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
e-mail: {griff,spano}@stats.ox.ac.uk

March 30, 2022

Abstract The frequencies $X_1, X_2, \ldots$ of an exchangeable Gibbs random partition $\Pi$ of $\mathbb{N} = \{1, 2, \ldots\}$ (Gnedin and Pitman (2006)) 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 $\Pi$. In particular, conditioning on $1 = i_1 < i_2 < \ldots$, a representation is shown to be

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

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

* Research supported in part by EPSRC grant GR/T21783/01.
Correspondence to: Dario Spanò, Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, United Kingdom.
1 Introduction.

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

We consider a family of consistent random partitions studied by Gnedin and Pitman [13] 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

\[
\frac{V_{n+1,k+1}}{V_{n,k}}
\]

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

\[
\frac{n_j - \alpha}{n - \alpha k} \left(1 - \frac{V_{n+1,k+1}}{V_{n,k}}\right)
\]

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

\[
V_{1,1} = 1; \quad V_{n,k} = (n - \alpha k)V_{n+1,k} + V_{n+1,k+1}.
\]

Every partition \(\Pi\) of \(\mathbb{N}\) so generated is called an exchangeable Gibbs partition with parameters \((\alpha, V)\) (EGP(\(\alpha, V)\)), where exchangeable means that, for every \(n\), the distribution of \(\Pi_n\) is a symmetric function of the vector \(\mathbf{n} = (n_1, \ldots, n_k)\) of its frequencies \((24)\) (see section \(2\) below). Actually, the whole family of EGP\(s\), treated in [13] 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 EGP\(s\) is Pitman’s two-parameter family, for which \(V\) is given by

\[
V_{n,k}^{(\alpha, \theta)} = \frac{\prod_{j=1}^{k} (\theta + \alpha(j - 1))}{\theta^{(n)}}
\]

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 EGP\(s\) with \(V\)-coefficients of the form

\[
V_{n,k}^{(\alpha, \theta)} = \frac{V_k}{c_n}
\]

for some sequence of constants \((c_n)\) \((13, Corollary 4)\). If we let \(\alpha = 0\) in \((4)\), we obtain the well known Ewens’ partition for which:

\[
V_{n,k}^{(0, \theta)} = \frac{\theta^k}{\theta^{(n)}}.
\]

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. \([31]\), \([20]\)) and became a paradigm for the modern developments of a theory of exchangeable random partitions \((1), [24], [11]\).

For every fixed \(\alpha\), the set of all EGP(\(\alpha, V)\) forms a convex set; Gnedin and Pitman proved it and gave a complete description of the extreme points \((13, Theorem 12)\). It turns out, in particular, that for every \(\alpha \leq 0\), the extreme set is given by 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 \((\alpha, s)\), \(s > 0\), whose \(V\)-coefficients are given by:

\[
V_{n,k}(s) = \alpha^k s^{n/\alpha} G_\alpha(n - \alpha k, s^{-1/\alpha}),
\]
with
\[ G_\alpha(q,t) := \frac{1}{T(q)f_\alpha(t)} \int_0^t f_\alpha(t-v)v^{\alpha-1}dv, \]
and \( f_\alpha \) is an \( \alpha \)-stable density (28, Theorem 4.5). The partition induced by (9) 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 \( f_\alpha \). The parameter \( s \) has the interpretation as the (a.s.) limit of the ratio \( K_n/n^\alpha \) as \( n \to \infty \), where \( K_n \) is the number of classes in the partition \( \Pi_n \) generated via \( V_{n,k}(s) \). \( K_n \) is shown in 13 to play a central role in determining the extreme set of \( V_{n,k} \) for every \( \alpha \); the distribution of \( K_n \) for every \( n \), turns out to be of the form
\[ \mathbb{P}(K_n = k) = V_{n,k} \left[ \frac{n}{k} \right]_\alpha. \]
where \([n/k]_\alpha\) are generalized Stirling numbers, defined as the coefficients of \( x^n \) in
\[ \frac{n!}{\alpha^k k!} (1 - (1-x)^\alpha)^k \]
(see 13 and reference therein). As \( n \to \infty \), \( K_n \) behaves differently for different choices of the parameter \( \alpha \): almost surely it will be finite for \( \alpha < 0 \), \( K_n \sim S \log n \) for \( \alpha = 0 \) and \( K_n \sim Sn^\alpha \) for positive \( \alpha \), 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 \( i = (1 = i_1 < i_2 < \ldots) \). To this purpose, we adopt a combinatorial approach proposed by Griffiths and Lessard 14 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. 31), whose equilibrium distribution is given by Ewens’ partition (3), 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 \( n = (n_1, \ldots, n_k) \), conditional on the record indices \( i_n = (1 = i_1 < i_2 < \ldots < i_k) \) of \( \Pi_n \), as well as the distribution of \( i_n \); (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 8 where we find that, conditional on \( i = (1, i_2, \ldots) \), for every \( j = 1, 2, \ldots \), \( X_j|i \equiv \xi_{j-1} \prod_{i=j}^\infty (1 - \xi_i) \)
almost surely, for an independent sequence \((\xi_0, \xi_1, \ldots) \in [0,1]^\infty \) such that \( \xi_0 = 1 \) and \( \xi_m \) 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 \( i = (i_1, i_2, \ldots) \) which is a non-homogenous Markov chain, starting at \( i_1 = 1 \), with transition probabilities
\[ P_j(i_{j+1}|i_j) = \frac{(i_j - \alpha j)(i_{j+1} - i_j - 1)}{V_{i_j,i_{j+1}}}, \quad j \geq 1. \]
The representation (9) extends Griffiths and Lessard’s result on Ewens’ partitions (14, 29), recovered just by letting \( \alpha = 0 \).

In section 5.2 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 \( N \). 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 $X_1$. In section 4.1 an alternative description of the conditional distribution of $-\log X_j$ given $i_j$ leads to a characterization of Ewens-Kingman’s partitions as the only class for which $-\log X_j$ can be expressed as an infinite sum of independent random variables.

Finally in section 4 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 ([24], [25]), for which:

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

(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 ([24]), as it will be clear in section 3.2. When $\alpha$ is strictly positive, the structure of the age-ordered frequencies in the extreme points lose such a simple structure. A description is available in [23].

