

# Characterizations of exchangeable partitions and random discrete distributions by deletion properties <sup>\*</sup>

Alexander Gnedin<sup>†</sup>      Chris Haulk<sup>‡</sup>      Jim Pitman<sup>§</sup>

November 20, 2009

## Abstract

We prove a long-standing conjecture which characterises the Ewens-Pitman two-parameter family of exchangeable random partitions, plus a short list of limit and exceptional cases, by the following property: for each  $n = 2, 3, \dots$ , if one of  $n$  individuals is chosen uniformly at random, independently of the random partition  $\pi_n$  of these individuals into various types, and all individuals of the same type as the chosen individual are deleted, then for each  $r > 0$ , given that  $r$  individuals remain, these individuals are partitioned according to  $\pi'_r$  for some sequence of random partitions  $(\pi'_r)$  which does not depend on  $n$ . An analogous result characterizes the associated Poisson-Dirichlet family of random discrete distributions by an independence property related to random deletion of a frequency chosen by a size-biased pick. We also survey the regenerative properties of members of the two-parameter family, and settle a question regarding the explicit arrangement of intervals with lengths given by the terms of the Poisson-Dirichlet random sequence into the interval partition induced by the range of a homogeneous neutral-to-the right process.

## Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>   | <b>2</b>  |
| <b>2</b> | <b>Partition Structures</b>                                   | <b>2</b>  |
| <b>3</b> | <b>Partially Exchangeable Partitions</b>                      | <b>9</b>  |
| <b>4</b> | <b>Exchangeable Partitions</b>                                | <b>12</b> |
| <b>5</b> | <b>The deletion property without the regularity condition</b> | <b>18</b> |
| <b>6</b> | <b>Regeneration and <math>\tau</math>-deletion</b>            | <b>19</b> |

---

<sup>\*</sup>This research is supported in parts by N.S.F. Award 0806118.

<sup>†</sup>University of Utrecht; email A.V.Gnedin@uu.nl

<sup>‡</sup>University of California at Berkeley; email haulk@stat.berkeley.edu

<sup>§</sup>University of California at Berkeley; email pitman@stat.berkeley.edu

# 1 Introduction

Kingman [14] introduced the concept of a *partition structure*, that is a family of probability distributions for random partitions  $\pi_n$  of a positive integer  $n$ , with a sampling consistency property as  $n$  varies. Kingman's work was motivated by applications in population genetics, where the partition of  $n$  may be the allelic partition generated by randomly sampling a set of  $n$  individuals from a population of size  $N \gg n$ , considered in a large  $N$  limit which implies sampling consistency. Subsequent authors have established the importance of Kingman's theory of partition structures, and representations of these structures in terms of exchangeable random partitions and random discrete distributions [21], in a number of other settings, which include the theory of species sampling [22], random trees and associated random processes of fragmentation and coalescence [23, 9, 1, 2], Bayesian statistics and machine learning [26, 27]. Kingman [13] showed that the Ewens sampling formula from population genetics defines a particular partition structure  $(\pi_n)$ , which he characterized by the following property, together with the regularity condition that  $\mathbb{P}(\pi_n = \lambda) > 0$  for every partition  $\lambda$  of  $n$ :

for each  $n = 2, 3, \dots$ , if an individual is chosen uniformly at random independently of a random partitioning of these individuals into various types according to  $\pi_n$ , and all individuals of the same type as the chosen individual are deleted, then conditionally given that the number of remaining individuals is  $r > 0$ , these individuals are partitioned according to a copy of  $\pi_r$ .

We establish here a conjecture of Pitman [20] that if this property is weakened by replacing  $\pi_r$  by  $\pi'_r$  for some sequence of random partitions  $(\pi'_r)$ , and a suitable regularity condition is imposed, then  $(\pi_n)$  belongs to the two-parameter family of partition structures introduced in [20]. Theorem 3 below provides a more careful statement. We also present a corollary of this result, to characterize the two-parameter family of Poisson-Dirichlet distributions by an independence property of a single size-biased pick, thus improving upon [21].

Kingman's characterization of the Ewens family of partition structures by deletion of a type has been extended in another direction by allowing other deletion algorithms but continuing to require that the distribution of the partition structure be preserved. The resulting theory of *regenerative partition structures* [6], is connected to the theory of regenerative sets, including Kingman's regenerative phenomenon [12], on a multiplicative scale. In the last section of the paper we review such deletion properties of the two-parameter family of partition structures, and offer a new proof of a result of Pitman and Winkel [25] regarding the explicit arrangement of intervals with lengths given by the terms of the Poisson-Dirichlet random sequence into the interval partition induced by a multiplicatively regenerative set.

## 2 Partition Structures

This section briefly reviews Kingman's theory of partition structures, which provides the general context of this article. To establish some terminology and notation for use throughout the paper, recall that a *composition*  $\lambda$  of a positive integer  $n$  is a sequence

of positive integers  $\lambda = (\lambda_1, \dots, \lambda_k)$ , with  $\sum_{i=1}^k \lambda_i = n$ . Both  $k = k_\lambda$  and  $n = n_\lambda$  may be regarded as functions of  $\lambda$ . Each term  $\lambda_i$  is called a *part* of  $\lambda$ . A *partition*  $\lambda$  of  $n$  is a multiset of positive integers whose sum is  $n$ , commonly identified with the composition of  $n$  obtained by putting its positive integer parts in decreasing order, or with the infinite sequence of non-negative integers obtained by appending an infinite string of zeros to this composition of  $n$ . So

$$\lambda = (\lambda_1, \lambda_2, \dots) \text{ with } \lambda_1 \geq \lambda_2 \geq \dots \geq 0$$

represents a partition of  $n = n_\lambda$  into  $k = k_\lambda$  parts, where

$$n_\lambda := \sum_i \lambda_i \text{ and } k_\lambda := \max\{i : \lambda_i > 0\}.$$

Informally, a partition  $\lambda$  describes an unordered collection of  $n_\lambda$  balls of  $k_\lambda$  different colors, with  $\lambda_i$  balls of the  $i$ th most frequent color. A *random partition of  $n$*  is a random variable  $\pi_n$  with values in the finite set of all partitions  $\lambda$  of  $n$ . Kingman [14] defined a *partition structure* to be a sequence of random partitions  $(\pi_n)_{n \in \mathbb{N}}$  which is *sampling consistent* in the following sense:

if a ball is picked uniformly at random and deleted from  $n$  balls randomly colored according to  $\pi_n$ , then the random coloring of the remaining  $n - 1$  balls is distributed according to  $\pi_{n-1}$ .

As shown by Kingman [15], the theory of partition structures and associated partition-valued processes is best developed in terms of random partitions of the set of positive integers. Our treatment here follows [20]. If we regard a random partition  $\pi_n$  of a positive integer  $n$  as a random coloring of  $n$  unordered balls, an associated random partition  $\Pi_n$  of the set  $[n] := \{1, \dots, n\}$  may be obtained by placement of the colored balls in a row. We will assume for the rest of this introduction that this placement is made by a random permutation which given  $\pi_n$  is uniformly distributed over all  $n!$  possible orderings of  $n$  distinct balls.

Formally, a partition of  $[n]$  is a collection of disjoint non-empty blocks  $\{B_1, \dots, B_k\}$  with  $\cup_{i=1}^k B_i = [n]$  for some  $1 \leq k \leq n$ , where each  $B_i \subseteq [n]$  represents the set of places occupied by balls of some particular color. We adopt the convention that the blocks  $B_i$  are listed *in order of appearance*, meaning that  $B_i$  is the set of places in the row occupied by balls of the  $i$ th color to appear. So  $1 \in B_1$ , and if  $k \geq 2$  the least element of  $B_2$  is the least element of  $[n] \setminus B_1$ , if  $k \geq 3$  the least element of  $B_3$  is the least element of  $[n] \setminus (B_1 \cup B_2)$ , and so on. This enumeration of blocks identifies each partition of  $[n]$  with an *ordered partition*  $(B_1, \dots, B_k)$ , subject to these constraints. The sizes of parts  $(|B_1|, \dots, |B_k|)$  of this partition form a composition of  $n$ . The notation  $\Pi_n = (B_1, \dots, B_k)$  is used to signify that  $\Pi_n = \{B_1, \dots, B_k\}$  for some particular sequence of blocks  $(B_1, \dots, B_k)$  listed in order of appearance. If  $\Pi_n$  is derived from  $\pi_n$  by uniform random placement of balls in a row, then  $\Pi_n$  is *exchangeable*, meaning that its distribution is invariant under every deterministic rearrangement of places by a permutation of  $[n]$ . Put another way, for each partition  $(B_1, \dots, B_k)$  of  $[n]$ , with blocks in order of appearance,

$$\mathbb{P}(\Pi_n = (B_1, \dots, B_k)) = p(|B_1|, \dots, |B_k|) \tag{1}$$

for a function  $p = p(\lambda)$  of compositions  $\lambda$  of  $n$  which is a symmetric function of its  $k$  arguments for each  $1 \leq k \leq n$ . Then  $p$  is called the *exchangeable partition probability function* (EPPF) associated with  $\Pi_n$ , or with  $\pi_n$ , the partition of  $n$  defined by the unordered sizes of blocks of  $\Pi_n$ .

As observed by Kingman [15],  $(\pi_n)$  is sampling consistent if and only if the sequence of partitions  $(\Pi_n)$  can be constructed to be consistent in the sense that for  $m < n$  the restriction of  $\Pi_n$  to  $[m]$  is  $\Pi_m$ . This amounts to a simple recursion formula satisfied by  $p$ , recalled later as (20). The sequence  $\Pi = (\Pi_n)$  can then be interpreted as a random partition of the set  $\mathbb{N}$  of all positive integers, whose restriction to  $[n]$  is  $\Pi_n$  for every  $n$ . Such  $\Pi$  consists of a sequence of blocks  $\mathcal{B}_1, \mathcal{B}_2, \dots$ , which may be identified as random disjoint subsets of  $\mathbb{N}$ , with  $\cup_{i=1}^{\infty} \mathcal{B}_i = \mathbb{N}$ , where the nonempty blocks are arranged by increase of their minimal elements, and if the number of nonempty blocks is some  $K < \infty$ , then by convention  $\mathcal{B}_i = \emptyset$  for  $i > K$ . Similarly,  $\Pi_n$  consists of a sequence of blocks  $\mathcal{B}_{ni} := \mathcal{B}_i \cap [n]$ , where  $\cup_i \mathcal{B}_{ni} = [n]$ , and the nonempty blocks are consistently arranged by increase of their minimal elements, for all  $n$ .

These considerations are summarized by the following proposition:

**Proposition 1** (Kingman [15]) *The most general partition structure, defined by a sampling consistent collection of distributions for partitions  $\pi_n$  of integers  $n$ , is associated with a unique probability distribution of an exchangeable partition of positive integers  $\Pi = (\Pi_n)$ , as determined by an EPPF  $p$  according to (1).*

We now recall a form of *Kingman's paintbox construction* of such an exchangeable random partition  $\Pi$  of positive integers. Regard the unit interval  $[0, 1)$  as a continuous spectrum of distinct colors, and suppose given a sequence of random variables  $(P_1^\downarrow, P_2^\downarrow, \dots)$  called *ranked frequencies*, subject to the constraints

$$1 \geq P_1^\downarrow \geq P_2^\downarrow \geq \dots \geq 0, \quad P_* := 1 - \sum_{j=1}^{\infty} P_j^\downarrow \geq 0. \quad (2)$$

The color spectrum is partitioned into a sequence of intervals  $[l_i, r_i)$  of lengths  $P_i^\downarrow$ , and in case  $P_* > 0$  a further interval  $[1 - P_*, 1)$  of length  $P_*$ . Each point  $u$  of  $[0, 1)$  is assigned the color  $c(u) = l_i$  if  $u \in [l_i, r_i)$  for some  $i = 1, 2, \dots$ , and  $c(u) = u$  if  $u \in [1 - P_*, 1)$ . This coloring of points of  $[0, 1)$ , called *Kingman's paintbox* associated with  $(P_1^\downarrow, P_2^\downarrow, \dots)$ , is sampled by an infinite sequence of independent uniform $[0, 1]$  variables  $U_i$ , to assign a color  $c(U_i)$  to the  $i$ th ball in a row of balls indexed by  $i = 1, 2, \dots$ . The associated *color partition* of  $\mathbb{N}$  is generated by the random equivalence relation  $\sim$  defined by  $i \sim j$  if and only if  $c(U_i) = c(U_j)$ , meaning that either  $U_i$  and  $U_j$  fall in the same compartment of the paintbox, or that  $i = j$  and  $U_i$  falls in  $[1 - P_*, 1)$ .

