0, \infty]$ into intervals making up the complement $\mathcal{R}^c = [0, \infty]\backslash\mathcal{R}$, and the points E_j group within the intervals; in these terms the composition C_n becomes a record of all non-zero cluster sizes, from left to right.

The *Markov chain representation* of C_n stems from the following first-part deletion property of C_n. Given that the first part of C_n is m, the composition of $n - m$ obtained by removing this part has the same distribution as C_{n-m}. This property is obvious in the renewal context: it follows from the regenerative property of $\mathcal{R}$ (applied at the leftmost point of $\mathcal{R}$ to the right of E_1) taken together with the memorylessness property of the exponential distribution. The deletion property implies that the parts of C_n can be viewed as decrements of a decreasing Markov chain Q_n which has state space $\{0, 1, \ldots, n\}$, starts at state n and eventually gets absorbed at 0. The one-step transition probability from n to $n - m$ is

$$q(n, m) = \frac{w(n, m)}{1 - w(n, 0)}, \qquad m = 1, \ldots, n. \tag{1}$$

where

$$w(n, m) = \binom{n}{m} \int_0^1 x^m (1-x)^{n-m} \omega(\mathrm{d}x), \qquad m = 0, 1, \ldots, n,$$

are the binomial moments of ω. A similar expression can be given in terms of Ω, with $1 - \mathrm{e}^{-z}$ in place of x. The quantity $1 - w(n, 0) = w(n, 1) + \ldots + (n, n)$ will appear throughout as a normalizing factor, so we write $W(n) = 1 - w(n, 0)$. In other words,

$$W(n) = \int_0^\infty (1 - \mathrm{e}^{-nz}) \Omega(\mathrm{d}z)$$

is the characteristic exponent of the measure Ω thought of as a Lévy measure associated with $\mathcal{R}$.

For a given composition $(n_1, \ldots, n_k)$ of n, the probability that C_n assumes this value is of the product form

$$p(n_1, \ldots, n_k) = q(n_1 + \ldots + n_k, n_2 + \ldots + n_k) q(n_2 + \ldots + n_k, n_3 + \ldots + n_k) \cdots q(n_k, n_k), \tag{2}$$

because this is the probability that the chain Q_n has decrements $n_1, \ldots, n_k$ before absorption in 0.

2. Basic recursions

We will be interested in the first instance in the number of parts K_n of the composition C_n. It follows from $\omega\{1\} = 0$ that $q(n, m) > 0$ for all $n \geqslant m \geqslant 1$ and K_n goes to infinity with n.

Observe that the sizes of intervals comprising the partition of [0, 1] are $Y_j = (1 - X_1) \ldots (1 - X_{j-1}) X_j$. Rephrasing the stick-breaking interpretation, K_n is the number of boxes occupied by at least one of n balls, with probability Y_j of being in the jth box. Karlin (1967) is a basic reference on the model with infinitely many boxes and non-random frequencies, and some information on K_n can be extracted from Karlin's results by conditioning on (Y_j).

Consider two conditions on ω which limit concentration of mass near 1 and 0:

$$\mu := \int_0^1 |\log(1 - x)| \omega(\mathrm{d}x) < \infty, \tag{3}$$

$$\int_0^1 |\log x| \omega(\mathrm{d}x) < \infty. \tag{4}$$

Reformulated, condition (3) says that the first moment of Ω is finite:

$$\mu = \int_0^\infty z\Omega(\mathrm{d}z) < \infty. \tag{5}$$

The unconditional law of large numbers from Karlin (1967) implies the following proposition:

Proposition 1. *If conditions* (3) *and* (4) *are satisfied then, as* $n \to \infty$,

$$K_n \sim \frac{1}{\mu} \log n$$

with probability one.

Proof. By the strong law of large numbers we have, for $j \to \infty$,

$$-\frac{1}{j} \log Y_j = -\frac{1}{j} \sum_{i=1}^{j-1} \log(1 - X_j) \frac{1}{j} \log X_j \to \frac{1}{\mu}$$

(condition (4) is necessary and sufficient for the second term to be negligible). From this relation we have, for Karlin's (1967, p. 376) function α,

$$\#\left\{ j : Y_j > \frac{1}{x} \right\} \sim \frac{1}{\mu} \log x, \qquad \text{as } x \to \infty$$

almost surely. By Theorem 1$'$ in Karlin (1967),

$$\mathrm{E}(K_n | (Y_j)) \sim \frac{1}{\mu} \log n,$$

and by Theorem 8 in that paper the statement holds conditionally on (Y_j), hence also unconditionally. □

Note that 'deconditioning' itself does not allow one to draw any conclusions about the asymptotics for $\mathrm{E}K_n$ (see Proposition 3 below). The methods of Karlin (1967) could be used further to show that the conditional variance of K_n converges to $\mu^{-1} \log 2$ almost surely. We will not dwell on converting these results into their unconditional counterparts, rather we will take an approach based on the renewal features of our model.

Let F_n be the first part of C_n, with distribution $P(F_n = m) = q(n, m)$. The Markov property of the composition implies that K_n satisfies a distributional equation

$$K_n \overset{d}{=} 1 + K'_{n-F_\mathrm{n}}, \tag{6}$$

where $F_n, K'_1, K'_2, \ldots$ are independent and each K'_j has the same distribution as K_j. Averaging in (6), we see that $a_n = \mathrm{E}K_n$ satisfies a linear recursion

$$a_n = 1 + \sum_{m=1}^{n} q(n, m) a_{n-m} \tag{7}$$

with boundary value $a_0 = 0$.

