Записки научных
семинаров ПОМИ
Том 326, 2005 г.

A. Gnedin, J. Pitman

SELF-SIMILAR AND MARKOV COMPOSITION STRUCTURES

Abstract. The bijection between composition structures and random closed subsets of the unit interval implies that the composition structures associated with $S \cap [0,1]$ for a self-similar random set $S \subset \mathbb{R}_+$ are those which are consistent with respect to a simple truncation operation. Using the standard coding of compositions by finite strings of binary digits starting with a 1, the random composition of n is defined by the first n terms of a random binary sequence of infinite length. The locations of 1s in the sequence are the places visited by an increasing time-homogeneous Markov chain on the positive integers if and only if $S = \exp(-W)$ for some stationary regenerative random subset W of the real line. Complementing our study in previous papers, we identify self-similar Markovian composition structures associated with the two-parameter family of partition structures.

1. Introduction

A *composition of* n is a sequence $\lambda = (\lambda_1, \dots, \lambda_\ell)$ of some number ℓ of positive integer *parts* λ_i with $\sum_{i=1}^{\ell} \lambda_i = n$. We may regard λ as a distribution of n identical balls in a row of ℓ boxes, with λ_i the number of balls in the ith box from the left end of the row. Thus the composition $(2, 4, 1, 2)$ of 9 may be represented in *balls-in-boxes notation* as

$$(2,4,1,2) \leftrightarrow [00]\ [0000]\ [0]\ [00] \tag{1}$$

or recoded in *binary notation* by replacing each "[0" in the balls-in-boxes notation by 1 and ignoring each "]", to obtain in this example

$$(2,4,1,2) \leftrightarrow 101000110.$$

In general, the first digit in the binary notation of a composition must be a 1, but the remaining digits can be chosen freely, so there are 2^{n-1} different compositions of n. Two other notations will be useful. We write

Research supported in part by N.S.F. Grant DMS-0405779.

$\lambda^{\leftarrow}$ for the *reversal* of λ and $\lambda^{\downarrow}$ for the *decreasing rearrangement of* λ, also called the *partition derived from* λ. For instance

$$(2,4,1,2)^{\leftarrow} = (2,1,4,2) \leftrightarrow [00]\ [0]\ [0000]\ [00] \leftrightarrow 101100010$$

$$(2,4,1,2)^{\downarrow} = (4,2,2,1) \leftrightarrow [0000]\ [00]\ [00]\ [0] \leftrightarrow 100010101$$

A *random composition of* n is a random variable $\mathcal{C}_n$ with values in the set of all compositions of n. We are interested in sequences $(\mathcal{C}_n)$ of random compositions which are consistent as n varies with respect to various *reduction operations*. In the balls-in-boxes description, let the places of the n balls be indexed from left to right by the set $[n] := \{1, \dots, n\}$. Let (Y_n) be a sequence of random variables with $Y_n \in [n]$ for each n, with Y_n independent of $\mathcal{C}_n$. Let $\mathcal{C}_n^-$ be the composition of $n-1$ obtained by deleting the ball in place Y_n from the balls-in-boxes representation of $\mathcal{C}_n$. For instance, if

$$\mathcal{C}_9 = (2,4,1,2) \leftrightarrow [00]\ [0000]\ [0]\ [00] \leftrightarrow 101000110$$

as above, and $Y_9 = 6$, then

$$\mathcal{C}_9^- = (2,3,1,2) \leftrightarrow [00]\ [000]\ [0]\ [00] \leftrightarrow 10100110.$$

Whereas if instead $Y_9 = 7$, then

$$\mathcal{C}_9^- = (2,4,2) \leftrightarrow [00]\ [0000]\ [00] \leftrightarrow 10100010.$$

We say that the sequence of compositions $(\mathcal{C}_n)$ is (Y_n)*-consistent* if there is the equality in distribution

$$\mathcal{C}_n^- \stackrel{d}{=} \mathcal{C}_{n-1} \quad \text{for every} \quad n = 2, 3, \dots \tag{2}$$

where $\mathcal{C}_n^-$ is $\mathcal{C}_n$ reduced by deletion of the ball in place Y_n. Then, by Kolmogorov's extension theorem, the sequence $(\mathcal{C}_n)$ can be realised jointly with (Y_n) on a common probability space, so that the equality in (2) also holds almost surely. We say that such a realisation of $(\mathcal{C}_n)$ is *strong* (Y_n)*-consistent*. We are particularly concerned with the operations of *uniform, left* and *right* reduction corresponding to Y_n with uniform distribution on $[n]$, to $Y_n \equiv 1$, and to $Y_n \equiv n$. That is, removal of a ball picked uniformly at random, or the left-most ball, or the right-most ball. So we may call $(\mathcal{C}_n)$ *uniform-, left-* or *right-consistent* as the case may be. Note that $(\mathcal{C}_n)$ is left-consistent iff $(\mathcal{C}_n^{\leftarrow})$ is right-consistent.

There is an obvious bijection between sequences of distributions of $\mathcal{C}_n$ which are right-consistent and probability distributions of infinite binary sequences $(\xi_1, \xi_2, \dots)$ with $\xi_1 = 1$: a strong right-consistent realisation of $\mathcal{C}_n$ in binary notation is the truncation $(\xi_1, \xi_2, \dots, \xi_n)$ of the infinite binary sequence. An alternate representation is obtained by replacing (ξ_i) by the random set of positive integers $\{i \geqslant 1 : \xi_i = 1\}$. Thus right-consistent sequences of compositions may be identified with random subsets of positive integers which contain 1.

A uniform consistent sequence of random compositions $(\mathcal{C}_n)$ is also called a *composition structure* [6, 9]. The corresponding sequence of random partitions $(\mathcal{C}_n^{\downarrow})$ is then a *partition structure* in the sense of Kingman (see [15] for a survey and background). That is to say,

$$(\mathcal{C}_n^{\downarrow -})^{\downarrow} \overset{d}{=} \mathcal{C}_{n-1}^{\downarrow} \quad \text{for every} \quad n = 2, 3, \dots . \tag{3}$$

where the left side is the decreasing rearrangement of a reduction of $\mathcal{C}_n^{\downarrow}$ by removal of uniformly chosen random ball. Kingman gave a representation of partition structures which Gnedin refined as follows:

Theorem 1 [6]**.** *Let $(\mathcal{C}_n)$ be a composition structure. Then there exists a random closed subset Z of $[0,1]$ such that a strong uniform consistent realisation of $(\mathcal{C}_n)$ can be constructed as follows: let (U_i) be a sequence of independent uniform $[0,1]$ variables, independent of Z, and let $\mathcal{C}_n$ be the sequence of sizes of equivalence classes among $U_1, \dots, U_n$, listed left to right, as these points are classified by the random equivalence relation $\sim$ induced by $U_i \sim U_j$ for $i \neq j$ if and only if U_i and U_j fall in the same interval component of $[0,1] \backslash Z$.*

Remark. For consistency with further considerations in this paper we include the point 1 in Z only if 1 is not an isolated point in Z. Thus, if the rightmost interval of $[0,1] \setminus Z$ exists, it is semiopen.

We are most interested in the case when Z is *light*, meaning that the Lebesgue measure of Z equals 0 almost surely. The collection of component intervals of $[0,1] \backslash Z$ then defines a *random interval partition* of $[0,1]$, that is, a collection of open subintervals of $[0,1]$, the sum of whose lengths is 1. The collection of lengths of component intervals of $[0,1] \backslash Z$, suitably indexed, is then a *random discrete distribution*, as studied in [17]. We regard the composition structure $(\mathcal{C}_n)$ as a combinatorial representation of either Z or its associated interval partition, just as the partition structure $(\mathcal{C}_n^{\downarrow})$ may be regarded as a combinatorial representation of the unordered collection of interval lengths.

In a series of previous papers [3, 6–12, 14, 17], we have studied the composition structures, partition structures, random interval partitions, and random discrete distributions corresponding to various random subsets of $[0,1]$ of particular interest. Here we tie together some threads from these previous studies, to show how various analytic properties of the random subset Z of $[0,1]$, which are natural from the perspective of continuous parameter stochastic processes, correspond to various combinatorial properties of the associated composition structure$(\mathcal{C}_n)$.

