Electronic Journal of Probability

Vol. 12 (2007), Paper no. 40, pages 1101–1130.

Journal URL
http://www.math.washington.edu/~ejpecp/

Record indices and age-ordered frequencies in Exchangeable Gibbs Partitions

Robert C. Griffiths
Dario Spanò*
University of Oxford
Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, UK
{griff,spano}@stats.ox.ac.uk

Abstract

The frequencies $X_1, X_2, \ldots$ of an exchangeable Gibbs random partition Π of $\mathbb{N} = \{1, 2, \ldots\}$ (Gnedin and Pitman (2005)) are considered in their *age-order*, i.e. their size-biased order. We study their dependence on the sequence $i_1, i_2, \ldots$ of least elements of the blocks of Π. In particular, conditioning on $1 = i_1 < i_2 < \ldots$, a representation is shown to be

$$X_j = \xi_{j-1} \prod_{i=j}^{\infty} (1 - \xi_i) \qquad\qquad j = 1, 2, \ldots$$

where $\{\xi_j : j = 1, 2, \ldots\}$ is a sequence of independent Beta random variables. Sequences with such a product form are called *neutral to the left.* We show that the property of conditional left-neutrality in fact characterizes the Gibbs family among all exchangeable partitions, and leads to further interesting results on: (i) the conditional Mellin transform of X_k, given i_k, and (ii) the conditional distribution of the first k normalized frequencies, given $\sum_{j=1}^{k} X_j$ and i_k; the latter turns out to be a mixture of Dirichlet distributions. Many of the mentioned representations are extensions of Griffiths and Lessard (2005) results on Ewens' partitions.

Key words: Exchangeable Gibbs Partitions, GEM distribution, Age-ordered frequencies,

*Research supported in part by EPSRC grant GR/T21783/01.

Beta-Stacy distribution, Neutral distributions, Record indices..

AMS 2000 Subject Classification: Primary 60C05, 60G57, 62E10.

Submitted to EJP on August 25, 2006, final version accepted July 30, 2007.

1 Introduction.

A random partition of $[n] = \{1, \ldots, n\}$ is a random collection $\Pi_n = \{\Pi_{n1}, \ldots, \Pi_{nK_n}\}$ of disjoint nonempty subsets of $[n]$ whose union is $[n]$. The K_n classes of Π_n, where K_n is a random integer in $[n]$, are conventionally ordered by their least elements $1 = i_1 < i_2 < \ldots < i_{K_n} \leq n$. We call $\{i_j\}$ the sequence of *record indices* of Π_n, and define the *age-ordered frequencies* of Π_n to be the vector $\mathbf{n} = (n_1, \ldots, n_k)$ such that n_j is the cardinality of Π_{nj}. Consistent Markov partitions $\Pi = (\Pi_1, \Pi_2, \ldots)$ can be generated by a set of predictive distributions specifying, for each n, how Π_{n+1} is likely to extend Π_n, that is: given Π_n, a conditional probability is assigned for the integer $(n+1)$ to join any particular class of Π_n or to start a new class.

We consider a family of consistent random partitions studied by Gnedin and Pitman [14] which can be defined by the following prediction rule: (i) set $\Pi_1 = (\{1\})$; (ii) for each $n \geq 1$, conditional on $\Pi_n = (\pi_{n1}, \ldots, \pi_{nk})$, the probability that $(n+1)$ starts a new class is

$$\frac{V_{n+1,k+1}}{V_{n,k}} \tag{1}$$

otherwise, if n_j is the cardinality of π_{nj} $(j = 1, \ldots, k)$, the probability that $(n+1)$ falls in the j-th "old"class π_{nj} is

$$\frac{n_j - \alpha}{n - \alpha k}\left(1 - \frac{V_{n+1,k+1}}{V_{n,k}}\right), \tag{2}$$

for some $\alpha \in (-\infty, 1]$ and a sequence of coefficients $V = (V_{n,k} : k \leq n = 1, 2, \ldots)$ satisfying the recursion:

$$V_{1,1} = 1; \qquad V_{n,k} = (n - \alpha k)V_{n+1,k} + V_{n+1,k+1}. \tag{3}$$

Every partition Π of $\mathbb{N}$ so generated is called an *exchangeable Gibbs partition with parameters* (α, V) (EGP(α, V)), where exchangeable means that, for every n, the distribution of Π_n is a symmetric function of the vector $\mathbf{n} = (n_1, \ldots, n_k)$ of its frequencies ([25]) (see section 2 below). Actually, the whole family of EGPs, treated in [14] includes also the value $\alpha = -\infty$, for which the definition (1)-(2) should be modified; this case will not be treated in the present paper.

A special subfamily of EGPs is *Pitman's two-parameter family*, for which V is given by

$$V_{n,k}^{(\alpha,\theta)} = \frac{\prod_{j=1}^{k}(\theta + \alpha(j-1))}{\theta_{(n)}} \tag{4}$$

where either $\alpha \in [0, 1]$ and $\theta \geq -\alpha$ or $\alpha < 0$ and $\theta = m|\alpha|$ for some integer m. Here and in the following sections, $a_{(x)}$ will denote the generalized increasing factorial i.e. $a_{(x)} = \Gamma(a + x)/\Gamma(a)$, where $\Gamma(\cdot)$ is the Gamma function.

Pitman's family is characterized as the unique class of EGPs with V-coefficients of the form

$$V_{n,k} = \frac{V_k^*}{c_n}$$

for some sequence of constants (c_n) ([14], Corollary 4). If we let $\alpha = 0$ in (4), we obtain the well known *Ewens' partition* for which:

$$V_{n,k}^{(0,\theta)} = \frac{\theta^k}{\theta_{(n)}}. \tag{5}$$

Ewens' family arose in the context of Population Genetics to describe the properties of a population of genes under the so-called infinitely-many-alleles model with parent-independent mutation (see e.g. [32], [21]) and became a paradigm for the modern developments of a theory of exchangeable random partitions ([1], [25], [11]).

For every fixed α, the set of all EGP(α, V) forms a convex set; Gnedin and Pitman proved it and gave a complete description of the extreme points ([14], Theorem 12). It turns out, in particular, that for every $\alpha \leq 0$, the extreme set is given by partitions from Pitman's two-parameter family. For each $\alpha \in (0,1)$, the extreme points are all partitions of the so-called Poisson-Kingman type with parameters (α, s), $s > 0$, whose V-coefficients are given by:

$$V_{n,k}(s) = \alpha^k s^{n/\alpha} G_\alpha(n - \alpha k, s^{-1/\alpha}), \tag{6}$$

with

$$G_\alpha(q,t) := \frac{1}{\Gamma(q) f_\alpha(t)} \int_0^t f_\alpha(t-v) v^{q-1} dv, \tag{7}$$

where f_α is an α-stable density ([29], Theorem 4.5). The partition induced by (6) has limit frequencies (ranked in a decreasing order) equal in distribution to the jump-sizes of the process $(S_t/S_1 : t \in [0,1])$, conditioned on $S_1 = s$, where $(S_t : t > 0)$ is a stable subordinator with density f_α. The parameter s has the interpretation as the (a.s.) limit of the ratio K_n/n^α as $n \to \infty$, where K_n is the number of classes in the partition Π_n generated via $V_{n,k}(s)$. K_n is shown in [14] to play a central role in determining the extreme set of $V_{n,k}$ for every α; the distribution of K_n, for every n, turns out to be of the form

$$\mathbb{P}(K_n = k) = V_{n,k} \begin{bmatrix} n \\ k \end{bmatrix}_\alpha . \tag{8}$$

where $\begin{bmatrix} n \\ k \end{bmatrix}_\alpha$ are generalized Stirling numbers, defined as the coefficients of x^n in

$$\frac{n!}{\alpha^k k!}(1-(1-x)^\alpha)^k$$

(see [14] and reference therein). As $n \to \infty$, K_n behaves differently for different choices of the parameter α: almost surely it will be finite for $\alpha < 0$, $K_n \sim S \log n$ for $\alpha = 0$ and $K_n \sim S n^\alpha$ for positive α, for some positive random variable S.

In this paper we want to study how the distribution of the limit age-ordered frequencies $X_j = \lim_{n\to\infty} n_j/n$ $(j = 1, 2, \ldots)$ in an Exchangeable Gibbs partition depends on its record indices $\mathbf{i} = (1 = i_1 < i_2 < \ldots)$. To this purpose, we adopt a combinatorial approach proposed by Griffiths and Lessard [15] to study the distribution of the age-ordered allele frequencies $X_1, X_2, \ldots$ in a population corresponding to the so-called coalescent process with mutation (see e.g. [32]), whose equilibrium distribution is given by Ewens' partition (5), for some mutation parameter $\theta > 0$. In such a context, the record index i_j has the interpretation as the number of ancestral lineages surviving back in the past, just before the last gene of the j-th oldest type, observed in the current generation, is lost by mutation.

Following Griffiths and Lessard's steps we will (i) find, for every n, the distribution of the age ordered frequencies $\mathbf{n} = (n_1, \ldots, n_k)$, conditional on the record indices $\mathbf{i}_n = (1 = i_1 < i_2 < \ldots <$

i_k) of Π_n, as well as the distribution of $\mathbf{i}_n$; (ii) take their limits as $n \to \infty$; (iii) for $m = 1, 2, \ldots$, describe the distribution of the m-th age-ordered frequency conditional on i_m alone. We will follow such steps, respectively, in sections 3.1, 3.2, 4. In addition, we will derive in section 5 a representation for the distribution of the first k age-ordered frequencies, conditional on their cumulative sum and on i_k.

In our investigation of EGPs, the key result is relative to the step (ii), stated in Proposition 3.2, where we find that, conditional on $\mathbf{i} = (1, i_2, \ldots)$, for every $j = 1, 2, \ldots$,

$$X_j|\mathbf{i} \overset{d}{=} \xi_{j-1} \prod_{i=j}^{\infty} (1 - \xi_i), \tag{9}$$

almost surely, for an independent sequence $(\xi_0, \xi_1, \ldots) \in [0, 1]^\infty$ such that $\xi_0 \equiv 1$ and ξ_m has a Beta density with parameters $(1 - \alpha, i_{m+1} - \alpha m - 1)$ for each $m \geq 1$. The representation (9) does not depend on V. The parameter V affects only the distribution of the record indices $\mathbf{i} = (i_1, i_2, \ldots)$ which is a non-homogenous Markov chain, starting at $i_1 \equiv 1$, with transition probabilities

$$P_j(i_{j+1}|i_j) = (i_j - \alpha j)_{(i_{j+1} - i_j - 1)} \frac{V_{i_{j+1}, j+1}}{V_{i_j, j}}, \qquad j \geq 1. \tag{10}$$

The representation (9)-(10) extends Griffiths and Lessard's result on Ewens' partitions ([15], (29)), recovered just by letting $\alpha = 0$.

In section 3.2 we stress the connection between the representation (9) and a wide class of random discrete distributions, known in the literature of Bayesian nonparametric statistics as Neutral to the Left (NTL) processes ([10], [5]) and use such a connection to show that the structure (9) with independent $\{\xi_j\}$ actually characterizes EGP's among all exchangeable partitions of $\mathbb{N}$.

The representation (9) is useful to find the moments of both X_j and $\sum_{i=1}^{j} X_i$, conditional on the j-th record index i_j alone, as shown in section 4. In the same section a recursive formula is found for the Mellin transform of both random quantities, in terms of the Mellin transform of the size-biased pick X_1.