Remark. Recursions akin to (7) are common in the average-case analysis of algorithms; see Rösler and Rüschendorf (2001). The dissertation by Bruhn (1996) is devoted solely to them. Some of Bruhn's results are reproduced in Rösler (2001) along with distributional analysis of equations more general than (6). The class of recursions treated in the work cited relates to the assumption that the weights $q(n, \cdot)$, considered as measures with support $\{1/n, 2/n, \ldots, n/n\}$, satisfy an equiboundedness condition and converge weakly to some measure on [0, 1].

In our case the convergence of $q(n, \cdot)$ to ω is clear from the convergence of moments (which amounts to Bernstein's trick used to prove the Weierstrass uniform approximation theorem). Beyond that, the Bruhn–Rösler conditions certainly hold when ω has a smooth density. However, we have been unable to check their (very technical) conditions for the general measure ω and will rely on the special structure (1). A specific feature of the class of recursions studied here is that we have a canonical renewal process as a part of the model, while Bruhn and Rösler needed to construct an auxiliary renewal process to 'mimic' the recursion.

We formulate the next fact as $L^1([0, 1], \omega)$-approximability of the logarithm by the (ordinary) Bernstein polynomials, but we will also make use of the formula (8). There is a variety of closely related results in the literature: the best known is the aforementioned argument due to Bernstein, then there are a number of L^1-results on generalized Bernstein polynomials (see Lorenz 1953), and pointwise asymptotic expansions are found in Flajolet (1999) and Jacquet and Szpankowski (1999). Still, the summation formulae (8) and (18) below seem to be new.

The Bernstein polynomial of degree n for $\log(1-x)$ is

$$B_n(x) = \sum_{m=1}^{n-1} \binom{n}{m} x^m (1-x)^{n-m} \log\left(1 - \frac{m}{n}\right).$$

Lemma 2. *If ω satisfies* (3)*, then*

$$\lim_{n\to\infty} \int_0^1 |B_n(x) - \log(1-x)|\omega(\mathrm{d}x) = 0.$$

Proof. There is no simple formula for the expectation of the logarithm of binomial random variable, but replacing the logarithms with the harmonic numbers and noting that $\log(1 - m/n) = h_{n-m} - h_n + o((n-m)^{-1})$, $h_n = 1 + n^{-1} + \ldots + 3^{-1} + 2^{-1}$ being the nth harmonic number, we have the explicit summation formula

$$\sum_{m=0}^{n-1} \binom{n}{m} x^m (1-x)^{n-m} (h_{n-m} - h_n) - x^n h_n = -\frac{x}{1} - \frac{x^2}{2} - \ldots - \frac{x^n}{n}. \tag{8}$$

By monotone convergence the series on the right-hand side approaches $\log(1-x)$ in the sense of $L^1(\omega, [0, 1])$ whatever ω. This easily yields the claim. □

Proposition 1 strongly suggests the logarithmic asymptotics for a_n. Our proof of this fact will rely on the following simple observation. Given $n_0 > 1$, suppose that (a_n) satisfies (7) for $n \geqslant n_0$; then $a_n + c$ also satisfies the recursion for $n \geqslant n_0$, whatever the constant c.

Proposition 3. *If ω satisfies* (3) *then any sequence (a_n) satisfying* (7) *for $n \geqslant n_0 \geqslant 1$ has asymptotics*

$$a_n \sim \frac{\log n}{\mu}.$$

In particular, this holds for the sequence $a_n = \mathrm{E}K_n$ which is the unique solution which satisfies (7) *for $n > 0$ and has the boundary value $a_0 = 0$.*

Proof. Assume that there exists $\epsilon > 0$ such that $a_n > (1+\epsilon)\mu^{-1}\log n$ for infinitely many values of n. Our proof is by contradiction. Selecting ϵ smaller, for any fixed c we can obtain the inequality $a_n > (1+\epsilon)\mu^{-1}\log n + c$ for infinitely many values of n. Let $n(c)$ be the minimum such n; then $n(c) \to \infty$ as $c \to \infty$. Thus, for $n < n(c)$ we have $a_n < (1+\epsilon)\mu^{-1}\log n + c$, which implies

$$1 + \sum_{m=1}^{n(c)} q(n(c), m) a_{n(c)-m} < 1 + c + \frac{(1+\epsilon)}{\mu} \sum_{m=1}^{n(c)} q(n(c), m) \log(n(c) - m).$$

Now from (7) and the definition of $n(c)$ we derive

$$(1+\epsilon)\frac{\log n(c)}{\mu} + c < 1 + c + \frac{1+\epsilon}{\mu} \sum_{m=1}^{n(c)} q(n(c), m) \log(n(c) - m), \tag{9}$$

where c itself cancels but $n(c)$ can be taken arbitrarily large by the choice of c.

