Regenerative Composition Structures *

Alexander Gnedin[†] and Jim Pitman[‡]

Technical Report No. 644

Department of Statistics
University of California
367 Evans Hall # 3860
Berkeley, CA 94720-3860

July 21, 2003; revised December 8, 2003

Abstract

A new class of random composition structures (the ordered analog of Kingman's partition structures) is defined by a regenerative description of component sizes. Each regenerative composition structure is represented by a process of random sampling of points from an exponential distribution on the positive halfline, and separating the points into clusters by an independent regenerative random set. Examples are composition structures derived from residual allocation models, including one associated with the Ewens sampling formula, and composition structures derived from the zero set of a Brownian motion or Bessel process. We provide characterisation results and formulas relating the distribution of the regenerative composition to the Lévy parameters of a subordinator whose range is the corresponding regenerative set. In particular, the only reversible regenerative composition structures are those associated with the interval partition of $[0,1]$ generated by excursions of a standard Bessel bridge of dimension $2-2\alpha$ for some $\alpha \in [0,1]$.

AMS 2000 *subject classifications.* Primary 60G09, 60C05. Keywords: exchangeability, composition structure, regenerative set, sampling formula, subordinator

*Research supported in part by N.S.F. Grant DMS-0071448
†Utrecht University; e-mail gnedin@math.uu.nl
‡University of California, Berkeley; e-mail pitman@stat.Berkeley.EDU

1 Introduction

A *composition* of a positive integer n is a sequence of positive integers $\lambda = (n_1, \dots, n_k)$ with sum $\Sigma_j n_j = n$. Each n_i may be called a *part* of the composition. We will use the notation $\lambda \models n$ to say that λ is a composition of n. A *random composition* of n is a random variable $\mathcal{C}_n$ with values in the set of all 2^{n-1} compositions of n. A *composition structure* $(\mathcal{C}_n)$ is a Markovian sequence of random compositions of n, for $n = 1, 2, \dots$, whose cotransition probabilities are determined by the following property of *sampling consistency* [12], [17]: if n identical balls are distributed into an ordered series of boxes according to $(\mathcal{C}_n)$, then $\mathcal{C}_{n-1}$ is obtained by discarding one of the balls picked uniformly at random, and then deleting an empty box in case one is created. We study composition structures with the following further property:

Definition 1.1 A composition structure $(\mathcal{C}_n)$ is *regenerative* if for all $n > m \geq 1$, given that the first part of $\mathcal{C}_n$ is m, the remaining composition of $n - m$ is distributed like $\mathcal{C}_{n-m}$.

According to our main result (Theorem 5.3), each regenerative composition structure can be represented by a process of random sampling of points from the exponential distribution on $[0, \infty[$, and separating the sample points into clusters by points of an independent regenerative random closed subset $\mathcal{R}$ of $[0, \infty[$. We recall in Theorem 5.1 the fundamental result of Maisonneuve [29] that every such $\mathcal{R}$ can be represented as the closed range of a *subordinator* (S_t), that is an increasing process with stationary independent increments. Each possible distribution of a regenerative composition structure is thereby described in terms of the drift coefficient $\mathtt{d}$ and Lévy measure ν of an associated subordinator.

Alternatively, we can transform $\mathcal{R}$ into $\widetilde{\mathcal{R}} := 1 - \exp(-\mathcal{R}) \subset [0, 1]$ and replace the exponential sample by a sample from the uniform distribution on $[0, 1]$. In this form the construction is an instance of the *ordered paintbox representation* of composition structures, developed in [12], [17],[32]. Keeping track of only the sizes of parts, and not their order, every composition structure induces a *partition structure*, that is a sequence of sampling consistent *partitions* of integers, as studied by Kingman [27, 28]. Passing from compositions to partitions is equivalent to passing from the ordered paintbox $\widetilde{\mathcal{R}}^c = [0, 1] \setminus \widetilde{\mathcal{R}}$ to *Kingman's paintbox* defined by the decreasing sequence of lengths of interval components of $\widetilde{\mathcal{R}}^c$. A partition structure is thereby associated with a typically infinite collection of composition structures, each corresponding to a different way of ordering interval components of given lengths. We show that if one of these composition structures is regenerative, it is unique in distribution (Corollary 9.3). In Section 9.1 we also discuss necessary and sufficient conditions for the existence of such a regenerative rearrangement.

Known examples of regenerative composition structures include the compositions associated with the ordered Ewens sampling formula [12], and those derived from the zero set of a recurrent Bessel process in [32]. The partition structures corresponding to these examples are instances of the two parameter family of partition structures studied in [31, 34]. We show in Section 10 that each member of this family, with positive values of parameters, corresponds to a unique regenerative composition structure. Also (Theorem 12.1), the only reversible regenerative composition structures are the members of this family associated with the interval partition of $[0, 1]$ generated by excursions of a standard Bessel bridge of dimension $2 - 2\alpha$ for some $\alpha \in [0, 1]$. See also Section 4 and [16], [15] for further examples of regenerative composition structures.

Our definition of regenerative composition stuctures is reminiscent of Kingman's characterisation of the one-parameter Ewens partition structure by invariance with respect to deletion of a random part, selected in a size-biased fashion. This property is called *species noninterference* or *neutrality* in the setting of population genetics. We refer to [3], [13], [34], for background on partition structures, exchangeability and related matters. As shown by James [22], another closely related concept, developed in the setting of Bayesian nonparametric statistics, is Doksum's [11] notion of a random discrete probability distribution that is *neutral to the right*.

From an algebraic viewpoint, our representation of regenerative composition structures is equivalent to solving a nonlinear recurrence (Proposition 3.3). The nonlinearity of the recursion reflects the fact that the family of probability laws of regenerative compositions is not closed under mixtures. So unlike the problems of characterising all partition or composition structures, the problem of characterising all regenerative composition structure, is not just a problem of identifying the extreme points of a convex set. Still, we show in Section 8 that it can be reduced to such a problem (the Hausdorff moment problem), by a suitable non-linear transformation. The Lévy data $(\mathtt{d}, \nu)$ of the associated subordinator are thereby encoded in a finite measure on $[0,1]$.

2 Compositions and partitions

This section recalls briefly some background material on composition structures and their associated partition structures. See [17], [12], [32], [31], [34] for a fuller account. For a composition structure $(\mathcal{C}_n)$, and a composition $\lambda = (n_1, \ldots, n_k)$ of n, define the *composition probability function* p by

$$p(\lambda) := \mathbb{P}(\mathcal{C}_n = \lambda). \tag{1}$$

For each fixed n, this function defines a probability distribution on the set of compositions $\lambda \models n$, and these distributions are subject to the following linear relation describing the sampling consistency. For $\lambda = (n_1, \ldots, n_k) \models n$ and $\mu \models n+1$ we say that μ *extends* λ and write $\mu \searrow \lambda$ if μ is obtained from λ by either increasing a part n_j by one or by inserting a part 1 in the sequence λ. The sampling consistency amounts to the recursion

$$p(\lambda) = \sum_{\mu \searrow \lambda} \kappa(\lambda, \mu) p(\mu)\,, \qquad p(1) = 1 \tag{2}$$

where the coefficient $\kappa(\lambda, \mu)$ equals $(n_j + 1)/(n+1)$ if μ is obtained by increasing a part n_j, and equals $(j+1)/(n+1)$ if μ is obtained by inserting a 1 into a row of consecutive ones $1, 1, \ldots, 1$ of length $j \geq 0$.

Regard $\mathcal{C}_n$ as a way to partition a row of n identical balls into an ordered series of non-empty boxes, and independently of $\mathcal{C}_n$ let the balls be labelled by a uniform random permutation of the set $[n] := \{1, \ldots, n\}$. This defines a random *exchangeable ordered partition* $\mathcal{C}_n^*$ of the set $[n]$ whose distribution is defined as follows. For each *particular* ordered partition of $[n]$ into k classes of sizes $n_1, \ldots, n_k$, say c^*,

$$\mathbb{P}(\mathcal{C}_n^* = c^*) = \binom{n}{n_1, \ldots, n_k}^{-1} p(n_1, \ldots, n_k) \tag{3}$$

since the multinomial coefficient is the number of such ordered partitions of $[n]$, and these are equally likely. The sampling consistency property of a composition structure $(\mathcal{C}_n)$ means that $(\mathcal{C}_n^*)$ can be constructed *consistently*, in the sense that $\mathcal{C}_{n-1}^*$ is the restriction of $\mathcal{C}_n^*$ obtained by deleting element n. Then $\mathcal{C}_n$ is the ordered record of sizes of classes of $\mathcal{C}_n^*$, and the entire sequence $(\mathcal{C}_n^*)$ defines an exchangeable ordered partition of the set $\mathbb{N}$ of all positive integers.

Ignoring the order of classes yields a random *exchangeable partition* Π of the set $\mathbb{N}$. The restriction Π_n of Π to $[n]$ is obtained by ignoring the order of classes of $\mathcal{C}_n^*$. So for each *particular* partition π of $[n]$ into k classes whose sizes in some order are $n_1, \ldots, n_k$,

$$\mathbb{P}(\Pi_n = \pi) = \binom{n}{n_1, \ldots, n_k}^{-1} \sum_{\sigma} p(n_{\sigma(1)}, \ldots, n_{\sigma(k)}) \tag{4}$$

where the sum is over the $k!$ permutations of $[k]$, corresponding to the $k!$ different ordered partitions c^* of $[n]$ derived from the given partition π of $[n]$. This symmetric function of $(n_1, \ldots, n_k)$ is the *exchangeable partition probability function (EPPF)* of [31, 34]. Note by construction that the partition of n defined by the decreasing rearrangement of sizes of classes of Π_n, or of $\mathcal{C}_n^*$, is identical to the decreasing rearrangement of the parts of $\mathcal{C}_n$. Such a sequence of random partitions of n, subject to a consistency constraint, is called a *partition structure*.

3 Regenerative composition structures

Let $(\mathcal{C}_n)$ be a composition structure with composition probability function p. Let F_n denote the size of the first part of $\mathcal{C}_n$, and denote the distribution of F_n by

$$q(n:m) := \mathbb{P}(F_n = m) = \sum_{(n_1, \ldots, n_k)} 1(n_1 = m) p(n_1, \ldots, n_k), \qquad 1 \leq m \leq n, \tag{5}$$

where the sum is over all compositions $(n_1, \ldots, n_k)$ of n, and $1(\cdots)$ denotes the indicator function which equals 1 if $\cdots$ and 0 else. We call q the *decrement matrix* of the composition structure $(\mathcal{C}_n)$.

Proposition 3.1 *A composition structure $(\mathcal{C}_n)$ is regenerative in the sense of Definition* 1.1 *iff for each $n = 1, 2, \ldots$ the distribution of $\mathcal{C}_n$ is determined by the* product formula

$$p(n_1, \ldots, n_k) = \prod_{j=1}^{k} q(N_j : n_j) \tag{6}$$

for each composition $(n_1, \ldots, n_k)$ of n, where $N_j := n_j + \cdots + n_k$ and $q(n : m)$ is the decrement matrix defined by (5). *Thus the law of a regenerative composition structure is uniquely determined by its decrement matrix.*

Proof. This is easily shown by induction on the number of parts of a composition. □

Note that if $q(2 : 1) = 1$ then $q(n : m) = 1(m = 1)$, meaning that each $\mathcal{C}_n$ is a pure singleton composition, with $p(1, 1, \ldots, 1) \equiv 1$. Whereas if $q(2 : 2) = 1$ then $q(n : m) = 1(m = n)$ meaning that each $\mathcal{C}_n$ is a trivial one-part composition with $p(n) \equiv 1$. These facts are easy to check using (2) and $p \geq 0$, and they are intuitively obvious: $q(2 : 1) = 1$ (respectively $q(2 : 1) = 0$) means that two randomly sampled balls never come from the same box (respectively from different boxes). In all other cases $0 < q(n : m) < 1$ for all $n > 1$ and therefore $0 < p(\lambda) < 1$ for $\lambda \models n > 1$.

The product formula (6) identifies $\mathcal{C}_n$ with the sequence of decrements of a transient Markov chain $Q_n := Q_n(0), Q_n(1), \ldots$ with values in $\{0, \ldots, n\}$. This chain has decreasing paths starting from the state $Q_n(0) = n$, with the terminal state 0 and time-homogeneous triangular transition matrix $(q(n : n-m), 1 \leq m \leq n < \infty)$. In this interpretation the parts of a composition $n_1, \ldots, n_k$ are the magnitudes of jumps of the chain, while $(N_1, \ldots, N_k)$ is the path of Q_n prior to absorbtion. For example, if $\mathcal{C}_8 = (3, 2, 1, 2)$, the path of Q_8 is

$$(Q_8(0), Q_8(1), \ldots) = (8, 5, 3, 2, 0, 0, \ldots).$$

Consider now the joint law of two compositions derived from a regenerative composition $\mathcal{C}_n$ by a random splitting, say $\mathcal{C}_n = (\mathcal{C}_n^<, \mathcal{C}_n^>)$, where $\mathcal{C}_n^<$ is a composition of $m(\mathcal{C}_n^<) \in \{1, \ldots, n\}$, and $\mathcal{C}_n^>$ is the remaining composition of $n - m(\mathcal{C}_n^<)$, regarded as a trivial sequence with no elements if $m(\mathcal{C}_n^<) = n$. Suppose that the number of parts of $\mathcal{C}_n^<$ is a *randomised stopping time* of the chain Q_n, meaning [36] that for each $1 \leq k \leq n$, given $\mathcal{C}_n$ with at least k parts, the conditional probability that $\mathcal{C}_n^<$ has exactly k parts depends only on the first k parts of $\mathcal{C}_n$. Equivalently, for each $\lambda = (n_1, \ldots, n_\ell) \models n$ and each $\lambda^< = (n_1, \ldots, n_k)$ for some $1 \leq k \leq \ell$,

$$\mathbb{P}(\mathcal{C}_n^< = \lambda^< \,|\, \mathcal{C}_n = \lambda) = f_n(\lambda^<) \tag{7}$$

for some function f_n of compositions of m for $1 \leq m \leq n$. The strong Markov property of Q_n then implies that

(i) the compositions $\mathcal{C}_n^<$ and $\mathcal{C}_n^>$ are conditionally independent given $m(\mathcal{C}_n^<)$, and

(ii) for each $1 \leq m < n$, given $m(\mathcal{C}_n^<) = m$ the remaining composition $\mathcal{C}_n^>$ of $n - m$ is distributed like $\mathcal{C}_{n-m}$.

Conversely, we record the following proposition which applies in particular to the splitting scheme defined by (7) with $f_n(n_1, \ldots, n_k) = n_k/n$. In terms of balls in boxes, such a split is made just to the right of the box containing a ball picked uniformly at random.