A random closed subset S of $\mathbb{R}_+$ is called *self-similar* (or *scale-invariant*) if

$$S \stackrel{d}{=} c\,S \text{ for all } c > 0. \tag{4}$$

In Sec. 2, we establish the following result, which generalizes a construction introduced in [14] in the case discussed in Example 2 below.

Theorem 2. *For a sequence of distributions of random compositions* $(\mathcal{C}_n)$*, the following two conditions are equivalent:*

- $(\mathcal{C}_n)$ *is both uniform consistent and right-consistent.*
- $(\mathcal{C}_n)$ *can be derived by uniform sampling from* $S \cap [0,1]$ *for some self-similar random closed subset* S *of* $\mathbb{R}_+$.

When these conditions hold, a strong right-consistent version of $(\mathcal{C}_n)$ *can be constructed as follows: independent of* S*, let* $\epsilon_1 < \epsilon_2 < \ldots$ *be the atoms of a homogeneous Poisson point process (henceforth PPP) on* $\mathbb{R}_+$*, and let the binary representation of* $\mathcal{C}_n$ *be the first* n *digits of* (ξ_i) *defined by* $\xi_1 = 1$ *and for* $j > 1$

$$\xi_j = 1([\epsilon_{j-1}\,,\,\epsilon_j] \cap S \neq \emptyset).$$

Note that there are two quite different realisations of $(\mathcal{C}_n)$, one that is strong uniform consistent, obtained by uniform sampling from $S \cap [0,1]$ as described in Theorem 1, and one that is strong right-consistent, obtained by Poisson sampling from $S \subseteq [0,\infty[$ as described in Theorem 2. Obviously, it is impossible to construct $(\mathcal{C}_n)$ to be simultaneously strong uniform consistent and strong right-consistent.

Example 1 [2, 5, 8, 9]**.** Let $(\xi_1, \xi_2, \ldots)$ be a random Bernoulli string with independent digits and distribution

$$\mathbb{P}(\xi_j = 1) = 1 - \mathbb{P}(\xi_j = 0) = \theta/(j + \theta - 1),$$

where $0 < \theta < \infty$ is a parameter. Let $\mathcal{C}_n$ be encoded by the first n digits, so $(\mathcal{C}_n)$ is strong right-consistent by construction. It is elementary that

$\mathcal{C}_n$ is distributed according to the formula

$$\mathbb{P}(\mathcal{C}_n = (\lambda_1, \dots, \lambda_\ell)) = \frac{\theta^\ell n!}{(\theta)_n} \prod_{j=1}^{\ell} \frac{1}{\Lambda_j}, \tag{5}$$

where $(\theta)_n = \theta(\theta+1)\cdots(\theta+n-1)$ and $\Lambda_j = \lambda_1 + \cdots + \lambda_j$. This is a variant of the Ewens sampling formula [5, 8], which gives the distribution of the sizes of blocks, in reverse size-biased order, of a Ewens partition of n with parameter θ. In the case $\theta = 1$, the sequence (ξ_i) results from encoding the cycle partition of a uniform random permutation of $[n]$ by Feller coupling [2]. This sequence also appears in the theory of extremes as the sequence of record indicators of independent identically distributed observations with continuous distribution [4]. The uniform consistency of $(\mathcal{C}_n)$ was observed in [5]. It is known that the random set Z in Kingman's representation is the restriction to $[0,1]$ of the self-similar random set S which is the union of $\{0\}$ and the set of points of a scale-invariant Poisson process on $[0, \infty[$, with intensity $\theta\, dx/x$, $x > 0$. Properties of this scale-invariant Poisson process are reviewed in [1]. Two trivial composition structures appear as limiting cases for $\theta \downarrow 0$ and $\theta \uparrow \infty$.

Example 2 [14, 9]**.** Let $\xi_j = 1$ $(R_k = j$ for some $k \geqslant 0)$ where $R_0 = 1$ and $R_k = 1 + X_1 + \cdots + X_k$ is the discrete renewal process derived from independent and identically distributed X_j with

$$\mathbb{P}(X_j = r) = (-1)^{r-1}\binom{\alpha}{r}, \quad r = 1, 2, \dots,$$

where $0 < \alpha < 1$. The corresponding right-consistent sequence of compositions has distribution

$$\mathbb{P}(\mathcal{C}_n = \lambda) = \lambda_\ell\, \alpha^{\ell-1} \prod_{j=1}^{\ell} \frac{(1-\alpha)_{\lambda_j - 1}}{\lambda_j!}. \tag{6}$$

That this sequence of compositions $(\mathcal{C}_n)$ is both right-consistent and uniform consistent was shown in [14], where the two different strong consistent constructions of Theorems 2 and 1 were given for S the self-similar zero set of a Bessel process of dimension $2 - 2\alpha$, and $Z = S \cap [0,1]$. The trivial cases appear again as limits for $\alpha = 0$ or 1.

It was shown by J. Young [19] that no other choice of distribution either for a sequence of independent Bernoulli variables (as in Example 1), or for a renewal sequence with independent spacing between 1's (as in

Example 2) yields a right-consistent sequence of compositions $(\mathcal{C}_n)$ such that $(\mathcal{C}_n^{\downarrow})$ is a partition structure. As the latter condition is weaker than the uniform consistency of $(\mathcal{C}_n)$, no more right-consistent composition structures can be obtained from these constructions. However, we will show that a construction adopted from [17, 19, 9] allows an interesting extrapolation of the above examples to obtain a right-consistent composition structure with two parameters (α, θ) with $0 \leqslant \alpha < 1$ and $\theta > -\alpha$, corresponding to the two-parameter Ewens–Pitman family of partition structures.

To emphasise the general correspondence between composition structures and random sets provided by Theorem 1, we use terminology for composition structures to reflect properties of their associated random sets. So we prefer the term *self-similar* rather than right-consistent for the composition structure $(\mathcal{C}_n)$ obtained by uniform sampling from $S \cap [0,1]$ for a self-similar random set S. In [9] we described the random sets associated with composition structures $(\mathcal{C}_n)$ with the following *left-regenerative* property: for every n and $1 \leqslant x \leqslant n$, conditionally given the leftmost part of $\mathcal{C}_n$ is x, the remaining composition of $n-x$ is a distributional copy of $\mathcal{C}_{n-x}$. Here we find it more convenient to work with the *right-regenerative* property, defined in the same way with the rightmost part instead of the leftmost part. Evidently, $(\mathcal{C}_n)$ derived by uniform sampling from Z is right-regenerative if and only if $(\mathcal{C}_n^{\leftarrow})$ derived by uniform sampling from $1-Z$ is left-regenerative. So the main result of [9] can be restated as follows: a composition structure $(\mathcal{C}_n)$ is right-regenerative if and onle if $(\mathcal{C}_n)$ is derived by uniform sampling from e^{-W} for W a regenerative random subset of $[0,\infty[$. Here we distinguish a class of *Markov* composition structures such that the binary representation of $\mathcal{C}_n$ has 1's at the places visited by a decreasing Markov chain on $[n]$ with some transition matrix q which does not depend on n, and some initial distribution $q_*(n,\cdot)$ on $[n]$. These turn out to be derived by uniform sampling from e^{-W} for W a delayed regenerative random subset of $[0,\infty[$. In the special case when W is a *stationary* regenerative set, e^{-W} is the restriction to $[0,1]$ of a self-similar random subset of $[0,\infty[$. The *self-similar Markov* compositions so obtained turn out to be those whose infinite binary representation has 1's at the places visited by an increasing Markov chain on $\mathbb{N}$. Finally, extending our study in [9], we introduce self-similar Markov composition structures associated with the two-parameter Ewens–Pitman family of partition structures.