From Lemma 2 we see that

$$\sum_{m=1}^{n-1} q(n, m) \log(n-m) = \log n - \mu + o(1),$$

and substituting this formula into (9) and letting $c \to \infty$ yields $0 < -\epsilon$, which is the contradiction promised. Thus the assumption was false, and because ϵ was arbitrary we have

$$\limsup \frac{a_n}{\mu^{-1}\log n} \leqslant 1.$$

A symmetric argument proves the analogous lower bound, and the claim follows. □

Turning to the variance of the number of parts $v_n = \operatorname{var} K_n$, we derive from (6) a recursion

$$v_n = \left(2a_n - 1 - a_n^2 + \sum_{m=1}^{n} q(n, m) a_{n-m}^2\right) + \sum_{m=1}^{n} q(n, m) v_{n-m}, \qquad v_0 = 0, \tag{10}$$

which involves $a_n = \mathrm{E}K_n$. Both (7) and (10) are instances of the general equation

$$b_n = r_n + \sum_{m=1}^{n} q(n, m) b_{n-m}, \qquad b_0 = 0, \tag{11}$$

where (b_n) are unknowns, and (r_n) is given. The proof of Proposition 3 is easily extended to obtain the following corollary:

Corollary 4. *Assume* (3). *For any* n_0 *and* $r \neq 0$, *if* (b_n) *satisfies* (11) *for* $n > n_0$ *and if* $r_n \to r$, *then* $b_n \sim r\mu^{-1} \log n$ *an* $n \to \infty$.

With a logarithmic asymptotics for v_n in mind, we aim to show the convergence of the bracketed inhomogeneous term in (10). It is easily seen that for this purpose we need more than just the principal-term asymptotics of the expectation, and this is exactly where the renewal theory provides indispensable tools.

3. Renewal approximation

It is well known that a renewal process starting at 0 admits a delayed version which has expected number of renewals within $[0, z]$ (the potential measure) growing linearly with z; see Feller (1971). It turns out that the stationary renewal process induces a 'stationary' version of the Markov chain Q_n, which can be used for the asymptotic analysis of (6).

Let $g(n, m)$ be the probability that Q_n ever visits state m (which means that at some round of the game there are exactly m players left). Since Q_n can visit each non-absorbing state at most once, $g(n, m)$ is also the potential function, that is, the expected number of visits to m. Interpreting r_m as a 'reward' collected on visiting state m, we can think of b_n satisfying (11) as the total expected reward of Q_n. The interpretation implies

$$b_n = \sum_{m=1}^{n-1} g(n, m) r_m \tag{12}$$

and reduces solving (7) to computation of the potential function. An explicit formula is complicated (see Gnedin and Pitman (2003)). Fortunately, there is a simple asymptotic formula.

Suppose Ω is not supported by a lattice and has finite first moment, so that (5) holds. For ω this means that (3) holds and that the support is not a geometric sequence like $1 - x^j$ (in particular, the case of geometric frequencies, when ω is supported by a single point, is excluded). Switching to the renewal representation, we introduce a probability distribution

$$\Omega_0[0, z] = \frac{1}{\mu} \int_0^z \Omega[\zeta, \infty] \mathrm{d}\zeta.$$

Let the *overshoot* $B(z)$ be the distance from z to the leftmost point of $\mathcal{R}$ to the right of z ($B(z)$ is sometimes called the forward process, or forward recurrence time, or residual lifetime, etc.). The renewal theory, as presented in Feller (1971), says that Ω_0 is the limiting distribution of the overshoot as $z \to \infty$. Observe that Q_n visits m when there is a point of $\mathcal{R}$

between E_{n-m-1} and E_{n-m} or, equivalently, when the overshoot at E_{n-m-1} does not exceed $E_{n-m} - E_{n-m-1}$. The spacing between the two order statistics is independent of E_{n-m-1} and its distribution is Exponential(m). By the renewal theorem the distribution of $B(E_{n-m-1})$ converges to Ω_0 as $n \to \infty$ because $E_{n-m-1} \to \infty$ (in probability), thus

$$g(n, m) = P(B(E_{n-m-1}) < E_{n-m} - E_{n-m-1}) \to \int_0^\infty \mathrm{e}^{-mz}\Omega_0(\mathrm{d}z) = \frac{1}{\mu m}\int_0^\infty (1 - \mathrm{e}^{-mz})\Omega(\mathrm{d}z),$$

where the last step follows by integrating by parts. Changing measure back to ω we obtain the following result:

Proposition 5. *If ω is not supported by a geometric sequence and satisfies* (3) *then, for any m,*

$$\lim_{n\to\infty} g(n, m) = \frac{W(m)}{\mu m}.$$

The proposition suggests modifying the chain Q_n so that the potential function becomes exactly

$$g_0(m) := \frac{W(m)}{\mu m}, \qquad m = 1, \ldots, n-1.$$

We shall do this by assuming a special distribution for the first transition (which can be thought of as a qualifying round before the game).

Remark. Another possibility would be to introduce a proper initial distribution on $\{0, 1, \ldots, n\}$ so that the formula for the potential function was also valid for $m = n$. But this would correspond to composition of a random integer, a model we wish to avoid.

Renewal theory offers construction of a stationary version of $\mathcal{R}$. Take Z_0 independent of $\mathcal{R}$ and with distribution Ω_0. The shifted set $\mathcal{R}_0 = Z_0 + \mathcal{R}$ is the range of the stationary (delayed) renewal process. For any $z \geqslant 0$ the overshoot distribution for $\mathcal{R}_0$ at z coincides with Ω_0.

The points of $\mathcal{R}_0$ induce an interval partition of $[0, \infty]$, thus also a partition of the sequence of order statistics $E_1, \ldots, E_n$. Recording the sizes of blocks, we obtain a *stationary* composition C_{0n} of n. The parts of C_{0n} are considered as decrements of a new Markov chain Q_{0n}. Repeating the argument which led us to Proposition 5, we derive from invariance of the distribution of $B(z)$ that g_0 is the potential function of Q_{0n}.