**Theorem 2** (Kingman's paintbox representation of exchangeable partitions [14]) *Each exchangeable partition  $\Pi$  of  $\mathbb{N}$  generates a sequence of ranked frequencies  $(P_1^\downarrow, P_2^\downarrow, \dots)$  such that the conditional distribution of  $\Pi$  given these frequencies is that of the color partition of  $\mathbb{N}$  derived from  $(P_1^\downarrow, P_2^\downarrow, \dots)$  by Kingman's paintbox construction. The exchangeable partition probability function  $p$  associated with  $\Pi$  determines the distribution of  $(P_1^\downarrow, P_2^\downarrow, \dots)$ , and vice versa.*

The distributions of ranked frequencies  $(P_1^\downarrow, P_2^\downarrow, \dots)$  associated with naturally arising partition structures  $(\pi_n)$  are quite difficult to deal with analytically. See for instance [24]. Still,  $(P_1^\downarrow, P_2^\downarrow, \dots)$  can be constructed as the decreasing rearrangement of the frequencies  $P_i$  of blocks  $\mathcal{B}_i$  of  $\Pi$  defined as the almost sure limits

$$P_i = \lim_{n \rightarrow \infty} n^{-1} |\mathcal{B}_{ni}| \quad (3)$$

where  $i = 1, 2, \dots$  indexes the blocks in order of appearance, while

$$P_* = 1 - \sum_{i=1}^{\infty} P_i = 1 - \sum_{i=1}^{\infty} P_i^\downarrow$$

is the asymptotic frequency of the union of singleton blocks

$$\mathcal{B}_* := \cup_{\{i: |\mathcal{B}_i|=1\}} \mathcal{B}_i,$$

so that (3) holds also for  $i = *$ . The frequencies are called *proper* if  $P_* = 0$  a.s.; then almost surely every nonempty block  $\mathcal{B}_i$  of  $\Pi$  has a strictly positive frequency, hence  $|\mathcal{B}_i| = \infty$ , while every block  $\mathcal{B}_i$  with  $0 < |\mathcal{B}_i| < \infty$  is a singleton block.

The ranked frequencies  $P_1^\downarrow, P_2^\downarrow, \dots$  appear in the sequence  $(P_j)$  in the order in which intervals of these lengths are discovered by a process of uniform random sampling, as in Kingman's paintbox construction. If  $P_* > 0$  then in addition to the strictly positive terms of  $P_1^\downarrow, P_2^\downarrow, \dots$  the sequence  $(P_i)$  also contains infinitely many zeros which correspond to singletons in  $\Pi$ . The conditional distribution of  $(P_j)$  given  $(P_j^\downarrow)$  can also be described in terms of iteration of a single size-biased pick, defined as follows. For a sequence of non-negative random variables  $(X_i)$  with  $\sum_i X_i \leq 1$  and a random index  $J \in \{1, 2, \dots, \infty\}$ , call  $X_J$  a *size-biased pick* from  $(X_i)$  if  $X_J$  has value  $X_j$  if  $J = j < \infty$  and  $X_J = 0$  if  $J = \infty$ , with

$$\mathbb{P}(J = j | (X_i, i \in \mathbb{N})) = X_j \quad (0 < j < \infty) \quad (4)$$

(see [7] for this and another definition of size-biased pick in the case of improper frequencies). The *sequence derived from  $(X_i)$  by deletion of  $X_J$  and renormalization* refers to the sequence  $(Y_i)$  obtained from  $(X_i)$  by first deleting the  $J$ th term  $X_J$ , then closing up the gap if  $J \neq \infty$ , and finally normalizing each term by  $1 - X_J$ . Here by convention,  $(Y_i) = (X_i)$  if  $X_J = 0$  and  $(Y_i)$  is the zero sequence if  $X_J = 1$ . Then  $P_1$  is a size-biased pick from  $(P_j^\downarrow)$ ,  $P_2$  is a size-biased pick from the sequence derived from  $(P_j^\downarrow)$  by deletion of  $P_1$  and renormalization, and so on. For this reason,  $(P_i)$  is said to be a *size-biased permutation* of  $(P_i^\downarrow)$ .

**The two-parameter family** It was shown in [20] that for each pair of real parameters  $(\alpha, \theta)$  with

$$0 \leq \alpha < 1, \theta > -\alpha \quad (5)$$

the formula

$$p_{\alpha, \theta}(\lambda) := \frac{\prod_{i=1}^{k-1} (\theta + i\alpha)}{(\theta + 1)_{n-1}} \prod_{j=1}^k (1 - \alpha)_{\lambda_j - 1} \quad (6)$$

where  $k = k_\lambda$ ,  $n = n_\lambda$ , and

$$(x)_n := x(x+1)\dots(x+n-1) = \frac{\Gamma(x+n)}{\Gamma(x)}$$

is a rising factorial, defines the EPPF of an exchangeable random partition of positive integers whose block frequencies  $(P_i)$  in order of appearance admit the *stick-breaking representation*

$$P_i = W_i \prod_{j=1}^{i-1} (1 - W_j) \quad (7)$$

for random variables  $W_j$  such that

$$W_1, W_2, \dots \text{ are mutually independent} \quad (8)$$

with

$$W_k \stackrel{d}{=} \beta_{1-\alpha, \theta+k\alpha} \quad (9)$$

where  $\stackrel{d}{=}$  indicates equality in distribution, and  $\beta_{a,b}$  for  $a, b > 0$  denotes a random variable with the beta( $a, b$ ) density

$$\mathbb{P}(\beta_{a,b} \in du) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1} (1-u)^{b-1} du \quad (0 < u < 1) \quad (10)$$

which is also characterized by the moments

$$\mathbb{E}[\beta_{a,b}^i (1 - \beta_{a,b})^j] = \frac{(a)_i (b)_j}{(a+b)_{i+j}} \quad (i, j = 0, 1, 2, \dots). \quad (11)$$

Formula (6) also defines an EPPF for  $(\alpha, \theta)$  in the range

$$\alpha < 0, \theta = -M\alpha \text{ for some } M \in \mathbb{N}, \quad (12)$$

in which case the stick-breaking representation (7) with factors as in (9) makes sense for  $1 \leq k \leq M$ , with the last factor  $W_M = 1$ . The frequencies  $(P_1, \dots, P_M)$  in this case are a size-biased random permutation of  $(Q_1, \dots, Q_M)$  with the symmetric Dirichlet distribution with  $M$  parameters equal to  $\nu := -\alpha > 0$ . It is well known that the  $Q_i$  can be constructed as  $Q_i = \gamma_\nu^{(i)} / \Sigma, 1 \leq i \leq M$ , where  $\Sigma = \sum_{i=1}^M \gamma_\nu^{(i)}$  and the  $\gamma_\nu^{(i)}$  are independent and identically distributed copies of a gamma variable  $\gamma_\nu$  with density

$$\mathbb{P}(\gamma_\nu \in dx) = \Gamma(\nu)^{-1} x^{\nu-1} e^{-x} dx \quad (x > 0). \quad (13)$$

As shown by Kingman [13], the  $(0, \theta)$  EPPF (6) for  $\alpha = 0, \theta > 0$  arises in the limit of random sampling from such symmetric Dirichlet frequencies as  $\nu = -\alpha \downarrow 0$  and  $M \uparrow \infty$  with  $\nu M = \theta$  held fixed. In this case, the distribution of the partition  $\pi_n$  is that determined by the Ewens sampling formula with parameter  $\theta$ , the residual fractions  $W_i$  in the stick-breaking representation are identically distributed like  $\beta_{1,\theta}$ , and the ranked frequencies  $P_i^\downarrow$  can be obtained by normalization of the jumps of a gamma process with stationary independent increments  $(\gamma_\nu, 0 \leq \nu \leq \theta)$ . Perman, Pitman and Yor [17] gave extensions

of this description to the case  $0 < \alpha < 1$  when the distribution of ranked frequencies can be derived from the jumps of a stable subordinator of index  $\alpha$ . See also [24, 18, 19] for further discussion and applications to the description of ranked lengths of excursion intervals of Brownian motion and Bessel processes.

In the limit case when  $\nu = -\alpha \rightarrow \infty$  and  $\theta = M\nu \rightarrow \infty$ , for a fixed positive integer  $M$ , the EPPF (6) converges to

$$p_M(\lambda) := \frac{M(M-1)\cdots(M-k+1)}{M^n}, \quad (14)$$

corresponding to sampling from  $M$  equal frequencies

$$P_1 = P_2 = \cdots = P_M = 1/M$$

as in the classical coupon collector's problem with some fixed number  $M$  of equally frequent types of coupon. We refer to the collection of partition structures defined by (6) for the parameter ranges (5) and (12), as well as the limit cases (14), as the *extended two-parameter family*.

The partition  $\mathbf{0}$  of  $\mathbb{N}$  into singletons and the partition  $\mathbf{1}$  of  $\mathbb{N}$  into a single block both belong to the closure of the two-parameter family. As noticed by Kerov [11], a mixture of these two trivial partitions with mixing proportions  $t$  and  $1-t$  also belongs to the closure, as is seen from (6) by letting  $\alpha \rightarrow 1$  and  $\theta \rightarrow -1$  in such a way that  $(1-\alpha)/(\theta+1) \rightarrow t$  and  $(\theta+\alpha)/(\theta+1) \rightarrow 1-t$ .

**Characterizations by deletion properties** The main focus of this paper is the following result, which was first conjectured by Pitman [20]. For convenience in presenting this result, we impose the following mild *regularity condition* on the EPPF  $p$  associated with a partition structure  $(\pi_n)$ :

$$p(2, 2, 1) > 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} p(n) = 0. \quad (15)$$

Equivalently, in terms of the frequencies  $P_i$  in order of appearance,

$$\mathbb{P}(0 < P_1 < P_1 + P_2 < 1) > 0 \quad \text{and} \quad \mathbb{P}(P_1 = 1) = 0, \quad (16)$$

or again, in terms of the ranked frequencies  $P_i^\downarrow$ ,

$$\mathbb{P}(0 < P_2^\downarrow, P_1^\downarrow + P_2^\downarrow < 1) > 0 \quad \text{and} \quad \mathbb{P}(P_1^\downarrow = 1) = 0. \quad (17)$$

Note that this regularity condition does not rule out the case of improper frequencies. See Section 5 for discussion of how the following results can be modified to accommodate partition structures not satisfying the regularity condition.

**Theorem 3** *Among all partition structures  $(\pi_n)$  with EPPF  $p$  subject to (15), the extended two-parameter family is characterized by the following property:*

*if one of  $n$  balls is chosen uniformly at random, independently of a random coloring of these balls according to  $\pi_n$ , then given the number of other balls of the same color as the chosen ball is  $m-1$ , for some  $1 \leq m < n$ , the coloring of the remaining  $n-m$  balls is distributed according to  $\pi'_{n-m}$  for some sequence of partitions  $(\pi'_1, \pi'_2, \dots)$  which does not depend on  $n$ .*

Moreover, if  $(\pi_n)$  has the  $(\alpha, \theta)$  EPPF (6), then  $(\pi'_n)$  has the  $(\alpha, \theta + \alpha)$  EPPF (6), whereas if  $(\pi_n)$  has the EPPF (14) for some  $M$ , then the EPPF of  $(\pi'_n)$  has the same form except with  $M$  decremented by 1.

Note that it is not assumed as part of the property that  $(\pi'_n)$  is a partition structure. Rather, this is implied by the conclusion. Our formulation of Theorem 3 was inspired by Kingman [13] who assumed also that  $\pi'_n \stackrel{d}{=} \pi_n$  for all  $n$ . The conclusion then holds with  $\alpha = 0$ , in which case the distribution of  $\pi_n$  is that determined by the Ewens sampling formula from population genetics.