Proposition 3.2 *Suppose a composition structure $(\mathcal{C}_n)$ admits a random splitting $\mathcal{C}_n = (\mathcal{C}_n^<, \mathcal{C}_n^>)$ for each n, such that* (7) *holds with $f_n(m) > 0$ for all $1 \leq m < n$, and* (ii) *holds. Then $(\mathcal{C}_n)$ is regenerative.*

Proof. Let p denote the composition probability function of $(\mathcal{C}_n)$, as in (1). By definition, $(\mathcal{C}_n)$ is regenerative iff for all $1 \leq m < n$ and all compositions $\lambda^>$ of $n-m$

$$p(m, \lambda^>) = q(n:m)p(\lambda^>) \tag{8}$$

for some matrix $q(n:m)$, which is then the decrement matrix of $(\mathcal{C}_n)$. Whereas (ii) holds iff for all $1 \leq m < n$ and all compositions $\lambda^<$ of m and $\lambda^>$ of $n-m$

$$\sum_{\lambda^< \models m} f_n(\lambda^<)p(\lambda^<, \lambda^>) = \hat{q}(n:m)p(\lambda^>) \tag{9}$$

for some matrix $\hat{q}(n:m)$, in which case $\hat{q}(n:m) = \mathbb{P}(m(\mathcal{C}_n^<) = m)$. Assuming that (9) holds, (8) is obvious for $m = 1$ with $q(n:1) = \hat{q}(n:1)/f_n(1)$. Proceeding by induction on m, suppose that (9) holds for all $1 \leq m < n$, and that (8) has been established with m' instead of m for all $1 \leq m' < m < n$. Apart from the term $f_n(m)p(m, \lambda^>)$, all terms of the sum in (9) involve compositions $\lambda^<$ all of whose parts are smaller than m. So the inductive hypothesis allows us to write these terms as $f_n(\lambda^<)h_n(\lambda^<)p(\lambda^>)$ where $h_n(\lambda^<)$ is a product of entries of the decrement matrix q. Now rearrange (9) to isolate the term $f_n(m)p(m, \lambda^>)$ on the left, and observe that $p(\lambda^>)$ is a common factor on the right, to complete the induction. □

Our aim now is to describe as explicitly as possible all matrices q which define a composition structure by means of (6). We start with an algebraic description:

Proposition 3.3 *A non-negative matrix q is the decrement matrix of some regenerative composition structure iff $q(1:1) = 1$ and*

$$q(n:m) = \frac{m+1}{n+1}\, q(n+1:m+1) + \frac{n+1-m}{n+1}\, q(n+1:m) + \frac{1}{n+1}\, q(n+1:1)\, q(n:m) \tag{10}$$

for $1 \leq m \leq n$.

Proof. We will show first that the condition (10) is sufficient, that is (10) and (6) imply (2). Indeed, assuming (10) and (6)

$$q(n:n) = q(n+1:n+1) + \frac{1}{n+1}q(n+1:n) + \frac{1}{n+1}q(n+1:1)q(n:n)$$

implies readily

$$p(n) = p(n+1) + \frac{1}{n+1}p(n,1) + \frac{1}{n+1}p(1,n)$$

which means (2) for all one-part compositions. Now suppose (2) holds for all compositions with less than k parts, and let $\lambda \models n$ be a composition with k parts. Write λ in the form $\lambda = (m, \lambda')$ where $\lambda' \models n-m$. We have by the induction hypothesis and (6)

$$\sum_{\mu \searrow \lambda} \kappa(\lambda, \mu)p(\mu) = \frac{1}{n+1}p(1,\lambda) + \frac{m+1}{n+1}p(m+1, \lambda') + \frac{n-m+1}{n+1} \sum_{\mu' \searrow \lambda'} \kappa(\lambda', \mu')p(m, \mu') =$$
$$\frac{1}{n+1}q(n+1:1)q(n:m)p(\lambda') + \frac{m+1}{n+1}q(n+1:m+1)p(\lambda') + \frac{n-m+1}{n+1}q(n+1:m)p(\lambda')$$

which by (10) and (6) is equal to $q(n:m)p(\lambda') = p(\lambda)$ and the induction step is completed.

Conversely, assuming (2) and (6) the recursion (10) follows by a similar argument with $k = 2$. □

4 First examples

Example 1 (Geometric sampling [8, 23]). Imagine infinitely many players labeled $1, 2, \ldots$ who flip repeatedly the same coin with fixed probability $x \in]0,1]$ for tails. In the first round, each of the players tosses the coin and those who flip tails drop out. In the second round each of the remaining players must toss again and those who flip tails drop out, and so on. If we restrict consideration to players labeled $1, \ldots, n$, a composition $\mathcal{C}_n$ arises by arranging the players into groups as they drop out. These compositions are sampling consistent by exchangeability among the players and they form a regenerative composition structure because 'all rounds are the same'. Equivalently, we could attribute to each player j an individual value ξ_j, the number of rounds the player remains in the game, and tie the players into blocks by equality of their individual values. The ξ_j are independent with same geometric distribution. The probability that of n players exactly m tie for the minimum value $\min(\xi_1, \ldots, \xi_n)$ is equal to

$$q(n:m) = \frac{\binom{n}{m} x^m (1-x)^{n-m}}{1-(1-x)^n}\,, \qquad m = 1, \ldots, n$$

which is the binomial distribution conditioned on a positive value. Note that the one-part or the pure singleton compositions appear for $x = 1$ or $x \downarrow 0$, respectively.

It is the memoryless property which makes the geometric distribution work, and sampling from any other *fixed* distribution on integers would not produce a regenerative composition. Still, it is possible to preserve the regenerative feature by randomising the distribution in a very special way.

Example 2 (Stick-breaking compositions [40, 22, 15, 16, 12, 21]). Let (X_k) be independent copies of some random variable X with $0 < X \leq 1$. Think of X_k as the probability of tails for the kth coin. Modify the algorithm in the previous example by requiring that at round k each of the remaining players must toss the kth coin. It is easily seen that the resulting composition structure is regenerative. Fixing a group of n players and conditioning on the number of players that drop out at the first coin-tossing trial we obtain the recurrence

$$q(n:m) = \binom{n}{m} \mathbb{E}\left(X^m (1-X)^{n-m}\right) + \mathbb{E}\,(1-X)^n\, q(n:m)$$

resulting in the decrement matrix

$$q(n:m) = \frac{\binom{n}{m} \mathbb{E}\,(X^m(1-X)^{n-m})}{\mathbb{E}\,(1-(1-X)^n)} \qquad m = 1, \ldots, n \tag{11}$$

which says that $q(n:\cdot)$ is a mixture of binomial distributions conditioned on a positive value.

For example, if X is uniform on $[0,1]$, then $q(n:m) = n^{-1}$, that is a discrete uniform distribution for each n. More generally, if X has a beta distribution with parameters $(1, \theta)$, $\theta > 0$, the decrement matrix becomes

$$q(n:m) = \binom{n}{m} \frac{[\theta]_{n-m}\, m!}{[\theta+1]_{n-1}\, n}\,, \tag{12}$$

where

$$[\theta]_n := \theta(\theta+1)\cdots(\theta+n-1) \tag{13}$$

is a rising factorial. The corresponding partition structure is well known to be that defined by the Ewens sampling formula [13]. The individual values of the players are now only conditionally i.i.d., with conditional distribution

$$\mathbb{P}(\xi_j = i \,|\, X_1, X_2, \ldots)) = (1-X_1)\cdots(1-X_{i-1})X_i\,.$$

Additional randomisation allows the same composition structure to be defined in another way. Mark the players by independent uniform [0,1] random variables (u_j), also independent of (X_k). Consider a random partition of $[0,1]$ into intervals by points

$$Y_k = 1 - \prod_{i=1}^{k} (1-X_i),\ \ k = 1, 2, \ldots. \tag{14}$$

The number of intervals is finite if $\mathbb{P}(X=1)>0$ or infinite otherwise. Group together those players whose individual marks fall in the same *component* $]Y_{k-1},Y_k[$, and maintain the order of groups from the left to the right. This sequential algorithm of random interval division is often referred to as *stick-breaking* or as a *residual allocation model*. Note that in the stick-breaking case the partition of $[0,1]$ has a first (leftmost) interval, a second interval, and so on.

Example 3 (Brownian bridge [32]). Consider the partition of $[0,1]$ by the set of zeros of a Brownian bridge. This set is perfect, i.e. a compact set with no isolated points. Given a uniform sample (u_j) group together all sample points which fall into same excursion interval. This defines a composition structure which is regenerative, by a self-similarity property of the set of zeros. The decrement matrix is described later by formula (46) for $\alpha=\theta=1/2$. Unlike the stick-breaking case there is no leftmost interval.

Example 4 (Brownian motion, meander case [32]). Same as Example 3 but we take the set of zeros of a Brownian motion on $[0,1]$. The collection of intervals is not simply ordered, but there is a definite last (i.e. rightmost) interval, known as the *meander* interval, whose right endpoint is 1. The decrement matrix is described by formula (46) for $\alpha=1/2, \theta=0$.

Example 5 (Myriads of singletons). Fix $\mathtt{d}>0$ and a distribution of X on $]0,1]$. Modify the stick-breaking partition of Example 2 by assuming two types of independent residual allocations. At each odd step the stick is broken with residual measure beta$(1,\mathtt{d}^{-1})$, and at each even step the stick is broken according to X. That is, consider independent random variables $Z_1,X_1,Z_2,X_2,\dots$ with $Z_i\stackrel{d}{=}\text{beta}(1,\mathtt{d}^{-1})$ and $X_i\stackrel{d}{=}X$, and define

$$Y_{2k+1}=1-(1-Z_{k+1})\prod_{j=1}^{k}(1-Z_j)(1-X_j)\,,\qquad Y_{2k}=1-(1-X_k)(1-Z_k)\prod_{j=1}^{k-1}(1-Z_j)(1-X_j).$$

Consider a random closed set $\widetilde{\mathcal{R}}$ which includes endpoints $Y_0:=0$ and 1 and the union of intervals $[Y_{2k},Y_{2k+1}]$, $k=0,1,\dots$. If $\mathbb{P}(X=1)=0$ the interval partition has infinitely many components.

Draw an independent sample of uniform points (u_j) and define a composition by requiring that the sample points which hit components $[Y_{2k},Y_{2k+1}]$ of $\widetilde{\mathcal{R}}$ become singletons, while all those which fall in a particular gap $]Y_{2k+1},Y_{2k+2}[$ are grouped together. For n large, a typical composition of n will start with a *myriad* of singleton parts $1,1,\dots,1$ whose number is of the order of n, followed by one part whose size is of the order of n, followed by a myriad, etc.

For $m>1$ conditioning on the number of sample points out of n which fall into $]Y_1,Y_2[$ leads to a recursion

$$q(n:m)=\binom{n}{m}\mathbb{E}\left((1-Z)^nX^m(1-X)^{n-m}\right)+\mathbb{E}\left((1-Z)^n(1-X)^n\right)q(n:m)$$

which implies q as in (11) but with additional term $n\mathtt{d}$ in the denominator.

The total asymptotic frequency of myriads, say f, is equal to the Lebesgue measure of $\widetilde{\mathcal{R}}$ and satisfies a distributional equation

$$f\stackrel{d}{=}Z_1+(1-Z_1)(1-X_1)f' \tag{15}$$

where f',Z_1,X_1 are independent and $f'\stackrel{d}{=}f$. Analysis of this equation shows that the moments of f are given by a simple formula which we record later in (31).

5 General representation

Background on subordinators and regenerative sets Let $\mathtt{d}\geq 0$ and ν be a measure on $]0,\infty]$ satisfying

$$\int_0^\infty \min(1,z)\,\nu(\mathrm{d}z)<\infty\,. \tag{16}$$

Here and henceforth the integral is over the closed interval $[0,\infty]$. There is no mass at 0 but we allow the case when ν gives a positive mass to $z=\infty$. We also require that either $\mathtt{d}$ or ν be nonzero. Consider a Poisson point process (PPP) on $[0,\infty[\times[0,\infty]$ with intensity measure Lebesgue$\times\nu$. Denoting by (τ_j,Δ_j) the generic atom of the PPP, define the process

$$S_t=\mathtt{d}\,t+\sum_{\tau_j\le t}\Delta_j,\qquad t\ge 0. \tag{17}$$

The process (S_t) is a *subordinator*, that is a Lévy process with increasing càdlàg paths, with $S_0=0$ and $S_t\uparrow\infty$. For $\rho>0$ let $\Phi(\rho)$ be the Laplace exponent of the subordinator defined for $\rho\ge 0$ by

$$\mathbb{E}[\exp(-\rho S_t)]=\exp[-t\Phi(\rho)].$$

That is, according to the Lévy-Khintchine formula,

$$\begin{aligned}
\Phi(\rho) &= \int_0^\infty(1-e^{-\rho z})\nu(\mathrm{d}z)+\rho\mathtt{d} && (18)\\
&= \int_0^1(1-(1-x)^\rho)\,\widetilde{\nu}(\,\mathrm{d}x)+\rho\mathtt{d} && (19)\\
&= \int_0^1\rho(1-x)^{\rho-1}\widetilde{\nu}[x,1]\,\mathrm{d}x+\rho\mathtt{d} && (20)
\end{aligned}$$

where $\nu(\mathrm{d}z)$ is the usual Lévy measure associated with the subordinator, and $\widetilde{\nu}(\mathrm{d}x)$ is the image of ν via the transformation $x=1-e^{-z}$. Let

$$\mathcal{R}=\{S_t,t\ge 0\}^{\mathrm{cl}}$$

be the *closed range* of the subordinator. For a random closed subset $\mathcal{R}$ of $[0,\infty]$ let

$$G(\mathcal{R},t):=\sup\mathcal{R}\cap[0,t]\ \text{and}\ D(\mathcal{R},t):=\inf\mathcal{R}\cap]t,\infty] \tag{21}$$

with the usual conventions $\sup\emptyset=0$ and $\inf\emptyset=\infty$. Following Maisonneuve [29] and Bertoin [6], call $\mathcal{R}$ *regenerative* if for each $t\in[0,\infty[$, conditionally on $\{D(\mathcal{R},t)<\infty\}$, the random set $(\mathcal{R}-D(\mathcal{R},t))\cap[0,\infty]$ is distributed like $\mathcal{R}$ and is independent of $[0\,,\,D(\mathcal{R},t)]\cap\mathcal{R}$. The following representation of regenerative sets is fundamental:

Theorem 5.1 (Maisonneuve [29]) *The closed range $\mathcal{R}$ of a subordinator (S_t) is a regenerative random subset of $[0,\infty]$. Moreover, every regenerative random subset $\mathcal{R}$ of $[0,\infty]$ has the same distribution as the closed range of some subordinator $(S_t,t\ge 0)$, whose Laplace exponent Φ is uniquely determined up to constant multiples.*

Standard exponential sampling We will exploit some known facts about the passage of a subordinator across an independent exponential level, which we recall in the following lemma.

Lemma 5.2 [32] *Let ϵ be an exponential random variable with rate ρ, independent of $\mathcal{R}$ which is the closed range of a subordinator (S_t) with Laplace exponent Φ. Let $G_\epsilon:=G(\mathcal{R},\epsilon)$, $D_\epsilon:=D(\mathcal{R},\epsilon)$, and $\Delta_\epsilon:=D_\epsilon-G_\epsilon$, so that almost surely Δ_ϵ is the length of the interval component of $[0,\infty]\setminus\mathcal{R}$ which covers ϵ, with $\Delta_\epsilon=0$ if $\epsilon\in\mathcal{R}$. The random variables G_ϵ and Δ_ϵ are independent, with Laplace transforms*

$$\mathbb{E}\exp(-s\,G_\epsilon)=\frac{\Phi(\rho)}{\Phi(s+\rho)}\,,\qquad \mathbb{E}\exp\big(-s\,\Delta_\epsilon\big)=\frac{\Phi(s+\rho)-\Phi(s)}{\Phi(\rho)}\,. \tag{22}$$

Note that the second formula in (22) is equivalent to

$$\mathbb{P}(\Delta_\epsilon\in\mathrm{d}z)=\frac{(1-e^{-\rho z})\,\nu(\mathrm{d}z))+\rho\mathtt{d}\,\delta_0(\mathrm{d}z)}{\Phi(\rho)} \tag{23}$$

where δ_0 is a unit mass at 0.