For any reward function the solution of (11) satisfies

$$\sum_{m=1}^{n-1} g_0(m) r_m = \sum_{m=1}^{n-1} q_0(n, m) b_{n-m}, \tag{13}$$

where $q_0(n, \cdot)$ is the distribution of the first part of C_{0n}. This formula follows by computing the total expected reward of Q_{0n} upon departure from state n. Including state n leads to

$$r_n + \sum_{m=1}^{n-1} g_0(m) r_m = \sum_{m=0}^{n} w_0(n, m) q_{n-m}, \tag{14}$$

where $w_0(n, \cdot)$ is the distribution of the number of E_js to the left of Z_0. Explicitly,

$$w_0(n, m) = \binom{n}{m} \int_0^\infty (1 - \mathrm{e}^{-z})^m \mathrm{e}^{-(n-m)z} \Omega_0(\mathrm{d}z)$$

and

$$q_0(n, m) = \binom{n}{m} w_0(n, m) + w_0(n, 0) q(n, m).$$

When expressed via binomial moments of ω, this becomes

$$q_0(n, m) = \frac{1}{\mu} \binom{n}{m} \left(\sum_{k=0}^{m} (-1)^{m-k} \frac{W(n-k)}{n-k} + \frac{w(n, m)}{n} \right)$$

for $1 \leqslant m \leqslant n$ (with the convention $W(0)/0 = \mu$ needed for $k = m = n$).

Remarks. The relation between compositions C_{0n} and C_n is that they are identically distributed given the size of the first part. The distribution of C_{0n} is of the form (2) with the first factor replaced by $q_0(n, n_1)$.

The distributional identity $(C_n) \overset{d}{=} (C_{0n})$ holds if and only if $\Omega = \Omega_0$, in which case Ω is an exponential distribution, $\mathcal{R}$ is a homogeneous Poisson point process and therefore C_n is governed by the ordered ESF. This explains, to an extent, the role of ESF as a 'central limit' because superposition of many rare renewal processes approaches the Poisson process.

For suitable choice of Ω the sums on the right-hand side of (13) or (14) become Cesàro or Euler averages. The left-hand side is easy to analyse, but drawing conclusions directly from these relations about the behaviour of (a_n) is only possible when (a_n) is known to satisfy certain regularity conditions (the Tauberian conditions). The direct approach seems hard to realize because the regularity conditions are very sensitive to the summability method.

For $r_n \equiv 1$ the left-hand side of (14) is the expected number of parts of the stationary composition, which is equal to

$$1 + \sum_{m=1}^{n-1} \frac{W(m)}{m\mu}$$

and, quite expectedly, is asymptotic to $\mu^{-1} \log n$.

Example. For ESF(1) we have $W(n) = 1 - (n+1)^{-1}$ and $\mu = 1$, whence the expected number of parts is h_n as is well known (Ewens and Tavaré 1997).

We have seen that $g(n, m) \to g_0(n, m)$ for $n \to \infty$ and wish to obtain the asymptotics of (12) by substituting g_0 for g. To this end we need a stronger assumption on ω,

$$\nu := \int_0^1 (\log(1-x))^2 \omega(\mathrm{d}x) < \infty,$$

which in terms of Ω means finiteness of the second moment, $\nu = \int_0^\infty x^2 \Omega(\mathrm{d}x)$.

Proposition 6. *Suppose that ω is not supported by a geometric sequence and also $\nu < \infty$. Suppose that (r_n) is such that $|r_n| < r'_n$, where (r'_n) is a decreasing sequence satisfying $\sum r'_n/n < \infty$. Then for (b_n) solving* (12) *with such (r_n), we have*

$$\lim_{n\to\infty} b_n = \frac{1}{\mu} \sum_{n=1}^{\infty} \frac{W(n) r_n}{n}.$$

Proof. Given integer J, suppose that (r_n) is such that $r_n = 0$ for $n < J$, is decreasing for $n \geqslant J$ and satisfies $\sum r_n/n < \infty$. We wish to show that for the sequence (b_n) solving (12) with such (r_n) there is a bound

$$\limsup b_n < \frac{1}{\mu} \sum_{n=1}^{\infty} \frac{r_n}{n} + \frac{r_J \nu}{\mu^2}. \tag{15}$$

To this end we will make use of the renewal representation.

Recall that Q_n collects reward r_m is the chain visits state m. This occurs when $\mathcal{R}$ has at least one point between E_{n-m} and E_{n-m+1}, in which case let us assign reward r_m to the rightmost such point (equivalently, given E_{n-m} is the leftmost point in a cluster, the point of $\mathcal{R}$ in question is the left end-point of the component interval $\subset [0,\infty]\backslash\mathcal{R}$ containing E_{n-m}). Let U be the potential measure of $\mathcal{R}$, so that $U[0,z]$ is the expected cardinality of $\mathcal{R} \cap [0,z]$. The total expected reward of Q_n may be written as

$$r_n W(n) + \int_0^\infty \Phi_n(z) U(\mathrm{d}z),$$

where the first term stands for the reward at $0 \in \mathcal{R}$, which is only due in the event that the first jump of the renewal process exceeds E_1, and the integrand is

$$\Phi_n(z) = \sum_{m=1}^{n} \binom{n}{m} \mathrm{e}^{-zm} (1-\mathrm{e}^{-z})^{n-m} r_m W(m). \tag{16}$$

In the same manner, we associate the rewards collected by the stationary chain Q_{0n} with the separating points of $\mathcal{R}_0$ and with 0 (an exceptional point, not in $\mathcal{R}_0$). The expected reward becomes

$$r_n \frac{W(n)}{n\mu} + \int_0^\infty \Phi_n(z) U_0(\mathrm{d}z),$$

where the first term stands for the event that $E_1 < Z_0$ (i.e., E_1 falls to the left of $\mathcal{R}_0$). This is, of course, yet another expression for the right-hand side of (14).

