Regeneration in random combinatorial structures

Alexander V. Gnedin

Department of Mathematics, Utrecht University, Postbus 80010, 3508 TA Utrecht
e-mail: A.V.Gnedin@uu.nl

Abstract: Kingman’s theory of partition structures relates, via a natural sampling procedure, finite partitions to hypothetical infinite populations. Explicit formulas for distributions of such partitions are rare, the most notable exception being the Ewens sampling formula, and its two-parameter extension by Pitman. When one adds an extra structure to the partitions like a linear order on the set of blocks and regenerative properties, some representation theorems allow to get more precise information on the distribution. In these notes we survey recent developments of the theory of regenerative partitions and compositions. In particular, we discuss connection between ordered and unordered structures, regenerative properties of the Ewens-Pitman partitions, and asymptotics of the number of components.

AMS 2000 subject classifications: Primary 60G09, 60C05.
Keywords and phrases: Random partitions and compositions, regeneration, exchangeability.

Received April 2010.

Contents

1 Preface .................................................. 106
2 The paintbox and the two-parameter family ...................... 107
  2.1 The ordered paintbox .................................. 107
  2.2 The Ewens-Pitman partition structures .................. 113
3 Regenerative composition structures .......................... 114
  3.1 Compositions derived from stick-breaking ................ 115
  3.2 Composition structure derived from the zero set of BM ..... 119
  3.3 Regenerative sets and subordinators .................... 119
  3.4 Regenerative compositions from the two-parameter family .... 125
4 Regenerative partition structures and the problem of arrangement ........................................ 128
  4.1 Structural distributions .................................. 128
  4.2 Partition structures invariant under deletion of a part .... 129
  4.3 Deletion kernels for the two-parameter family ............ 131
  4.4 Back to regenerative compositions ...................... 131
  4.5 The arrangement problem for the two-parameter family .... 133
  4.6 More on \((\alpha, \alpha)\) compositions: reversibility ........... 135
5 Self-similarity and stationarity ................................ 136

*This is an original survey paper.
The kind of discrete regenerative phenomenon discussed here is present in the cycle patterns of random permutations. To describe this instance, first recall that every permutation of \( [n] := \{1, \ldots, n\} \) is decomposable in a product of disjoint integer parts. For instance, permutation \((1 3)(2)\) of the set \([3]\) corresponds to the partition of integer 3 with parts 2 and 1. Permutations of different degrees \(n\) are connected in a natural way. Starting with a permutation of \([n]\), a permutation of the smaller set \([n - 1]\) is created by removing element \(n\) from its cycle. This reduction is a surjective \(n\)-to-1 mapping. For instance, three permutations \((1 3)(2), (1)(2 3), (1)(2)(3)\) are mapped to \((1)(2)\).

Now suppose the permutation is chosen uniformly at random from the set of all \(n!\) permutations of \([n]\). The collection of cycle-sizes is then a certain random partition \(\pi_n\) of integer \(n\). By the \(n\)-to-1 property of the projection, the permutation reduced by element \(n\) is the uniformly distributed permutation of \([n - 1]\), with the cycle partition \(\pi_{n-1}\). The transition from \(\pi_n\) to \(\pi_{n-1}\) is easy to describe directly, without reference to underlying permutations: choose a random part of \(\pi_n\) by a size-biased pick, i.e. when the parts of partition are \(n_1, \ldots, n_k\), choose each \(n_j\) with probability \(n_j/n\), then reduce the chosen part by 1. This connection suggests to view the random partitions with varying \(n\) altogether as components of an infinite partition structure \((\pi_n, n = 1, 2, \ldots)\), which is a partition-valued Markov chain with the just specified backward transition rule, from \(\pi_n\) to \(\pi_{n-1}\).

Apart from the consistency property inherent to any partition structure, there is another recursive self-reproduction property of the partitions derived from the cycle patterns of uniform permutations. Fix \(n\) and suppose a part is chosen by a size-biased pick from \(\pi_n\) and completely deleted. Given the part was \(n\), the partition reduced by this part will be a distributional copy of \(\pi_{n-1}\). In this sense the partition structure \((\pi_n, n = 1, 2, \ldots)\) regenerates with respect to the deletion of a part chosen by the size-biased pick. For instance, we can identify the size-biased pick with the cycle containing 1, then the regeneration means that
deleting the cycle with 1 and re-labeling the remaining elements by increasing bijection with \([n - m]\) yields a uniform permutation of \([n - m]\).

The size of the cycle containing element \(n\) has uniform distribution on \(1, \ldots, n\), thus, for large \(n\), this size is about \(nU\), for \(U\) a random variable with uniform distribution on the unit interval. In the same way, the iterated deletion of parts by size-biased picking becomes similar to the splitting of \([0, 1]\) at points representable via products \(U_1 \cdots U_k\) of independent uniform variables. The latter is a special case of the multiplicative renewal process sometimes called stick-breaking. Remarkably, the connection between the splitting of \([0, 1]\) and the \(\pi_n\)'s is invertible: a version of Kingman’s ‘paintbox algorithm’ constructs the uniform permutations from the stick-breaking splitting of \([0, 1]\) in subintervals.

In these notes we consider sequences of partitions and ordered partitions, which are consistent in the same sense as the cycle patterns of permutations for various \(n\). In contrast to that, the assumption about the regeneration property of such structures will be fairly general. For instance, we may require regeneration with respect to the deletion of a part chosen uniformly at random from the set of parts; the corresponding splitting of \([0, 1]\) is then generated by a random Cantor set. The connection between combinatorial partitions and random splittings of the unit interval is central in the theory and will be analysed in detail in the general context of regenerative structures.

Our key topics include connections of the regenerative structures to increasing Lévy processes, relations between ordered and unordered structures, additional symmetries like reversibility and self-similarity, regenerative features of the Ewens-Pitman family, and limit distributions for the component counts.

2. The paintbox and the two-parameter family

2.1. The ordered paintbox

A composition of integer \(n\) is an ordered sequence \(\lambda^o = (\lambda_1, \ldots, \lambda_k)\) of positive integer parts with sum \(|\lambda^o| := \sum_j \lambda_j = n\). We shall think of composition as a model of occupancy, meaning \(n\) 'balls' separated by 'walls' into some number of nonempty 'boxes', like in this diagram

```
|   • • • |   •  •  • |
```

representing composition \((3, 1, 2)\). A wall \(\mid\) is either placed between two consecutive \(\bullet\)'s or not, hence there are \(2^{n-1}\) compositions of \(n\). Sometimes we shall also use encoding of the compositions into binary sequences, in which a 1 followed by some \(m - 1\) zeroes corresponds to part \(m\), like the code 100110 for composition \((3, 1, 2)\).

A related labeled object is an ordered partition of the set \([n] := \{1, \ldots, n\}\), which may be obtained by some enumeration of the balls by integers \(1, \ldots, n\), like

```
|   2 4 5 |   3 1 6 |
```
The number of such labelings, that is the number of ordered set partitions with shape \((\lambda_1, \ldots, \lambda_k)\), is equal to the multinomial coefficient

\[
f^n(\lambda_1, \ldots, \lambda_k) := \frac{n!}{\lambda_1! \cdots \lambda_k!}.
\]

Throughout, symbol \(\circ\) will denote a function of composition, also when the function is not sensitive to the permutation of parts.

Discarding the order of parts in a composition \((\lambda_1, \ldots, \lambda_k)\) yields a partition of integer \(|\lambda|\), usually written as a ranked sequence of nondecreasing parts. For instance, the ranking maps compositions \((3, 1, 2)\) and \((1, 3, 2)\) to the same partition \((3, 2, 1)\) ↓, where ↓ will be used to denote the operation of ranking and to also indicate that the arrangement of parts in sequence is immaterial.

Sometimes we use notation like \(2 \in (4, 2, 2, 1)\) ↓ to say that \(2\) is a part of partition. The number of partitions of the set \([n]\) with the same shape \(\lambda \downarrow = (\lambda_1, \ldots, \lambda_k)\) ↓ is equal to

\[
f(\lambda^\downarrow) := n! \prod_{r=1}^{k_{\lambda^\downarrow}^\downarrow} \frac{1}{(r!)^{k_r} k_r !}
\]

where \(k_r = \#\{ j : \lambda_j = r \} \geq 0\) is the number of parts of \(\lambda^\downarrow\) of size \(r\). Observe that the counts \(k_1, \ldots, k_n\) uniquely determine \(\lambda^\downarrow\) and satisfy \(\sum_r r k_r = n\).

A random composition/partition of \(n\) is simply a random variable with values in the finite set of compositions/partitions of \(n\). One statistical context where these combinatorial objects appear is the species sampling problem. Imagine an alien who has no idea of the mammals. Suppose the first six mammals she observes are tiger, giraffe, elephant, elephant, elephant and giraffe, appearing in this sequence. Most frequent – three of these – have long trunks, two are distinctively taller than the others, and one is striped. She records this as partition \((3, 2, 1)\) ↓ into three distinct species. Composition \((1, 3, 2)\) could appear as the record of species abundance by more delicate classification according to typical height, from the lowest to the tallest\(^1\). Enumerating the animals in the order of observation gives a labeled object, a partition/ordered-partition of the set \([6] = \{1, \ldots, 6\}\).

There are many ways to introduce random partitions or compositions. The method adopted here is intrinsically related to the species sampling problem. This is the following ordered version of Kingman’s paintbox (see [9, 16, 28, 54]).

**Ordered paintbox** Let \(\mathcal{R}\) be a random closed subset of \([0, 1]\). The complement open set \(\mathcal{R}^c := (0, 1) \setminus \mathcal{R}\) has a canonical representation as a disjoint union of countably many open interval components, which we shall call the gaps of \(\mathcal{R}\). Independently of \(\mathcal{R}\), sample points \(U_1, U_2, \ldots\) from the uniform distribution on \([0, 1]\) and group the points in clusters by the rule: \(U_i, U_j\) belong to the same cluster if they hit the same gap of \(\mathcal{R}\). If \(U_i\) falls in \(\mathcal{R}\) let \(U_i\) be a singleton. For

\(^1\)If her guidebook would describe four species, e.g. these three and the cows, her records would be \((3, 2, 1, 0)^\downarrow, (1, 0, 3, 2)\) (weak partitions, respectively, weak compositions), but we assumed that she knew apriori really nothing of the mammals.
each \( n \), count the representatives of clusters among \( U_1, \ldots, U_n \) and define \( \kappa_n \), a random composition of integer \( n \), to be the record of positive counts in the left-to-right order of the gaps.

For instance, \( \kappa_n \) assumes the value \((3, 1, 2)\) if, in the left-to-right order, there is a gap hit by three points out of \( U_1, \ldots, U_6 \), a singleton cluster resulting from either some gap or from some \( U_j \in \mathcal{R} \), and a gap hit by two of \( U_1, \ldots, U_6 \).

In the proper case \( \mathcal{R} \) has Lebesgue measure zero almost surely, hence the event \( U_j \in \mathcal{R} \) occurs only with probability zero. We may think then of points of \( \mathcal{R} \) as possible locations of walls | and of the points of \([0, 1]\) as possible locations of balls •. In a particular realisation, the balls appear at locations \( U_j \), and the walls bound the gaps hit by at least one ball. In the improper case, \( \mathcal{R} \) may have positive measure with nonzero probability. If \( U_j \in \mathcal{R} \) we can imagine a box with walls coming so close together that no further ball will fit in this box, so \( U_j \) will forever remain a singleton, no matter how many balls are added.

Sometimes we shall identify \( \mathcal{R} \) with the splitting of \([0, 1]\) it induces, and just call \( \mathcal{R} \) itself the paintbox. Molchanov 51 gives extensive exposition of the theory of random sets, although an intuitive idea will suffice for most of our purposes. This can be a set of some fixed cardinality, e.g. splittings of \([0, 1]\) following a Dirichlet distribution (see 17, 42), or complicated random Cantor-type sets like the set of zeroes of the Brownian motion. It will be also convenient to make no difference between two closed subsets of \([0, 1]\) if they only differ by endpoints 0 or 1. If 1 or 0 is not accumulation point for \( \mathcal{R} \), the gap adjacent to the boundary will be called right or left meander.

The paintbox with random \( \mathcal{R} \) is a kind of canonical representation of ‘non-parametric priors’ in the species sampling problem. View \( \mathbb{R} \) as an ordered space of distinct types. Originally by Kingman 47, the types were colours making up a paintbox. Consider a random probability measure \( F \) on reals as a model of infinite ordered population. Let \( \xi_1, \xi_2, \ldots \) be a sample from \( F \), which means that conditionally given \( F \), the \( \xi_j \)'s are i.i.d. with distribution \( F \). An ordered partition of the sample is defined by grouping \( j \)'s with the same value of \( \xi_j \), with the order on the groups maintained by increase of the values. The case of diffuse (nonatomic) \( F \) is trivial – then ties among \( \xi_j \)'s have probability zero and the partition has only singletons, so the interesting case is \( F \) with atoms, when the partition will have nontrivial blocks. The same ordered partition is induced by any other distribution obtained from \( F \) by a suitable monotonic transformation, which may be random. To achieve the uniqueness, we view \( F \) as a random distribution function and we observe that \( \xi_i \leq \xi_j \) iff \( F(\xi_i) \leq F(\xi_j) \). Conditioning on \( F \) and applying the quantile transform \( y \to F(y) \) to the sample produces another sample \( \tilde{\xi}_1, \tilde{\xi}_2, \ldots \) from the transformed distribution \( \tilde{F} \) supported by \([0, 1]\). In the diffuse case, \( \tilde{F} \) is well known to be the uniform distribution, and in general the distribution function \( \tilde{F} \) is of special kind: it satisfies \( \tilde{F}(x) \leq x \) for \( x \in [0, 1] \) and \( \tilde{F}(x) = x \) \( F \)-a.s. Moreover, each jump location of \( \tilde{F} \) is preceded by a flat (where \( \tilde{F} \) is constant), whose length is equal to the size of the jump. The latter implies that the composition derived from \( \tilde{F} \) by grouping equal \( \tilde{\xi}_j \)'s in clusters is the same as the composition obtained via the paintbox construction.
from $R = \text{support}(\tilde{F})$. The identification with the paintbox construction can be shown more directly, i.e. without appealing to $\tilde{F}$, by taking for $R$ the range of the random function $F$ (note that $\text{support}(\tilde{F})$ with 0 attached to it coincides with the range of $F$).

Note further important features inherent to the paintbox construction:

- The unlabeled object, $\kappa_n$, is determined by $R$ and the uniform order statistics $U_{n:1} < \cdots < U_{n:n}$, i.e. the ranks of $U_1, \ldots, U_n$ do not matter for the composition.
- Attaching label $j$ to the ball corresponding to $U_j$, we obtain, for each $n$, an ordered partition $K_n$ of the set $[n]$, with shape $\kappa_n$. This ordered set partition is exchangeable, meaning that a permutation of the labels does not change the distribution of $K_n$, thus all ordered partitions of $[n]$ with the same shape have the same probability.
- The ordered partitions $K_n$ are consistent as $n$ varies. Removing ball $n$ (and deleting an empty box in case one is created) reduces $K_n$ to $K_{n-1}$. The infinite sequence $K = (K_n)$ of consistent ordered partitions of $[1], [2], \ldots$ defines therefore an exchangeable ordered partition of the infinite set $\mathbb{N}$ into some collection of nonempty blocks.

Translating the consistency in terms of compositions $\kappa_n$ we arrive at

**Definition 2.1.** A sequence $\kappa = (\kappa_n)$ of random compositions of $n = 1, 2, \ldots$ is called a composition structure if these are sampling consistent: for each $n > 1$, conditionally given $\kappa_n = (\lambda_1, \ldots, \lambda_k)$ the composition $\kappa_{n-1}$ has the same distribution as the composition obtained by reducing by 1 each part $\lambda_j$ with probability $\lambda_j/n$.

A size-biased part of composition $\lambda^o$ is a random part which coincides with every part $\lambda_j$ with probability $\lambda_j/|\lambda^o|$. A size-biased part of a random composition $\kappa_n$ is defined conditionally on the value $\kappa_n = \lambda^o$. The sampling consistency condition amounts to the transition from $\kappa_n$ to $\kappa_{n-1}$ by reducing a size-biased part. This special reduction rule in Definition 2.1 is a trace of the exchangeability in $K_n$ that remains when the labels of balls are erased: indeed, given the sizes of the blocks, the ball with label $n$ belongs to a particular block of size $\lambda_j$ with probability $\lambda_j/n$.

Keep in mind that the consistency of ordered set partitions $K_n$ is understood in the strong sense, as a property of random objects defined on the same probability space, while Definition 2.1 only requires weak consistency in terms of the distributions of $\kappa_n$’s. By the measure extension theorem, however, the correspondence between the probability laws of exchangeable ordered partitions of $\mathbb{N}$ and the laws of composition structures is one-to-one, and any composition structure can be realised through an exchangeable ordered partition of $\mathbb{N}$. In view of this correspondence, dealing with labeled or unlabeled objects is just the matter of convenience, and we shall freely switch from one model to another.

A central result about the general composition structures says that these can be uniquely represented by a paintbox [16]. This extends Kingman’s [47] representation of partition structures.
Theorem 2.2. For every composition structure $\kappa = (\kappa_n)$ there exists a unique distribution for a random closed set $\mathcal{R}$ which by means of the paintbox construction yields, for each $n$, a distributional copy of $\kappa_n$.

Sketch of proof The line of the proof is analogous to modern proofs of de Finetti’s theorem which asserts that a sequence of exchangeable random variables is conditionally i.i.d. given the limiting empirical distribution of the sequence (see Aldous [1, 2]). To this end, we need to make the concept of a random closed set more precise by introducing a Borel structure on the space of closed sets. One way to do this is to introduce topology on the space of closed subsets of $[0, 1]$ by means of the Hausdorff distance. Recall that for $R_1, R_2 \subset [0, 1]$ (with boundary points 0, 1 adhered to the sets) the distance is equal to the smallest $\epsilon$ such that the $\epsilon$-inflation of $R_1$ covers $R_2$ and the same holds with the roles of the sets swapped, so the distance is small when the sizes and positions of a few biggest gaps are approximately the same for both sets.

Realise all $\kappa_n$’s on the same probability space through some exchangeable $K$. Encode each composition $(\lambda_1, \ldots, \lambda_k)$ into a finite set $\{0, \Lambda_1/n, \ldots, \Lambda_{k-1}/n, 1\}$ where $\Lambda_j = \lambda_1 + \cdots + \lambda_j$. This maps $\kappa_n$ to a finite random set $\mathcal{R}_n \subset [0, 1]$. By a martingale argument it can be shown that the law of the large numbers holds: as $n \to \infty$ the sets $\mathcal{R}_n$ converge almost surely to a random closed set $\mathcal{R}$. The limit set $\mathcal{R}$ directs the paintbox representation of $\kappa$.