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 [24]. 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 [18], Kerov and Tsilevich [19], and more recently by Gnedin [12], and Nacu [22], 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 [24] and [28] and reference therein. Let $\mu$ 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:

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

(12)

The function $q_\mu$ is called the Partially exchangeable probability function (PEPF) of $\mu$, and has the interpretation as the probability distribution of a random partition $\Pi_n = (\Pi_{n1}, \ldots, \Pi_{nk})$, for which a sufficient statistic is given by its age-ordered frequencies $(n_1, \ldots, n_k)$, that is:

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

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 $\Pi_n$ an exchangeable random partition.

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

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

(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 \(\Pi_n\) is given, because of the symmetry of its EPPF, by its unordered frequencies (i.e. the count of how many frequencies in \(\Pi_n\) are equal to 1, ..., to \(n\)), whose distribution is given by their (unordered) sampling formula:

\[
\hat{\mu}(n) = \left(\frac{|n|}{n}\right) \frac{1}{\prod_{i=1}^{n} b_i!} q_{\mu}(n),
\]

where

\[
\left(\frac{|n|}{n}\right) = \frac{|n|!}{\prod_{j=1}^{k} n_j!}
\]

and \(b_i\) is the number of \(n_j\)'s in \(n\) 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

\[
\hat{\mu}(n) = \left(\frac{|n|}{n}\right) a(n)q_{\mu}(n),
\]

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

\[
a(n) = \prod_{j=1}^{k} \frac{n_j}{n - \sum_{i=1}^{j-1} n_i}
\]

is the distribution of the size-biased permutation of \(n\).

If \(\Pi = (\Pi_n)\) is a (partially) exchangeable partition with PEPF \(q_{\mu}\), then the vector \(n^{-1}(n_1, n_2, \ldots)\) of the relative frequencies, in age-order, of \(\Pi_n\), converges a.s. to a random point \(P = (p_1, p_2, \ldots) \in \Delta\) with distribution \(\mu\): 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 \(q_{\mu}\) 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) \(\Pi_{n+1}\), given (those of) \(\Pi_n\). 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 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.

Proposition 1. \(\Pi_{n+1}\) given \(\Pi_n\).

The use of an urn scheme of the form \((1)-(2)\) in Population Genetics is due to Hoppe \((15)\) 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 [32] and [26]. Examples of recent applications of Exchangeable Gibbs partitions in Bayesian nonparametric Statistics are in [21], [16]. 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 $\Pi$ be a partially exchangeable random partition. Since, for every $n$, its age-ordered frequencies $n = (n_1, \ldots, n_k)$ are a sufficient statistic for $\Pi_n$, then all realizations $\pi_n$ with the same $n$ and the same record indices $i_n = 1 < i_2 < \ldots < i_k \leq n$ must have equal probability. To evaluate the joint probability of the pair $(n, i_n)$, we only need to replace $a(n)$ 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 [14] to be equal to

$$a(n, i_n) = \frac{n!}{\prod_{j=1}^k n_j!}.$$ 

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

$$\tilde{\mu}(n, i_n) = \frac{n!}{\prod_{j=1}^k n_j!} a(n, i_n) q_n(n).$$

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

$$\tilde{\mu}(i_n) = \sum_{n \in B(i_n)} \frac{n!}{\prod_{j=1}^k n_j!} a(n, i_n) q_n(n).$$

where

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

is the set of all possible $n$ compatible with $i$. In [14] such a formula is derived for the particular case of Ewens’ partitions. For general random partitions see also [22], section 2.

Notice that, for every $n$ such that $|n| = n$,

$$a(n) = \sum_{i_n \in C(n)} a(n, i_n) = \prod_{j=1}^k \frac{n_j}{n - \sum_{j=1}^{S_{j-1} + 1} n_i},$$

where

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

is the set of all possible $i_n$ compatible with $n$. Then the marginal distribution of the age-ordered frequencies (15) is recovered by summing (19) over $C(n)$.

This observation incidentally links a classical combinatorial result to partially exchangeable random partitions.
Proposition 1 Let $\Pi = (\Pi_n)$ be a partially exchangeable partition with PEPF $q_n$.

(i) Given the frequencies $\mathbf{n} = (n_1, \ldots, n_k)$ in age-order, the probability that the least elements of the classes of $\Pi_n$ are $i_n = (i_1, \ldots, i_k)$, does not depend on $q_n$ and is given by

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

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

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

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 (22). Part two can be proved by using a representation theorem due to Pitman (24, 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 $i$, the right-hand side of (21) is equal to $a(\mathbf{n}, i_n)/a(\mathbf{n})$. Then, for every $\mathbf{n}$,

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

and

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

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

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

Also define, for $j = 1, \ldots, k-l+1, n^*_j := S_j^* - S_{j-1}^*$ with $S_j^* := S_{j+l+1} - i_j + 1$, and $i_j^* := i_{j+l+1} - i_j + 1$. Then $C_{[i_l]}(\mathbf{n}) = C(\mathbf{n}^*)$ so that, for a fixed $l \leq k$, the conditional probability of $i_2, \ldots, i_l$, given $\mathbf{n} = (n_1, \ldots, n_k)$, is

$$\mathbb{P}(i_2, \ldots, i_l|\mathbf{n}) = \frac{1}{n!} \sum_{C_{[i_l]}(\mathbf{n})} \prod_{j=1}^{k} \frac{(S_j - i_j)!}{(S_{j-1} - i_j + 1)!} (n - S_{j-1})$$

$$= \frac{(n - i_l)!}{n!} \prod_{j=1}^{l-1} \frac{(S_j - i_j)!}{(S_{j-1} - i_j + 1)!} (n - S_{j-1}) \frac{(n - S_{l-1})}{(S_{l-1} - i_l + 1)!}$$

$$\times \sum_{C(\mathbf{n}^*)} \frac{a(\mathbf{n}^*, i_2^{*}, \ldots, i_l^*)}{a(\mathbf{n}^*)},$$

where $n^* = n - i_l + 1$. The sum in (22) is 1; multiply and divide the remaining part by $\frac{(S_{l-1}^* - (S_l - i_1)!)/(|\mathbf{n}^*|)}$. The probability can therefore be rewritten as

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

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

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

then, for $l$ fixed,

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

$$= \prod_{j=2}^{l} W_{j-1}^{i_j - i_{j-1} - 1} (1 - W_{j-1}).$$
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 11 one recovers an identity due to Nacu (22, (7)).