2. Self-similar composition structures

2.1. Composition probability function

The distribution of a random composition $\mathcal{C}_n$ of integer n is a nonnegative function

$$p(\lambda) = \mathbb{P}(\mathcal{C}_n = \lambda)$$

on compositions λ of n which satisfies $\sum_{\lambda:|\lambda|=n} p(\lambda) = 1$. Here and henceforth $|\lambda|$ denotes the sum of parts of a composition λ. For a general sequence of random compositions $(\mathcal{C}_n)$ these marginal distributions are described by a *composition probability function* (CPF) defined for all compositions of integers. A sequence of compositions is (Y_n)-consistent if and only if the CPF satisfies a linear recurrence of the form

$$p(\lambda) = \sum_{\mu:|\mu|=|\lambda|+1} p(\mu)\, \varkappa(\mu, \lambda), \tag{7}$$

where $\varkappa(\mu, \lambda)$ for μ with $|\mu| = n$ is a matrix describing the transition probabilies from compositions of n to compositions of $n-1$ determined in the balls-in-boxes representation by removal of a ball from place Y_n. See [6, 9] for details in the case of uniform consistency when Y_n has uniform distribution on $[n]$. For $(\mathcal{C}_n)$ that is right-consistent, the recurrence is just linear relation

$$p(\lambda_1, \ldots, \lambda_\ell) = p(\lambda_1, \ldots, \lambda_\ell + 1) + p(\lambda_1, \ldots, \lambda_\ell, 1). \tag{8}$$

2.2. Proof of Theorem 2

We start by remarking that the distribution of a self-similar set S is uniquely determined by the distribution of its restriction Z to $[0,1]$, which satisfies the condition equivalent to (4):

$$c\,(S \cap [0,1]) \stackrel{d}{=} S \cap [0,c], \quad \text{for} \quad 0 < c < 1. \tag{9}$$

This follows from the known fact that the distribution of a *stationary* set $W \subset \mathbb{R}$ (invariant under shifts) is determined by the distribution of $W \cap \mathbb{R}_+$, and we can transform a self-similar S into a stationary set $W := -\log S$.

For $n = 1, 2, \ldots$, let $Z_n \subset [0,1[$ be a finite set encoding $\mathcal{C}_n$ via the correspondence $(\lambda_1, \ldots, \lambda_\ell) \to \{0, \Lambda_1/n, \ldots, \Lambda_{\ell-1}/n\}$ for $\Lambda_j = \lambda_1 + \ldots + \lambda_j$. Assuming now that we are working with a strong uniform consistent realisation of $(\mathcal{C}_n)$, by the law of large numbers [6] the Hausdorff distance between $Z_n \cup \{1\}$ and $Z \cup \{1\}$ goes to 0 with probability 1. The Hausdorff distance between Z_n and Z also goes to 0. This can be shown by considering the last block of the composition, which has a positive frequency if and only if 1 is not an accumulation point for Z. In this sense, $Z_n \to Z$ a.s., hence also $Z_{\lfloor nx \rfloor} \to Z$ a.s. for every $x \in]0,1[$. Translating the truncation property in terms of Z_n's, we obtain

$$Z_n \cap \left[0\,,\,\frac{\lfloor nx \rfloor}{n}\right[\stackrel{d}{=} \frac{\lfloor nx \rfloor}{n} Z_{\lfloor nx \rfloor}.$$

For $n \to \infty$, the left side converges to $Z \cap [0,x[$ a.s., while the right side converges to $x\,Z$ a.s., hence the limits must have the same distribution. This means that Z is self-similar, by (9). The strong right-consistent representation is obtained by noting that the scaling $(S, \epsilon_1, \ldots, \epsilon_n) \to (S/\epsilon_{n+1}, \epsilon_1/\epsilon_{n+1}, \ldots, \epsilon_n/\epsilon_{n+1})$ transforms S to a copy of itself (by self-similarity) and maps the first n Poisson points to the increasing sequence of n uniform order statistics. $\square$

2.3. Some definitions

We call Z *heavy* if Z has positive Lebesgue measure with nonzero probability, and we call Z *light* otherwise. The set Z can be discrete (as in Example 1) or perfect (as in Example 2) or neither discrete nor perfect.

For $x \in \mathbb{R}_+$, introduce

$$G_x := \sup(Z \cap [0,x]), \quad A_x := x - G_x, \quad D_x := \inf(Z \cap]x, \infty[).$$

The *age process* $(A_x\,,\, x \geq 0)$ uniquely determines Z. In the event $x \in Z$ we have $G_x = x$ and $A_x = 0$, while in the event $x \notin Z$ the point x is covered by an open *gap* $]G_x, D_x[\subset \mathbb{R}_+ \setminus Z$. The interval $]G_1, 1[$ of length A_1 is called the *meander*. If Z is heavy, the meander may be empty with positive probability (then $A_1 = 0$), while for light Z the meander is nondegenerate (and $A_1 > 0$ a.s.).

3. Block counts, meander, and the tagged interval

3.1. The structural distribution

For a random closed set $Z \subset [0,1]$ and a uniform random point U independent of Z, let V be the size of the gap in Z covering U in the case $U \in [0,1] \setminus Z$, and let $V = 0$ in the case $U \in Z$. The gap covering a random point is sometimes called the *tagged* interval.

Let (V_j) be the decreasing sequence of lengths of gaps comprising $[0,1] \setminus Z$, so that $\sum_j V_j \leq 1$ and $1 - \sum_j V_j$ is the Lebesgue measure of Z. If all the V_j's are pairwise distinct with probability one, then

$$\mathbb{P}(V = V_j \,|\, V_1, V_2, \ldots) = V_j, \quad \mathbb{P}(V = 0 \,|\, V_1, V_2, \ldots) = 1 - \sum_j V_j\,.$$

So V may be called a *size-biased pick* from the sequence of lengths.

Suppose now that the composition structure $(\mathcal{C}_n)$ is derived by uniform sampling from Z. The distribution of V is called the *structural distribution* (of $(\mathcal{C}_n)$, or of the associated partitition structure, or of the associated random discrete distribution of interval lengths). Recall that p denotes the CPF of $(\mathcal{C}_n)$. Observe that

$$p(n) = \mathbb{E}V^{n-1}$$

because $\mathcal{C}_n$ equals the one-part composition (n) when $n-1$ further uniform points hit the interval of length V found by U. Other relations of this type are

$$\mu_{n,r} := \mathbb{E}K_{n,r} = \binom{n}{r}\mathbb{E}\left(V^{r-1}(1-V)^{n-r}\right), \quad 1 \le r \le n, \tag{10}$$

$$\mu_n := \mathbb{E}K_n = \sum_r \mu_{n,r} = \mathbb{E}\left(\frac{1-(1-V)^n}{V} 1(V>0)\right) + n\,\mathbb{P}(V=0), \tag{11}$$

where $K_{n,r}$ is the number of parts of $\mathcal{C}_n$ of size r, and $K_n = \sum_r K_{n,r}$ is the number of parts of $\mathcal{C}_n$. Note that $(K_{n,r}, 1 \leqslant r \leqslant n)$ is a standard encoding of $\mathcal{C}_n^{\downarrow}$, the random partition of n induced by $\mathcal{C}_n$.

Theorem 3 [17]. *Suppose that $S \subset \mathbb{R}_+$ is self-similar, and let $Z := S \cap [0,1]$. Let A_1 be the length of the meander interval of Z, that is the rightmost gap in $[0,1]\backslash Z$, with $A_1 = 0$ if $1 \in Z$, and let V be the length of the component interval of $[0,1]\backslash Z$ which contains U independent of Z, with $V = 0$ if $U \in Z$. Then A_1 has the same distribution as V.*

Proof. Let ρ_n be the rightmost in the sample of n uniform points. Given A_{ρ_n}, the remaining $n-1$ sample points with probability $(A_{\rho_n}/\rho_n)^{n-1}$ fall in the same gap of Z as ρ_n. By self-similarity, $A_{\rho_n}/\rho_n \overset{d}{=} A_1$. So, the probability that all n points fall in the same gap is

$$p(n) = \mathbb{E}(A_w/w)^{n-1} = \mathbb{E}A_1^{n-1}.$$

Comparing with $p(n) = \mathbb{E}V^{n-1}$, we arrive at the conclusion, since a probability distribution on $[0,1]$ is determined by its moments. □

Note that the event $(A_1 = V)$ has probability $\mathbb{E}A_1 = \mathbb{E}V = p(2)$. Pitman and Yor (see [17]) went further to distinguish a *strong sampling property* for the meander

$$\mathbb{P}(A_1 = V_k \mid V_1, V_2, \dots) = V_k\,, \ \ k = 1, 2, \dots \tag{12}$$

meaning the condition that the meander length is a size-biased pick from all lengths. This property holds in some cases (e.g., for Z in Examples 1 and 2, and in the setup of Theorem 14 below), but does not hold in general.

