DIRICHLET-KINGMAN PARTITION REVISITED

Thierry Huillet, Servet Martinez

DIRICHLET-KINGMAN PARTITION REVISITED

Thierry HUILLET [1], Servet MARTINEZ [2]
[1] Laboratoire de Physique Théorique et Modélisation,
CNRS-UMR 8089 et Université de Cergy-Pontoise,
Site de Saint Martin, 2 avenue Adolphe-Chauvin,
95032, Cergy-Pontoise, FRANCE
[2] Departamento de Ingeniería Matemática,
Centro Modelamiento Matemático,
UMI 2807, UCHILE-CNRS,
Casilla 170-3 Correo 3, Santiago, CHILE.

January 9, 2007

Abstract

We reconsider the Dirichlet model for the random division of an interval. This model is parameterized by the number $n > 1$ of fragments, together with a set of positive parameters $(\theta_1, ..., \theta_n)$. Its main remarkable properties are recalled, developed and illustrated.

Explicit results on the statistical structure of its size-biased permutation are next supplied. This distribution appears in the sorting of items problem under the move-to-front rule. Assuming the parameters satisfy $\sum_{m=1}^{n} \theta_m \to \gamma < \infty$ as $n \uparrow \infty$, it is shown that the Dirichlet distribution has a Dirichlet-Kingman non-degenerate weak limit whose properties are briefly outlined.

KEYWORDS: Random discrete distribution, asymmetric Dirichlet partition, size-biased permutation, gamma subordinator, Dirichlet-Kingman partition.

Running title: Dirichlet-Kingman partition.

AMS 1991: Primary 60G57, 62E17, Secondary: 60K99, 62E15, 62E20

1 Introduction

Consider the random Dirichlet partition of the unit interval into n fragments with parameters $\boldsymbol{\theta}_n := (\theta_1, ..., \theta_n) > 0$. In this model, the random fragment sizes $\mathbf{S}_n := (S_1, ..., S_n)$, satisfying $\sum_{m=1}^n S_m = 1$, have Dirichlet distribution, say $D_n(\boldsymbol{\theta}_n)$, whose definition and main properties are discussed in Section 2; in particular, we recall here the RAM structure of $D_n(\boldsymbol{\theta}_n)$ (in Theorem 1) and a formula which allows to compute many spacings functionals in closed form (Theorem 2), making $D_n(\boldsymbol{\theta}_n)$ an exactly soluble model. As an illustration of this formula, we compute among other things the distribution of the pair $\left(S_{(1)}, S_{(n)}\right)$ arising in the order statistics $\mathbf{S}_{(n)} := \left(S_{(1)}, ..., S_{(n)}\right)$, satisfying $S_{(1)} > ... > S_{(n)}$ and $\sum_{m=1}^n S_{(m)} = 1$. We also draw attention on a stability under scaling property of Dirichlet partitions, together with some questions related to sampling problems from these partitions.

Explicit results on the law of the size-biased permutation $\mathbf{L}_n := \mathrm{SBP}(\mathbf{S}_n)$ of $\mathbf{S}_n$ are then given in Section 3. A size-biased permutation (SBP) of the fragment sizes is the one obtained in a size-biased sampling process without replacement from a Dirichlet partition $D_n(\boldsymbol{\theta}_n)$. It appears as the limiting law of a sampling process from $D_n(\boldsymbol{\theta}_n)$ when size-biased sampled fragments are iteratively moved to the front of the list in the heaps process. The main points which we deal with are the following: In Proposition 3, we derive the length of the first size-biased randomly chosen fragment, together with a stochastic comparison property with the lengths of $\mathbf{S}_n$. In Lemma 4, the order in which the consecutive fragments are being visited is considered. In Lemma 5, the residual allocation model (RAM) mixture structure of the SBP distribution is derived. In Theorem 6, we compute the joint law of the size-biased permutation fragment sizes explicitly. Using this, it is shown in Corollary 7 that, under some conditions, consecutive fragments in the size-biased permuted partition are arranged in stochastic descending order.

In Section 4, the limiting Dirichlet-Kingman partition on the infinite dimensional simplex is considered. Assuming the parameters $\boldsymbol{\theta}_n$ are such that $\gamma_n := \sum_{m=1}^n \theta_m$ converges to some finite limit $\gamma > 0$ as $n \uparrow \infty$, it is shown that $\mathbf{S}_n \overset{d}{\sim} D_n(\boldsymbol{\theta}_n)$ has a Dirichlet-Kingman non-degenerate weak limit. Some of its properties are outlined.

2 Asymmetric Dirichlet partition of the interval: Definition and main properties

2.1 The Dirichlet model

We shall consider the following random partition of the unit interval into n fragments. Let $\boldsymbol{\theta}_n := (\theta_m, m = 1, ..., n)$ be some set of n positive parameters and introduce $\gamma_m := \sum_{l=1}^m \theta_l,\ m = 1, ..., n$. With $\Gamma(\cdot)$ the Euler gamma

function, assume that the random fragment sizes $\mathbf{S}_n := (S_1, ..., S_n)$, satisfying $\sum_{m=1}^n S_m = 1$, is distributed according to the Dirichlet $D_n(\boldsymbol{\theta}_n)$ density function with respect to the uniform distribution on the simplex

$$f_{S_1,...,S_n}(s_1, ..., s_n) = \frac{\Gamma(\gamma_n)}{\prod_{m=1}^n \Gamma(\theta_m)} \prod_{m=1}^n s_m^{\theta_m - 1} \cdot \delta_{(\sum_{m=1}^n s_m - 1)}. \tag{2.1}$$

We shall put $\mathbf{S}_n \overset{d}{\sim} D_n(\boldsymbol{\theta}_n)$ if $\mathbf{S}_n$ is Dirichlet distributed with parameters $\boldsymbol{\theta}_n$ as above. Note that $\mathbf{S}_n$ also can be interpreted as spacings between consecutive points on the interval located at $\mathcal{S}_m := \sum_{k=1}^m S_k$, $m = 1, ..., n$, with $\mathcal{S}_0 := 0$.

Alternatively, the law of $\mathbf{S}_n := (S_1, ..., S_n)$ can easily be shown to be characterized by its joint moment function $(q_m > -\theta_m, m = 1, ..., n)$

$$\mathbf{E}\left[\prod_{m=1}^n S_m^{q_m}\right] = \frac{\Gamma(\gamma_n)}{\Gamma(\gamma_n + \sum_{m=1}^n q_m)} \prod_{m=1}^n \frac{\Gamma(\theta_m + q_m)}{\Gamma(\theta_m)}. \tag{2.2}$$

This expression is symmetric in the parameters and we shall therefore assume without loss of generality that $\theta_1 \geq ...\theta_1 \geq \theta_n$.

We first recall that if a $(0,1)-$valued random variable B has beta distribution with parameters $\alpha, \beta > 0$ (say $B \overset{d}{\sim} \text{beta}(\alpha, \beta)$) then its density is

$$f_B(s) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} s^{\alpha - 1}(1 - s)^{\beta - 1}, \ s \in (0, 1),$$

so that its moment function is

$$\mathbf{E}[B^q] = \frac{\Gamma(\alpha + q)}{\Gamma(\alpha)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha + \beta + q)}, \ q > -\alpha,$$

with $\Gamma(.)$ the Euler-gamma function. Further, $B \overset{d}{\to} 1$ when $\beta \downarrow 0$ and, since $1 - B \overset{d}{\sim} \text{beta}(\beta, \alpha)$, $B \overset{d}{\to} 0$ when $\alpha \downarrow 0$. By convention, we shall therefore identify $\text{beta}(\alpha, 0)$ (respectively $\text{beta}(0, \beta)$) to a Dirac measure at point 1 (respectively 0).

Similarly, we recall that a positive random variable X with gamma(θ) distribution (say $X \overset{d}{\sim} \text{gamma}(\theta)$, $\theta > 0$) has density

$$f_X(x) = \frac{1}{\Gamma(\theta)} x^{\theta - 1} e^{-x}, \ x > 0,$$

moment function

$$\mathbf{E}[X^q] = \Gamma(\theta + q)/\Gamma(\theta), \ q > -\theta$$

and Laplace-Stieltjes transform $\mathbf{E}\left[e^{-pX}\right] = (1 + p)^{-\theta}$, $p > -1$.

We shall also make use of the following notations in the sequel:

Let $m_1 \neq ... \neq m_k$ be a $k-$sequence of distinct integers taken from $\{1, ..., n\}$; then $\gamma_{n\backslash m_1,..,m_k} := \gamma_n - \sum_{l=1}^{k} \theta_{m_l}$. By convention, when $k = 0$, we shall assume that $\gamma_{n\backslash 1,..,k} := \gamma_n$.

Making use of the above notational convenience, it follows from Eq. (2.2) that $S_m \overset{d}{\sim} \text{beta}\left(\theta_m, \gamma_{n\backslash m}\right)$, $m = 1, ..., n$; the individual fragment sizes are not, in general, identically distributed, unless all θ_m are equal which we shall rule out in the sequel as it was studied in detail elsewhere. Under our hypothesis $\theta_1 \geq ... \geq \theta_n$, it may be checked that $S_1 \succeq ... \succeq S_n$ i.e. that the fragment sizes are arranged in stochastically decreasing order (the likelihood ratio between adjacent pairs S_{m-1}, S_m being monotone).

Let $\sigma_m := \mathbf{E}(S_m) = \theta_m/\gamma_n; m = 1, ..., n$.

Then the parameter set $\boldsymbol{\theta}_n := (\theta_m, m = 1, ..., n)$ can also be mapped into $(\sigma_m; m = 1, ..., n-1, \gamma_n)$, the transformation $\boldsymbol{\theta}_n \rightarrow (\sigma_m; m = 1, ..., n-1, \gamma_n)$ being one-to-one with $\theta_m = \sigma_m \gamma_n; m = 1, ..., n-1$ and $\theta_n = \left(1 - \sum_1^{n-1} \sigma_m\right) \gamma_n$. Parameter γ_n is a "precision" parameter that indicates how concentrated the distribution of $\mathbf{S}_n$ is around its mean $\boldsymbol{\sigma}_n := (\sigma_1, ..., \sigma_n)$: The larger γ_n is, the more the distribution of $\mathbf{S}_n$ is concentrated around $\boldsymbol{\sigma}_n$ (as one can check by observing that univariately $\mathbf{E}(S_m) = \sigma_m$ and $\sigma^2(S_m) = \sigma_m(1 - \sigma_m)/(\gamma_n + 1)$).