We now modify the reward processes for chains Q_n and Q_{0n} by deleting the first term (reward at 0) and by replacing Φ_n with another function $\tilde{\Phi}_n$, defined by (16) but with factors $W(m)$ deleted. Deleting the first term has no asymptotic effect because r_n goes to 0 as $n \to \infty$. We also have $\Phi_n(z) \leqslant \tilde{\Phi}_n(z)$, thus $\Phi_n(z)$ corresponds to a more generous reward structure, with reward at z being r_m if $E_{n-m} < z < E_{n-m+1}$ (thus there is no other constraint on z except that $z \in \mathcal{R}$ or $z \in \mathcal{R}_0$). The modified reward associated with $\mathcal{R}_0$ is the sum on the left-hand side of (15).

The function $\tilde{\Phi}_n(z)$ is unimodal, with unique maximum attained at z^*, which is the unique positive solution of the equation

$$-r_J\binom{n-1}{J} + \sum_{m=J}^{n-1}(r_m - r_{m+1})\binom{n-1}{m}\left(\frac{\mathrm{e}^{-z}}{1-\mathrm{e}^{-z}}\right)^{m-J} = 0$$

(the uniqueness follows from the monotonicity of (r_j), $j > J$). For $n \to \infty$, Poisson approximation provides asymptotics $n\mathrm{e}^{-z^*} \to \zeta$, where ζ is the unique positive root of the transcendental equation

$$-\frac{r_J}{J!} + \sum_{m=J}^{\infty}(r_m - r_{m+1})\frac{\zeta^{m-J}}{m!} = 0.$$

In the following argument it is only important that $z^* \to \infty$ as $n \to \infty$.

Because $\mathcal{R}_0$ is $\mathcal{R}$ shifted to the right, $\mathcal{R}_0 = Z_0 + \mathcal{R}$, there is a one-to-one correspondence between the sets $\mathcal{R} \cap](z^* - Z_0)_+, z^*]$ and $\mathcal{R}_0 \cap]0, z^*]$. Furthermore, because $\tilde{\Phi}_n$ is increasing on $[0, z^*]$, the total (modified) reward of $\mathcal{R}_0$ over $]0, z^*]$ is larger than that of $\mathcal{R}$ on $]0, (z^* - Z_0)_+]$. On the other hand, the expected reward of $\mathcal{R}$ on $](z^* - Z_0)_+, z^*]$ has an asymptotic bound $r_J \nu/(2\mu^2)$; indeed, $r_J = \max r_j$ is an upper bound for the instantaneous reward and the potential $U](z^* - Z_0)_+, z^*]$ is asymptotic to

$$\mathrm{E}\frac{Z_0}{\mu} = \frac{1}{\mu}\int_0^{\infty} z\Omega[z, \infty]\,\mathrm{d}z = \frac{1}{2\mu}\int_0^{\infty} z^2\Omega(\mathrm{d}z),$$

as follows from the two-term expansion in the renewal theorem, in the case $\nu < \infty$ (see Feller 1971, Section XI.4).

Since the function $\tilde{\Phi}_n$ is decreasing, to the right of z^* the relation is reversed. Shifting the origin to the leftmost point of $\mathcal{R}_0 \cap [z^*, \infty]$ enables one to view $\mathcal{R}$ on the new scale as the range of a delayed renewal sequence. Thus the expected reward of $\mathcal{R}_0$ on $[z^*, \infty]$ is larger than that of $\mathcal{R}$, up to a term estimated by $r_J\nu/(2\mu^2)$, exactly as above. Putting the two parts together shows that the expected modified reward is bounded by the right-hand side of (15). The unmodified reward is smaller, hence (15).

Now suppose that (r_n) is decreasing and satisfies $\sum r_n/n < \infty$. We split the sequence at J and decompose it into $r_n = r_n 1_{\{n<J\}} + r_n 1_{\{n \geqslant J\}}$. Since recursion (12) is linear, the decomposition forces the representation of the solution to take the form, say, $b_n = b'_n + b''_n$. Applying the renewal theorem, we get $b'_n \to \mu^{-1}\sum_{n=1}^{J} r_n W(n)/n$. As for the second part, $\limsup b''_n$ is estimated with the help of (15) and approaches zero when $J \to \infty$, because both r_J and the tail sum of the series vanish.

For an arbitrary sequence satisfying the condition of the proposition, splitting at J yields one part converging to $\mu^{-1}\sum_{n=1}^{J} r_n W(n)/n$ and another part estimated by a solution with reward sequence decreasing for $n > J$, thus going to 0 as J grows. □

4. Asymptotics of moments

We are now in a position to improve on the asymptotics of $a_n = \mathrm{E} K_n$. Recall that the asymptotic expansion of the harmonic number starts with $h_n = \log n + \gamma + O(n^{-1})$.