Finally in section 5 we obtain an expression for the density of the first k age-ordered frequencies $X_1, \ldots, X_k$, conditional on $\sum_{i=1}^{k} X_i$ and i_k, as a mixture of Dirichlet distributions on the $(k - 1)$-dimensional simplex ($k = 1, 2, \ldots$). Such a result leads to a self-contained proof for the marginal distribution of i_k, whose formula is closely related to Gnedin and Pitman's result (8).

As a completion to our results, it should be noticed that a representation of the *unconditional* distribution of the age-ordered frequencies of an EGP can be derived as a mixture of the age-ordered distributions of their extreme points, which are known: for $\alpha \leq 0$, the extreme age-ordered distribution is the celebrated two-parameter GEM distribution ([25],[26]), for which:

$$X_j = B_j \prod_{i=1}^{j-1} (1 - B_j), \quad j = 1, 2, \ldots \tag{11}$$

for a sequence $(B_j : j = 1, 2, \ldots)$ of independent Beta random variables with parameters, respectively $\{(1 - \alpha, \theta + j\alpha) : j = 1, 2, \ldots\}$. Such a representation reflects a property of *right-neutrality*, which in a sense is the inverse of (9), as it will be clear in section 3.2. When α is strictly positive, the structure of the age-ordered frequencies in the extreme points lose such a simple structure. A description is available in [24].

We want to embed Griffiths and Lessard's method in the general setting of Pitman's theory of exchangeable and partially exchangeable random partitions, for which our main reference is [25]. Pitman's theory will be summarized in section 2 . The key role played by record indices in the study of random partitions has been emphasized by several authors, among which Kerov [19], Kerov and Tsilevich [20], and more recently by Gnedin [13], and Nacu [23], who showed that the law of a partially exchangeable random partition is completely determined by that of its record indices. We are indebted to an anonymous referee for signalling the last two references, whose findings have an intrinsic connection with many formulae in our section 2.

2 Exchangeable and partially exchangeable random partitions.

We complete the introductory part with a short review of Pitman's theory of exchangeable and partially exchangeable random partitions, and stress the connection with the distribution of their record indices. For more details we refer the reader to [25] and [29] and reference therein. Let μ be a distribution on $\Delta = \{x = (x_1, x_2, \ldots) \in [0,1]^\infty : |x| \leq 1\}$, endowed with a Borel sigma-field. Consider the function:

$$q_\mu(n_1, \ldots, n_k) = \int_\Delta \left(\prod_{j=1}^{k} p_j^{n_j - 1} \right) \prod_{1}^{k-1} \left(1 - \sum_{i=1}^{j} p_i \right) \mu(dp). \tag{12}$$

The function q_μ is called the *partially exchangeable probability function (PEPF)* of μ, and has the interpretation as the probability distribution of a random partition $\Pi_n = (\Pi_{n1}, \ldots, \Pi_{nK_n})$, for which a sufficient statistic is given by its age-ordered frequencies, that is:

$$\mathbb{P}_\mu(\Pi_n = (\pi_{n1}, \ldots, \pi_{nk})) = q_\mu(n_1, \ldots, n_k)$$

for every partition $(\pi_{n1}, \ldots, \pi_{nk})$ such that $|\pi_{nj}| = n_j$ $(j = 1, \ldots, k \leq n)$.

If $q_\mu(n_1, \ldots, n_k)$ is symmetric with respect to permutations of its arguments, it is called an *exchangeable partition probability function (EPPF)*, and the corresponding partition Π_n an *exchangeable random partition.*

For exchangeable Gibbs partitions, the EPPF is, for $\alpha \in (-\infty, 1]$,

$$q_{\alpha,V}(n_1, \ldots, n_k) = V_{n,k} \prod_{j=1}^{k} (1-\alpha)_{(n_j - 1)}, \tag{13}$$

with $(V_{n,k})$ defined as in (3). This can be obtained by repeated application of (1)-(2)-(3).

A minimal sufficient statistic for an exchangeable Π_n is given, because of the symmetry of its EPPF, by its *unordered frequencies* (i.e. the count of how many frequencies in Π_n are equal to 1, ..., to n), whose distribution is given by their (unordered) sampling formula:

$$\tilde{\mu}(\mathbf{n}) = \binom{|\mathbf{n}|}{\mathbf{n}} \frac{1}{\prod_1^n b_i!} q_\mu(\mathbf{n}), \tag{14}$$

where

$$\binom{|\mathbf{n}|}{\mathbf{n}} = \frac{|\mathbf{n}|!}{\prod_{j=1}^{k} n_j!}$$

and b_i is the number of n_j's in $\mathbf{n}$ equal to i $(i = 1, \ldots, n)$.

It is easy to see that for a Ewens' partition (whose EPPF is (13) with $\alpha = 0$ and V given by (5)), formula (14) returns the celebrated *Ewens' sampling formula.*

The distribution of the age-ordered frequencies

$$\bar{\mu}(\mathbf{n}) = \binom{|\mathbf{n}|}{\mathbf{n}} a(\mathbf{n}) q_\mu(\mathbf{n}), \tag{15}$$

differs from (14) only by a counting factor, where

$$a(\mathbf{n}) = \prod_{j=1}^{k} \frac{n_j}{n - \sum_{i=1}^{j-1} n_i} \tag{16}$$

is the distribution of the size-biased permutation of $\mathbf{n}$.

If $\Pi = (\Pi_n)$ is a (partially) exchangeable partition with PEPF q_μ, then the vector $n^{-1}(n_1, n_2, \ldots)$ of the relative frequencies, in age-order, of Π_n, converges a.s. to a random point $P = (P_1, P_2, \ldots) \in \Delta$ with distribution μ: thus the integrand in (12) has the interpretation as the conditional PEPF of a partially exchangeable random partition, given its limit age-ordered frequencies $(p_1, p_2, \ldots)$. If q_μ is an EPPF, then the measure $d\mu$ is *invariant under size-biased permutation.*

The notion of PEPF gives a generalized version of Hoppe's urn scheme, i.e. a predictive distribution for (the age-ordered frequencies of) Π_{n+1}, given (those of) Π_n. In an urn of Hoppe's type there are colored balls and a black ball. Every time we draw a black ball, we return it in the urn with a new ball of a new distinct color. Otherwise, we add in the urn a ball of the same color as the ball just drawn.

Pitman's extended urn scheme works as follows. Let q be a PEPF, and assume that initially in the urn there is only the black ball. Label with j the j-th distinct color appearing in the sample. After $n \geq 1$ samples, suppose we have put in the urn $\mathbf{n} = (n_1, \ldots, n_k)$ balls of colors $1, \ldots, k$, respectively, with colors labeled by their order of appearance. The probability that the next ball is of color j is

$$\mathbb{P}(\mathbf{n} + \mathbf{e}_j | \mathbf{n}) = \frac{q(\mathbf{n} + \mathbf{e}_j)}{q(\mathbf{n})} \mathbb{I}(j \leq k) + \left(1 - \sum_{j=1}^{k} \frac{q(\mathbf{n} + \mathbf{e}_j)}{q(\mathbf{n})}\right) \mathbb{I}(j = k+1), \tag{17}$$

where $\mathbf{e}_j = (\delta_{ij} : i = 1, \ldots, k)$ and δ_{xy} is the Kronecker delta. The event $(j = k + 1)$, in the last term of the right-hand side of (17), corresponds to a new distinct color being added to the urn.

The predictive distribution of a Gibbs partition is obtained from its EPPF by substituting (13) into (17):

$$\mathbb{P}(\mathbf{n}+\mathbf{e}_j|\mathbf{n}) = \frac{n_j-\alpha}{n-\alpha k}\left(1-\frac{V_{n+1,k+1}}{V_{n,k}}\right)\mathbb{I}(j\leq k) + \frac{V_{n+1,k+1}}{V_{n,k}}\mathbb{I}(j=k+1), \tag{18}$$

which gives back our definition (1)-(2) of an EGP.

The use of an urn scheme of the form (1)-(2) in Population Genetics is due to Hoppe [16] in the context of Ewens' partitions (infinitely-many-alleles model), for which the connection between order of appearance in a sample and age-order of alleles is shown by [8]. In [10] an extended version of Hoppe's approach is suggested for more complicated, still exchangeable population models (where e.g. mutation can be recurrent). Outside Population Genetics, the use of (1)-(2) for generating trees leading to Pitman's two-parameter GEM frequencies, can be found in the literature of random recursive trees (see e.g. [7]). Urn schemes of the form (17) are a most natural tool to express one's *a priori* opinions in a Bayesian statistical context, as pointed out by [33] and [27]. Examples of recent applications of Exchangeable Gibbs partitions in Bayesian nonparametric statistics are in [22], [17]. The connection between (not necessarily infinite) Gibbs partitions and coagulation-fragmentation processes is explored by [3] (see also [2] and reference therein).

2.1 Distribution of record indices in partially exchangeable partitions.

Let Π be a partially exchangeable random partition. Since, for every n, its collection of age-ordered frequencies $\mathbf{n}=(n_1,\ldots,n_k)$ is a sufficient statistic for Π_n, all realizations π_n with the same $\mathbf{n}$ *and* the same record indices $\mathbf{i}_n = 1 < i_2 < \ldots < i_k \leq n$ must have equal probability. To evaluate the joint probability of the pair $(\mathbf{n},\mathbf{i}_n)$, we only need to replace $a(\mathbf{n})$ in (15) by an appropriate counting factor. This is equal to the number of arrangements of n balls, labelled from 1 to n, in k boxes with the constraint that exactly n_j balls fall in the same box as the ball i_j. Such a number was shown by [15] to be equal to

$$\binom{|\mathbf{n}|}{\mathbf{n}} a(\mathbf{n},\mathbf{i}_n)$$

where

$$\binom{|\mathbf{n}|}{\mathbf{n}} = \frac{n!}{\prod_{j=1}^k n_j!} \quad \text{and} \quad a(\mathbf{n},\mathbf{i}_n) = \frac{\prod_{j=1}^k \binom{S_j - i_j}{n_j - 1}}{\binom{|\mathbf{n}|}{\mathbf{n}}}$$

with $S_j := \sum_{i=1}^j n_i$. Thus, if $\Pi = (\Pi_n)$ is a partially exchangeable random partition with PEPF q_n, then the joint probability of age-ordered frequencies and record indices is

$$\bar{\mu}(\mathbf{n},\mathbf{i}_n) = \binom{|\mathbf{n}|}{\mathbf{n}} a(\mathbf{n},\mathbf{i}_n) q_\mu(\mathbf{n}). \tag{19}$$

The distribution of the record indices can be easily derived by marginalizing:

$$\bar{\mu}(\mathbf{i}_n) = \sum_{\mathbf{n}\in B(\mathbf{i}_n)} \binom{|\mathbf{n}|}{\mathbf{n}} a(\mathbf{n},\mathbf{i}_n) q_\mu(\mathbf{n}). \tag{20}$$

where

$$B_n(\mathbf{i}_n) = \{(n_1,\ldots,n_k) : \sum_{i=1}^k n_i = n;\ S_{j-1} \geq i_j - 1, j = 1,\ldots,k\}$$

is the set of all possible $\mathbf{n}$ compatible with $\mathbf{i}$. In [15] such a formula is derived for the particular case of Ewens' partitions. For general random partitions see also [23], section 2.