**Corollary 1** (22, Proposition 5) For every sequence \( 1 = i_1 < i_2 < \ldots < i_k + 1 = n \), and every point \( x = (x_1, x_2, \ldots) \in \Delta \),

\[
\prod_{j=1}^{k-1} w_j^{i_{j+1} - i_j - 1} = \sum_{B(i)} \binom{n}{i} a(i, n) \prod_{j=1}^{k} x_j^{n_j - 1}
\]  

(23)

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

**Proof** Multiply both sizes by \( \prod_{1}^{k} (1 - w_{j-1}) \); by Proposition 1 and (12), formula (23) is just the equality (20) with the choice \( d\mu = \delta_x \).

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(\( \alpha, V \)). 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 \( n \) 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 \( I_n \), as explained in the following Lemma.

**Lemma 1** Let \( \Pi = (\Pi_n) \) be an EGP(\( \alpha, V \)), 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 \( \Pi_n \) are \( I_n = (i_1, \ldots, i_k) \) is

\[
\mu_{\alpha,V}(n) = \psi_{\alpha,n,k}^{-1}(i_n) V_{n,k}.
\]

(24)

where

\[
\psi_{\alpha,n,k}(i_n) = \frac{\Gamma(1 - \alpha)}{\Gamma(n - \alpha k)} \prod_{j=1}^{k} \frac{\Gamma(i_j - j\alpha)}{\Gamma(i_j - j\alpha - (1 - \alpha))}.
\]

(25)

Then the sequence \( i_1, i_2, \ldots \) forms a non-homogeneous Markov chain starting at \( i_1 = 1 \) and with transition probabilities given by

\[
P_j(i_{j+1} | i_j) = (i_j - \alpha j)(i_{j+1} - i_j - 1) V_{i_{j+1}, j+1} V_{i_j, j}, \quad j \geq 1.
\]

(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:

\[
P(K_{n+1} = k_{n+1} | K_n = k) = \left( 1 - \frac{V_{n+1,k}}{V_{n,k}} \right) I(k_{n+1} = k) + \frac{V_{n+1,k+1}}{V_{n,k}} I(k_{n+1} = k + 1).
\]

(27)

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

\[
\{K_{n+1} = K_n + 1 | K_n = k\} = \{i_k + 1 = n + 1\}.
\]

Therefore, from (27).
Remark 1 Notice that, as for $P_i$, vector $1_j$ of indices, is independent of $n$. For each $n$ \( V_{n,k} \) of indices, is independent of $V$ and is equal to

\[
\bar{\mu}_{\alpha,V}(i_n) = \prod_{j=1}^{k} \frac{V_{i,j}}{V_{i,j-1}} \prod_{l \notin i_n} (l - 1 - K_{l-1}\alpha) \frac{V_{i,K_{l-1}}}{V_{l-1,K_{l-1}}}
\]

\[
= V_{n,k} \prod_{l \notin i_n} (l - 1 - K_{l-1}\alpha).
\]

The last product in (28) is equal to

\[
\prod_{l \notin i_n} (l - K_{l-1}\alpha - 1) = \frac{\Gamma(n - \alpha k)}{\Gamma(1 - \alpha)} \prod_{j=2}^{k} \frac{\Gamma(i_{j} - j\alpha - (1 - \alpha))}{\Gamma(i_{j} - j\alpha)}.
\]

\[
= \psi^{-1}_{\alpha,n,k}(i_n).
\]

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

\[
\bar{\mu}_{\alpha,V}(i_{ik}) = \prod_{j=1}^{k-1} \mu_j(i_{j+1}|i_j),
\]

for $P_j$ satisfying (26) for every $j$.

The distribution of the age-ordered frequencies in an EGP $\Pi_n$, conditional on the record indices, can be easily obtained from Lemma 1 and (19).

**Proposition 2** Let $\Pi = (\Pi_n)$ be an EGP($\alpha,V$), 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 $n$ in age-order, given the vector $i_n$ of indices, is independent of $V$ and is equal to

\[
\bar{\mu}_\alpha(n|i_n) = \psi_{\alpha,n,k}(i_n) \left( \prod_{j=1}^{k} \frac{(S_j - i_j)}{n_j - 1} (1 - \alpha)^{(n_j-1)} \right).
\]

(30)

**Remark 1** Notice that, as $\alpha \to 0$, formula (30) reduces to that for Ewens’ partitions, proved in [14]:

\[
\bar{\mu}_0(n|i_n) = \prod_{j=1}^{k} \frac{(i_j - 1)}{2} \prod_{j=1}^{k} \frac{(S_j - i_j)!}{(S_j - i_j + i_j - 1)!}.
\]

**Proof** Recall that the probability of a pair $\langle n, i_n \rangle$ is given by

\[
\bar{\mu}_{\alpha,V}(n,i_n) = \binom{n}{n} a(n,i_n) q_{\alpha,V}(n).
\]

(31)

Now it is easy to derive the conditional distribution of a configuration given a sequence $i_n$, as:

\[
\bar{\mu}_\alpha(n|i_n) = \frac{\bar{\mu}_{\alpha,V}(n,i_n)}{\bar{\mu}_{\alpha,V}(i_n)}
\]

\[
= \binom{n}{n} \left( \prod_{j=1}^{k} \frac{(S_j - i_j)}{n_j - 1} (1 - \alpha)^{(n_j-1)} \right) \frac{V_{n,k}}{\bar{\mu}_{\alpha,V}(i_n)}
\]

\[
= \psi_{\alpha,n,k}(i_n) \left( \prod_{j=1}^{k} \frac{(S_j - i_j)}{n_j - 1} (1 - \alpha)^{(n_j-1)} \right),
\]

(32)

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 \( i = (i_1 < i_2 < \ldots) \) generated by an EGP(\( \alpha, V \)).