Let (ϵ_j) be a sequence of independent standard exponential variables, independent of (S_t), and let $\epsilon_{1n}, \ldots, \epsilon_{nn}$ be the first n sample points $\epsilon_1, \ldots, \epsilon_n$ arranged in increasing order. Define a partition of the set $\{1, \ldots, n\}$ into blocks of consecutive integers by letting j and $j+1$ belong to different blocks iff the closed interval $[\epsilon_{jn}\,,\,\epsilon_{j+1,n}]$ contains some point of $\mathcal{R}$, for $j < n$. Note in particular that $\{j\}$ is a singleton block if $\epsilon_{jn} \in \mathcal{R}$. Define a composition $\mathcal{C}_n$ of n by the sequence of counts of block-sizes of this random partition of $\{1, \ldots, n\}$ into blocks of consecutive integers, from the left to the right. It is obvious by construction that $(\mathcal{C}_n)$ is a composition structure, call it the *composition structure derived from the subordinator by standard exponential sampling.*

Introduce the binomial moments

$$\Phi(n:m) = \binom{n}{m} \int_0^\infty (1-e^{-z})^m \, e^{-(n-m)z} \, \nu(\mathrm{d}z) + n\mathtt{d}\, 1(m=1) \tag{24}$$

$$= \binom{n}{m} \int_0^1 x^m (1-x)^{n-m} \, \widetilde{\nu}(\,\mathrm{d}x) + \, n\mathtt{d}\, 1(m=1) \tag{25}$$

for $\widetilde{\nu}(\mathrm{d}x)$ the image of $\nu(\mathrm{d}z)$ via $x = 1-e^{-z}$, as in (18)-(19). Note by (16) that the integrals are finite for $1 \le m \le n$, and that these quantities are linearly related to the Laplace exponent Φ by the elementary identities

$$\Phi(n) = \sum_{m=1}^{n} \Phi(n:m)\,, \qquad n = 1, 2, \ldots \tag{26}$$

$$\Phi(n:m) = \binom{n}{m} \sum_{j=0}^{m} (-1)^{j+1} \binom{m}{j} \Phi(n-m+j)\,, \quad 1 \le m \le n \tag{27}$$

where $\Phi(0) = 0$.

Theorem 5.3 (i) *The composition structure derived from a subordinator by standard exponential sampling is regenerative, with decrement matrix*

$$q(n:m) = \frac{\Phi(n:m)}{\Phi(n)}\,. \tag{28}$$

(ii) *Every regenerative composition structure can be so derived from some subordinator.*

(iii) *The Lévy data* $(\mathtt{d}, \nu)$ *of the subordinator is determined uniquely up to a positive factor by the regenerative composition structure.*

Proof of (i). The regenerative property of the composition structure derived from a subordinator follows easily from the memoryless property of exponential distribution and the regenerative property of $\mathcal{R}$ at time $D_{1n} := D(\mathcal{R}, \epsilon_{1n})$. To derive (28), observe that ϵ_{1n} is exponential with rate n and, by the construction,

$$q(n:m) = \mathbb{P}(D_{1n} \in [\epsilon_{mn}\,,\,\epsilon_{m+1,n}])$$

(with the convention $\epsilon_{n+1,n} = \infty$). Let $G_{1n} := G(\mathcal{R}, \epsilon_{1n})$ and $\Delta_{1n} := D_{1n} - G_{1n}$. By Lemma 5.2, Δ_{1n} has distribution (23) for $\rho = n$. Moreover, given $\Delta_{1n} = z$ with $z > 0$, the random variable $\epsilon_{1n} - G_{1n}$ is distributed like exponential variable $\epsilon(n)$ with rate n conditioned on $\epsilon(n) < z$. So the probability that ϵ_{1n} hits the closed range $\mathcal{R}$ of the subordinator (causing a singleton) is

$$\mathbb{P}(D_{1n} = \epsilon_{1n}) = \mathbb{P}(\Delta_{1n} = 0) = \frac{n\mathtt{d}}{\Phi(n)} \tag{29}$$

and given the complementary event that ϵ_{1n} misses $\mathcal{R}$, with $\epsilon_{1n} - G_{1n} = x > 0$ and $\Delta_{1n} = z > x$, the conditional probability that $D_{1n} \in [\epsilon_{mn}\,,\, \epsilon_{m+1,n}]$ equals

$$\binom{n-1}{m-1}(1-e^{-(z-x)})^{m-1}e^{-(z-x)(n-m)}.$$

So the probability that ϵ_{1n} finds a gap in $\mathcal{R}$, and exactly m of the n exponential variables $\epsilon_1, \ldots, \epsilon_n$ fall in that gap, is

$$\frac{1}{\Phi(n)}\int_0^\infty \nu(\mathrm{d}z)\int_0^z ne^{-nx}\,\mathrm{d}x\binom{n-1}{m-1}(1-e^{-(z-x)})^{m-1}e^{-(z-x)(n-m)}$$
$$= \frac{1}{\Phi(n)}\binom{n}{m}\int_0^\infty e^{-(n-m)z}(1-e^{-z})^m\,\nu(\mathrm{d}z)$$

by application of the formula $\int_0^z me^{-mx}(1-e^{z-x})^{m-1}\mathrm{d}x = (1-e^{-z})^m$ which has an immediate interpretation in terms of the order statistics of m independent exponential variables. Now (28) follows because $q(n:m)$ is given by the above formula for $m > 1$ and has the additional term $n\mathtt{d}/\Phi(n)$ from (29) for $m = 1$.

Proofs of (ii). We offer two completely different proofs of (ii): one by probabilistic analysis of random closed sets in Section 7, and one by analytic methods in Section 8.

Proofs of (iii). Probabilistically, (iii) follows from Theorem 5.1 and a general fact about composition structures [17, Corollary 12]. Analytically, (iii) can be read from the results of Section 8. □

If the regenerative composition structure $(\mathcal{C}_n)$ is derived from a subordinator by standard exponential sampling, the associated composition $\mathcal{C}^*$ of the infinite set $\mathbb{N}$ is simply constructed by assigning i and j to different classes iff the closed interval with endpoints ϵ_i and ϵ_j intersects $\mathcal{R}$. The ordering of classes is maintained according to the order of the ϵ_j associated with the classes. The random set of positive integers j whose ϵ_j falls in a particular interval component of $\mathcal{R}^c := [0,\infty]\setminus\mathcal{R}$ forms a *positive* class, while each j whose ϵ_j hits $\mathcal{R}$ forms a singleton class. By the law of large numbers, the probability assigned to an interval component of $\mathcal{R}^c$ by the standard exponential distribution is the *frequency* of the corresponding class of $\mathcal{C}^*$, that is the almost sure limit as $n \to \infty$ of the proportion of elements of $[n]$ which belong to the class. For instance, if $]a,b[\subset \mathcal{R}^c$ is the interval component which covers ϵ_1, then for large n the class of $\mathcal{C}_n^*$ containing element 1 will have approximately $n(e^{-a}-e^{-b})$ elements, so there will be some part of $\mathcal{C}_n$ of this size. We note the following Corollary of Theorem 5.3:

Corollary 5.4 *Let f denote the random frequency of the union of all singleton classes in the exchangeable random partition of $\mathbb{N}$ associated with a regenerative composition structure with decrement matrix* (28). *Then*

$$f = \mathtt{d}\int_0^\infty \exp(-S_t)\,\mathrm{d}t \tag{30}$$

where (S_t) is the associated subordinator with Laplace exponent Φ and $\mathtt{d}$ is the drift coefficient of (S_t), and the distribution of f on $[0,1]$ is determined by the moments

$$\mathbb{E}(f^n) = \frac{n!\,\mathtt{d}^n}{\prod_{i=1}^n \Phi(i)} \qquad (n = 1,2,\ldots). \tag{31}$$

Proof. The derivation from (S_t) by standard exponential sampling gives

$$f = \int_0^\infty e^{-z}\;1(S_t = z \text{ for some } t \geq 0)\,\mathrm{d}z$$

and (30) follows by the change of variable $z = S_t$. This change of variable follows by noting that the function $t \mapsto S_t$ is almost everywhere differentiable with derivative $\mathtt{d}$. Formula (31) can now be read from the work of Carmona, Petit and Yor [9, Prop. 3.3], or derived from (15). □

Extensive discussion of the exponential functional $\int_0^\infty \exp(-S_t)\,\mathrm{d}t$ is found in [7, 22]. See [19] for further applications to regenerative composition structures.

6 Topological preliminaries

This section deals with some preliminaries for an approximation argument in the next section. Let $\mathcal{M}$ be the collection of all subsets $R \subset [0,1]$ which are closed and contain the points 0 and 1. For $R \subset [0,1]$ and $d > 0$ define the *d-inflation* of R to be the union of all open intervals of length $2d$ centered at some $z \in R$. For two elements of $\mathcal{M}$, the *Hausdorff distance* between them is the infinum of all d's such that the d-inflation of either of the two sets covers the other one. With the Hausdorff distance, $\mathcal{M}$ becomes a compact, complete, separable metric space. For $R \in \mathcal{M}$ the complementary open set $R^c = [0,1] \setminus R$ is a union of finitely or countably many disjoint open interval components. We can describe $\mathcal{M}$ by the positions of these interval components: closeness of two sets means their complements have the same number of 'large' interval components and the location of these interval components is approximately the same for both sets. This idea is made precise by the following convergence criterion. The proofs of this and the following two lemmas are straightforward, and left to the reader.

Lemma 6.1 *$R_n \to R$ in $\mathcal{M}$ iff for each $\epsilon > 0$, as $n \to \infty$ the following two conditions are satisfied:*

(i) *for all sufficiently large n, the number of interval components of R_n^c larger than ϵ equals the number of interval components of R^c larger than ϵ, say $k(\epsilon)$,*

(ii) *the vector of $2k(\epsilon)$ endpoints of these components of R_n^c, recorded in increasing order, converges to the analogous vector for R^c.*

For $R \in \mathcal{M}$ and $z \in [0,1[$ let $D(R,z) := \inf R \cap]z,1]$. We extend the domain of definition of $D(R,\cdot)$ by setting $D(R,z) = 0$ for $z \le 0$ and $D(R,z) = 1$ for $z \ge 1$. The function $D(R,\cdot)$ carries complete information about R and is a useful tool to describe many functionals on $\mathcal{M}$.

Lemma 6.2 *For $R \in \mathcal{M}$, the function $D(z) := D(R,z)$ on $[0,1[$ has the following two properties:*

(i) *$D(z) = z$ iff $z \in R$ and z is not isolated in R on the right,*

(ii) *For $]a,b[$ an open interval component of R^c, we have $D(z) \equiv b$ for all $z \in [a,b[$.*

These properties determine the function $D(\cdot)$ uniquely and further imply that

(iii) *$D(z)$ is a nondecreasing càdlàg function satisfying $D(z) \ge z$,*

(iv) *$D(1-) = 1$ and $D(a-) = a$ for each interval component $]a,b[$ of R^c.*

On each interval component of R^c the function $D(R,\cdot)$ is constant, and it coincides with the identity almost everywhere on R. It follows that the correspondence $R \mapsto D(R,\cdot)$ is injective, because R^c can be uniquely recovered as the union of open 'flats' of $D(R,\cdot)$. On the other hand, (iii) and (iv) imply that $D(R,\cdot)$ is a distribution function of a probability measure on $[0,1]$ associated with R. This distribution is constructed from Lebesgue measure by sweeping out the total mass $b-a$ of each interval component $]a,b[\subset R^c$ to the point a. These representations of closed sets as functions or measures provide alternative descriptions of the topology on $\mathcal{M}$.

Lemma 6.3 *Three topologies on $\mathcal{M}$ coincide:*

(i) *the topology induced by Hausdorff metric on $\mathcal{M}$,*

(ii) *Skorokhod topology restricted to the space of càdlàg functions $D(R,\cdot), R \in \mathcal{M}$,*

(iii) *the weak topology restricted to the space of probability distributions associated with $R \in \mathcal{M}$.*

Suppose $R \in \mathcal{M}$ and $z \in\ [0,1[$ are such that $R \cap\ ]z,1[\neq \emptyset$. Then $D(R,z) < 1$ and we can define another closed set

$$R(z) := \left\{ \frac{y - D(R,z)}{1 - D(R,z)} : y \in R \cap [D(R,z), 1] \right\} \tag{32}$$

which is the part of R strictly to the right of $D(R,z)$, scaled back to $[0,1]$.

Lemma 6.4 *Suppose $R \in \mathcal{M}$ and $R \cap]z,1[\neq \emptyset$ for a given $z \in [0,1[$. Let (z_n) and (R_n) be sequences of elements of $]z,1[$ and $\mathcal{M}$, respectively. If either*

(i) *$z_n \to z$, $R_n \to R$ and z is not a point of R isolated on the right, or*

(ii) *$z_n \downarrow z$ and $R_n = R$ for all n,*

then

$$D(R_n, z_n) \to D(R,z), \quad R_n \cap [0,\, D(R_n, z_n)] \to R \cap [0,\, D(R,z)] \ \text{ and } \ R_n(z_n) \to R(z).$$

(We will formally adjoin point 1 to such intersections in order to treat them as elements of $\mathcal{M}$.)

Proof. If (i) holds and $z \notin R$ then there is an interval component of R^c (respectively R_n^c) which covers z and has $D(R,z)$ (respectively $D(R_n,z)$) as the right endpoint. Taking ϵ in Lemma 6.1 small enough we see that $D(R_n,z_n) = D(R_n,z) \to D(R,z)$ and thus $R_n(z_n) \to R(z)$. When $z \in R$ but z is not isolated in R on the right, $D(R,z)$ coincides with z and for any $z' \in R \cap]z,1]$ we have $D(R_n,z_n) \in [z,z']$ for n large, therefore by letting $z' \downarrow z$ we still get $D(R_n,z_n) \to D(R,z)$, and again the argument can be completed by applying Lemma 6.1. In case (ii) the assertion follows along same lines by the right-continuity of $D(R,\cdot)$.
□

Example. In case $z \in R$ is isolated on the right (that is, when z is a left endpoint of some interval component of R^c) the convergence can fail. For instance, taking $R = \{0, 1/3, 2/3, 1\}$ and $R_n = \{0, 1/3 - 1/n, 1/3 + 2/n, 2/3, 1\}$ we have $R_n \to R$ and observe that $R_n(1/3 + 2/n) \to R(1/3) = \{0,1\}$ as it follows from $D(R_n, 1/3 + 2/n) = D(R, 1/3) = 2/3$, but for other z the limit could be different, e.g. $R_n(1/3 - 1/n) = R_n(1/3) = R_n(1/3 + 1/n) \to \{0, 1/2, 1\}$ in accord with $D(R_n, 1/3 - 1/n) = D(R_n, 1/3) = D(R_n, 1/3 + 1/n) = 1/3 + 2/n \to 1/3$.

Since $\mathcal{M}$ is a metric space we can speak of random closed sets and of their convergence almost surely or in distribution, which we denote $\stackrel{a.s.}{\to}$ and $\stackrel{d}{\to}$ respectively. For a random closed set $\widetilde{\mathcal{R}}$ in $\mathcal{M}$ the function $D(\widetilde{\mathcal{R}}, z)$ becomes a random process with parameter z, with càdlàg paths.