There are various equivalent formulations of the result in terms of (i) the exchangeable quasi-orders on $\mathbb{N}$ [43], (ii) the entrance Martin boundary for the time-reversed Markov chain $(\kappa_n, n = \ldots, 2, 1)$ and (iii) certain functionals on the infinite-dimensional algebra of quasisymmetric functions [28].

We define the composition probability function (CPF for shorthand) $p^\circ$ of a composition structure $\kappa$ as

$$p^\circ(\lambda^\circ) := \mathbb{P}(\kappa_n = \lambda^\circ), \quad |\lambda^\circ| = n, \ n = 1, 2, \ldots$$

For fixed $|\lambda^\circ| = n$ this is the distribution of $\kappa_n$. To avoid confusion with the distribution of $K_n$ we stress that the probability of any particular value of the set partition $K_n$ with shape $\lambda^\circ$ is equal to $p^\circ(\lambda^\circ)/f^\circ(\lambda^\circ)$. Sampling consistency translates as a backward recursion

$$p^\circ(\lambda^\circ) = \sum_{\mu^\circ} c(\lambda^\circ, \mu^\circ)p^\circ(\mu^\circ), \quad (2.1)$$

where $\mu^\circ$ runs over all shapes of extensions of any fixed ordered partition of $[n]$ with shape $\lambda^\circ$ to some ordered partition of $[n+1]$. For instance, taking $\lambda^\circ = (2, 3)$, $\mu^\circ$ assumes the values $(1, 2, 3), (2, 1, 3), (2, 3, 1), (3, 3), (2, 4)$. The coefficient $c(\lambda^\circ, \mu^\circ)$ is the probability to obtain $\lambda^\circ$ from $\mu^\circ$ by reducing a part of $\mu^\circ$ selected by the random size-biased pick.

For fixed $n$, if $p^\circ(\lambda^\circ)$ is known for compositions $\lambda^\circ$ with $|\lambda^\circ| = n$, then solving (2.1) backwards gives the values of CPF for all compositions with $|\lambda^\circ| \leq n$. By linearity of the recursion, every such partial solution, with $n' \leq n$, is a convex
A.V. Gnedin/Regeneration in random combinatorial structures

112

combination of $2^{n-1}$ solutions obtained by taking delta measures on the level $n$. Similarly, without restricting $n$, the set of CPF’s is convex and compact in the weak topology of functions on a countable set; this convex set is an infinite-dimensional Choquet simplex, i.e. the set has the property of uniqueness of barycentric decomposition in terms of its extreme elements (a general property of projective limits of classical simplices, see e.g. Proposition 10.21 in [38]).

The extreme CPF’s are precisely those derived from nonrandom paintboxes. The correspondence between extreme solutions and closed subsets of $[0,1]$ is a homeomorphism, which extends to the homeomorphism between all CPF’s and distributions for random closed $\mathcal{R}$.

Discarding the order of parts in each $\kappa_n$ we obtain Kingman’s partition structure $\pi = (\pi_n)$ with $\pi_n = \kappa_n^\downarrow$. Partition structures satisfy the same sampling consistency condition as in Definition 2.1. The corresponding labeled object is an exchangeable partition $\Pi = (\Pi_n)$ of the infinite set $\mathbb{N}$. The law of large numbers for partition structures says that, as $n \to \infty$, the vector $n^{-1}\pi_n$ padded by infinitely many zeroes converges (weakly for $\pi_n$, strongly for $\Pi_n$) to a random element $S$ of the infinite-dimensional simplex

$$\nabla = \{(s_i) : s_1 \geq s_2 \ldots \geq 0, \sum_i s_i \leq 1\},$$

so the components of $S$ are the asymptotic frequencies of the ranked parts of $\kappa_n$. The partition probability function (PPF)

$$p(\lambda^\downarrow) := P(\pi_n = \lambda^\downarrow), \quad |\lambda^\downarrow| = n, \quad n = 1, 2, \ldots,$$

specifies distributions of $\pi_n$’s and satisfies a recurrence analogous to (2.1). Our definition of PPF differs by the standard combinatorial factor $f(\lambda^\downarrow)$ from the definition in [56]. The correspondence between PPF’s and distributions for unordered paintbox $\mathcal{S}$ is bijective. Note that the possibility of strict inequality $\sum_j s_j < 1$ occurs in the improper case, where the diffuse mass $1 - \sum_j s_j$, sometimes also called dust [9], is equal to the cumulative frequency of singleton blocks of $\Pi$ given $\mathcal{S} = (s_j)$.

Discarding order is a relatively easy operation. In terms of ordered and unordered paintboxes $\mathcal{R}$ and $\mathcal{S}$ the connection is expressed by the formula

$$\mathcal{S} = (\mathcal{R}^\circ)^\downarrow,$$

(2.2)

where the ranking $\downarrow$ means that the gap-sizes of $\mathcal{R}$ are recorded in nonincreasing order. The operation $\downarrow$ is a continuous mapping from the space of closed subsets of $[0,1]$ to $\nabla$. In terms of distributions, passing from CPF to PPF is expressed by the symmetrisation formula

$$p(\lambda^\downarrow) = \sum_\sigma p^\circ(\lambda^\sigma),$$

(2.3)

where $\lambda^\sigma$ runs over all distinct arrangements of parts of $\lambda^\downarrow$ in a composition (e.g. for partition $(2,1,1)^\downarrow$ there are three such compositions $(2,1,1), (1,2,1), (1,1,2)$).
In the other direction, there is one universal way to introduce the order. With every partition structure one can associate a unique symmetric composition structure, for which any of the following three equivalent conditions holds:

(i) all terms in the RHS of (2.3) are equal,
(ii) conditionally given $\Pi_n$ with $k$ blocks, any arrangement of the blocks in $K_n$ has the same probability,
(iii) the gaps of $\mathcal{R}$ appear in the exchangeable random order.

The last property (iii) means that, conditionally given $S = (s_j)$, if $s_k > 0$ then every relative order of the first $k$ largest gaps (labeled by $[k]$) of sizes $s_1, \ldots, s_k$ has probability $1/k!$. This rule defines $\mathcal{R}$ unambiguously in the proper case, and extension to the improper case follows by continuity [17]. A simple example of symmetric $\mathcal{R}$ is associated with splitting $[0, 1]$ according to the symmetric Dirichlet distribution on a finite-dimensional simplex.

Besides from the symmetric composition structure, there are many other composition structures associated with a given partition structure. Understanding the connection in the direction from unordered to ordered structures is a difficult problem of arrangement. To outline some facets of the problem, suppose we have a rule to compute $p^\circ$ from $p$, what does it mean then in terms of passing from $S$ to $\mathcal{R}$? Specifically, given $S = (s_j)$, in which order the intervals of sizes $s_1, s_2, \ldots$ should be arranged in an open set? Other way round, suppose we have a formula for $p$ and know that (2.2) is true, how then can we compute the probability that given, say, $\pi_5 = (3, 2)\downarrow$ the parts appear in the composition as $(2, 3)$? Most questions like that cannot have universal answers, because random sets and random series are objects of high complexity, and the paintbox correspondence cannot be expressed by simple formulas.

### 2.2. The Ewens-Pitman partition structures

In the theory of partition structures and their applications to species sampling, processes of fragmentation and coagulation, random trees etc., a major role is played by the Ewens-Pitman two-parameter family of partitions, with PPF

$$p_{\alpha, \theta}(\lambda^\downarrow) = f(\lambda^\downarrow) \frac{\prod_{j=1}^{k-1} (\theta + \alpha j)}{(1 + \theta)^{n-1}} \prod_{j=1}^{k} (1 - \alpha)_{\lambda_j - 1}, \quad \lambda^\downarrow = (\lambda_1, \ldots, \lambda_k)^\downarrow, \quad \lambda^\downarrow$$

where and henceforth $(z)_n := z(z + 1) \cdots (z + n - 1)$ is a rising factorial. The principal range of the parameters is

$$\{ (\alpha, \theta) : 0 \leq \alpha < 1, \theta > -\alpha \} \cup \{ (\alpha, \theta) : \alpha < 0, -\theta/\alpha \in \mathbb{N} \}, \quad (2.5)$$

and there are also a few degenerate boundary cases defined by continuity.

One of many remarkable features of these partitions is the following explicit ‘rule of succession’ (due to Dubins and Pitman) for generating the corresponding exchangeable partition $\Pi = (\Pi_n)$. Start with the one-element partition $\Pi_1$.

Inductively, suppose $\Pi_n$ has been constructed then, given that the shape of $\Pi_n$
is \((\lambda_1, \ldots, \lambda_k)^i\), the ball \(n + 1\) is placed in the existing box \(i\) with probability \((\lambda_i - \alpha)/(n + \theta)\) for \(i = 1, \ldots, k\), and starts a new box with probability \((\theta + k\alpha)/(n + \theta)\).

In one popular interpretation known as ‘Chinese restaurant process’ the balls correspond to customers arriving in the restaurant, and boxes are circular tables. With account of the circular ordering of customers at the tables, and subject to uniform random placement at each particular table, the process also defines a consistent sequence of random permutations for \(n = 1, 2, \ldots\). In the case \(\alpha = 0, \theta = 1\) we obtain a forward algorithm to construct uniform random permutations: after the permutation of \([n]\) is constructed, element \(n + 1\) is inserted with the same probability clockwise next to each of \(1, 2, \ldots, n\) in one of the ‘old’ cycles, and a ‘new’ singleton cycle is created with probability \(1/(n + 1)\).

The two-parameter family has numerous connections to basic types of random processes like the Poisson process and the Brownian motion, see Pitman’s lecture notes [56] for a summary. It also provides an exciting framework for the problem of arrangement [23]. Explicit formula (2.4) and the simple sequential construction make these partitions very attractive for applications.

One direction in which (2.4) was generalised is the theory of Gibbs exchangeable partitions [21, 31, 45, 50]. When the ball \(n + 1\) is placed in one of the ‘old’ boxes \(i = 1, \ldots, k\), it is placed in box \(i\) with probability still proportional to \(n_i - \alpha\) (under assumption of a product form for PPF this rule is forced by exchangeability [31]). But the succession rule for the ‘new’ box can be more general. For \(\alpha = -1\) there exists another two-parameter family of Gibbs partitions with simple succession rule and PPF, see [21] for the characterisation of this and the Ewens-Pitman families.

3. Regenerative composition structures

Every \(\kappa_n\) in a composition structure may be regarded as a reduced copy of \(\kappa_{n+1}\). We complement this now by another type of self-reproduction property, related to the reduction by a whole box.

**Definition 3.1.** A composition structure \(\kappa = (\kappa_n)\) is called regenerative if for all \(n > m \geq 1\), the following deletion property holds. Conditionally given the first part of \(\kappa_n\) is \(m\), if the part is deleted then the remaining composition of \(n - m\) is distributed like \(\kappa_{n-m}\).

Similar definition applies to exchangeable ordered partition \(K = (K_n)\), with the amendment that when the first block of size \(m\) of \(K_n\) is deleted, the elements of the remaining blocks should be re-labeled by an increasing bijection with \([n - m]\).

Denote \(F_n\) the first part of \(\kappa_n\) and consider its distribution

\[
q(n : m) := \sum_{|\lambda^0| = n, \lambda_1 = m} p^0(\lambda^0).
\]

It follows immediately from the definition that \(\kappa\) is regenerative iff the CPF has the product form
A.V. Gnedin/Regeneration in random combinatorial structures

\[ p^\circ(\lambda_1, \ldots, \lambda_k) = \prod_{j=1}^{k} q(\Lambda_j : \lambda_j), \quad (3.1) \]

where \( \Lambda_j = \lambda_j + \cdots + \lambda_k \) for \( 1 \leq j \leq k \).

For each \( n \), the formula identifies \( \kappa_n \) with the sequence of decrements of a decreasing Markov chain \( Q^\downarrow_n = (Q^\downarrow_n(t), \ t = 0, 1, \ldots) \) on \( 0, \ldots, n \). The chain starts at \( n \), terminates at \( 0 \), and jumps from \( n' \leq n \) to \( n' - m \) with probability \( q(n' : m) \).

The binary code of \( \kappa_n \) is obtained by writing 1's in positions \( n - Q^\downarrow_n(t) + 1 \), \( t = 0, 1, \ldots \), and writing 0's in all other positions, with the convention that the last 1 in position \( n + 1 \) is not included in the code. In view of this Markov chain interpretation, we call \( q(n : m), 1 \leq m \leq n, n \in \mathbb{N} \) the decrement matrix of \( \kappa \). Since \( p^\circ \) is computable from \( q \), the decrement matrix determines completely the distributions of \( \kappa_n \)'s and the distribution of the associated exchangeable ordered partition \( K \).

For a given regenerative \( \kappa \) let \( \pi = (\pi_n) \), with \( \pi_n = \kappa_n^\downarrow \), be the related partition structure. Think of \( \kappa_n \) as an arrangement of parts of \( \pi_n \) in some order. For partition \( \lambda^\downarrow \) of \( n \) and each \( m \in \lambda^\downarrow \) define the deletion kernel

\[ d(\lambda^\downarrow, m) = \mathbb{P}(F_n = m \mid \pi_n = \lambda^\downarrow), \]

which specifies the conditional probability, given the unordered multiset of parts, to place a part of size \( m \) in the first position in \( \kappa_n \) (so \( d(\lambda^\downarrow, m) = 0 \) if \( m \notin \lambda \)).

The deletion property of \( \kappa \) implies that the PPF of \( \pi \) satisfies the identity

\[ p(\lambda^\downarrow) d(\lambda^\downarrow, m) = q(n : m) p(\lambda^\downarrow \setminus \{m\}), \quad (3.2) \]

where \( q(n : \cdot) \), the distribution of \( F_n \), may be written in terms of the deletion kernel as

\[ q(n : m) = \sum_{\{\lambda^\downarrow : |\lambda^\downarrow| = n, m \in \lambda^\downarrow\}} d(\lambda^\downarrow, m) p(\lambda^\downarrow). \quad (3.3) \]

Intuitively, the deletion kernel is a stochastic algorithm of choosing a part of partition \( \pi_n \) to place it in the first position of composition \( \kappa_n \). Iterated choices arrange all parts of each \( \pi_n \) in \( \kappa_n \), hence the deletion kernel may be used to describe the arrangement on the level of finite partitions. The partition structure \( \pi \) inherits from \( \kappa \) the property of invariance under deletion of a part chosen by some random rule, expressed formally as (3.2) and (3.3). This is, of course, a subtle property when compared with more obvious invariance of \( \kappa \) under the first-part deletion, as specified in Definition 3.1.

### 3.1. Compositions derived from stick-breaking

Exploiting the paintbox construction we shall give a large family of examples of regenerative composition structures. The method is called stick-breaking, and it is also known under many other names like e.g. residual allocation model or, deeper in history, random alms [41].
Let \((W_i)\) be independent copies of some random variable \(W\) with range \(0 < W \leq 1\). A value of \(W\) is chosen, and the unit stick \([0, 1]\) is broken at location \(W\) in two pieces, then the left piece of size \(W\) is frozen, and the right piece of size \(1 - W\) is broken again in proportions determined by another copy of \(W\), and so on ad infinitum. The locations of breaks make up a random set \(R\) with points

\[ Y_k = 1 - \prod_{i=1}^{k} (1 - W_i), \quad k = 1, 2, \ldots, \]

so the gaps are \(R^c = \bigcup_{k=0}^{\infty} (Y_k, Y_{k+1})\). The cardinality of \(R\) is finite if \(P(W = 1) > 0\), but otherwise infinite, with points accumulating only at the right endpoint of the unit interval. By the i.i.d. property of the proportions, the part of \(R\) to the right of \(Y_1\) is a scaled copy of the whole set,

\[ \frac{(R \cap [Y_1, 1]) - Y_1}{1 - Y_1} \overset{d}{=} R, \]

and this re-scaled part of \(R\) is independent of \(Y_1\).

Suppose a composition structure \(\kappa\) is derived from the paintbox \(R = \{Y_j, j = 0, 1, \ldots\}\). If \((0, Y_1)\) contains at least one of the first \(n\) uniform points \(U_j\), then the first part of the composition \(\kappa_n\) is equal to the number of uniforms hitting this interval. Otherwise, conditionally given \(Y_1\), the sample comes from the uniform distribution on \([Y_1, 1]\). Together with the property (3.5) of \(R\) this implies

\[ q(n : m) = \binom{n}{m} \mathbb{E}(W^m(1-W)^{n-m}) + \mathbb{E}(1-W)^n q(n : m), \]

whence the law of the first part of \(\kappa_n\) is

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

which is a mixture of binomial distributions conditioned on a positive value. The key property (3.5) we exploited can be generalised for every \(Y_k \in R\), from which iterating the argument we obtain the product formula (3.1).

In [19] the composition was introduced as the outcome of the following iterative Bernoulli sieve procedure. Start with \(n\) balls. In the first round choose a value for \(W\) and given this value mark the balls independently heads, respectively tails, with probabilities \(W\) and \(1 - W\). The ‘heads’ are removed to make up the first block of composition. If the number of heads is zero the round is ignored. Choose another independent value for \(W\) and iterate with remaining balls.

Concrete examples are obtained by choosing a distribution for \(W\). For instance, taking delta measure \(\delta_x\) with some \(x \in (0, 1)\) yields \(R = \{1-(1-x)^k, k = 0, \ldots, \infty\}\), which induces the same composition structure as the one associated with sampling from the geometric distribution on the set of integers. This composition structure was studied in many contexts, including theory of records and random search algorithms.
Expectations involved in (3.6) may be computed explicitly only in some cases, e.g. for $W$ with polynomial density, but even then the product formula (3.1) rarely simplifies.

**Example** Here is an example of a relatively simple decrement matrix. Taking $W$ with the general two-parameter beta density

$$\nu(dx) = \frac{x^{\gamma-1}(1-x)^{\theta-1}dx}{B(\gamma, \theta)}, \ (\gamma, \theta > 0)$$

we arrive at

$$q(n : m) = \binom{n}{m} \frac{(\gamma)_m(\theta)_{n-m}}{(\gamma + \theta)_n - (\theta)_n}. \ (3.8)$$

The product formula (3.1) simplifies moderately for general integer $\gamma$, and massively in the following case $\gamma = 1$.