**Proposition 3** Let \( \Pi = (\Pi_n)_{n \geq 1} \) be an Exchangeable Gibbs Partition with index \( \alpha > -\infty \) for some \( V \). Let \( i = (i_1 < i_2 < \ldots) \) 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

\[
X_j \overset{d}{=} \xi_{j-1} \prod_{m=j}^{\infty} (1 - \xi_m), \quad j \geq 1, \tag{33}
\]

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

**Remark 2** Proposition 3 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 \( \alpha \), as a consequence of the limit behavior of \( K_n \), the number of blocks of an EGP \( \Pi_n \), as recalled in the introduction. For \( \alpha < 0 \), \( \mathbf{i} \) is almost surely a finite sequence; for nonnegative \( \alpha \), the length \( l \) of \( \mathbf{i} \) will be a.s. either \( l \sim s \log n \) (for \( \alpha = 0 \)) or \( l \sim n^{\alpha} \) (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

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

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 \( \mathbf{i}_n' = (i'_1, \ldots, i'_r) \) where

\[
i'_j = i_j + \sum_{i=1}^{j} a_i. \quad (j = 1, \ldots, k).
\]

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

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

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

\[
\frac{\psi_{\alpha,n,k}(\mathbf{i}_n)}{\psi_{\alpha,n,k}(\mathbf{i}_n')} = \prod_{j=2}^{k} \frac{[i_j - j\alpha - (1 - \alpha)](\sum_{i=1}^{j} a_i)}{(i_j - j\alpha)(\sum_{i=1}^{j} a_i)}
\]

\[= \prod_{j=1}^{k-1} \mathbb{E}((1 - \xi_j)^{\sum_{i=1}^{j} a_i}), \tag{36}\]

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 \((36)\), \( S_0 = 0, S_k = n \), so the product is equal to

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

\[
E \left( \prod_{j=1}^{r-1} W_j^{a_{j+1}} \right) = \prod_{j=2}^{\infty} E((1 - \xi_j) \sum_{i=1}^{\infty} \alpha_i)
\]

which gives the limit distribution of the cumulative sums:

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

But

\[
X_j = W_j - W_{j-1} = \prod_{i=j}^{\infty} (1 - \xi_j) - \prod_{i=j-1}^{\infty} (1 - \xi_j) = \xi_{j-1} \prod_{i=j}^{\infty} (1 - \xi_i), \quad j = 1, 2, \ldots
\]

and the proof is complete.

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

Proposition 3 says that, conditional on all the record indices \( i_1, i_2, \ldots \), the sequence of relative increments of an EGP(\( \alpha, V \))

\[
\xi = \left( \frac{X_2}{W_2}, \frac{X_3}{W_3}, \ldots \right).
\]

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 [29], [30] for a modern account).

It is possible to show that such an independence property of the \( \xi \) 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 [30].

**Definition 1** Let \( k \) be any fixed positive integer (non necessarily finite).

(i) Let \( P = (P_1, P_2, \ldots, P_k) \) be a random point in \([0, 1]^k\) such that \( \sum_{i=1}^{k} P_i \leq 1 \) and, for every \( j = 1, \ldots, k - 1 \) denote \( F_j = \sum_{i=1}^{j} P_i \). \( P \) is called a Neutral to the Right (NTR) sequence if the vector \( B_1, B_2, \ldots, B_{k-1} \) of relative increments

\[
B_j = \frac{P_j}{1 - F_{j-1}}, \quad j = 1, \ldots, k - 1
\]

is a sequence of independent random variables in \([0, 1]\).

Let \((\alpha, \beta)\) be a point in \([0, \infty)^{k-1}\). A NTR vector such that every increment \( B_j \) is a Beta \((\alpha_j, \beta_j)\) \((j = 1, \ldots, k - 1)\), is called a Beta-Stacy distribution with parameter \((\alpha, \beta)\).
An equivalent characterization of ISBP measure is:

A Left-Beta-Stacy distribution with parameter \((\alpha, \beta)\) is a NTL vector \(P\) such that \(P^* := (P_k, P_{k-1}, \ldots, P_1)\) is NTR.

A known result due to [25] 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\((\alpha, V)\) with V-coefficients given by [1]. 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 shows that, conditional on the record indices \(i_1, i_2, \ldots, i_k\) and on \(W_k\) they are all NTL distributions. For a fixed \(k\) set

\[ Y_j = \frac{X_{k-j+1}}{W_k}, \quad 1 \leq j \leq k. \]

Then

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

and

\[ \xi_{k-j} = \frac{Y_j}{1 - F_{j-1}}. \]

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 \(W_k \to 1\) where \(K_\infty = \lim_{n \to \infty} K_n\) ).

The following proposition is a converse of Proposition 3

**Proposition 4** Let \(X = (X_1, X_2, \ldots) \in \Delta\) be the age-ordered frequencies of an infinite exchangeable random partition \(\Pi\) of \(\mathbb{N}\). Assume, conditionally on the record indices of \(\Pi\), \(X\) is a NTL sequence. Then \(\Pi\) is an exchangeable Gibbs partition for some parameters \((\alpha, V)\).

**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], [25]). 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 2** Invariance under size-biased permutation ([25], Theorem 4). Let \(X\) be a random point of \([0, 1]^\infty\) such that \(|P| \leq 1\) almost surely with respect to a probability measure \(dp\). For every \(k\), let \(\mu_k\) denote the distribution of \(X_1, \ldots, X_k\), and \(G_k\) the measure on \([0, 1]^k\), absolutely continuous with respect to \(\mu_k\) 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 \(\Pi\), and denote with \(\mathbb{P}_\mu\) the marginal law of the record indices of \(\Pi\). Consider the measure \(G_k\) of Lemma 2. By Proposition 4 (ii), for every \(k\)

\[ G_k(dx_1 \times \cdots \times dx_k) = \mu_k(dx_1 \times \cdots \times dx_k) \prod_{i=1}^{k-1} (1 - w_{j}) = \mu_k(dx_1 \times \cdots \times dx_k) \mathbb{P}(i_1 = 1, i_2 = 2, \ldots, i_k = k \mid x_1, \ldots, x_k) = \mu_k(dx_1 \times \cdots \times dx_k) \mathbb{P}_\mu(i_k = k) \]

An equivalent characterization of ISBP measure is:
Corollary 2 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

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

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

The other result we recall is about Dirichlet distributions.