Examples of sequences $\theta_m; m = 1, ..., n$:

1/ Assume $\theta_m = n - m + 1$, $m = 1, ..., n$. Then $\gamma_n = (n(n+1))/2$ and $\sigma_m = 2(n - m + 1)/(n(n+1))$.

2/ Assume $\theta_m = 1/m$, $m = 1, ..., n$. Then $\gamma_n = \sum_1^n 1/m$ is the $n - th$ harmonic number.

3/ Assume $\theta_m = p^m$, $m = 1, ..., n$ and $p \in (0, 1)$. Then $\gamma_n = \sum_1^n p^m = p(1 - p^n)/(1 - p)$.

Remarks (ubiquity of Dirichlet partitions):

(i) In population genetics, Dirichlet distributions derive their importance from the fact that they are limit laws of certain diffusion processes on the simplex, whose state-space can be interpreted as gene frequencies at a selectively neutral locus under the finitely-many alleles model of mutation. These are properly rescaled versions of the Wright-Fisher, Moran or Cannings models (see Ewens (1990) and references therein for review). A completely different occurrence of Dirichlet partitions as limit laws under certain dilution or erosion events is also described in Vlad et al (2002).

(ii) Next, let $\boldsymbol{\mu}_n := (\mu_m := \mathbf{E}(-\log S_m),\ m = 1, ..., n)$ denote the observable. We note that $\mu_m = \psi(\gamma_m) - \psi(\gamma_n)$ where ψ is the classical digamma function. As can easily be checked (again see Vlad et al (2002) for example), $\boldsymbol{\theta}_n$

are Legendre conjugates of $\boldsymbol{\mu}_n$ because $D_n(\boldsymbol{\theta}_n)$ is the distribution maximizing entropy under the constraints $\boldsymbol{\mu}_n$. The involved partition function is

$$Z_n(\boldsymbol{\theta}_n) = \prod_{m=1}^{n} \left[\Gamma(\theta_m)/\Gamma(\gamma_n)\right].$$

(*iii*) In the random version of the Rényi car-parking problem with no stopping rule, there is an occurrence of an asymmetric Dirichlet partition in the counting-of-intervals-packed problem (Baryshnikov, Gnedin 2001). Here indeed, some implicit use is made of $D_3(1, \alpha - 1, 1)$ with $\alpha > 1$. Equivalently, the probability law of a random interval I launched on $[0,1]$ is given by

$$\mathbf{P}(I \subset [s_1, 1 - s_3]) = (1 - s_1 - s_3)^{\alpha}$$

on $\Delta = \{s_1 \geq 0, s_3 \geq 0, s_1 + s_3 \leq 1\}$. It depends only on the length $1 - s_1 - s_3$ of $[s_1, 1 - s_3]$, not on its location (translational invariance) and possesses the scaling property

$$\phi_{s_1,s_3}(I) \mid I \subset [s_1, 1 - s_3] \stackrel{d}{=} I$$

where $\phi_{s_1,s_3}(.)$ is the increasing affine transformation mapping $[s_1, 1 - s_3]$ onto $[0,1]$. In this model, the middle interval will not further split (the launched interval is packed), whereas the two side-ones are allowed to further split (they constitute gaps where additional intervals can be inserted if small enough to fit).

These properties allow to treat the packing of random intervals problem as a non-conservative self-similar fragmentation process where fragmentation of size-x fragments occur at rate x^{α} (Bertoin, Gnedin 2004).

2.2 Main statistical features

We now proceed with further properties.

- *The RAM structure of the Dirichlet partition*

We recall here a fundamental property of the Dirichlet partition $\mathbf{S}_n \stackrel{d}{\sim} D_n(\boldsymbol{\theta}_n)$ which appears as an exercise (without proof) in the book of Devroye (1986), on page 585; see also Ewens (1990) and Haas and Formery (2002) where this property is used and discussed respectively in biological and geological applications. The term RAM stands for residual allocation model which is sometimes used in this context. This property has been known for long to Bayesian statisticians (see e.g. Freedman (1963), Fabius (1964) and Ferguson (1974)).

Theorem 1 *Let $(B_1, ..., B_{n-1})$ be independent random variables with distribution $B_k \stackrel{d}{\sim} beta(\theta_k, \gamma_{n\setminus 1,..,k})$, $k = 1, ..., n-1$. With $\overline{B}_i := 1 - B_i \stackrel{d}{\sim} beta(\gamma_{n\setminus 1,..,i}, \theta_i)$*

and $\prod_{i=1}^{0} \overline{B}_i := 1$*, define*

$$S_k := \left(\prod_{i=1}^{k-1} \overline{B}_i \right) B_k \ , \ k = 1, ..., n-1, \tag{2.3}$$

$$S_n = 1 - \sum_{k=1}^{n-1} S_k = \prod_{k=1}^{n-1} \overline{B}_k. \tag{2.4}$$

Then $\mathbf{S}_n \overset{d}{\sim} D_n(\boldsymbol{\theta}_n)$.

This representation of $D_n(\boldsymbol{\theta}_n)$ *is called the stick-breaking scheme (or RAM) representation of* $\mathbf{S}_n$.

Proof. Using independence and the expression of the moment function for beta-distributed random variables,

$$\begin{aligned}
\mathbf{E}(S_k^q) &= \left[\prod_{i=1}^{k-1} \mathbf{E}\left(\overline{B}_i^q \right) \right] \mathbf{E}(B_k^q) \\
&= \left[\prod_{i=1}^{k-1} \frac{\Gamma\left(\gamma_{n\setminus 1,..,i} + q\right) \Gamma\left(\gamma_{n\setminus 1,..,i-1}\right)}{\Gamma\left(\gamma_{n\setminus 1,..,i}\right) \Gamma\left(\gamma_{n\setminus 1,..,i-1} + q\right)} \right] \frac{\Gamma(\theta_k + q) \Gamma\left(\gamma_{n\setminus 1,..,k-1}\right)}{\Gamma(\theta_k) \Gamma\left(\gamma_{n\setminus 1,..,k-1} + q\right)} \\
&= \frac{\Gamma(\gamma_n) \Gamma(\theta_k + q)}{\Gamma(\gamma_n + q) \Gamma(\theta_k)}, \text{ for } k = 1, ..., n-1.
\end{aligned}$$

This shows that $S_k \overset{d}{\sim} \text{beta}\left(\theta_k, \gamma_{n\setminus k}\right)$, $k = 1, ..., n-1$.

Next, $\mathbf{E}(S_n^q) = \prod_{k=1}^{n-1} \mathbf{E}\left(\overline{B}_k^q \right) = \frac{\Gamma(\gamma_n)\Gamma(\theta_n+q)}{\Gamma(\gamma_n+q)\Gamma(\theta_n)}$ and $S_n \overset{d}{\sim} \text{beta}\left(\theta_n, \gamma_{n\setminus n}\right)$. The joint distribution of the S_m can be treated similarly. □

• *The gamma distribution and Dirichlet partition*

As is well-known, the Dirichlet partition $\mathbf{S}_n \overset{d}{\sim} D_n(\boldsymbol{\theta}_n)$, defined by (2.1), can be generated by $S_m = X_m / \mathcal{X}_n$, with $\mathcal{X}_n := \sum_{m=1}^{n} X_m$ the sum of n independent gamma(θ_m) distributed random variables. In addition, $(X_1, ..., X_{m-1})$ is independent of $\mathcal{X}_n$. As a result, $\mathbf{S}_n$ can also be defined conditionally by

$$\mathbf{S}_n \overset{d}{=} (X_1, ..., X_n \mid \mathcal{X}_n = 1).$$

More generally, considering the same construction starting with an interval of length $x > 0$ gives consecutive spacings, say $\mathbf{S}_n(x) := (S_m(x) : \ m = 1, ..., n)$ generated by

$$\mathbf{S}_n(x) \overset{d}{=} (X_1, ..., X_n \mid \mathcal{X}_n = x).$$

From this, one can check the scaling property $S_m(x) \overset{d}{=} x S_m(1) := x S_m$, $m = 1, ..., n$, univariately and multivariately and $\sum_{m=1}^{n} S_m(x) = x$.

• *A constructive formula for computing with Dirichlet*

Furthermore, the following result which results directly from the above properties can be useful (see Huillet and Martinez (2003))

Theorem 2 *Consider the Dirichlet partitioning model* $\mathbf{S}_n \overset{d}{\sim} D_n(\boldsymbol{\theta}_n)$.

(i) *Let* f *be any Borel-measurable function for which*

$$\int_0^\infty \mathbf{E}\left(|f(\mathbf{S}_n(x))|\right) x^{\gamma_n - 1} e^{-px} dx < \infty.$$

Then, with $\mathbf{X}_n(p) := (X_m(p); m = 1, ..., n)$, n *independent random variables defined by* $X_m(p) = \frac{1}{p} X_m$, $p > 0$, $m = 1, ..., n$ *where* $X_m \overset{d}{\sim}$ *gamma*(θ_m), *we have*

$$\int_0^\infty \mathbf{E}\left(f(\mathbf{S}_n(x))\right) x^{\gamma_n - 1} e^{-px} dx = \frac{\Gamma(\gamma_n)}{p^{\gamma_n}} \mathbf{E}\left(f(\mathbf{X}_n(p))\right). \tag{2.5}$$

(ii) *If* f *is homogeneous of degree* d, *i.e. if* $f(x\mathbf{s}_n) = x^d f(\mathbf{s}_n)$, $x > 0$, $\mathbf{s}_n := (s_1, ..., s_n) \in \mathbf{R}^n$, *and if* $\mathbf{E}\left(|f(\mathbf{S}_n)|\right) < \infty$ *then, with* $\mathbf{X}_n := (X_1, ..., X_n)$

$$\mathbf{E}\left(f(\mathbf{S}_n)\right) = \frac{\Gamma(\gamma_n)}{\Gamma(\gamma_n + d)} \mathbf{E}\left(f(\mathbf{X}_n)\right). \tag{2.6}$$

This Theorem 2 allows to compute many spacings functionals in terms of simpler functionals of independent gamma random variables or processes. We list below some of its applications. It generalizes to the full asymmetric Dirichlet partition a formula first given by Steutel (1967) in the context of uniform partitions.