Definition 6.5 For a random set $\widetilde{\mathcal{R}} \in \mathcal{M}$ and $z \in\ [0,1[$ we say that z is singular if

$$\mathbb{P}(z \in \widetilde{\mathcal{R}} \ \text{ and } \ D(\widetilde{\mathcal{R}}, z) > z) > 0.$$

Lemma 6.6 *For any random set $\widetilde{\mathcal{R}} \in \mathcal{M}$ the collection of singular points is at most countable.*

Proof. Consider the product of the basic probability space and $[0,1]$ with Lebesgue measure. For each singular z define $U_z := \bigcup (\{\omega\} \times [z, D(\widetilde{\mathcal{R}}, z)[\,)$ where the union is over all elementary events ω such that z is isolated on the right. Each U_z has a positive measure, and for distinct z's the U_z's must be disjoint.
□

7 Multiplicatively regenerative sets

By mapping $[0,\infty]$ onto $[0,1]$ via $z \mapsto 1 - e^{-z}$ we transform a subordinator (S_t) into a *multiplicative subordinator* $\widetilde{S}_t := 1 - \exp(-S_t)$: for $t' > t$ the ratio $(1 - \widetilde{S}_{t'})/(1 - \widetilde{S}_t)$ has same distribution as $1 - \widetilde{S}_{t'-t}$

and is independent of $(S_u, 0 \le u \le t)$. This construction appears also in [11], [15], [22]. The counterpart of (17) is

$$\widetilde{S}_t = 1 - e^{-\mathtt{d}t} \prod_{\tau_j \le t} (1 - \widetilde{\Delta}_j)$$

where $\widetilde{\Delta}_j = 1 - \exp(-\Delta_j)$ and the product is over the atoms $(\tau_j, \widetilde{\Delta}_j)$ of a PPP in the strip $[0, \infty[\times [0,1]$, with intensity measure Lebesgue$\times \widetilde{\nu}$ where $\widetilde{\nu}$ is the image of the measure ν via $z \mapsto 1 - e^{-z}$. Note that the mapping preserves order, so that $(\widetilde{S}_t)$ increases from 0 to 1.

Let $\widetilde{\mathcal{R}} := 1 - \exp(-\mathcal{R})$ be the range of the multiplicative subordinator. The transformation $z \mapsto 1 - e^{-z}$ takes an exponential sample (ϵ_j) into a uniform sample (u_j). The regenerative composition structure $(\mathcal{C}_n)$ derived from the subordinator (S_t) by exponential sampling can now be described as follows: $\mathcal{C}_n$ is induced by separating the first n uniform variables u_j by the points of $\widetilde{\mathcal{R}}$. Note that the frequencies of positive classes derived from $(\mathcal{C}_n)$ now coincide with the lengths of open interval components of $\widetilde{\mathcal{R}}^c = [0,1] \setminus \widetilde{\mathcal{R}}$, and remaining frequency of singletons f, as in Corollary 5.4, is the Lebesgue measure of $\widetilde{\mathcal{R}}$. The following lemma is easily checked:

Lemma 7.1 *For random closed sets $\widetilde{\mathcal{R}} \in \mathcal{M}$ and $\mathcal{R} \subset [0, \infty]$ related by $\widetilde{\mathcal{R}} = 1 - \exp(-\mathcal{R})$, the random set $\mathcal{R}$ is regenerative iff $\widetilde{\mathcal{R}}$ is multiplicatively regenerative according to the following definition.*

Definition 7.2 A random closed set $\widetilde{\mathcal{R}} \in \mathcal{M}$ is called *multiplicatively regenerative* if, for each $z \in [0,1[$, conditionally on $\{D(\widetilde{\mathcal{R}}, z) < 1\}$ the random set $\widetilde{\mathcal{R}}(z)$, defined as in (32), is independent of $[0\,,\, D(\widetilde{\mathcal{R}}, z)] \cap \widetilde{\mathcal{R}}$, and has the same distribution as $\widetilde{\mathcal{R}}$.

It is convenient to represent this multiplicative regeneration property in terms of the following operation on closed sets. For $R \in \mathcal{M}$ let $G(R,1) = \sup R \cap [0,1[$. For $R_1, R_2 \in \mathcal{M}$ we construct a new set $R_1 \rhd R_2$ by scaling R_2 and fitting it into $[G(R_1\,,1), 1]$:

$$R_1 \rhd R_2 := R_1 \cup \{G + (1-G)z : \, z \in R_2\}, \quad \text{where } G = G(R_1\,,1).$$

Let $\widetilde{\mathcal{R}}'$ denote an independent replica of $\widetilde{\mathcal{R}}$. Then $\widetilde{\mathcal{R}}$ is multiplicatively regenerative iff for each $z \in]0,1[$

$$\widetilde{\mathcal{R}} \stackrel{d}{=} (\widetilde{\mathcal{R}} \cap [0\,,\, D(\widetilde{\mathcal{R}}, z)]) \rhd \widetilde{\mathcal{R}}'. \tag{33}$$

To proceed we need to adapt a law of large numbers from [17]. We associate each composition $(n_1, \ldots, n_k)$ of n with the finite closed set whose points are partial sums of the parts of $n_1, \ldots, n_k$ divided by n; e.g. the composition $(4,2,3,1)$ of 10 is associated with the set $\{0,\ 0.4,\ 0.6,\ 0.9,\ 1\}$. Thus a composition structure $(\mathcal{C}_n)$ is associated with a sequence of random sets $(\widetilde{\mathcal{R}}_n)$.

Lemma 7.3 [17] *Let $(\mathcal{C}_n)$ be a composition structure and let $(\widetilde{\mathcal{R}}_n)$ be the associated sequence of random elements of $\mathcal{M}$. Then $\widetilde{\mathcal{R}}_n \stackrel{a.s.}{\to} \widetilde{\mathcal{R}}$, for some $\widetilde{\mathcal{R}} \in \mathcal{M}$, and $(\mathcal{C}_n)$ is distributed as if by using $\widetilde{\mathcal{R}}$ to separate the points in a random sample of uniform $[0,1]$ variables independent of $\widetilde{\mathcal{R}}$.*

The proof of Theorem 5.3 is completed by Lemma 7.1 and the following lemma:

Lemma 7.4 *If the composition structure $(\mathcal{C}_n)$ is regenerative, then $\widetilde{\mathcal{R}}$ is multiplicatively regenerative.*

Proof. Since $\mathcal{C}_n$ may be constructed from $\widetilde{\mathcal{R}}$ and independent uniform r.v.'s (u_j), the associated random set $\widetilde{\mathcal{R}}_n$ becomes a 'paintbox' function of this data, say $\widetilde{\mathcal{R}}_n = \mathrm{PB}(n, \widetilde{\mathcal{R}}, (u_j))$. For each fixed $z \in]0,1[$, the truncated set $(\widetilde{\mathcal{R}}_n \cap [0\,,\, D(\widetilde{\mathcal{R}}_n, z)]) \cup \{1\}$ is another measurable function of same data, say $\mathrm{PB}_z(n, \widetilde{\mathcal{R}}, (u_j))$. Without loss of generality, we may assume that the underlying probability space supports the data $(\widetilde{\mathcal{R}}, (u_j))$ and an independent copy $(\widetilde{\mathcal{R}}', (u'_j))$. This allows us to couple these two representations as in the following argument.

Because $(\mathcal{C}_n)$ is regenerative, each $\widetilde{\mathcal{R}}_n$ is a transformed random path of the Markov chain Q_n, that is $\widetilde{\mathcal{R}}_n = \{1-n^{-1}Q_n(j),\ j=0,1,\ldots\}$. The strong Markov property and time-homogeneity of Q_n imply that for each stopping time τ, given $Q_n(\tau)=m$ the sequence $(Q_n(\tau+j), j=0,1,\ldots)$ has the same distribution as $(Q_m(j), j=0,1,\ldots)$, and is conditionally independent of τ and $(Q_n(j), j=0,1,\ldots,\tau-1)$. Considering the particular stopping time $\tau=\min\{j : Q_n(j) < n(1-z)\}$, we reformulate this Markov property as a distributional identity

$$\widetilde{\mathcal{R}}_n \stackrel{d}{=} \widetilde{\mathcal{R}}^*_n := \mathrm{PB}_z(n,\widetilde{\mathcal{R}},(u_j)) \rhd \mathrm{PB}(\mu_n,\widetilde{\mathcal{R}}',(u'_j)) \quad \text{where} \;\; \mu_n = n(1-D(\widetilde{\mathcal{R}}_n,z)) \tag{34}$$

with the convention that $\widetilde{\mathcal{R}}^*_n = \widetilde{\mathcal{R}}_n$ on $\{\mu_n=0\}$. The idea is to analyse the asymptotics of $\widetilde{\mathcal{R}}^*_n$.

Suppose first that z is not a singular point for $\widetilde{\mathcal{R}}$ in the sense of Definition 6.5. Then, by Lemma 6.4 (i), $D(\widetilde{\mathcal{R}}_n,z) \stackrel{a.s.}{\to} D(\widetilde{\mathcal{R}},z)$ and by the strong law of large numbers $\mu_n \sim n(1-D(\widetilde{\mathcal{R}}_n,z)) \to \infty$ on $\{D(\widetilde{\mathcal{R}},z)<1\}$, hence by Lemma 7.3 we have $\mathrm{PB}(\mu_n,\widetilde{\mathcal{R}}',(u'_j)) \stackrel{a.s.}{\to} \widetilde{\mathcal{R}}'$ almost everywhere on $\{D(\widetilde{\mathcal{R}},z)<1\}$. For a similar reason, $\mathrm{PB}_z(n,\widetilde{\mathcal{R}},(u_j)) \stackrel{a.s.}{\to} \widetilde{\mathcal{R}} \cap [0\,,\,D(\widetilde{\mathcal{R}},z)]$. It follows that $\widetilde{\mathcal{R}}^*_n$ has an almost sure limit, say $\widetilde{\mathcal{R}}^*$, and hence that

$$\widetilde{\mathcal{R}} \stackrel{d}{=} \widetilde{\mathcal{R}}^* \stackrel{d}{=} (\widetilde{\mathcal{R}} \cap [0\,,\,D(\widetilde{\mathcal{R}},z)]) \rhd \widetilde{\mathcal{R}}'$$

which is the desired decomposition (33).

In case of singular z, we still have $\mu_n \to \infty$ thus $\widetilde{\mathcal{R}}(z) \stackrel{d}{=} \widetilde{\mathcal{R}}$, as above. To show the independence choose a nonsingular point $z'>z$, then by the above $\widetilde{\mathcal{R}}(z')$ is independent of $\widetilde{\mathcal{R}} \cap [0\,,\,D(\widetilde{\mathcal{R}},z')]$ (given $\{D(\widetilde{\mathcal{R}},z')<1\}$) and because $D(\widetilde{\mathcal{R}},z') \geq D(\widetilde{\mathcal{R}},z)$ it is also independent of $\widetilde{\mathcal{R}} \cap [0\,,\,D(\widetilde{\mathcal{R}},z)]$. By Lemma 6.6 we can let nonsingular z' approach z from the right, and by Lemma 6.4 (ii) we have then $\widetilde{\mathcal{R}}(z') \to \widetilde{\mathcal{R}}(z)$, which implies readily the required independence for $\widetilde{\mathcal{R}}(z)$. $\square$

Remarks

Singular points We note that these cannot appear when $\widetilde{\nu}[0,1]=\infty$, by the well known result of Kesten [26], [5, Theorem III.2.4] that in this case $\mathbb{P}(x \in \widetilde{\mathcal{R}})=0$ for each $x \in]0,1[$. But if $\widetilde{\nu}[0,1]<\infty$, singular points appear when $\mathtt{d}=0$ and $\widetilde{\nu}$ has atoms, as in Example 1.

Discussion of cases In the case of a finite measure $\widetilde{\nu}$ and $\mathtt{d}=0$ the range $\widetilde{\mathcal{R}}$ is an increasing sequence obtained by stick-breaking with i.i.d. factors. In this case 1 is the only possible accumulation point for $\widetilde{\mathcal{R}}$, and 1 is always such a point when $\widetilde{\nu}\{1\}=0$ and thus $\widetilde{\mathcal{R}}$ has infinitely many points. By analogy, the more complex case $\widetilde{\nu}[0,1]=\infty$ can be interpreted as a generalised stick-breaking procedure embedded in continuous time, so that infinitely many breaks are performed within any arbitrarily small period (prior to possible absorbtion), in this case the compact set $\widetilde{\mathcal{R}} \subset [0,1]$ is perfect (except that 1 is an isolated point in case $\widetilde{\nu}\{1\}>0$). Note that the Lebesgue measure of $\widetilde{\mathcal{R}}$, which is positive only in case $\mathtt{d}>0$, equals the total frequency of all singleton classes of the associated exchangeable random partition of $\mathbb{N}$.

Topology The space of all composition structures is an infinite dimensional Choquet simplex with the product topology (of pointwise convergence of $p(n_1,\ldots,n_k)$ for all positive integer compositions $(n_1,\ldots,n_k)$). Regenerative compositions constitute a smaller compact with the relative topology neatly described in terms of the Lévy data, as follows. Normalising the parameters so that $\mathtt{d}+\int_0^\infty x\,\widetilde{\nu}(\mathrm{d}x)=1$, the convergence of regenerative composition structures corresponds to convergence of $\mathtt{d}$'s and weak convergence of $\widetilde{\nu}$'s away from 0 (vague convergence on $]0,1]$). The set of stick-breaking composition structures is dense in the space of all regenerative composition structures.

Two kinds of regeneration According to Lemma 7.1, $\mathcal{R}$ is regenerative iff $\widetilde{\mathcal{R}} := 1-\exp(-\mathcal{R})$ is multiplicatively regenerative. In a special case, the restriction $\widehat{\mathcal{R}} := (\mathcal{R} \cap [0,1[\,) \cup \{1\}$ is also a multiplicatively regenerative set, of course different from $\widetilde{\mathcal{R}}$. This occurs if $\mathcal{R}$ has the self-similarity property $c\,(\mathcal{R} \cap [0,1[\,) \stackrel{d}{=} \mathcal{R} \cap [0,c[$ for $0<c<1$. That is to say, if $\mathcal{R}=[0,\infty]$ or $\mathcal{R}$ has no points in $]0,1[$, or $\mathcal{R}$ is the range of a stable subordinator of index α for some $\alpha \in]0,1[$. In that case $\widehat{\mathcal{R}} = 1 - \exp(-\mathcal{R}')$ where $\mathcal{R}'$ is the range of a killed subordinator which is described explicitly in Section 10.3.

A sufficient condition for regeneration We note that in the usual definition of a regenerative random subset $\mathcal{R}$ of $[0,\infty]$, as in Section 5, the independence of the two random sets $\mathcal{R}_t := (\mathcal{R} - D(\mathcal{R},t)) \cap [0,\infty]$ and $[0\,,\,D(\mathcal{R},t)] \cap \mathcal{R}$ for all t can be replaced by the apparently weaker condition of independence of the random set $\mathcal{R}_t$ and the random variable $D(\mathcal{R},t)$ for all t. This is due to the following consequence of Proposition 3.2, and Theorem 5.3, which does not seem easy to prove by other methods:

Corollary 7.5 *Let $\mathcal{R}$ be a random closed subset of $[0,\infty]$, let ϵ be an exponential random variable with rate 1 independent of $\mathcal{R}$, and let $\mathcal{R}_\epsilon := (\mathcal{R} - D(\mathcal{R},\epsilon)) \cap [0,\infty]$. If $\mathcal{R}_\epsilon \overset{d}{=} \mathcal{R}$ and $\mathcal{R}_\epsilon$ is independent of $D(\mathcal{R},\epsilon)$ then $\mathcal{R}$ is regenerative.*

A similar condition involving a single independent uniform variable can be given for multiplicative regeneration of a random closed subset $\widetilde{\mathcal{R}}$ of $[0,1]$.