Notice that, for every $\mathbf{n}$ such that $|\mathbf{n}| = n$,

$$a(\mathbf{n}) = \sum_{\mathbf{i}_n \in C(\mathbf{n})} a(\mathbf{n}, \mathbf{i}_n) = \prod_{j=1}^{k} \frac{n_j}{n - \sum_{i=1}^{j-1} n_i},$$

where

$$C(\mathbf{n}) = \{(1 < i_2 < \ldots < i_k \le n) : k \le n,\ i_j \le S_{j-1} + 1\}$$

is the set of all possible $\mathbf{i}_n$ compatible with $\mathbf{n}$. Then the marginal distribution of the age-ordered frequencies (15) is recovered by summing (19) over $C(\mathbf{n})$.

This observation incidentally links a classical combinatorial result to partially exchangeable random partitions.

Proposition 2.1. *Let $\Pi = (\Pi_n)$ be a partially exchangeable partition with PEPF q_μ.*
(i) Given the frequencies $\mathbf{n} = (n_1, \ldots, n_k)$ in age-order, the probability that the least elements of the classes of Π_n are $\mathbf{i}_n = (i_1, \ldots, i_k)$, does not depend on q_μ and is given by

$$\mathbb{P}(\mathbf{i}_n|\mathbf{n}) = \frac{1}{n!} \prod_{j=1}^{k} \frac{(S_j - i_j)!}{(S_{j-1} - i_j + 1)!} (n - S_{j-1}). \tag{21}$$

(ii) Let $W_j = \lim_{n\to\infty} S_j/n$. Conditional on $\{W_j : j = 1, 2, \ldots\}$, the waiting times

$$T_j = i_j - i_{j-1} - 1 \qquad\qquad (j = 2, 3, \ldots)$$

are independent geometric random variables, each with parameter $(1 - W_{j-1})$, respectively.

Proof. Part (i) can be obtained by a manipulation of a standard result on uniform random permutations of $[n]$. Part two can be proved by using a representation theorem due to Pitman ([25], Theorem 8). We prefer to give a direct proof of both parts to make clear their connection. Simply notice that, for every $\mathbf{n}$ and $\mathbf{i}$, the right-hand side of (21) is equal to $a(\mathbf{n}, \mathbf{i}_n)/a(\mathbf{n})$. Then, for every $\mathbf{n}$,

$$\sum_{\mathbf{i}_n} \mathbb{P}(\mathbf{i}_n|\mathbf{n}) = \sum_{\mathbf{i}_n \in C(\mathbf{n})} \frac{a(\mathbf{n}, \mathbf{i}_n)}{a(\mathbf{n})} = 1$$

and

$$\sum_{\mathbf{n}} \sum_{C(\mathbf{n})} \mathbb{P}(\mathbf{i}_n|\mathbf{n}) \bar{\mu}(\mathbf{n}) = 1,$$

where $\bar{\mu}(\mathbf{n})$ is as in (15), hence $\mathbb{P}(\mathbf{i}_n|\mathbf{n})$ is a regular conditional probability and (i) is proved. Now, consider the set

$$C_{[i_l, l]}(\mathbf{n}) := \{i_l < i_{l+1} < \ldots < i_k \le n : i_j \le S_{j-1} + 1,\ j = l+1, \ldots, k\}.$$

Also define, for $j = 1, \ldots, k - l + 1$, $n_j^* := S_j^* - S_{j-1}^*$ with $S_j^* := S_{j+l-1} - i_l + 1$, and $i_j^* := i_{j+l-1} - i_l + 1$. Then $C_{[i_l,l]}(\mathbf{n}) = C(\mathbf{n}^*)$ so that, for a fixed $l \le k$, the conditional probability of

$i_2, \ldots, i_l$, given $\mathbf{n} = (n_1, \ldots, n_k)$, is

$$\begin{aligned}\mathbb{P}(i_2, \ldots, i_l|\mathbf{n}) &= \frac{1}{n!} \sum_{C_{[i_l,l]}(\mathbf{n})} \prod_{j=1}^{k} \frac{(S_j - i_j)!}{(S_{j-1} - i_j + 1)!}(n - S_{j-1}) \\ &= \frac{(n-i_l)!}{n!} \left[\prod_{j=1}^{l-1} \frac{(S_j - i_j)!}{(S_{j-1} - i_j + 1)!}(n - S_{j-1}) \right] \frac{(n - S_{l-1})}{(S_{l-1} - i_l + 1)!} \\ &\quad \times \sum_{C(\mathbf{n}^*)} \frac{a(\mathbf{n}^*, \mathbf{i}^*_{n^*})}{a(\mathbf{n}^*)}, \end{aligned} \tag{22}$$

where $n^* = n - i_l + 1$. The sum in (22) is 1; multiply and divide the remaining part by $[S_l^{l-1}(S_l - i_l)!]/(S_l - 1)!$. The probability can therefore be rewritten as

$$\mathbb{P}(i_2, \ldots, i_l|\mathbf{n}) = \frac{(S_l - 1)_{[i_l - 1]}}{(n-1)_{[i_l-1]}} \left(\frac{S_l}{n}\right)^{-(l-1)} \prod_{j=1}^{l} \left(1 - \frac{S_{j-1}}{n}\right) \left[\frac{S_l^{l-1}}{(S_l - 1)!} \prod_{j=1}^{l} \frac{(S_j - i_j)!}{(S_{j-1} - i_j + 1)!} \right],$$

where $a_{[r]} = a(a-1)\cdots(a-r+1)$ is the falling factorial. Now, define

$$(W_j : j = 1, 2, \ldots) = \lim_{n\to\infty} (S_j/n : j = 1, 2, \ldots);$$

then, for l fixed,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(i_2, \ldots, i_l|\mathbf{n}) &= W_l^{i_l - l} \prod_{j=1}^{l} (1 - W_{j-1}) \left[\prod_{j=2}^{l} \left(\frac{W_{j-1}}{W_l}\right)^{i_j - i_{j-1} - 1} \right] \\ &= \prod_{j=2}^{l} W_{j-1}^{i_j - i_{j-1} - 1}(1 - W_{j-1}),\end{aligned}$$

which is the distribution of $k-1$ independent geometric random variables, each with parameter $(1 - W_j)$, and the proof is complete. □

By combining the definition (12) of PEPF and Proposition 2.1, one recovers an identity due to Nacu ([23], (7)).

Corollary 2.1. ([23], Proposition 5) *For every sequence* $1 = i_1 < i_2 < \ldots < i_k + 1 = n$, *and every point* $p = (p_1, p_2, \ldots) \in \Delta$,

$$\prod_{j=1}^{k-1} w_j^{i_{j+1} - i_j - 1} = \sum_{B(\mathbf{i})} \binom{|\mathbf{n}|}{\mathbf{n}} a(\mathbf{i}, \mathbf{n}) \prod_{j=1}^{k} p_j^{n_j - 1} \tag{23}$$

where $w_j = \sum_{i=1}^{j} p_i$ $(j = 1, 2, \ldots)$.

Proof. Multiply both sizes by $\prod_1^k (1 - w_{j-1})$: by Proposition 2.1 and (12), formula (23) is just the equality (20) with the choice $d\mu = \delta_p$. □

3 Age-ordered frequencies conditional on the record indices in Exchangeable Gibbs partitions

3.1 Conditional distribution of sample frequencies.

From now on we will focus only on EGP(α, V). We have seen that the conditional distribution of the record indices, given the age-ordered frequencies of a partially exchangeable random partition, is purely combinatorial as it does not depend on its PEPF. We will now find the conditional distribution of the age-ordered frequencies $\mathbf{n}$ given the record indices, i.e. the step (i) of the plan outlined in the introduction. We show that such a distribution does not depend on the parameter V, which in fact affects only the marginal distribution of $\mathbf{i}_n$, as explained in the following Lemma.

Lemma 3.1. *Let $\Pi = (\Pi_n)$ be an EGP(α, V), for some $\alpha \in (-\infty, 1)$ and $V = (V_{n,k} : k \leq n = 1, 2, \ldots)$. For each n, the probability that the record indices in Π_n are $\mathbf{i}_n = (i_1, \ldots, i_k)$ is*

$$\bar{\mu}_{\alpha,V}(\mathbf{i}_n) = \psi^{-1}_{\alpha,n}(\mathbf{i}_n) V_{n,k}. \tag{24}$$

where

$$\psi_{\alpha,n}(\mathbf{i}_n) = \frac{\Gamma(1-\alpha)}{\Gamma(n-\alpha k)} \prod_{2}^{k} \frac{\Gamma(i_j - j\alpha)}{\Gamma(i_j - j\alpha - (1-\alpha))}. \tag{25}$$

The sequence $i_1, i_2, \ldots$ forms a non-homogeneous Markov chain starting at $i_1 = 1$ and with transition probabilities given by

$$P_j(i_{j+1}|i_j) = (i_j - \alpha j)_{(i_{j+1}-i_j-1)} \frac{V_{i_{j+1},j+1}}{V_{i_j,j}}, \qquad j \geq 1. \tag{26}$$

Proof. The proof can be carried out by using the urn scheme (18). For every n, let K_n be the number of distinct colors which appeared before the $n+1$-th ball was picked. From (18), the sequence $(K_n : n \geq 1)$ starts from $K_1 = 1$ and obeys, for every n, the prediction rule:

$$\mathbb{P}(K_{n+1} = k_{n+1}|K_n = k) = \left(1 - \frac{V_{n+1,k+1}}{V_{n,k}}\right) \mathbb{I}(k_{n+1} = k) + \frac{V_{n+1,k+1}}{V_{n,k}} \mathbb{I}(k_{n+1} = k+1). \tag{27}$$

By definition, K_n jumps at points $1 < i_2 < \ldots$, due to the equivalence

$$\{K_{n+1} = K_n + 1 | K_n = k\} = \{i_{k+1} = n+1\}.$$

Thus, for every $n, k \leq n$ every sequence $\mathbf{i}_n = (1 = i_1 < \ldots < i_k \leq n)$ corresponds to a sequence $(k_1, \ldots, k_n)$ such that $k_{i_j} = j,\ j = 1, \ldots, k$ and $k_m = k_{m-1}\ \forall m \in [n] : m \notin \mathbf{i}_n$. From (27),

$$\begin{aligned} \bar{\mu}_{\alpha,V}(\mathbf{i}_n) &= \prod_{j=1}^{k} \frac{V_{i_j,j}}{V_{i_{j-1},j-1}} \prod_{l \notin \mathbf{i}_n} (l - 1 - k_{l-1}\alpha) \frac{V_{l,k_{l-1}}}{V_{l-1,k_{l-1}}} \\ &= V_{n,k} \prod_{l \notin \mathbf{i}_n} (l - 1 - k_{l-1}\alpha). \end{aligned} \tag{28}$$

The last product in (28) is equal to

$$\begin{aligned}\prod_{l\notin \mathbf{i}_n}(l-k_{l-1}\alpha-1) &= \frac{\Gamma(n-\alpha k)}{\Gamma(1-\alpha)}\prod_{j=2}^{k}\frac{\Gamma(i_j-j\alpha-(1-\alpha))}{\Gamma(i_j-j\alpha)}. \\ &= \psi_{\alpha,n}^{-1}(\mathbf{i}_n). \end{aligned} \tag{29}$$

and this proves (24). The second part of the Lemma (i.e. the transition probabilities (26)) follow immediately just by replacing, in (24), n with i_k, for every k, to show that

$$\bar{\mu}_{\alpha,V}(\mathbf{i}_{i_k}) = \prod_{j=1}^{k-1} P_j(i_{j+1}|i_j),$$

for P_j satisfying (26) for every j. □

The distribution of the age-ordered frequencies in an EGP Π_n, conditional on the record indices, can be easily obtained from Lemma 3.1 and (19).