A series of direct applications of Theorem 2.

Consider the statement (i) of Theorem 2. The right hand-side quantity

$$\Gamma(\gamma_n)\, p^{-\gamma_n} \mathbf{E}\left(f(\mathbf{X}_n(p))\right)$$

may be interpreted as the Laplace transform in the variable p of $\mathbf{E}\left(f(\mathbf{S}_n(x))\right) x^{\gamma_n - 1}$. Inverting this Laplace transform and putting $x = 1$ yields $\mathbf{E}\left(f(\mathbf{S}_n)\right)$.

1. As an illustration, we shall use this to compute the joint distribution of the largest and smallest spacing in a Dirichlet$(\boldsymbol{\theta}_n)$ partition. Suppose $1 \geq b > a \geq 0$ and consider the spacings' functional

$$f(S_1, ..., S_n) = \prod_{m=1}^n \mathbf{I}(a < S_m \leq b).$$

Then $\mathbf{E}f(S_1, ..., S_n) = \mathbf{P}\left(S_{(n)} > a, S_{(1)} \leq b\right)$ is the required probability, assuming $S_{(1)} > ... > S_{(n)}$ to be the order statistics of $(S_1, ..., S_n)$. The case $a = 0$ $(b = 1)$ gives the probability $\mathbf{P}\left(S_{(1)} \leq b\right)$, respectively $\mathbf{P}\left(S_{(n)} > a\right)$.

From statement (i) of Theorem 2 indeed, the quantity

$$\Gamma(\gamma_n)\, p^{-\gamma_n} \prod_{m=1}^{n} \mathbf{P}(a < X_m(p) \leq b) = \Gamma(\gamma_n) \prod_{m=1}^{n} \left[\frac{1}{\Gamma(\theta_m)} \int_a^b x^{\theta_m - 1} e^{-px} dx \right]$$

interprets as the Laplace transform of $\mathbf{P}\left(S_{(n)}(x) > a, S_{(1)}(x) \leq b\right) x^{\gamma_n - 1}$. Inverting this Laplace transform and putting $x = 1$ yields $\mathbf{P}\left(S_{(n)} > a, S_{(1)} \leq b\right)$. From this, we obtain directly

$$\mathbf{P}\left(S_{(n)} > a, S_{(1)} \leq b\right) = \frac{\Gamma(\gamma_n)}{\prod_{m=1}^{n} \Gamma(\theta_m)} *_{m=1}^{n} h_{\theta_m}(1),$$

where $*_{m=1}^{n} h_{\theta_m}(1)$ is the n-fold convolution of the functions $x \to h_{\theta_m}(x) = x^{\theta_m - 1} \mathbf{I}(b \geq x > a)$, $m = 1, ..., n$ evaluated at $x = 1$. Putting $a = s$, $b = 2s$, the above probability turns out to be the probability of a s-parking configuration with n cars, when Dirichlet-parked cars of size s avoid overlap with no room left to insert a new car within gaps.

2. From part (ii) of the Theorem 2, any homogeneous functional of Dirichlet spacings can be directly computed from the simpler one of independent gamma variables each with specific means θ_m which leads to considerable simplification.

2.*a*. For example, considering the function $f(S_1, ..., S_n) = \prod_{m=1}^{n} S_m^{q_m}$, we get that f is homogeneous with degree $d = \sum_{m=1}^{n} q_m$. Application of (ii) in this particular case gives Eq. (2.2).

2.*b*. Let $m_1 \neq ... \neq m_k \in \{1, ..., n\}$ be k distinct integers. Considering the function $f(S_1, ..., S_n) = \left(\sum_{l=1}^{k} S_{m_l}\right)^q$, we see that f is homogeneous with degree $d = q$. Application of (ii) in this particular case gives

$$\mathbf{E}\left[\left(\sum_{l=1}^{k} S_{m_l}\right)^q\right] = \frac{\Gamma(\gamma_n)}{\Gamma(\gamma_n + q)} \frac{\Gamma\left(\sum_{l=1}^{k} \theta_{m_l} + q\right)}{\Gamma\left(\sum_{l=1}^{k} \theta_{m_l}\right)},$$

showing that $\sum_{l=1}^{k} S_{m_l} \overset{d}{\sim} \text{beta}\left(\sum_{l=1}^{k} \theta_{m_l}, \gamma_{n \setminus m_1, .., m_k}\right)$.

2.*c*. (stability under scaling operations). Consider the random variables

$$\widetilde{S}_{m_l} := S_{m_l} / \sum_{l=1}^{k} S_{m_l},\ l = 1, .., k.$$

They constitute a new partition of the unit interval and

$$(i)\quad \left(\widetilde{S}_{m_l}; l = 1, .., k\right) \overset{d}{\sim} D_k(\theta_{m_1}, ..., \theta_{m_k}),$$

$$(ii)\quad \left(\widetilde{S}_{m_l}; l=1,..,k\right) \text{ and } \sum_{l=1}^{k} S_{m_l} \text{ are independent.}$$

To see this, we observe that $f(S_1,...,S_n) := \left(\sum_{l=1}^k S_{m_l}\right)^{q_0} \prod_{l=1}^k \widetilde{S}_{m_l}^{q_l}$ is homogeneous with degree q_0 resulting in

$$\begin{aligned}\mathbf{E}f(S_1,...,S_n) &= \frac{\Gamma(\gamma_n)}{\Gamma(\gamma_n+q_0)}\mathbf{E}\left[\left(\sum_{l=1}^k X_{m_l}\right)^{q_0}\prod_{l=1}^k \widetilde{X}_{m_l}^{q_l}\right] \\ &= \frac{\Gamma\left(\sum_{l=1}^k \theta_{m_l}+q_0\right)\Gamma(\gamma_n)}{\Gamma\left(\sum_{l=1}^k \theta_{m_l}\right)\Gamma(\gamma_n+q_0)}\frac{\Gamma\left(\sum_{l=1}^k \theta_{m_l}\right)}{\Gamma\left(\sum_{l=1}^k (\theta_{m_l}+q_l)\right)}\prod_{l=1}^k \frac{\Gamma(\theta_{m_l}+q_l)}{\Gamma(\theta_{m_l})}\end{aligned}$$

because $\widetilde{X}_{m_l} := X_{m_l}/\sum_{l=1}^k X_{m_l}$, $X_{m_l} \overset{d}{\sim} \text{gamma}(\theta_{m_l})$, $l=1,...,k$, is independent of $\sum_{l=1}^k X_{m_l}$ and $\left(\widetilde{X}_{m_l}; l=1,...,k\right) \overset{d}{\sim} D_k(\theta_{m_1},...,\theta_{m_k})$.

2.*d*. The distribution of the partition function $\sum_{m=1}^n S_m^q$ is sometimes of interest. In particular, its mean value $\mathbf{E}\left(\sum_{m=1}^n S_m^q\right)$ but also its full moment function $\mathbf{E}\left(\left(\sum_{m=1}^n S_m^q\right)^\lambda\right)$ are worth being considered. For general partitions, these quantities are hardly computable. When considering the Dirichlet partition model, significant simplifications are expected since the spacing functional

$$f(S_1,...,S_n) = \left(\sum_{m=1}^n S_m^q\right)^\lambda$$

is homogeneous with degree $d = q\lambda$ and so part (ii) of Theorem 2 applies.

3. We finally briefly outline that these tools are also useful in the computation of simple sampling formulae.

3.*a*. Let $(U_1,...,U_k)$ be k iid uniform throws on $\mathbf{S}_n$. Let $\mathbf{K}_n := (K_1,...,K_n)$ be an integral-valued random vector which counts the number of visits to the different fragments in a $k-$sample. Hence, if N_l is the random fragment number in which the $l-$th trial falls, then $K_m := \sum_{l=1}^k \mathbf{I}(N_l = m)$, $m=1,...,n$.

With $\sum_{m=1}^n k_m = k$ and $\mathbf{k}_n := (k_1,...,k_n)$ we have the multinomial distribution:

$$\mathbf{P}(\mathbf{K}_n = \mathbf{k}_n \mid \mathbf{S}_n) = \frac{k!}{\prod_{m=1}^n k_m!}\prod_{m=1}^n S_m^{k_m}.$$

Averaging over $\mathbf{S}_n$, applying (ii) of Theorem 2, we find

$$\mathbf{P}(\mathbf{K}_n = \mathbf{k}_n) = \mathbf{E}\mathbf{P}(\mathbf{K}_n = \mathbf{k}_n \mid \mathbf{S}_n) = \frac{\prod_{m=1}^n \{\theta_m\}_{k_m}}{\{\gamma_n\}_k},$$

where $\{\theta\}_k := (\theta)_k / k!$ and $(\theta)_k := \theta(\theta+1)...(\theta+k-1)$, $k \geq 1$, $(\theta)_0 := 1$. This distribution is known as the Dirichlet multinomial distribution.