3.2. The last part and the tagged part of composition

Theorem 3 implies that for self-similar composition structure, as $n \to \infty$, the frequency of the last block of $\mathcal{C}_n$ has approximately the same distribution as the frequency of the block selected by a size-biased pick. A stronger fact is true: a similar identity holds for each n, and not only asymptotically. This was already observed in [14, Proposition 11 (i)] in the case of renewal strings in Example 2. Intuitively, since both uniform and right reduction transform $\mathcal{C}_n$ into a composition with the same distribution, it is natural to expect that the sizes of the reduced parts have the same distribution.

Theorem 4. *For a composition structure $(\mathcal{C}_n)$, let P_n denote the size of a random part of $\mathcal{C}_n$ which, conditionally given $\mathcal{C}_n$, is selected with probability proportional to the size. Let L_n be the size of the last part of $\mathcal{C}_n$. If $(\mathcal{C}_n)$ is right-consistent, then $P_n \stackrel{d}{=} L_n$ for all n.*

This follows immediately from the following lemma.

Lemma 5. *Let $\mathcal{C}_n$ and $\mathcal{C}_{n-1}$ be two random compositions of n and $n-1$, respectively, defined on a common probability space in such a way that $\mathcal{C}_{n-1}$ is obtained from $\mathcal{C}_n$ by the removal of a single ball in the balls-in-boxes representation. Let $(\omega_{n,r}, 1 \le r \le n)$ be the distribution of the number of balls in the same box of $\mathcal{C}_n$ as the ball removed, and let $\mu_{n,r}$ and $\mu_{n-1,r}$ be the expected numbers of boxes containing r balls for $\mathcal{C}_n$ and $\mathcal{C}_{n-1}$, respectively, as above in (10). Then the distribution $(\omega_{n,r}, 1 \le r \le n)$ is determined by the distributions of the partitions generated by $\mathcal{C}_{n-1}$ and $\mathcal{C}_n$ according to the formulas $\omega_{n,n} = \mu_{n,n}$ and*

$$\omega_{n,r} - \omega_{n,r+1} = \mu_{n,r} - \mu_{n-1,r} \quad (1 \le r \le n-1).$$

Proof. Follow the evolution of $K_{n,r}$ as n varies. This variable increases by 1 when a ball is chosen in a box of $r+1$ balls (which is impossible for $r = n$), and decreases by 1 when a ball is chosen in a box of r balls. The probabilities of these events are $\omega_{n,r+1}$ and $\omega_{n,r}$, respectively. In all other cases the sampling does not affect $K_{n,r}$. The formula for expected increments follows. □

Note that Theorem 3 follows from Theorem 4 by the law of large numbers. A discrete analogue of (12) holds for compositions in Examples 1 and 2: conditionally given $\mathcal{C}_n^{\downarrow}$, the last part L_n of $\mathcal{C}_n$ is a size-biased pick from all parts.

3.3. A characterisation of structural distributions

The following characterisation of structural distributions is a minor extension of [17, Condition 1] to include the heavy case.

Theorem 6. *The structural distribution of the interval partition derived from a self-similar random set S has the form*

$$\mathbb{P}(V \in \mathrm{d}x) = \frac{\widetilde{\nu}[x,1]}{(\mathtt{d}+\mathtt{m})\,(1-x)}\,\mathrm{d}x + \frac{\mathtt{d}}{\mathtt{d}+\mathtt{m}}\,\delta_0(\mathrm{d}x), \tag{13}$$

where $\widetilde{\nu}$ is a measure on $]0,1]$ satisfying

$$\mathtt{m} := \int_0^1 |\log(1-x)|\widetilde{\nu}(\mathrm{d}x) < \infty$$

and $\mathtt{d}$ is a nonnegative constant. Thus, the structural distribution may have an atom at 0, and otherwise has a density $\phi(x)$, $0 < x \le 1$, such that $(1-x)\phi(x)$ is decreasing. The data $(\mathtt{d}, \widetilde{\nu})$ are determined uniquely up to a positive factor.

Proof. Assume the normalisation $\mathtt{m} = 1$. Let $W = -\log S$ be stationary and X be a random variable whose distribution coincides with the conditional distribution of the size of the gap of W covering 0 given this size is positive. By the ergodic theorem, the part of the gap on the positive halfline is distributed like XU, with U uniform $[0,1]$ independent of X. The conditional distribution of A_1 given $A_1 > 0$ is then the same as for e^{-XU}, which implies along the lines of the argument in [17, Section 4] that A_1 has a density written as $\widetilde{\nu}[x,1]/(1-x)$. The unconditional distribution in the form (13) follows by defining $\mathtt{d}$ from

$$\mathbb{P}(A_1 = 0) = \mathbb{P}(0 \in W) = \frac{\mathtt{d}}{1+\mathtt{d}}$$

(where the middle term is the the long-run Lebesgue measure of W per unit length). □

The structural distribution also accounts for some functionals of self-similar composition structures which involve the ordering of parts. For a self-similar composition structure $(\mathcal{C}_n)$ with binary representation $(\xi_1\xi_2\ldots)$, define the *potential function*

$$g(j) = \mathbb{P}(\xi_j = 1).$$

In terms of balls-in-boxes, this is the probability, for each $n \geqslant j$, that the jth ball of $\mathcal{C}_n$ falls in a different box from its predecessor. Note that for a composition structure which was not right-consistent, the analogous quantity would typically depend on n as well as j. In terms of $\mu_n := \mathbb{E}K_n$ and the moments of the structural distribution, we read from (11) that

$$g(j) = \mu_j - \mu_{j-1} = \mathbb{E}(1-V)^{j-1}, \quad \mu_n = \sum_{j=1}^{n} g(j)$$

(where $\mu_0 = 0$).

Nacu [13], proved that the probability law of the general exchangeable partition of $\mathbb{N}$ is uniquely determined by the distribution of the sequence (I_j) of indicators of minimal elements of the blocks (so $I_7 = 1$ means that 7 is the minimal element in some block). In terms of Kingman's representation, $I_j = 1$ each time U_j discovers a new gap in Z or hits Z. For the Ewens composition structure of Example 1, (I_j) has the same distribution as (ξ_j). For a general self-similar composition structure, the sequences are differently distributed (as, *e.g.*, in Example 2), but the right-consistency of $\mathcal{C}_n$ implies that

$$\mathbb{P}(I_j = 1) = \mathbb{P}(\xi_j = 1) = g(j)$$

because

$$I_1 + \cdots + I_n = \xi_1 + \cdots + \xi_n = K_n,$$

the number of parts of $\mathcal{C}_n$.

3.4. A fragmentation product

The following operation on self-similar sets generalises the one found in [9, 17, 19]. Let $Z \subset \mathbb{R}_+$ be self-similar and independent of Z. Let (M_j) be independent copies of the same random closed set $M \subset [0,1]$. For each gap in Z with left-point $z_j \in Z$ and size s_j fit the set $z_j + s_j M_j$ in this gap, and take the union of Z and all these scaled shifted copies of M.

Then the result $Z \otimes M$ (read *Z fragmented by M*) is easily shown to be self-similar. For example, when $M = \{1/2\}$, the set $Z \otimes M$ is obtained by adding the midpoint for each gap in Z.

The operation has an analogue in terms of composition structures (as in [19]). For two composition structures $(\mathcal{C}_n)$ and $(\mathcal{C}'_n)$, for each n, break the generic part of $\mathcal{C}_n$, say r, into smaller parts according to an independent copy of $\mathcal{C}'_r$. The resulting sequence of compositions is a right-consistent composition structure provided $(\mathcal{C}_n)$ is so, and the corresponding self-similar random set is the fragmentation product $Z \otimes M$ of the sets in Kingman's representation of $(\mathcal{C}_n)$ and $(\mathcal{C}'_n)$.

4. Markovian composition structures

4.1. Decrement matrices

The following extension of the concept of a *regenerative composition structure* introduced in [9] extends our study in that paper and prepares for the results in the next section.

Definition 7. A composition structure $(\mathcal{C}_n)$ is called *Markovian* if for some infinite transition probability matrices

$$(q(n:m),\, 1 \le m \le n < \infty) \quad \textit{and} \quad (q_*(n:m),\, 1 \le m \le n < \infty)$$

the distribution of each $\mathcal{C}_n$ is given by the product formula

$$p(\lambda) = q_*(n:\lambda_\ell) \prod_{k=1}^{\ell-1} q(\Lambda_k \,:\, \lambda_k), \tag{14}$$

where $\lambda = (\lambda_1, \ldots, \lambda_\ell)$ is a composition of n, and $\Lambda_k = \lambda_1 + \ldots + \lambda_k$ for $k \le \ell$.