In Section 4 we offer a proof of Theorem 3 by purely combinatorial methods. Some preliminary results which we develop in Section 3 allow Theorem 3 to be reformulated in terms of frequencies as in the following Corollary:

**Corollary 4** *Let the asymptotic frequencies  $(P_i)$  of an exchangeable random partition of positive integers  $\Pi$  be represented in the stick-breaking form (7) for some sequence of random variables  $W_1 < 1, W_2, \dots$ . The condition*

$$W_1 \text{ is independent of } (W_2, W_3, \dots) \tag{18}$$

*obtains if and only if*

$$\text{the } W_i \text{ are mutually independent.} \tag{19}$$

*If in addition to (18) the regularity condition (15) holds, then  $\Pi$  is governed by the extended two-parameter family, either with  $W_i \stackrel{d}{=} \beta_{1-\alpha, \theta+i\alpha}$ , or with  $W_i = 1/(M - i + 1)$  for  $1 \leq i \leq M$ , as in the limit case (14), for some  $M = 3, 4, \dots$*

The characterization of the two-parameter family using (19) rather than the weaker condition (18) was provided by Pitman [21]. As we show in Section 4, it is possible to derive (19) directly from (18), without passing via Theorem 3.

The law of frequencies  $(P_i)$  defined by the stick-breaking scheme (7) for independent factors  $W_i$  with  $W_i \stackrel{d}{=} \beta_{1-\alpha, \theta+i\alpha}$  is known as the two-parameter Griffiths-Engen-McCloskey distribution, denoted  $\text{GEM}(\alpha, \theta)$ . The property of the independence of residual proportions  $W_i$ , also known as *complete neutrality*, has also been studied extensively in connection with finite-dimensional Dirichlet distributions [3].

The above results can also be expressed in terms of ranked frequencies. Recall that the distribution of ranked frequencies  $(P_k^\downarrow)$  of an  $(\alpha, \theta)$ -partition is known as the *two-parameter Poisson-Dirichlet distribution*  $\text{PD}(\alpha, \theta)$ . According to the previous discussion, a random sequence  $(P_k^\downarrow)$  with  $\text{PD}(\alpha, \theta)$  distribution is obtained by ranking a sequence  $(P_i)$  with  $\text{GEM}(\alpha, \theta)$  distribution. The  $\text{PD}(\alpha, \theta)$  distribution was systematically studied in [24], and has found numerous further applications to random trees and associated processes of fragmentation and coagulation [23, 9, 2].

**Corollary 5** *Let  $(P_k^\downarrow)$  be a decreasing sequence of ranked frequencies subject to the regularity condition (2) and (17). For  $P_j^\downarrow$ , a size-biased pick from  $(P_k^\downarrow)$ , let  $(Q_k^\downarrow)$  be derived from  $(P_k^\downarrow)$  by deletion of  $P_j^\downarrow$  and renormalization. The random variable  $P_j^\downarrow$  is independent of the sequence  $(Q_k^\downarrow)$  if and only if either the distribution of  $(P_k^\downarrow)$  is  $\text{PD}(\alpha, \theta)$  for some*

$(\alpha, \theta)$ , or  $P_k^\downarrow = 1/M$  for all  $1 \leq k \leq M$ , for some  $M \geq 3$ . In the former case, the distribution of  $(Q_k^\downarrow)$  is  $\text{PD}(\alpha, \theta + \alpha)$ , whereas in the latter case, the deletion and renormalization simply decrements  $M$  by one.

The ‘if’ part of this Corollary is Proposition 34 of Pitman-Yor [24], while the ‘only if’ part follows easily from Corollary 4, using Kingman’s paintbox representation.

### 3 Partially Exchangeable Partitions

We start by recalling from [20] some basic properties of *partially exchangeable partitions of positive integers*, which are consistent sequences  $\Pi = (\Pi_n)$ , where  $\Pi_n$  is a partition of  $[n]$  whose probability distribution is of the form (1) for some function  $p = p(\lambda)$  of compositions  $\lambda$  of positive integers. The consistency of  $\Pi_n$  as  $n$  varies amounts to the *addition rule*

$$p(\lambda) = \sum_{j=1}^{k+1} p(\lambda^{(j)}), \quad (20)$$

where  $k = k_\lambda$  is the number of parts of  $\lambda$ , and  $\lambda^{(j)}$  is the composition of  $n_\lambda + 1$  derived from  $\lambda$  by incrementing  $\lambda_j$  to  $\lambda_j + 1$ , and leaving all other components of  $\lambda$  fixed. In particular, for  $j = k_\lambda + 1$  this means appending a 1 to  $\lambda$ . There is also the normalization condition  $p(1) = 1$ . To illustrate (20) for  $\lambda = (3, 1, 2)$ :

$$p(3, 1, 2) = p(4, 1, 2) + p(3, 2, 2) + p(3, 1, 3) + p(3, 1, 2, 1).$$

The following proposition recalls the analog of Kingman’s representation for partially exchangeable partitions:

**Proposition 6** (Corollary 7 from [20]) *Every partially exchangeable partition of positive integers  $\Pi$  is such that for each  $k \geq 1$ , the  $k$ th block  $\mathcal{B}_k$  has an almost sure limit frequency  $P_k$ . The partition probability function  $p$  can then be presented as*

$$p(\lambda) = \mathbb{E} \left[ \prod_{i=1}^k P_i^{\lambda_i - 1} \prod_{j=1}^{k-1} R_j \right], \quad (21)$$

where  $k = k_\lambda$  and  $R_j := (1 - P_1 - \dots - P_j)$ . Alternatively, in terms of the residual fractions  $W_k$  in the stick-breaking representation (7):

$$p(\lambda) = \mathbb{E} \left[ \prod_{i=1}^k W_i^{\lambda_i - 1} \overline{W}_i^{\Lambda_{i+1}} \right], \quad (22)$$

where  $\overline{W}_i := 1 - W_i$ ,  $\Lambda_j := \sum_{i \geq j} \lambda_i$ . This formula sets up a correspondence between the probability distribution of  $\Pi$ , encoded by the partition probability function  $p$ , and an arbitrary joint distribution of a sequence of random variables  $(W_1, W_2, \dots)$  with  $0 \leq W_i \leq 1$  for all  $i$ .

In terms of randomly coloring a row of  $n_\lambda$  balls, the product whose expectation appears in (22) is the conditional probability given  $W_1, W_2, \dots$  of the event that the first  $\lambda_1$  balls are colored one color, the next  $\lambda_2$  balls another color, and so on. So (22) reflects the fact that conditionally given  $W_1, W_2, \dots$  the process of random coloring of integers occurs according to the following residual allocation scheme [20, Construction 16]:

Ball 1 is painted a first color, and so is each subsequent ball according to a sequence of independent trials with probability  $W_1$  of painting with color 1. The set of balls so painted defines the first block  $\mathcal{B}_1$  of  $\Pi$ . Conditionally given  $\mathcal{B}_1$ , the first unpainted ball is painted a second color, and so is each subsequent unpainted ball according to a sequence of independent trials with probability  $W_2$  of painting with color 2. The balls colored 2 define  $\mathcal{B}_2$ , and so on. Given an arbitrary sequence of random variables  $(W_k)$  with  $0 \leq W_k \leq 1$ , this coloring scheme shows how to construct a partially exchangeable partition of  $\mathbb{N}$  whose asymptotic block frequencies are given by the stick-breaking scheme (7).

Note that the residual allocation scheme terminates at the first  $k$ , if any, such that  $W_k = 1$ , by painting all remaining balls color  $k$ . The values of  $W_i$  for  $i$  larger than such a  $k$  have no effect on the construction of  $\Pi$ , so cannot be recovered from its almost sure limit frequencies. To ensure that a unique joint distribution of  $(W_1, W_2, \dots)$  is associated with each  $p$ , the convention may be adopted that the sequence  $(W_i)$  terminates at the first  $k$  if any such that  $W_k = 1$ . This convention will be adopted in the following discussion.

For  $W_i$  which are independent, formula (22) factorizes as

$$p(\lambda) = \prod_{i=1}^k \mathbb{E}(W_i^{\lambda_i-1} \overline{W}_i^{\Lambda_i+1}). \quad (23)$$

In particular, for independent  $W_i$  with the beta distributions (9), this formula is readily evaluated using (11) to obtain (6). Inspection of (6) shows that this function of compositions  $\lambda$  is a symmetric function of its parts. Hence the associated random partition  $\Pi$  is exchangeable.

There is an alternate sequential construction of the two-parameter family of partitions which has become known as the ‘‘Chinese Restaurant Process’’ (see [19], Chapter 3). Instead of coloring rows of balls, imagine customers entering a restaurant with an unlimited number of tables. Initially customer 1 sits at table 1. At stage  $n$ , if there are  $k$  occupied tables, the  $i$ th of them occupied by  $\lambda_i$  customers for  $1 \leq i \leq k$ , customer  $n+1$  sits at one of the previously occupied tables with probability  $(\lambda_i - \alpha)/(n + \theta)$ , and occupies a new table  $k+1$  with probability  $(\theta + k\alpha)/(n + \theta)$ . It is then readily checked that for each partition of  $[n]$  into blocks  $B_i$  with  $|B_i| = \lambda_i$ , after  $n$  customers labeled by  $[n]$  have entered the restaurant, the probability that those customers labeled by  $B_i$  sat at table  $i$  for each  $1 \leq i \leq k_\lambda$  is given by the product formula (6). Moreover, the stick-breaking description of the limit frequencies  $P_i$  is readily derived from the Pólya urn-scheme description of exchangeable trials which given a beta( $a, b$ )-distributed variable  $S$ , are independent with success probability  $S$ .

Continuing the consideration of a partially exchangeable partition  $\Pi$  of positive integers, we record the following Lemma.

**Lemma 7** *Let  $\Pi$  be a partially exchangeable random partition of  $\mathbb{N}$  with partition probability function  $p$ , and with blocks  $\mathcal{B}_1, \mathcal{B}_2, \dots$  and residual frequencies  $W_1, W_2, \dots$  such that  $W_1 < 1$  almost surely. Let  $\Pi'$  denote the partition of  $\mathbb{N}$  derived from  $\Pi$  by deletion of the block  $\mathcal{B}_1$  containing 1 and re-labeling of  $\mathbb{N} - \mathcal{B}_1$  by the increasing bijection with  $\mathbb{N}$ . Then the following hold:*

(i) *The partition  $\Pi'$  is partially exchangeable, with partition probability function*

$$p'(\lambda_2, \dots, \lambda_k) = \sum_{\lambda_1=1}^{\infty} \binom{\lambda_1 + \dots + \lambda_k - 2}{\lambda_1 - 1} p(\lambda_1, \lambda_2, \dots, \lambda_k) \quad (24)$$

*and residual frequencies  $W_2, W_3, \dots$*

(ii) *If  $\Pi$  is exchangeable, then so is  $\Pi'$ .*

(iii) *For  $1 \leq m \leq n$*

$$q(n : m) := \mathbb{P}(\mathcal{B}_1 \cap [n] = [m]) = \mathbb{E}(W_1^{m-1} \overline{W}_1^{n-m}), \quad (25)$$