Applying Bayes formula, the posterior distribution of $\mathbf{S}_n$ given $\mathbf{K}_n = \mathbf{k}_n$ is determined by its density at point $\mathbf{s}_n$ on the simplex as

$$f_{\mathbf{S}_n}(\mathbf{s}_n \mid \mathbf{K}_n = \mathbf{k}_n) = \frac{\Gamma(\gamma_n + k)}{\prod_{m=1}^n \Gamma(\theta_m + k_m)} \prod_{m=1}^n s_m^{(\theta_m + k_m)-1} \cdot \delta_{(\sum_{m=1}^n s_m - 1)}.$$

This shows, as is well-known, that $\mathbf{S}_n \mid \mathbf{K}_n = \mathbf{k}_n \overset{d}{\sim} D_n(\boldsymbol{\theta}_n + \mathbf{k}_n)$, where $\boldsymbol{\theta}_n + \mathbf{k}_n = (\theta_1 + k_1, ..., \theta_n + k_n)$ is obtained by shifting $\boldsymbol{\theta}_n$ and hence that

$$\mathbf{E}(S_m \mid \mathbf{K}_n = \mathbf{k}_n) = \frac{\theta_m + k_m}{\gamma_n + k},\ m = 1, ..., n.$$

This suggests the following recursive approach to the sampling formula where successive samples are now drawn from the corresponding iterative posterior distributions. More specifically, let $(N_1, ...N_k) \in \{1, ..., n\}^k$ be the labels of the successive fragments thus drawn. Then,

$$\mathbf{P}(N_1 = n_1) = \mathbf{E}(\mathbf{P}(N_1 = n_1) \mid \mathbf{S}_n) = \mathbf{E}(S_{n_1}) = \frac{\theta_{n_1}}{\gamma_n},$$

$$\mathbf{P}(N_2 = n_2 \mid N_1) = \frac{\theta_{n_2} + \mathbf{I}(N_1 = n_2)}{\gamma_n + 1}, ...,$$

$$\mathbf{P}(N_k = n_k \mid N_1, ..., N_{k-1}) = \frac{\theta_{n_k} + \sum_{l=1}^{k-1} \mathbf{I}(N_l = n_k)}{\gamma_n + k - 1}.$$

Proceeding in this way, the joint distribution of $(N_1, ..., N_k)$ reads

$$\begin{aligned} \mathbf{P}(N_1 = n_1, ..., N_k = n_k) &= \frac{\theta_{n_1}}{\gamma_n} \prod_{l=1}^{k-1} \frac{\theta_{n_{l+1}} + \sum_{j=1}^{l} \mathbf{I}(N_j = n_{l+1})}{\gamma_n + l} \\ &= \frac{\prod_{m=1}^n (\theta_m)_{k_m}}{(\gamma_n)_k}, \end{aligned}$$

where $k_m := \sum_{l=1}^k \mathbf{I}(n_l = m)$. Being invariant under permutations of the entries, this distribution is exchangeable. The sequence $N_1, ..., N_k$ is a Pòlya urn sequence.

3.*b*. The joint conditional generating function of $\mathbf{K}_n$ reads

$$\mathbf{E}\left(\prod_{m=1}^n u_m^{K_m} \mid \mathbf{S}_n\right) = \left(\sum_{m=1}^n u_m S_m\right)^k,$$

which is homogeneous with degree $d = k$ allowing to compute $\mathbf{E}\left(\prod_{m=1}^{n} u_m^{K_m}\right)$. Further, with $\widetilde{X}_m := X_m / \sum_{m=1}^{n} X_m$, $X_m \overset{d}{\sim} \text{gamma}(\theta_m)$, $m = 1, ..., n$, as above, using independence between $\left(\widetilde{X}_m,\ m = 1, ..., n\right)$ and $\sum_{m=1}^{n} X_m$ and recalling $\left(\widetilde{X}_m,\ m = 1, ..., n\right) \overset{d}{\sim} D_n(\boldsymbol{\theta}_n)$

$$\begin{aligned}
\mathbf{E}\left(\prod_{m=1}^{n} u_m^{K_m/k}\right) &= \frac{\Gamma(\gamma_n)}{\Gamma(\gamma_n + k)} \mathbf{E}\left[\left(\sum_{m=1}^{n} u_m^{1/k} X_m\right)^k\right] \\
&\underset{k\uparrow\infty}{\sim} \frac{\Gamma(\gamma_n)}{\Gamma(\gamma_n + k)} \mathbf{E}\left[\left(\sum_{m=1}^{n} X_m\right)^k \left(1 + \frac{1}{k}\sum_{m=1}^{n} \widetilde{X}_m \log u_m\right)^k\right] \\
&\underset{k\uparrow\infty}{\sim} \mathbf{E}\left(\prod_{m=1}^{n} u_m^{\widetilde{X}_m}\right) = \mathbf{E}\left(\prod_{m=1}^{n} u_m^{S_m}\right).
\end{aligned}$$

This shows that

$$\mathbf{K}_n / k \xrightarrow{d} \mathbf{S}_n \text{ as } k \uparrow \infty.$$

Note that, applying the strong law of large numbers (conditionally given $\mathbf{S}_n$), the above convergence in law also holds almost surely.

• *Representation of $D_n(\boldsymbol{\theta}_n)$ in terms of a conditioned Moran subordinator*

Let us first recall some well-known facts from infinitely divisible random variables and processes [See Bertoin (1996), and Steutel and van Harn (2004), for general monographs on infinite-divisibility].

Let $\mathcal{X} \geq 0$ be an infinitely divisible (ID) random variable with Laplace-Stieltjes transform $\phi(p) := \mathbf{E}\left[e^{-p\mathcal{X}}\right]$, $p \geq 0$ and Laplace exponent

$$\psi(p) := -\log \phi(p) = \int_0^\infty \left(1 - e^{-px}\right) \pi(dx). \tag{2.7}$$

Here, $\pi(dx)$ is a positive Radon measure on $(0, \infty)$, the Lévy measure for jumps of $\mathcal{X}$. Let $\overline{\pi}(x) := \pi(x, \infty)$ be its continuous and decreasing tail function. It is assumed that $\overline{\pi}(0) = \infty$ and $\int_{(0,\infty)} (1 \wedge x)\, \pi(dx) < \infty$.

Let $(\Gamma_k,\ k \geq 1)$ be a Poisson point process on the half line $(0, \infty)$ meaning that $(\Gamma_k - \Gamma_{k-1},\ k \geq 1)$ are iid and $\exp(1)$ −distributed. From the Lévy-Itô decomposition for ID random variables in terms of jumps, we have

$$\mathcal{X} \overset{d}{=} \sum_{k=1}^{\infty} \overline{\pi}^{-1}(\Gamma_k). \tag{2.8}$$

Normalizing, i.e. defining $\left(\zeta_{(k)} := \overline{\pi}^{-1}(\Gamma_k) / \mathcal{X},\ k \geq 1\right)$, we are left with an infinite partition of the unit interval into random fragments with descending sizes

$\zeta_{(1)} > ... > \zeta_{(k)} > ...$ for which $1 = \sum_{k\geq 1} \zeta_{(k)}$. The distribution of $\left(\zeta_{(k)},\, k \geq 1\right)$ is called a Poisson-Kingman (PK) distribution, see Pitman, (2003).

Assume now $\pi(dx) = \gamma x^{-1} e^{-x} dx$, $x > 0$. Then $\mathcal{X} =: \mathcal{X}_\gamma$ has gamma(γ) distribution for which $\phi(p) = (1+p)^{-\gamma}$.

The normalized partition $\left(\zeta_{(k)} := \overline{\pi}^{-1}(\Gamma_k)/\mathcal{X}_\gamma,\, k \geq 1\right)$ in this particular case is Poisson-Dirichlet partition, say $PD(\gamma)$.

A closely related point of view is the following; the random variable $\mathcal{X}_\gamma$ induces a gamma (or Moran) subordinator process $(\mathcal{X}_t; t \geq 0)$ with $\mathcal{X}_0 := 0$ and Lévy measure for everywhere dense jumps $x^{-1}e^{-x}dx$. Let $\gamma > 0$ be such that $\mathcal{X}_\gamma = 1$ and consider the conditioned process $(\mathcal{X}_t; t \geq 0 \mid \mathcal{X}_\gamma = 1)$. Then the rank jumps of this conditioned process again have Poisson-Dirichlet distribution.

In addition, see Kingman (1993), with $\gamma_m := \sum_{k=1}^m \theta_k$, $\gamma_0 = 0$

$$\left(\mathcal{X}_{\gamma_m} - \mathcal{X}_{\gamma_{m-1}}; m = 1, ..., n \mid \mathcal{X}_{\gamma_n} = 1\right) \overset{d}{\sim} D_n(\boldsymbol{\theta}_n).$$

In some applications (see Kingman (1975), Donnelly (1991) in the context of the heaps process), S_m, $m = 1, ..., n$, interpret as the random popularities of a collection of n books arranged on a shelf. If instead of a collection of books, a population of animals from n different species were considered, popularities verbatim interpret as species abundance; see Kingman (1978) and Ewens (1990) for such interpretations.

3 Size-biased permutation from asymmetric Dirichlet partitions

The results on size-biased permutation of the asymmetric Dirichlet distributions $D_n(\boldsymbol{\theta}_n)$ presented in this Section seem to be new. They extend the results presented in Barrera et al (2005) in the particular case of symmetric Dirichlet partitions $D_n(\theta)$, assuming $\theta_1 = ... = \theta_n =: \theta$. These were used to derive the search-cost distribution in heaps processes at equilibrium.

Assume some observer is sampling the unit interval as follows: Drop points at random onto the randomly broken interval $\mathbf{S}_n \overset{d}{\sim} D_n(\boldsymbol{\theta}_n)$ and record the corresponding labels of visited fragments. We shall consider the problem of determining the order in which the various fragments are discovered in such a sampling process. To avoid revisiting the same fragment many times, once it has been discovered, we need to remove it from the population as soon as it has been met in the sampling process. But to do that, the law of its size is needed. Once this is done, after renormalizing the remaining fragments' sizes, we are left with a population of $n-1$ fragments, the sampling of which will necessarily supply a so far undiscovered fragment. The distribution of its size can itself be computed and so forth, renormalizing again, until the whole available fragments population has been visited. In this way, not only the visiting order of the different fragments should be understood but also their sizes. The

purpose of this Section is to describe the statistical structure of the size-biased permutation of the fragment sizes as those obtained while avoiding the ones previously encountered in a sampling process from Dirichlet partition $D_n(\boldsymbol{\theta}_n)$.