8 Reduction to a moment problem

This section offers a completely different approach to the basic integral representation of decrement matrices of regenerative composition structures implied by Theorem 5.3, meaning the formula

$$q(n:m) = \frac{\Phi(n:m)}{\Phi(n)} \tag{35}$$

where

$$\Phi(n:m) = \binom{n}{m} \sum_{j=0}^{m} (-1)^{j+1} \binom{m}{j} \Phi(n-m+j)\,, \quad 1 \le m \le n \tag{36}$$

with

$$\Phi(n) = \int_0^1 (1-(1-x)^n)\,\tilde{\nu}(\mathrm{d}x) + n\,\mathtt{d}\,. \tag{37}$$

for some measure $\tilde{\nu}(\mathrm{d}x)$ on $]0,1]$ and $\mathtt{d} \ge 0$. This approach, more algebraic and analytic than probabilistic, is by a direct analysis of the recursion (10) using the substitution (35), where $\Phi(n:m)$ and $\Phi(n)$ are treated as real variables related by (36) with $\Phi(0) = 0$. Note that then

$$\Phi(n) = \Phi(n:1) + \cdots + \Phi(n:m) \tag{38}$$

as recorded earlier in (26). This leads to a recursion which is solved by the known integral representation (37) of *completely alternating sequences* [4], that is sequences Φ such that $\Phi(n:m)$ defined by (36) is non-negative for all n and m. For completeness, we offer an elementary derivation of the integral representation.

Lemma 8.1 *Suppose that a sequence of numbers $\Phi(n), n = 0,2,\ldots$ with $\Phi(0) = 0$ and $\Phi(n) > 0$ for $n \ge 0$ is such that each entry $\Phi(n:m)$ of the matrix* (36) *is non-negative. Then the matrix* (35) *is the decrement matrix of some regenerative composition structure.*

Proof. Observe that (36) implies the recursion

$$\Phi(n:m) = \frac{m+1}{n+1}\Phi(n+1:m+1) + \frac{n-m+1}{n+1}\Phi(n+1:m), \qquad 1 \le m \le n < \infty \tag{39}$$

Dividing this identity by $\Phi(n+1)$, and substituting it in the to-be-checked (10), we transform it by elementary algebra to

$$\Phi(n+1:1) = (n+1)(\Phi(n+1) - \Phi(n))$$

which is true as a special case of (36). $\square$

Lemma 8.2 *The decrement matrix of a regenerative composition structure can be represented in the form* (35)*, by a matrix* $(\Phi(n:m), 1 \le m \le n < \infty)$ *with non-negative entries satisfying* (39) *and* (38)*. The matrix* Φ *is determined by* q *uniquely up to a positive factor.*

Proof. The statement is only nontrivial when $0 < p(n) < 1, n \ge 2$. So let us consider a decrement matrix with entries $0 < q(n:m) < 1$. Fix n and set by definition $\Phi(n:m) := q(n:m)$ for $m = 1, \dots, n$. Consider the unique solution $(\Phi(j:m), 1 \le m \le j \le n)$ to (39) with the values $q(n:m)$ at level n. Because $q(n:m) > 0$, it is easily seen that $\Phi(j:m) > 0$ for $1 \le m \le j \le n$ and therefore $\Phi(j) := \Phi(j:1) + \cdots + \Phi(j:j) > 0$ for $j < n$ (and $\Phi(n) = 1$). By Lemma 8.1 the elements $\Phi(j:m)/\Phi(j)$ satisfy the recursion (10) for $j < n$ and for $j = n$ they coincide with $q(n:m)$. Thus by the uniqueness of solutions to (10) for $j < n$ with given values at level n we conclude that $q(j:m)$ coincides with $\Phi(j:m)/\Phi(j)$ for all $1 \le m \le j \le n$.

Keeping n fixed, suppose there is another representation $q(j:m) = \Phi'(j:m)/\Phi'(n)$ $j \le n$, then $\Phi'(n:m) = \Phi'(n)q(n:m)$, thus arguing as above and using linearity we get $\Phi'(j:m) = \Phi'(n)\Phi(j:m)$ for $1 \le m \le j \le n$. Thus the representation for given n is unique up to a multiple, and it becomes unique subject to a normalisation constraint.

Assuming the normalisation $\Phi(1:1) = 1$, the finite matrices $(\Phi(j:n), 1 \le m \le j \le n)$ constructed for each n are consistent as n varies, by the uniqueness for each particular n, thus they constitute an infinite matrix and the desired representation follows. □

Our main characterisation of regenerative composition structures and the uniqueness (parts (ii) and (iii) of Theorem 5.3) are implied by Lemma 8.2 and the following known result:

Proposition 8.3 [4, Proposition 6.12 for $k = 1$, p. 134] *A sequence* $\Phi(n), n = 0, 1, 2, \dots$ *with* $\Phi(0) = 0$ *and* $\Phi(n) > 0$ *for* $n > 0$ *is such that all entries* $\Phi(n:m)$ *defined by* (36) *are non-negative if and only if there is the integral representation* (37) *for some measure* $\widetilde{\nu}$ *on* $]0,1]$ *and* $\mathtt{d} \ge 0$*. Moreover* $\widetilde{\nu}$ *and* $\mathtt{d}$ *are uniquely determined by* Φ*.*

Proof. Define a matrix ϕ in terms of Φ by

$$\binom{n}{m}\phi(n:m) = \Phi(n:m), \quad 1 \le m \le n < \infty. \tag{40}$$

Then ϕ satisfies the recursion

$$\phi(n:m) = \phi(n+1:m+1) + \phi(n+1:m), \quad \text{for } \ 1 \le m \le n\,.$$

Let $\mu(i,j) := \phi(i+j+1:j+1)$. Then the recursion becomes

$$\mu(i,j) = \mu(i+1,j) + \mu(i,j+1) \qquad \text{for } \ i \ge 0,\, j \ge 0.$$

It is a known consequence of de Finetti's representation of infinite exchangeable sequences of 0's and 1's, or the Hausdorff moment problem (see e.g. [14], [4], [25], [33]), that each non-negative solution of this recursion can be represented as $\mu(i,j) = \int_0^1 x^i(1-x)^j\,\tilde{\eta}(\mathrm{d}x)$ for a unique finite measure $\tilde{\eta}$ on $[0,1]$. That is,

$$\phi(n:m) = \int_{]0,1]} x^{m-1}(1-x)^{n-m}\,\tilde{\eta}(\mathrm{d}x) + \tilde{\eta}(\{0\})\ 1(m=1),$$

and the corresponding solution to (39) is read from (40). To pass to the form (25) denote $\mathtt{d} = \tilde{\eta}(\{0\})$ and introduce another measure $\widetilde{\nu}$ on $]0,1]$ which is absolutely continuous with respect to $\tilde{\eta}$ with density $\tilde{\nu}(\mathrm{d}x) = x^{-1}\,\tilde{\eta}(\mathrm{d}x)$ for $x \in]0,1]$. In general, $\tilde{\nu}$ need not be a finite measure, but it has finite mean because $\tilde{\eta}$ is finite. Finally (37) can be read from (25) and (38). □

We note in passing a slightly different formulation of Proposition 8.3. For Φ as in (37) with $\Phi(1) = 1$, integration by parts as in (20) gives

$$\Phi(n+1)/(n+1) = \int_0^1 u^n\,\tilde{\nu}[1-u,1]\,\mathrm{d}u + \mathtt{d} = \mathbb{E}(X^n), \qquad n \ge 0$$

where X is a random variable with increasing density $\tilde{\nu}[1-u,1]$ for $0<u<1$ and $\mathbb{P}(X=1)=\mathtt{d}$. Then (26) can be rewritten as

$$\Phi(n+1:n+1-j)=(n+1)(\mu_{n,j}-\mu_{n,j-1}), \qquad 0\le j\le n$$

where $\mu_{n,j}:=\binom{n}{j}\mathbb{E}(X^j(1-X)^{n-j})$ for $0\le j\le n$ and $\mu_{n,-1}:=0$. So Proposition 8.3 can be reformulated as follows:

Corollary 8.4 (Diaconis and Freedman [10, Theorem 10]) *Given a sequence of real numbers μ_n with $\mu_0=1$, let*

$$\mu_{n,j}:=\binom{n}{j}\sum_{i=0}^{n-j}(-1)^i\binom{n-j}{i}\mu_{j+i}.$$

There exists a probability distribution of X on $[0,1]$ such that $\mathbb{E}(X^n)=\mu_n$ and the distribution of X is absolutely continuous with an increasing density on $[0,1[$, allowing the possibility that $\mathbb{P}(X=1)>0$, if and only if $\mu_{n,0}\ge 0$ for all $n\ge 0$ and $\mu_{n,j}-\mu_{n,j-1}\ge 0$ for all $1\le j\le n$. The distribution of X is then unique, with $\mu_{n,j}=\binom{n}{j}\mathbb{E}[X^j(1-X)^{n-j}]$.

9 Parametrisation of decrement matrices

The representation $q(n:m)=\Phi(n:m)/\Phi(n)$ provides one parametrisation of the regenerative composition structures in terms of a sequence $\Phi(n), n\ge 1$. To be probabilistically meaningful, this must be the sequence of evaluations of some Laplace exponent at positive integer values. But we may also regard the expressions for $q(n:m)$ as a collection of rational functions in variables $\Phi(n), n\ge 1$. This section presents some alternative parameterisations of regenerative composition structures, and discusses their probabilistic and algebraic relations to each other.

9.1 Structural moments

One meaningful collection of parameters is the sequence of diagonal entries

$$p(n)=q(n:n)$$

which starts with $p(1)=1$. We call these diagonal entries of the decrement matrix the *structural moments* of composition structure, as they coincide with moments of the *structural distribution* Σ:

$$p(n)=\int_0^1 x^{n-1}\,\Sigma(\mathrm{d}x)$$

where Σ is the distribution of the length of the interval component of $\widetilde{\mathcal{R}}^c$ containing a given uniform sample point, say u_1. This random length is the frequency of the class of $\mathcal{C}^*$ containing element 1, that is a size-biased pick from the collection of frequencies [34]. Note from (29) and (19), or from Corollary 5.4, that the expectation of the total frequency of singletons $f=\mathrm{Lebesgue}(\widetilde{\mathcal{R}})$ is the measure assigned by Σ to 0:

$$\mathbb{E}(f)=\Sigma(\{0\})=\mathtt{d}/\Phi(1)=\mathtt{d}/\left(\mathtt{d}+\textstyle\int_0^1 x\,\widetilde{\nu}(\mathrm{d}x)\right).$$

From $p(n)=\Phi(n:n)/\Phi(n)$, by expanding the numerator by (27) we obtain a relation

$$\Phi(n)(p(n)+(-1)^n)=\sum_{j=1}^{n-1}(-1)^{n+1}\binom{n}{j}\Phi(j)\,, \tag{41}$$

which may be seen as a recursion for $\Phi(n), n=1,2,\ldots$. Assuming the initial value $\Phi(1)=1$ the recursion has a unique solution, which is necessarily positive by Lemma 8.2. Thus the recursion (41) allows q to be recovered from $p(n), n=1,2,\ldots$, by first recursively computing $\Phi(n), n=1,2,\ldots$, then $\Phi(n:m)$ from (27) and finally using (35). Thus we have proved

Proposition 9.1 *A regenerative composition structure is uniquely determined by the structural moments $p(n) = q(n:n)$ for $n = 1, 2, \ldots$. Each $q(n:m)$ for $1 \le m \le n$ is expressible as a rational function in the variables $p(1) = 1, p(2), \ldots, p(n)$.*

To illustrate the result, the first few entries are

$$\begin{aligned}
q(2:1) &= 1 - p(2) \\
q(3:1) &= \frac{1 - 3p(2) + 2p(3)}{1 - p(2)} \\
q(3:2) &= \frac{2p(2) - 3p(3) + p(2)p(3)}{1 - p(2)} \\
q(4:1) &= \frac{1 - 5p(2) + 8p(3) - 4p(2)p(3) - 3p(4) + 3p(2)p(4)}{1 - 2p(2) + 2p(3) - p(2)p(3)} \\
q(4:2) &= \frac{3p(2) - 9p(3) + 6p(2)p(3) + 6p(4) - 9p(2)p(4) + 3p(3)p(4)}{1 - 2p(2) + 2p(3) - p(2)p(3)} \\
q(4:3) &= \frac{3p(3) - 3p(2)p(3) - 4p(4) + 8p(2)p(4) - 5p(3)p(4) + p(2)p(3)p(4)}{1 - 2p(2) + 2p(3) - p(2)p(3)}
\end{aligned}$$

The complexity of such formulas increases rapidly with n.

In general, structural moments do not determine a composition structure uniquely, because they do not even determine the associated partition structure. See [34] for further discussion. Since uniqueness does hold in the special case of regenerative composition structures, it is natural to seek a characterisation of structural moments in this case. There is the following immediate consequence of Proposition 9.1 and Lemma 8.1:

Corollary 9.2 *A sequence $p(n), n = 1, 2, \ldots$ with $p(1) = 1$ and $0 < p(n) < 1$ for $n > 1$ is a sequence of structural moments of some regenerative composition structure if and only if the following conditions are fulfilled:*

(i) *the sequence $\Phi(n), n = 1, 2, \ldots$ defined by the recursion (41) with $\Phi(1) = 1$ is positive, and*

(ii) *each $\Phi(n:m), 1 \le m \le n < \infty$ defined by (27) is non-negative.*

If this is the case,

$$p(n) = \frac{\int_0^1 x^n \, \widetilde{\nu}(\mathrm{d}x)}{\int_0^1 (1 - (1-x)^n) \, \widetilde{\nu}(\mathrm{d}x) + n \, \mathtt{d}} \qquad n > 1$$

for some $\mathtt{d} \ge 0$ and some measure $\widetilde{\nu}$ on $]0, 1]$ with finite first moment.

Remark. We know that $p(n), n = 1, 2, \ldots$ is a moment sequence from the general facts about partition structures, or from the interpretation of $p(n)$ as the probability that n balls fall in the same box. From an analytical perspective, it does not seem obvious that the nonlinear tranform given by $p(n) = \Phi(n:n)/\Phi(n), n = 1, 2, \ldots$ indeed yields a completely monotonic sequence for arbitrary Laplace exponent.

Because the structural moments are determined by the (unordered) partition structure, Proposition 9.1 and Kingman's representation of partition structures [27] imply:

Corollary 9.3 *Each distribution of an infinite exchangeable partition of $\mathbb{N}$ (which can be identified with a partition structure) corresponds to at most one regenerative composition structure. Equivalently, for each distribution of a decreasing sequence (Y_j) with $Y_j \ge 0$ and $\sum Y_j \le 1$, there exists at most one distribution for a multiplicatively regenerative set $\widetilde{\mathcal{R}} \in \mathcal{M}$ such that the ranked lengths of interval components of $\widetilde{\mathcal{R}}^c$ are distributed like (Y_j).*

A constructive method to verify if a given exchangeable partition of $\mathbb{N}$ is induced by a regenerative composition structure amounts to computing q from the structural moments, and then checking that the given EPPF coincides with the EPPF computed by formulas (6) and (4).