Proposition 3.1. *Let $\Pi = (\Pi_n)$ be an EGP(α, V), for some $\alpha \in (-\infty, 1)$ and $V = (V_{n,k} : k \leq n = 1, 2, \ldots)$. For each n, the conditional distribution of the sample frequencies $\mathbf{n}$ in age-order, given the vector $\mathbf{i}_n$ of indices, is independent of V and is equal to*

$$\bar{\mu}_\alpha(\mathbf{n}|\mathbf{i}_n) = \psi_{\alpha,n}(\mathbf{i}_n)\left(\prod_{j=1}^{k}\binom{S_j-i_j}{n_j-1}(1-\alpha)_{(n_j-1)}\right). \tag{30}$$

Remark 3.1. *Notice that, as $\alpha \to 0$, formula (30) reduces to:*

$$\bar{\mu}_0(\mathbf{n}|\mathbf{i}_n) = \frac{\prod_{l=2}^{k}(i_l-1)}{(n-1)!}\prod_{j=1}^{k}\frac{(S_j-i_j)!}{(S_{j-1}+i_j-1)!}.$$

This is known as the law of cycle partitions of a permutation given the minimal elements of cycles, derived in the context of Ewens' partitions in [15].

Proof. Recall that the probability of a pair $(\mathbf{n}, \mathbf{i}_n)$ is given by

$$\bar{\mu}_{\alpha,V}(\mathbf{n}, \mathbf{i}_n) = \binom{|\mathbf{n}|}{\mathbf{n}} a(\mathbf{n}, \mathbf{i}_n) q_{\alpha,V}(\mathbf{n}). \tag{31}$$

Now it is easy to derive the conditional distribution of a configuration given a sequence $\mathbf{i}_n$, as:

$$
\begin{aligned}
\bar{\mu}_\alpha(\mathbf{n}|\mathbf{i}_n) &= \frac{\bar{\mu}_{\alpha,V}(\mathbf{n},\mathbf{i}_n)}{\bar{\mu}_{\alpha,V}(\mathbf{i}_n)} \\
&= \binom{|\mathbf{n}|}{\mathbf{n}} \left(\frac{\prod_{j=1}^k \binom{S_j - i_j}{n_j - 1}(1-\alpha)_{(n_j-1)}}{\binom{|\mathbf{n}|}{\mathbf{n}}} \right) \frac{V_{n,k}}{\bar{\mu}_{\alpha,V}(\mathbf{i})} \\
&= \psi_{\alpha,n}(\mathbf{i}_n) \left(\prod_{j=1}^k \binom{S_j - i_j}{n_j - 1} (1-\alpha)_{(n_j-1)} \right), \qquad (32)
\end{aligned}
$$

and the proof is complete. □

3.2 The distribution of the limit frequencies given the record indices.

We now have all elements to derive a representation for the limit relative frequencies in age-order, conditional on the limit sequence of record indices $\mathbf{i} = (i_1 < i_2 < \ldots)$ generated by an EGP(α, V).

Proposition 3.2. *Let $\Pi = (\Pi_n)_{n \geq 1}$ be an Exchangeable Gibbs Partition with index $\alpha > -\infty$ for some V. Let $\mathbf{i} = (i_1 < i_2 < \ldots)$ be its limit sequence of record indices and $X_1, X_2, \ldots$ be the age-ordered limit frequencies as $n \to \infty$.*

A regular conditional distribution of $X_1, X_2, \ldots$ given the record indices is given by

$$
X_j \overset{d}{=} \xi_{j-1} \prod_{m=j}^{\infty} (1 - \xi_m), \qquad j \geq 1, \qquad (33)
$$

a.s., where $\xi_0 \equiv 1$ and, for $j \geq 1$, ξ_j is a Beta random variable in $[0,1]$ with parameters $(1-\alpha, i_{j+1} - j\alpha - 1)$.

Remark 3.2. *Proposition 3.2 is a statement about a regular conditional distribution. The question about the existence of a limit conditional distribution of $X|\mathbf{i}$ as a function of $\mathbf{i} = \lim_n \mathbf{i}_n$ has different answer according to the choice of α, as a consequence of the limit behavior of K_n, the number of blocks of an EGP Π_n, as recalled in the introduction. For $\alpha < 0$, $\mathbf{i}$ is almost surely a finite sequence; for nonnegative α, the length k of $\mathbf{i}$ will be a.s. either $k \sim s \log n$ (for $\alpha = 0$) or $k \sim s n^\alpha$ (for $\alpha > 0$), for some $s \in [0, \infty]$. The infinite product representation (33) still holds in any case if we adopt the convention $i_k \equiv \infty$ for every $k > K_\infty$ where $K_\infty := \lim_{n\to\infty} K_n$.*

Proof. The form (30) of the conditional density $\mu_\alpha(\mathbf{n}|\mathbf{i}_n)$ implies

$$
\sum_{|\mathbf{n}|=n} \prod_{j=1}^{k} \binom{S_j - i_j}{n_j - 1} (1-\alpha)_{(n_j-1)} = \psi^{-1}_{\alpha,n}(\mathbf{i}_n). \qquad (34)
$$

For some $r < k$, let $a_2, \ldots, a_r$ be positive integers and set $a_1 = 0$ and $a_{r+1} = \ldots = a_k = 0$. Define $\mathbf{i}'_n = (i'_1, \ldots, i'_k)$ where

$$i'_j = i_j + \sum_1^j a_i \qquad (j = 1, \ldots, k).$$

Now take the sum (34) with $\mathbf{i}_n$ replaced by $\mathbf{i}'_n$, and multiply it by $\psi_{\alpha,n}(\mathbf{i}_n)$. We obtain

$$\frac{\psi_{\alpha,n}(\mathbf{i}_n)}{\psi_{\alpha,n}(\mathbf{i}'_n)} = \mathbb{E}\left(\prod_{j=1}^k \frac{(S_j - i'_j)!}{(S_j - i_j)!} \frac{(S_{j-1} - i_j + 1)!}{(S_{j-1} - i'_j + 1)!}\right) \tag{35}$$

where the expectation is taken with respect to $\bar{\mu}_\alpha(\cdot|\mathbf{i}_n)$. The left hand side of (35) is

$$\begin{aligned}\frac{\psi_{\alpha,n}(\mathbf{i}_n)}{\psi_{\alpha,n}(\mathbf{i}'_n)} &= \prod_{j=2}^k \frac{[i_j - j\alpha - (1-\alpha)]_{(\sum_{i=1}^j a_i)}}{(i_j - j\alpha)_{(\sum_{i=1}^j a_i)}} \\ &= \prod_{j=1}^{k-1} \mathbb{E}((1-\xi_j)^{\sum_{i=1}^j a_i}), \end{aligned} \tag{36}$$

where $\xi_1, \ldots, \xi_{k-1}$ are independent Beta random variables, each with parameters $(1-\alpha, i_{j+1} - j\alpha - 1)$.

Let $b_j = \sum_{i=1}^j a_i$. On the right hand side of (35), $S_0 = 0, S_k = n$, so the product is equal to

$$\prod_{j=1}^k \left(\frac{S_{j-1}}{S_j}\right)^{b_j} \left[\prod_{l=0}^{b_j - 1} \left(\frac{1 - \frac{i_j + l}{S_j}}{1 - \frac{i_j - 1 + l}{S_{j-1}}}\right)\right] \tag{37}$$

Since $a_j = 0$ for $j = 1$ and $j > r$, as $k, n \to \infty$ the product inside square brackets converges to 1 so the limit of (37) is

$$\prod_{j=1}^{r-1} W_j^{a_{j+1}}$$

where $W_j = \lim_{n\to\infty} S_j/n$. Hence from (35) it follows that in the limit

$$\mathbb{E}\left(\prod_{j=1}^{r-1} W_j^{a_j}\right) = \prod_{j=2}^\infty E((1-\xi_j)^{\sum_{i=1}^j a_i})$$

which gives the limit distribution of the cumulative sums:

$$W_j \overset{d}{=} \prod_{i=j}^\infty (1-\xi_i), \qquad j = 1, 2, \ldots$$

But

$$\begin{aligned} X_j &= W_j - W_{j-1} \\ &\overset{d}{=} \prod_{i=j}^{\infty}(1-\xi_j) - \prod_{i=j-1}^{\infty}(1-\xi_j) \\ &= \xi_{j-1}\prod_{i=j}^{\infty}(1-\xi_i), \qquad\qquad j = 1, 2, \dots \end{aligned}$$

and the proof is complete. □

3.3 Conditional Gibbs frequencies, Neutral distributions and invariance under size-biased permutation.

Proposition 3.2 says that, conditional on all the record indices $i_1, i_2, \dots$, the sequence of relative increments of an EGP(α, V)

$$\xi = \left(\frac{X_2}{W_2}, \frac{X_3}{W_3}, \dots\right). \tag{38}$$

is a sequence of independent coordinates. In fact, such a process can be interpreted as the negative, time-reversed version of a so-called *Beta-Stacy* process, a particular class of random discrete distributions, introduced in the context of Bayesian nonparametric statistics as a useful tool to make inference for right-censored data (see [30], [31] for a modern account).

It is possible to show that such an independence property of the ξ sequence (conditional on the indices) actually characterizes the family of EGP partitions. To make clear such a statement we recall a concept of *neutrality* for random $[0,1]$-valued sequences, introduced by Connor and Mosimann [4] in 1962 and refined in 1974 by Doksum [6] in the context of nonparametric inference and, more recently, by Walker and Muliere [31].

Definition 3.1. *Let k be any fixed positive integer (non necessarily finite).*

(i) Let $P = (P_1, P_2, \dots, P_k)$ be a random point in $[0,1]^k$ such that $|P| = \sum_{i=1}^k P_i \leq 1$ and, for every $j = 1, \dots, k-1$ denote $F_j = \sum_{i=1}^j P_i$. P is called a Neutral to the Right (NTR) *sequence if the vector $(B_j : 1 \leq j < k)$ of relative increments*

$$B_j = \frac{P_j}{1 - F_{j-1}} \qquad\qquad j < k$$

is a sequence of independent random variables in $[0,1]$.

Let (α, β) be a point in $[0,\infty]^{k-1}$ (or in $[0,\infty]^\infty$ if $k = \infty$). A NTR vector P such that $|P| = 1$ almost surely and, for every $j < k$, every increment B_j is a Beta (α_j, β_j), is called a Beta-Stacy distribution *with parameter (α, β).*

(ii) A Neutral to the left (NTL) *vector $P = (P_1, P_2, \dots, P_k)$ is a vector such that $P^* := (P_k, P_{k-1}, \dots, P_1)$ is NTR.*

A Left-Beta-Stacy *distribution is a NTL vector P such that P^* is Beta-Stacy.*

A known result due to [26] is that the only class of exchangeable random partitions whose limit age-ordered frequencies are (unconditionally) a NTR distribution, is Pitman's two-parameter family, i.e. the EGP(α, V) with V-coefficients given by (4). In this case, the age-ordered frequencies follow the so-called two-parameter GEM distribution, a special case of Beta-Stacy distribution with each B_j being a Beta$(1-\alpha, \theta + j\alpha)$ random variable.
The age-ordered frequencies of all other Gibbs partitions are not NTR; on the other side, Proposition 3.2 shows that, conditional on the record indices $i_1, i_2, \ldots$, and on W_k they are all NTL distributions. For a fixed k set

$$Y_j = \frac{X_{k-j+1}}{W_k}, \qquad 1 \le j \le k.$$

Then

$$1 - F_j = \frac{W_{k-j}}{W_k}$$

and

$$\xi_{k-j} = \frac{Y_j}{1 - F_{j-1}}. \tag{39}$$

By construction, the sequence $Y_1, \ldots, Y_k$ is a Beta-Stacy sequence with parameters $\alpha_{k,j} = 1 - \alpha$ and $\beta_{k,j} = i_{k-j+1} - (k-j+1)\alpha - (1-\alpha)$ $(j = 1, \ldots, k)$. The property of (conditional) left-neutrality is maintained as $k \to \infty$ (just condition on $W_{K_\infty} = 1$ where $K_\infty = \lim_{n\to\infty} K_n$).