Remark: We shall for example borrow the physical image to the heaps process; see Kingman (1975), Flajolet et al (1992), Fill (1996), Fill and Holst (1996) and Jelenkovìc (1999). Books' popularities are assumed to satisfy $\mathbf{S}_n \overset{d}{\sim} D_n(\boldsymbol{\theta}_n)$. When a book is demanded, it is removed and replaced (before a next demand) to the top of the shelf, other books being shifted accordingly; successive demands are independent. Iterating this heaps process (as a recurrent positive Markov chain over the set of permutations), there is intuitively a tendency, when the system has reached equilibrium, to find more popular books to the top of the heap. At equilibrium indeed (see Donnelly (1991) and references therein to Dies, Hendricks and Letac' works), books' popularities are given by $\mathbf{L}_n := \mathrm{SBP}(\mathbf{S}_n) \overset{d}{\sim} SBD_n(\boldsymbol{\theta}_n)$ and result (iii) in Corollary 7 stating that $L_1 \succeq ... \succeq L_n$ confirms and gives statistical sense to this intuition. Note from this that $\mathbf{L}_n = \mathrm{SBP}(\mathbf{L}_n)$ ($\mathbf{L}_n$ is invariant under size-biased permutation) and that $\mathbf{L}_n = \mathrm{SBP}\big(\mathbf{S}_{(n)}\big)$ since $\mathbf{S}_{(n)}$ is simply obtained from $\mathbf{S}_n$ by rearranging its components in descending order, observing that the sampling process is blind to the mutual fragments' positions, being only sensitive to their sizes.

• *The length of the first size-biased randomly chosen fragment*

From the size-biased picking construction, it follows (see Engen (1978), for example, for similar treatment) that for any non-negative measurable function φ on $[0,1]$,

(3.1)
$$\mathbf{E}[\varphi(L_1)] = \mathbf{E}[\mathbf{E}[\varphi(L_1) \mid \mathbf{S}_n]] =$$
$$\sum_{m=1}^{n} \mathbf{E}[(\varphi(S_m))\,\mathbf{P}(L_1 = S_m \mid \mathbf{S}_n)] = \sum_{m=1}^{n} \mathbf{E}[S_m \varphi(S_m)].$$

Taking in particular $\varphi(x) = \mathbf{I}(x > s)$ in Eq. (3.1), we get the so-called structural distribution $\overline{F}_{L_1}(s) := \mathbf{P}[L_1 > s]$ in the form

(3.2)
$$\overline{F}_{L_1}(s) = \sum_{m=1}^{n} \mathbf{E}[S_m \mathbf{I}(S_m > s)] = \sum_{m=1}^{n} \int_s^1 t \cdot dF_{S_m}(t).$$

Proposition 3 *We have:* $L_1 = B_{M_1;n\setminus M_1}$ *where* $B_{M_1;n\setminus M_1} \overset{d}{\sim} beta\big(1+\theta_{M_1}, \gamma_{n\setminus M_1}\big)$ *is a* M_1-*mixture of* $beta\big(1+\theta_{M_1}, \gamma_{n\setminus M_1}\big)$ *distributed random variables and it holds that*

(3.3)
$$L_1 \succeq S_n.$$

Proof. Recalling that $S_m \overset{d}{\sim} \mathrm{beta}\big(\theta_m; \gamma_{n\setminus m}\big)$, $m = 1, ..., n$, one can check

directly from Eq. (3.2) that

$$\overline{F}_{L_1}(s) = \sum_{m=1}^{n} \int_s^1 \frac{\Gamma(\gamma_n)}{\Gamma(\theta_m)\,\Gamma\left(\gamma_{n\setminus m}\right)} t^{(\theta_m+1)-1}(1-t)^{\gamma_n-\theta_m-1}\,dt$$

$$= \sum_{m=1}^{n} \frac{\theta_m}{\gamma_n} \int_s^1 \frac{\Gamma(1+\gamma_n)}{\Gamma(1+\theta_m)\,\Gamma\left(\gamma_{n\setminus m}\right)} t^{(\theta_m+1)-1}(1-t)^{\gamma_n-\theta_m-1}\,dt,$$

which is the tail distribution function of a M_1-mixture of beta$\left(1+\theta_{M_1}, \gamma_{n\setminus M_1}\right)-$ distributed random variables with $\mathbf{P}(M_1 = m) = \frac{\theta_m}{\gamma_n}$, $m = 1, ..., n$. Equivalently,

$$\mathbf{E}\left(L_1^q\right) = \sum_{m=1}^{n} \frac{\theta_m}{\gamma_n} \frac{\Gamma(1+\theta_m+q)\,\Gamma(1+\gamma_n)}{\Gamma(1+\theta_m)\,\Gamma(1+\gamma_n+q)}$$

is the moment function of L_1, $q > -(1+\theta_n)$.

The likelihood ratio between the two distributions of L_1 and S_n being monotone, the stochastic domination property follows. □

The telling feature of this last observation is that although $\mathbf{S}_n := (S_1, ..., S_n)$ is such that $S_1 \succeq ... \succeq S_n$ with largest fragment at the top of the list, the first size-biased sampled fragment has size L_1 which is at least stochastically larger than the smallest of the S_m, namely S_n. (Note that, as required, when all θ_m are equal, $L_1 \succeq S_n$ means that L_1 is stochastically larger than any of the S_m).

• *The visiting order of the fragments in the SBP process.*

Before proceeding with the computation of the distribution of the full size-biased permutation partition, we need to consider the order in which the fragments are being visited.

For any permutation $m_1, ..., m_n$ of $\{1, ..., n\}$, with $M_1, ..., M_k$, $k = 1, ..., n$, the first k *distinct* fragments labels which have been visited in the SBP sampling process, we have

$$\mathbf{P}(M_1 = m_1, ..., M_k = m_k \mid \mathbf{S}_n) = \left(\prod_{i=1}^{k-1} \frac{S_{m_i}}{1-\sum_{l=1}^{i} S_{m_l}}\right) S_{m_k}, \tag{3.4}$$

so that

$$\mathbf{P}(M_k = m_k \mid \mathbf{S}_n, M_1 = m_1, ..., M_{k-1} = m_{k-1}) = \frac{S_{m_k}}{1-\sum_{l=1}^{k-1} S_{m_l}}. \tag{3.5}$$

As a result,

$$\mathbf{P}(M_k = m \mid \mathbf{S}_n) = S_m \sum_{(m_1 \neq ... \neq m_{k-1}) \neq m} \prod_{i=1}^{k-1} \frac{S_{m_i}}{1-\sum_{l=1}^{i} S_{m_l}} \tag{3.6}$$

is the conditional probability (given $\mathbf{S}_n$) that the $k-$th visited fragment is fragment number m from $D_n(\boldsymbol{\theta}_n)$.

Lemma 4 *Given $M_1 = m_1, ..., M_{k-1} = m_{k-1}$, the distribution of M_k is given by*

$$\mathbf{P}(M_k = m_k \mid M_1 = m_1, ..., M_{k-1} = m_{k-1}) = \frac{\theta_{m_k}}{\gamma_{n\setminus m_1,..,m_{k-1}}}$$

on the set $m \in \{1, ..., n\} \setminus \{m_1 \neq ... \neq m_{k-1}\}$. The joint probability distribution of $M_1, ..., M_k$, $k = 1, ..., n$, is

$$\mathbf{P}(M_1 = m_1, ..., M_k = m_k) = \prod_{i=1}^{k} \frac{\theta_{m_i}}{\gamma_{n\setminus m_1,..,m_{i-1}}}, \tag{3.7}$$

with $m_1 \neq ... \neq m_k \in \{1, ..., n\}$.

Proof. Although this result is immediate, we shall supply a short proof of it. From Eq. (3.5), the function $\mathbf{S}_n \to \frac{S_{m_k}}{1-\sum_{l=1}^{k-1} S_{m_l}} = \frac{S_{m_k}}{\sum_{m\neq\{m_1,...,m_{k-1}\}} S_m}$ is homogeneous with degree 0. Applying (ii) of Theorem 2, with $\{X_1, ..., X_n\}$ independent random variables satisfying $X_m \overset{d}{\sim} \text{gamma}(\theta_m)$, we get

$$\mathbf{P}(M_k = m_k \mid M_1 = m_1, ..., M_{k-1} = m_{k-1}) = \mathbf{E}\left[\frac{S_{m_k}}{\sum_{m\neq\{m_1,...,m_{k-1}\}} S_m}\right]$$

$$= \mathbf{E}\left[\frac{X_{m_k}}{\sum_{m\neq\{m_1,...,m_{k-1}\}} X_m}\right] = \frac{\theta_{m_k}}{\gamma_{n\setminus m_1,..,m_{k-1}}}.$$

The joint distribution of $M_1, ..., M_k$ results from the definition of the conditional probability. □

One can check that $M_1 \succ ... \succ M_n$: The fragments labels of the SBP sampling process are arranged in stochastically decreasing order (the likelihood ratio between adjacent pairs M_{k-1}, M_k being monotone).

• *The RAM structure of the size-biased permutation*

Let $\mathbf{S}_n := (S_1, ..., S_n)$ be the random partition of the interval $[0, 1]$ considered here. Let L_1 be the length of the first randomly chosen fragment M_1, so with $L_1 := S_{M_1}$. We have $\mathbf{P}(M_1 = m_1 \mid \mathbf{S}_n) = S_{m_1}$ and

$$\mathbf{P}(M_1 = m_1) = \frac{\theta_{m_1}}{\gamma_n}.$$

As was shown in Proposition 3, we have $L_1 \overset{d}{=} B_{M_1;n\setminus M_1}$ where $B_{M_1;n\setminus M_1} \overset{d}{\sim} \text{beta}\big(1+\theta_{M_1}, \gamma_{n\setminus M_1}\big)$ is a M_1-mixture of beta distributed random variables. A

standard problem is to iterate the size-biased picking procedure, by avoiding the fragments already encountered: By doing so, a size-biased permutation (SBP) of the fragments is obtained. It turns out that SBP($\mathbf{S}_n$) has a residual allocation model (RAM) structure.