The general problem of characterising structural distributions of partition structures was posed by Pitman and Yor [37]. The characterisation of structural distributions of regenerative composition structures provided by Corollary 9.2 leaves open the following question: given the collection of structural moments of a regenerative composition, or given its Laplace exponent Φ, describe in some way how the classes of the associated unordered partition should be arranged to produce the composition. We answer some restricted forms of this question in the next section, but do not see how to answer it in any generality.

9.2 Singleton probabilities

Instead of the event 'n balls fall in same box', consider the event 'n balls fall in n different boxes'. Let $e(n)$ be the probability of this event, that is

$$e(n) := p(1, 1, \ldots, 1) = q(n : 1)q(n - 1 : 1) \cdots q(2 : 1).$$

By the definition and from the representation (35) we derive

$$\frac{e(n)}{e(n-1)} = q(n : 1) = n\left(1 - \frac{\Phi(n-1)}{\Phi(n)}\right)$$

which can be read as

$$\frac{\Phi(1)}{\Phi(n)} = \prod_{j=2}^{n}\left(1 - \frac{e(j)}{j\, e(j-1)}\right). \tag{42}$$

This shows that any one of the sequences $(e(n), n > 0)$, $(q(n : 1), n > 0)$ or $(\Phi(n)/Phi(1), n > 0)$ uniquely determines each of the other two sequences.

In the variables $q(n : 1), n = 1, 2, \ldots$ the elements of decrement matrix become polynomials

$$q(n : m) = \binom{n}{m}\sum_{j=0}^{m}(-1)^{m-j+1}\binom{m}{j}\prod_{k=0}^{j-1}\left(1 - \frac{q(n-k : 1)}{n-k}\right), \tag{43}$$

to be compared with the rational functions of structural moments considered in Subsection 9.1. For example

$$q(4 : 2) = 2\, q(3 : 1) - \frac{3}{2}\, q(4, 1) - \frac{1}{2}\, q(3 : 1)\, q(4 : 1).$$

The definition of $e(n)$ makes sense for a general partition structure. Thus to check if a given partition structure is induced by a regenerative composition structure, we can use the above formulas to translate $e(n), n > 0$, into q and then compare the EPPF resulting from (6), (4) with the given EPPF. In particular, if a regenerative rearrangement is possible, the sequences $p(n), n > 0$ and $e(n), n > 0$ must be computable from each other, as appears by eliminating the variables Φ from $p(n) = \Phi(n : n)/\Phi(n)$ and (42).

10 The two-parameter family

10.1 General setup

Consider the (α, θ)-*partition structure* determined by following formula of [31, 34] for the distribution of Π_n, an exchangeable partition of $[n]$: for each particular partition π of $[n]$ into k classes of sizes $n_1, \ldots, n_k$

$$\mathbb{P}(\Pi_n = \pi) = \frac{\prod_{i=1}^{k-1}(\theta + \alpha i)}{[1+\theta]_{n-1}}\prod_{i=1}^{k}[1-\alpha]_{n_i - 1} \tag{44}$$

where the notation (13) is used for rising factorials. This formula defines a partition structure for $0 \le \alpha < 1$ and $\theta \ge 0$, and also for some (α, θ) with either $\alpha < 0$ or $\theta < 0$. We wish to establish if this partition structure can be associated with some regenerative composition structure.

Following the method in Section 9 we first compute $e(n)$ as a special case of (44):

$$e(n) = p(1,1,\ldots,1) = \prod_{j=0}^{n-1} \frac{\theta+\alpha j}{\theta+j}$$

which leads by application of (42) to

$$\frac{\Phi(n)}{\Phi(1)} = \frac{n[\theta+1]_{n-1}}{[2+\theta-\alpha]_{n-1}}\,. \tag{45}$$

This yields, by virture of (39) or (27), the formula

$$\frac{\Phi(n:m)}{\Phi(1)} = \binom{n}{m} \frac{[1-\alpha]_{m-1}}{[2+\theta-\alpha]_{n-1}} \frac{[\theta+1]_{n-1}}{[\theta+n-m]_m} \left((n-m)\alpha + m\theta\right).$$

Therefore

$$q(n:m) = \frac{\Phi(n:m)}{\Phi(n)} = \binom{n}{m} \frac{[1-\alpha]_{m-1}}{[\theta+n-m]_m} \frac{((n-m)\alpha+m\theta)}{n}\,. \tag{46}$$

Since q in (46) is non-negative exactly when $0 \le \alpha < 1$ and $\theta \ge 0$ we conclude that q is the decrement matrix of a regenerative composition structure for precisely this range of parameters.

Observe that the resulting formula

$$p(n) = q(n:n) = \frac{[1-\alpha]_{n-1}}{[1+\theta]_{n-1}} \tag{47}$$

yields the moments of beta$(1-\alpha, \alpha+\theta)$, which is the structural distribution for *all* members of the two-parameter family of partition structures.

Adopting the normalisation $\Phi(1) = \mathrm{B}(1-\alpha, 1+\theta)$, where

$$\mathrm{B}(a,b) := \Gamma(a)\Gamma(b)/\Gamma(a+b)$$

the Laplace exponent extending (45) becomes

$$\Phi(s) = s\mathrm{B}(1-\alpha, s+\theta). \tag{48}$$

The corresponding measure is determined by the formula

$$\widetilde{\nu}[x,1] = x^{-\alpha}(1-x)^{\theta}, \qquad 0 < x < 1. \tag{49}$$

It remains to check that the partition structure induced by this regenerative composition structure is given by (44). This is done in the following theorem:

Theorem 10.1 *For $0 \le \alpha < 1$ and $\theta \ge 0$ the distribution of the exchangeable random partition Π_n of $[n]$ derived from the regenerative composition structure with Laplace exponent (48) is that of an (α,θ) partition defined by formula (44). For other values of (α,θ), besides the limiting case $(1,\theta)$ for $\theta \ge 0$ which generates the pure singleton partition, there is no regenerative composition structure which generates an (α,θ)-partition structure.*

Proof. By the above discussion we can restrict consideration to the case $0 \le \alpha < 1$ and $\theta \ge 0$. By application of formulas (4), (6) and (46), the EPPF derived from the regenerative composition structure with Laplace exponent (48) is a sum of $k!$ terms of the form

$$\frac{1}{[\theta]_n} \prod_{i=1}^{k} [1-\alpha]_{n_i-1} \frac{(N_i - n_i)\alpha + n_i\theta}{N_i}$$

where the sequence $(n_1, \dots, n_k)$ and its tail sums $N_i = \sum_{j=i}^{k} n_j$ must be replaced by permutations of the sequence and correspondingly transformed tail sums. To match up with (44), it just has to be checked that the corrresponding sum of $k!$ terms derived from

$$\prod_{i=1}^{k} \frac{(N_i - n_i)\alpha + n_i\theta}{N_i((k-i)\alpha + \theta)} \tag{50}$$

equals 1. But this is easily verified together with the probabilistic interpretation given in the following corollary. □

Corollary 10.2 *In the setting of the previous theorem, given that the blocks of Π_n are of sizes $n_1, \dots, n_k$ when put in some arbitrary order, and given that the first $i-1$ of these blocks are the first $i-1$ blocks of the ordered partition $\mathcal{C}_n^*$, the conditional probability that this coincidence continues for one more step is the ith factor in* (50).

Put another way, given block sizes $n_1, \dots, n_k$ and that the first $i-1$ blocks have been picked to leave blocks of sizes n_j for $i \le j \le k$, the next block is the block of index j with probability proportional to $(N_i - n_j)\alpha + n_j\theta$.

Several particular instances of the above results are known, as indicated in the following discussion of special cases.

10.2 Case $(0, \theta)$ for $\theta \ge 0$

In this case the measure $\widetilde{\nu}$ in (49) is a probability measure, the beta$(1, \theta)$ distribution. So the above theorem and its corollary reduce to the well known fact that the ordered Ewens formula associated with beta$(1, \theta)$ stick-breaking puts its parts in a size-biased random order [12].

10.3 Case $(\alpha, 0)$ for $0 < \alpha < 1$

In this case

$$\widetilde{\nu}(\mathrm{d}x) = \alpha x^{-\alpha-1}\mathrm{d}x + \delta_1(\mathrm{d}x)$$

is a measure with a beta density on $]0,1[$ and a unit atom at 1. The product formula (6) reduces to

$$p(n_1, \dots, n_k) = n_k \alpha^{k-1} \prod_{j=1}^{k} \frac{[1-\alpha]_{n_j - 1}}{n_j!},$$

which is identical to the formula in [32, Equation (28)]. By comparision of these two formulas, the random composition is in this case is identical in distribution to that generated by $\mathcal{R}_\alpha \cap [0,1]$ where $\mathcal{R}_\alpha$ is the range of a stable subordinator of index α. In particular, $\mathcal{R}_\alpha$ can be realised as the zero set of a Bessel process of dimension $2 - 2\alpha$. For $\alpha = 1/2$ this is the zero set of a standard Brownian motion.

The decrement matrix q in this case has the special property that there is a probability distribution f on the positive integers such that

$$q(n:m) = f(m) \text{ if } m < n \text{ and } q(n:n) = 1 - \sum_{m=1}^{n-1} f(m). \tag{51}$$

Specifically,

$$f(m) = \frac{\alpha[1-\alpha]_{m-1}}{m!} \tag{52}$$

and hence $q(n:n) = [1-\alpha]_{n-1}/(n-1)!$. The work of Young [41] shows that the only non-degenerate regenerative composition structures with a decrement matrix of the form (51), for some probability distribution f on the positive integers, are those with f of the form (52), obtained by uniform sampling from $\mathcal{R}_\alpha \cap [0,1]$ for some $0 < \alpha < 1$.

The multiplicative regeneration property of $\mathcal{R}_\alpha \cap [0,1]$ is an immediate consequence of the standard regeneration and self-similarity properties of $\mathcal{R}_\alpha$ as a subset of $[0,\infty]$. It implies that $\mathcal{R}_\alpha \cap [0,1]$ has the same distribution as the closure of $\{1-\exp(-S_t), t \geq 0\}$ where (S_t) is a subordinator with no drift and Lévy measure

$$\nu(\mathrm{d}z) = \alpha(1-e^{-z})^{-\alpha-1}e^{-z}\mathrm{d}z + \delta_\infty(\mathrm{d}z)$$

on $[0,\infty]$ which is the image of $\widetilde{\nu}$ via $x \mapsto -\log(1-x)$, so ν has an atom of mass 1 at ∞.

As a check, let $\tau := \inf\{t : S_t = \infty\}$, which is the exponential time with rate 1 when the subordinator jumps to ∞. Then, by application of the transformation and the Lévy-Khintchine formula, if we let $G := \sup \mathcal{R}_\alpha \cap [0,1[$, then we find for $s > 0$

$$\mathbb{E}(1-G)^s = \mathbb{E}\left(\exp(-sS_{\tau-})\right) = \frac{1}{\Phi(s)} = \frac{B(1-\alpha+s,\alpha)}{B(1-\alpha,\alpha)}.$$

This confirms the well known fact that the distribution of $1-G$ is beta$(1-\alpha,\alpha)$. It may also be observed, using properties of the local time process $(L_t, t \geq 0)$ associated with $\mathcal{R}_\alpha$, as discussed in [30], that the exponential time τ can be represented as

$$\tau = c_\alpha \int_0^1 (1-t)^{-\alpha}\mathrm{d}L_t$$

for some constant c_α depending on the normalisation of the local time process. The fact that this local time integral has an exponential distribution was derived by an analytic argument in [20, Corollary 3.4].

As discussed in [32], the length of the last interval component $]G,1[$ of the complement to $\mathcal{R}_\alpha \cap [0,1]$ is a size-biased pick from the collection of the interval lengths, and conditionally on G the remaining interval components are in symmetric order; moreover these properties are inherited by the compositions of n for every n. Corollary 10.2 in this case is new. It makes precise another sense in which given the partition of n generated by $\mathcal{R}_\alpha \cap [0,1[$, the smaller blocks tend to come first in the composition of n.

10.4 Case (α,α) for $0 < \alpha < 1$

Passing to the variable $z = -\log(1-x)$ we see from (49) that the associated regenerative subset of $[0,\infty]$ has zero drift and Lévy measure

$$\nu(\mathrm{d}z) = \alpha(1-e^{-z})^{-\alpha-1}e^{-\alpha z}\,\mathrm{d}z \qquad z \in]0,\infty[\,.$$

It can be read from [38] that such a regenerative set is generated as the zero set of the squared Ornstein-Uhlenbeck process (X_t) of dimension $2-2\alpha$ driven by the stochastic differential equation $\mathrm{d}X_t = 2\sqrt{X_t}\,\mathrm{d}B_t + (2-2\alpha-X_t)\mathrm{d}t$ where (B_t) is a standard Brownian motion, and that the image of this regenerative set via $x = 1-e^{-z}$ is the zero set of a Bessel bridge of dimension $2-2\alpha$. In case $\alpha = 1/2$ this is a Brownian bridge, as in Example 3. In the notation introduced in the discussion of the previous case, this corresponds to conditioning $\mathcal{R}_\alpha \cap [0,1]$ on the event $1 \in \mathcal{R}_\alpha$. This can be rigorously understood by first conditioning on $G \in [1-\epsilon,1]$ and then taking a weak limit as $\epsilon \downarrow 0$.

The decrement matrix in this case has the special property that

$$q(n:m) = \frac{f(m)\,r(n-m)}{r(n)} \tag{53}$$

where f is given by (52) and $r(n) = [\alpha]_m/m!$ is the probability that a random walk on positive integers with step distribution f visits n. Equivalently, the composition probability function is

$$p\,(n_1,\ldots,n_k) = \frac{\prod_{j=1}^k f(n_j)}{r(n)} \tag{54}$$

or more explicitly

$$p\,(n_1,\ldots,n_k) = \frac{n!}{[\alpha]_n}\alpha^k \prod_{j=1}^k \frac{[1-\alpha]_{n_i-1}}{n_i!}. \tag{55}$$

It follows from a result of Kerov [24] that the decrement matrix of a non-degenerate regenerative composition structure can be expressed in the form (53) for some functions f and r iff it is of the form (55) for some $\alpha \in]0,1[$. The same conclusion is also a consequence Theorem 12.1 in the next section. The conclusion of Corollary 10.2 in this case is that given the partition of $[n]$ the block sizes appear in $\mathcal{C}_n$ in a uniform random order. This can be seen directly from the symmetry of formula (54) as a function of $(n_1, \ldots, n_k)$.

10.5 Case (α, θ) for $0 < \alpha < 1$, $\theta > 0$

It is known [37, 39, 33] that an (α, θ) partition of $\mathbb{N}$ can be constructed as follows. First construct a $(0, \theta)$ partition of $\mathbb{N}$, then shatter each class of this partition according to an independent $(\alpha, 0)$ partition. This operation restricts naturally to $[n]$ for each n, and can be interpreted in terms of a fragmentation operation on the frequencies of classes. This result can be lifted to the level of regenerative composition structures as follows.