Formula (14) has the following interpretation. Imagine a *decreasing* time-homogeneous Markov chain $Q_n^{\downarrow} = (Q_{n,t}^{\downarrow},\, t = 0, 1, \ldots)$ with state-space $\{1, 2, \ldots, n\}$ and terminal absorbing state 1. The chain has initial distribution

$$\mathbb{P}(Q_{n,0}^{\downarrow} = j) = q_*(n : n - j + 1)\,, \quad j = 1, \ldots, n,$$

and it jumps from state j $(2 \le j \le n)$ to i $(1 \le i < j)$ with probability $q(j-1 : j-i)$. We call q_* and q *decrement matrices.* In these terms, a random composition of n can be identified with a path of $Q_n^{\downarrow}$, i.e., the binary representation of $\mathcal{C}_n$ (for fixed n) is obtained by writing 1's in the positions visited by $Q_n^{\downarrow}$.

In the case $q = q_*$, the formula (14) defines a regenerative composition structure, as introduced in [9]. As mentioned in the Introduction, to fit in the present framework, the convention in that paper regarding the ordering of blocks should be reversed.

Lemma 8. *For a Markovian composition structure, we have*

$$q(n:r) = \frac{r+1}{n+1}\, q(n+1:r+1) + \frac{n+1-r}{n+1}\, q(n+1:r) + \frac{1}{n+1}\, q(n+1:1)\, q(n:r), \tag{15}$$

$$q_*(n:r) = \frac{r+1}{n+1}\, q_*(n+1:r+1) + \frac{n+1-r}{n+1}\, q_*(n+1:r) + \frac{1}{n+1}\, q_*(n+1:1)\, q(n:r). \tag{16}$$

Conversely, if two nonnegative matrices q_* and q satisfy these recursions and $q_*(1:1) = q(1:1) = 1$, then they define a Markovian composition structure via (14).

Proof. The recursions follow from (14) and uniform consistency. When a composition is reduced by sampling, the last block either remains unaltered or, in the case the last block is a singleton and gets deleted, coincides with the second-last block, whence (16).

The first recursion is familiar from [9, 10], but proving it under the more general assumption (14) requires more algebra. For (a, b, c) a composition of n, use uniform consistency to obtain

$$p(a,b,c) = \frac{1}{n+1}p(1,a,b,c) + \frac{a+1}{n+1}p(a+1,b,c) + \frac{1}{n+1}p(a,1,b,c) + \frac{b+1}{n+1}p(a,b+1,c) + \frac{1}{n+1}p(a,b,1,c) + \frac{c+1}{n+1}p(a,b,c+1) + \frac{1}{n+1}p(a,b,c,1).$$

Group the first three terms in the right-hand side as

$$\frac{a+1}{n+1}q_*(n+1:c)q(n+1-c:b)p(a)$$

and factor all other terms using (14). Factor the left side as

$$p(a,b,c) = q_*(n:c)q(a+b:b)p(a)$$

and express $q_*(n:c)$ through $q(n+1:\cdot)$ using (16). Cancelling common terms and factors yields (15).

The converse is checked as in [9, Proposition 3.3]. □

4.2. Kingman's representation

Let $(Y_t, t \geq 0)$ be a *subordinator* (with $Y_0 = 0$), meaning an increasing Lévy process. Let $0 \leq X \leq \infty$ be a random variable, independent of (Y_t) and satisfying $\mathbb{P}(X < \infty) > 0$. We call the process $(X + Y_t,\, t \geq 0)$ a *delayed subordinator*, and call its closed range W a *delayed regenerative set*. The distribution of W determines that of X (because $X = \min W$) and determines the Lévy parameters $(\nu, \mathtt{d})$ up to a positive factor (since given $X < \infty$, the set $W - X$ is regenerative). Introduce the Lévy–Khintchine exponent

$$\Phi(s) = \mathtt{d}s + \int_0^\infty (1 - e^{-sy})\,\nu(\mathrm{d}y),\ \ s \geq 0, \tag{17}$$

its two-parameter extension

$$\Phi(n:m) = \binom{n}{m} \sum_{j=0}^{m} (-1)^{j+1\binom{m}{j}} \Phi(n-m+j),\ \ 1 \leq m \leq n, \tag{18}$$

and the moments

$$\Psi(n:m) = \binom{n}{m} \mathbb{E}\left((A_1^m(1-A_1)^{n-m}\right),\ \ 0 \leq m \leq n, \tag{19}$$

where $A_1 = 1 - \exp(-X)$.

Theorem 9. *A composition structure $(\mathcal{C}_n)$ is Markovian if and only if it can be derived by uniform sampling from $Z = \exp(-W)$, with W a delayed regenerative set. Explicitly, the distribution of $(\mathcal{C}_n)$ is given by the product formula with decrement matrices*

$$q(n:m) = \frac{\Phi(n:m)}{\Phi(n)}, \tag{20}$$

$$q_*(n:m) = \Psi(n:0)q(n:m) + \Psi(n:m). \tag{21}$$

Proof. The argument for the "if" part follows the same line as in [9, Theorem 5.2 (i)]. For the "only if" part let $\mathcal{C}_n$ be derived by uniform sampling from the random closed set $Z \subset [0,1]$. Assume first that $G_1 <$

1 a.s. for $G_1 = \sup Z \cap [0,1[$. Let L_n be the last part of $\mathcal{C}_n$. Given $n - L_n = m$, let $\mathcal{C}'_m$ be a composition of m obtained by deleting the last part of $\mathcal{C}_n$. Note that this definition does not depend on $n > m$ and that by Lemma 8 and [9, Proposition 3.3], hence $(\mathcal{C}'_m)$ is a regenerative composition structure.

Let Z'_m be discrete random sets encoding $\mathcal{C}'_m$, as in the proof of Theorem 2, but with 1 appended to Z'_m. The set Z_n (not containing 1) encoding $\mathcal{C}_n$ can be represented as

$$Z_n \stackrel{d}{=} \left(\frac{n - L_n}{n}\right) Z'_{n-L_n}$$

where L_n and (Z'_m) are independent. By the law of large numbers [6], Z_n converge to Z, while by [9, Theorem 5.2 (ii)] Z'_m converge to some set $Z' = \exp(-W')$ with regenerative $W' \subset [0,\infty]$. As $n \to \infty$, the law of large numbers ensures that $1 - L_n/n \to G_1$ a.s., hence in the limit we have

$$Z \stackrel{d}{=} G_1 \, Z'$$

where G_1 and the set Z' are independent. Hence the set $W = -\log Z$ is delayed regenerative.

The case $\mathbb{P}(G_1 = 1) > 0$ is treated similarly. This can be viewed as a mixture (over q_*) of the trivial one-block composition structure and another Markovian one. □

5. Self-similar Markov composition structures

5.1. Markov sequences

Let $Q^\uparrow = (Q^\uparrow_t, t = 0, 1, \dots)$ be a time-homogeneous *increasing* Markov chain with the state-space $\{1, 2, \dots\}$ and the initial state $Q_0 = 1$. Define a string $\xi_1\xi_2\dots$ by identifying the positions of 1's with the sequence of sites visited by $Q^\uparrow$:

$$\xi_j = 1(Q_t = j \quad \text{for some} \quad t).$$

This defines a right-consistent sequence of compositions $(\mathcal{C}_n)$, so that each $\mathcal{C}_n$ the encodes the path of $Q^\uparrow$ killed before crossing the level n. We will consider such compositions which are also uniform consistent, in which case (in view of Theorem 2) we will call $(\mathcal{C}_n)$ a *self-similar Markov* composition structure.

Bernoulli sequences in Example 1 yield self-similar Markov composition structures. Another instance is the renewal sequence in Example 2, with $Q^\uparrow$ a discrete renewal process.