*and there is the addition rule*

$$q(n : m) = q(n + 1 : m + 1) + q(n + 1 : m). \quad (26)$$

(iv) *Let  $T_n := \inf\{m : |[n + m] \setminus \mathcal{B}_1| = n\}$  which is the number of balls of the first color preceding the  $n$ th ball not of the first color. Then*

$$\mathbb{P}(T_n = m) = \binom{m + n - 2}{m - 1} q(n + m : m), \quad (27)$$

*and consequently*

$$\sum_{m=1}^{\infty} \binom{m + n - 2}{m - 1} q(n + m : m) = 1. \quad (28)$$

*Proof.* Formula (25) is read from the general construction of  $\mathcal{B}_1$  given  $W_1$  by assigning each  $i \geq 2$  to  $\mathcal{B}_1$  independently with the same probability  $W_1$ . The formulas (24) and (27) are then seen to be marginalizations of the following expression for the joint distribution of  $T_n$  and  $\Pi'_n$ , the restriction of  $\Pi'$  to  $[n]$ :

$$\mathbb{P}(T_n = m, \Pi'_n = (C_1, \dots, C_{k-1})) = \binom{m + n - 2}{m - 1} q(n + m : m) p(m, |C_1|, \dots, |C_{k-1}|) \quad (29)$$

for every partition  $(C_1, \dots, C_{k-1})$  of  $[n]$ . To check (29), observe that the event in question occurs if and only if  $\Pi_{n+m} = (B_1, \dots, B_k)$  for some blocks  $B_i$  with  $|B_1| = m$  and  $|B_i| = |C_{i-1}|$  for  $2 \leq i \leq k$ . Once  $B_1$  is chosen, each  $B_i$  for  $2 \leq i \leq k$  is the image of  $C_{i-1}$  via the increasing bijection from  $[n]$  to  $[n + m] \setminus B_1$ . For prescribed  $C_{i-1}, 2 \leq i \leq k$ , the choice of  $B_1 \subset [n + m]$  is arbitrary subject to the constraint that  $1 \in B_1$  and  $n + m \notin B_1$ . The number of choices is the binomial coefficient in (29), so the conclusion is evident.  $\square$

The connection between Theorem 3 and Corollary 4 is established by the following Lemma:

**Lemma 8** *Let  $\Pi$  be a partially exchangeable partition of  $\mathbb{N}$  with residual frequencies  $W_i$  such that  $\mathbb{P}(W_1 < 1) = 1$ , with the convention that the sequence terminates at the first  $k$  (if any) such that  $W_k = 1$ , so the joint distribution of  $(W_i)$  is determined uniquely by the partition probability function  $p$  of  $\Pi$ , and vice versa, according to formula (22). For  $\mathcal{B}_1$  the first block of  $\Pi$  with frequency  $W_1$ , let  $\Pi'$  be derived from  $\Pi$  by deleting block  $\mathcal{B}_1$  and relabeling the remaining elements as in Lemma 7. The following four conditions on  $\Pi$  are equivalent:*

- (i)  $W_1$  is independent of  $(W_2, W_3, \dots)$ .
- (ii) The partition probability function  $p$  of  $\Pi$  admits a factorization of the following form, for all compositions  $\lambda$  of positive integers with  $k \geq 2$  parts:

$$p(\lambda) = q(n_\lambda : \lambda_1)p'(\lambda_2, \dots, \lambda_k) \quad (30)$$

for some non-negative functions  $q(n : m)$  and  $p'(\lambda_2, \dots, \lambda_k)$ .

- (iii) For each  $1 \leq m < n$ , the conditional distribution of  $\Pi'_{n-m}$  given  $|\mathcal{B}_1 \cap [n]| = m$  depends only on  $n - m$ .
- (iv) The random set  $\mathcal{B}_1$  is independent of the random partition  $\Pi'$  of  $\mathbb{N}$ .

Finally, if these conditions hold, then (ii) holds in particular for  $q(n : m)$  as in (25) and  $p'(\lambda_2, \dots, \lambda_k)$  the partition probability function of  $\Pi'$ .

*Proof.* That (i) implies (ii) is immediate by combination of the moment formula (22), (24) and (25). Conversely, if (ii) holds for some  $q(n : m)$  and  $p'(\lambda_2, \dots, \lambda_k)$ , Lemma 7 implies easily that (ii) holds for  $q$  and  $p'$  as in that Lemma. So (ii) gives a formula of the form

$$\mathbb{E}[f(W_1)g(W_2, W_3, \dots)] = \mathbb{E}[f(W_1)]\mathbb{E}[g(W_2, W_3, \dots)], \quad (31)$$

where  $g$  ranges over a collection of bounded measurable functions whose expectations determine the law of  $W_2, W_3, \dots$ , and for the  $g$  associated with  $\lambda_2, \dots, \lambda_k$ , the function  $f(w)$  ranges over the polynomials  $w^{m-1}(1-w)^n$  where  $m = \lambda_1 \in \mathbb{N}$  and  $n = n_\lambda - \lambda_1 = \sum_{j=2}^k \lambda_j$ . But linear combinations of these polynomials can be used to uniformly approximate any bounded continuous function of  $w$  on  $[0, 1]$  which vanishes in a neighbourhood of 1. It follows that (31) holds for all such  $f$ , for each  $g$ , hence the full independence condition (i). Lastly, the equivalence of (ii), (iii) and (iv) is easily verified.  $\square$

## 4 Exchangeable Partitions

For a block  $B$  of a random partition  $\Pi_n$  of  $[n]$  with  $|B| = m$ , let  $\Pi_n \setminus B$  denote the partition of  $[n - m]$  obtained by first deleting the block  $B$  of  $\Pi_n$ , then mapping the restriction of  $\Pi_n$  to  $[n] \setminus B$  to a partition of  $[n - m]$  via the increasing bijection between  $[n] \setminus B$  and  $[n - m]$ . In terms of a coloring of  $n$  balls in a row, this means deleting all  $m$  balls of some color, then closing up the gaps between remaining balls, to obtain a coloring of  $n - m$  balls in a row. Theorem 3 can be formulated a little more sharply as follows:

**Theorem 9** *Among all exchangeable partitions  $(\Pi_n)$  of positive integers with EPPF  $p$  subject to (15), the extended two-parameter family is characterized by the following property:*

*if  $\mathcal{B}_{n1}$  denotes the random block of  $\Pi_n$  containing 1, then for each  $1 \leq m < n$ , conditionally given  $\mathcal{B}_{n1}$  with  $|\mathcal{B}_{n1}| = m$ , the partition  $\Pi_n \setminus \mathcal{B}_{n1}$  has the same distribution as  $\Pi'_{n-m}$  for some sequence of partitions  $\Pi'_1, \Pi'_2, \dots$  which does not depend on  $n$ .*

*Moreover, if  $(\Pi_n)$  is an  $(\alpha, \theta)$  partition, then we can take for  $(\Pi'_n)$  the exchangeable  $(\alpha, \theta + \alpha)$  partition of  $\mathbb{N}$ .*

For an arbitrary partition  $\Pi_n$  of  $[n]$  with blocks listed in the order of appearance, define  $J_n$  as the index of the block containing an element chosen from  $[n]$  uniformly at random, independently of  $\Pi_n$ . We call the block  $\mathcal{B}_{nJ_n}$  a *size-biased pick* from the sequence of blocks. Note that this definition agrees with (4) in the sense that the number  $|\mathcal{B}_{nJ_n}|/n$  is a size-biased pick from the numerical sequence  $(|\mathcal{B}_{nj}|/n, j = 1, 2, \dots)$ , because given a sequence of blocks of partition  $\Pi_n$  the value  $J_n = j$  is taken with probability  $|\mathcal{B}_{nj}|/n$ . Assuming  $\Pi_n$  exchangeable, the size of the block  $|\mathcal{B}_{n1}|$  has the same distribution as  $|\mathcal{B}_{nJ_n}|$  conditionally given the ranked sequence of block-sizes, and the reduced partitions  $\Pi_n \setminus \mathcal{B}_{n1}$  and  $\Pi_n \setminus \mathcal{B}_{nJ_n}$  also have the same distributions. The equivalence of Theorem 3 and Theorem 9 is evident from these considerations.

We turn to the proof of Theorem 9. The condition considered in Theorem 9 is just that considered in Lemma 8(iii), so we can work with the equivalent factorization condition (30). We now invoke the symmetry of the EPPF for an exchangeable  $\Pi$ . Suppose that an EPPF  $p$  admits the factorization (30), and re-write the identity (30) in the form

$$p(m, \lambda) = \frac{q(|\lambda| + m : m)}{q(|\lambda| + 1 : 1)} p(1, \lambda).$$

For this expression we must have non-zero denominator, but this is assured by  $\mathbb{P}(0 < W_1 < 1) > 0$ , which is implied by the regularity condition (15). Instead of part  $m$  in  $p(m, \lambda)$ , we have now 1 in  $p(1, \lambda)$ . But  $p$  is symmetric, hence we can iterate, eventually reducing each part to 1.

Let  $\lambda = (\lambda_1, \dots, \lambda_k)$  be a generic composition, and denote  $\Lambda_j = \lambda_j + \dots + \lambda_k$  the tail sums, thus  $\Lambda_1 = |\lambda|$ . Iteration yields

$$p(\lambda) = \frac{q(\Lambda_1 : \lambda_1)}{q(1 + \Lambda_2 : 1)} \frac{q(1 + \Lambda_2 : \lambda_2)}{q(2 + \Lambda_3 : 1)} \dots \frac{q(k - 2 + \Lambda_{k-1} : \lambda_{k-1})}{q(k - 1 + \Lambda_k : 1)} \frac{q(k - 1 + \Lambda_k : \lambda_k)}{q(k : 1)} p(1^k), \quad (32)$$

where  $p(1^k)$  is the probability of the singleton partition of  $[k]$ . This leads to the following lemma, which is a simplification of [21, Lemma 12]:

**Lemma 10** *Suppose that an EPPF  $p$  satisfies the factorization condition (30) and the regularity condition (15). Then*

(i) either

$$q(n : m) = \frac{(a)_{m-1}(b)_{n-m}}{(a+b)_{n-1}}$$

for some  $a, b > 0$ , corresponding to  $W_1$  with beta( $a, b$ ) distribution,

(ii) or

$$q(n : m) = c^{m-1}(1-c)^{n-m}$$

for some  $0 < c < 1$ , corresponding to  $W_1 = c$ , in which case necessarily  $c = 1/M$  for some  $M \geq 3$ .

*Proof.* By symmetry and the assumption that  $p(2, 2, 1) > 0$ , it is easily seen from Kingman's paintbox representation that for each  $m = 1, 2, \dots$  there is some composition  $\mu$  of  $m$  such that

$$p(3, 2, \mu) = p(2, 3, \mu) > 0,$$

where for instance  $(3, 2, \mu)$  means the composition of  $5 + m$  obtained by concatenation of  $(3, 2)$  and  $\mu$ . Indeed, it is clear that one can take either  $\mu = 1^m$  or  $\mu$  to be a single part of size  $m$ , according to whether the probability of at least three non-zero frequencies is zero or greater than zero. Applying (32) for suitable  $k \geq 3$  with  $p(1^k) > 0$ , and cancelling some common factors of the form  $q(n', m')$ , which are all strictly positive because  $p(2, 2, 1) > 0$  implies  $\mathbb{P}(0 < W_1 < 1) > 0$ , we see that for every  $m = 1, 2, \dots$

$$\frac{q(m+5 : 3)q(m+3 : 2)}{q(m+3 : 1)} = \frac{q(m+5 : 2)q(m+4 : 3)}{q(m+4 : 1)}. \quad (33)$$

We have by the addition rule (26)

$$q(m+1 : 2) = q(m : 1) - q(m+1 : 1), \quad q(m+2 : 3) = q(m : 1) - 2q(m+1 : 1) + q(m+2 : 1),$$

and introducing variables  $x_m = q(m : 1)$ ,  $n = m + 2$

$$\frac{(x_{n+1} - 2x_{n+2} + x_{n+3})(x_n - x_{n+1})}{x_{n+1}} = \frac{(x_{n+2} - x_{n+3})(x_n - 2x_{n+1} + x_{n+2})}{x_{n+2}}.$$

The recursion is homogeneous, to pass to inhomogeneous variables divide both sides of the equality by  $x_n$ , then set  $y_n := x_{n+1}/x_n$  and rewrite as

$$(1 - 2y_{n+1} + y_{n+2}y_{n+1})(1 - y_n) = (1 - y_{n+2})(1 - 2y_n + y_n y_{n+1}),$$

which simplifies as

$$-2y_{n+1} + y_{n+1}y_{n+2} + y_n y_{n+1} = -y_n - y_{n+2} + 2y_n y_{n+2}.$$

Finally, use substitution

$$y_n = 1 - \frac{1}{z_n}$$

to arrive at

$$\frac{z_n - 2z_{n+1} + z_{n+2}}{z_n z_{n+1} z_{n+2}} = 0.$$

From this,  $z_n$  is a linear function of  $n$ , which must be nondecreasing to agree with  $0 < y_n < 1$ .