The following proposition is a converse of Proposition 3.2.

Proposition 3.3. *Let $X = (X_1, X_2, \ldots) \in \Delta$ be the age-ordered frequencies of an infinite exchangeable random partition Π of $\mathbb{N}$. Assume, conditionally on the record indices of Π, X is a NTL sequence. Then Π is an exchangeable Gibbs partition for some parameters (α, V).*

Proof. The frequencies of an exchangeable random partition of $\mathbb{N}$ are in age-order if and only if their distribution is *invariant under size-biased permutation* (ISBP, see [9], [26], [12]). To prove the proposition, we combine two known results: the first is a characterization of ISBP distributions; the second is a characterization of the Dirichlet distribution in terms of NTR processes. We recall such results in two lemmas.

Lemma 3.2. Invariance under size-biased permutation ([26], Theorem 4). *Let X be a random point of $[0,1]^\infty$ such that $|X| = 1$ almost surely with respect to a probability measure $d\mu$. For every k, let μ_k denote the distribution of $X_1, \ldots, X_k$, and G_k the measure on $[0,1]^k$, absolutely continuous with respect to μ_k with density*

$$\frac{dG_k}{d\mu_k}(x_1, \ldots, x_k) = \prod_{i=1}^{k-1}(1 - w_j)$$

*where $w_j = \sum_{i=1}^{j} x_i$, $j = 1, 2, \ldots$
X is invariant under size-biased permutation if and only if G_k is symmetric with respect to permutations of the coordinates in $\mathbb{R}^k$.*

Let X be the frequencies of an exchangeable partition Π, and denote with $\mathbb{P}_\mu$ the marginal law of the record indices of Π. Consider the measure G_k of Lemma 3.2. By Proposition 2.1 (ii), for every k

$$\begin{aligned} G_k(dx_1 \times \cdots \times dx_k) &= \mu_k(dx_1 \times \cdots \times dx_k) \prod_{i=1}^{k-1}(1-w_j) \\ &= \mu_k(dx_1 \times \cdots \times dx_k)\mathbb{P}(i_1 = 1, i_2 = 2, \ldots, i_k = k \,|x_1, \ldots, x_k) \\ &= \mu_k(dx_1 \times \ldots \times dx_k | i_k = k)\mathbb{P}_\mu(i_k = k) \end{aligned} \tag{40}$$

An equivalent characterization of ISBP measure is:

Corollary 3.1. *The law of X is invariant under size-biased permutation if and only if, for every k, there is a version of the conditional distribution*

$$\mu_k(dx_1 \times \ldots \times dx_k | i_k = k)$$

which is invariant under permutations of coordinates in $\mathbb{R}^k$.

The other result we recall is about Dirichlet distributions.

Lemma 3.3. Dirichlet and neutrality ([5], Theorem 7). *Let P be a random k-dimensional vector with positive components such that their sum equals 1. If P is NTR and P_k does not depend on $(1-P_k)^{-1}(P_1, \ldots, P_{k-1})$. Then P has the Dirichlet distribution.*

Now we have all elements to prove Proposition 3.3. Let $\mu(\cdot|i_1, i_2 \ldots)$ be the distribution of a NTL vector X such that the distribution of $\xi_j := X_{j+1}/W_{j+1}$, (with $W_j = \sum_{i=1}^j X_i$) has marginal law γ_j for $j = 1, 2, \ldots$. For every k, given $i_1, \ldots, i_k$, the vector $(X_2/W_2, \ldots, X_k/W_k)$ is conditionally independent of W_k and

$$\mu_k(dx_1 \times \cdots \times dx_k | i_k = k) = \left(\prod_{j=1}^{k-1} \gamma_j(d\xi_j) \right) \zeta_k(dw_k)$$

where ζ_k is the conditional law of W_k given $i_k = k$.
For X to be ISBP, corollary 3.1 implies that the product

$$\prod_{j=1}^{k-1} \gamma_j(d\xi_j)$$

must be a symmetric function of $x_1, \ldots, x_k$. Then, for every k, the vector $(X_1/W_k, \ldots, X_k/W_k)$ is both NTL and NTR, which implies in particular that X_k/W_k is independent of $W_{k-1}^{-1}(X_1, \ldots, X_{k-1})$. Therefore, by Lemma 3.3 and symmetry, $(\frac{X_1}{W_k}, \ldots, \frac{X_k}{W_k})$ is, conditionally on W_k and $\{i_k = k\}$, a symmetric Dirichlet distribution, with parameter, say, $1 - \alpha > 0$. By (40), the EPPF corresponding to $d\mu$ is equal to

$$\mathbb{E}\left[\prod_{j=1}^{k} X_j^{n_j - 1}(1 - W_{j-1})\right] = \mathbb{P}_\mu(i_k = k)\mathbb{E}\left[\prod_{j=1}^{k} X_j^{n_j - 1} | i_k = k\right]. \tag{41}$$

By the NTL assumption, we can write

$$\begin{aligned}\prod_{j=1}^{k} X_j^{n_j-1} &= \prod_{j=2}^{k}\left(\frac{X_j}{W_j}\right)^{n_j-1}\left(\frac{W_j}{W_k}\right)^{n_j-1} W_k^{n-k}, \\ &\overset{d}{=} \left(\prod_{j=2}^{k}\xi_j^{n_{j+1}-1}(1-\xi_j)^{S_j-j}\right) W_k^{n-k}, \qquad (42)\end{aligned}$$

where $S_j = \sum_{i=1}^{j} n_i$ $(j = 1, \dots, k)$. The last equality is due to

$$\prod_{j=1}^{k}\frac{W_j}{W_k} = \prod_{j=1}^{k-1}\left(\frac{W_j}{W_{j+1}}\right)^j.$$

Now, set

$$V_{n,k} = \frac{\mathbb{P}_\mu(i_k = k)\mathbb{E}(W_k^{n-k}|i_k = k)}{[k(1-\alpha)]_{(n-k)}}.$$

Equality (42) implies

$$\begin{aligned}\mathbb{E}\left[\prod_{j=1}^{k} X_j^{n_j-1}(1-W_{j-1})\right] &= \mathbb{P}_\mu(i_k = k)\int\left(\int\prod_{j=1}^{k-1}\xi_j^{n_{j+1}-1}(1-\xi_j)^{S_j-j}\gamma_j(\xi_j)d\xi_j\right)w_k^{n-k}\zeta_k(w_k)dw_k \\ &= \mathbb{P}_\mu(i_k = k)\mathbb{E}(W_k^{n-k}|i_k = k)\prod_{j=1}^{k-1}\frac{(1-\alpha)_{(n_{j+1}-1)}[j(1-\alpha)]_{(S_j-j)}}{[(j+1)(1-\alpha)]_{(S_{j+1}-(j+1))}} \\ &= V_{n,k}\prod_{j=1}^{k}(1-\alpha)_{(n_j-1)},\end{aligned}$$

which completes the proof. □

4 Age-ordered frequencies conditional on a single record index.

A representation for the Mellin transform of the m-th age-ordered cumulative frequencies W_m, conditional on i_m alone $(m = 1, 2, \dots)$ can be derived by using Proposition 3.2. We first point out a characterization for the moments of W_m, stated in the following Lemma.

Lemma 4.1. *Let $X_1, X_2, \dots$ be the limit age-ordered frequencies generated by a Gibbs partition with parameters α, V. For every $m = 1, 2, \dots$ and nonnegative integer n*

$$\mathbb{E}(W_m^n|i_m) = (i_m - \alpha m)_{(n)}\frac{V_{i_m+n,m}}{V_{i_m,m}} \qquad (43)$$

and

$$\mathbb{E}(X_m^n|i_m) = (1-\alpha)_{(n)}\frac{V_{i_m+n,m}}{V_{i_m,m}}. \qquad (44)$$

Proof. Let Π be an EGP (α, V) and denote $Y_j : j = 1, 2, \ldots$ the sequence of indicators $\{0,1\}$ such that $Y_j = 1$ if j is a record index of Π. Then $Y_1 = 1$ and, for every $l, m \leq l$,

$$\begin{aligned}\mathbb{P}(Y_{l+1} = 0 \mid \sum_{i=1}^{l} Y_i = m) &= (l - \alpha m)\frac{V_{l+1,m}}{V_{l,m}} \\ &= 1 - \mathbb{P}(Y_{l+1} = 1 \mid \sum_{i=1}^{l} Y_i = m).\end{aligned}$$

By proposition 2.1 and formula (13), given the cumulative frequencies $W = W_1, W_2, \ldots$,

$$\mathbb{P}(Y_{l+1} = 0 | \sum_{i=1}^{l} Y_i = m, W) = 1 - \mathbb{P}(Y_{l+1} = 1 \mid \sum_{i=1}^{l} Y_i = m, W) = W_m.$$

(see also [25], Theorem 6). Obviously this also implies that, conditional on W, the random sequence $K_l := \sum_{i=1}^{l} Y_i$, $(l = 1, 2, \ldots)$ is Markov, so we can write, for every l, m

$$\mathbb{P}(Y_{l+1} = 0 \,|\, K_l = m,\ K_l{-}1 = m{-}e, W) = \mathbb{P}(Y_{l+1} = 0 \,|\, K_l = m, W) = W_m, \qquad e = 0, 1. \quad (45)$$

Hence

$$\begin{aligned}\mathbb{E}(W_m | i_m) &= \mathbb{E}\left[\mathbb{P}(Y_{i_m+1} = 0 \mid K_{i_m} = m, K_{i_m-1} = m-1, W)\right] \\ &= \mathbb{E}\left[\mathbb{P}(Y_{i_m+1} = 0 \mid K_{i_m} = m, W)\right] \\ &= (i_m - \alpha m)\frac{V_{i_m+1,m}}{V_{i_m,m}}, \qquad (46)\end{aligned}$$

which proves the proposition for $m = 1$. The Markov property of K_n and (45) also lead, for every m, to

$$\begin{aligned}\mathbb{E}(W_m^n | i_j) &= \mathbb{E}\left[\mathbb{P}(Y_{i_m+1} = \ldots = Y_n = 0 \mid K_{i_m} = m, K_{i_m-1} = m-1, W)\right] \\ &= \prod_{i=1}^{n} \mathbb{E}\left[\mathbb{P}(Y_{i_m+i} = 0 \mid K_{i_m+i-1} = m, W)\right] \\ &= (i_m - \alpha m)_{(n)}\frac{V_{i_m+n,m}}{V_{i_m,m}},\end{aligned}$$

where the last equality is obtained as an n-fold iteration of (46).

The second part of the Lemma (formula (44)) follows from Proposition 3.2:

$$\begin{aligned}\mathbb{E}(X_m^n | i_m) &= \mathbb{E}(\xi_{m-1}^n | i_m)\mathbb{E}(W_m^n | i_m) \\ &= \frac{(1-\alpha)_{(n)}}{(i_m - \alpha m)_{(n)}}\mathbb{E}(W_m^n | i_m)\end{aligned}$$

which combined with (43) completes the proof. □

Given the coefficients $\{V_{n,k}\}$, analogous formulas to (43) and (44) can be obtained to describe the conditional Mellin transforms of W_m and X_m (respectively), in terms of the Mellin transform of the size-biased pick X_1.