Theorem 10.3 *For $0 < \alpha < 1$ and $\theta > 0$, let $Y_0 = 0$ and let $0 < Y_1 < Y_2 < \ldots$ be defined by the independent stick-breaking scheme* (14) *for X with beta$(1, \theta)$ distribution, for each $i = 1, 2, \ldots$ let $\mathcal{R}_\alpha(i)$ for $i = 1, 2, \ldots$ be a sequence of independent copies of the range $\mathcal{R}_\alpha$ of a stable subordinator, and define a random closed subset $\widetilde{\mathcal{R}}_{(\alpha,\theta)}$ of $[0,1]$ by*

$$\widetilde{\mathcal{R}}_{(\alpha,\theta)} = \{1\} \cup \bigcup_{i=1}^{\infty} \left([Y_{i-1}, Y_i] \cap [Y_{i-1} + \mathcal{R}_\alpha(i)] \right).$$

Then $\mathcal{R}_{(\alpha,\theta)}$ is a multiplicatively regenerative random subset of $[0,1]$, which can be represented as $\widetilde{\mathcal{R}}_{(\alpha,\theta)} = 1 - \exp(-\mathcal{R}_{(\alpha,\theta)})$ where $\mathcal{R}_{(\alpha,\theta)}$ is the range of a subordinator with Laplace exponent (48)*, and the composition structure obtained by uniform random sampling from $\mathcal{R}_{(\alpha,\theta)}$ is regenerative with decrement matrix* (46).

Proof. It is easily checked, using the muliplicative regeneration of the stick-breaking scheme, and the self-similarity of $\mathcal{R}_\alpha$, that $\widetilde{\mathcal{R}}_{(\alpha,\theta)}$ is multiplicatively regenerative. The description of the Laplace exponent then follows from Proposition 9.1, since the structural distribution is easily identified. $\square$

The particular case $\alpha = \theta$ of Theorem 10.3 is largely contained in the work of Aldous and Pitman [2]. In particular, for $\alpha = \theta = 1/2$ this construction of the zero set of a Brownian bridge plays a key role in the asymptotic theory of random mappings developed in [1] and [2].

11 The Green matrix

For a given composition probability function (1), the *Green matrix* is defined by the formula

$$g(n,j) = \sum_{\lambda \models n,\, j \in \{N_i\}} p(\lambda), \qquad 1 \le j \le n < \infty$$

where the summation is over all compositions $\lambda = (n_1, \ldots, n_k) \models n$ which have integer j among tail sums $N_j = n - n_1 - \ldots - n_{j-1}$ (where we set $n_0 = 0$). Recalling the interpretation of a regenerative composition structure as a consistent family of Markov chains $Q_n, n = 1, 2, \ldots$, as in Section 3, $g(n,j)$ is the chance that Q_n with transition matrix q and initial state n ever visits state j. In particular, $g(n,n) = 1$.

Example. For the 2-parameter family we have for $1 \le j \le n$

(i) for $(0, \theta)$

$$g(n,j) = \frac{\theta}{j + \theta}$$

as is well known;

(ii) for $(\alpha, 0)$
$$g(n,j) = \frac{[\alpha]_{n-j}}{(n-j)!}$$
which by (51)-(52) is the probability that a particular random walk with negative increments started at level n ever visits state j;

(iii) for (α, α)
$$g(n,j) = \frac{\binom{n}{j}[\alpha]_j}{(\alpha+n-j)\cdots(\alpha+n-1)}$$
which is the probability of the same event for the random walk of the previous case conditioned to hit 0.

Lemma 11.1 *The Green matrix of a regenerative composition structure is the unique solution of the recursion*
$$g(n,j) = \frac{j+1-q(j+1:1)}{n+1}\, g(n+1,j+1) + \frac{n+1-j}{n+1}\, g(n+1,j) \tag{56}$$
with boundary condition $g(n,n) = 1$.

Proof. The path of the chain Q_n, defining a composition of n, is obtained via random deletion of a state from $1, 2, \ldots, n+1$, then restricting a path of Q_{n+1} to the undeleted states and re-labeling the states by ranking then from 1 to n. The event 'Q_n visits j' occurs when either Q_{n+1} visits j and one of the states $j+1, \ldots, n+1$ is deleted (in which case state j retains the label) or Q_{n+1} visits $j+1$ and one of the states $1, \ldots, j+1$ is deleted (if state $j+1$ is not deleted it changes the label to j). The first event has probability $g(n+1,j)(n+1-j)/(n+1)$ and the second $g(n+1,j+1)(j+1)/(n+1)$. The events are not disjoint and their intersection is the event 'Q_{n+1} visits both $j+1$ and j, and state $j+1$ is deleted' which has probability $g(n+1,j+1)q(j+1:1)/(n+1)$. The uniqueness claim is obvious from the recursion. □

Next result gives an explicit formula for the Green matrix in terms of the representation (35) via Laplace exponent.

Theorem 11.2 *The Green matrix of a regenerative composition structure is*
$$g(n,j) = \Phi(j)\binom{n}{j}\sum_{a=0}^{n-j}\binom{n-j}{a}\frac{(-1)^a}{\Phi(j+a)}\,. \tag{57}$$

Proof. In view of
$$q(j+1:1) = (j+1)\left(1 - \frac{\Phi(j)}{\Phi(j+1)}\right)$$
the first factor in the right side of (56) equals $(j+1)\Phi(j)/((n+1)\Phi(j+1))$. Substituting this in (56), and canceling the common factor $\binom{n}{j}\Phi(j)$ the to-be-checked recursion follows from the identity
$$\Delta^{n-j+1}s(j) = \Delta^{n-j}s(j+1) - \Delta^{n-j}s(j)$$
where Δ is the forward difference operator $\Delta s(i) := s(i+1) - s(i)$ and s is the sequence $s(i) = 1/\Phi(i)$ for $i \geq 1$. □

We give one application of the formula. Let L_n be the *last* part of $\mathcal{C}_n$. In the event $\{L_n = j\}$ the chain Q_n visits state j and then has the last positive decrement j. The distribution of the last part follows from this observation and (57):
$$\mathbb{P}(L_n = j) = g(n,j)q(j:j) = \Phi(j:j)\binom{n}{j}\sum_{a=0}^{n-j}\binom{n-j}{a}\frac{(-1)^a}{\Phi(j+a)}\,. \tag{58}$$

In particular, normalising by $\Phi(1) = 1$ for simplicity,

$$\mathbb{P}(L_n = 1) = n\left[1 - \sum_{k=2}^{n}\binom{n-1}{k-1}\frac{(-1)^k}{\Phi(k)}\right]. \tag{59}$$

12 Symmetry

Each composition structure $(\mathcal{C}_n)$ has a dual $(\hat{\mathcal{C}}_n)$, where $\hat{\mathcal{C}}_n$ is the sequence of parts of $\mathcal{C}_n$ in reverse order. If $(\mathcal{C}_n)$ is derived by uniform sampling from a random closed set $\widetilde{\mathcal{R}} \subseteq [0,1]$, then $\hat{\mathcal{C}}_n$ is derived similarly from $1 - \widetilde{\mathcal{R}}$. If $(\mathcal{C}_n)$ is regenerative, and so is $(\hat{\mathcal{C}}_n)$, then $(\mathcal{C}_n)$ and $(\hat{\mathcal{C}}_n)$ must be identical in distribution, by Corollary 9.3. Equivalently, $\widetilde{\mathcal{R}} \stackrel{d}{=} 1 - \widetilde{\mathcal{R}}$, in which case we call the composition structure *reversible*. Two degenerate examples are provided by $\widetilde{\mathcal{R}} = \{0\} \cup \{1\}$ and $\widetilde{\mathcal{R}} = [0,1]$. The existence of regenerative composition structures which are non-degenerate and reversible is quite surprising and counter-intuitive, because the ideas of stick-breaking and multiplicative regeneration suggest that typical interval sizes should decay in some sense from the left to the right. However, it is evident from the formula (54) that for every $0 < \alpha < 1$ the regenerative composition structure associated with an (α, α) partition is reversible. Indeed, this composition structure is *symmetric*, meaning that the probability of a composition $(n_1, \ldots, n_k)$ is a function of $(n_1, \ldots, n_k)$ which is symmetric with respect to all permutations of the arguments. The equivalent condition on $\widetilde{\mathcal{R}}$ is that the interval components of the complement of $\widetilde{\mathcal{R}}$ form an exchangeable interval partition of $[0,1]$, as defined in [3]. We note in passing that a large family of symmetric composition structures was derived from the jumps of a subordinator in [35]. See also [18].

Theorem 12.1 *Let $(\mathcal{C}_n)$ be the regenerative composition structure derived by uniform sampling from a random closed set $\widetilde{\mathcal{R}} \subseteq [0,1]$. Let F_n be the size of the first part of $\mathcal{C}_n$, and L_n be the size of the last part of $\mathcal{C}_n$. The following conditions are equivalent:*
(i) $\mathbb{P}(F_n = 1) = \mathbb{P}(L_n = 1)$ *for all n;*
(ii) $F_n \stackrel{d}{=} L_n$ *for all n;*
(iii) $(\mathcal{C}_n)$ *is reversible;*
(iv) $(\mathcal{C}_n)$ *is symmetric;*
(v) $(\mathcal{C}_n)$ *is the regenerative composition structure with EPPF* (55), *associated with an (α,α) partition as in Section* 10.4 *for some $\alpha \in [0,1]$.*

Before the proof of this result, we read from Theorem 5.3 and the discussion of Section 10.4 the following restatement of the equivalence of conditions (iii) and (v):

Corollary 12.2 *For a random closed subset $\widetilde{\mathcal{R}}$ of* $[0,1]$*, the following two conditions are equivalent:*
(i) $\widetilde{\mathcal{R}}$ *is multiplicatively regenerative and* $\widetilde{\mathcal{R}} \stackrel{d}{=} 1 - \widetilde{\mathcal{R}}$.
(ii) $\widetilde{\mathcal{R}}$ *is distributed like the zero set of a standard Bessel bridge of dimension* $2-2\alpha$*, for some* $\alpha \in [0,1]$.

Proof of Theorem 12.1. According to formula (28), for any regenerative composition structure

$$\mathbb{P}(F_n = 1) = q(n:1) = \frac{\Phi(n) - \Phi(n-1)}{\Phi(n)/n} \tag{60}$$

and the expressions (60) and (59) are obviously equal if $n = 1$ or $n = 2$. We know that the (α,α) regenerative composition structure is symmetric, hence reversible. So for $\Phi_\alpha(n) := [1+\alpha]_{n-1}/(n-1)!$, the identity $\mathbb{P}(F_n = 1) = \mathbb{P}(L_n = 1)$ together with (60) and (59) yields

$$\frac{n\alpha}{n-1+\alpha} = n - n\frac{(-1)^n}{\Phi_\alpha(n)} - n\sum_{k=2}^{n-1}\binom{n-1}{k-1}\frac{(-1)^k}{\Phi_\alpha(k)}. \tag{61}$$

Suppose now that a regenerative composition structure is such that $\mathbb{P}(F_n = 1) = \mathbb{P}(L_n = 1)$ for all $n = 1, 2, 3, \ldots$, and let us prove by induction that its Laplace exponent Φ normalised by $\Phi(1) = 1$ is such that

$$\Phi(s) = \Phi_\alpha(s) \tag{62}$$

for all $s = 1, 2, \ldots$, where $\alpha \in [0, 1]$ is defined by (62) for $s = 2$, that is $\Phi(2) = 1 + \alpha$. According to (60) and (59), we have for all $n = 2, 3, \ldots$ that

$$\frac{\Phi(n) - \Phi(n-1)}{\Phi(n)/n} = n - n\frac{(-1)^n}{\Phi(n)} - n\sum_{k=2}^{n-1}\binom{n-1}{k-1}\frac{(-1)^k}{\Phi(k)} \tag{63}$$

so if we make the inductive hypothesis that (62) holds for all $s \leq n-1$ then we read from (61) and (63) that

$$\frac{\Phi(n) - \Phi(n-1)}{\Phi(n)/n} = \frac{n\alpha}{n-1+\alpha} + n(-1)^n\left[\frac{1}{\Phi_\alpha(n)} - \frac{1}{\Phi(n)}\right]$$

which yields the expression

$$\Phi(n) = (\Phi_\alpha(n-1) - (-1)^n)/(1 - \alpha(n-1-\alpha) - (-1)^n/\Phi_\alpha(n)).$$

But we know this formula holds for $\Phi(n) = \Phi_\alpha(n)$, so this must be the unique solution of the recursion, and the inductive step is established. Finally, the sequence $\Phi(1), \Phi(2), \ldots$ determines $\Phi(s)$ for all $s \geq 0$, by consideration of the second formula in (19), and the fact that a finite measure on $[0, 1]$ is determined by its moments. $\square$

13 Transition probabilities

Transition probabilities describing the succession of random compositions $(\mathcal{C}_n)$ or ordered partitions $(\mathcal{C}_n^*)$ as n grows follow at once from the product formula (6) for the composition probability function. For ordered partitions of $[n]$ these transition probabilities can be read immediately from (3), as indicated in James [22, §5.4].

Assuming that $\mathcal{C}_n^* = (A_1, \ldots, A_k)$, an ordered partition $\mathcal{C}_{n+1}^*$ of $[n+1]$ is obtained either by inserting singleton block $\{n+1\}$ into the sequence $A_1, \ldots, A_k$ or by adjoining the element $n+1$ to one of the blocks. It is easy to compute that $n+1$ is inserted before A_1 with probability

$$\frac{q(n+1:1)}{n+1}$$

or adjoined to A_1 with probability

$$\frac{n_1+1}{n+1}\,\frac{q(n+1:n_1+1)}{q(n:n_1)}.$$

Inductively, with probability

$$\prod_{i=1}^{j}\left(1 - \frac{q(N_i+1:1)}{N_i+1} - \frac{n_i+1}{N_i+1}\,\frac{q(N_i+1:n_i+1)}{q(N_i:n_i)}\right)$$

$n+1$ is neither inserted immediately before nor adjoined to one of the blocks $A_1, \ldots, A_j$, and conditionally on this event (and given $(A_1, \ldots, A_k)$) this element is inserted as a singleton immediately following A_j with probability

$$\frac{q(N_{j+1}+1:1)}{N_{j+1}+1}$$

or adjoined to A_{j+1} (for $j < k$) with probability

$$\frac{n_{j+1}+1}{N_{j+1}+1}\,\frac{q(N_{j+1}+1:n_{j+1}+1)}{q(N_{j+1}:n_{j+1})}.$$

Here, the n_i are the sizes of the A_i and the N_i are as in (6).

A transition law for integer compositions follows from the above. It is exactly the same as for the analogous ordered set partitions with the exception of the case when a composition of n is changed by appending a 1 to a series of unit parts like $1, 1, \dots, 1$, in which case the transition probability is obtained by summation of individual probabilities of all possible singleton insertions into the series.

14 Interval partitions

The above probabilities of the two kinds of transition (insertion and joining) are equal to the expected sizes of intervals of a partition of $[0, 1]$ induced by a uniform sample of n points and $\widetilde{\mathcal{R}}$. From this viewpoint, a better prediction of the 'future' compositions arising when more points are added to the sample is obtained by conditioning on the actual sizes of intervals.

At first we shall describe a somewhat simpler distribution of the interval sizes for the $[0, \infty]$-partition, which can be seen as discretisation of a subordinator in the spirit of [32], Sections 3 and 4. For each n, a random set $\mathcal{R}$ and exponential order statistics $\epsilon_{1n}, \dots, \epsilon_{nn}$ induce a partition of $[0, \infty]$ associated with finite composition $\mathcal{C}_n$. The partition is comprised of two kinds of parts: those containing some sample points or not. The parts of the first kind are either open interval components of $\mathcal{R}^c$ which contain at least one of the ϵ_{jn}'s, or one-point parts $\{\epsilon_{jn}\}$ corresponding to $\epsilon_{jn} \in \mathcal{R}$ and appearing with positive probability only for $\mathtt{d} > 0$. The parts of the second kind are the connected components (intervals or separate points) of the set resulting from removing parts of the first kind. The parts of different kinds interlace and if $\mathcal{C}_n$ has K_n classes there are $2K_n + 1$ pieces of the partition, say $J_{1n}, I_{1n}, \dots, J_{K_n-1,n}, I_{K_n,n}, J_{K_n+1,n}$, which can be open or semiopen intervals or one-point sets. Let $G_{1n}, H_{1n}, \dots, G_{K_n-1,n}, H_{K_n,n}, G_{K_n+1,n}$ be the sizes of the parts, with slight abuse of language we will call them 'intervals', with understanding that some of them can degenerate into a point.