In the first step of this size-biased picking procedure indeed, partition $\mathbf{S}_n =: \mathbf{S}_n^{(0)}$ is changed into

$$(L_1, S_1, ..., S_{M_1-1}, S_{M_1+1}, ..., S_n),$$

where, conditionally on M_1, $(L_1, S_1, ..., S_{M_1-1}, S_{M_1+1}, ..., S_n)$ plainly has Dirichlet distribution

$$D_n(1+\theta_{M_1}, \theta_1, ..., \theta_{M_1-1}, \theta_{M_1+1}, ..., \theta_n).$$

Rescaling, this may be written as $\mathbf{S}_n^{(0)} \to \left(L_1, (1-L_1)\,\mathbf{S}_{n-1}^{(1)}\right)$, where $\mathbf{S}_{n-1}^{(1)} := \left(S_1^{(1)}, ..., S_{M_1-1}^{(1)}, S_{M_1+1}^{(1)}, ..., S_n^{(1)}\right)$ is a new random partition of the unit interval into $n-1$ random fragments.

Given $L_1 \overset{d}{=} B_{M_1;n\setminus M_1} \overset{d}{\sim} \text{beta}\big(1+\theta_{M_1}, \gamma_{n\setminus M_1}\big)$, the conditional joint distribution of the remaining components of $\mathbf{S}_n^{(0)}$ is the same as that of $(1-L_1)\,\mathbf{S}_{n-1}^{(1)}$ where the $(n-1)-$vector $\mathbf{S}_{n-1}^{(1)} \overset{d}{\sim} D_{n-1}(\boldsymbol{\theta}_n \setminus \theta_{M_1})$ has the distribution of a Dirichlet random partition into $n-1$ fragments with parameters $\boldsymbol{\theta}_n \setminus \theta_{M_1}$ (using the stability under scaling property 2.*c.* described in section 2, see also Kingman (1993), Chapter 9). Furthermore, given M_1, $\mathbf{S}_{n-1}^{(1)}$ is independent of $1-L_1$. Pick next at random an interval in $\mathbf{S}_{n-1}^{(1)}$ and call V_2 its length, now with distribution $\text{beta}\big(1+\theta_{M_2}, \gamma_{n\setminus M_1,M_2}\big)$, and iterate until all fragments have been exhausted.

With $V_1 := L_1$, the length of the second fragment by avoiding the first reads $L_2 = (1-V_1)\,V_2$. Iterating, the final SBP of $\mathbf{S}_n$ is $\mathbf{L}_n := (L_1, ..., L_n)$ and we shall put $\mathbf{L}_n = \text{SBP}(\mathbf{S}_n)$. From this construction, we easily get

Lemma 5 *Let* $\mathbf{L}_n = SBP(\mathbf{S}_n)$. *Given* $M_1, ..., M_n$, *assume* $(V_1, ..., V_{n-1})$ *are independent random variables with distribution* $V_k \overset{d}{\sim} beta\big(1+\theta_{M_k}, \gamma_{n\setminus M_1,..,M_k}\big)$, $k = 1, ..., n-1$. *If* $\overline{V}_k := 1 - V_k$, *then,*

$$L_k = \left(\prod_{i=1}^{k-1} \overline{V}_i\right) V_k \ , \ k = 1, ..., n-1, \tag{3.8}$$

$$L_n = 1 - \sum_{k=1}^{n-1} L_k = \prod_{k=1}^{n-1} \overline{V}_i, \tag{3.9}$$

is the conditional RAM representation of the size-biased permutation $\mathbf{L}_n$ *of* $\mathbf{S}_n$.

Note that $\overline{V}_i := 1 - V_i \overset{d}{\sim} \text{beta}\big(\gamma_{n\setminus M_1,..,M_i}, 1+\theta_{M_i}\big)$ and that V_n should be set to 1. The random variables V_k, $k = 1, ..., n-1$ are not stricto sensu

independent, rather they are conditionally independent given $M_1, ..., M_k$. These are well-known construction and properties; see Kingman (1993), Chapter 9.6 and Donnelly and Tavaré (1986).

This conditional RAM representation allows to compute the joint distribution of the size-biased permutation $\mathbf{L}_n$ of $\mathbf{S}_n$. We shall say in the sequel that, if $\mathbf{L}_n = \text{SBP}(\mathbf{S}_n)$, then $\mathbf{L}_n \overset{d}{\sim} SBD_n(\boldsymbol{\theta}_n)$ assuming that $\mathbf{S}_n \overset{d}{\sim} D_n(\boldsymbol{\theta}_n)$.

• *The joint distribution of the size-biased permutation*

The SBP of $\mathbf{S}_n$ is $\mathbf{L}_n$ with $\mathbf{L}_n \overset{d}{\sim} SBD_n(\boldsymbol{\theta}_n)$ and $\mathbf{S}_n \overset{d}{\sim} D_n(\boldsymbol{\theta}_n)$. First, we have

$$(L_1, ..., L_n) = (S_{M_1}, ..., S_{M_n}), \tag{3.10}$$

and consequently

$$\mathbf{P}(L_1 = S_{m_1}, ..., L_n = S_{m_n} \mid \mathbf{S}_n) = \left(\prod_{k=1}^{n-1} \frac{S_{m_k}}{1 - \sum_{l=1}^{k} S_{m_l}} \right) S_{m_n}. \tag{3.11}$$

Consider now the joint moment function of the random size-biased permutation $\mathbf{L}_n = (L_1, ..., L_n)$. The following result holds

Theorem 6 *Given $M_1, ..., M_n$, the joint moment function of the SBP $\mathbf{L}_n = (L_1, ..., L_n) \overset{d}{\sim} SBD_n(\boldsymbol{\theta}_n)$ reads*

$$\mathbf{E}\left[\prod_{k=1}^{n} L_k^{q_k} \mid M_1, ..., M_n \right] = \mathbf{E}\left[\left(\prod_{k=1}^{n-1} \frac{S_{M_k}^{q_k+1}}{1 - \sum_{l=1}^{k} S_{M_l}} \right) S_{M_n}^{q_n+1} \right] \tag{3.12}$$

$$= \prod_{k=1}^{n-1} \frac{\Gamma\left(1 + \gamma_{n\setminus M_1,..,M_{k-1}}\right) \Gamma\left(1 + \theta_{M_k} + q_k\right) \Gamma\left(\gamma_{n\setminus M_1,..,M_k} + q_{k+1} + ... + q_n\right)}{\Gamma\left(1 + \theta_{M_k}\right) \Gamma\left(\gamma_{n\setminus M_1,..,M_k}\right) \Gamma\left(1 + \gamma_{n\setminus M_1,..,M_{k-1}} + q_k + ... + q_n\right)}.$$

Averaging over $M_1, ..., M_n$ whose law is given from Eq. (3.7) by

$$\mathbf{P}(M_1 = m_1, ..., M_n = m_n) = \prod_{k=1}^{n} \frac{\theta_{m_k}}{\gamma_{n\setminus m_1,..,m_{k-1}}},$$

gives the exact joint distribution of $\mathbf{L}_n$.

Proof. Let $V \overset{d}{\sim} \text{beta}(a, b)$. Then, with $\overline{V} := 1 - V$, it holds that

$$\mathbf{E}\left[V^{q_1} \overline{V}^{q_2} \right] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 v^{a+q_1-1} (1-v)^{b+q_2-1} dv \tag{3.13}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+q_1)\Gamma(b+q_2)}{\Gamma(a+b+q_1+q_2)}.$$

Adapting this computation, recalling that, given $M_1, ..., M_k$, we have $V_k \overset{d}{\sim} \text{beta}\big(1+\theta_{M_k}, \gamma_{n\setminus M_1,..,M_k}\big)$, the quantity $\mathbf{E}\left[V_k^{q_k}\overline{V}_k^{q_{k+1}+...+q_{n-1}}\right]$ which appears in $\mathbf{E}\left[\prod_{k=1}^n L_k^{q_k} \mid M_1, ..., M_n\right]$ using Eq. (3.8) has the expression displayed inside the product from Eq. (3.13). □

• *One-dimensional marginals*

From Theorem 6, we get the one-dimensional law of the L_k, $k = 1, ..., n$. Furthermore, one may check that, under some condition, the L_k are arranged in stochastically decreasing order (denoted by $\succeq$). More precisely

Corollary 7 *(i) Given $M_1, ..., M_k$, the law of L_k, for $k = 1, ..., n$, is characterized by*

$$\mathbf{E}\left[L_k^q \mid M_1, ..., M_k\right] = \mathbf{E}\left[V_k^q\right] \prod_{i=1}^{k-1} \mathbf{E}\left[\overline{V}_i^q\right] = \tag{3.14}$$

$$\frac{\Gamma\left(1+\theta_{M_k}+q\right)\Gamma\left(1+\gamma_{n\setminus M_1,..,M_{k-1}}\right)}{\Gamma\left(1+\theta_{M_k}\right)\Gamma\left(1+\gamma_{n\setminus M_1,..,M_{k-1}}+q\right)} \prod_{i=1}^{k-1} \frac{\Gamma\left(\gamma_{n\setminus M_1,..,M_i}+q\right)\Gamma\left(1+\gamma_{n\setminus M_1,..,M_{i-1}}\right)}{\Gamma\left(\gamma_{n\setminus M_1,..,M_i}\right)\Gamma\left(1+\gamma_{n\setminus M_1,..,M_{i-1}}+q\right)}.$$

Averaging over $M_1, ..., M_k$ whose law is given by Eq. (3.7) gives the exact distribution of L_k.