Proposition 7. *Suppose that ω is not supported by a geometric seqeunce and satisfies* (4), *and that $\nu < \infty$. Then*

$$a_n = \frac{\log n}{\mu} + \frac{\gamma}{\mu} + b + o(1),$$

where γ is the Euler constant and

$$b = \frac{1}{\mu} \int_0^1 \log x \, \omega(\mathrm{d}x) + \frac{\nu}{2\mu^2}.$$

Proof. Writing $a_n = \mu^{-1} h_n + b_n$, substituting this into (7) and using the summation formula (8), we find that (b_n) satisfies (11) with

$$r_n = 1 - \frac{1}{\mu W(n)} \int_0^1 \left(\frac{x}{1} + \frac{x^2}{2} + \ldots + \frac{x^n}{n} \right) \omega(\mathrm{d}x),$$

which can also be written as

$$r_n W(n) = -\int_0^1 (1-x)^n \omega(\mathrm{d}x) + \frac{1}{\mu} \int_0^1 \left(\frac{x^{n+1}}{n+1} + \frac{x^{n+2}}{n+1} + \ldots \right) \omega(\mathrm{d}x).$$

Using monotone convergence and manipulating the series, we find that

$$\sum_{n=1}^{\infty} \frac{r_n W(n)}{n\mu} = \frac{1}{\mu} \int_0^1 \log x \omega(\mathrm{d}x) + \frac{1}{2\mu^2} \int_0^1 (\log(1-x))^2 \omega(\mathrm{d}x) = b.$$

Since $W(n) \to 1$ and $r_n W(n)$ is the difference between two terms which decrease in n, application of Proposition 6 yields $b_n \to b$. □

With no additional assumptions we will derive asymptotics of the variance $v_n = \operatorname{var} K_n$. The key issue is the asymptotic evaluation of the inhomogeneous term of the recursion.

Lemma 8. *Under the assumptions of Proposition 7, the expectation $a_n = \mathrm{E} K_n$ satisfies*

$$\lim_{n \to \infty} \left(2a_n - 1 - a_n^2 + \sum_{m=1}^{n-1} q(n, m) a_{n-m}^2 \right) = \frac{\nu}{\mu^2} - 1.$$

Proof. For b_n, r_n having the same meaning as in Proposition 7, we have

$$b_n \to b, \qquad W(n) \to 1, \qquad b_n - \sum_{m=1}^{n} q(n, m) b_{n-m} = r_n,$$

and integrating by parts yields

$$r_n W(n) = -n \int_0^1 \omega[0, x](1-x)^{n-1} \mathrm{d}x + \frac{1}{\mu} \int_0^1 \frac{\omega[x, 1] x^n}{1-x} \mathrm{d}x.$$

Further useful estimates follow from the $n \to \infty$ asymptotics

$$\int_0^1 x^n \omega(\mathrm{d}x) = o\left(\frac{1}{\log n}\right), \qquad \int_0^1 (1-x)^n \omega(\mathrm{d}x) = o\left(\frac{1}{\log n}\right), \qquad r_n = o\left(\frac{1}{\log n}\right). \tag{17}$$

To justify the first relation, observe that integrability and monotonicity of $\log(1-x)$ imply that $\omega[x, 1] = o(|\log(1-x)|^{-1})$ for $x \uparrow 1$ (in fact, the relation is equivalent to the integrability). Integrating by parts and using monotonicity, we have

$$\int_0^1 x^n \omega(\mathrm{d}x) = \int_0^1 n x^{n-1} \omega[x, 1] \mathrm{d}x < \text{const.} \int_0^1 n x^{n-1} |\log x|^{-1} \, \mathrm{d}x,$$

and by a Tauberian argument this is $o(|\log n|^{-1})$. The second relation follows in the same way from

$$\int_0^1 (1-x)^n \omega(\mathrm{d}x) = n \int_0^1 (1-x)^{n-1} \omega[0, x] \, \mathrm{d}x < \text{const.} \int_0^1 n(1-x)^{n-1} |\log(1-x)|^{-1} \, \mathrm{d}x$$

and $\omega[0, x] = o(|\log x|^{-1})$ for $x \downarrow 0$. The third relation follows from the first two.

Substituting $a_n = \mu^{-1} h_n + b_n$ and grouping terms, we have

$$2a_n - 1 - a_n^2 + \sum_{m=1}^n q(n, m) a_{n-m}^2 = T_1 + T_2 + T_3 - 1$$

with three terms to be evaluated:

$$T_1 = -b_n^2 + \sum_{m=1}^n q(n, m) b_{n-m}^2,$$

$$T_2 = 2b_n - 2b_n \frac{h_n}{\mu} - \frac{2}{\mu} \sum_{m=1}^n q(n, m) b_{n-m} h_{n-m},$$

$$T_3 = \frac{2h_n}{\mu} - \frac{h_n^2}{\mu^2} + \frac{1}{\mu^2} \sum_{m=1}^n q(n, m) h_{n-m}^2.$$

From $b_n \to b$ it is obvious that $T_1 \to 0$ as $n \to \infty$. To see that T_2 also vanishes, write

$$b_{n-m} h_{n-m} = b_{n-m} h_n + (b_{n-m} - b)(h_{n-m} - h_n) + b(h_{n-m} - h_n);$$

then from (8) and (17) we obtain

$$\sum_{m=1}^{n-1} q(n, m)(h_{n-m} - h_n) = \frac{1}{W(n)} \int_0^1 \left(h_n x^n - \sum_{j=1}^n \frac{x^j}{j} \right) \omega \, \mathrm{d}x,$$

hence by (17) and Lemma 2,

$$b\sum_{m=1}^{n} q(n,\, m)h_{n-m} - h_n) \to -b\mu,$$

$$\sum_{m=1}^{n} q(n,\, m)(b_{n-m} - b)(h_{n-m} - h_n) \to 0,$$

$$\sum_{m=1}^{n} q(n,\, m)b_{n-m}h_n = (b_n - r_n)h_n = b_n h_n + o(1),$$

which indeed implies $T_2 \to 0$. To evaluate T_3 we need a summation formula similar to (8), but this time we should take a combinatorial analogue of $\log^2$ in place of log. To this end, introduce