**Regenerative composition structures associated with Ewens’ partitions** Now suppose $W$ has a beta$(1, \theta)$ density

$$\nu(dx) = \theta(1-x)^{\theta-1}dx, \ x \in (0, 1).$$

Evaluating beta integrals in (3.6) we find the decrement matrix

$$q(n : m) = \binom{n}{m} \frac{(\theta)_{n-m} m!}{(\theta + 1)_{n-1} n}, \ (3.9)$$

and massive cancellation in (3.1) gives the CPF

$$p_{0,\theta}(\lambda_1, \ldots, \lambda_k) = \frac{\theta^k n!}{(\theta)_n} \prod_{j=1}^k \frac{1}{\Lambda_j}, \ (3.10)$$

with $\Lambda_j = \lambda_j + \cdots + \lambda_k$. Symmetrisation (2.3) gives the PPF known as the **Ewens sampling formula** (ESF)

$$p_{0,\theta}(\lambda^i) = f(\lambda^i) \frac{\theta^k}{(\theta)_n} \prod_{i=1}^k (n_i - 1)!, \ (3.11)$$

which is a special case of (2.4) with $\alpha = 0$. Recall that $f(\lambda^i)$ is the number of set partitions of $[n]$ with given shape. The range of parameter is $\theta \in [0, \infty]$, with the boundary cases defined by continuity.

For $\theta = 1$, the distribution of $W$ is uniform$[0, 1]$ and $q(n : m) = n^{-1}$ is a discrete uniform distribution for each $n$: the associated partition $\pi_n$ is the same as the cycle partition of a uniform random permutation of $[n]$. For general $\theta$, the ESF corresponds to a biased permutation, which for each $n$ takes a particular value with probability $\theta^\#cycles/(\theta)_n$. See [3] for properties of these permutations.

We shall call $\kappa$ with CPF (3.10) Ewens’ regenerative composition structure. The problem of arrangement has in this case a simple explicit solution. For
partition \((\lambda_1, \ldots, \lambda_k)\) the size-biased permutation is the random arrangement of parts obtained by the iterated size-biased picking without replacement. For nonnegative \((s_j) \in \nabla\) with \(\sum_j s_j = 1\) define a size-biased permutation in a similar way: a generic term \(s_i\) is placed in position 1 with probability proportional to \(s_i\), then another term \(s_j\) \((j \neq i)\) is chosen from the remaining terms with probability proportional to \(s_j\) and placed in position 2, etc. The resulting random sequence is then in the size-biased order, with distribution of the sequence invariant under the size-biased permutation \(^2\).

**Theorem 3.2.** Ewens’ composition structure \((3.10)\) has parts in the size-biased order, for every \(n\). Conversely, if a regenerative composition structure has parts in the size-biased order, then its CPF is \((3.10)\) for some \(\theta \in [0, \infty]\).

The paintbox also has a similar property: the intervals \((Y_j, Y_{j+1})\) are in the size-biased order. In particular, given the sizes of intervals, a generic gap of size \(s_j\) is the left meander with probability \(s_j\). The law of ranked frequencies \(S\) is known as Poisson-Dirichlet distribution. The law of the gap-sizes \((Y_1 - Y_0, Y_2 - Y_1, \ldots)\) is called the GEM distribution.

**Remark** For set partitions, size-biased ordering is sometimes understood as the arranging of blocks of \(\Pi_n\) by increase of their minimal elements (other often used names: age ordering, sampling ordering). This creates an ordered partition for each \(n\), but this ordered partition is not exchangeable, since e.g. element 1 is always in the first block. In Ewens’ case, but not in general, the unlabeled compositions associated with the arrangement by increase of the minimal elements of blocks are sampling consistent as in Definition 2.1 (this observation is due to Donnelly and Joyce [15]). The last assertion is just another formulation of Theorem 3.2.

Combining Theorem 3.2 with Definition 3.1 we conclude for (unordered) Ewens’ partition structure a property which may be called invariance under deletion of a size-biased part:

\[ \pi_n \setminus J_n \overset{d}{=} \pi_{n - J_n'}, \quad n = 1, 2, \ldots, \quad (3.12) \]

where in the LHS \(J_n \in \pi_n\) is the part chosen by size-biased pick, \(\setminus\) means removing a part, and in the RHS \(J_n'\) is independent of \((\pi_k)\) and has the same distribution as \(J_n\). Kingman [48] showed that \((3.12)\) is characteristic for the ESF and interpreted the property as noninterference of species.

A generalisation of the distributional relation \((3.12)\) was studied in [23]. If for some partition structure \((\pi_n)\) the identity \((3.12)\) holds but with \((\pi_k)\) in the RHS replaced by arbitrary partitions \(\rho_k\) of \(k = 1, 2, \ldots\) which need not be consistent in any sense, then \((\pi_n)\) must be one of the Ewens-Pitman partition structures (subject to minor regularity conditions to exclude a short list of degenerate exceptional cases).

\(^2\)To define a size-biased arrangement in the improper case \(\sum_j s_j \leq 1\) consider any closed set \(R \subset [0, 1]\) with ranked sequence of gap-sizes \((s_j)\). Sample uniformly balls \(U_j\) and record the gap-sizes by increase of the minimal labels of balls, with understanding the points of \(R\) as zero-size gaps.
3.2. Composition structure derived from the zero set of BM

Consider the process \((B_t, t \geq 0)\) of Brownian motion (BM), and let \(\mathcal{Z} = \{t : B_t = 0\}\) be the zero set of BM. The complement \(\mathcal{R} \setminus \mathcal{Z}\) is the union of the excursion intervals, where the BM is away from zero. Define \(\mathcal{R}\) as \(\mathcal{Z}\) restricted to \([0, 1]\). There is a meander gap between the last zero of BM and 1, caused by an incomplete excursion abruptly ended at \(t = 1\), but to the left of the meander the set \(\mathcal{R}\) is of the Cantor-type, without isolated points. Thus, in contrast to the stick-breaking case, the gaps cannot be simply enumerated from left to right. Since the BM is a recurrent process with the strong Markov property, the set of zeroes to the right of the generic excursion interval is a shifted distributional copy of the whole \(\mathcal{Z}\), independent of the part of \(\mathcal{Z}\) to the left of (and including) the excursion interval. This implies that \(\mathcal{Z}\) is a regenerative set, a property familiar from the elementary renewal theory. The scaling property of the BM, \((c^{-1/2} B_{ct}) \overset{d}{=} (B_t)\), implies the self-similarity, \(c \mathcal{Z} \overset{d}{=} \mathcal{Z}\) for \(c > 0\), i.e. the invariance of \(\mathcal{Z}\) under homotheties.

Following Pitman \([54]\), consider the composition structure \(\kappa\) derived from \(\mathcal{R} = \mathcal{Z} \cap [0, 1]\). To check the deletion property in Definition 2.1 it is convenient to modify the paintbox model in a way accounting for the self-similarity.

**Modified sampling scheme** Let \(\mathcal{Z}\) be a random self-similar subset of \(\mathcal{R}\). Fix \(n\) and let \(X_1 < X_2 < \cdots\) be the points of a unit Poisson process, independent of \(\mathcal{Z}\). The interval \([0, X_{n+1}]\) is split in gaps at points of \(\mathcal{Z}\), so we can define a composition \(\kappa_n\) of \(n\) by grouping \(X_1, \ldots, X_n\) in clusters within the gaps in \([0, X_{n+1}]\). As \(n\) varies, these compositions comprise the same composition structure, as the one induced by the standard paintbox construction with \(\mathcal{Z} \cap [0, 1]\), because (i) the vector \((X_1/X_{n+1}, \ldots, X_n/X_{n+1})\) is distributed like the vector of \(n\) uniform order statistics \((U_{1:n}, \ldots, U_{n:n})\) and (ii) by self-similarity, \(\mathcal{Z} \setminus X_{n+1} \overset{d}{=} \mathcal{Z}\).

Note that, because the locations of ‘balls’ vary with \(n\), the model secures a weak consistency of \(\kappa_n\)’s, but does not produce strongly consistent ordered set partitions \(K_n\). Applying the modified scheme in the BM case, the deletion property is obvious from the regeneration of \(\mathcal{Z}\), homogeneity of the Poisson process and the self-similarity of \(\mathcal{Z}\).

We shall return to this BM example on several occasions.

3.3. Regenerative sets and subordinators

In the stick-breaking case the regeneration property of the induced composition structure \(\kappa\) followed from the observation that \(\mathcal{R}\) remains in a sense the same when its left meander is truncated. This could not be applied in the case of composition structure derived from the BM, since the leftmost gap does not exist. By a closer look it is seen that a weaker property of \(\mathcal{R}\) would suffice. For a given closed \(\mathcal{R} \subset [0, 1]\) define the ‘droite’ point \(Z_x := \min\{\mathcal{R} \cap [x, 1]\}, x \in [0, 1]\), which is the right endpoint of the gap covering \(x\) (or \(Z_x = x\) in the event \(x \in \mathcal{R}\)).
Definition 3.3. A random closed set $\mathcal{R} \subset [0,1]$ is called \textit{multiplicatively regenerative} (m-regenerative for short) if $Z_x$ is independent of $(1 - Z_x)^{-1}((\mathcal{R} \cap [Z_x,1]) - Z_x)$ and, given $Z_x < 1$, the distributional identity is fulfilled

$$
(\mathcal{R} \cap [Z_x,1]) - Z_x \overset{d}{=} \mathcal{R}
$$

for every $x \in [0,1)$.

\textbf{Remark} We do not require explicitly the independence of $[0,Z_x] \cap \mathcal{R}$ and $(1 - Z_x)^{-1}((\mathcal{R} \cap [Z_x,1]) - Z_x)$, which would correspond to the conventional regeneration property in the additive theory. In fact, this apparently stronger property follows from the weaker independence property in Definition 3.3 due to connection with the bulk-deletion properties of composition structures. See [30] for details.

For m-regenerative paintbox $\mathcal{R}$ the deletion property of $\kappa$ follows by considering the gap that covers $U_{n1} = \min(U_1, \ldots, U_n)$. Then $q(n : \cdot)$ is the distribution of the rank of the largest order statistic in this gap.

To relate Definition 3.3 with the familiar (additive) concept of regenerative set, recall that a \textit{subordinator} $(S_t, t \geq 0)$ is an increasing right-continuous process with $S_0 = 0$ and stationary independent increments (Lévy process) [8]. The fundamental characteristics of subordinator are the Lévy measure $\hat{\nu}$ on $(0, \infty]$, which controls the intensity and sizes of jumps, and the drift coefficient $d \geq 0$ responsible for a deterministic linear drift component. The distribution is determined by means of the Laplace transform

$$
\mathbb{E}[\exp(-\rho S_t)] = \exp[-t\Phi(\rho)], \quad \rho \geq 0,
$$

where the Laplace exponent is given by the Lévy-Khintchine formula

$$
\Phi(\rho) = \rho d + \int_{(0,\infty]} (1 - e^{-\rho y})\hat{\nu}(dy), \quad (3.13)
$$

which is also the general representation of Bernstein functions [60]. The Lévy measure must satisfy the condition $\Phi(1) < \infty$ which implies $\nu(y,\infty] < \infty$ and also restricts the mass near 0, to avoid immediate passage of the subordinator to $\infty$. A positive mass at $\infty$ is allowed, in which case $(S_t)$ is a killed subordinator, that is jumping to the absorbing state $\infty$ at some random exponential time with rate $\hat{\nu}\{\infty\}$. Two standard examples of subordinators are

1. Stable subordinators with parameter $0 < \alpha < 1$, characterised by

$$
\hat{\nu}(dy) = \frac{c\alpha}{\Gamma(1-\alpha)}y^{-\alpha-1}dy, \quad d = 0, \quad \Phi(\rho) = c\rho^\alpha.
$$

2. Gamma subordinators with parameter $\theta > 0$, characterised by

$$
\hat{\nu}(dy) = cy^{-1}e^{-\theta y}dy, \quad d = 0, \quad \Phi(\rho) = c\log(1 + \rho/\theta).
$$

The constant $c > 0$ can be always taken equal to 1 by a linear time-change.
Let $\tilde{R} = \{ S_t, t \geq 0 \}$ be the closed range of a subordinator. By properties of the increments, $\tilde{R}$ is regenerative: for $Z_y$, the ‘droite’ point at $y > 0$, conditionally given $Z_y < \infty$, the random set $(\tilde{R} - Z_y) \cap [0, \infty]$ is distributed like $\tilde{R}$ and is independent of $[0, Z_y] \cap \tilde{R}$ and $Z_y$. Also the converse is true: by a result of Maisonneuve [51] every regenerative set is the closed range of some subordinator, with $(\tilde{\nu}, d)$ determined uniquely up to a positive multiple.

Call the increasing process $(1 - \exp(-S_t), t \geq 0)$ multiplicative subordinator, and let $R = 1 - \exp(-\tilde{R})$ be its range. The regeneration property of $\tilde{R}$ readily implies that $R$ is $m$-regenerative. As time passes, the multiplicative subordinator proceeds from 0 to 1, thus it is natural to adjust the Lévy measure to the multiplicative framework by transforming $\tilde{\nu}$, by the virtue of $y \to 1 - e^{-y}$, in some measure $\nu$ on $(0, 1]$. Measure $\nu$ accounts for a kind of continuous-time stick-breaking: when at some time the remaining part of a stick is of length $z$, a piece $z(1-y)$ is split away at rate $\nu(dx)$.

We shall still call $\nu$ the Lévy measure where there is no ambiguity. In these terms the Lévy-Khintchine formula becomes

$$\Phi(\rho) = \rho d + \int_0^1 \{1 - (1 - x)\rho\} \nu(dx).$$

(3.14)

For integer $1 \leq m \leq n$ introduce also the binomial moments of $\nu$

$$\Phi(n : m) = \binom{n}{m} \int_0^1 x^m (1 - x)^{n-m} \nu(dx) + 1(m = 1)\, nd$$

(where $1(\cdots)$ stands for indicator), so that $\Phi(n) = \sum_{m=1}^n \Phi(n : m)$. According to one interpretation of (3.13), $\Phi(\rho)$ is the probability rate at which the subordinator passes through independent exponential level with mean $1/\rho$. Similarly, $\Phi(n)$ is the rate at which the multiplicative subordinator passes through $U_{n:1}$ and $\Phi(n : m)$ is the rate to jump from below $U_{n:1}$ to a value between $U_{n:m}$ and $U_{n:m+1}$. From this, the probability that the first passage through $U_{n:1}$ covers $m$ out of $n$ uniform points is equal to

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

(3.15)

which is the general representation for decrement matrix of a regenerative composition structure associated with $m$-regenerative set. The proper case corresponds to the zero drift, $d = 0$, then passage through a level can only occur by a jump.

In the case of finite $\tilde{\nu}$ and $d = 0$ the subordinator is a compound Poisson process with no drift. Scaling $\nu$ to a probability measure, the range of $(1 - \exp(-S_t), t \geq 0)$ is a stick-breaking set with the generic factor $W$ distributed according to $\nu$; then (3.15) becomes (3.6).

The connection between regenerative compositions structures and regenerative sets also goes in the opposite direction.
Theorem 3.4. [30] Every regenerative composition structure can be derived by the paintbox construction from the range of a multiplicative subordinator, whose parameters \((\nu, d)\) are determined uniquely up to a positive multiple.

Sketch of proof Sampling consistency together with the regeneration imply that the first \(n\) rows of the minor \((q(n' : \cdot), n' \leq n)\) are uniquely determined by the last row \(q(n : \cdot)\) via formulas

\[
q(n' : m') = \frac{q_0(n' : m')}{{1 - q_0(n' : 0)}}, \quad 1 \leq m' \leq n', \tag{3.16}
\]

\[
q_0(n' : m') = \sum_{m=1}^{n} q(n : m) \binom{n-m}{n'} \frac{m^n}{m}, \quad 0 \leq m' \leq n'. \tag{3.17}
\]

Think of \(\kappa_n\) as allocation of \(F_n\) balls in box labeled \(B\), and \(n - F_n\) balls in other boxes. Formula (3.17) gives the distribution of the number of balls remaining in \(B\) after \(n - n'\) balls have been removed at random without replacement, with account of the possibility \(m' = 0\) that \(B\) may become empty. Formula (3.16) says that the distribution of \(F_{n'}\) is the same as that of the number of balls which remain in \(B\) conditionally given that at least one ball remains. This relation of \(F_{n'}\) is counter-intuitive, because sampling may eliminate the first block of \(\kappa_n\) completely (equivalently, the first block of \(K_n\) may have no representatives in \([n']\)).

Invoking the simplest instance of (3.16), with \(n' = n - 1\), we have

\[
\frac{q(n : m)}{1 - q(n + 1 : 1)/(n + 1)} = \frac{m + 1}{n + 1} q(n + 1 : m + 1) + \frac{n + 1 - m}{n + 1} q(n + 1 : m).
\]

This is a nonlinear recursion, but passing to formal homogeneous variables \(\Phi(n : m)\) and using the substitution \(q(n : m) = \Phi(n : m)/\Phi(n)\) with \(\Phi(n) := \sum_{m=1}^{n} \Phi(n : m)\) result in the linear relation

\[
\Phi(n : m) = \frac{m + 1}{n + 1} \Phi(n + 1 : m + 1) + \frac{n - m + 1}{n + 1} \Phi(n + 1 : m).
\]

Equivalently, in terms of the iterated differences

\[
\Phi(n : m) = \binom{n}{m} \sum_{j=0}^{m} (-1)^{j+1} \binom{m}{j} \Phi(n - m + j), \quad 1 \leq m \leq n.
\]

The positivity condition \(\Phi(n : m) \geq 0\) implies that the sequence \((\Phi(n), n \geq 0)\) (where \(\Phi(0) = 0\)) must be completely alternating [7], i.e. its iterated differences have alternating signs. The latter also means that the difference sequence \((\Phi(n + 1) - \Phi(n), n \geq 0)\) is completely monotone, hence by the famous Hausdorff theorem \(\Phi(n)\)'s are representable as moments of some finite measure on \([0, 1]\). From this (3.14) follows for integer values of \(\rho\) with some \((\nu, d)\). The latter secures (3.14) for arbitrary \(\rho > 0\) by the uniqueness of interpolation.
Interestingly, the argument only exploits a recursion on \( q \), hence avoids explicit limit transition from \( x \) to \( \mathcal{R} \), as one could expect by analogy with Theorem 2.2. See [14, 30, 33] for variations.