As the terminology is meant to suggest, self-similar Markov composition structures are Markov in the sense of Sec. 4. To see this, for each n consider a Markov chain $Q_n^{\uparrow}$ with state-space $[n] \cup \{\infty\}$ such that $Q_n^{\uparrow}$ coincides with $Q^{\uparrow}$ as long as the latter stays in $[n]$, but jumps to ∞ at the moments when $Q^{\uparrow}$ exits $[n]$. Let $Q_n^{\downarrow}$ be a time-reversal of $Q_n^{\uparrow}$, so that $Q_{n,0}^{\downarrow}$ has the same distribution as the value of $Q_n^{\uparrow}$ immediately before exiting $[n]$. The chains $Q_n^{\downarrow}$, $n = 1, 2, \ldots$, are *coherent* in the sense that, for $m \le n$, when $Q_n^{\downarrow}$ enters $[m]$ its state has the same distribution as $Q_{m,0}^{\downarrow}$. Conversely, if the chains $(Q_n^{\downarrow})$ are coherent, their reversals $(Q_n^{\uparrow})$ can be organised in a single "super-chain" $Q^{\uparrow}$ with state-space $\{1, 2, \ldots\}$. So, this property distinguishes the self-similar Markov case within the general Markov case. Another feature characterising the coherent sequence is that there is a common potential function: for all $n \ge m$ the probability that $Q_n^{\downarrow}$ visits state m does not depend on n.

We recall that a *stationary regenerative set* [18] is the range of a process $(X + Y_t\,,\, t \ge 0)$ where (Y_t) is a subordinator with Lévy measure satisfying

$$\mathtt{m} = \int_0^\infty y\,\nu(\mathrm{d}y) < \infty, \tag{22}$$

$\mathtt{d} \ge 0$ is some drift coefficient, X is independent of (Y_t) and has distribution

$$\mathbb{P}(X \in \mathrm{d}y) = \frac{\nu[y,\infty]}{\mathtt{d}+\mathtt{m}}\,\mathrm{d}y + \frac{\mathtt{d}}{\mathtt{d}+\mathtt{m}}\delta_0(\mathrm{d}y). \tag{23}$$

Thus, $(X + Y_t)$ is a delayed subordinator with a special choice of distribution for X, to make the range stationary.

Theorem 10. *A composition structure $\mathcal{C}$ is self-similar Markov if and only if its associated self-similar set Z can be presented as $Z = \exp(-W)$ where W is a regenerative set with stationary delay.*

Proof. Follows by combining Theorems 2 and 9. □

We see that self-similarity of Markov composition structures imposes further constraints on the decrement matrices q and q_* in the product formula (14). Thus, q can be associated only with a finite-mean subordinator, and is given then by (20) with Φ as in (17) and (18). Similarly, q_* is given by (21) for Ψ as in (19), and A_1 having distribution

$$\mathbb{P}(A_1 \in \mathrm{d}x) = \frac{\widetilde{\nu}[x,1]\mathrm{d}x}{(\mathtt{d}+\mathtt{m})(1-x)} + \frac{\mathtt{d}}{\mathtt{d}+\mathtt{m}}\delta_0(\mathrm{d}x),$$

where $\widetilde{\nu}$ is the image of ν under $y \mapsto 1 - e^{-y}$. By Theorem 3, this is also the structural distribution of Z, and comparing with (13), we observe that the distribution is of exactly the same type as for the general self-similar Z according to Theorem 6. For the potential function there is a simple formula

$$g(j) = \frac{1}{\mathtt{d}+\mathtt{m}} \frac{\Phi(j-1)}{j-1} \quad \text{for} \quad j > 1, \quad g(1) = 1, \tag{24}$$

which appeared in [7, p. 86] in a special case, and the transition probabilities f of $Q^{\uparrow}$ are recovered from

$$q(j-1 : j-i) = \frac{f(j \mid i) g(i)}{g(j)}, \; 1 \le i < j.$$

The relation between a regenerative composition with decrement matrix q and the associated self-similar Markov composition structure with matrices q and q_* (with q_* given by (21)) is the combinatorial counterpart of the relation between a subordinator and its stationary version.

5.2. Arrangements

A difficult and interesting question is the relation between partition structures and their possible *arrangements* as composition structures with certain properties. Some aspects of this problem were treated in [9, 10]. Although we do not know a simple algorithm to check if the blocks of a given partition may be ordered to produce a self-similar Markov composition structure, we can show the uniqueness.

Proposition 11. *If a partition structure admits an arrangement as a self-similar Markov composition structure, then such arrangement is unique in distribution.*

The claim follows from the next lemma by recalling that the moments of the structural distribution $p(n)$ are determined by the associated partition structure.

Lemma 12. *For $(\mathcal{C}_n)$ a self-similar Markov composition structure, for each n the distribution of $\mathcal{C}_n$ is uniquely determined by the structural moments $p(1) = 1, p(2), \ldots, p(n+1)$.*

Proof. A binomial expansion in (10) shows that $\mu_{n,r}, 1 \le r \le n$, are computable from $p(1), \ldots, p(n)$. By Theorem 4 and the formula $q_*(n : r) = r\,\mu_{n,r}/n$, also $(q_*(n' : r), 1 \le r \le n' \le n+1)$ is computable from $p(1), \ldots, p(n+1)$. Applying (16), we see by induction that the minor of the decrement matrix $(q(n' : r), 1 \leq r \leq n' \leq n)$ is computable from $p(1), \ldots, p(n+1)$, which taken together with (14) proves the claim. $\Box$

For the regenerative case (when $q = q_*$) we have shown that only the moments $p(2), \ldots, p(n)$ are needed to recover the distribution of the composition of order n [9, Proposition 7.1]. The explicit formulas are rather involved already in that case.

While a self-similar Markov arrangement (if any) of a partition structure is unique, many self-similar composition structures may project onto the same partition structure. For example, for self-similar Z, $M \subset [0,1]$, and M' the reflection of M about $1/2$, both fragmentation products $Z \otimes M$ and $Z \otimes M'$ induce the same partition structure, but the composition structures are different unless $M \stackrel{d}{=} M'$. This implies nonuniqueness in the problem of binary representability of partition structures studied in [19].

6. The two-parameter family

We are interested in self-similar composition structures associated with the members of the two-parameter family of partition structures [15]. For the range of parameters $\theta > -\alpha$, $0 \leq \alpha < 1$, these partition structures may be introduced as follows.

Let (V_i) or $(V_i, i \in I)$ denote a *random discrete distribution*, that is, a collection of random variables indexed by i in some finite or countably infinite set I, with

$$V_i \geqslant 0 \quad \text{and} \quad \sum_i V_i = 1 \text{ almost surely} .$$

We use $\{V_i\}$ as an informal notation for multi-set of all nonzero values of V_i, without regard to how they are indexed by I. Formally, $\{V_i\}$ is encoded by the sequence $(\hat{V}_j, j = 1, 2, \ldots) := \text{RANK}(V_i)$ meaning that $(\hat{V}_j, j = 1, 2, \ldots)$ is the decreasing rearrangement of (V_i) with padding by zeros if necessary. Let us write simply

$$\{V_i\} \sim (\alpha, \theta) \tag{25}$$

if $\text{RANK}(V_i)$ has the Poisson–Dirichlet distribution with two parameters (α, θ), defined following [16, 15] as the distribution of $\text{RANK}(\widetilde{V}_i)$ where

$$\widetilde{V}_1 := W_1 \tag{26}$$

has beta$(1-\alpha, \alpha+\theta)$ distribution, and, for $i \geqslant 1$,

$$\widetilde{V}_i := (1-W_1)\cdots(1-W_{i-1})W_i$$

where W_i has beta$(1-\alpha, \alpha+i\theta)$ distribution and the W_i are independent. It is known [16] that if $\{V_i\} \sim (\alpha,\theta)$, then such $\widetilde{V}_i$ can be constructed by size-biased random permutation of $\{V_i\}$. Then $\widetilde{V}_1 = V_J$ for a random index J with

$$\mathbb{P}(J = j \,|\, \{V_i\}) = V_j$$

while

$$\{V_k^{\#}\} := \left\{\frac{V_i - V_N}{1-V_N},\ i \neq J\right\} \sim (\alpha, \alpha+\theta) \tag{27}$$

and

$$\{V_k^{\#}\} \text{ is independent of } \widetilde{V}_1.$$

Since $\{V_i\}$ can be measurably recovered from $\widetilde{V}_1$ and $\{V_k^{\#}\}$ as

$$\{V_i\} = \{\widetilde{V}_1 \cup \{1-\widetilde{V}_1)V_k^{\#}\}, \tag{28}$$

an immediate consequence is

Lemma 13 [16, Proposition 35]. *If $\widetilde{V}_1$ has beta$(1-\alpha,\alpha+\theta)$ distribution and $\{V_k^{\#}\} \sim (\alpha,\theta)$, and $\{V_i\}$ is defined by* (28), *then $(V_i) \sim (\alpha,\theta)$ and $\widetilde{V}_1$ is a size-biased pick from $\{V_k^{\#}\}$.*