If  $z_n$  is not constant, then going back to  $x_n$ 's we obtain

$$q(n : 1) = c_0 \frac{(b)_{n-1}}{(a+b)_{n-1}}, \quad n \geq 3,$$

for some  $a, b, c_0$ , where the factor  $c_0$  appears since the relation (33) is homogeneous. It is seen from the moments representation

$$q(n : 1) = \int_{[0,1]} (1-x)^{n-1} \mathbb{P}(P_1 \in dx), \quad n \geq 3,$$

that when  $a, b$  are fixed, the factor  $c_0$  is determined from the normalization by choosing a value of  $\mathbb{P}(P_1 = 1)$ . The condition  $p(n) \rightarrow 0$  means that  $\mathbb{P}(P_1 = 1) = 0$ , in which case  $c_0 = 1$  and the distribution of  $P_1$  is beta( $a, b$ ) with some positive  $a, b$ .

If  $(z_n, n \geq 3)$  is a constant sequence, then  $q(n : 1)$  is a geometric progression, and a similar argument shows that the case (ii) prevails. That  $c = 1/M$  for some  $M \geq 3$  is quite obvious: the only way that a size-biased choice of a frequency can be constant is if there are  $M$  equal frequencies for some  $M \geq 1$ . The regularity assumption (15) rules out the cases  $M = 1, 2$ .

□

**Proof of Theorem 9** In the case (i) of Lemma 10, substituting in (32) yields

$$\frac{p(\lambda)}{p(1^k)} = \frac{(a)_{\lambda_1-1} (b)_{\Lambda_2} (a+b)_{\Lambda_2} (a)_{\lambda_2-1} (b)_{\Lambda_3+1} (a+b)_{\Lambda_3+1} \dots (a)_{\lambda_k-1} (b)_{k-1} (a+b)_{k-1}}{(a+b)_{\Lambda_1-1} (b)_{\Lambda_2} (a+b)_{\Lambda_2} (b)_{\Lambda_3+1} \dots (a+b)_{\Lambda_k+k-2} (b)_{k-1}},$$

provided  $p(1^k) > 0$ . After cancellation this becomes

$$\frac{p(\lambda)}{p(1^k)} = \frac{(a+b)_{k-1}}{(a+b)_{n-1}} \prod_{j=1}^k (a)_{\lambda_j-1},$$

where  $n = \Lambda_1 = \lambda_1 + \dots + \lambda_k = |\lambda|$ . Specializing,

$$\frac{p(2, 1^{k-1})}{p(1^k)} = \frac{a}{a+b+k-1}$$

and using the addition rule (20)

$$p(1^k) = kp(2, 1^{k-1}) + p(1^{k+1}),$$

we obtain the recursion

$$\frac{p(1^{k+1})}{p(1^k)} = \frac{a+b+k(1-a)-1}{a+b+k-1}, \quad p(1) = 1.$$

Now (6) follows readily by re-parametrisation  $\theta = a+b-1$ ,  $\alpha = 1-a$ .

The case (ii) of Lemma 10 is even simpler, as it is immediate that  $W_1 = 1/M$  implies that the partition is generated as if by coupon collecting with  $M$  equally frequent coupons.

□

**Proof of Corollary 4** As observed earlier, Corollary 4 characterizing the extended two-parameter family by the condition that

$$W_1 \text{ and } (W_2, W_3, \dots) \text{ are independent} \quad (34)$$

can be read from Theorem 9 and Lemma 8. We find it interesting nonetheless to provide another proof of Corollary 4 based on analysis of the limit frequencies rather than the EPPF. This was in fact the first argument we found, without which we might not have persisted with the algebraic approach of the previous section.

Suppose then that  $W_1, W_2, W_3, \dots$  is the sequence of residual fractions associated with an EPPF  $p$ , and that (34) holds. The symmetry condition  $p(r+1, s+1) = p(s+1, r+1)$  and the moment formula (22) give

$$\mathbb{E}(W_1^r \overline{W_1}^{s+1}) \mathbb{E}(W_2^s) = \mathbb{E}(W_1^s \overline{W_1}^{r+1}) \mathbb{E}(W_2^r) \quad (35)$$

for non-negative integers  $r$  and  $s$ . Setting  $r = 0$ , this expresses moments of  $W_2$  in terms of the moments of  $W_1$ . So the distribution of  $W_1$  determines that of  $W_2$ . Assume now the regularity condition (15). According to Lemma 10 we are reduced either to the case with  $M$  equal frequencies with sum 1, or to the case where  $W_1$  has a beta distribution, and hence so does  $W_2$ , by consideration of (35). There is nothing more to discuss in the first case, so we assume for the rest of this section that

each of  $W_1$  and  $W_2$  has a non-degenerate beta distribution, with possibly different parameters. (36)

Recall that

$$P_1 = W_1 \text{ and } P_2 = (1 - W_1)W_2.$$

As observed in [21],

the conditional distribution of  $(P_3, P_4, \dots)$  given  $P_1$  and  $P_2$  depends symmetrically on  $P_1$  and  $P_2$ .

This can be seen from Kingman's paintbox representation, which implies that conditionally given  $P_1^\downarrow, P_2^\downarrow, \dots$ , as well as  $P_1$  and  $P_2$ , the sequence  $(P_3, P_4, \dots)$  is derived by a process of random sampling from the frequencies  $(P_i^\downarrow)$  with the terms  $P_1$  and  $P_2$  deleted. No matter what  $(P_i^\downarrow)$  this process depends symmetrically on  $P_1$  and  $P_2$ , so the same is true without the extra conditioning on  $(P_i^\downarrow)$ .

Since  $P_1 + P_2$  is a symmetric function of  $P_1$  and  $P_2$ , and  $(W_3, W_4, \dots)$  is a measurable function of  $P_1 + P_2$  and  $(P_3, P_4, \dots)$ ,

the conditional distribution of  $W_3, W_4, \dots$  given  $(P_1, P_2)$  depends symmetrically on  $P_1$  and  $P_2$ .

The condition that  $W_1$  is independent of  $(W_2, W_3, W_4, \dots)$  implies easily that

$W_1$  is conditionally independent of  $(W_3, W_4, \dots)$  given  $W_2$ .

Otherwise put:

$P_1$  is conditionally independent of  $(W_3, W_4, \dots)$  given  $P_2/(1 - P_1)$ ,

hence by the symmetry discussed above

$P_2$  is conditionally independent of  $(W_3, W_4, \dots)$  given  $P_1/(1 - P_2)$ .

Let  $X := P_2/(1 - P_1)$ ,  $Y := P_1/(1 - P_2)$  and  $Z := (W_3, W_4, \dots)$ . Then we have both

$$X \text{ is conditionally independent of } Z \text{ given } Y, \quad (37)$$

and

$$Y \text{ is conditionally independent of } Z \text{ given } X, \quad (38)$$

from which it follows under suitable regularity conditions (see Lemma 11 below) that

$$(X, Y) \text{ is independent of } Z, \quad (39)$$

meaning in the present context that

$$W_1, W_2 \text{ and } (W_3, W_4, \dots) \text{ are independent.} \quad (40)$$

Lauritzen [16, Proposition 3.1] shows that (37) and (38) imply (39) under the assumption that  $(X, Y, Z)$  has a positive and continuous joint density relative to a product measure. From (36) and strict positivity of the beta densities on  $(0, 1)$ , we see that  $(X, Y)$  has a strictly positive and continuous density relative to Lebesgue measure on  $(0, 1)^2$ . We are not in a position to assume that  $(X, Y, Z)$  has a density relative to a product measure. However, the passage from (37) and (38) to (39) is justified by Lemma 11 below without need for a trivariate density. So we deduce that (40) holds. By Lemma 7,  $(W_2, W_3, \dots)$  is the sequence of residual fractions of an exchangeable partition  $\Pi'$ , and  $W_2$  has a beta density. So either  $W_3 = 1$  and we are in the case (12) with  $M = 3$ , or  $W_3$  has a beta density, and the previous argument applies to show that

$$W_1, W_2, W_3 \text{ and } (W_4, W_5, \dots) \text{ are independent.}$$

Continue by induction to conclude the independence of  $W_1, W_2, \dots, W_k$  for all  $k$  such that  $p(1^k) > 0$ .  $\square$

**Lemma 11** *Let  $X, Y$  and  $Z$  denote random variables with values in arbitrary measurable spaces, all defined on a common probability space, such that (37) and (38) hold. If the joint distribution of the pair  $(X, Y)$  has a strictly positive probability density relative to some product probability measure, then (39) holds.*

*Proof.* Let  $p(X, Y)$  be a version of  $\mathbb{P}(Z \in B \mid X, Y)$  for  $B$  a measurable set in the range of  $Z$ . By standard measure theory (e.g. Kallenberg [10, 6.8]) the first conditional independence assumption gives  $\mathbb{P}(Z \in B \mid X, Y) = \mathbb{P}(Z \in B \mid X)$  a.s. so that

$$p(X, Y) = g(X) \text{ a.s. for some measurable function } g.$$

Similarly from the second conditional independence assumption,

$$p(X, Y) = h(Y) \text{ a.s. for some measurable function } h,$$

and we wish to conclude that

$$p(X, Y) = c \text{ a.s. for some constant } c.$$

To complete the argument it suffices to draw this conclusion from the above two assumptions about a jointly measurable function  $p$ , with  $(X, Y)$  the identity map on the product space of pairs  $\mathcal{X} \times \mathcal{Y}$ , and the two almost sure equalities holding with respect to some probability measure  $P$  on this space, with  $P$  having a strictly positive density relative to a product probability measure  $\mu \otimes \nu$ . Fix  $u \in (0, 1)$ , from the previous assumptions it follows that

$$\{p(X, Y) > u\} = \{X \in A_u\} = \{Y \in C_u\} \quad \text{a.s.} \quad (41)$$

for some measurable sets  $A_u, C_u$ , whence

$$\{p(X, Y) > u\} = \{X \in A_u\} \cap \{Y \in C_u\} \quad \text{a.s.}, \quad (42)$$

where the almost sure equalities hold both with respect to the joint distribution  $P$  of  $(X, Y)$ , and with respect to a product probability measure  $\mu \otimes \nu$  governing  $(X, Y)$ . But under  $\mu \otimes \nu$  the random variables  $X$  and  $Y$  are independent. So if  $q := (\mu \otimes \nu)(p(X, Y) > u)$ , then (41) and (42) imply that  $q = q^2$ , so  $q = 0$  or  $q = 1$ . Thus  $p(X, Y)$  is constant a.s. with respect to  $\mu \otimes \nu$ , hence also constant with respect to  $P$ .  $\square$

## 5 The deletion property without the regularity condition

Observe that the property required in Theorem 3 is void if  $\pi_n$  happens to be the one-block partition  $(n)$ . This readily implies that mixing with the trivial one-block partition  $\mathbf{1}$  does not destroy the property. Therefore the  $\mathbf{1}$ -component may be excluded from the consideration, meaning that it is enough to focus on the case

$$P_1 < 1 \text{ a.s.}, \text{ or equivalently } P_1^\downarrow < 1 \text{ a.s.}, \text{ or equivalently } \lim_{n \rightarrow \infty} p(n) = 0. \quad (43)$$

Suppose then that this condition holds, but that the first condition in (15) does not hold, so that  $p(2, 2, 1) = 0$ . Then

$$\mathbb{P}(P_2^\downarrow = 1 - P_1^\downarrow > 0) + \mathbb{P}(P_1^\downarrow < 1, P_2^\downarrow = 0) = 1.$$

If both terms have positive probability then  $\mathbb{P}(W_2 = 1 | W_1 = 0) = 0$  but  $\mathbb{P}(W_2 = 1 | W_1 > 0) > 0$ , so the independence of  $W_1$  and  $W_2$  fails. Thus the independence forces either  $\mathbb{P}(P_2^\downarrow = 1 - P_1^\downarrow > 0) = 1$  or  $\mathbb{P}(P_1^\downarrow < 1, P_2^\downarrow = 0) = 1$ . The two cases are readily treated:

- (i) If  $\mathbb{P}(P_2^\downarrow = 1 - P_1^\downarrow > 0) = 1$  then  $W_2 = 1$  a.s. and the independence trivially holds. This is the case when  $\Pi$  has two blocks almost surely.
- (ii) If  $\mathbb{P}(P_1^\downarrow < 1, P_2^\downarrow = 0) = 1$  and  $\mathbb{P}(P_1^\downarrow > 0) > 0$  then  $\mathbb{P}(W_2 > 0 | W_1 > 0) = 0$  but  $\mathbb{P}(W_2 > 0 | W_1 = 0) > 0$ , hence  $W_1$  and  $W_2$  are not independent. Therefore  $\mathbb{P}(P_1^\downarrow < 1, P_2^\downarrow = 0) = 1$  and the independence imply  $P_1^\downarrow = 0$  a.s., meaning that  $\Pi = \mathbf{0}$ .