We can also view \( F(t) = 1 - \exp(-S_t) \) as a random distribution function on \( \mathbb{R}_+ \) used to construct a composition by sampling from \( F \), as in the species sampling problem described in Section 1. These so-called neutral to the right priors have extensive applications in Bayesian statistics [44].

**Additive paintbox** It is sometimes convenient to induce regenerative composition structures using a subordinator \((S_t)\) to create the gaps. Then, independent unit-rate exponential variables \( E_1, E_2, \ldots \) should be used in the role of balls, instead of uniform \( U_j \)’s in the multiplicative framework.

Formula (3.15) can be re-derived by appealing to the potential measure \( \nabla \) of subordinator. Heuristically, think of \( \nabla(dy) \) as of probability to visit location \( y \) at some time, and of \( \nu \) as distribution of size of a generic jump of the subordinator. The probability that the first part \( m \) of composition is created by visiting \( y \) by a jump of given size \( z \) is then the product of \( 1 - e^{-ny} \nabla(dy) \) and \( (1 - e^{-z})^m e^{-(n-m)\tilde{\nu}(dz)} \). Taking into account the formula for Laplace transform of the potential measure [8]

\[
\int_0^\infty (1 - e^{-\rho y})\nabla(dy) = \frac{1}{\Phi(\rho)},
\]

we arrive at (3.15) by integration. The compensation formula for Poisson processes [8, p. 76] is needed to make this argument rigorous.

The advantage of working with \( \mathbb{R}_+ \) is that the regeneration property involves no scaling. A disadvantage is that the asymptotic frequency of balls within the walls \((a, b)\) is the exponential probability \( e^{-b} - e^{-a} \), as compared to the size of gap in the multiplicative representation on \([0, 1]\).

In particular, for Ewens’ composition structures the subordinator \((S_t)\) is a compound Poisson process with the jump distribution exponential(\( \theta \)), so the range \( \tilde{R} \) of \((S_t)\) is a homogeneous Poisson point process with rate \( \theta \), and \( R \) is an inhomogeneous Poisson point process with rate function \( \theta/(1 - x) \) on \([0, 1]\).

**Example 3.5.** Consider the infinite Lévy measure on \([0, 1]\) with density \( \nu(dx) = x^{-1}(1-x)^{\theta-1}dx \). Denoting \( h_\theta(n) = \sum_{k=1}^n \frac{1}{\theta+k-1} \) the generalised harmonic numbers we compute

\[
q(n : m) = \frac{n!(\theta)_{n-m}}{m(n-m)!(\theta)_nh_\theta(n)}, \quad p^\circ(\lambda_1, \ldots, \lambda_k) = \frac{f^\circ(\lambda)}{(\theta)_n} \prod_{j=1}^k \frac{1}{h_\theta(\Lambda_j)}.
\]

This composition structure appears as the limit of stick-breaking composition structures (3.8) as \( \gamma \to 0 \). Although the CPF looks very similar to Ewens’ (3.10), there is no simple product formula for the associated partition structure, even in the case \( \theta = 1 \).

**Example 3.6.** (Regenerative hook compositions) Hook composition structures are induced by killed pure-drift subordinators with \( \nu(dx) = \delta_1(dx) \) and \( d \in \)
[0, ∞]. They have decrement matrices with the only nonzero entries

\[ q(n : n) = \frac{1}{1 + nd}, \quad q(n : 1) = \frac{nd}{1 + nd}. \]

The compositions \( \kappa_n \) only assume values like \((1, 1, \ldots, 1, m)\). Ferrer’s diagrams of the associated partitions \((m, 1, \ldots, 1)^{\downarrow}\), are \(\Gamma\)-shaped hooks.

The hook compositions bridge between the pure-singleton composition (with \( R = [0, 1] \)) and the trivial one-block composition (with \( R = \{0, 1\} \)). For arbitrary composition structure with some \( \Lambda \)-Levy exponent \( \Phi \) we can construct a similar deformation by adding atomic component \( \beta \delta_1(dx) \) to the \( \Lambda \)-Levy measure; this results in a family of decrement matrices

\[ \Phi(n : m) + \beta \delta_1(m = n), \]

\[ \Phi(n) + \beta, \quad (3.18) \]

with the one-block composition appearing in the limit \( \beta \to \infty \).

**Sliced splitting** We introduce another kind of parametric deformation of a subordinator. Let \( S = (S_t) \) be a subordinator with range \( \tilde{\mathbb{R}} \), and let \( X_1 < X_2 < \cdots \) be the points of homogeneous Poisson process with density \( \theta \). Take \( S(j), \) \( j \geq 0 \), to be independent copies of \( S \), also independent of the Poisson process. We construct the path of interrupted subordinator \( S(\theta) \) by shifting and glueing pieces of \( S(j) \)’s in one path.

Run \( S(0) \) until the passage through level \( X_1 \) at some time \( T_1 \), so \( S_{T_1-}(0) < X_1 \leq S_{T_1}(0) \). Leave the path of the process south-west of the point \((T_1, X_1)\) as it is, and cut the rest north-east part of the path. At time \( T_1 \) start the process \((S_t(1) + X_1, t \geq T_1)\) and let it running until passage through \( X_2 \). Iterate, creating partial paths running from \((T_j, X_j)\) to \((T_{j+1}, X_{j+1})\). From the properties of subordinators and memoryless property of overshoots of the Poisson process, one sees that \( S(\theta) \) is indeed a subordinator.

The range of \( S(\theta) \),

\[ \tilde{\mathbb{R}}(\theta) := \bigcup_{j \geq 0} [X_j, X_{j+1}) \cap (X_j + \tilde{\mathbb{R}}(j)), \]

can be called sliced splitting. First \( \mathbb{R}_+ \) is split at locations \( X_j \), then each gap \((X_j, X_{j+1})\) is further split at points of \( \tilde{\mathbb{R}}(j) + X_j \cap (X_j, X_{j+1}) \) where \( \tilde{\mathbb{R}}(j) \overset{d}{=} \tilde{\mathbb{R}} \) are i.i.d. The range of \( 1 - \exp(-S(\theta)) \) can be constructed by a similar splitting of the gaps between the points \( 1 - \exp(-X_j) \), which are the atoms of a Poisson point process with density \( \theta/(1 - x) \).

Denote, as usual, \( \Phi, \tilde{\nu} \) the characteristics of \( S \), and \( \Phi_\theta, \tilde{\nu}_\theta \) the characteristics of \( S(\theta) \). Then we have

\[ \Phi_\theta(\rho) = \frac{\rho}{\rho + \theta} \Phi(\rho + \theta), \quad \tilde{\nu}_\theta[y, \infty] = e^{-\theta y} \tilde{\nu}(y, \infty). \quad (3.19) \]

To see this, a heuristics is helpful to guess the passage rate through exponential level. Denote \( E_\rho, E_\theta \) independent exponential variables with parameters \( \rho, \theta \).
The process $S^\theta$ passes the level $E_\rho$ within infinitesimal time interval $(0, t)$ when $S_t > \min(E_\rho, E_\theta) = E_\rho$. The inequality $S_t > \min(E_\rho, E_\theta) \overset{d}{=} E_{\rho+\theta}$ occurs with probability $\Phi(\rho + \theta)t + o(t)$, and probability of the event $E_\theta < E_\rho$ is $\rho/(\rho + \theta)$.

A generalisation of (3.19) for spectrally negative Lévy processes appears in [49, Lemma 2.1]. Other examples of factorisation of Bernstein functions are found in [60].

The Green matrix

For a sequence of compositions $\kappa = (\kappa_n)$ which, in principle, need not be consistent in any sense we can define $g(n, j)$ as the probability that a ‘1’ stays in position $j$ of the binary code of $\kappa_n$. That is to say, $g(n, j)$ is the probability that the parts of $\kappa_n$ satisfy $\lambda_1 + \cdots + \lambda_i - 1 = j - 1$ for some $i \geq 1$. Call $(g(n, j), 1 \leq j \leq n, n \in \mathbb{N})$ the Green matrix of $\kappa_n$. For $\kappa$ a regenerative composition structure, $g(n, j)$ is the probability that the Markov chain $Q^{\downarrow}n$ ever visits state $n + 1 - j$, and we have an explicit formula in terms of the Laplace exponent (see [30])

$$g(n, j) = \Phi(j) \binom{n}{j} \sum_{i=0}^{j-1} \binom{j-1}{i} \frac{(-1)^i}{\Phi(j+i)}.$$  

(3.20)

3.4. Regenerative compositions from the two-parameter family

Let $\pi$ be the two-parameter partition structure with PPF (2.4). Sometimes notation PD($\alpha, \theta$) is used for the law of frequencies $\mathcal{S}$, where PD stands for Poisson-Dirichlet, and sometimes in applications this law is called ‘the Pitman-Yor prior’ after [59]. Formulas for PD($\alpha, \theta$) are difficult, but the sequence of frequencies in size-biased order can be obtained by inhomogeneous stick-breaking scheme (3.4) with independent $W_j \overset{d}{=} \text{beta}(1 - \alpha, \theta + j\alpha)$. The law of this size-biased arrangement of $\mathcal{S} \overset{d}{=} \text{PD}(\alpha, \theta)$ is a probability measure on the simplex $\{(x_1, x_2, \ldots) : \sum x_j = 1, x_j > 0 \}$ known as the two-parameter Griffiths-Engen-McClosey distribution GEM($\alpha, \theta$).

We will see that for $0 \leq \alpha < 1$ and $\theta \geq 0$ and only for these values of the parameters the parts of $\pi_n$ can be arranged in a regenerative composition structure for $n = 1, 2, \ldots$.

Define a (multiplicative) Lévy measure $\nu$ on $[0, 1]$ by the formula for its right tail

$$\nu[x, 1] = x^{-\alpha}(1 - x)^{\theta}. \tag{3.21}$$

The density of this measure is a mixture of two beta-type densities, and in the case $\theta = 0$ there is a unit atom at 1. The associated Laplace exponent is

$$\Phi(\rho) = \rho B(1 - \alpha, \rho + \theta) = \frac{\rho \Gamma(1 - \alpha) \Gamma(\rho + \theta)}{\Gamma(\rho + 1 - \alpha + \theta)}, \tag{3.22}$$

and the binomial moments are

$$\Phi(n : m) = \binom{n}{m} (\alpha B(m - \alpha, n - m + 1 + \theta) + \theta B(m + 1 - \alpha, n - m + \theta)).$$
so there exists a regenerative composition structure $\kappa$ with the decrement matrix

$$q(n : m) = \frac{\Phi(n : m)}{\Phi(n)} = \binom{n}{m} \frac{(1 - \alpha)_{m-1} \cdot (n - m)\alpha + m\theta}{(\theta + n - m)_m n}. \quad (3.23)$$

It is a good exercise in algebra to show that the symmetrisation (2.3) of the product-form CPF with decrement matrix (3.23) is indeed the two-parameter PPF (2.4).

Like their unordered counterparts, the two-parameter regenerative compositions have many interesting features. Three subfamilies are of special interest and, as the experience shows, should be always analysed first.

**Case** $(0, \theta)$ for $\theta \geq 0$. This is the ESF case (3.10), with $\nu$ being the beta$(1, \theta)$ distribution. The parts of composition appear in the size-biased order, the gaps of $R_c$ too.

**Case** $(\alpha, 0)$ for $0 < \alpha < 1$ is a remarkable instance which has features of both multiplicative and additive regeneration. In this case

$$\nu(dx) = \alpha x^{-\alpha-1} dx + \delta_1(dx)$$

is an infinite measure with a unit atom at 1. The composition structure is directed by $R = Z \cap [0, 1]$, where $Z$ is the range of stable subordinator. On the other hand, $R$ can be also obtained as the range of multiplicative subordinator $(1 - \exp(-S_t), t \geq 0)$, where $(S_t)$ is the subordinator with Lévy measure

$$\tilde{\nu}(dy) = \alpha (1 - e^{-y})^{-\alpha-1} e^{-y} dy + \delta_\infty(dy).$$

The product formula (3.1) specialises to

$$p^\circ(\lambda_1, \ldots, \lambda_k) = f^\circ(\lambda^0)\lambda_k\alpha^{k-1} \prod_{j=1}^{k} \frac{(1 - \alpha)_{\lambda_j-1}}{\lambda_j!}. \quad (3.24)$$

This composition structure was introduced in [54], where $Z$ was realised as the zero set of a Bessel process of dimension $2 - 2\alpha$. For $\alpha = 1/2$ this is the zero set of BM.

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

$$q(n : m) = h(m) \text{ if } m < n \text{ and } q(n : n) = 1 - \sum_{m=1}^{n-1} h(m). \quad (3.24)$$

This means that 1’s in the binary code of $\kappa_n$ can be identified with the set of sites within $1, \ldots, n$ visited by a positive random walk on integers (lattice renewal process), with the initial state 1. Specifically,

$$h(m) = \frac{\alpha(1 - \alpha)_{m-1}}{m!}, \quad (3.25)$$

and $q(n : n) = (1 - \alpha)_{n-1}/(n - 1)!$. 

The arrangement of parts of $\pi_n$ in a composition is obtained by placing a size-biased part of $\pi_n$ in the last position in $\kappa_n$, then by shuffling the remaining parts uniformly at random to occupy all other positions. Exactly the same rule applies on the paintbox level: for $S$ following PD$(\alpha, 0)$, a term is chosen by the size-biased pick and attached to 1 as the meander, then the remaining gaps are arranged in the exchangeable order.

**Case** $(\alpha, \alpha)$ for $0 < \alpha < 1$. The associated regenerative set has zero drift and the Lévy measure

$$\tilde{\nu}(dy) = \alpha(1 - e^{-y}) - \alpha^{-1}e^{-\alpha y} dy \quad y \geq 0,$$

this is the zero set of an Ornstein-Uhlenbeck process. The corresponding range of multiplicative subordinator can be realised as the zero set of a Bessel bridge of dimension $2 - 2\alpha$; in the case $\alpha = 1/2$ this is the Brownian bridge.

The parts of $\kappa_n$ are identifiable with the increments of a random walk with the same step distribution $h$ as in (3.25) for the $(\alpha, 0)$ case, but now conditioned on visiting the state $n + 1$. The CPF is

$$p_{\alpha, \alpha}^o(\lambda_1, \ldots, \lambda_k) = f^o(\lambda^o) \frac{\alpha^k}{(\alpha)_n} \prod_{j=1}^k (1 - \alpha)_{\lambda_i - 1}. \quad (3.26)$$

This function is symmetric for each $k$, which implies that the parts of each $\kappa_n$ are in the exchangeable random order. This confirms the known fact that the excursion intervals of a Bessel bridge appear in completely random order.

Due to symmetry, the transition rule from $\kappa_n$ to $\kappa_{n+1}$ is a simple variation of the Chinese restaurant scheme. Now the tables are ordered in a row. Given $\kappa_n = (\lambda_1, \ldots, \lambda_k)$, ball $n + 1$ is placed in one of the existing boxes with chance $(\lambda_j - \alpha)/(n + \alpha)$ as in the unordered scheme, and when a new table is to be occupied, the box is placed with equal probability to the right, to the left or in-between any two of $k$ boxes occupied so far.

In the case $(\alpha, 0)$, there is a right meander appearing in consequence of killing at rate $\tilde{\nu}(\infty) = 1$. Removing the atom at $\infty$ yields another $m$-regenerative set (not in the two-parameter family) obtained by (i) splitting $[0, 1]$ using beta$(1, \theta)$ stick-breaking, (ii) fitting in each gap $(Y_{j-1}, Y_j)$ a scaled copy of the $(\alpha, 0)$ $m$-regenerative set. The decrement matrix is (3.18), with $\Phi$ like for the $(\alpha, 0)$ $m$-regenerative set and $\beta = -1$. A discrete counterpart, composition $\kappa_n$, is realisable as a path of a random walk with a barrier, but the CPF for this composition structure has no simple formula.

**The $m$-regenerative set with parameters** $0 < \alpha < 1$, $\theta > 0$ is constructable from the sets $(0, \theta)$ and $(\alpha, 0)$ by the sliced splitting method underlying (3.19). To define a multiplicative version of the two-level paintbox, and to have a relation like (3.19), first split $[0, 1]$ at points $Y_i$ of the Poisson process with density $\theta/(1 - x)$ as in the Ewens case, (recall that this is the same as stick-breaking with beta$(1, \theta)$ factor $W$). Then for each $j$ choose an independent copy of the
α-stable regenerative set starting at $Y_{j-1}$ and abruptly at $Y_j$, and use this copy to split $(Y_{j-1}, Y_j)$.

The resulting m-regenerative set corresponds to the $(\alpha, \theta)$ composition structure, so (3.19) becomes

$$\Phi_{\alpha, \theta}(\rho) = \frac{\rho}{\rho + \theta} \Phi_{\alpha, 0}(\rho + \theta),$$

which is trivial to check. As another check, observe that the structural distribution beta$(1 - \alpha, \alpha + \theta)$ is the Mellin convolution of beta$(1, \theta)$ and beta$(1 - \alpha, \alpha)$, as it must be for the two-level splitting scheme.

The construction is literally the same on the level of finite compositions [32, 61]. First a regenerative Ewens $(0, \theta)$ composition of $n$ is constructed, then each part is independently split in a sequence of parts according to the rules of the regenerative $(\alpha, 0)$-composition.

4. Regenerative partition structures and the problem of arrangement

We discuss next connections between regenerative composition structures and their associated partition structures. One important issue is the uniqueness of the correspondence.

4.1. Structural distributions

For arbitrary random sequence of frequencies $S$, let $R$ be any arrangement in a paintbox, i.e. with $R^c$ related to $S$ via (2.2). Let $\tilde{P}$ be the size of the gap covering the independent uniform point $U_1$, with the convention that $\tilde{P} = 0$ in the event $U_1 \in R$. We shall understand $\tilde{P}$ as a size-biased pick from $S$, this agrees with the (unambiguous) definition in the proper case and extends it when the sum of positive frequencies may be less than 1. Obviously, the particular choice of $R$ with gap-sizes $S$ is not important.