Lemma 3 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 $P_n$ does not depend on $(1 - X_k)^{-1}(P_1, \ldots, P_k)$. Then $X$ has the Dirichlet distribution.

Now we have all elements to prove Proposition 4. Let $\mu(\cdot | i_1, i_2, \ldots)$ be the distribution of a NTL vector such that the distribution of $\xi_j := X_{j+1}/W_{j+1}$ has marginal law $d\gamma_j$ 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 $\zeta_k$ is the conditional law of $W_k$ given $i_k = k$.

For $X$ to be ISBP, corollary 2 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 $W_k^{-1}(X_1, \ldots, X_{k-1})$. Therefore, by Lemma 3 and symmetry, $(X_1/W_k, \ldots, X_k/W_k)$ 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

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

(41)

By the NTL assumption, we can write

$$\prod_{j=1}^{k} X_j^{n_j-1} = \prod_{j=2}^{k} \left( \frac{X_j}{W_j} \right)^{n_j-1} \left( \frac{W_j}{W_k} \right)^{n_j-1} W_k^{n_j-k},$$

$$\prod_{j=2}^{k} \left( \frac{X_j}{W_j} \right)^{n_j-1} W_k^{n_j-k-1} (1 - \xi_j)^{S_j-j} W_k^{n_j-k},$$

(42)

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

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

Now, set

$$V_{n,k} = \frac{\mathbb{P}_\mu(i_k = k) \mathbb{E}(W_k^{n_k-k} | i_k = k)}{[k(1 - \alpha)](n_k - k)}.$$
Equality (42) implies
\[
\mathbb{E} \left[ \prod_{j=1}^{k} X_j^{n_j-1} (1 - W_{j-1}) \right] = \mathbb{P}(i_k = k) \int \left( \prod_{j=1}^{k-1} \xi_j^{n_j+1-1} (1 - \xi_j)^{S_j - y_j} \xi_j d\xi_j \right) \int \frac{w_{k}^{n_k-1} \zeta_k(w_1) d\omega_k}{\omega_k^{n_k-1}}
\]
\[
= \mathbb{P}(i_k = k) \mathbb{E}(W_k^{n_k-1} | i_k = k) \prod_{j=1}^{k-1} \frac{(1 - \alpha)(n_{j+1} - 1)}{[j + 1](1 - \alpha)(S_j - j)}
\]
\[
= V_{n,k} \prod_{j=1}^{k} (1 - \alpha)(n_j - 1),
\]
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 (\(m = 1, 2, \ldots\)) can be derived by using Proposition 3. We first point out a characterization for the moments of \(W_m\), stated in the following Lemma.

**Lemma 4** Let \(X_1, X_2, \ldots\) be the limit age-ordered frequencies generated by a Gibbs partition with parameters \(\alpha, V\). For every \(m = 1, 2, \ldots\) and nonnegative integer \(n\)

\[
\mathbb{E}(W_m^n | i_m) = (i_m - \alpha m)_n \frac{V_{n+m,m}}{V_{m,m}}
\]

and

\[
\mathbb{E}(X_m^n | i_m) = (1 - \alpha)_n \frac{V_{n+m,m}}{V_{m,m}}.
\]

**Proof** Let \(\Pi\) be an EGP \((\alpha, V)\) 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 \(\Pi\). Then \(Y_1 = 1\) and, for every \(l, m \leq l\),

\[
\mathbb{P}(Y_{l+1} = 0 | \sum_{i=1}^{l} Y_i = m) = (l - \alpha m) \frac{V_{l+1,m}}{V_{l,m}}
\]

\[
= 1 - \mathbb{P}(Y_{l+1} = 1 | \sum_{i=1}^{l} Y_i = m).
\]

By proposition 1 and formula (13), given the cumulative frequencies \(W = W_1, W_2, \ldots\),

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

(see also [24], 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 \(l, m\)

\[
\mathbb{P}(Y_{l+1} = 0 | K_l = m, K_{l-1} = m - e, W) = \mathbb{P}(Y_{l+1} = 0 | K_l = m, W) = W_m, \quad e = 0, 1.
\]

Hence

\[
\mathbb{E}(W_m | i_m) = \mathbb{E} \left[ \mathbb{P}(Y_{m+1} = 0 | K_m = m, K_{m-1} = m - 1, W) \right]
\]

\[
= \mathbb{E} \left[ \mathbb{P}(Y_{m+1} = 0 | K_m = m, W) \right]
\]

\[
= (i_m - \alpha m) \frac{V_{m+1,m}}{V_{m,m}},
\]

(46)
which proves the proposition for \( m = 1 \). The Markov property of \( K_n \) and \( \{45\} \) also lead, for every \( m \), to
\[
\mathbb{E}(W_m^\phi | i_m) = \mathbb{E} [ \mathbb{P}(Y_{i_m+1} = \ldots = Y_n = 0 \mid K_{i_m} = m, K_{i_m-1} = m-1, W) ] \\
= \prod_{i=1}^{n} \mathbb{E} [ \mathbb{P}(Y_{i+m+i} = 0 \mid K_{i+m+i-1} = m, W) ] \\
= (i_m - \alpha m) \frac{V_{i_m+n,m}}{V_{i_m,m}},
\]
where the last equality is obtained as an \( n \)-fold iteration of \( \{46\} \).