Proposition 4.1. *Let $X_1, X_2, \ldots$ be the limit age-ordered frequencies generated by a Gibbs partition with parameters α, V. For every $m = 1, 2, \ldots$ and $\phi \geq 0$*

$$\mathbb{E}(W_m^\phi | i_m) = (i_m - \alpha m)_{(\phi)} \frac{V_{i_m,m}[\phi]}{V_{i_m,m}} \tag{47}$$

and

$$\mathbb{E}(X_m^\phi | i_m) = (1 - \alpha)_{(\phi)} \frac{V_{i_m,m}[\phi]}{V_{i_m,m}}, \tag{48}$$

for a sequence of functions $(V_{n,k}[\cdot] : \mathbb{R} \to \mathbb{R}; k, n = 1, 2, \ldots)$, uniquely determined by $V_{n,k} \equiv V_{n,k}[0]$, such that, for every $\phi \geq 0$,

$$V_{1,1}[\phi] = \frac{\mathbb{E}(X_1^\phi)}{(1-\alpha)_{(\phi)}}; \tag{49}$$

$$V_{n,k}[\phi] = (n + \phi - \alpha k) V_{n+1,k}[\phi] + V_{n+1,k+1}[\phi], \qquad n, k = 1, 2, \ldots; \tag{50}$$

$$V_{n,k}[\phi + 1] = V_{n+1,k}[\phi] \qquad n, k = 1, 2, \ldots. \tag{51}$$

Remark 4.1. *To complete the representation given in Proposition 4.1, notice that, for every α, the distribution of X_1 (the so-called* structural distribution*) is known for the extreme points of the Gibbs(α, V) family.*
In particular, for $\alpha \leq 0$ X_1 has a Beta$(1 - \alpha, \theta + \alpha)$ density ($\theta > 0$), where $\theta = m|\alpha|$ for some integer m when $\alpha < 0$ (see e.g. [27]). In this case,

$$V_{1,1}[\phi](\theta) = \frac{1}{(1-\alpha)_{(\phi)}} \left(\frac{(1-\alpha)_{(\phi)}}{(\theta+1)_{(\phi)}} \right) = \frac{1}{(\theta+1)_{(\phi)}}.$$

When $\alpha > 0$, we saw in the introduction that every extreme point in the Gibbs family is a Poisson-Kingman (α, s) partition for some $s > 0$; in this case the density of X_1 is

$$f_{1,\alpha}(x|s) = \frac{\alpha s^{-1}(x)^{-\alpha}}{\Gamma(1-\alpha)} \frac{f_\alpha((1-x)s^{1/\alpha})}{f_\alpha(s^{1/\alpha})} \qquad 0 < x < 1,$$

for an α-stable density f_α ([28], (57)), leading to

$$V_{1,1}[\phi](s) = \alpha s^{\frac{\phi-1}{\alpha}} G_\alpha(\phi - \alpha - 1, s^{-1/\alpha})$$

where G_α is as in (7).
Thus, for every α, the structural distribution of a Gibbs(α, V) partition, which defines $V_{1,1}[\phi]$ in (49), can be obtained as mixture of the corresponding extreme structural distributions.

Proof. Note that, for $\phi = 0, 1, 2, \ldots$, the proposition holds by Lemma 4.1 with $V_{n,m}[\phi] \equiv V_{n+\phi,m}$. For general $\phi \geq 0$ observe that, for every $m, n \in \mathbb{N}$,

$$\mathbb{E}(W_m^{\phi+n-m}|i_m = m) = \mathbb{E}(W_m^{\phi}|i_m = n)\mathbb{E}(W_m^{n-m}|i_m = m). \tag{52}$$

To see this, consider the random sequences Y_n, K_n defined in the proof of Lemma 4.1. By (45),

$$\begin{aligned}
\mathbb{E}(W_m^{\phi+n-m}|i_m = m) &= \mathbb{E}\left[W_m^{\phi}\mathbb{P}(K_n = m|K_m = m, W) \mid K_m = m\right] \\
&= \mathbb{E}\left[W_m^{\phi}\mathbb{P}(K_n = m|\ W) \mid K_m = m\right] \\
&= \mathbb{E}\left[W_m^{\phi} \mid K_n = m, K_m = m\right]\mathbb{E}\left[\mathbb{P}(K_n = m|\ W) \mid K_m = m\right] \\
&= \mathbb{E}\left[W_m^{\phi} \mid K_n = m\right]\mathbb{E}\left[W_m^{n-m} \mid K_m = m\right] \\
&= \mathbb{E}\left[W_m^{\phi} \mid i_m = n\right]\mathbb{E}\left[W_m^{n-m} \mid K_m = m\right]
\end{aligned}$$

where the last two equalities follow from (45), the Markov property of K_n and the exchangeability of the $Y's$.

From Lemma 4.1, we can rewrite (52) as

$$\mathbb{E}(W_m^{\phi}|i_m = n) = \frac{\mathbb{E}(W_m^{\phi+n-m}|i_m = m)}{[m(1-\alpha)]_{(n-m)}}\frac{V_{m,m}}{V_{n,m}}. \tag{53}$$

Now define

$$M_m(\phi) = \mathbb{E}(W_m^{\phi}|i_m = m), \qquad \phi \geq 0, m = 1, 2, \ldots. \tag{54}$$

and

$$V_{n,m}[\phi] = \frac{V_{m,m}M_m(\phi + n - m)}{[m(1-\alpha)]_{(n+\phi-m)}} \qquad \phi \geq 0, n, m = 1, 2, \ldots. \tag{55}$$

Notice that, with such a definition, Lemma 1 implies that $V_{n,m}[0] = V_{n,m}$. Moreover, $V_{n,m}[\phi + 1] = V_{n+1,m}[\phi]$ so (51) is satisfied; then (53) reads

$$\mathbb{E}(W_m^{\phi}|i_m) = (i_m - \alpha m)_{(\phi)}\frac{V_{i_m,m}[\phi]}{V_{i_m,m}}, \tag{56}$$

that is: (49),(47) and are satisfied.

Now it only remains to prove that such choice of $V_{n,m}[\phi]$ obeys the recursion (50) for every n, m, ϕ. By the same arguments leading to (52),

$$\begin{aligned}
M_m(\phi) - M_m(\phi+1) &= \mathbb{E}[W_m^{\phi}(1 - W_m)|K_m = m] \\
&= \mathbb{E}\left[W_m^{\phi}\big(1 - \mathbb{P}(K_{m+1} = m|W)\big) \mid K_m = m\right] \\
&= \mathbb{E}\left[W_m^{\phi} \mid K_{m+1} = m+1, K_m = m\right]\mathbb{E}\left[1 - \mathbb{P}(K_{m+1} = m|W) \mid K_m = m\right] \\
&= \mathbb{E}\left[(1-\xi_m)^{\phi} \mid i_{m+1} = m+1\right]\mathbb{E}\left[W_{m+1}^{\phi} \mid K_{m+1} = m+1\right]\frac{V_{m+1,m+1}}{V_{m,m}}
\end{aligned} \tag{57}$$

The last equality is a consequence of Proposition 3.2, for which, conditional on $K_{m+1} = m+1$ (which is equivalent to $\{i_{m+1} = m+1\}$), $W_m = (1-\xi_m)W_{m+1}$ for ξ_m (independent of W_{m+1}) having a Beta$(1-\alpha, m(1-\alpha))$ distribution. Therefore (57) can be rewritten as

$$M_m(\phi) - M_m(\phi+1) = \frac{[m(1-\alpha)]_{(\phi)}}{[(m+1)(1-\alpha)]_{(\phi)}} M_{m+1}(\phi) \frac{V_{m+1,m+1}}{V_{m,m}}, \tag{58}$$

and (50) follows from (58),(51), after some simple algebra, by comparing the definition (55) of $V_{n,k}[\phi]$, and the recursion (3) for the V-coefficients of an EGP(α, V). In particular, (58) shows that the functions $V_{n,k}[\phi]$ are uniquely determined by $V = (V_{n,k})$.

The equality (48) can now be proved in the same way as the moment formula (44). □

5 A representation for normalized age-ordered frequencies in an exchangeable Gibbs partition.

In this section we provide a characterization of the density of the first k (normalized) age-ordered frequencies, given i_k and W_k, and an explicit formula for the marginal distribution of i_k. We give a direct proof, obtained by comparison of the unconditional distribution of $X_1, \ldots, X_k$, $(k = 1, 2, \ldots)$, in a general Gibbs partition, with its analogue in Pitman's two-parameter model. Such a comparison is naturally induced by proposition 3.2, which says that, conditional on the record indices, the distribution of the age-ordered frequencies is the same for every Gibbs partition. Remember that the limit (unconditional) age-ordered frequencies in such a family are described by the two-parameter GEM distribution, for which

$$X_j \overset{d}{=} B_j \prod_{i=1}^{j-1} (1-B_i)$$

for a sequence $B_1, B_2, \ldots$ of independent Beta random variables with parameters, respectively, $(1-\alpha, \theta + j\alpha)$ (see e.g. [27]).

Proposition 5.1. *Let $X_1, X_2, \ldots$ be the age-ordered frequencies of a EGP(α, V) and, for every k let $W_k = \sum_{i=1}^k X_i$.*

(i) Conditional on $W_k = w$ and on $i_k = k+i$, the law of the vector $X_1, \ldots, X_k$ is

$$d\mu_{\alpha,V}(x_1, \ldots, x_k | w, k+i) = \frac{V_{k+i,k}}{V_{k+i-1,k-1}} w^{-(k-1)} \sum_{\mathbf{m} \in \mathbb{N}^{k-1} : |\mathbf{m}| = k+i-1} \bar{\mu}_{\alpha,V}(\mathbf{m}) \mathcal{D}_{(\mathbf{m}-\alpha)}\left(\frac{x_1}{w}, \ldots, \frac{x_k}{w}\right) dx_1, \ldots, dx_{k-1} \tag{59}$$

where: $\mathcal{D}_{(\mathbf{m}-\alpha)}$ is the Dirichlet density with parameters $(m_1 - \alpha, \ldots, m_{k-1} - \alpha, 1-\alpha)$, and $\mu_{\alpha,V}(\mathbf{m})$ is the age-ordered sampling formula (15) with Gibbs' EPPF $q_{\alpha,V}$ as in (13):

$$\bar{\mu}_{\alpha,V}(\mathbf{m}) = \left(\frac{(k+i-1)!}{\prod_{j=1}^{k-1} (m_j)! (k+i-1-\sum_{i=1}^{j-1} m_i)} \right) V_{k+i-1,k-1} \prod_{j=1}^{k-1} (1-\alpha)_{(m_j - 1)}.$$

(ii) The marginal distribution of i_k is

$$\mathbb{P}(i_k = k+i) = V_{k+i,k} \frac{\alpha^{-(k-1)}}{(k+i-1)!} \sum_{j=0}^{k-1} \frac{(-1)^{j+k+i-1}}{j!(k-j-1)!} (\alpha j)_{[k+i-1]}, \tag{60}$$

where $a_{[n]} = a(a-1)\cdots(a-n+1)$.