The law of $\tilde{P}$ is known as the structural distribution of $S$. Most properties of this distribution readily follow from the fact that it is a mixture of delta measures $\sum_j s_j \delta_{s_j}(dx) + \left(1 - \sum_j s_j\right) \delta_0(dx)$. In particular, the $(n - 1)$st moment of $\tilde{P}$ is the probability that $\pi_n$ is the trivial one-block composition $(n)$ or, what is the same, that $\pi_n = (n)^4$:

$$p(n) = \mathbb{E}[\tilde{P}^{n-1}].$$

In general, there can be many partition structures which share the same structural distribution, but for the regenerative composition structures the correspondence is one-to-one. Indeed, we have

$$p(n) = q(n : n) = \frac{\Phi(n : n)}{\Phi(n)}.$$
With some algebra one obtains a recursion for the Laplace exponent

\[ \Phi(n)(p(n) + (-1)^n) = \sum_{j=1}^{n-1} (-1)^{j+1} \binom{n}{j} \Phi(j), \]

which shows that the moments sequence \((p(n), n \in \mathbb{N})\) determines \((\Phi(n), n \in \mathbb{N})\) uniquely up to a positive multiple, hence determines the decrement matrix \(q\).

Explicit expressions of the entries of \(q\) through the \(p(n)\)'s are some rational functions in \(p(n)\)'s, for instance \(q(3:2) = \frac{2p(2) - 3p(3) + p(2)p(3)}{1 - p(2)}\).

Because the moments \(p(n)\) are determined by the sizes of gaps and not by their arrangement, we conclude that

**Theorem 4.1.** [30] Each partition structure corresponds to at most one regenerative composition structure. Equivalently, for random frequencies \(S\) there exists at most one distribution for a \(m\)-regenerative set \(R\) with \((R^c)^\downarrow \overset{d}{=} S\).

In principle, one can determine if some PPF \(p\) corresponds to a regenerative CPF \(p^\circ\) by computing \(q\) formally from the one-block probabilities \(p(n)\)'s, then checking positivity of \(q\), and if it is positive then comparing the symmetrised PPF \((3.1)\) corresponding to \(q\) with \(p\). This method works smoothly in the case of Ewens-Pitman two-parameter partitions. For the \((\alpha, \theta)\) partition structures the structural distribution is \(\beta(1-\alpha, \theta + \alpha)\) and \(p(n)\) is equal to

\[ E[\tilde{P}^{n-1}] = \frac{\theta(1-\alpha)_{n-1}}{(\theta)_n}. \]

Computing formally \(q\) from the \(p(n)\)'s we arrive at \(q\) coinciding with \((3.23)\). However, a decrement matrix must be nonnegative, which is not the case for some values of the parameters:

**Theorem 4.2.** [30] Every Ewens-Pitman partition structure with parameters in the range \(0 \leq \alpha \leq 1, \theta \geq 0\) has a unique arrangement as a regenerative composition structure. For other values of the parameters such arrangement does not exist.

Actually, it is evident that any partition structure of the ‘discrete subfamily’ with \(\alpha < 0\) in \((2.5)\) cannot be regenerative just because the number of parts in each \(\pi_n\) is bounded by \(-\theta/\alpha\).

### 4.2. Partition structures invariant under deletion of a part

Relations \((3.2)\) and \((3.3)\) say that partition structures inherit a deletion property from the parent regenerative compositions. In this section we discuss the reverse side of this connection, which puts the regeneration property in the new light. The main idea is that if a partition structure \(\pi\) has a part-deletion property, then the iterated deletion creates order in a way consistent for all \(n\), thus canonically associating with \(\pi\) a regenerative composition structure. The
arrangement problem may be seen as connection between the intrinsic order of composition and the sampling (size-biased) order.

Let \( \pi \) be a partition structure. A random part of \( \pi_n \) is an integer random variable \( P_n \) which satisfies \( P_n \in \pi_n \). The joint distribution of \( \pi_n \) and \( P_n \) is determined by the PPF and some deletion kernel \( d(\lambda^1, m) \), which specifies the conditional distribution of \( P_n \) given partition \( \pi_n \)

\[
p(\lambda^1) d(\lambda^1, m) = \mathbb{P}(\pi_n = \lambda^1, P_n = m), \quad |\lambda^1| = n. \tag{4.1}
\]

For each \( n = 1, 2, \ldots \) the distribution of \( P_n \) is then

\[
q(n : m) = \mathbb{P}(P_n = m) = \sum_{\{\lambda^1 : |\lambda^1| = n, m \in \lambda^1\}} d(\lambda^1, m) p^1(\lambda^1), \quad 1 \leq m \leq n. \tag{4.2}
\]

The formulas differ from (3.2) and (3.3) in that now they refer to some abstract ‘random part’ \( P_n \) of unordered structure. The requirement that \( P_n \) is a part of \( \pi_n \) makes

\[
\sum_{\text{distinct } m \in \lambda^1} d(\lambda^1, m) = 1.
\]

**Definition 4.3.** Call a partition structure \( \pi = (\pi_n) \) regenerative if, for each \( n \), there exists a joint distribution for \( \pi_n \) and its random part \( P_n \) such that for each \( 1 \leq m < n \) conditionally given \( P_n = m \) the remaining partition \( \pi_n \setminus m \) of \( n - m \) has the same distribution as \( \pi_{n-m} \). Call \( \pi \) regenerative w.r.t. \( q \) if \( q(n : \cdot) \) is the law of \( P_n \), which means that

\[
p(\lambda^1) d(\lambda^1, m) = q(n : m) p(\lambda^1 \setminus \{m\}), \quad n = 1, 2, \ldots, \tag{4.3}
\]

The regeneration means that \( \pi \) satisfies the stochastic recursion

\[
\pi_n \setminus P_n \overset{d}{=} \pi_{n-P_n}, \quad n = 1, 2, \ldots, \tag{4.4}
\]

with the same conventions as in (3.12), but with the size-biased pick replaced by \( P_n \overset{d}{=} P'_n \).

**Example** (Hook partition structures) This is a continuation of Example 3.6. Call \( \lambda^1 \) a hook partition if only \( \lambda_1 \) may be larger than 1, for instance \((4, 1, 1, 1)^4\). For every deletion kernel with the property

\[
d(\lambda^1, 1) = 1 \quad \text{if} \quad 1 \in \lambda^1,
\]

can be shown that the only partition structures regenerative w.r.t. such \( d \) are those supported by hook partitions, and they have \( q(n : n) = 1/(1 + nd) \), \( q(n : 1) = nd/(1 + nd) \) for some \( d \in [0, \infty] \). A regenerative partition structure is of the hook type if and only if \( p((2, 2)^4) = 0 \).

**Theorem 4.4.** [29] If a partition structure is regenerative and satisfies the condition \( p((2, 2)^4) > 0 \), then \( q \) uniquely determines \( p \) and \( d \), and \( p \) uniquely determines \( q \) and \( d \). Equivalently, if a regenerative partition structure is not of the hook type then the corresponding deletion kernel is unique.
Observe that all hook partitions share the same deletion kernel(s). We will see next that for each deletion kernel associated with the two-parameter family there is a one-parameter subfamily regenerative with respect to this kernel. One can wonder if it is always the case that a nontrivial regenerative partition structure has a deletion kernel common to a family of such structures (the problem is due to Jim Pitman).

4.3. Deletion kernels for the two-parameter family

For Ewens’ composition structures (3.10) the deletion kernel is the size-biased pick

\[ d_0(\lambda^i, m) = \frac{k_m m}{n}, \text{ where } k_m = \# \{ j : \lambda_j = m \}, \ n = |\lambda^i|. \]

The factor \( k_m \) appears since the kernel specifies the chance to choose one of the parts of given size \( m \), rather than a particular part of size \( m \). Rephrasing Kingman’s characterisation \[48\]: if a partition structure is regenerative w.r.t. \( d_0 \), then the PPF is the the ESF (3.11) with some \( \theta \in [0, \infty) \).

For the regenerative composition structures of the two-parameter family (with nonnegative \( \alpha, \theta \)) the deletion kernel is one of

\[ d_\tau(\lambda^i, m) := \frac{k_m (n - m) \tau + m(1 - \tau)}{(1 - \tau + (k - 1)\tau)}, \quad \tau \in [0, 1], \]

where \( k = \sum_m k_m \) and \( n = |\lambda^i| \). Kingman’s characterisation of the ESF is a special case of a more general result:

**Theorem 4.5.** [29] Fix \( \tau \in [0, 1] \). The only partition structures that are regenerative w.r.t. deletion kernel \( d_\tau \) are the \((\alpha, \theta)\) partition structures with

\[ 0 \leq \alpha \leq 1, \ \theta \geq 0 \quad \text{and} \quad \alpha/(\theta + \alpha) = \tau. \]

Three distinguished subfamilies are characterised as follows:

1. The kernel \( d_0 \) is the size-biased choice; only \((0, \theta)\) partition structures are regenerative w.r.t. \( d_0 \).
2. The kernel \( d_{1/2} \) is a uniform random choice of a part; only \((\alpha, \alpha)\) partition structures are regenerative w.r.t. \( d_{1/2} \).
3. The kernel \( d_1 \) can be called cosize-biased deletion, as each (particular) part \( m \in \lambda^i \) is selected with probability proportional to \( |\lambda^i| - m \); only \((\alpha, 0)\) partitions are regenerative w.r.t. \( d_1 \).

4.4. Back to regenerative compositions

The framework of regenerative partitions suggests to study three objects: the PPF \( p \), the deletion kernel \( d \) and the distribution of deleted part \( q \). Naively, it might seem that \( d \), which tells us how a part is deleted, is the right object to start with, like in Kingman’s characterisation of the ESF via the size-biased
deletion. However, apart from the deletion kernels $d_{1, \cdot}$ for the two-parameter family, and kernels related to hook partitions we do not know examples where the approach based on the kernels could be made explicit. Strangely enough, to understand the regeneration mechanism for partitions, one should ignore for a while the question how a part is deleted, and only focus on $q$ which tells us what is deleted.

Fix $n$ and let $q(n : \cdot)$ be an arbitrary distribution on $[n]$. Consider a Markov $q(n : \cdot)$-chain on the set of partitions of $n$ by which a partition $\lambda^\downarrow$ (thought of as allocation of balls in boxes) is transformed by the rules:

- choose a value of $P_n$ from the distribution $q(n : \cdot)$,
- given $P_n = m$ sample without replacement $m$ balls and discard the boxes becoming empty,
- put these $m$ balls in a newly created box.

Similarly, define a Markov $q(n : \cdot)$-chain on compositions $\kappa^\downarrow$ of $n$ with the only difference that the newly created box is placed in the first position. Obviously, the $q(n : \cdot)$-chain on compositions projects to the $q(n : \cdot)$-chain on partitions when the order of boxes is discarded.

**Lemma 4.6.** If (4.3) holds for some fixed $n$ and distribution $q(n : \cdot)$ then the law of $\pi_n$ is a stationary distribution for the $q(n : \cdot)$-chain on partitions.

**Sketch of proof** The condition (4.3) may be written as a stochastic fixed-point equation (4.4). The relation is inverted as $\pi_n - P_n \cup \{P_n\} \equiv \pi_n$.

There is an obvious parallel assertion about a random composition $\kappa_n$, which satisfies

$$\kappa_n \setminus F_n \equiv \kappa_n - F_n,$$

where $\setminus$ stands for the deletion of the first part $F_n$ with distribution $q(n : \cdot)$, and in the RHS ($\kappa_n$) is independent of $F_n \equiv F_n$.

**Lemma 4.7.** The unique stationary distribution of the $q(n : \cdot)$-chain on compositions is the one by which $\kappa_n$ follows the product formula for $1 \leq n' \leq n$ with $q(n' : \cdot)$ given by (3.16). Symmetrisation of the law of $\kappa_n$ by (2.3) gives the unique stationary distribution of the $q(n : \cdot)$-chain on partitions.

It follows that if (4.3) holds for some $n$ then it holds for all $n' \leq n$, with all $p(\lambda^\downarrow), d(\lambda^\downarrow, \cdot)$ for $|\lambda| = n'$ uniquely determined by $q(n : \cdot)$ via sampling consistency. Thus, in principle, for partitions of $n' \leq n$ the regeneration property is uniquely determined by arbitrary discrete distribution $q(n : \cdot)$ through the following steps: find first $(q(n' : \cdot), n' \leq n)$ from sampling consistency (3.16), then use the product formula for compositions (3.1), then the symmetrisation (2.3). With all this at hand, the deletion kernel can be determined from (4.1). Letting $n$ vary, the sampling consistency of all $q(n : \cdot)$’s implies that $q$ is a decrement matrix of a regenerative composition structure.

Starting with $\pi_n$, the deletion kernel determines a Markov chain on subpartitions of $\pi_n$. A part $P_n$ is chosen according to the kernel $d$ and deleted, from
the remaining partition $\pi_n \setminus P_n$ another part is chosen according to $d$ etc. This brings the parts of $\pi_n$ in the deletion order.

**Theorem 4.8.** [29] Suppose a partition structure $\pi = (\pi_n)$ is regenerative w.r.t. $q$, then

(i) $q$ is a decrement matrix of some regenerative composition structure $\kappa$,
(ii) $\pi$ is the symmetrisation of $\kappa$,
(iii) $\kappa$ is obtained from $\pi$ by arranging, for each $n$, the parts of $\pi_n$ in the deletion order.

Thus the regeneration concepts for partition and composition structures coincide. It is not clear, however, how to formulate the regeneration property in terms of the unordered frequencies $S$. The only obvious way is to compute PPF and then check if the PPF corresponds to a regenerative CPF. Moreover, the deletion kernel may have no well-defined continuous analogue. For instance, in the $(\alpha, \alpha)$ case $d_{1/2}$ is a uniformly random choice of a part from $\pi_n$, but what is a ‘random choice of a term’ from the infinite random series $S$ under PD($\alpha, \alpha$)?

### 4.5. The arrangement problem for the two-parameter family

The selection operation defined by the kernel (4.5) can be introduced in more general terms. Let $(x_1, \ldots, x_k)$ be a fixed sequence of positive numbers with sum $\sigma = \sum_{j=1}^{k} x_j$. For a fixed parameter $\tau \in [0, 1]$, define a random variable $T$ with values in $[k]$ by

$$
P(T = j \mid (x_1, \ldots, x_k)) = \frac{(1 - \tau)x_j + \tau(\sigma - x_j)}{\sigma(1 - \tau + \tau(k - 1))}.
$$

(4.6)

The random variable $x_T$ is called a $\tau$-biased pick from $x_1, \ldots, x_k$. The law of $x_T$ does not depend on the order of the sequence $(x_1, \ldots, x_k)$, and there is also a scaling invariance: $\sigma^{-1}x_T$ is a $\tau$-biased pick from $(\sigma^{-1}x_1, \ldots, \sigma^{-1}x_k)$. Thus $0$-biased pick is a size-biased pick from $(x_1, \ldots, x_k)$, choosing any particular element with probability proportional to its size. A $1/2$-biased pick is a uniform random choice from the list, as (4.6) then equals $1/k$ for all $j$. And a $1$-biased pick is the co-size biased pick, as it chooses $j$ with probability proportional to its co-size $\sigma - x_j$.

Define the $\tau$-biased permutation of $x_1, \ldots, x_k$, denoted $\text{perm}_\tau(x_1, \ldots, x_k)$, by iterating a single $\tau$-biased pick, as follows. A number $x_T$ is chosen from $x_1, \ldots, x_k$ without replacement, with $T$ distributed on $[k]$ according to (4.6), and $x_T$ is placed in position 1. Then the next number is chosen from $k - 1$ remaining numbers using again the rule of $\tau$-biased pick, and placed in position 2, etc.

The deletion order stated in Theorem 4.8 (iii) specialises for the $(\alpha, \theta)$-partitions as

$$
\kappa_n \overset{d}{=} \text{perm}_\tau(\pi_n), \quad n = 1, 2, \ldots
$$

(4.7)
where $\tau = \alpha/(\alpha + \theta)$. This solves the arrangement problem on the level of finite partitions.

The instance $\text{perm}_0$ is the size-biased permutation, which is defined more widely for finite or infinite summable sequences $(x_1, x_2, \ldots)$, and shuffles them in the same way as it shuffles $(\sigma^{-1}x_1, \sigma^{-1}x_2, \ldots)$ where $\sigma = \sum_j x_j$. In more detail, each $x_j$ is placed in position 1 with probability $x_j/\sigma$; given $T = j$, for each $i \neq j$, $x_i$ is placed in position 2 with probability $x_i/(\sigma - x_j)$, etc. For Ewens’ partitions we have the relation $\mathcal{R}^d = \text{perm}_0(\mathcal{S})$,

with the convention that $\mathcal{R}$ is defined through the complement $[0, 1] \setminus \mathcal{R}^c$, and $\mathcal{R}^c$ is obtained conditionally on $\mathcal{S} = (s_j)$ by stacking the intervals of sizes $s_j$ contiguously, in the size-biased order.

We turn now to the general $(\alpha, \theta)$ partitions. The main difficulty is that the $\tau$-biased permutation cannot be directly extended to infinite positive summable sequence $(x_1, x_2, \ldots)$, since the 'k = $\infty$' instance of (4.6) is not a proper distribution for $\tau \neq 0$. We will find, nevertheless, a construction of $\text{perm}_\tau$ which has such extension, hence can be applied to the sequence of PD/GEM($\alpha, \theta$) frequencies.

A linear order $\triangleright$ on $\mathbb{N}$ is conveniently described by a sequence of the initial ranks $(\rho_j) \in \{1\} \times \{2\} \times \cdots$, with $\rho_j = i$ if and only if $j$ is ranked $i$th smallest in the order $\triangleright$ among the integers $1, \ldots, j$. For instance, the initial ranks $1, 2, 1, 3, \ldots$ appear when $3 \triangleright 1 \triangleright 4 \triangleright 2$. For $\xi \in [0, \infty]$ define a random order $\triangleright_{\xi}$ on $\mathbb{N}$ by assuming that the initial ranks $\rho_k, k \in \mathbb{N}$, are independent, with distribution

$$P(\rho_k = j) = \frac{1}{k + \xi - 1}1(0 < j < k) + \frac{\xi}{k + \xi - 1}1(j = k), \quad k > 1.$$ 

The edge cases $\xi = 0, \infty$ are defined by continuity. The order $\triangleright_1$ is a ‘uniformly random order’, in the sense that restricting to $[n]$ we have all $n!$ orders equally likely, for every $n$. The order $\triangleright_\infty$ coincides with the standard order $<$ almost surely. For every permutation $i_1, \ldots, i_n$ of $[n]$, we have

$$P(i_1 \triangleright_{\xi} \cdots \triangleright_{\xi} i_n) = \frac{\xi^r}{\xi(\xi + 1)\ldots(\xi + n - 1)}$$

where $r$ is the number of upper records in the permutation. See [20] for this and more general random permutations, that are conditionally uniform given their record statistics. Denote by $\triangleright_{\xi}(x_1, x_2, \ldots)$ the arrangement of $x_1, x_2, \ldots$ in succession according to the $\triangleright_{\xi}$-order on the set of indices $\mathbb{N}$.