The second part of the Lemma (formula \( \{44\} \)) follows from Proposition \( \{3\} \):
\[
\mathbb{E}(X_m^n | i_m) = \mathbb{E}(X_m^{n-1} | i_m) \mathbb{E}(W_m^\phi | i_m) \\
= \frac{(1 - \alpha)(\phi)}{(i_m - \alpha m)(\phi)} \mathbb{E}(W_m^\phi | i_m)
\]
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 5** Let \( X_1, X_2, \ldots \) be the limit age-ordered frequencies generated by a Gibbs partition with parameters \( \alpha, V \). For every \( m = 1, 2, \ldots \) and \( \phi \geq 0 \)
\[
\mathbb{E}(W_m^\phi | i_m) = (i_m - \alpha m) \frac{V_{i_m,m}[\phi]}{V_{i_m,m}}
\]
and
\[
\mathbb{E}(X_m^\phi | i_m) = (1 - \alpha) \frac{V_{i_m,m}[\phi]}{V_{i_m,m}}
\]
for a sequence of functions \( \{V_{n,k}[\cdot] : \mathbb{R} \rightarrow \mathbb{R}; k, n = 1, 2, \ldots\} \), uniquely determined by \( V_{n,k} \equiv V_{n,k}[0] \), such that, for every \( \phi \geq 0 \),
\[
V_{1,1}[\phi] = \frac{\mathbb{E}(X_1^\phi)}{(1 - \alpha)(\phi)}; \\
V_{n,k}[\phi] = (n + \phi - \alpha k)V_{n+1,k}[\phi] + V_{n+1,k+1}[\phi], \quad n, k = 1, 2, \ldots; \\
V_{n,k}[\phi + 1] = V_{n+1,k}[\phi], \quad n, k = 1, 2, \ldots.
\]

**Remark 3** To complete the representation given in Proposition \( \{5\} \) notice that, for every \( \alpha \), the distribution of \( X_1 \) (the so-called *structural distribution*) is known for the extreme points of the Gibbs(\( \alpha, V \)) 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. \( \{26\} \)). 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 (\( \alpha, s \)) partition for some \( s > 0 \); in this case the density of \( X_1 \) is
\[
f_{1,\alpha}(x|s) = \frac{\alpha s^{-1}(x-\alpha)^{-\alpha} f_\alpha((1-x)s^{1/\alpha})}{\Gamma(1 - \alpha) f_\alpha(s^{1/\alpha})}, \quad 0 < x < 1,
\]
for an \( \alpha \)-stable density \( f_\alpha \) (\( \{27\}, \{57\} \)), leading to
\[
V_{1,1}[\phi](s) = \alpha s^{\frac{\phi - \alpha}{\alpha}} G_\alpha(\phi - \alpha - 1, s^{-1/\alpha})
\]
where \( G_\alpha \) is as in \( \{7\} \).

Thus, for every \( \alpha \), the structural distribution of a Gibbs(\( \alpha, V \)) 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 with $V_{n,m}[\phi] \equiv V_{n+\phi,m}$. For general $\phi \geq 0$ observe that, for every $m,n \in \mathbb{N}$,

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

(52)

To see this, consider the random sequences $Y_n, K_n$ defined in the proof of Lemma 4. By (48),

$$
E(W_m^{\phi+n-m}|i_m = m) = E[W_m^{\phi}1_{\{K_m = m\}|K_m = m, W}| K_m = m]
$$

$$
= E[W_m^{\phi}1_{\{K_m = m\}}| W, K_m = m]
$$

$$
= E[W_m^{\phi}1_{\{K_m = m\}}| W, K_m = m, | E[\mathbb{P}(K_n = m|W)|K_m = m]
$$

$$
= E[W_m^{\phi}1_{\{K_m = m\}}| W, i_m = m]E[W_m^{n-m}|K_m = m]
$$

where the last two equalities follow from (55), the Markov property of $K_n$ and the exchangeability of the $Y$'s.

From Lemma 4, we can rewrite (52) as

$$
E(W_m^{\phi}|i_m = n) = \frac{E(W_m^{\phi+n-m}|i_m = m)}{E(W_m^{\phi+n-m}|i_m = m)} V_m V_n.
$$

(53)

Now define

$$
M_m(\phi) = E(W_{m+1}^{\phi}|i_m = m), \quad \phi \geq 0, m = 1, 2, \ldots
$$

(54)

and

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

(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

$$
E(W_m^{\phi}|i_m) = (i_m - \alpha m)(\phi) V_{n,m}[\phi],
$$

(56)

that is: (55), (56) and are satisfied.

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

$$
M_m(\phi) - M_m(\phi + 1) = E[W_m^{\phi}(1 - W_m)|K_m = m]
$$

$$
= E[W_m^{\phi}(1 - \mathbb{P}(K_{m+1} = m|W))| K_m = m]
$$

$$
= E[W_m^{\phi}1_{\{K_m + 1 = m + 1, K_m = m\}}|1 - \mathbb{P}(K_{m+1} = m|W)| K_m = m]
$$

$$
= E[(1 - \xi_m)^{\phi}|i_m + 1 = m + 1] E[W_m^{\phi}1_{K_m + 1 = m + 1}] V_{m+1,m+1} V_{m,m}.
$$

(57)

The last equality is a consequence of Proposition 3 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 $\xi_m$ (independent of $W_{m+1}$) having a Beta$(1 - \alpha, m(1 - \alpha))$ distribution. Therefore (57) can be rewritten as

$$
M_m(\phi) - M_m(\phi + 1) = \frac{m(1-\alpha)(\phi)}{(m+1)(1-\alpha)(\phi)} M_m(\phi) V_{m+1,m+1} V_{m,m}.
$$

(58)

and (50) follows from (58), (51), after some simple algebra, by comparing the definition (55) of $V_{n,k}[\phi]$, and the recursion (53) for the $V$-coefficients of an EGP($\alpha, V$). In particular, (58) shows that the functions $V_{n,k}[\phi]$ are uniquely determined by $V = (V_{n,k})$.

The equality (58) can now be proved in the same way as the moment formula (44).
4.1 Convolution structure.

An alternative representation of $-\log X_m$ can be given in terms of an infinite sum of random variables, as done by [14] for Kingman’s model (Ewens’ partition). Formula (42) in [14] shows that, conditional on 

$$ -\log(X_j) = \sum_{i=2}^{i_j} Y_i + \sum_{m=i_j+1}^{\infty} Z_m, $$

(59)

where $\{Y_i\}$ is a sequence of independent exponential random variables such that the $i$-th component has rate $i - 1$, and $\{Z_m\}$ is a sequence of mutually independent random variables with density continuous everywhere in the non-negative reals except one atom at zero:

$$ f_i(w) = \frac{1}{1 + \rho_i} I(w = 0) + \frac{\rho_i}{1 - \rho_i} (i - 1)e^{-(i-1)w}I(w > 0), $$