Theorem 14.1 *The distribution of the random sequence $G_{1n}, H_{1n}, \dots, G_{K_n-1,n}, H_{K_n,n}, G_{K_n+1,n}$ of interval sizes has the following properties:*

(i) *given the composition $\mathcal{C}_n$ all interval sizes are conditionally independent,*

(ii) *G_{1n} is independent of $\mathcal{C}_n$ and also independent of other interval sizes, and has Laplace transform*

$$\mathbb{E} \exp(-s\, G_{1n}) = \frac{\Phi(n)}{\Phi(n+s)}\,, \tag{64}$$

(iii) *the unconditional distribution of H_{1n} is given by*

$$\mathbb{P}\,(H_{1n} \in \mathrm{d}z) = \frac{1 - e^{-nz}}{\Phi(n)}\,\nu(\mathrm{d}z) + \frac{n\mathtt{d}}{\Phi(n)}\,\delta_0(\mathrm{d}z), \tag{65}$$

and given $\mathcal{C}_n$ the analogous conditional distribution of H_{1n} is

$$\frac{\binom{n}{m}(1 - e^{-z})^m\, e^{-(n-m)z}\,\nu(\mathrm{d}z) + n\,\mathtt{d}\,1(m = 1)\,\delta_0(\mathrm{d}z)}{\Phi(n : m)}$$

where m is the first part of $\mathcal{C}_n$,

(iv) *conditionally on the event that the first $j - 1$ parts of $\mathcal{C}_n$ sum up to m, the truncated sequence $G_{jn}, H_{jn}, \dots, H_{K_n,n}, G_{K_n+1,n}$ is independent of the variables $G_{1n}, H_{1n}, \dots, G_{j-1,n}, H_{j-1,n}$ and of the first $j - 1$ parts of composition $\mathcal{C}_n$, and has the same distribution as the interlacing sequence*

$$G_{1,n-m}\,, H_{1,n-m}\,, \dots, H_{K_{n-m}-j,n-m}\,, G_{K_{n-m}-j+1,n-m}$$

of interval sizes associated with the composition $\mathcal{C}_{n-m}$ of integer $n - m$.

Proof. The independence claims involved in (i) and (iv) follow from the memoryless property of the exponential distribution and the strong Markov property of $\widetilde{\mathcal{R}}$ applied at the right endpoints of intervals I_j or J_j. Formulas (64), (65) follow from Lemma 5.2 and the second formula in (iii) follows by routine conditioning. □

Mapping $[0,\infty]$ to $[0,1]$ by $z \mapsto 1-e^{-z}$ sends the partition of $[0,\infty]$ to a partition of the unit interval, say $\widetilde{J}_{1n}, \widetilde{I}_{1n}, \ldots, \widetilde{I}_{K_n}, \widetilde{J}_{K_n+1}$, which is the partition induced by a uniform sample and a multiplicatively regenerative set $\widetilde{\mathcal{R}}$. The probability law of the partition of $[0,1]$ follows from Theorem 14.1. Thus, by virtue of the identity $\mathbb{E}(1-\widetilde{G}_{1n})^s = \mathbb{E}\exp(-s\,G_{1n})$ the Laplace transform (64) becomes a Mellin transform. Similarly, the ratio $\widetilde{H}_{1n}/(1-\widetilde{G}_{1n})$ is independent of $\widetilde{G}_{1n}$ and has distribution

$$\mathbb{P}\left(\frac{\widetilde{H}_{1n}}{1-\widetilde{G}_{1n}} \in \mathrm{d}x\right) = \frac{1-(1-x)^n}{\Phi(n)}\,\widetilde{\nu}(\mathrm{d}x) + \frac{n\mathtt{d}}{\Phi(n)}\,\delta_0(\mathrm{d}x)\,.$$

The distribution of the rest intervals follows recursively, by scaling with factor $(1-\widetilde{G}_{1n}-\widetilde{H}_{1n})^{-1}$.

The sizes of these $2K_n+1$ intervals, say $\widetilde{G}_{jn}$ and $\widetilde{H}_{jn}$, determine the law of the extended composition when adding new sample points. For example,

$$\mathbb{E}\widetilde{G}_{1n} = 1-\mathbb{E}\,(1-\widetilde{G}_{1n}) = 1-\frac{\Phi(n)}{\Phi(n+1)} = \frac{\Phi(n+1:1)}{(n+1)\Phi(n+1)}$$

which by (28) is equal to $q(n+1:1)/(n+1)$ in accord with Section 13. The sizes also have a transparent frequency interpretation in terms of the infinite composition $\mathcal{C}$. For example, $\widetilde{G}_{1n}$ is the total frequency of the classes of $\mathcal{C}^*$ strictly preceding the first class represented in $\mathcal{C}_n^*$, and $\widetilde{H}_{1n}$ is the frequency of the first class represented in $\mathcal{C}_n^*$.

Tripartite decomposition of $[0,1]$**.** For $n=1$ the partition consists of three intervals $\widetilde{J}_{11}, \widetilde{I}_{11}, \widetilde{J}_{21}$ of sizes $G := \widetilde{G}_{11}, H := \widetilde{H}_{11}, D := \widetilde{G}_{21}$. The variable H is the frequency of the class of element 1 and its distribution is the structural distribution. Similarly, G is the total frequency of classes strictly preceding the class of 1 in $\mathcal{C}^*$, and D is the total frequency of classes strictly following the class of 1.

Moments of G, H and D have clear interpretation in terms of finite compositions. Thus

$$\mathbb{E}\,(1-G)^{n-1} = \sum_{m=1}^{n}\frac{m}{n}\,q(n:m) = \frac{\Phi(1)}{\Phi(n)} \tag{66}$$

is the probability that element 1 is in the first block of $\mathcal{C}_n^*$ or, what is the same, that a size-biased pick of a part from $\mathcal{C}_n$ yields the first part. Similarly,

$$\mathbb{E}\,D^{n-1} = \frac{q(n:1)}{n} = \frac{\Phi(n:1)}{n\,\Phi(n)} \tag{67}$$

is the probability that $\{1\}$ is the first block of $\mathcal{C}_n^*$.

Furthermore, the random variable H can be written as a product of two independent variables $1-G$ and $H/(1-G)$, hence

$$\mathbb{E}\left(\frac{H}{1-G}\right)^{n-1} = \frac{\mathbb{E}H^{n-1}}{\mathbb{E}(1-G)^{n-1}} = \frac{\Phi(n:n)}{\Phi(1)} \tag{68}$$

which is the conditional probability that the composition $\mathcal{C}_n^*$ is trivial given 1 is in the first block.

For joint moments we have the formula

$$\mathbb{E}G^i\,H^{j-1}\,D^k = \left(\sum_{a=0}^{i}\binom{i}{a}\frac{(-1)^a}{\Phi(a+j+k)}\right)\left(\sum_{b=0}^{k}(-1)^b\binom{j}{b}\Phi(j+b:j+b)\right) \tag{69}$$

(the second sum may be further converted to variables $\Phi(1), \Phi(2), \ldots$) which follows from (66), (68) and $\mathbb{E}H^n = p(n) = \Phi(n:n)/\Phi(n)$ by the binomial expansion of

$$G^i\, H^j\, D^k = (1-(1-G))^j (1-G)^{j+k} \left(\frac{H}{1-G}\right)^j \left(1 - \frac{H}{1-G}\right)^k .$$

The joint moments have the following interpretation. Let (A_1, A_2, A_3) be an ordered partition of $[n]$, $n = i+j+k$, such that $1 \in A_2$ and the blocks are of sizes i, j and k, respectively, with $i \geq 0$, $j \geq 1$ and $k \geq 0$. Then (69) is the probability that A_2 is a block of $\mathcal{C}_n^*$ and (A_1, A_2, A_3) is coarser than $\mathcal{C}_n^*$.

It follows that

$$\binom{n-1}{i, j-1, k} \mathbb{E}\, G^i\, H^{j-1}\, D^k$$

is the probability that a size-biased pick of a part of $\mathcal{C}_n$ is j, and this part is preceded by a composition of i and followed by a composition of k (with the obvious meaning when i or k is zero). For $k = 0$ this probability is equal to $(j/n)\mathbb{P}(L_n = j)$ where L_n is the last part of $\mathcal{C}_n$, computing this yields an alternative proof for (58) and the formula for the Green matrix (57).

Acknowledgment Thanks to the referee for a careful reading of the paper, and for a number of suggestions which helped to improve the exposition.

References

[1] D. Aldous and J. Pitman. Brownian bridge asymptotics for random mappings. *Random Structures and Algorithms*, 5:487–512, 1994.

[2] D. Aldous and J. Pitman. Two recursive decompositions of Brownian bridge related to the asymptotics of random mappings. Technical Report 595, Dept. Statistics, U.C. Berkeley, 2002. Available via www.stat.berkeley.edu.

[3] D. J. Aldous. Exchangeability and related topics. In P.L. Hennequin, editor, *École d'été de probabilités de Saint-Flour, XIII—1983*, pages 1–198. Springer, Berlin, 1985. Lecture Notes in Mathematics 1117.

[4] C. Berg, J.P.R. Christensen and P. Ressel. *Harmonic analysis on semigroups. Theory of positive definite and related functions.* Springer Graduate Texts in Mathematics vol. 100, N.Y., 1984.

[5] J. Bertoin. *Lévy processes*. Cambridge University Press, Cambridge, 1996.

[6] J. Bertoin. Regenerative embedding of Markov sets. *Probab. Theory Related Fields*, 108(4):559–571, 1997.

[7] J. Bertoin and M. Yor. On subordinators, self-similar Markov processes and some factorizations of the exponential variable. *Electron. Comm. Probab.*, 6:95–106 (electronic), 2001.

[8] F. T. Bruss and C. A. O'Cinneide. On the maximum and its uniqueness for geometric random samples. *J. Appl. Probab.*, 27(3):598–610, 1990.

[9] P. Carmona, F. Petit, and M. Yor. On the distribution and asymptotic results for exponential functionals of Lévy processes. In *Exponential functionals and principal values related to Brownian motion*, pages 73–130. Rev. Mat. Iberoam., Madrid, 1997.

[10] P. Diaconis and D. A. Freedman. The Markov Moment Problem and de Finettti's Theorem: Parts I and II. Technical Report 631, Dept. Statistics, U.C. Berkeley, 2003. http://www.stat.berkeley.edu/tech-reports/631.pdf.

[11] K. Doksum. Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, 2:183 – 201, 1974.

[12] P. Donnelly and P. Joyce. Consistent ordered sampling distributions: characterization and convergence. *Adv. Appl. Prob.*, 23:229–258, 1991.

[13] W. J. Ewens and S. Tavaré. The Ewens sampling formula. In N. S. Johnson, S. Kotz, and N. Balakrishnan, editors, *Multivariate Discrete Distributions.* Wiley, New York, 1995.

[14] W. Feller. *An Introduction to Probability Theory and its Applications*, volume II. Wiley, 2nd edition, 1971.

[15] A. Gnedin. Three sampling formulas. *Combinatorics, Probability and Computing.*

[16] A. Gnedin. Bernoulli sieve. *Bernoulli*

[17] A. V. Gnedin. The representation of composition structures. *Ann. Probab.*, 25(3):1437–1450, 1997.

[18] A. V. Gnedin. On the Poisson-Dirichlet limit. *J. Multivariate Anal.*, 67(1):90–98, 1998.

[19] A. V. Gnedin, J. Pitman and M. Yor. Asymptotic laws for regenerative composition structures I: the regular variation case. Preprint.

[20] M. Gradinaru, B. Roynette, P. Vallois and M. Yor. Abel transform and integrals of Bessel local times. *Ann. Inst. H. Poincaré Probab. Statist.*, 35(4):531–572, 1999.

[21] F. M. Hoppe. The sampling theory of neutral alleles and an urn model in population genetics. *Journal of Mathematical Biology*, 25:123 – 159, 1987.

[22] L. F. James. Poisson calculus for spatial neutral to the right processes. arXiv:math.PR/0305053, 2003.

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

[24] S. Kerov. Coherent random allocations and the Ewens-Pitman formula. PDMI Preprint, Steklov Math. Institute, St. Petersburg, 1995.

[25] S. Kerov. The boundary of Young lattice and random Young tabeaux. *DIMACS Ser. Discr. Math. Theor. Comp. Sci.*, 24: 133–158, 1996.

[26] H. Kesten. *Hitting probabilities of single points for processes with stationary independent increments.* Memoirs of the American Mathematical Society, No. 93. American Mathematical Society, Providence, R.I., 1969.

[27] J. F. C. Kingman. The representation of partition structures. *J. London Math. Soc.*, 18:374–380, 1978.

[28] J. F. C. Kingman. *The Mathematics of Genetic Diversity.* SIAM, 1980.

[29] B. Maisonneuve. Ensembles régénératifs de la droite. *Z. Wahrsch. Verw. Gebiete*, 63:501 – 510, 1983.

[30] Yu. A. Neretin. The group of diffeomorphisms of a ray, and random Cantor sets. *Mat. Sb.*, 187(6):73–84, 1996. Translation in *Sbornik Math.* 187 (1996), no. 6, 857–868.

[31] J. Pitman. Exchangeable and partially exchangeable random partitions. *Probab. Th. Rel. Fields*, 102:145–158, 1995.

[32] J. Pitman. Partition structures derived from Brownian motion and stable subordinators. *Bernoulli*, 3:79–96, 1997.

[33] J. Pitman. Coalescents with multiple collisions. *Ann. Probab.*, 27:1870–1902, 1999.

[34] J. Pitman. Combinatorial stochastic processes. Technical Report 621, Dept. Statistics, U.C. Berkeley, 2002. Lecture notes for St. Flour course, July 2002. Available via www.stat.berkeley.edu.

[35] J. Pitman. Poisson-Kingman partitions. In D. R. Goldstein, editor, *Science and Statistics: A Festschrift for Terry Speed*, volume 30 of *Lecture Notes-Monograph Series*, pages 1–34. Institute of Mathematical Statistics, Hayward, California, 2003.

[36] J. Pitman and T.P. Speed. A note on random times. *Stoch. Proc. Appl.*, 1:369–374, 1973.

[37] J. Pitman and M. Yor. Random discrete distributions derived from self-similar random sets. *Electron. J. Probab.*, 1:Paper 4, 1–28, 1996.

[38] J. Pitman and M. Yor. On the lengths of excursions of some Markov processes. In *Séminaire de Probabilités XXXI*, pages 272–286. Springer, 1997. Lecture Notes in Math. 1655.

[39] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, 25:855–900, 1997.

[40] S. Sawyer and D. Hartl. A sampling theory for local selection. *J. Genet.*, 64:21–29, 1985.

[41] J. E. Young. *Partition-valued stochastic processes with applications.* Ph. d. thesis, University of California, Berkeley, 1995.