**Theorem 4.9.** For $0 \leq \alpha < 1$, $\theta \geq 0$ the arrangement of GEM($\alpha, \theta$) frequencies represented as open intervals in an independent random order $\triangleright_{\theta/\alpha}$ yields a paintbox $\mathcal{R}$ representable as the range of a multiplicative subordinator determined by (3.21), (3.22).

This solution to the arrangement problem appeared without proof in [57, Corollary 7] in a context where the regenerative ordering of frequencies was motivated
by an application to a tree growth process. A complete proof was given in [23], with the following alternative construction of the \( \tau \)-biased permutation being the main ingredient.

**Lemma 4.10.** For \( \xi = (1 - \tau)/\tau \) there is the compositional formula
\[
\text{perm}_{\tau}(x_1, \ldots, x_k) \overset{d}{=} \xi(\text{perm}_0(x_1, \ldots, x_k)),
\]
where on the right-hand side \( \xi \) and \( \text{perm}_0 \) are independent.

Summarising, for partitions of \( n \) the RHS of (4.7) coincides in distribution with \( \overset{\tau}{\theta}/\alpha(\text{perm}_0(\pi_n)) \). In the same way, the \((\alpha, \theta)\) \( m \)-regenerative set satisfies \( R \overset{d}{=} \overset{\theta}{\alpha}(\text{perm}_0(S)), \) where \( S \overset{d}{=} \text{PD}(\alpha, \theta) \) and \( \text{perm}_0(S) \overset{d}{=} \text{GEM}(\alpha, \theta) \).

**4.6. More on \((\alpha, \alpha)\) compositions: reversibility**

We have seen that the \((\alpha, \alpha)\) composition structures are the only regenerative compositions which have parts in the exchangeable order. We show now that these structures can be characterised by some weaker properties of reversibility.

Every composition structure \( \kappa \) has a dual \( \hat{\kappa} \), where each \( \hat{\kappa}_n \) is the sequence of parts of \( \kappa_n \) read in the right-to-left order. For example, the value \((3, 2)\) of \( \kappa_5 \) corresponds to the value \((2, 3)\) of \( \hat{\kappa}_5 \). If \( \kappa \) is derived from \( R \), then \( \hat{\kappa} \) is derived from the reflected paintbox \( 1 - R \). If both \( \kappa \) and \( \hat{\kappa} \) are regenerative then by the uniqueness (Theorem 4.1) they must have the same distribution. If \( \kappa \) is reversible, i.e. \( \kappa \overset{d}{=} \hat{\kappa} \), then the first part of \( \kappa_n \) must have the same distribution as its last part.

**Theorem 4.11.** [30] Let \( \kappa \) be a regenerative composition structure. Let \( F_n \) denote the first and \( L_n \) the last part of \( \kappa_n \). The following conditions are equivalent:

(i) \( \mathbb{P}(F_n = 1) = \mathbb{P}(L_n = 1) \) for all \( n \);
(ii) \( F_n \overset{d}{=} L_n \) for all \( n \);
(iii) \( \kappa_n \overset{d}{=} \hat{\kappa}_n \) for all \( n \) (reversibility),
(iv) \( \kappa \) is an \((\alpha, \alpha)\)-composition structure with some \( 0 \leq \alpha \leq 1 \).

**Sketch of proof** Some manipulations with finite differences yield
\[
\mathbb{P}(F_n = 1) = q(n : 1) = \frac{\Phi(n) - \Phi(n - 1)}{\Phi(n)/n},
\]
\[
\mathbb{P}(L_n = 1) = n \left[ 1 - \sum_{k=2}^{n} \binom{n-1}{k-1} (-1)^k \frac{\Phi(k)}{\Phi(n)} \right].
\]
Equating these probabilities, one arrives at \( \Phi(n) = (1 + \alpha)_{n-1}/(n - 1)! \) where \( \Phi(2) := 1 + \alpha \) and the normalisation \( \Phi(1) = 1 \) is assumed. The latter is the Laplace exponent corresponding to the \((\alpha, \alpha)\) composition. \( \square \)

Invoking the paintbox correspondence, the result implies
Corollary 4.12. For a random closed subset $R$ of $[0,1]$, the following two conditions are equivalent:

(i) $R$ is m-regenerative and $R \overset{d}{=} 1 - R$.
(ii) $R$ is distributed like the zero set of a Bessel bridge of dimension $2 - 2\alpha$, for some $0 \leq \alpha \leq 1$.

The degenerate boundary cases with $\alpha = 0$ or 1 are defined by continuity.

The reversibility property stated in Theorem 4.11 (iii) was used in [40] to characterise a class of random continuous fragmentation trees.

5. Self-similarity and stationarity

A random closed set $Z \subset [0,\infty)$ is called self-similar if it is ‘weakly homothetical about zero’, that is satisfies $cZ \overset{d}{=} Z$, $c > 0$. The property is a multiplicative analogue of the stationarity property (translation invariance) of a random subset of $\mathbb{R}$ (see [51] for a general account). We encountered self-similarity in connection with paintboxes for $(\alpha,0)$ compositions.

Regenerative $(0,\theta)$-compositions can be also embedded in the self-similar framework by passing to duals, which places the center of self-similarity at 0. The mirrored paintbox for the dual Ewens’ composition structure is the stick-breaking set $R = \{V_1 \cdot \cdots \cdot V_i, i = 0,1,\ldots\}$ with i.i.d. $V_i \overset{d}{=} \text{beta}(\theta,1)$. This set is the restriction to $[0,1]$ of a self-similar Poisson point process with variable density $\theta/y$, $y > 0$.

Introduce the operation of right reduction as cutting the last symbol of the binary code of composition. For instance, the right reduction maps 100110 to 10011.

Definition 5.1. A sequence of random compositions $\kappa = (\kappa_n)$ is called right-consistent if the right reduction maps $\kappa_{n+1}$ in a stochastic copy of $\kappa_n$. If $\kappa$ is a composition structure, we call it self-similar if it is right-consistent.

If a sequence of compositions $\kappa = (\kappa_n)$ is right-consistent, it can be realised on the same probability space as a single infinite random binary string $\eta_1,\eta_2,\ldots$, with $\kappa_n$ being the composition encoded in the first $n$ digits $\eta_1,\ldots,\eta_n$. For right-consistent $\kappa$ the Green matrix is of the form

$$g(n,j) = g(j) := \mathbb{P}(\eta_j = 1), \quad 1 \leq j \leq n, \quad n = 1,2,\ldots$$

Theorem 5.2. [32] A composition structure $\kappa$ is self-similar iff the paintbox $R$ is the restriction to $[0,1]$ of a selfsimilar set $Z$. In this case $\kappa$ can be encoded in an infinite binary string.

Sketch of proof The ‘if’ part is easily shown using the modified sampling scheme, as in the BM example. The ‘only if’ part exploits convergence of random sets as in Theorem 2.2.

Arbitrary infinite binary string $\eta_1,\eta_2,\ldots$ (starting from 1) do not correspond to a composition structure, unless the sampling consistency is satisfied. Let us
review the \((0, \theta)\) and \((\alpha, 0)\) compositions from the standpoint of binary representation.

**Example.** For \(\theta > 0\) let \(\eta_1, \eta_2, \ldots\) be a Bernoulli string with independent digits and
\[
g(j) = \Pr(\eta_j = 1) = \frac{\theta}{j + \theta - 1}.
\]
This encodes the dual Ewens’ composition structure, with the last-part deletion property. In the modified sampling scheme, the role of balls is taken by a homogeneous Poisson point process, and the boxes are created by points of an independent self-similar Poisson process \(Z_\theta\) of intensity \(\theta/x\).

The family of composition structures can be included in a Markov process with \(\theta \geq 0\) considered as a continuous time parameter \([34]\). On the level of paintboxes the dynamics amounts to intensifying Poisson processes, so that within time \(d\theta\) the Poisson process \(Z_\theta\) is superposed with another independent Poisson process with density \(d\theta/x\). This is an instance of sliced splitting, so (3.19) is in force. From this viewpoint, a special feature is that the \(\theta\)-splittings are consistently defined, also in terms of the interrupted subordinators, which are here compound Poisson processes with exponential jumps.

Remarkably, the splitting process remains Markovian in terms of the binary codes, and has the dynamics in which every ‘0’ eventually turns in ‘1’ by the rule: at time \(\theta\), a ‘0’ in the generic position \(j\) of the code is switching at rate \(1/(\theta + j - 1)\) to a ‘1’, independently of digits in all other positions.

**Example.** For \(\alpha \in (0, 1)\) let \((T_k)\) be a lattice renewal process with \(T_0 = 1\) and independent increments with distribution
\[
\Pr(T_{k+1} - T_k = m) = (-1)^{m-1}\binom{\alpha}{m}
\]
(for \(\alpha = 1/2\) this is the distribution for the recurrence time of the standard random walk). For \(\eta_j = 1(\cap_{k \geq 0}\{T_k = j\})\) the sequence \(\eta_1, \eta_2, \ldots\) encodes the regenerative \((\alpha, 0)\) composition structure. The Green matrix is \(g(j) = (\alpha)_{j-1}/(j-1)\).

It is known \([61]\) that no other Bernoulli or renewal random 0–1 strings are sampling consistent, i.e. produce composition structures. We shall turn to a larger class of strings with a Markov property, but first let us review a few general features of the self-similar compositions.

Let \(P_n\) be the size-biased pick from \(\kappa_n\), and let \(L_n\) be the last part of the composition. Similarly, let \(\tilde{P}\) be the size-biased gap-length of \(R\), and \(L\) be the size of the meander gap adjacent to 1.

**Theorem 5.3.** \([32]\) Let \(\kappa\) be a self-similar composition structure, thus derived from some self-similar set \(Z\). Then

(i) \(\tilde{P} \overset{d}{=} L\), and (ii) \(\tilde{P}_n \overset{d}{=} L_n\),

and the Green matrix is \(g(j) = (1 - \tilde{P})^{j-1}\).
Sketch of proof Since reducing the last box by one ball has the same effect as reducing the box chosen by the size-biased pick, the sizes of the boxes must have the same distribution. This yields (ii), and (i) follows as $n \to \infty$. Alternatively, inspecting the gap covering $U_{n:n}$ it is seen that $E[L^{n-1}] = \rho^2(n)$, the probability of one-block composition, so the moments of $\tilde{P}$ and $\tilde{L}$ coincide. Similarly, $\eta_j = 1$ in the event $U_{n:1} > \max(Z \cap [0, U_{n:j}])$.

The identity (ii) together with a generalisation of a result by Pitman and Yor [58] yields a characterisation of certain structural distributions, and shows that $\tilde{P}$ has a decreasing density on $(0, 1].$

**Theorem 5.4.** [32] The structural distribution for self-similar composition structure is of the form

\[
\text{P}(\tilde{P} \in dx) = \frac{\nu[x, 1]}{(d + m)(1 - x)} dx + \frac{d}{d + m} \delta_0(dx), \quad x \in [0, 1],
\]

where $d \geq 0$ and $\nu$ is a measure on $(0, 1]$ with

\[
m := \int_0^1 |\log(1 - x)|\nu(dx) < \infty.
\]

There is no atom at 0 iff $d = 0$, that is iff $\mathcal{Z}$ has Lebesgue measure zero.

**5.1. Markovian composition structures**

For a time being we switch to regeneration in the right-to-left order of parts, starting from the last part, like for the dual Ewens' composition. It is more convenient in the self-similar context to have the homotheties centered at 0. We first modify the deletion property of compositions by allowing a special distribution for the first deleted part, which is now the last, i.e. the rightmost part of composition.

**Definition 5.5.** A composition structure is called *Markovian* if the CPF is of the product form

\[
p^\circ(\lambda^\circ) = q^{(0)}(n : \lambda_k) \prod_{j=1}^{k-1} q(\Lambda_j : \lambda_j), \quad \Lambda_j = \lambda_1 + \cdots + \lambda_j.
\]

where $q^{(0)}$ and $q$ are two decrement matrices.

Similarly to (3.1), formula (5.2) says that 1’s in the binary code of $\zeta$ appear at sites $Q^*_k(t) + 1$ visited by a decreasing Markov chain, with the new feature that the distribution of the first decrement is determined by $q^{(0)}$, and not by $q$.

The counterpart of Theorem 3.4 for (5.2) is straightforward. For $(S_t)$ a subordinator, consider the process $(V \cdot \exp(-S_t), \ t \geq 0)$, where $V$ takes values in $(0, 1)$ and is independent of $(S_t)$. The range of this process is a $m$-regenerative set (now with right-to-left regeneration) scaled by the random factor $V$. Taking this
A.V. Gnedin/Regeneration in random combinatorial structures

set for paintbox $R$, thus with the meander gap $[V, 1]$, a Markovian composition structure is induced with $q(n : m) = \Phi(n : m)/\Phi(n)$ as in (3.15), and

$$
q^{(0)}(n : m) = \Phi^{(0)}(n : 0)q(n : m) + \Phi^{(0)}(n : m),
$$

$$
\Phi^{(0)}(n : m) = \left( \begin{array}{c} n \\ m \end{array} \right) E \{ V^{n-m}(1-V)^m \}.
$$

For every Markovian composition structure the paintbox is of the described form.

### 5.2. Self-similar Markov composition structures

Let $Q^\uparrow = (Q^\uparrow(t), t = 0, 1, \ldots)$ be a time-homogeneous increasing Markov chain on $\mathbb{N}$ with $Q^\uparrow(0) = 1$. An infinite string $\eta_1, \eta_2, \ldots$ is defined as the sequence of sites visited by $Q^\uparrow$

$$
\eta_j = 1(Q^\uparrow(t) = j \text{ for some } t).
$$

If the string determines some composition structure $\asymp$, then $\asymp$ is self-similar.

A composition structure is called self-similar Markov if it has such a binary representation generated by an increasing Markov chain.

A stationary regenerative set (or stationary Markov [51]) is the range of a process $(X + S_t, t \geq 0)$ where $(S_t)$ is a finite mean-subordinator, with Lévy measure satisfying

$$
m = \int_0^{\infty} y\tilde{\nu}(dy) < \infty,
$$

drift $d \geq 0$ and the initial value $X$ whose distribution is

$$
P(X \in dy) = \frac{\tilde{\nu}[y, \infty]}{d + m} dy + \frac{d}{d + m} \delta_0(dy).
$$

Unlike $\nu$ in (5.1), the measure $\tilde{\nu}$ lives on $(0, \infty)$, as in (3.13).

**Theorem 5.6.** [32] A composition structure $\asymp$ is self-similar Markov if and only if $R = \exp(-\tilde{R})$, where $\tilde{R}$ is a stationary regenerative set.

The distribution of size-biased pick is then (5.1) with $\nu$ the image of $\tilde{\nu}$ under $y \to 1 - e^{-y}$. The Green matrix can be written in terms of the Laplace exponent

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

The relation between this and (3.20) is that the RHS of (3.20) converges to $g(j)$ as $n \to \infty$. This fact is analogous to the elementary renewal theorem.

Like in the regenerative case, the decrement matrices are determined, in principle, by the probabilities $(p(n), n \geq 0)$, which are moments of the structural distribution, whence the analogue of Theorem 4.1:

**Theorem 5.7.** If a partition structure admits arrangement as a self-similar Markov composition structure, then such arrangement is unique in distribution.
Application to the two-parameter family

For $0 \leq \alpha < 1$ and $\theta > 0$ let $\mathcal{R}_{\alpha,\theta}$ be the $m$-regenerative set associated with $(\alpha, \theta)$ regenerative composition structure, and let $V$ be an independent variable whose distribution is beta$(\theta + \alpha, 1 - \alpha)$. Then the scaled reflected set $V \cdot (1 - \mathcal{R}_{\alpha,\theta})$ is associated with a self-similar Markov composition structure corresponding to $(\alpha, \theta - \alpha)$ partition structure. This follows from the stick-breaking representation of the frequencies in size-biased order, with independent factors beta$(\theta + j\alpha, 1 - \alpha)$, $j = 1, 2, \ldots$. The Green function $g$ and transition probabilities for $Q^T$ can be readily determined.

A ‘stationary’ version of the regenerative $(\alpha, \theta)$ composition is the self-similar Markov arrangement of the $(\alpha, \theta - \alpha)$ partition. The structural distribution is beta$(1 - \alpha, \theta + \alpha)$, which is also the law of the meander size $1 - V$. Note that $\theta - \alpha$ may assume negative values, hence every partition with $\theta > \alpha$ has a self-similar Markov arrangement. This ‘rehabilitates’ $(\alpha, \theta)$ partitions with $-\alpha < \theta < 0$ that lack regeneration literally, rather the property appears in a modified form, as stationary regeneration. If $\theta \geq 0$ then both types of regeneration are valid.

The $(\alpha, 0)$ composition with the left-to-right regeneration is also self-similar Markov, i.e. has the ‘stationary’ right-to-left regeneration property. This combination of regeneration properties is characteristic for this class.

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

The arrangement of $(\alpha, \theta)$ partition in a self-similar Markov composition structure is the same on both paintbox and finite $-n$ level. The size-biased pick is placed at the end, then the rest parts are arranged to the left of it as for the dual $(\alpha, \theta + \alpha)$ regenerative structure, see Section 3.4. Property (i) in Theorem 5.3 holds in the strong sense: conditionally given the unordered frequencies $S$, the length of the meander is a size-biased pick (see [58]).

6. Asymptotics of the block counts

Every partition structure is in essence an infinite occupancy model, in which frequencies of boxes $S$ are random and, given $S = (s_j) \in \mathcal{S}$, the balls are dropped independently, with probability $s_j$ to hit box $j$. Questions of traditional interest concern the number of boxes occupied by the first $n$ balls, the number of boxes occupied by $r$ out of $n$ balls, the most occupied box etc. Starting from 60’s such questions were intensively studied for the models with nonrandom frequencies (also depending on $n$), but there are relatively few results for the models with random frequencies. For regenerative partition/composition structures the study is facilitated by the Markovian character of partitions for each $n$ and connections with the underlying regenerative paintbox.

For $\kappa = (\kappa_n)$ a regenerative composition structure, let $K_n$ be the number of parts in $\kappa_n$ and let $K_{n,r}$ be the number of parts equal $r$, so that $\sum_r rK_{n,r} = n$, $\sum_r K_r = K_n$. For instance, in the event $\kappa_{10} = (2, 4, 2, 1, 1)$ we have $K_{10} = 5$, $K_{10,1} = 2$, $K_{10,2} = 2$, $K_{10,3} = 0$ etc. The full vector $(K_{n,1}, \ldots, K_{n,n})$ is one of the ways to record the partition associated with $\kappa_n$. In the species sampling
context, $K_n$ is the number of distinct species represented in a sample, hence it is often considered as a measure of diversity.