(60)

where $\rho_i = \frac{d}{i - 1}$ ([14], p. 173). With the help of (53) it is possible to show that a similar representation holds true for general Gibbs partitions, but that the collection of the $Y$’s is split in two subsequences. It turns out that the independence property of the $Z$’s actually characterizes Kingman’s model: for general Gibbs partitions with parameters $\alpha, V$, the $Z$’s are stochastically linked by the sequence $K = (K_n : n = 1, 2, \ldots)$ as defined in the proof of Proposition 2. Remember that $K$ is a Markov process on $N$ such that $K_1 = 1$ and for $j > 1$

$$ p_{kj} = P(K_j = k + 1|K_{j-1} = k) = 1 - P(K_j = k|K_{j-1} = k) $$

$$ = \frac{V_{j,k+1}}{V_{j-1,k}}, $$

(61)

**Proposition 6** Let $X = X_1, X_2, \ldots$ be the limit relative frequencies generated by a Gibbs partition with parameters $\alpha, V$. Conditionally on $i_m$,

$$ -\log X_m = \sum_{j=2}^{m} Y_j^* + \sum_{i=m+1}^{i_m} Y_i^{(m)} + \sum_{l=i_m+1}^{\infty} Z_l $$

(62)

where: $\{Y_j^*\}$ are independent random variables, with distribution, respectively,

$$ f_j^{(\alpha)}(y)dy = \frac{e^{-y(j-1-a(j-1))(1-e^{-y})^{-\alpha}}}{B(j-1-a(j-1), 1-\alpha)} \mathbb{I}(y > 0)dy; $$

(63)

$\{Y_i^{(m)}\}$’s are independent random variables, with distribution, respectively,

$$ f_i^{(\alpha, m)}(y)dy = \frac{e^{-y(i-1-a\delta_m)}}{i-\alpha\delta_m - 1} \mathbb{I}(y > 0)dy, $$

(64)

and $\{Z_l\}$ are such that such that: (i) $Z_l$ is conditionally independent of $Z_{l-1}$, given $K_{l-1}$; (ii) conditionally on $\{K_{l-1} = k, K_l = v\}$, the density of $Z_l$ is

$$ g_{i,k,v}^{\alpha,V}(z)dz = \delta_{k,v} \delta_{0,1} dz + (1 - \delta_{k,v}) \frac{e^{-z(l-1-a\delta_m)}(1-e^{-z})^{-\alpha}}{B(l-1-a\delta, 1-\alpha)} \mathbb{I}(z > 0)dz. $$

(65)

**Proof** From Proposition (53), and the properties of Beta distributions,

$$ \mathbb{E}(e^{\phi} \log X_m | i_m) = \mathbb{E}(e^{\phi} \prod_{i=m}^{\infty} (1 - \xi_i)^{\phi_i} | i_m) $$

$$ = \frac{(1-\alpha)^{\phi}}{(i_m - \alpha\delta_m)^{\phi}} \mathbb{E} \left( \prod_{k=m}^{\infty} (i_{k+1} - ak - 1)^{\phi_i} | i_m \right). $$

(66)
The fraction on the left-hand side can be decomposed as:

\[
\frac{(1 - \alpha)_{(i_m - a m)_{(\phi)}}}{(i_m - a m)_{(\phi)}} = \left( \prod_{j=2}^{m} \frac{(j - 1 - \alpha(j - 1))_{(\phi)}}{(j - 1)_{(\phi)}} \right) \left( \prod_{i=m+1}^{m} \frac{(i - 1 - a m)_{(\phi)}}{(i - a m)_{(\phi)}} \right) \\
= \left( \prod_{j=2}^{m} E(\bar{Y}_{j}^{*}) \right) \left( \prod_{i=m+1}^{m} E(\bar{Y}_{i}^{(m)\phi}) \right),
\]

where, for each \( j \),

\[
\bar{Y}_{j}^{*} \sim \text{Beta}(j - 1 - \alpha(j - 1), 1 - \alpha),
\]

and for each \( i \),

\[
\bar{Y}_{i}^{(m)} \sim \text{Beta}(j - 1 - a m, 1).
\]

Now simply set \( Y_{j}^{*} := -\log \bar{Y}_{j}^{*} \) and \( Y_{i}^{(m)} := -\log \bar{Y}_{i}^{(m)} \) to see that

\[
\frac{(1 - \alpha)_{(i_m - a m)_{(\phi)}}}{(i_m - a m)_{(\phi)}} = E(e^{-\phi\left(\sum_{j=2}^{m} Y_{j}^{*} + \sum_{i=m+1}^{m} Y_{i}^{(m)}\right)}),
\]

which provides the first two sums in the representation (62). Define \( \chi_{l} = K_{l} - K_{l-1} \) where \( K_{l} \) is described by (61). Then

\[
E\left( \prod_{k=m}^{\infty} \frac{(i_{k+1} - a k - 1)_{(\phi)}}{(i_{k+1} - a k)_{(\phi)}} | i_{m} \right) = E\left( \prod_{l=m+1}^{\infty} \frac{(l - \alpha K_{l} - (1 - \alpha)\phi\chi_{l})_{(\phi)}}{(l - \alpha(k)\phi\chi_{l})_{(\phi)}} | i_{m} \right) \\
= E\left( \prod_{l=m+1}^{\infty} E(Z_{l}^{(\phi)\chi_k} | K_{l}, K_{l-1}) | i_{m} \right),
\]

where each \( Z_{l} \) is a Beta random variable with parameters \( (l - \alpha K_{l} - (1 - \alpha), 1 - \alpha) \). Now define

\[
Z_{l} = -\chi_{l} \log \bar{Z}_{l}.
\]

Then \( Z_{l} \) has the required distribution (63) and the last expectation can be written as

\[
E\left( \prod_{l=m+1}^{\infty} E(e^{-\phi Z_{l} K_{l}, K_{l-1}) | i_{m} \right) = E\left( e^{-\phi \sum_{l=m+1}^{\infty} Z_{l} | i_{m} \right)
\]

which completes the proof.

The characterization of Kingman’s model follows immediately.

**Corollary 3** Let \( X = X_{1}, X_{2}, \ldots \) be the limit relative frequencies generated by a Gibbs partition with parameters \( \alpha, V \). The random variables involved in the three sums in (62) are all mutually independent if and only if \( \alpha = 0 \) and \( V \) is such that, for every \( n, k \leq n, V_{n,k} = \theta^{k}/\theta(n) \) for some \( \theta > 0 \).