(ii) Let $B_{n\setminus M_1,..,M_{k-1},1} \overset{d}{\sim} beta\big(\gamma_{n\setminus M_1,..,M_{k-1}}, 1\big)$. Assume $M_k > M_{k-1}$ and let

$$C_{n\setminus M_1,..,M_k,1} := B_{n\setminus M_1,..,M_{k-1},1} \cdot B_{M_{k-1},M_k}$$

where $B_{M_{k-1},M_k} \overset{d}{\sim} beta\big(1+\theta_{M_k}, \theta_{M_{k-1}}-\theta_{M_k}\big)$ is independent of $B_{n\setminus M_1,..,M_{k-1},1}$.

Then, given $M_1, ..., M_k$ and $M_k > M_{k-1}$

$$L_k \overset{d}{=} C_{n\setminus M_1,..,M_k,1} \cdot L_{k-1},\ k = 2, ..., n, \tag{3.15}$$

where pairs $C_{n\setminus M_1,..,M_k,1}$ and L_{k-1} are conditionally mutually independent for $k = 2, ..., n$.

(iii) Given $M_k > M_{k-1}$, $L_k \succeq L_{k-1}$, $k = 2, ..., n$.

Proof. (i) is a direct consequence of the construction, since $\overline{V}_i := 1 - V_i \overset{d}{\sim} \text{beta}\big(\gamma_{n\setminus M_1,..,M_i}, 1+\theta_{M_i}\big)$, $i = 1, ..., k-1$, $V_k \overset{d}{\sim} \text{beta}\big(1+\theta_{M_k}, \gamma_{n\setminus M_1,..,M_k}\big)$ are mutually independent, conditionally given $M_1, ..., M_k$. Recalling the expression of the moment function for beta distributions, the corresponding expression of $\mathbf{E}\left[L_k^q \mid M_1, ..., M_k\right]$ follows up.

(ii) Regrouping terms directly from Eq. (3.14), we have $\mathbf{E}\left[L_k^q \mid M_1, ..., M_k\right] = \mathbf{E}\left[L_{k-1}^q \mid M_1, ..., M_{k-1}\right]\mathbf{E}\left[B_k^q \mid M_1, ..., M_k\right]$ with

$$\begin{aligned}\mathbf{E}\left[B_k^q \mid M_1, ..., M_k\right] \;=\; & \frac{\Gamma\left(\gamma_{n\setminus M_1,..,M_{k-1}}+q\right)}{\Gamma\left(\gamma_{n\setminus M_1,..,M_{k-1}}\right)} \frac{\Gamma\left(1+\gamma_{n\setminus M_1,..,M_{k-1}}\right)}{\Gamma\left(1+\gamma_{n\setminus M_1,..,M_{k-1}}+q\right)} \times \\ & \frac{\Gamma\left(1+\theta_{M_k}+q\right)}{\Gamma\left(1+\theta_{M_k}\right)} \frac{\Gamma\left(1+\theta_{M_{k-1}}\right)}{\Gamma\left(1+\theta_{M_{k-1}}+q\right)}.\end{aligned}$$

If $M_k > M_{k-1}$ (an event with probability larger than 1/2), this is the moment function of the product of a beta$\left(\gamma_{n\setminus M_1,..,M_{k-1}}, 1\right)$ distributed random variable times an independent beta$\left(1+\theta_{M_k}, \theta_{M_{k-1}} - \theta_{M_k}\right)$ distributed random variable. This makes sense since, recalling the sequence $\{\theta_m\}$ is decreasing, $\theta_{M_{k-1}} - \theta_{M_k} \geq 0$ (recalling a beta$(\alpha, 0)$ −distributed random variable degenerates to 1).

(iii) Given $M_1, ..., M_k$ and $M_k > M_{k-1}$, we have from (ii): $L_k \succeq L_{k-1}$, $k = 2, ..., n$. Averaging over $M_1, ..., M_k \cap \{M_k > M_{k-1}\}$, $L_k \succeq L_{k-1}$, $k = 2, ..., n$, unconditionally. □

From part (i) of corollary 7, we obtain in particular the average value of L_k. Indeed, putting $q = 1$ in Eq. (3.14) and averaging, using Eq. (3.7), elementary regrouping of terms gives

$$\mathbf{E}(L_k) = \frac{1}{\gamma_n(1+\gamma_n)} \sum_{m_1 \neq ... \neq m_k} \theta_{m_k}(1+\theta_{m_k}) \prod_{i=1}^{k-1} \frac{\theta_{m_i}}{1+\gamma_{n\setminus m_1,..,m_i}} \tag{3.16}$$

$$= \sum_{m=1}^{n} \frac{\theta_m(1+\theta_m)}{\gamma_n(1+\gamma_n)} \sum_{(m_1 \neq ... \neq m_{k-1}) \neq m} \prod_{i=1}^{k-1} \frac{\theta_{m_i}}{1+\gamma_{n\setminus m_1,..,m_i}},$$

for $k = 1, ..., n$. Note, from normalization, that $\sum_{k=1}^{n} \mathbf{E}(L_k) = 1$.

This result is useful in the context of heaps processes for the following reason. Let C_n be the search-cost of an item in the library having reached equilibrium given by $\mathbf{L}_n$. The search cost of an item in a library is the number of items above it in the heap; averaging over the items' popularities gives the search cost of a typical item. From this definition, C_n plainly is the random variable taking the value $k-1$ with probability $\mathbf{E}(L_k)$. From this, we immediately obtain

Theorem 8 *With $\mathbf{E}(L_k)$ given by Eq. (3.16), the exact law of the search-cost C_n within $\mathbf{L}_n$, is given by*

$$\mathbf{P}(C_n = k) = \mathbf{E}(L_{k+1}), \ k = 0, .., n-1. \tag{3.17}$$

4 The Dirichlet-Kingman limit

First, reconsider the symmetric Dirichlet $D_n(\theta)$ distribution, hence with a finite number n of fragments in the partition and when all θ_m are identical. The distribution of $\mathbf{L}_n :=$SBP$(\mathbf{S}_n)$ when $\mathbf{S}_n \overset{d}{\sim} D_n(\theta)$ has been considered in Barrera et al (2005). It was noted $\mathbf{L}_n \overset{d}{\sim} SBD_n(\theta)$ and the above construction of $SBD_n(\boldsymbol{\theta}_n)$ particularizes to $SBD_n(\theta)$ when all θ_m are identical. In the context of the symmetric Dirichlet model, Kingman considered the following limit $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$. Although $\mathbf{S}_n$ itself has a degenerate weak limit when $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$, this situation is worth being considered because many interesting statistical features emerge. This was first noted

by Kingman (1975). Indeed, considering the ordered version $\mathbf{S}_{(n)}$ of $\mathbf{S}_n$ with $\mathbf{S}_{(1)} > ... > \mathbf{S}_{(n)}$, the weak limit of $\mathbf{S}_{(n)}$ is well-defined and well-known to be the Poisson-Dirichlet $PD(\gamma)$ distribution. Furthermore, the weak limit in the sense of Kingman of $\mathbf{L}_n \overset{d}{\sim} SBD_n(\theta)$ is also well-defined and is the Griffiths-Engen-McCloskey, or $GEM(\gamma)$, distribution. The $GEM(\gamma)$ distribution turns out to be also the size-biased permutation of the Poisson-Dirichlet partition (see Kingman (1993), Chapter 9 and Pitman (2002) for additional results). It is of course invariant under an additional action of the size-biased-permutation operation (see Pitman, (1996)).

$GEM(\gamma)-$partitions exhibit many fundamental invariance properties (for a review of these results and applications to Computer Science, Combinatorial Structures, Physics, Biology.., see Tavaré and Ewens (1997) and the references therein for example; this model and related ones are also fundamental in Probability Theory; see Pitman (1996, 1999, 2002).

The purpose of this Section is to consider similar problems in the context of the full asymmetric Dirichlet partition. In this case, we obviously have

Theorem 9 *Assume the parameter set $(\theta_1, ..., \theta_n)$ is such that $\gamma_n := \sum_{m=1}^n \theta_m$ sums to some finite limit $\gamma > 0$ as $n \uparrow \infty$. Then $\mathbf{S}_n \overset{d}{\sim} D_n(\boldsymbol{\theta}_n)$ has a non-degenerate weak limit $\mathbf{S}_\infty$, where $\mathbf{S}_\infty$ is the random partition on the infinite-dimensional simplex characterized by its joint moment function*

$$\mathbf{E}\left[\prod_{m\geq 1} S_m^{q_m}\right] = \frac{\Gamma(\gamma)}{\Gamma\left(\gamma + \sum_{m\geq 1} q_m\right)} \prod_{m\geq 1} \frac{\Gamma(\theta_m + q_m)}{\Gamma(\theta_m)}. \tag{4.1}$$

Proof. The proof follows from the fact that a random vector with bounded support has distribution characterized by its moments and from the continuity of the Euler-gamma function. □

We shall call such a random partition of the interval $\mathbf{S}_\infty$ a Dirichlet-Kingman partition and we shall write $\mathbf{S}_\infty \overset{d}{\sim} DK_\gamma(\boldsymbol{\theta}_\infty)$ where $\boldsymbol{\theta}_\infty := (\theta_1, ..., \theta_m, ...)$ satisfies $\sum_{m\geq 1} \theta_m = \gamma$. Note that for such countable partition $\mathbf{S}_\infty$ of the unit interval, any finite-dimensional distribution $(S_{m_1}, ..., S_{m_k})$ with $m_1 \neq ... \neq m_k \in \mathbb{N}\backslash\{0\}^k$ has Dirichlet distribution $D_k(\theta_{m_1}, ..., \theta_{m_k})$. Although this partition is not explicitly discussed in Kingman (1993), page 93, its construction is tacitly suggested therein.

Corollary 10 *The partition $\mathbf{S}_\infty \overset{d}{\sim} DK_\gamma(\boldsymbol{\theta}_\infty)$ is in the RAM class with*

$$S_k = \left(\prod_{i=1}^{k-1} \overline{B}_i\right) B_k,\ k \geq 1, \tag{4.2}$$

where $(B_k, k \geq 1)$ are independent with respective law $B_k \overset{d}{\sim} beta\left(\theta_k, \gamma_{\infty\backslash 1,..,k}\right)$.