We are interested in the large-$n$ asymptotics of $K_n$ and $K_{n,r}$ for $r = 1, 2, \ldots$. This can be called the small-blocks problem, as opposed to study of the largest part of $\kappa_n$ or a few largest parts. Typically the composition will have a relatively small number of large parts of some size of order $n$ and many parts of size $r \ll n$, the latter making the principal contribution to $K_n$.

Unless indicated otherwise, we assume that $d = 0$ (proper case, no drift) and that $\nu(\infty) = 0$ (no killing, no right meander). Then the order of growth of $K_n$ is sublinear, $K_n \ll n$, and $K_n \uparrow \infty$ almost surely. Moreover, by a very general law of large numbers $K_n \sim \mathbb{E}(K_n \mid S)$ a.s. [22, 46]. The latter suggests that a major part of the variability of $K_n$ just comes from the conditional expectation, to the extent that the asymptotic behaviour of both random variables is the same. It seems, however, there are no general results which either give conditions under which the limiting distributions (if any) of the random variables $K_n$ and $\mathbb{E}(K_n \mid S)$ coincide, or results which in some way distinguish the randomness in frequencies from the randomness in sampling the balls. For fixed frequencies the limiting law of $K_n$ and each $K_{n,r}$ is normal subject to the sole condition that the variance goes to $\infty$ (see [4] for the most recent account), but the randomness in ‘environment’ $S$ leads to a much wider spectrum of possibilities.

One general tool is the structural distribution $\sigma(dx) = \mathbb{P}(P \in dx)$ of the size-biased pick $\tilde{P}$, which can be used to compute the expectations via

\[
\mathbb{E}[K_n] = \int_{0}^{1} \frac{1 - (1 - x)^n}{x} \sigma(dx), \quad \mathbb{E}[K_{n,r}] = \binom{n-1}{r-1} \int_{0}^{1} x^{r-1} (1-x)^{n-r} \sigma(dx).
\]

It is clear from these formulas that the asymptotics of the moments are determined by the behaviour of $\sigma$ near 0, because $(1 - x)^n$ decays exponentially fast on any interval $[\epsilon, 1]$.

The block counts $K_n, K_{n,r}$ depend only on the partition, and not on the order of the parts. Nevertheless, the Markovian character of regenerative compositions and the connection with subordinators can be efficiently exploited to study these functionals by methods of the renewal theory. This may be compared with other classes of partitions studied with the help of local limit theorems: partitions obtained by conditioning random sums of independent integer variables [3], and Poisson-Kingman partitions derived from conditioned subordinators [55].

For Ewens’ partitions it is well known that $K_n$ is asymptotically normal, with both mean and variance of the order of $\theta \log n$ (see [3, 56]); the $\theta = 1$ case is also familiar in connection with records and permutations. In contrast to that, for $(\alpha, \theta)$ partitions with $\alpha > 0$ the right scale for $K_n$ is $n^\alpha$ ($\alpha$-diversity [56]). These known facts will be embedded in a much more general consideration.

The number of parts satisfies a distributional fixed-point equation

\[
K_n \overset{d}{=} 1 + K'_{n-F_n}
\]

where $K'_m, m \leq n - 1$, are independent of the first part $F_n$ with distribution $q(n : \cdot)$, and satisfy $K'_m \overset{d}{=} K_m$. Known asymptotics (e.g. [13, 52]) derived from
such identities do not cover the full range of possibilities and require very restrictive moment conditions which are not easy to provide (see however [36] for one application of this approach). In what follows we report on the asymptotics which were obtained by different methods, based on the connection with subordinators, poissonisation, methods of the renewal theory, and Mellin transform [5, 25, 35, 36]. See [25, 37, 39] for related work on the absorption times of decreasing Markov chains on integers.

We assume as before the paintbox construction with balls $U_1, \ldots, U_n$ and $R$ the closed range of a multiplicative subordinator $(1 - \exp(-S_t), t \geq 0)$. In these terms, $K_{n,r}$ is the number of gaps in the range hit by exactly $r$ out of $n$ uniform points, and $K_n$ is the total number of nonempty gaps.

**Remark** If the subordinator has positive drift $d > 0$, then $K_n \sim K_{n,1} \sim n \operatorname{meas}(R)$ a.s., so singletons make a leading contribution to $K_n$. The Lebesgue measure of $R$ is a random variable proportional to the exponential functional of the subordinator,

$$\operatorname{meas}(R) = d \int_0^\infty \exp(-S_t)dt.$$  

It is informative to consider the number of parts $K_n$ as the terminal value of the increasing process $K_n := (K_n(t), t \geq 0)$, where $K_n(t)$ is the number of parts of the subcomposition derived from the configuration of uniform points not exceeding $1 - \exp(-S_t)$, i.e. produced by the subordinator within the time $[0, t]$. The number of $r$-parts $K_{n,r}$ is the terminal value of another process $K_{n,r} := (K_{n,r}(t), t \geq 0)$ which counts $r$-parts, but this process is not monotone as $n$ increases (although $\sum_{r' \geq r} K_{n,r'}(t)$ is).

We can think of the subordinator representation of a regenerative composition structure as a coagulation process in which, if at time $t$ there are $n'$ particles, every $m$-tuple of them is merging to form a single particle at rate $\Phi(n':m)$. The particle emerging from the coalescence is immediately frozen. Starting with $n$ particles, $K_n(t)$ counts the number of frozen particles at time $t$. The dynamics is analogous to that of Pitman-Sagitov $\Lambda$-coalescents, with the difference that in the $\Lambda$-coalescents the mergers remain active and keep coagulating with other existing particles [53]. For a class of $\Lambda$-coalescents a coupling with compositions was used to explore asymptotics of the coalescent processes [25].

The asymptotics in the small-block problem largely depend on the behaviour of the right tail of the Lévy measure near 0. If $\tilde{\nu}$ is finite, then simply $\tilde{\nu}[y, \infty] \to \tilde{\nu}[0, \infty]$ as $y \to 0$, but if $\tilde{\nu}$ is infinite it seems difficult if at all possible to make any conclusions without the following assumption.

**Assumption of regular variation** We shall suppose that $\tilde{\nu}$ satisfies the condition of regular variation

$$\tilde{\nu}[y, \infty] \sim \ell(1/y)y^{-\alpha} \quad y \downarrow 0,$$

where the index satisfies $0 \leq \alpha \leq 1$ and $\ell$ is a function of slow variation at $\infty$, i.e. $\ell$ satisfies $\ell(t/y)/\ell(1/y) \to 1$ as $y \to 0$ for all $t > 0$.  


Note that the assumption is satisfied in the case of finite \( \nu \). By the monotone density version of Karamata’s Tauberian theorem \[11\], for \( 0 \leq \alpha < 1 \) the condition (6.1) is equivalent to the asymptotics of the Laplace exponent

\[ \Phi(\rho) \sim \Gamma(1 - \alpha) \rho^\alpha \ell(\rho), \quad \rho \to \infty. \]

Qualitatively different asymptotics are possible. Very roughly, the whole spectrum can be divided in the following cases, each requiring separate analysis.

- The finite Lévy measure case. This is the case of stick-breaking compositions, with \((S_t)\) a nondecreasing compound Poisson process. Further subdivision is determined by the moments of the distribution of jumps.
- The slow variation case with \( \alpha = 0 \) and \( \ell(y) \to \infty \) as \( y \to \infty \). Typical example: regenerative compositions associated with gamma subordinators. A variety of limit behaviours appears for distinct \( \ell \).
- The proper regular variation case with \( 0 < \alpha \leq 1 \). Typical example: composition associated with \( \alpha \)-stable subordinator (with \( 0 < \alpha < 1 \)). The case \( \alpha = 1 \) is somewhat special.

One principal difference between the cases of (proper) regular and slow variation is in the time scales at which major growth and variability of \( K_n \) occur. In the case \( \alpha > 0 \) all \( K_n(t), K_{n,r}(t) \) are of the same order as \( K_n \), whereas in the case \( \alpha = 0 \) we have \( K_n(t) \ll K_n \) and \( K_{n,r} \ll K_n \).

### 6.1. Stick-breaking compositions

In the case of finite Lévy measure we scale \( \nu \) to a probability measure. Then \( \nu \) is the distribution of \( -\log(1 - W) \), where \( W \) is the generic stick-breaking factor. Introduce the moments

\[ m := \mathbb{E}[-\log(1 - W)], \quad \sigma^2 := \text{Var}[\log(1 - W)], \quad m_1 := \mathbb{E}[-\log W], \quad (6.2) \]

which may be finite or infinite.

Let \( M_n \) be the index of the rightmost occupied gap, which contains the maximum order statistic \( U_{n:n} \). Roughly speaking, stick-breaking implies a fast exponential decay of the sizes of gaps, hence one can anticipate a cutoff phenomenon: empty gaps can occur only in a range close to \( U_{n:n} \). From the extreme-value theory we know that \( -\log(1 - M_n) - \log n \) has a limit distribution of the Gumbel type, thus \( M_n \) can be approximated by the number of jumps of \((S_t)\) before crossing level \( \log n \).

It should be noted that exponential decay of nonrandom frequencies, like for the geometric distribution, implies oscillatory asymptotics in the occupancy problem \[6, 12\]. By stick-breaking the oscillations do not appear since the main variability comes due to randomness in frequencies themselves, so the variability coming from sampling is dominated.

Consider a renewal process with distribution for spacings like that of \( -\log(1 - W) \). If the moments are finite, \( m < \infty, \sigma^2 < \infty \), then a standard result from the
renewal theory implies that the number of renewals on $[0, \log n]$ is approximately normal for large $n$, with the expected value asymptotic to $(\log n)/m$. The same is valid for $M_n$, and under the additional assumption $m_1 < \infty$ also for $K_n$ (see [19]). Under weaker assumptions on the moments, the possible asymptotics correspond to other limit theorems of renewal theory.

**Theorem 6.1.** [26] Suppose the distribution of $-\log(1-W)$ is nonlattice with $m_1 < \infty$. The following assertions are equivalent.

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

(ii) The distribution $\tilde{\nu}$ of $-\log(1-W)$ either belongs to the domain of attraction of a stable law, or the function $\tilde{\nu}[x, \infty]$ slowly varies as $x \to \infty$.

Furthermore, this limiting distribution of $(K_n - b_n)/a_n$ is as follows.

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

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

$$\int_{x}^{1} (\log y)^2 \mathbb{P}(1-W \in dy) \sim \ell(-\log x) \quad \text{as} \quad x \to 0,$$

for some $\ell$ slowly varying at $\infty$, then for $b_n = m^{-1} \log n$, $a_n = m^{-3/2} c_{\lfloor \log n \rfloor}$ and $c_n$ any sequence satisfying $\lim_{n \to \infty} n\ell(c_n)/c_n^2 = 1$, the limiting distribution is standard normal.

(c) Assume that the relation

$$\mathbb{P}(1-W \leq x) \sim (-\log x)^{-\gamma} \ell(-\log x) \quad \text{as} \quad x \to 0, \quad (6.3)$$

holds with $\ell$ slowly varying at $\infty$ and $\gamma \in [1, 2)$, and assume that $m < \infty$ if $\gamma = 1$, then for $b_n = m^{-1} \log n$, $a_n = m^{-(\gamma+1)/2} c_{\lfloor \log n \rfloor}$ and $c_n$ any sequence satisfying $\lim_{n \to \infty} n\ell(c_n)/c_n^2 = 1$, the limiting distribution is $\gamma$-stable with characteristic function

$$\tau \mapsto \exp\{-|\tau|^\gamma \Gamma(1-\gamma)(\cos(\pi \gamma/2) + i \sin(\pi \gamma/2) \text{sgn}(\tau))\}, \quad \tau \in \mathbb{R}. \quad (6.4)$$

(d) Assume that $m = \infty$ and the relation (6.3) holds with $\gamma = 1$. Let $c$ be any positive function satisfying $\lim_{x \to \infty} x\ell(c(x))/c(x) = 1$ and set

$$\psi(x) := x \int_{\exp(-c(x))}^{1} \mathbb{P}(1-W \leq y)/y \, dy.$$

Let $b$ be any positive function satisfying $b(\psi(x)) \sim \psi(b(x)) \sim x$ (asymptotic inverse to $\psi$). Then, with $b_n = b(\log n)$ and $a_n = b(\log n)c(b(\log n))/\log n$, the limiting distribution is $1$-stable with characteristic function

$$\tau \mapsto \exp\{-|\tau| \gamma/2 - i \log |\tau| \text{sgn}(\tau)\}, \quad \tau \in \mathbb{R}. \quad (6.4)$$
(c) If the relation (6.3) holds with $\gamma \in [0,1)$ then, for $b_n = 0$ and $a_n := \log^+ n/\ell(\log n)$, the limiting distribution is the scaled Mittag-Leffler law $\theta_\gamma$ (exponential, if $\gamma = 0$) characterised by the moments
\[
\int_0^\infty x^n \theta_\gamma(dx) = \frac{n!}{\Gamma^n(1 - \gamma)\Gamma(1 + n\gamma)}, \quad n \in \mathbb{N}.
\]

**Sketch of proof** The results are first derived for $M_n$ by adopting asymptotics from the renewal theory. To pass to $K_n$ it is shown, under the condition $m_1 < \infty$, that the variable $M_n - K_n$ (the number of empty boxes to the left of $U_{n,n}$) converges in distribution and in the mean to a random variable with expected value $m_1/m$. \qed

**Example** Suppose $W$ has a beta density (3.7). The moments are easily computable as $m = \Psi(\theta + \gamma) - \Psi(\theta), \ m_1 = \Psi(\theta + \gamma) - \Psi(\gamma), \ \sigma^2 = \Psi'(\theta) - \Psi'(\theta + \gamma)$ (with $\Psi = \Gamma'/\Gamma$ denoting the logarithmic derivative of the gamma function). We are therefore in the case (a) of Theorem 6.1, hence $K_n$ is asymptotically normal with $E[K_n] \sim m_1/\gamma n$ and $\text{Var}[K_n]$ of the same order.

The instance $\gamma = 1$ recovers well-known asymptotics of Ewens’ partitions, which have $K_n \sim \theta \log n$ a.s. In this case the limit law of the number of empty gaps $M_n - K_n$ has probability generating function $z \mapsto \Gamma(\theta + 1)\Gamma(\theta + 1 - z\theta)/\Gamma(1 + 2\theta - z\theta)$ (which identifies a mixture of Poisson distributions, see [26]).

**Example** Suppose the law of $W$ is given by $P(1 - W \leq x) = (1 - \log x)^{-1}, \ x \in (0,1)$. It can be checked that $m_1 < \infty$, hence the case (c) applies and
\[
\frac{(\log \log n)^2}{\log n}K_n - \log \log n - \log \log \log n
\]
converges to a 1-stable law with characteristic function (6.4).

The number of empty boxes $M_n - K_n$ converges in probability to 0.

Under assumptions $m < \infty, \ m_1 < \infty$ the limit behaviour of $K_{n,r}$’s is read from a limiting occupancy model [27]. To describe the limit we pass to the dual composition, generated by right-to-left stick-breaking $\mathcal{R} = \{V_1 \cdots V_i : i \geq 1\}$ with independent $1 - V_i \overset{d}{=} W$. Let $(X_{n,1}, X_{n,2}, \ldots)$ be the occupancy numbers of the gaps in the left-to-right order, this is a random weak composition (0’s allowed) of $n$ with $X_{n,1} > 0$ and $X_{n,j} \geq 0$ for $j > 1$. By inflating $[0,1]$ with factor $n$, the uniform sample converges as a point process to a unit Poisson process (balls). On the other hand, $n\mathcal{R}$ converges to a self-similar point process $\mathcal{Z}$, whose gaps play the role of boxes. From this, the occupancy vector $(X_{n,1}, X_{n,2}, \ldots)$ acquires a limit, which is an occupancy vector $(X_1, X_2, \ldots)$ derived from the limiting point processes. The limit distribution of the occupancy vector is
\[
P(X_1 = \lambda_1, \ldots, X_k = \lambda_k) = \frac{1}{m(\lambda_1 + \cdots + \lambda_k)} \prod_{i=1}^k \tilde{q}(\lambda_i : \lambda_i)
\]
where $\lambda_1 > 0, \lambda_i \geq 0, \Lambda_i = \lambda_1 + \cdots + \lambda_i$ and $\tilde{q}(n : m) = \binom{n}{m} E[W^m(1 - W)^{n-m}]$. Correspondingly, $K_{n,r}$’s jointly converge in distribution to $\#\{i : X_i = r\}, r = \ldots$.
The convergence also holds for $K_{n,0}$, which we define as the number of empty gaps preceding the last occupied gap.

If $W \overset{d}{=} \text{beta}(1, \theta)$ then $Z$ is Poisson process with density $\theta/x$. Then $K_{n,r}$’s converge in distribution to independent Poisson variables with mean $\theta/r$, which is a well known property of Ewens’ partitions \cite{3}. It is a challenging open problem to identify the limit laws of the $K_{n,r}$’s for general distribution of $W$.

Assumption $m_1 < \infty$ is a way to keep the number of empty boxes bounded, in view of the relation $\lim_{n \to \infty} \mathbb{E}K_{n,0} = m_1/m$. The reduction to the asymptotics of $M_n$ is then possible since $M_n = K_n + K_{n,0}$. The case $m_1 = \infty$ is more delicate.

The technique of perturbed random walks to cover the case $m_1 = \infty$ was applied in \cite{24}, where it was shown that the limit laws for $K_n$ are still as in Theorem 6.1, but $K_n$ requires more sophisticated centering accounting for the ‘empty boxes’. Next example illustrates the situation.

**Example** For $\gamma \in (0, 1/2)$ consider $W$ with distribution

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

Then $\sigma^2 < \infty$ and

$$\mathbb{P}(|\log W| > y) \sim y^{-\gamma} \quad \text{as } y \to \infty.$$ We have $m_1 = \infty$ and, in fact,

$$\mathbb{E}K_{n,0} \sim \frac{1}{m} \sum_{k=1}^{n} \frac{\mathbb{E}(1 - W)^k}{k} \sim b_n - m^{-1}\log n \sim \frac{1}{m(1 - \gamma)}\log^{1-\gamma} n.$$ The distribution of $(K_n - a_n)/b_n$ is asymptotically normal with $a_n$ as in Theorem 6.1 but

$$b_n = m^{-1}(\log n - (1 - \gamma)^{-1}\log^{1-\gamma} n + o(\log^{1-\gamma} n)),$$ where the additional term of centering $b_n - m^{-1}\log n$ is not killed by scaling, rather appears to compensate for the influence of $K_{n,0}$.