$$s_n = \sum_{1\leqslant i\leqslant j\leqslant n} \frac{1}{ij},$$

then there is a summation formula

$$\sum_{m=0}^{n-1} \binom{n}{m} x^m (1-x)^{n-m} s_{n-m} = s_n - \sum_{j=1}^{n} \frac{x^j}{j}(h_n - h_{j-1}) \tag{18}$$

where we recognize a partial sum of the Taylor series

$$\frac{1}{2}(\log(1-x))^2 = \sum_{j=1}^{\infty} \frac{x^j}{j} h_{j-1}.$$

It follows that

$$\sum_{m=1}^{n-1} q(n,\, m)s_{n-m} = s_n - \frac{1}{W(n)}\int_0^1 \left(\sum_{j=1}^{n} \frac{x^j}{j}(h_n - h_{j-1})\right)\omega(\mathrm{d}x),$$

and because h_n^2 differs from $2s_n$ by the partial sum of a converging series,

$$h_n^2 = 2s_n - \sum_{j=1}^{n} \frac{1}{j^2},$$

we conclude that

$$\sum_{m=1}^{n-1} q(n,\, m)h_{n-m}^2 = h_n^2 - \frac{2}{W(n)}\int_0^1 \left(\sum_{j=1}^{n} \frac{x^j}{j}(h_n - h_{j-1})\right) w(\mathrm{d}t) + o(1)$$

$$= h_n^2 - \frac{2\mu h_n}{W(n)} + \frac{\nu}{W(n)} + o(1) = h_n^2 - 2\mu h_n + \nu + o(1),$$

where we have exploited monotone convergence and (17). Now it is easily seen that $T_3 \to \nu/\mu^2$.

Putting the terms together, we arrive at $T_1 + T_2 + T_3 - 1 \to 1 - \nu/\mu^2$. □

Remark. The summation formula (18) implies an analogue of Lemma 2: for arbitrary normalized weight ω, the square of logarithm is $L^2([0, 1], \omega)$-approximable by its Bernstein polynomial.

Appealing to Corollary 4, we obtain the desired asymptotics of variance. Define $\sigma^2 = \nu - \mu^2$, that is, $\sigma^2 = \int_{[0,\infty}(z - \mu)^2\Omega(\mathrm{d}z)$ is the variance of distribution Ω.

Proposition 9. *Under the assumptions of Proposition 7,*

$$\mathrm{var}\, K_n \sim \frac{\sigma^2}{\mu^3} \log n.$$

5. A central limit theorem

We turn next to the central limit theorem (CLT) for K_n. Neininger and Rüschendorf (2002) derived a general CLT for solutions of equations such as (6) (with an error estimate in a suitable probability metric). In our context, the assumptions of their Theorem 2.1 are easily checked, with the only exception that their result requires some expansion $\mathrm{var}\, K_n = \mu^{-1} \log n + O((\log n)^{1-\epsilon})$, which is not guaranteed by the integrability of $(\log(1 - x))^2$ but rather relies on integrability of a higher power of the logarithm. We shall see that in our situation no additional assumptions are necessary and the CLT follows by a simple comparison with the number of renewals.

Given n, define a *cell* to be a component interval of $[0, \infty]\backslash\mathcal{R}$ containing at least one E_j, $j \leqslant n$. Clearly, the total number of cells is K_n. Let L_n be the number of cells which have left end-point smaller than $\log n$, and let R_n be the number of renewals on $[0, \log n]$ (including 0), that is $R_n = \#(\mathcal{R} \cap\ [0, \log n])$. It is an easy matter to see that $L_n \leqslant R_n$ and $L_n \leqslant K_n$. Moreover, since ther expected number of order statistics that exceed $\log n$ is 1, we have $\mathrm{E}(K_n - L_n) < 1$.

Proposition 10. *Under the assumptions of Proposition 7,*

$$\frac{K_n - \mu^{-1} \log n}{(\sigma^2\mu^{-3} \log n)^{1/2}} \tag{19}$$

converges weakly to the standard normal randon variable.

Proof. From Feller (1971, Section XI.5) we know that R_n is asymptotically normal with expectation $\mu^{-1} \log n$ and variance $\sigma^2\mu^{-3} \log n$, and from Feller (1971, Section XI.3) that $\mathrm{E}R_n = \mu^{-1} \log n + \nu(2\mu^2)^{-1} + o(1)$. By asymptotics of moments (Propositions 7 and 9) and the above inequalities, the L^1-distance between any two of the three random variables $(K_n - a_n)v_n^{-1/2}$, $(L_n - a_n)v_n^{-1/2}$ and $(R_n - a_n)v_n^{-1/2}$ goes to zero. It follows that L_n and K_n are also asymptotically normal. □

In fact, the renewal theorem, taken together with a Poisson limit for the number of E_js exceeding $\log n$, implies weak convergence of $K_n - L_n$. Asymptotics of the expectation involves the exponential integral function

$$I(z) = \int_z^\infty \mathrm{e}^{-y} y^{-1}\, \mathrm{d}y.$$

Proposition 11. *We have*

$$\lim_{n\to\infty} \mathrm{E}(K_n - L_n) = \frac{\gamma}{\mu} + \frac{1}{\mu}\int_0^1 \log x\,\omega(\mathrm{d}x) + \frac{1}{\mu}\int_0^1 I(x)\omega(\mathrm{d}x).$$