We conclude that the most general exchangeable partition  $\Pi$  which has the property in Theorem 9 is a two-component mixture, in which the first component is either a partition from the extended two-parameter family, or a two-block partition as in (i) above, or  $\mathbf{0}$ , and the second component is the trivial partition  $\mathbf{1}$ .

## 6 Regeneration and $\tau$ -deletion

In this section we partly survey and partly extend the results from [5, 6] concerning characterizations of  $(\alpha, \theta)$  partitions by regeneration properties. As in Kingman's study of the regenerative processes [12], subordinators (increasing Lévy processes) appear naturally in our framework of *multiplicative* regenerative phenomena. Following [6], we call a partition structure  $(\pi_n)$  *regenerative* if

for each  $n$  it is possible to delete a randomly chosen part of  $\pi_n$  in such a way that for each  $0 < m < n$ , given the deleted part is of size  $m$ , the remaining parts form a partition of  $n - m$  with the same distribution as  $\pi_{n-m}$ .

In terms of an exchangeable partition  $\Pi = (\Pi_n)$  of  $\mathbb{N}$ , the associated partition structure  $(\pi_n)$  is regenerative if and only if

for each  $n$  it is possible to select a random block  $\mathcal{B}_{nJ_n}$  of  $\Pi_n$  in such a way that for each  $0 < m < n$ , conditionally given that  $|\mathcal{B}_{nJ_n}| = m$  the partition  $\Pi_n \setminus \mathcal{B}_{nJ_n}$  of  $[n - m]$  is distributed according to the unconditional distribution of  $\Pi_{n-m}$ :

$$(\Pi_n \setminus \mathcal{B}_{nJ_n} \text{ given } |\mathcal{B}_{nJ_n}| = m) \stackrel{d}{=} \Pi_{n-m} \quad (44)$$

where  $\Pi_n \setminus \mathcal{B}_{nJ_n}$  is defined as in the discussion preceding Theorem 9. Moreover, there is no loss of generality in supposing further that the conditional distribution of  $J_n$  given  $\Pi_n$  is of the form

$$\mathbb{P}(J_n = j \mid \Pi_n = \{B_1, \dots, B_k\}) = d(|B_1|, \dots, |B_k|; j) \quad (45)$$

for some *symmetric deletion kernel*  $d$ , meaning a non-negative function of a composition  $\lambda$  of  $n$  and  $1 \leq j \leq k_\lambda$  such that

$$d(\lambda_1, \lambda_2, \dots, \lambda_k; j) = d(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \dots, \lambda_{\sigma(k)}; 1) \quad (46)$$

for every permutation  $\sigma$  of  $[k]$  with  $\sigma(1) = j$ . To determine a symmetric deletion kernel, it suffices to specify  $d(\lambda; 1)$ , which is the conditional probability, given blocks of sizes  $\lambda_1, \lambda_2, \dots, \lambda_k$ , of picking the first of these blocks. This is a non-negative symmetric function of  $(\lambda_2, \dots, \lambda_k)$ , subject to the further constraint that its extension to arguments  $j \neq 1$  via (46) satisfies

$$\sum_{j=1}^{k_\lambda} d(\lambda; j) = 1$$

for every composition  $\lambda$  of  $n$ . The regeneration condition can now be reformulated in terms of the EPPF  $p$  of  $\Pi$  in a manner similar to (30):

**Lemma 12** *An exchangeable random partition  $\Pi$  with EPPF  $p$  is regenerative if and only if there exists a symmetric deletion kernel  $d$  such that*

$$p(\lambda)d(\lambda; 1) = q(n, \lambda_1) \binom{n}{\lambda_1}^{-1} p(\lambda_2, \dots, \lambda_k) \quad (47)$$

for every composition  $\lambda$  of  $n$  into at least two parts and some non-negative function  $q$ . Then

$$q(n, m) = \mathbb{P}(|\mathcal{B}_{n, J_n}| = m) \quad (m \in [n]) \quad (48)$$

for  $J_n$  as in (45).

*Proof.* Formula (47) offers two different ways of computing the probability of the event that  $\Pi_n = \{B_1, \dots, B_k\}$  and  $J_n = 1$  for an arbitrary partition  $\{B_1, \dots, B_k\}$  of  $[n]$  with  $|B_i| = \lambda_i$  for  $i \in [k]$ : on the left side, by definition of the symmetric deletion kernel, and on the right side by conditioning on the event  $\mathcal{B}_{n, J_n} = B_1$  and appealing to the regeneration property and exchangeability.  $\square$

Consider now the question of whether an  $(\alpha, \theta)$  partition with EPPF  $p = p_{\alpha, \theta}$  as in (6) is regenerative with respect to some deletion kernel. By the previous lemma and cancellation of common factors, the question is whether there exists a symmetric deletion kernel  $d(\lambda; j)$  such that the value of

$$q(n, \lambda_1) = d(\lambda; 1) \binom{n}{\lambda_1} \frac{(1 - \alpha)_{\lambda_1 - 1} (\theta + (k - 1)\alpha)}{(\theta + n - \lambda_1)_{\lambda_1}} \quad (49)$$

is the same for all compositions  $\lambda$  of  $n$  with  $k$  parts and a prescribed value of  $\lambda_1$ . But it is easily checked that the formula

$$d_{\alpha, \theta}(\lambda; j) = \frac{\theta \lambda_j + \alpha(n - \lambda_j)}{n(\theta + \alpha(k - 1))} \quad (50)$$

provides just such a symmetric deletion kernel. Note that the kernel depends on  $(\alpha, \theta)$  only through the ratio  $\tau := \alpha/(\alpha + \theta)$ , and that the kernel is non-negative for all compositions  $\lambda$  only if both  $\alpha$  and  $\theta$  are non-negative.

To provide a more general context for this and later discussions, let  $(x_1, \dots, x_k)$  be a fixed sequence of positive numbers with sum  $s = \sum_{j=1}^k x_j$ . For a fixed parameter  $\tau \in [0, 1]$ , define a random variable  $T$  with values in  $[k]$  by

$$\mathbb{P}(T = j \mid (x_1, \dots, x_k)) = \frac{(1 - \tau)x_j + \tau(s - x_j)}{s(1 - \tau + \tau(k - 1))}, \quad (51)$$

The random variable  $x_T$  is called a  $\tau$ -biased pick from  $x_1, \dots, x_k$ . The law of  $x_T$  does not depend on the order of the sequence  $(x_1, \dots, x_k)$ , and there is also a scaling invariance:  $s^{-1}x_T$  is a  $\tau$ -biased pick from  $(s^{-1}x_1, \dots, s^{-1}x_k)$ . Note that a 0-biased pick is a size-biased pick from  $(x_1, \dots, x_k)$ , choosing any particular element with probability proportional to its size. A 1/2-biased pick is a uniform random choice from the list, as (51) then equals  $1/k$  for all  $j$ . And a 1-biased pick may be called a co-size biased pick, as it chooses  $j$  with probability proportional to its co-size  $s - x_j$ .

These definitions are now applied to the sequence of block sizes  $x_j$  of the restriction to  $[n]$  of an exchangeable partition  $\Pi$  of  $\mathbb{N}$ . We denote by  $T_n$  a random variable whose conditional distribution given  $\Pi_n$  with  $k$  blocks and  $|\mathcal{B}_{nj}| = x_j$  for  $j \in [k]$  is defined by (51), and denote by  $\mathcal{B}_{nT_n}$  the  $\tau$ -biased pick from the sequence of blocks of  $\Pi_n$ . We call  $\Pi$   $\tau$ -regenerative if  $\Pi_n$  is regenerative with respect to deletion of the  $\tau$ -biased pick  $\mathcal{B}_{nT_n}$ .

**Theorem 13** [5, 6] *For each  $\tau \in [0, 1]$ , apart from the constant partitions  $\mathbf{0}$  and  $\mathbf{1}$ , the only exchangeable partitions of  $\mathbb{N}$  that are  $\tau$ -regenerative are the members of the two parameter family with parameters in the range*

$$\{(\alpha, \theta) \in [0, 1] \times [0, \infty] : \alpha/(\alpha + \theta) = \tau\}.$$

*Explicitly, the distribution of the  $\tau$ -biased pick for such  $(\alpha, \theta)$  partitions of  $[n]$  is*

$$\mathbb{P}(|\mathcal{B}_{nT_n}| = m) = \binom{n}{m} \frac{(1 - \alpha)_{m-1}}{(\theta + n - m)_m} \frac{(n - m)\alpha + m\theta}{n}, \quad m \in [n]. \quad (52)$$

*Proof.* The preceding discussion around (49) and (50) shows that members of the two parameter family with parameters in the indicated range are  $\tau$ -regenerative, and gives the formula (52) for the decrement matrix. See [6] for the proof that these are the only non-degenerate exchangeable partitions of  $\mathbb{N}$  that are  $\tau$ -regenerative.  $\square$

In particular, each  $(\alpha, \alpha)$  partition is 1/2-regenerative, meaning regenerative with respect to deletion of a block chosen uniformly at random. The constant partitions  $\mathbf{0}$  and  $\mathbf{1}$  are obviously  $\tau$  regenerative for every  $\tau \in [0, 1]$ . This is consistent with the characterization above because the  $(1, \theta)$  partition is the  $\mathbf{0}$  partition for every  $\theta \geq 0$ , and because the partition  $\mathbf{1}$  can be reached as a limit of  $(\alpha, \theta)$  partitions as  $\alpha, \theta \downarrow 0$  with  $\alpha(\alpha + \theta)^{-1}$  held fixed.

**Multiplicative regeneration** By Corollary 4, if  $(P_i)$  is the sequence of limit frequencies for a  $(0, \theta)$  partition for some  $\theta > 0$  and if the first limit frequency  $P_1$  is deleted and the other frequencies renormalized to sum to 1, then the resulting sequence  $(Q_j)$  is independent of  $P_1$  and has the same distribution as  $(P_j)$ . Because  $P_1$  is a size-biased pick from the sequence  $(P_i)$ , this regenerative property of the frequencies  $(P_i)$  can be seen as an analogue of the 0-regeneration property of the  $(0, \theta)$  partitions.

If  $(P_i)$  is instead the sequence of limit frequencies of an  $(\alpha, \theta)$  partition  $\Pi$  for parameters satisfying  $0 < \alpha < 1$ ,  $\alpha/(\alpha + \theta) = \tau$ , a question arises: does the regenerative property of  $\Pi_n$  with respect to a  $\tau$ -biased pick have an analogue in terms of a  $\tau$ -biased pick from the frequencies  $(P_i)$ ? This cannot be answered straightforwardly as in the  $\tau = 0$  case, because when  $\tau > 0$  the formula (51) defines a proper probability distribution only for series  $(x_j)$  with some finite number  $k$  of positive terms. For instance, in the case  $\tau = 1/2$  there is no such analogue of (51) as ‘uniform random choice’ from infinitely many terms.