Remark 5.1. *The density of i_k can be expressed in terms of generalized Stirling numbers $\begin{bmatrix} n \\ k \end{bmatrix}_\alpha$, defined as the coefficients of x^n in*

$$\frac{n!}{\alpha^k k!}(1-(1-x)^\alpha)^k$$

([18],[14]). Formula (60) can be re-expressed as:

$$\mathbb{P}(i_k) = V_{i_k,k} \begin{bmatrix} i_k - 1 \\ k-1 \end{bmatrix}_\alpha,$$

which makes clear the connection between the distribution of i_k and the distribution of K_n recalled in the introduction (formula (8)). In fact, (60) can be deduced simply from (8) by the Markov property of the sequence K_n as

$$\begin{aligned} \mathbb{P}(i_k) &= \mathbb{P}(K_{i_k} = k \mid K_{i_k - 1} = k-1)\ \mathbb{P}(K_{i_k-1} = k-1) \\ &= \frac{V_{i_k,k}}{V_{i_{k-1},k-1}} V_{i_{k-1},k-1} \begin{bmatrix} i_k - 1 \\ k-1 \end{bmatrix}_\alpha. \end{aligned}$$

However here we give a self-contained proof in order to show how (60) is implied by proposition 3.2 through (59).

Proof. From (10), we know that

$$P(i_2, \ldots, i_k) = V_{i_k,k} \frac{\Gamma(i_2 - \alpha - 1)\cdots\Gamma(i_k - \alpha(k-1) - 1)}{\Gamma(1-\alpha)\Gamma(i_2 - 2\alpha)\cdots\Gamma(i_{k-1} - (k-1)\alpha)};$$

By proposition 3.2 and Lemma 4.1,

$$\begin{aligned} \mathbb{E}(\prod_{j=1}^{k} X_j^{n_j} | i_1, \ldots, i_k) &= \mathbb{E}\left(\prod_{j=1}^{k-1} \xi_j^{n_{j+1}} (1-\xi_j)^{S_j} \prod_{i=k}^{\infty} (1-\xi_i)^n \;\middle|\; i_1, \ldots, i_k \right) \\ &= \prod_{j=1}^{k-1} \frac{(1-\alpha)_{(n_{j+1})} (i_{j+1} - j\alpha - 1)_{(S_j)}}{(i_{j+1} - (j+1)\alpha)_{(S_{j+1})}} \mathbb{E}(\prod_{i=k}^{\infty} (1-\xi_i)^n | i_k) \\ &= \left(\prod_{j=1}^{k-1} \frac{(1-\alpha)_{(n_{j+1})} (i_{j+1} - j\alpha - 1)_{(S_j)}}{(i_{j+1} - (j+1)\alpha)_{(S_{j+1})}} \right) \frac{\Gamma(i_k + n - \alpha k)}{\Gamma(i_k - \alpha k)} \frac{V_{n+i_k,k}}{V_{i_k,k}}, \end{aligned} \tag{61}$$

hence

$$\begin{aligned}
\mathbb{E} & \left(\prod_{j=1}^{k} X_j^{n_j} | 1 \equiv i_1, i_2, \ldots, i_k) P(i_2, \ldots, i_k) \right. \\
&= (i_k - \alpha k)_{(n)} V_{n+i_k,k} \prod_{j=1}^{k-1} \frac{(1-\alpha)_{n_{j+1}} \Gamma(i_{j+1} - j\alpha - 1 + S_j) \Gamma(i_{j+1} - \alpha(j+1))}{\Gamma(i_{j+1} - \alpha(j+1) + S_{j+1}) \Gamma(i_j - \alpha j)} \\
&= \prod_{j=1}^{k} (1-\alpha)_{(n_j)} \left[V_{i_k+n,k} \prod_{j=1}^{k-1} (i_j - \alpha j + S_j)_{(i_{j+1}-i_j-1)} \right], \qquad (62)
\end{aligned}$$

where the last equality follows after multiplying and dividing all terms by $(1-\alpha)_{n_1}$. Consequently, a moment formula for general Gibbs partitions is of the form:

$$\mathbb{E}_{(\alpha,V)} \left(\prod_{j=1}^{k} X_j^{n_j} \right) = \prod_{j=1}^{k} (1-\alpha)_{(n_j)} \sum_{1<i_2<\ldots<i_k} c_{(1,i_2)} \cdots c_{(i_{k-1},i_k)} V_{n+i_k,k}. \qquad (63)$$

where, for $1 < j \leq k-1$,

$$c_{(i_j,i_{j+1})} = (i_j - \alpha j + S_j)_{(i_{j+1}-i_j-1)}.$$

For fixed i_k denote

$$\lambda_{i_k} = \sum_{1<i_2<\cdots<i_k} c_{(i_1,i_2)} \cdots c_{(i_{k-1},i_k)}.$$

Then (63) reads

$$\mathbb{E}_{(\alpha,V)} \left(\prod_{j=1}^{k} X_j^{n_j} \right) = \prod_{j=1}^{k} (1-\alpha)_{(n_j)} \sum_{i=k}^{\infty} \lambda_i V_{n+i,k}. \qquad (64)$$

For Pitman's two-parameter family, this becomes

$$\mathbb{E}_{(\alpha,\theta)} \left(\prod_{j=1}^{k} X_j^{n_j} \right) = \prod_{j=1}^{k} (1-\alpha)_{(n_j)} \sum_{i=k}^{\infty} \lambda_i \frac{\prod_{j=1}^{k} (\theta + \alpha(j-1))}{\theta_{(n+i)}}. \qquad (65)$$

The two-parameter GEM distribution implies that

$$\begin{aligned}
\mathbb{E} \left(\prod_{j=1}^{k} X_j^{n_j} \right) &= \mathbb{E} \left(\prod_{j=1}^{k} B_j^{n_j} (1-B_j)^{n-S_j} \right) \\
&= \prod_{j=1}^{k} \frac{(1-\alpha)_{n_j} (\theta + j\alpha)_{(n-S_j)}}{(\theta + 1 + (j-1)\alpha)_{(n-S_{j-1})}} \\
&= \frac{1}{(\theta)_{(n)}} \prod_{j=1}^{k} \frac{(1-\alpha)_{(n_j)} (\theta + \alpha(j-1))}{(\theta + \alpha(j-1) + n - S_{j-1})}. \qquad (66)
\end{aligned}$$

therefore, from (65) we derive the identity:

$$\prod_{j=1}^{k} \frac{1}{(\theta + \alpha(j-1) + n - S_{j-1})} = \sum_{i=k}^{\infty} \lambda_i \frac{\theta_{(n)}}{\theta_{(n+i)}}, \qquad \theta > 0. \qquad (67)$$

For $\theta > 0$ replace θ by $\theta - n$ in (67), and denote $n_j^* = 1 - \alpha + n_j$. We now find an expansion of the left-hand side of (67) in terms of products of the type $\prod_1^k (n_j^*)_{(m_j)}$, for $m_1, \ldots, m_k \geq 0$. The left-hand side of (67) is now

$$\begin{aligned}
\prod_{j=1}^{k} \frac{1}{(\theta - n + \alpha(j-1) + n - S_{j-1}^*)} &= \prod_{j=1}^{k} \frac{1}{(\theta + j - 1 + S_{j-1}^*)} \\
&= \prod_{j=1}^{k} \int_0^1 t_{j-1}^{j-1+\theta-1-S_{j-1}^*} dt_{j-1} \\
&= \int \left(\prod_{j=1}^{k-1} t_j \right)^{\theta-1} (t_1 \cdots t_{k-1})^{-n_1^*} (t_2 \cdots t_{k-1})^{-n_2^*} \cdots t_{k-1}^{-n_{k-1}^*} \prod_{j=1}^{k-1} t_j^j \, dt_0 \cdots dt_{k-1}, \quad (68)
\end{aligned}$$

where $S_0^* = 0$ and $S_j^* = \sum_{i=1}^j n_j^*,\ j = 1, \ldots, k-1$. Make the change of variable

$$u_j = 1 - \prod_{i=j}^{k-1} t_i, \qquad j = 0, \ldots, k-1.$$

Then $0 < u_{k-1} < \ldots < u_0 < 1$. The absolute value of the Jacobian is

$$\begin{aligned}
\left| \frac{du}{dt} \right| &= (t_1 \cdots t_{k-1}) \times (t_2 \cdots t_{k-1}) \times t_{k-1} \times 1 \\
&= \prod_{j=1}^{k-1} t_j^j.
\end{aligned}$$

Thus (68) transforms to

$$\int_{0<u_{k-1}<\ldots<u_0<1} (1 - u_0)^{\theta-1} \prod_{j=1}^{k-1} (1-u)^{-n_j^*} du_1 \cdots du_{k-1} du_0.$$

Fix u_0 and consider

$$\begin{aligned}
&\int_{0<u_{k-1}<\ldots<u_0} \prod_{j=1}^{k-1} (1-u)^{m_j} du_1 \cdots du_{k-1} \\
&\qquad = u_0^{k-1+\sum_{i=1}^{k-1} m_i} \frac{1}{m_{j-1}+1} \cdot \frac{1}{m_{j-1}+m_{j-2}+1} \cdot \frac{1}{m_{j-1}+\ldots+m_1+1}.
\end{aligned}$$

The integral in (68) is thus

$$\sum_{m_1,\ldots,m_{k-1}\geq 0} \frac{c(\mathbf{m})}{\prod_{j=1}^{k-1}(1-\alpha)_{(m_j)}} \prod_{j=1}^{k-1} n_{j\,(m_j)}^* \int_0^1 (1-u_0)^{\theta-1} u_0^{k-1+\sum_{i=1}^{k-1} m_i} du_0, \qquad (69)$$

where

$$c(\mathbf{m}) = \prod_{j=1}^{k-1} \frac{1}{k - j + \sum_{i=j}^{k-1} m_i} \prod_{j=1}^{k-1} \frac{(1-\alpha)_{(m_j)}}{m_j!}. \qquad (70)$$

Now consider the right-hand side of (67), again with θ replaced by $\theta - n$:

$$\sum_{i=0}^{\infty} \frac{\lambda_{k+i}}{(k+i-1)!} \int_0^1 (1-x)^{\theta-1} x^{i+k-1} dx$$

and compare it with (69) to obtain a representation for the λ_i's in (64):

$$\frac{\lambda_{k+i}}{(k+i-1)!} = \sum_{\mathbf{m}\in\mathbb{N}_0^k:|\mathbf{m}|=i} \frac{c(\mathbf{m})}{(1-\alpha)_{(m_j)}} \prod_{j=1}^{k-1} n^*_{j\,(m_j)}, \tag{71}$$

where $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Recall that $n_j^* = 1 - \alpha + n_j$ and consider the identity

$$\frac{(1-\alpha)_{(n_j)} n^*_{j\,(m_j)}}{(1-\alpha)_{(m_j)}} = (1-\alpha+m_j)_{(n_j)}.$$

From (43) we also know that

$$V_{n+k+i,k} = \frac{V_{k+i,k}}{(k+i-\alpha k)_{(n)}} \mathbb{E}(W_k^n | i_k = k+i).$$

Thus (64) and (71) imply that

$$\begin{aligned} \mathbb{E}_{(\alpha,V)}\left(\prod_{j=1}^{k} X_j^{n_j}\right) &= \sum_{i=0}^{\infty} (k+i-1)! \mathbb{E}(W_k^n | i_k = k+i) V_{k+i,k} \\ &\quad \times \sum_{\mathbf{m}\in\mathbb{N}_0^{k-1}:|\mathbf{m}|=i} c(\mathbf{m}) \left[\frac{(1-\alpha)_{(n_k)} \prod_{j=1}^{k-1}(1-\alpha+m_j)_{(n_j)}}{(k+i-\alpha k)_{(n)}}\right]. \end{aligned} \tag{72}$$