### 6.2. Regular variation: $0 < \alpha \leq 1$

Suppose (6.1) holds with $0 < \alpha \leq 1$. The small-blocks asymptotics can be treated by reducing the occupancy problem to counting the gaps of given sizes, and applying strong laws available for nonrandom regularly varying frequencies (see [39] for another approach). For $x > 0$ let $N_x(t)$ be the number of gaps of size at least $x$, in the partial range of the multiplicative subordinator $(1 - \exp(-S_u), 0 \leq u < t)$. Introduce the exponential functionals

$$I_\alpha(t) := \int_0^t \exp(-\alpha S_u) \, dt, \quad I_\alpha := I_\alpha(\infty).$$
The distribution of $I_\alpha(\infty)$ is determined by the formula for the moments \[\mathbb{E}(I_\alpha)^k = \frac{k!}{\prod_{j=1}^{\infty} \Phi(\alpha j)},\] where $\Phi$ is the Laplace exponent of the subordinator $(S_t)$.

**Theorem 6.2.** [36] Suppose the Lévy measure fulfills (6.1). Then for $0 < t \leq \infty$

for $0 < \alpha < 1$ \[\frac{N_x(t)}{\ell(1/x)x^{-\alpha}} \to I_\alpha(t) \text{ a.s., } x \downarrow 0,\]

for $\alpha = 1$ \[\frac{N_x(t)}{\ell_1(1/x)x^{-\alpha}} \to I_\alpha(t) \text{ a.s., } x \downarrow 0,\]

where

\[\ell_1(z) = \int_z^\infty u^{-1}\ell(u)\text{du}\]

is another function of slow variation, satisfying $\ell_1(z) \gg \ell(z)$ as $z \to \infty$.

**Sketch of proof** Let $\tilde{N}_x(t)$ be the number of gaps in the range of the (additive) subordinator restricted to $[0,t]$. By the Lévy-Ito construction of $(S_t)$ from a Poisson process, we have the strong law $\tilde{N}_x(t) \sim \tilde{\nu}[y, \infty]$ t a.s. for $y \downarrow 0$. A small gap $(s,s+x)$ is mapped by the function $s \to 1 - e^{-s}$ in a gap of size $e^{-s}x$, from which the result for finite $t$ follows by integration. Special tail estimates are required to conclude that similar asymptotics hold with integration extended to $[0,\infty]$.

The instance $\alpha = 1$ may be called in this context the case of rapid variation. In this case $\ell$ in (6.1) must decay at $\infty$ sufficiently fast, in order to satisfy $\Phi(1) < \infty$.

Conditioning on the frequencies $S = (s_j)$ embeds the small-block problem in the framework of the classical occupancy problem: $n$ balls are thrown in an infinite series of boxes, with positive probability $s_j$ of hitting box $j$. By a result of Karlin [46], the number of occupied boxes is asymptotic to the expected number, from which $K_n \sim \mathbb{E}[K_n | R]$ a.s., and a similar result holds for $K_{n,r}$ under the regular variation with index $\alpha > 0$. Combining this with Theorem 6.2, we have (see [36])

**Theorem 6.3.** Suppose the Lévy measure fulfills (6.1). Then, uniformly in $0 < t \leq \infty$, as $n \to \infty$, the convergence holds almost surely and in the mean:

\[\frac{K_n(t)}{\Gamma(1-\alpha)n^\alpha\ell(n)} \to I_\alpha(t), \quad \frac{K_{n,r}(t)}{\Gamma(1-\alpha)n^\alpha\ell(n)} \to (-1)^{r-1}(\alpha \atop r)I_\alpha(t),\]

for $0 < \alpha < 1$ and $r \geq 1$, or $\alpha = 1$ and $r > 1$. Similarly, $K_n(t)/(n\ell_1(n)) \to I_1(t)$ for $\alpha = 1$.

Thus $K_n,K_{n,r}$ have the same order of growth if $0 < \alpha < 1$. In the case $\alpha = 1$ of rapid variation, singletons dominate, $K_{n,1} \sim K_n$, while all other $K_{n,r}$’s with $r > 1$ are of the same order of growth which is smaller than that of $K_n$. 

\[\]
Example The subordinator associated with the two-parameter family of compositions has $\Phi$ given by (3.22), hence

$$E(I_\alpha)^k = \frac{(\alpha + \theta)(2\alpha + \theta) \cdots ((k-1)\alpha + \theta)\Gamma(\theta + 1)}{\Gamma(k\alpha + \theta)\{a\Gamma(1-\alpha)\}^k},$$

which for $\theta = 0$ and $0 < \alpha < 1$ identifies the law of $I_\alpha$ as a Mittag-Leffler distribution. Theorem 6.3 recovers in this instance known asymptotics [56] related to the local times of Bessel bridges.

One generalisation of Theorems 6.2 and 6.3 is obtained by taking for paintbox the range of a process $\phi(S_t)$, where $\phi : \mathbb{R}_+ \to [0,1]$ is a smooth monotone function, with not too bad behaviour at $\infty$. The generalised power laws hold also for this class of partitions, with the only difference that the exponential functionals should be replaced by integrals $\int_0^t (\phi'(S_u))^{\alpha}du$, see [36].

6.3. Slow variation: $\alpha = 0$

The case of infinite Lévy measure with slowly varying tail $\hat{\nu}[y, \infty] \sim \ell(1/y)$ ($y \downarrow 0$) is intermediate between finite $\hat{\nu}$ and the case of proper regular variation. In this case $K_{n,r} \to \infty$ (like in the case $\alpha > 0$) but $K_{n,r} \ll K_n$ (like in the case of finite $\hat{\nu}$). Following Barbour and Gnedin [5] we will exhibit a further wealth of possible modes of asymptotic behaviour appearing in this transitional regime.

We assume that the first two moments of the subordinator are finite. The assumption about the moments is analogous to the instance (a) of Theorem 6.1 in the case of finite $\hat{\nu}$. The results will be formulated for the case

$$E[S_t] = t, \quad \text{Var}[S_t] = s^2 t,$$

which can be always achieved by a linear time scaling. Indeed, a general subordinator $S_t$ with

$$m := E[S_1] = \int_0^\infty x\hat{\nu}(dx), \quad \nu^2 := \text{Var}[S_1] = \int_0^\infty x^2\hat{\nu}(dx)$$

should be replaced by $S_{t/m}$, then $s^2 = \nu^2/m$. Because the linear time change does not affect the range of the process, it does not change the distribution of $K_n, K_{n,r}$.

For the sample (balls) we take a Poisson point process on $[0,1]$ with intensity $n > 0$. This is the same as assuming a Poisson($n$) number of uniform points thrown on $[0,1]$. To avoid new notations, we further understand $n$ as the intensity parameter, and use the old notation $K_n(t)$ to denote the number of blocks of the (poissonised) subcomposition on the interval $[0,1 - \exp(-S_t)]$. The convention for $K_n$ is the same. For large sample-sizes the poissonised quantities are very close to their fixed-$n$ counterparts, but the Poisson framework is easier to work with due to independence of the occupancy counts for distinct boxes (conditionally of the frequencies).
The total number of blocks is the terminal value $K_n = K_n(\infty)$ of the increasing process $K_n(t)$. Poissonisation makes the subcompositions within $[0, 1 - \exp(-S_t)]$ and $[1 - \exp(-S_t), 1]$ conditionally independent given $S_t$, hence $K_n(t)$ and $K_n(\infty) - K_n(t)$ are also conditionally independent. The consideration can be restricted to the time range $t < \tau_n$ where $\tau_n := \inf\{t : S_t > \log n\}$ is the passage time through $\log n$, since after this time the number of blocks produced is bounded by a Poisson(1) variable.

Define the poissonised Laplace exponent

$$\Phi_0(n) := \int_0^\infty \{1 - \exp(-n(1 - e^{-y}))\} \tilde{\nu}(dy).$$

For large $n$, we have $\Phi_0(n) \sim \Phi(n)$, but the former is more convenient to deal with, since it enters naturally the compensator of $(K_n(t), t \geq 0)$,

$$A_n(t) := \int_0^t \Phi_0(n \exp(-S_u))du.$$

Introduce

$$\Phi_k(n) := \int_0^\infty \{\Phi_0(ne^{-s})\}^k ds = \int_0^n \{\Phi_0(s)\}^k \frac{ds}{s}, \quad k = 1, 2.$$

By the assumption of slow variation and from $\Phi_0(n) \to \infty$ it follows that $\Phi_k$’s are also slowly varying, and satisfy

$$\Phi_2(n) \gg \Phi_1(n) \gg \Phi_0(n), \quad n \to \infty.$$

These functions give, asymptotically, the moments of $K_n$ and of the terminal value of the compensator

$$\mathbb{E}[K_n] = \mathbb{E}[A_n(\infty)] \sim \Phi_1(n), \quad \text{Var}[K_n] \sim \text{Var}[A_n(\infty)] \sim s^2 \Phi_2(n), \quad n \to \infty.$$

**Remark** In the stick-breaking case $\tilde{\nu}[0, \infty] = 1$ the asymptotics of $\text{Var}[A_n(\infty)]$ and $\text{Var}[K_n]$ are different, because the asymptotic relation $\Phi_1(n) \ll \Phi_2(n)$ is not valid. Instead, we have $\text{Var}[A_n(\infty)] \sim v^2 \bar{m}^{-3} \log n$, and $\text{Var}[K_n] \sim \sigma^2 \bar{m}^3 \log n$ with $\sigma^2 = v^2 - \bar{m}^2$.

The following approximation lemma reduces the study of $K_n(t)$ to the asymptotics of the compensator.

**Lemma 6.4.** We have, as $n \to \infty$,

$$\mathbb{E}[(K_n - A_n(\infty))^2] \sim \Phi_1(n),$$

and for any $b_n$ such that $\Phi_1(n)/b_n^2 \to 0$

$$\lim_{n \to \infty} \mathbb{P} \left[ \sup_{0 \leq t \leq \infty} |K_n(t) - A_n(t)| > b_n \right] = 0.$$
Sketch of proof Noting that $K_n(t) - A_n(t)$ is a square integrable martingale with unit jumps, we derive $\mathbb{E}[K_n - A_n(\infty)]^2 = \mathbb{E}[A_n(\infty)]$, from which the first claim follows. The second follows by application of Kolmogorov’s inequality.

From this the law of large number is derived:

**Theorem 6.5.** As $n \to \infty$, we have $K_n \sim A_n(\infty) \sim \Phi_1(n)$ almost surely and in the mean.

For more delicate results we need to keep fluctuations of the compensator under control. For this purpose we adopt one further assumption, akin to de Haan’s second order regular variation [11]. As in Karamata’s representation of slowly varying functions [11], write $\Phi_0$ as

$$\Phi_0(s) = \Phi_0(1) \exp \left( \int_0^s \frac{dz}{zL(z)} \right),$$

where

$$L(n) := \frac{\Phi_0(n)}{n \Phi_0'(n)}.$$

**The key assumption.** There exist constants $c_0, n_0 > 0$ such that

$$\left| \frac{nL'(n)}{L(n)} \right| < \frac{c_0}{\log n} \quad \text{for } n > n_0. \quad (6.5)$$

In particular, $L$ is itself slowly varying, which is equivalent to the slow variation of $n\Phi_0'(n)$ as $n \to \infty$. Note that the faster $L$, the slower $\Phi_0$. The assumption allows to limit local variations of $\Phi_0$, which makes possible approximating the compensator by a simpler process

$$A^*_n(t) := \int_0^{t \wedge \log n} \Phi_0(ne^{-u}) \left( 1 - \frac{S_u - u}{L(ne^{-u})} \right) du,$$

in which the subordinator enters linearly. This in turn allows to derive the limit behaviour of the compensator from the functional CLT for $(S_t)$ itself.

6.3.1. Moderate growth case

This is the case $L(n) \asymp \log n$. We shall state approximations in the sense of bounded Wasserstein distance on random elements of the Skorohod space $D_0(\mathbb{R}^+)$ of càdlàg functions with finite limits at $\infty$ (see [5] for details).

The time-changed scaled process

$$\mathcal{K}^{(1)}_n(u) := (\Phi_0(n)\sqrt{\log n})^{-1} \left( \mathcal{K}_n(u \log n) - \log n \int_0^{u \wedge 1} \Phi_0(n^{1-v}) dv \right)$$

is approximable by the process

$$Y^{(1)}_n(u) := s \int_0^{u \wedge 1} \hat{h}^{(1)}_n(v) B_v dv,$$
Where \((B_u)\) is the BM and
\[
\hat{h}^{(1)}_n(u) := \frac{\Phi_0(n^{1-u}) \log n}{\Phi_0(n)L(n^{1-u})}.
\]
In particular, if \(L(n) \sim \gamma \log n\) for some \(\gamma > 0\), then \(K^{(1)}_n\) weakly converges to
\[
Y^{(1)}(u) = s \int_0^{u \wedge 1} \gamma^{-1} (1 - v)^{(1 - \gamma)/\gamma} B_v dv.
\]

**Example** Consider a gamma-like subordinator with Laplace exponent \(\Phi(n) \sim c \log \frac{1}{n^{1/\gamma}}, m = E[S_1] = \Phi'(0), v^2 = Var[S_1] = \Phi''(0)\). A CLT for \(K_n\) holds with standard scaling and centering by the moments
\[
E^K_n \sim \frac{c \log^{1+1/\gamma} n}{m(1 + 1/\gamma)}, \quad Var^K_n \sim \frac{c^2 v^2 \log^{1+2/\gamma} n}{m^2(1 + 2/\gamma)}.
\]
A special case is the gamma subordinator with
\[
\tilde{\nu}(dy) = \theta e^{-\theta y} dy/y, \quad \Phi(n) = \theta \log(1 + n/\theta), \quad m = 1, \quad v^2 = 1/\theta.
\]
This instance and generalisations were studied in [35].

**6.3.2. Fast growth case**

This case is defined by the conditions \(L(n) \to \infty, L(n) \ll \log n\), then \(\Phi_0\) grows faster than any power of the logarithm. For instance \(\Phi_0(n) \asymp \exp(\gamma n)\) with \(0 < \gamma < 1\). The scaled process
\[
K^{(2)}_n(u) := (\Phi_0(n)\sqrt{L(n)})^{-1} \times \left( K_n(uL(n)) - L(n) \int_0^{u \wedge (\log n/L(n))} \Phi_0(n \exp(-vL(n))) dv \right)
\]
converges to
\[
Y^{(2)}(u) := s \int_0^u e^{-v} B_v dv.
\]

**6.3.3. Slow growth case**

Suppose that \(L(n) = c(n) \log n\), where \(c(n) \to \infty\) but slowly enough to have
\[
\int_2^\infty \frac{dn}{c(n)n \log n} = \infty
\]
(otherwise \(\tilde{\nu}\) is a finite measure). For instance we can have \(c(n) \asymp \log \log n\) (in which case \(\Phi(n) \asymp \log \log n\), but the growth \(c(n) \asymp \log^\gamma n\) with \(\gamma > 0\) is
excluded. Like in the case of finite $\tilde{\nu}$, almost all variability of $K_n$ comes from the range of times very close to the passage time $\tau_n$.

The key quantity describing the process $K_n(t)$ in this case is the family of integrals

$$\int_{(\tau_n-t)_+}^{\tau_n} \Phi_0(e^v)dv, \quad t \geq 0,$$

where $\tau_n$ is the passage time at level $\log n$. The randomness enters here only through $\tau_n$, which is approximately normal for large $n$ with $\mathbb{E}[\tau_n] \sim \log n$, $\text{Var}[\tau_n] \sim s^2 \log n$. The process

$$K_n^{(3)}(t) := (\Phi_0(n) \sqrt{\log n})^{-1} \left( K_n(t) - \int_{(\log n-t)_+}^{\log n} \Phi_0(e^v)dv \right)$$

is approximable by

$$Y^{(3)}(t) =: \eta - (\Phi_0(n) \sqrt{\log n})^{-1} \int_{(\log n-t+\eta \sqrt{\log n})_+}^{(\log n-t)_+} \Phi_0(e^v)dv,$$

where $\eta$ is a standard normal random variable.

6.3.4. Gamma subordinators and the like

We know more (see [35]) about the asymptotics of small-block counts $K_{n,r}$ for gamma-like subordinators with logarithmic behaviour $\tilde{\nu}[y, \infty] \sim -c \log y$ for $y \to 0$, under some additional assumptions on the tail of $\tilde{\nu}$ for $y$ near 0 and $\infty$. This class of subordinators is well suited for application of singularities analysis to formulas like

$$\int_0^\infty n^{r-1} \phi_r(n)dn = \frac{\Gamma(r+s)}{r!} \frac{\Phi(-s : -s)}{\Phi(-s)}, \quad -1 < \Re s < 0$$

for the Mellin transform of the expected value $\phi_r(n)$ of poissonised $K_{n,r}$. In this formula $\Phi(s)$ and $\Phi(-s : -s)$ are the analytical continuations in the complex domain of the Laplace exponent $\Phi(n)$ and the bivariate function $\Phi(n : n)$, respectively.

For the moments we have

$$\mathbb{E}[K_n] \sim \frac{\log^2 n}{2m}, \quad \text{Var}[K_n] \sim \frac{\nu^2 \log^3 n}{3m},$$

$$\mathbb{E}[K_{n,r}] \sim \frac{\log n}{rm}, \quad \text{Var}[K_{n,r}] \sim \left( \frac{\nu^2}{r^2 m^3} + \frac{1}{rm} \right) \log n.$$
converges in distribution to a multivariate Gaussian sequence with the covariance matrix
\[ \frac{\nu^2}{m^3} \frac{1}{i_j} + 1(i = j) \frac{1}{j m}, \quad i, j = 1, 2, \ldots. \]

See [35] for explicit assumptions on $\tilde{\nu}$ in this logarithmic case and further examples, including computations for the subordinator in Example 3.5.

The behaviour of $K_{n,r}$’s for regenerative composition structures with other slowly varying infinite Lévy measures $\tilde{\nu}$ remains an open problem.

Acknowledgements

These notes were presented in a series of lectures at the School on Information and Randomness 2008, in Santiago de Chile. The author is indebted to the organisers Servet Martinez, Alejandro Maass and Jaime San Martin for the invitation and hospitality.

References