Still, Ewens’ case provides a clue if we turn to a *bulk deletion*. Let  $P_J$  be a size-biased pick from the frequencies  $(P_j)$ , as defined by (4), and let  $(Q_j)$  be a sequence obtained from  $(P_j)$  by deleting all  $P_1, \dots, P_J$  and renormalizing. Then  $(Q_j)$  is independent of

$P_1, \dots, P_J$ , and  $(Q_j) \stackrel{d}{=} (P_j)$ . The latter assertion follows from the i.i.d. property of the residual fractions and by noting that (4) is identical with

$$\mathbb{P}(J = j \mid (W_i, i \in \mathbb{N})) = W_j \prod_{i=1}^{j-1} (1 - W_i).$$

A similar bulk deletion property holds for partitions in the Ewens' family, in the form:

$$(\Pi_n \setminus (\mathcal{B}_{n_1} \cup \dots \cup \mathcal{B}_{n_{J_n}}) \text{ given } |\mathcal{B}_1 \cup \dots \cup \mathcal{B}_{n_{J_n}}| = m) \stackrel{d}{=} \Pi_{n-m}$$

for all  $1 \leq m \leq n$ , where  $B_{n_{J_n}}$  is a size-biased pick from the blocks.

To make the ansatz of bulk deletion work for  $\tau \neq 0$  it is necessary to arrange the frequencies in a more complex manner. To start with, we modify the paintbox construction. Let  $\mathcal{U} \subset [0, 1]$  be a random open set canonically represented as the union of its disjoint open component intervals. We suppose that the Lebesgue measure of  $\mathcal{U}$ , equal to the sum of lengths of the components, is 1 almost surely. We associate with  $\mathcal{U}$  an exchangeable partition  $\Pi$  exactly as in Kingman's representation in Theorem 2. For each component interval  $G \subset \mathcal{U}$  there is an index  $i_G := \min\{n : U_n \in G\}$  that is the minimal index of a sequence  $(U_i)$  of iid uniform[0,1] points hitting the interval, and for all  $j$ ,  $P_j$  is the length of the  $j$ th component interval when the intervals are listed in order of increasing minimal indices. So  $(P_j)$  is a size-biased permutation of the lengths of interval components of  $\mathcal{U}$ .

Let  $\triangleleft$  be the linear order on  $\mathbb{N}$  induced by the *interval order* of the components of  $\mathcal{U}$ , so  $j \triangleleft k$  iff the interval of length  $P_j$ , which is the home interval of the  $j$ th block  $\mathcal{B}_j$  to appear in the process of uniform random sampling of intervals, lies to the left of the interval of length  $P_k$  associated with block  $\mathcal{B}_k$ . A convergence argument shows that  $\mathcal{U}$  is uniquely determined by  $(P_j)$  and  $\triangleleft$ . In loose terms,  $\mathcal{U}$  is an arrangement of a sequence of tiles of sizes  $P_j$  in the order on indices  $j$  prescribed by  $\triangleleft$ , and this arrangement is constructable by sequentially placing the tile  $j$  in the position prescribed by the order  $\triangleleft$  restricted to  $[j]$ .

For  $x \in [0, 1)$  let  $(a_x, b_x) \subset \mathcal{U}$  be the component interval containing  $x$ . Define  $\mathcal{V}_x$  as the open set obtained by deleting the bulk of component intervals to the left of  $b_x$ , then linearly rescaling the remaining set  $\mathcal{U} \cap [b_x, 1]$  to  $[0, 1]$ . We say that  $\mathcal{U}$  is *multiplicatively regenerative* if for each  $x \in [0, 1)$ ,  $\mathcal{V}_x$  is independent of  $\mathcal{U} \cap [0, b_x]$  and  $\mathcal{V}_x \stackrel{d}{=} \mathcal{U}$ .

An ordered version of the paintbox correspondence yields:

**Theorem 14** [5, 6] *An exchangeable partition  $\Pi$  is regenerative if and only if it has a paintbox representation in terms of some multiplicatively regenerative set  $\mathcal{U}$ . The deletion operation is then defined by classifying  $n$  independent uniform points from  $[0, 1]$  according to the intervals of  $\mathcal{U}$  into which they fall, and deleting the block of points in the leftmost occupied interval.*

A property of the frequencies  $(P_j)$  of an exchangeable regenerative partition  $\Pi$  of  $\mathbb{N}$  now emerges: there exists a strict total order  $\triangleleft$  on  $\mathbb{N}$ , which is a random order, which has some joint distribution with  $(P_j)$  such that arranging the intervals of sizes  $(P_j)$  in order  $\triangleleft$  yields a multiplicatively regenerative set  $\mathcal{U}$ . Equivalently, there exists a multiplicatively

regenerative set  $\mathcal{U}$  that induces a partition with frequencies  $(P_j)$  and an associated order  $\triangleleft$ . This set  $\mathcal{U}$  is then necessarily unique in distribution as a random element of the space of open subsets of  $[0, 1]$  equipped with the Hausdorff metric [5] on the complementary closed subsets. A subtle point here is that the joint distribution of  $(P_j)$  and  $\triangleleft$  is not unique, and neither is the joint distribution of  $(P_j)$  and  $\mathcal{U}$ , unless further conditions are imposed. For instance, one way to generate  $\triangleleft$  is to suppose that the  $(P_j)$  are generated by a process of uniform random from  $\mathcal{U}$ . But for a  $(0, \theta)$  partition, we know that another way is to construct  $\mathcal{U}$  from  $(P_j)$  by simply placing the intervals in deterministic order  $P_1, P_2, \dots$  from left to right. In the construction by uniform random sampling from  $\mathcal{U}$  the interval of length  $P_1$  discovered by the first sample point need not be the leftmost, and need not lie to the left of the second discovered interval  $P_2$ .

In [5] we showed that the multiplicative regeneration of  $\mathcal{U}$  follows from an apparently weaker property: if  $(a_U, b_U)$  is the component interval of  $\mathcal{U}$  containing an uniform  $[0, 1]$  sample  $U$  independent of  $\mathcal{U}$ , and if  $\mathcal{V}$  is defined as the open set obtained by deleting the component intervals to the left of  $b_U$  and linearly rescaling the remaining set  $\mathcal{U} \cap [b_U, 1]$  to  $[0, 1]$ , then given  $b_U < 1$ ,  $\mathcal{V}$  is independent of  $b_U$  (hence, as we proved, independent of  $\mathcal{U} \cap [0, b_U]$  too!) and has distribution equal to the unconditional distribution of  $\mathcal{U}$ . This independence is the desired analogue for more general regenerative partitions of the bulk-deletion property of Ewens' partitions.

The fundamental representation of multiplicatively regenerative sets involves a random process  $F_t$  known in statistics as a neutral-to-the right distribution function.

**Theorem 15** [5] *A random open set  $\mathcal{U}$  of Lebesgue measure 1 is multiplicatively regenerative if and only if there exists a drift-free subordinator  $S = (S_t, t \geq 0)$  with  $S_0 = 0$  such that  $\mathcal{U}$  is the complement to the closed range of the process  $F_t = 1 - \exp(-S_t)$ ,  $t \geq 0$ . The Lévy measure of  $S$  is determined uniquely up to a positive factor.*

According to Theorems 14 and 15, regenerative partition structures with proper frequencies are parameterised by a measure  $\tilde{\nu}(du)$  on  $(0, 1]$  with finite first moment, which is the image via the transformation from  $s$  to  $1 - \exp(-s)$  of the Lévy measure  $\nu(ds)$  on  $(0, \infty]$  associated with the subordinator  $S$ . The Laplace exponent  $\Phi$  of the subordinator, defined by the Lévy-Khintchine formula

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

determines the Lévy measure  $\nu(ds)$  on  $(0, \infty]$  and its image  $\tilde{\nu}(du)$  on  $(0, 1]$  via the formulae

$$\Phi(a) = \int_{(0, \infty]} (1 - e^{-ax})\nu(dx) = \int_{]0, 1]} (1 - (1 - x)^a)\tilde{\nu}(dx).$$

As shown in [5], the decrement matrix  $q$  of the regenerative partition structure, as in (48), is then

$$q(n, m) = \frac{\Phi(n, m)}{\Phi(n)}, \quad 1 \leq m \leq n, \quad n = 1, 2, \dots$$

where

$$\Phi(n, m) = \binom{n}{m} \int_{]0, 1]} x^m (1 - x)^{n-m} \tilde{\nu}(dx).$$

Uniqueness of the parameterisation is achieved by a normalisation condition, such as  $\Phi(1) = 1$ .

In [5] the subordinator  $S^{\alpha, \theta}$  which produces  $\mathcal{U}$  as in Theorem 15 for the  $(\alpha, \theta)$  partition was identified by the following formula for the right tail of its Lévy measure:

$$\nu(x, \infty] = (1 - e^{-x})^{-\alpha} e^{-x\theta}, \quad x > 0. \quad (53)$$

The subordinator  $S^{(0, \theta)}$  is a compound Poisson process whose jumps are exponentially distributed with rate  $\theta$ . For  $\theta = 0$  the Lévy measure has a unit mass at  $\infty$ , so the subordinator  $S^{(\alpha, 0)}$  is killed at unit rate. The  $S^{(\alpha, \alpha)}$  subordinator belongs to the class of Lamperti-stable processes recently studied in [4]. For positive parameters the subordinator  $S^{(\alpha, \theta)}$  can be constructed from the  $(0, \theta)$  and  $(\alpha, 0)$  cases, as follows. First split  $\mathbb{R}_+$  by the range of  $S^{(0, \theta)}$ , that is at points  $E_1 < E_2 < \dots$  of a Poisson process with rate  $\theta$ . Then run an independent copy of  $S^{(\alpha, 0)}$  up to the moment the process crosses  $E_1$  at some random time, say  $t_1$ . The level-overshooting value is neglected and the process is stopped. At the same time  $t_1$  a new independent copy of  $S^{(\alpha, 0)}$  is started at value  $E_1$  and run until crossing  $E_2$  at some random time  $t_2$ , and so on.

In terms of  $F_t = 1 - \exp(-S_t)$ , the range of the process in the  $(0, \theta)$  case is a stick-breaking set  $\{1 - \prod_{i=1}^{j-1} (1 - V_i), i = 0, 1, \dots\}$  with i.i.d. beta(1,  $\theta$ ) factors  $V_i$ . In the case  $(\alpha, 0)$  the range of  $(F_t)$  is the intersection of  $[0, 1]$  with the  $\alpha$ -stable set (the range of  $\alpha$ -stable subordinator). In other cases  $\mathcal{U}$  is constructable as a cross-breed of the cases  $(\theta, 0)$  and  $(0, \alpha)$ : first  $[0, 1]$  is partitioned in subintervals by the beta(1,  $\theta$ ) stick-breaking, then each subinterval  $(a, b)$  of this partition is further split by independent copy of the multiplicatively regenerative  $(\alpha, 0)$  set, shifted to start at  $a$  and truncated at  $b$ .

**Constructing the order** Following [6, 25], we shall describe an arrangement which allows us to pass from  $(\alpha, \theta)$  frequencies  $(P_j)$  to the multiplicatively regenerative set associated with the subordinator  $S^{(\alpha, \alpha)}$ . The connection between size-biased permutation with  $\tau$ -deletion (Lemma 17) is new.

A linear order  $\triangleleft$  on  $\mathbb{N}$  is conveniently described by a sequence of the initial ranks  $(\rho_j) \in [1] \times [2] \times \dots$ , with  $\rho_j = i$  if and only if  $j$  is ranked  $i$ th smallest in the order  $\triangleleft$  among the integers  $1, \dots, j$ . For instance, the initial ranks  $1, 2, 1, 3, \dots$  appear when  $3 \triangleleft 1 \triangleleft 4 \triangleleft 2$ .

For  $\xi \in [0, \infty]$  define a random order  $\triangleleft_\xi$  on  $\mathbb{N}$  by assuming that the initial ranks  $\rho_k, k \in \mathbb{N}$ , are independent, with distribution

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

The edge cases  $\xi = 0, \infty$  are defined by continuity. The order  $\triangleleft_1$  is a ‘uniformly random order’, in the sense that restricting to  $[n]$  we have all  $n!$  orders equally likely, for every  $n$ . The order  $\triangleleft_\infty$  coincides with the standard order  $<$  almost surely. For every permutation  $i_1, \dots, i_n$  of  $[n]$ , we have

$$\mathbb{P}(i_1 \triangleleft_\xi \dots \triangleleft_\xi i_n) = \frac{\xi^n}{\xi(\xi + 1) \dots (\xi + n - 1)}$$

where  $r$  is the number of upper records in the permutation. See [8] for this and more general permutations with tilted record statistics.