Proof. Using our notation, $\gamma_{\infty\backslash 1,..,k} := \sum_{m\geq 1}\theta_m - \sum_{m=1}^{k}\theta_m = \sum_{m\geq k+1}\theta_m$ and the proof is immediate. □

An example of such a $DK_\gamma(\boldsymbol{\theta}_\infty)$ distribution is when $\theta_m = p^m$, $m \geq 1$, for some $p \in (0,1)$ with $\gamma = p/(1-p)$.

Remark: Dirichlet-Kingman partitions suggest to study the following sequential colonizing process of space by some $n-$species population. Let $B_k \overset{d}{\sim} \text{beta}\big(\theta_k, \gamma_{\infty\backslash 1,..,k}\big)$, $k \geq 1$, be a sequence of independent random variables. Consider the following space-filling process: A first incoming species occupies a random fraction $S_1 = B_1$ of the available unit space and forthcoming species take independent random fractions of the remaining space left by the preceding ones, which is $S_k = \prod_{i=1}^{k-1}\overline{B}_i B_k$, $k \geq 2$. If this space-filling process terminates when each of the n species holds a portion of the space, then a free fraction remains, which is occupied by no species. It is: $S_{n+1} := 1 - \sum_{k=1}^{n} S_k = \prod_{i=1}^{n}\overline{B}_i$.

Clearly, as $n \uparrow \infty$, vacant space S_{n+1} converges almost surely to 0 and so $\mathbf{S}_n := (S_1, ..., S_n)$ converges weakly to the $DK_\gamma(\boldsymbol{\theta}_\infty)$ limit defined by Eq. (4.2).

Note that in the above particular example where $\theta_m = p^m$, $m \geq 1$, although B_k's law depends on k, its mean $\mathbf{E}(B_k) = \theta_k / \sum_{m\geq k}\theta_m = 1 - p$ is independent of the species number k. This is not true in general for other sequences $\boldsymbol{\theta}_\infty$.

From the above construction of $\mathbf{L}_n = \text{SBP}(\mathbf{S}_n)$, we easily get

Corollary 11 *Let $\mathbf{S}_\infty \overset{d}{\sim} DK_\gamma(\boldsymbol{\theta}_\infty)$ with $\sum_{m\geq 1}\theta_m = \gamma$. Then $\mathbf{L}_\infty = SBP(\mathbf{S}_\infty)$ is defined as follows; given the fragment numbers in their visiting order $M_1, ..., M_n$, with distribution*

$$\mathbf{P}(M_1 = m_1, ..., M_n = m_n) = \prod_{k=1}^{n}\frac{\theta_{m_k}}{\gamma_{\infty\backslash m_1,..,m_{k-1}}}, \tag{4.3}$$

assume $(V_1, ..., V_n)$ are independent random variables with distribution $V_k \overset{d}{\sim} beta\big(1+\theta_{M_k}, \gamma_{\infty\backslash M_1,..,M_k}\big)$, $k = 1, ..., n$.

If $\overline{V}_k := 1 - V_k$, then, for each $n \geq 1$

$$L_k = \left(\prod_{i=1}^{k-1}\overline{V}_i\right)V_k \ , \ k = 1, ..., n \tag{4.4}$$

defines the conditional RAM representation of the size-biased permutation $\mathbf{L}_\infty$ of $\mathbf{S}_\infty$.

The next Corollary is an immediate consequence of the above corresponding construction for $D_n(\boldsymbol{\theta}_n)$ in terms of the conditioned Moran subordinator $(\mathcal{X}_t; t \geq 0 \mid \mathcal{X}_\gamma = 1)$.

Corollary 12 *Let* $\mathbf{S}_\infty \overset{d}{\sim} DK_\gamma(\boldsymbol{\theta}_\infty)$ *with* $\sum_{m\geq 1}\theta_m = \gamma$. *With* $\gamma_m := \sum_{k=1}^m \theta_k$, $m \geq 1$, $\gamma_0 := 0$, *if* $(\mathcal{X}_t; t \geq 0)$ *is a Moran subordinator, then*

$$\left(\mathcal{X}_{\gamma_m} - \mathcal{X}_{\gamma_{m-1}}; m \geq 1 \mid \mathcal{X}_\gamma = 1\right) \overset{d}{\sim} DK_\gamma(\boldsymbol{\theta}_\infty).$$

Remark: When computing with Dirichlet-Kingman partitions, the following formula

$$\int_0^\infty \mathbf{E}\left(f(\mathbf{S}_\infty(x))\right) x^{\gamma-1} e^{-px} dx = \frac{\Gamma(\gamma)}{p^\gamma} \mathbf{E}\left(f(\mathbf{X}_\infty(p))\right),$$

extending Eq. (2.5), can be useful.

Acknowledgments: T. Huillet and S. Martínez are indebted for support to ECOS-Sud, to FONDAP in Applied Mathematics and to the Millenium Nucleus in Information and Randomness, Programa Científico Milenio $P01-005$. T. Huillet thanks the C.M.M. for its warm hospitality on a visiting occasion where this work was initiated.

References

[1] Baryshnikov Y., Gnedin A. (2001). Counting intervals in the packing process. *Annals of Applied Probability,* **11**, n° 3, 863–877.

[2] Barrera, J., Huillet, T., Paroissin, C. (2005). Size-biased permutation of Dirichlet partitions and search-cost distribution. *Probability in the Engineering & Informational Sciences*, **19**, n° 1, 83-97.

[3] Bertoin J., Gnedin A. (2004). Asymptotic laws for nonconservative self-similar fragmentations. *Electronic Journal of Probability,* **9**, n° 19, 575–593.

[4] Bertoin, J. (1996). *Lévy processes*. Cambridge University Press, Cambridge.

[5] Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.

[6] Donnelly, P. (1986). Partition structures, Pòlya urns, the Ewens sampling formula and the age of alleles. *Theoretical Population Biology*, **30**, 271-288.

[7] Donnelly, P., Tavaré, S. (1986). The age of alleles and a coalescent. *Advances in Applied Probability*, **18**, 1-19.

[8] Donnelly, P. (1991). The heaps process, libraries and size-biased permutation. *Journal of Applied Probability*, **28**, 321-335.

[9] Engen, S. (1978). *Stochastic abundance models*. Monographs on Applied Probability and Statistics, Chapman and Hall, London.

[10] Ewens, W.J. (1996). Some remarks on the law of succession. Athens Conference on Applied Probability and Time Series Analysis (1995), Vol. **I**, 229–244, *Lecture Notes in Statistics*, **114**, Springer, New York.

[11] Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, **3,** 87-112.

[12] Ewens, W.J. (1990). Population genetics theory - the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory*, Edt. S. Lessard, Kluwer, Dordrecht.

[13] Fabius, J. (1964). Asymptotic behavior of Bayes' estimates. *Annals of Mathematical Statistics*, **35**, 846–856.

[14] Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, **2**, 615–629.

[15] Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Annals of Mathematical Statistics*, **34**, 1386–1403.

[16] Fill, J.A. (1996). Limits and rates of convergence for the distribution of search cost under the move-to-front rule. *Theoretical Computer Science*, **164**, 185–206.

[17] Fill, J. A., Holst, L. (1996). On the distribution of search cost for the move-to-front rule. *Random Structures and Algorithms*, **8**, n° 3, 179–186.

[18] Flajolet, P., Gardy, D., Thimonier, L. (1992). Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, **39**, 207-229.

[19] Haas, A., Formery, P. (2002). Uncertainties in facies proportions estimation I. Theoretical framework: The Dirichlet distribution. *Journal of Mathematical Geology*, **34**, n° 6, 679-702.

[20] Huillet, T., Martinez, S. (2003). Sampling from finite random partitions. *Methodology and Computing in Applied Probability*, **5**, n° 4, 467-492.

[21] Jelenkovìc, P. R. (1999). Asymptotic approximation of the move-to-front search cost distribution and least-recently used caching fault probabilities. *Annals of Applied Probability*, **9**, n° 2, 430–464.

[22] Kingman, J.F.C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society. Series B*, **37**, 1–22.

[23] Kingman, J.F.C. (1978). Random partitions in population genetics. *Proceedings of the Royal Society.* London. Series A, **361**, n° 1704, 1–20.

[24] Kingman, J.F.C. (1993). *Poisson processes.* Clarendon Press, Oxford.

[25] Pitman, J. (1996). Random discrete distributions invariant under size-biased permutation. *Advances in Applied Probability,* **28**, 525-539.

[26] Pitman, J. (1999). Coalescents with multiple collisions. *Annals of Probability*, **27**, n° 4, 1870–1902.

[27] Pitman, J. (2002). Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformation of an interval partition. *Combinatorics, Probability and Computing*, **11**, n° 5, 501–514.

[28] Pitman, J. (2003). *Poisson-Kingman Partitions.* Statistics and Science: A Festschrift for Terry Speed, 1–34, IMS Lecture Notes Monogr. Ser., 40, Inst. Math. Statist., Beachwood, OH.

[29] Steutel, F.W. (1967). Random division of an interval. *Statistica Neerlandica*, **21**, 231-244.

[30] Steutel, F.W. and van Harn, K. (2004). *Infinite Divisibility of Probability Distributions on the Real Line.* Pure and Appl. Math. vol. **259**, Marcel Dekker Pub., New-York, Basel.

[31] Tavaré, S., Ewens, W.J. (1997). Multivariate Ewens distribution. Chapter **41** in *Discrete Multivariate Distributions*, Edts N.L. Johnson, S. Kotz and N. Balakrishnan, pages 232-246, (Wiley, New York).

[32] Vlad, M. O., Tsuchiya, M., Oefner, P., Ross, J. (2002). Bayesian analysis of systems with random chemical composition: Renormalization-group approach to Dirichlet distributions and the statistical theory of dilution. *Physical Review E (3)*, **65**, n° 1, Part 1, 011112, 8 pp.