In [9], we established that for $0 \leqslant \alpha < 1, \theta \geqslant 0$ a random discrete distribution $\{V_i\} \sim (\alpha,\theta)$ can be derived from the interval partition of $[0,1]$ associated with a unique regenerative composition structure. Specifically, the image of the Lévy measure of this (α,θ) regenerative composition structure under $y \mapsto 1-e^{-y}$ is the measure $\widetilde{\nu}$ on $]0,1]$ characterised by

$$\widetilde{\nu}[x,1] = x^{-\alpha}(1-x)^{\theta}, \tag{29}$$

and the decrement matrix is

$$q(n:r) = \binom{n}{r}\frac{(1-\alpha)_{r-1}}{(\theta+n-r)_r}\ \frac{(n-r)\alpha + r\theta}{n}. \tag{30}$$

By combining these known results we now obtain the following:

Theorem 14. *For $0 \leqslant \alpha < 1$, $\theta > 0$ let $Z = \exp(-W)$ where W is the stationary version of the regenerative set associated as above with an (α, θ) regenerative composition structure. Then Z is a self-similar Markov random set associated with an $(\alpha, \theta - \alpha)$ partition structure. The structural distribution of Z is beta$(1-\alpha, \theta)$, and Z has the strong sampling property* (12).

Proof. The structural distribution of Z is read from (29), (13), and (23). The construction of Z allows the application of Lemma 13, with θ replaced by $\theta - \alpha$, to deduce the other conclusions. □

Theorem 14 can also be derived more combinatorially as follows. Consider the Polya–Eggenberger distributions

$$q_{\alpha,\theta}(n:r) = \binom{n-1}{r-1} \frac{(\theta+\alpha)_{n-r}(1-\alpha)_{r-1}}{(\theta+1)_{n-1}}\ ,\ r = 1, \ldots, n$$

and define a function on compositions

$$\widehat{\pi}_{\alpha,\theta}(\lambda) = \prod_{k=1}^{\ell} q_{\alpha,\theta+(\ell-k)\alpha}(\Lambda_k : \lambda_k),\ \ \lambda = (\lambda_1, \ldots, \lambda_\ell), \tag{31}$$

where $\Lambda_k = \lambda_1 + \ldots + \lambda_k$. Formula (31) is the distribution of the (α, θ) partition structure with parts arranged from right to left in a size-biased order. The (α, θ)-partition structure is defined then by the partition probability function obtained by the symmetrisation of the CPF (see [10] for more details of this procedure):

$$\pi_{\alpha,\theta}(\lambda^{\downarrow}) = \sum_{\text{distinct } \sigma} \widehat{\pi}_{\alpha,\theta}(\lambda_\sigma), \tag{32}$$

where the summation extends over all *distinct* permutations $\lambda_\sigma = (\lambda_{\sigma(1)}, \ldots, \lambda_{\sigma(\ell)})$ of parts of composition λ. From (31) and (32) follows the recursion

$$\pi_{\alpha,\theta}(\lambda^{\downarrow}) = \sum_{\text{distinct } \lambda_j \in \lambda^{\downarrow}} q_{\alpha,\theta}(n : \lambda_j)\ \pi_{\alpha,\theta+\alpha}(\lambda^{\downarrow} - \lambda_j), \tag{33}$$

where $\lambda^{\downarrow}$ is a (ranked, unordered) partition of n and where $\lambda^{\downarrow} - \lambda_j$ is the partition $\lambda^{\downarrow}$ without part λ_j. Let $\mathcal{C}_n$ denote the self-similar Markov composition structure derived from $S = \exp(-W)$ as in the theorem.

Computing beta integrals to determine the distribution q_* of the last part of $\mathcal{C}_n$ according to (21), we obtain

$$q_*(n:r) = \binom{n}{r}\mathbb{E}A_1^r(1-A_1)^{n-r} + \mathbb{E}(1-A_1)^n q(n:r)$$

$$= \frac{\binom{n}{r}}{\mathrm{B}(1-\alpha,\theta)}\Big(\mathrm{B}(r-1+\alpha, n-r+\theta)$$

$$+\mathrm{B}(1-\alpha,n+\theta)\frac{(1-\alpha)_{r-1}}{(\theta+n-r)_r}\ \frac{(n-r)\alpha+\alpha\theta}{n}\Big)$$

which, upon simplification, shows that $q_* = q_{\alpha,\theta-\alpha}$. If the last part of $\mathcal{C}_n$ is r, then the rest of $\mathcal{C}_n$ must be a copy of the (α,θ) regenerative composition of $n-r$, hence the partition structure can be recovered from

$$\pi(\lambda^\downarrow) = \sum_{\text{distinct } \lambda_j\in\lambda^\downarrow} q_*(n:\lambda_j)\ \pi_{\alpha,\theta}(\lambda^\downarrow - \lambda_j)$$

which, by comparison with (33), shows that $\pi = \pi_{\alpha,\theta-\alpha}$, in accordance with the conclusion of the theorem.

Corollary 15. *For $0\le\alpha<1$, $\theta>-\alpha$, each (α,θ) partition structure has a distributionally unique arrangement as a self-similar Markov composition structure.*

There is an explicit stochastic algorithm which allows, for each n, arranging an unordered collection of parts of an (α,θ) partition into a Markovian self-similar composition. Given a partition λ, choose a part λ_j by a size-biased pick and declare it to the right end of the composition under construction. Then arrange the rest parts $\lambda-\lambda_j$ one-by-one, as for the regenerative $(\alpha,\theta+\alpha)$ composition (from right to left), using the appropriate deletion kernel [9, 10]. Specifically, when the rest partition is μ, the algorithm selects *each* part of size r of μ with probability

$$\frac{1}{n}\ \frac{(|\mu|-r)\tau+r(1-\tau)}{1-\tau+(k-1)\tau},$$

where $\tau=\alpha/(2\alpha+\theta)$ and k is the number of parts of μ; then the same procedure is applied to the reduced partition, etc. For example, consider the $(\alpha,0)$ partition of n, assumung it has ℓ parts; after placing a size-biased pick, the rest $\ell-1$ parts should be arranged in a random order, with all $(\ell-1)!$ orders being equally likely.

By the very construction, conditionally given the parts, the last part is a size-biased pick from all parts: this feature is a combinatorial analogue of the strong sampling property in Sect. 3 (as was stated for the $(\alpha, 0)$ case in [14, Proposition 11 (i)]). Algebraically, the combinatorial strong sampling property amounts to the identity

$$\frac{q_{\alpha,\theta}(n : \lambda_j)\ \pi_{\alpha,\alpha+\theta}(\lambda^{\downarrow} - \lambda_j)}{\sum_i q_{\alpha,\theta}(n : \lambda_i)\ \pi_{\alpha,\alpha+\theta}(\lambda^{\downarrow} - \lambda_i)} = \frac{\lambda_j}{n}.$$

6.1. Case $(\alpha = 0,\ \theta > 0)$

This is a case of Example 1, with independent digits and potential function

$$g(j) = \frac{\theta}{j + \theta - 1}.$$

Here S is $\mathrm{PPP}(\theta \mathrm{d}y/y)$. A characteristic feature is that it is the only self-similar Markov compostion structure which is right-regenerative. Indeed, if a random set is both regenerative and stationary regenerative, it is a homogeneous PPP.

The transition function for the $Q^{\uparrow}$ chain is

$$f(j \mid i) = \frac{\theta (j-2)!\,(\theta)_i}{(i-1)!\,(\theta)_j}.$$

Remark on records. The case $\theta = 1$ has classical interpretation in terms of indicators of records in a sequence of i.i.d. random variables with some continuous distribution. With reference to a question left open in [4, p. 297], a similar interpretation exists for any $\theta > 0$, but distributions of independent variables should be different. One possibility, based on a planar homogeneous Poisson process, is the following: divide the positive quadrant $\mathbb{R}^2_+$ into vertical strips of widths $\beta_j = (\theta)_{j-1}/(j-1)!$ and define the variables to be the heights of the lowest Poisson atoms in the strips, from left to right. Elementary algebra shows that, to agree with $\mathrm{ESF}(\theta)$, the collection of β_j's must be as above up to a common positive factor. The same distribution of record indicators appears for an independent sample from distributions $F^{\beta_1},\ F^{\beta_2}, \ldots$ where F is an arbitrary continuous distribution on $\mathbb{R}$.