Now, for every $\mathbf{m} \in \mathbb{N}_0^{k-1}$ such that $|\mathbf{m}| = i$, define $m'_j = m_j + 1$; then we can rewrite

$$\begin{aligned} (k+i-1)! c(\mathbf{m}) V_{k+i,k} &= \frac{V_{k+i,k}}{V_{k+i-1,k-1}} \left(\frac{(k+i-1)!}{\prod_{j=1}^{k-1} m'_j!(k+i-1-\sum_{l=1}^{j-1} m'_j)}\right) V_{k+i-1,k-1} \prod_{j=1}^{k-1} (1-\alpha)_{(m'_j-1)} \\ &= \frac{V_{k+i,k}}{V_{k+i-1,k-1}} \mu_{\alpha,V}(\mathbf{m}'). \end{aligned}$$

Thus the right-hand side of (72) becomes

$$= \sum_{i=0}^{\infty} \frac{V_{k+i,k}}{V_{k+i-1,k-1}} \sum_{\mathbf{m}'\in\mathbb{N}^{k-1}:|\mathbf{m}'|=k+i-1} \mu_{\alpha,V}(\mathbf{m}') \left[\frac{(1-\alpha)_{(n_k)} \prod_{j=1}^{k-1}(m_j'-\alpha)_{(n_j)}}{(k+i-\alpha k)_{(n)}} \mathbb{E}(W_k^n | i_k = k+i)\right] \tag{73}$$

The term between square brackets is the $n_1, \ldots, n_k$-th moment of k $[0,1]$-valued random variables $Y_1(\mathbf{m}'), \ldots, Y_k(\mathbf{m}')$ such that

$$\sum_{i=1}^{k} Y_i(\mathbf{m}') \overset{d}{=} (W_k | i_k = k+i)$$

and, conditional on $\sum_{i=1}^{k} Y_i(\mathbf{m}')$, the distribution of

$$\frac{Y_1(\mathbf{m}')}{\sum_{i=1}^{k} Y_i(\mathbf{m}')}, \dots, \frac{Y_k(\mathbf{m}')}{\sum_{i=1}^{k} Y_i(\mathbf{m}')}$$

is a Dirichlet distribution with parameters $(m_1{}'-\alpha, \dots, m_{k-1}{}'-\alpha, 1-\alpha)$, therefore (73) completes the proof of part (i).

To prove part (ii), we only have to notice that, from (73) it must follow that

$$\sum_{i=0}^{\infty} \frac{V_{k+i,k}}{V_{k+i-1,k-1}} \sum_{\mathbf{m}' \in \mathbb{N}^{k-1}:|\mathbf{m}'|=k+i-1} \mu_{\alpha,V}(\mathbf{m}') = 1$$

hence a version of the marginal probability of i_k is, for every k

$$\mathbb{P}(i_k = k+i) = \frac{V_{k+i,k}}{V_{k+i-1,k-1}} \sum_{\mathbf{m}' \in \mathbb{N}^{k-1}:|\mathbf{m}'|=k+i-1} \mu_{\alpha,V}(\mathbf{m}'). \tag{74}$$

This can be also argued directly, simply by noting that, in an EGP(α, V),

$$\sum_{\mathbf{m}' \in \mathbb{N}^{k-1}:|\mathbf{m}'|=k+i-1} \mu_{\alpha,V}(\mathbf{m}') = \mathbb{P}(i_{k-1} \leq k+i-1, i_k > k+i-1)$$

and that

$$\frac{V_{k+i,k}}{V_{k+i-1,k-1}} = \mathbb{P}(i_k = k+i | i_{k-1} \leq k+i-1, i_k > k+i-1).$$

We want to find an expression for the inner sum of (74). If we reconsider the term $c(\mathbf{m})$ as in (70) (for $\mathbf{m} \in \mathbb{N}_0^{k-1} : |\mathbf{m}| = i$), we see that

$$\sum_{\mathbf{m} \in \mathbb{N}_0^{k-1}:|\mathbf{m}|=i} c(\mathbf{m})$$

is the coefficient of ζ^i in

$$\begin{aligned} \frac{1}{(k-1)!}\left[\int_0^1 (1-u\zeta)^{\alpha-1} du\right]^{k-1} &= \frac{1}{(k-1)!}(\frac{1}{\zeta\alpha})^{k-1}[1-(1-\zeta)^{\alpha}]^{k-1} \\ &= (\zeta\alpha)^{-(k-1)} \sum_{j=0}^{k-1} \frac{(-1)^{k+i+j-1}}{j!(k-1-j)!}(1-\zeta)^{\alpha j}. \end{aligned}$$

Thus

$$\sum_{\mathbf{m} \in \mathbb{N}_0^{k-1}:|\mathbf{m}|=i} c(\mathbf{m}) = \frac{\alpha^{-(k-1)}}{(k+i-1)!} \sum_{j=0}^{k-1} \frac{(-1)^{k+i+j-1}}{j!(k-1-j)!}(j\alpha)_{[k+i-1]}. \tag{75}$$

Since

$$\sum_{\mathbf{m}' \in \mathbb{N}^{k-1}:|\mathbf{m}'|=k+i-1} \mu_{\alpha,V}(\mathbf{m}') = (k+i-1)!\; V_{k+i-1,k-1} \sum_{\mathbf{m} \in \mathbb{N}_0^{k-1}:|\mathbf{m}|=i} c(\mathbf{m})$$

then part (ii) is proved by comparison of (75) with (74). □

References

[1] D.J. Aldous *Exchangeability and related topics*, volume 1117 of *Lecture Notes in Mathematics*, pp. 1–198. Springer-Verlag, Berlin, 1985. Lecture notes from École d'été de Probabilités de Saint-Flour XIII - 1983. MR0883646

[2] J. Bertoin *Random fragmentation and coagulation processes.* Cambridge Studies in Advanced Mathematics, vol. 102. Cambridge University Press, Cambridge 2006. MR2253162

[3] N. Berestycki and J. Pitman. Gibbs distributions for random partitions generated by a fragmentation process, 2005. http://arXiv.org/abs/math/0512378. MR2314353

[4] R.J. Connor and J.E. Mosimann. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Am. Stat. Assoc.*, 64:194–206, 1969. MR0240895

[5] K. Bobecka and J. Wesołowski. The Dirichlet distribution and process through neutralities. *J. Theor. Probab.*, 20: 295Ű-308, 2007.

[6] K. Doksum. Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.*, 2:183–201, 1974. MR0373081

[7] R. Dong, C. Goldschmidt, and J. B. Martin. Coagulation-fragmentation duality, Poisson-Dirichlet distributions and random recursive trees, 2005. To appear in *Ann. Appl. Probab.*, http://arxiv.org/abs/math.PR/0507591. MR2288702

[8] P. Donnelly. Partition structures, Pólya urns, the Ewens sampling formula, and the ages of alleles. *Theoret. Population Biol.*, 30(2):271–288, 1986. MR0865115

[9] P. Donnelly and P. Joyce. Continuity and weak convergence of ranked and size-biased permutations on the infinite simplex. *Stoc. Proc. Appl.*,31(1):89–103, 1989. MR0996613

[10] P. Donnelly and T. G. Kurtz. Particle representations for measure-valued population models. *Ann. Probab.*, 27(1):166–205, 1999. MR1681126

[11] A. V. Gnedin. The representation of composition structures. *Ann. Probab.*, 25(3):1437–1450, 1997. MR1457625

[12] A. V. Gnedin. On convergence and extensions of size-biased permutations. *J. Appl. Probab.*, 35(3):642–650, 1998. MR1659532

[13] A. Gnedin. Constrained exchangeable partitions, 2006. To appear in *Discr. Math. Comp. Sci.*, http://arXiv:math/0608621v1 [math.PR]

[14] A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 325(Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 12):83–102, 244–245, 2005. MR2160320

[15] R. C. Griffiths and S. Lessard. Ewens' sampling formula and related formulae: Combinatorial proofs, extensions to variable population size and applications to ages of alleles. *Theor. Popul. Biol.*, 68:167–177, 2005.

[16] F. M. Hoppe. Pólya-like urns and the Ewens' sampling formula. *J. Math. Biol.*, 20(1):91–94, 1984. MR0758915

[17] H. Ishwaran and L.F. James. Some further developments for stick-breaking priors: finite and infinite clustering and classification, *Sankhyā. The Indian Journal of Statistics*, 65(3): 577–592, 2003. MR2060608

[18] S. V. Kerov. Combinatorial examples in the theory of AF-algebras. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)*, 172(Differentsialnaya Geom. Gruppy Li i Mekh. Vol. 10):55–67, 169–170, 1989. MR1015698

[19] S. V. Kerov. Subordinators and permutation actions with quasi-invariant measure. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 223(Teor. Predstav. Din. Sistemy, Kombin. i Algoritm. Metody. I):181–218, 340, 1995. MR1374320

[20] S. V. Kerov and N.V. Tsilevich. A random subdivision of an interval generates virtual permutations with the Ewens distribution. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 223(Teor. Predstav. Din. Sistemy, Kombin. i Algoritm. Metody. I):162–180, 339–340, 1995. MR1374319

[21] J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Probab.*, (Special Vol. 19A):27–43, 1982. Essays in statistical science. MR0633178

[22] A. Lijoi, I. Prünster, S.G. Walker. Bayesian nonparametric estimators derived from conditional Gibbs structures.*Tech. Report, Università degli Studi di Torino, 2007.*

[23] S. Nacu. Increments of random partitions, 2004. http://arXiv:math/0310091v2 [math.PR] MR2238047

[24] M. Perman, J. Pitman and M. Yor. Size-biased sampling of Poisson point processes and excursions *Probab. Theory Related Fields*, 92(1):21–39, 1992. MR1156448

[25] J. Pitman. Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields*, 102(2):145–158, 1995. MR1337249

[26] J. Pitman. Random discrete distributions invariant under size-biased permutation. *Adv. in Appl. Probab.*, 28(2):525–539, 1996. MR1387889

[27] J. Pitman. Some developments of the Blackwell-MacQueen urn scheme. *Statistics, probability and game theory*, volume 30 of *IMS Lecture Notes Monogr. Ser.*. Inst. Math. Statist., Hayward, CA, pp. 245–267, 1996. MR1481784

[28] J. Pitman. Poisson-Kingman partitions. *Statistics and science: a Festschrift for Terry Speed*, volume 40 of *IMS Lecture Notes Monogr. Ser.*. Inst. Math. Statist., Beachwood, OH, pp. 1–34, 2003. MR2004330

[29] J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics.* Springer-Verlag, Berlin Heidelberg, 2006. Lecture notes from École d'été de Probabilités de Saint-Flour XXXII - 2002, *ed.* J. Picard. MR2245368

[30] S. Walker and P. Muliere. Beta-Stacy processes and a generalization of the Pólya-urn scheme. *Ann. Statist.*, 25(4):1762–1780, 1997. MR1463574

[31] S. Walker and P. Muliere. A characterization of a neutral to the right prior via an extension of Johnson's sufficientness postulate. *Ann. Statist.*, 27(2):589–599, 1999. MR1714716

[32] G. Watterson. Lines of descent and the coalescent. *Theor. Popul. Biol.*, 26(1):77–92, 1984. MR0760232

[33] S. Zabell Predicting the unpredictable, *Synthese*, 90(2):205–232, 1992. MR1148566