Proof. Recalling (16), using Poisson approximation and the renewal theorem, and changing the variable of integration for $\zeta = n\mathrm{e}^{-z}$, we compute

$$\mathrm{E}(K_n - L_n) = \frac{1}{\mu}\int_{\log n}^\infty \Phi_n(z)\mathrm{d}z + o(1) = \int_0^1 \mathrm{e}^{-\xi} \sum_{m=1}^\infty \frac{\zeta^m W(m)}{m!}\frac{\mathrm{d}\zeta}{\mu\zeta} + o(1)$$

$$= \int_0^1\int_0^1 \frac{1-\mathrm{e}^{-zx}}{\mu z}\mathrm{d}z\omega(\mathrm{d}x) + o(1) = \frac{\gamma}{\mu} + \frac{1}{\mu}\int_0^1 \log x\,\omega(\mathrm{d}x) + \frac{1}{\mu}\int_0^1 I(x)\omega(\mathrm{d}x) + o(1),$$

where we have also used

$$\mathrm{e}^{-\zeta}\sum_{m=1}^\infty \frac{\zeta^m(1-(1-x)^m)}{m!} = 1 - \mathrm{e}^{-\zeta x}$$

and the well-known formula

$$\int_0^x \frac{1-\mathrm{e}^{-y}}{y}\mathrm{d}y = I(x) + \log x + \gamma.$$

□

Now recalling

$$\mathrm{E}K_n = \frac{\log n}{\mu} + \frac{\gamma}{\mu} + \frac{1}{\mu}\int_0^1 \log x\,\omega(\mathrm{d}x) + \frac{\nu}{2\mu^2} + o(1)$$

and comparing the expectations

$$\mathrm{E}L_n = \frac{\log n}{\mu} + \frac{\nu}{2\mu^2} - \frac{1}{\mu}\int_0^1 I(x)\omega(\mathrm{d}x) + o(1), \qquad \mathrm{E}R_n = \frac{\log n}{\mu} + \frac{\nu}{2\mu^2} + o(1),$$

we not only confirm ‘by computation’ the inequality $\mathrm{E}L_n \leqslant \mathrm{E}R_n$ $(= U[0, \log n])$ but also come to the conclusion that the number of interval components of $[0, \log n]\backslash\mathcal{R}$ which contain no E_js, $j \leqslant n$, remains bounded as $n \to \infty$. This conclusion is in good accord with a general point that the composition C_n is a proper combinatorial analogue of the regenerative set $\mathcal{R}$, as in Gnedin (2003), Gnedin and Pitman (2003) and Gnedin *et al.* (2003).

Acknowledgement

I would like to thank Andrey Levin for help with (18) and further Euler summation formulae.

References

Arratia, R., Barbour, A.G. and Tavaré, S. (2002) *Logarithmic Combinatorial Structures: A Probabilistic Approach.* To appear.

Bruhn, V. (1996) Eine Methode zur asymptotischen Behandlung einer Klasse von Rekursionsgleichungen mit einer Anwendung in der stochastischen Analyse des Quicksort-Algorithmus. Doctoral dissertation, Universität zu Kiel.

Bruss, T. and O'Cinneide, C. (1990) On the maximum and its uniqueness for geometric random samples. *J. Appl. Probab.*, **27**, 598–610.

Ewens, W.J. and Tavaré, S. (1997) The Ewens sampling formula. In N.S. Johnson, S. Kotz and N. Balakrishnan (eds), *Discrete Multivariate Distributions*. New York: Wiley.

Feller, W. (1971) *An Introduction to Probability Theory and its Applications*, Vol. 2, 2nd edn. New York: Wiley.

Flajolet, P. (1999) Singularity analysis and asymptotics of Bernoulli sums. *Theoret. Comput. Sci.*, **215**, 371–381.

Gnedin, A.V. (1997) The representation of composition structures. *Ann. Probab.*, **25**, 1437–1450.

Gnedin, A.V. (1998) On the Poisson–Dirichlet limit. *J. Multivariate Anal.*, **67**, 90–98.

Gnedin, A.V. (2003) Three sampling formulas. *Combin. Probab. Comput.* To appear.

Gnedin, A.V. and Pitman, J. (2003) Regenerative composition structures. Technical report 644, Dept. of Statistics, University of California at Berkeley.

Gnedin, A.V., Pitman, J. and Yor, M. (2003) Asymptotic laws for regenerative composition structures. In progress.

Jacquet, P. and Szpankowski, W. (1999) Entropy computations via analytic depoissonization. *IEEE Trans. Inform. Theory*, **45**, 1072–1081.

Karlin, S. (1967) Central limit theorems for certain infinite urn schemes. *J. Math. Mech.*, **17**, 373–401.

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

Lorenz, G. (1953) *Bernstein Polynomials*. Toronto: University of Toronto Press.

Neininger, R. and Rüschendorf, L. (2002) On the contraction method with degenerate limit equations. Preprint.

Pitman, J. (2002) Combinatorial stochastic processes. Technical report 621, Dept. of Statistics, University of California at Berkeley.

Rösler, U. (2001) On the analysis of stochastic divide and conquer algorithms. *Algorithmica*, **29**, 238–261.

Rösler, U. and Rüschendorf, L. (2001) The contraction method for recursive algorithms. *Algorithmica*, **29**, 3–33.

Received March 2003 and revised October 2003