**Theorem 16** [25, Corollary 7] *For  $0 \leq \alpha < 1, \theta \geq 0$  the arrangement of GEM( $\alpha, \theta$ ) frequencies  $(P_j)$  represented as open intervals in an independent random order  $\triangleleft_{\theta/\alpha}$  is a multiplicatively regenerative open set  $\mathcal{U} \subset [0, 1]$ , where  $\mathcal{U}$  is representable as the complement of the closed range of the process  $F_t = 1 - \exp(-S_t), t \geq 0$ , for  $S$  the subordinator with Lévy measure (53).*

This result was presented without proof as [25, Corollary 7], in a context where the regenerative ordering of frequencies was motivated by an application to a tree growth process. Here we offer a proof which exposes the combinatorial structure of the composition of size-biased permutation and a  $\triangleleft_{\theta/\alpha}$  ordering of frequencies.

For a sequence of positive reals  $(x_1, \dots, x_k)$ , define the  $\tau$ -biased permutation of this sequence, denoted  $\text{perm}_\tau(x_1, \dots, x_k)$ , by iterating a single  $\tau$ -biased pick, as follows. A number  $x_T$  is chosen from  $x_1, \dots, x_k$  without replacement, with  $T$  distributed on  $[k]$  according to (51), 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 instance  $\text{perm}_0$  is the size-biased permutation, which is defined more widely for finite or infinite summable sequences  $(x_1, x_2, \dots)$ , and shuffles them in the same way as it shuffles  $(s^{-1}x_1, s^{-1}x_2, \dots)$  where  $s = \sum_j x_j$ . Denote by  $\triangleleft_\xi(x_1, \dots, x_k)$  the arrangement of  $x_1, \dots, x_k$  in succession according to the  $\triangleleft_\xi$ -order on  $[k]$ .

**Lemma 17** *For  $\xi = (1 - \tau)/\tau$  there is the compositional formula*

$$\text{perm}_\tau(x_1, \dots, x_k) \stackrel{d}{=} \triangleleft_\xi(\text{perm}_0(x_1, \dots, x_k)), \quad (54)$$

where on the right-hand side  $\triangleleft_\xi$  and  $\text{perm}_0$  are independent.

*Proof.* On each side of this identity, the distribution of the random permutation remains the same if the sequence  $x_1, \dots, x_k$  is permuted. So it suffices to check that each scheme returns the identity permutation with the same probability. If on the right hand side we set

$$\text{perm}_0(x_1, \dots, x_k) = (x_{\sigma(1)}, \dots, x_{\sigma(k)})$$

then the right hand scheme generates the identity permutation with probability

$$\frac{\mathbb{E}\xi^R}{\xi(\xi + 1) \cdots (\xi + k - 1)} \quad (55)$$

where  $R$  is the number of upper records in the sequence of ranks which generated  $\sigma^{-1}$ , which equals the number of upper records in  $\sigma$ . Now  $R = \sum_{j=1}^k X_j$  where  $X_j$  is the indicator of the event  $A_j$  that  $j$  is an upper record level for  $\sigma$ , meaning that there is some  $1 \leq i \leq n$  such that

$$\sigma(i') < j \text{ for all } i' < i \text{ and } \sigma(i) = j.$$

Equivalently,  $A_j$  is the event that

$$\sigma^{-1}(j) < \sigma^{-1}(\ell) \text{ for each } j < \ell \leq k.$$

Or again, assuming for simplicity that the  $x_i$  are all distinct, which involves no loss of generality, because the probability in question depends continuously on  $(x_1, \dots, x_k)$ ,  $A_j$  is the event that  $x_j$  precedes  $x_\ell$  in the permutation  $(x_{\sigma(1)}, \dots, x_{\sigma(k)})$  for each  $j < \ell \leq k$ . Now it is easily shown that  $(x_{\sigma(1)}, \dots, x_{\sigma(k)})$  with  $x_1$  deleted is a size-biased permutation of  $(x_2, \dots, x_k)$ , and that the same is true conditionally given  $A_1$ . It follows by induction that the events  $A_j$  are mutually independent, with

$$\mathbb{P}(A_j) = x_j / (x_j + \dots + x_k) \text{ for } 1 \leq j \leq k.$$

This allows the probability in (55) to be evaluated as

$$\prod_{j=1}^k \frac{(\xi x_j + x_{j+1} + \dots + x_k)}{(x_j + x_{j+1} + \dots + x_k)(\xi + j - 1)}$$

This is evidently the probability that  $\text{perm}_\tau(x_1, \dots, x_k)$  generates the identity permutation, and the conclusion follows.  $\square$

The  $\tau$ -biased arrangement cannot be defined for infinite positive summable sequence  $(x_1, x_2, \dots)$ , since the ‘ $k = \infty$ ’ instance of (51) is not a proper distribution for  $\tau \neq 0$ . But the right-hand side of (54) is well-defined as arrangement of  $x_1, x_2, \dots$  in some total order, hence the composition  $\triangleleft_\xi \circ \text{perm}_0$  is the natural extension of the  $\tau$ -biased arrangement to infinite series.

*Proof of Theorem 16.* We represent a finite or infinite positive sequence  $(x_j)$  whose sum is 1 as an open subset of  $[0, 1]$  composed of contiguous intervals of sizes  $x_j$ . The space of open subsets of  $[0, 1]$  is endowed with the Hausdorff distance on the complementary compact sets. This topology is weaker than the product topology on positive series summable to 1. The limits below are understood as  $n \rightarrow \infty$ .

We know by a version of Kingman’s correspondence [20] that  $(|\mathcal{B}_{n_j}|/n, j \geq 1) \rightarrow (P_j)$  a.s. in the product topology. This readily implies  $\triangleleft_\xi(|\mathcal{B}_{n_j}|/n, j \geq 1) \rightarrow \triangleleft_\xi(P_j)$  a.s. in the Hausdorff topology, by looking at the  $M$  first terms for  $M$  such that these terms sum to at least  $1 - \epsilon$  with probability at least  $1 - \epsilon$ , then sending  $\epsilon \rightarrow 0$  and  $M \rightarrow \infty$ . In [5] we showed that  $\text{perm}_\tau(|\mathcal{B}_{n_j}|, j \geq 1) \rightarrow \mathcal{U}$  a.s. in the Hausdorff topology. (Here the definition of the  $\text{perm}_\tau$  is coupled with  $(|\mathcal{B}_{n_j}|, j \geq 1)$  by putting these blocks in the order determined by uniform sampling from  $\mathcal{U}$ ). The missing link is provided by Lemma 17, from which we obtain

$$\text{perm}_\tau(|\mathcal{B}_{n_j}|, j \geq 1) \stackrel{d}{=} \triangleleft_\xi(|\mathcal{B}_{n_j}|, j \geq 1),$$

with the  $\tau$ -biased permutation  $\text{perm}_\tau$  applied to the *finite* sequence of positive block-sizes  $(|\mathcal{B}_{n_j}|, j \geq 1)$ . Putting things together we conclude that  $\triangleleft_\xi(P_j, j \geq 1) \stackrel{d}{=} \mathcal{U}$ .  $\square$

In three special cases, already identified in the previous work [6], the arrangement of  $\text{PD}(\alpha, \theta)$  (or  $\text{GEM}(\alpha, \theta)$ ) frequencies in a multiplicatively regenerative set has a simpler

description: in the  $(0, \theta)$  case the frequencies are placed in the size-biased order; in the  $(\alpha, \alpha)$  case the frequencies are ‘uniformly randomly shuffled’; and in the  $(\alpha, 0)$  case a size-biased pick is placed contiguously to 1, while the other frequencies are ‘uniformly randomly shuffled’. The latter is an infinite analogue of the co-size biased arrangement  $\text{perm}_1$ .

We refer to [9, 25] for further recent developments related to ordered  $(\alpha, \theta)$  partitions and their regenerative properties.

## References

- [1] J. Bertoin, *Random fragmentation and coagulation processes*, volume 102 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2006.
- [2] J. Bertoin, Two-parameter Poisson-Dirichlet measures and reversible exchangeable fragmentation-coalescence processes. *Combin. Probab. Comput.*, 17(3):329–337, 2008.
- [3] K. Bobecka and J. Wesolowski, The Dirichlet distribution and process through neutralities, *J. Theor. Probab.* 20: 295–308, 2007.
- [4] M.E. Caballero, J.C. Pardo, and J.L. Pérez, On the Lamperti stable processes, 2008 <http://arxiv.org/abs/0802.0851>
- [5] A. Gnedin and J. Pitman, Regenerative composition structures, *Ann. Probab.*, 33 445–479, 2005.
- [6] A. Gnedin and J. Pitman, Regenerative partition structures. *The Electronic Journal of Combinatorics*, 11(2) R12:1–21, 2004.
- [7] A. Gnedin, On convergence and extensions of size-biased permutations. *J. Appl. Probab.*, 35(3):642–650, 1998.
- [8] A. Gnedin, Coherent random permutations with record statistics, *2007 Conference on Analysis of Algorithms, AofA 07 DMTCS Proc.* **AH** 147-158, <http://front.math.ucdavis.edu/0611.5538>
- [9] B. Haas, J. Pitman and M. Winkel, Spinal partitions and invariance under re-rooting of continuum random trees. *Ann. Probab.*, 36(5):1790–1837, 2008.
- [10] O. Kallenberg, *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.
- [11] S. Kerov, Coherent random allocations, and the Ewens-Pitman formula. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 325(Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 12):127–145, 246, 2005.
- [12] J. F. C. Kingman, *Regenerative Phenomena*, Wiley, 1972.

- [13] J. F. C. Kingman, Random partitions in population genetics. *Proc. Roy. Soc. London Ser. A*, 361(1704):1–20, 1978.
- [14] J. F. C. Kingman, The representation of partition structures. *J. London Math. Soc. (2)*, 18(2):374–380, 1978.
- [15] J. F. C. Kingman, The coalescent. *Stochastic Process. Appl.*, 13(3):235–248, 1982.
- [16] S. L. Lauritzen, *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.
- [17] M. Perman, J. Pitman and M. Yor, Size-biased sampling of Poisson point processes and excursions, *Probab. Th. Rel. Fields*, 92:21–39, 1992.
- [18] J. Pitman, Poisson-Kingman partitions. In D. R. Goldstein, editor, *Science and Statistics: A Festschrift for Terry Speed*, volume 30 of *Lecture Notes – Monograph Series*, pages 1–34. Institute of Mathematical Statistics, Beachwood, OH, 2003.
- [19] J. Pitman, *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard.
- [20] J. Pitman, Exchangeable and partially exchangeable random partitions. *Probab. Th. Rel. Fields*, 102:145–158, 1995.
- [21] J. Pitman, Random discrete distributions invariant under size-biased permutation. *Adv. Appl. Prob.*, 28:525–539, 1996.
- [22] J. Pitman, Some developments of the Blackwell-MacQueen urn scheme. In T.S. Ferguson et al., editor, *Statistics, Probability and Game Theory; Papers in honor of David Blackwell*, volume 30 of *Lecture Notes-Monograph Series*, pages 245–267. Institute of Mathematical Statistics, Hayward, California, 1996.
- [23] J. Pitman, Coalescents with multiple collisions. *Ann. Probab.*, 27:1870–1902, 1999.
- [24] J. Pitman and M. Yor, The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, 25:855–900, 1997.
- [25] J. Pitman and M. Winkel, Regenerative tree growth: binary self-similar continuum random trees and Poisson-Dirichlet compositions (2008) <http://front.math.ucdavis.edu/0803.3098>
- [26] Y. W. Teh, A hierarchical Bayesian language model based on Pitman-Yor processes. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 985–992, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

- [27] F. Wood, C. Archambeau, J. Gasthaus, L. James, and Y.W. Teh, A stochastic memoizer for sequence data. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1129–1136, New York, NY, USA, 2009. ACM.