6.2. Case $(0 < \alpha < 1,\ \theta = 0)$

The range S of an α-stable subordinator induces the renewal composition structure of Example 2. This is the self-similar version of the regenerative (α, α) composition structure, whose Lévy measure after the transform $x = 1 - e^{-y}$ is $\widetilde{\nu}_{\alpha,\alpha}$ defined by

$$\widetilde{\nu}_{\alpha,\alpha}[x, 1] = x^{-\alpha}(1 - x)^{\alpha},$$

hence the potential function is

$$g(j) = \frac{\Phi_{\alpha,\alpha}(j-1)}{\mathtt{m}_{\alpha,\alpha}(j-1)} = \frac{(\alpha)_{j-1}}{(j-1)!}.$$

(where $\mathtt{m}_{\alpha,\alpha}$ is the mean value as in (22)).

The induced composition structure is self-similar Markov as well as left-regenerative. Thus $S \cap [0, 1] = 1 - e^{-R}$ for R the range of another, killed, subordinator, as detailed in [9]. The combination of the two regeneration properties is characteristic:

Proposition 16. *If a composition structure $(\mathcal{C}_n)$ is both Markov self-similar and left regenerative, then $(\mathcal{C}_n)$ is the $(\alpha, 0)$ composition structure derived by sampling from the range of some α-stable subordinator.*

Proof. Let Z be the set in Kingman's representation of $(\mathcal{C}_n)$. The left regeneration property implies that Z is the range of a multiplicative subordinator $1 - e^{-A}$, where A is some subordinator. On the other hand, by Theorem 10, $Z = e^{-B}$ for B some stationary delayed subordinator, hence Z has a nontrivial meander with positive probability, which implies that A has a positive killing rate. Let Z_0 be the set Z conditioned on zero meander, which is the range of the multiplicative subordinator $1 - e^{-A_0}$, for A_0 the version of A without killing. Then, of course, $Z_0 = e^{-B_0}$ for B_0, the version of B but with zero delay. It follows that the composition structure induced by Z_0 is both left- and right-regenerative, that is, both sets Z_0 and $1 - Z_0$ are multiplicatively regenerative. By [9, Theorem 12.1 and Corollary 12.2], $Z_0 \stackrel{d}{=} 1 - Z_0$, Z_0 is the zero set of a Bessel bridge, and the composition structure induced by Z_0 is of type (α, α). By Theorem 14, the stationary version of this composition structure is of type $(\alpha, 0)$, and Z is the range (restricted to $[0, 1]$) of some α-stable subordinator. □

Remark. This result complements the characterisation of (α, α) regenerative composition structures in [9, Theorem 12.1]. Apparently, the assumption of the Markov property can be omitted, i.e., it seems sufficient to assume only that $(\mathcal{C}_n)$ is right-consistent. That the Markov property follows is not obvious, because the left-regeneration property of $(\mathcal{C}_n)$ does not imply the right Markov property of the composition in the sense of

Definition 7 (which requires time-homogeneity of the Markov chain). Still, a plausible argument is the following. As above, define left-regenerative (multiplicatively) Z_0 by conditioning on zero meander (a limiting procedure required to justify this definition is obvious). Fix $x \in]0,1[$ and condition on $x \in Z_0$, then, because Z is self-similar, $[0,x] \cap Z_0 \stackrel{d}{=} x\, Z_0$. But by the left-regeneration (multiplicative) property, $[0,x] \cap Z_0$ is independent of $[x,1] \cap Z_0$, whence the right-regeneration (multiplicative) property. Then the conclusion is above. A loose point in this argument is the conditioning on the zero event $x \in Z_0$.

6.3. Case (α, α)

For this partition there is a regenerative arrangement (the composition structure induced by the Bessel bridge) and another self-similar Markov arrangement. The latter is the self-similar version of the regenerative $(\alpha, 2\alpha)$ composition.

6.4. General $0 < \alpha < 1$, $\theta > -\alpha$

Explicit construction of the self-similar Markov composition structure associated with the (α, θ) partition structure exploits the fragmentation product introduced in Sec. 3.4. One ingredient is the Poisson process Z PPP$(\theta \mathrm{d}y/y)$, $\theta > 0$, restricted to $[0,1]$. Another factor is the set $M' = 1 - M \cap [0,1]$ obtained by reflecting the range M of the α-stable subordinator. The self-similar set is defined then as the fragmentation product $Z \otimes M'$, and the induced composition is the self-similar Markov version of partition $(\alpha, \theta - \alpha)$. Conditioning on zero meander will produce a set corresponding to (α, θ) regenerative composition, as in [9].

Unlike M, the set M' exploited here has the leftmost meander interval. The fragmentation product $Z \otimes M$ was introduced in [17]; the resulting composition structure is right-consistent but not Markovian.

References

1. R. Arratia, *On the central role of scale invariant Poisson processes on* $(0, \infty)$. In: Microsurveys in discrete probability (Princeton, NJ, 1997), 21–41, Amer. Math. Soc., Providence, RI.
2. R. Arratia, A.D. Barbour, and S. Tavaré, *Logarithmic combinatorial structures: a probabilistic approach.* EMS Monographs in Mathematics, European Mathematical Society, Zürich (2003).
3. A. D. Barbour and A. V. Gnedin, *Regenerative compositions in the case of slow variation.* (2005); arXiv:math.PR/0505171.
4. J. Bunge and C. Goldie, *Record sequences and their applications.* In: Handbook of Statistics **19** (D. N. Shanbhag and C.R. Rao eds.), Elsevier (2001), pp. 277–308.

5. P. Donnelly and P. Joyce, *Consistent ordered sampling distributions: characterization and convergence.* — Adv. Appl. Prob. **23** (1991), 229–258.
6. A. V. Gnedin, *The representation of composition structures.* — Ann. Probab. **25** (1997), 1437–1450.
7. A. V. Gnedin, *The Bernoulli sieve.* — Bernoulli **10** (2004), 79–96.
8. A. V. Gnedin, *Three sampling formulas.* — Combin. Probab. Comput. **13** (2004), 185–193.
9. A. V. Gnedin and J. Pitman, *Regenerative composition structures.* — Ann. Probab. **33** (2005), 445–479.
10. A. V. Gnedin and J. Pitman, *Regenerative partition structures.* — Elec. J. Comb. **11(2)** (2004/2005), R12.
11. A. V. Gnedin, J. Pitman, and M. Yor, *Asymptotic laws for composition structures derived from transformed subordinators.* — Ann. Probab. (to appear); arXiv:math.PR/0403438.
12. A. V. Gnedin, J. Pitman, and M. Yor, *Asymptotic laws for regenerative compositions: gamma subordinators and the like.* — (2004); arXiv:math.PR/0405440.
13. S. Nacu, *Increments of random partitions.* — Version 2 of Jan. (2004); arXiv:math.PR/0310091.
14. J. Pitman, *Partition structures derived from Brownian motion and stable subordinators.* — Bernoulli **3** (1997), 79–96.
15. J. Pitman, *Combinatorial stochastic processes.* — Lect. Notes for St. Flour course, July 2002, L. N. Math. (to appear). Available via www.stat.berkeley.edu.
16. J. Pitman and M. Yor, *The two-parameter Poisson–Dirichlet distribution derived from transformed subordinator.* — Ann. Probab. **25** (1996), 855–900.
17. J. Pitman and M. Yor, *Random discrete distributions derived from self-similar random sets.* — Electron. J. Probab. **1** (1996), 1–28.
18. M. I. Taksar, *Stationary Markov sets.* — L. N. Math. **1247** (1986), 302–340.
19. J. E. Young, *Partition-valued stochastic processes with applications.* PhD thesis, UC Berkeley (1995).

Utrecht University
E-mail: gnedin@math.uu.nl

Поступило 27 мая 2005 г.

University of California,
Berkeley
E-mail: pitman@stat.Berkeley.EDU