**Proof** We only have to verify the independence of the \( Z \)’s. In general, \( Z_{l} \) depends on \( Z_{l-1} \) only through \( K_{l-1} \) \( l > i_{m} \). By averaging over \( K_{l} \), one has that the conditional density of \( W_{l} \) given \( \{K_{l-1} = k\} \) is:

\[
(1 - p_{kl})\delta_{0}(dz) + p_{kl} e^{-z(l-1-\alpha k)(1-e^{-z})} \frac{1}{B(l-1-\alpha, 1-\alpha)} I(z > 0)dz.
\]

But remember that

\[
1 - p_{kl} = (l - \alpha k - 1) \frac{V_{l,k}}{V_{l-1,k}}
\]

does not depend on \( k \) only if \( \alpha = 0 \) and

\[
\frac{V_{l,k}}{V_{l-1,k}} = \frac{c_{l-1}}{c_{l}} \quad k, l = 1, 2, \ldots
\]
for some constant $c_i$. This implies that the $V$-coefficients are of the form

$$V_{n,k} = c_n^{-1}V^*_k;$$

but, as mentioned in the introduction, within the family of all EGP's, only Pitman's two-parameter partitions have such a form. Since we just saw that $\alpha$ is necessarily zero, then it must be

$$V_{n,k} = \frac{\theta^k}{\theta(n)}$$

for some $\theta > 0$, which corresponds exactly to Kingman's model. In this case, the density \[65\] reduces to

$$\frac{l - 1}{l + \theta - 1} \delta_{\theta} (dz) + \frac{\theta}{l + \theta - 1} e^{-z(l-1)}. \quad (69)$$

The converse is straightforward.

Remark 4 Notice that, when $\alpha = 0$, the $Y^*_j$'s and the $Y_j^{(m)}$ are all exponential random variables, each with parameter $(j - 1)$, respectively. This, together with \[69\], leads us back to Griffiths and Lessard's representation \[59\]-\[60\].

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 $g_k$ 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, 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. \[20\]).

Proposition 7 Let $X_1, X_2, \ldots$ be the age-ordered frequencies of a EGP($\alpha, V$) 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

$$d\mu_{\alpha,V}(x_1, \ldots, x_k | w, k+i) = \frac{V_{k+i,k}}{V_{k+i-1,k-1}} w^{-(k-1)} \sum_{m \in \mathbb{N}^{k-1}: \|m\|=k+i-1} \tilde{\mu}_{\alpha,V}(m) D_{(m-\alpha)} \frac{x_1}{w}, \ldots, \frac{x_k}{w} dx_1, \ldots, dx_{k-1}$$

where: $D_{(m-\alpha)}$ is the Dirichlet density with parameters $(m_1 - \alpha, \ldots, m_k - \alpha, 1 - \alpha)$, and $\mu_{\alpha,V}(m)$ is the age-ordered sampling formula \[13\] with Gibbs' EPPF $q_{\alpha,V}$ as in \[13\]:

$$\tilde{\mu}_{\alpha,V}(m) = \left( \frac{(k + i - 1)!}{\prod_{j=1}^{k-1} (m_j)! (k + i - 1 - \sum_{i=1}^{j-1} m_i)!} \right) \frac{V_{k+i-1,k-1}}{V_{k+i,k-1}} \prod_{j=1}^{k-1} (1 - \alpha)_{(m_j-1)}.$$  \quad (70)

(ii) The marginal distribution of $i_k$ is

$$P(i_k = k + i) = V_{k+i,k} \frac{\alpha^{-(k-1)}}{(k + i - 1)!} \sum_{j=0}^{k-1} \frac{(-1)^{j+i+1}}{j! (k - j - 1)! (\alpha)^{(j)}_{(k+i)}}, \quad (71)$$

where $a_{[n]} = a(a-1) \cdots (a-n+1)$. \[19\]
Remark 5 The density of \( i_k \) can be expressed in terms of generalized Stirling numbers \( \left[ \frac{n}{k} \right]_{\alpha} \), defined as the coefficients of \( x^n \) in

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

(17, 13). Formula (71) can be re-expressed as:

\[
\mathbb{P}(i_k) = V_{i_k, k} \left[ \begin{array}{c} i_k - 1 \\ k - 1 \end{array} \right]_\alpha,
\]

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, (71) can be deduced simply from (8) by the Markov property of the sequence \( K_n \) as

\[
\mathbb{P}(i_k) = \mathbb{P}(K_{i_k} = k | K_{i_k - 1} = k - 1) \mathbb{P}(K_{i_k - 1} = k - 1) = \frac{V_{i_k, k}}{V_{i_{k-1}, k-1}} \left[ \begin{array}{c} i_k - 1 \\ k - 1 \end{array} \right]_\alpha.
\]

However here we give a self-contained proof in order to show how (71) is implied by proposition 3 through (70).

Proof From (10), we know that

\[
P(i_2, \ldots, i_k) = V_{i_k, k} \frac{\Gamma(i_2 - \alpha - 1) \cdots \Gamma(i_k - \alpha(k - 1) - 1)}{\Gamma(1 - \alpha) \Gamma(i_2 - 2\alpha) \cdots \Gamma(i_k - (k - 1)\alpha)}.
\]

By proposition 3 and Lemma 4

\[
\mathbb{E}\left( \prod_{j=1}^k X_j^{n_j} | i_1, \ldots, i_k \right) = \mathbb{E}\left( \prod_{j=1}^{k-1} \xi_j^{n_{j+1}} (1 - \xi_j)^{S_j} \prod_{i=k}^{\infty} (1 - \xi_i) | i_1, \ldots, i_k \right)
\]

\[
= \prod_{j=1}^{k-1} \frac{(1 - \alpha)(n_{j+1} - j\alpha - 1)(s_{j+1})}{(i_{j+1} - (j + 1)\alpha)(s_{j+1})} \mathbb{E}\left( \prod_{i=k}^{\infty} (1 - \xi_i) | i_k \right)
\]

\[
= \left( \prod_{j=1}^{k-1} \frac{(1 - \alpha)(n_{j+1} - j\alpha - 1)(s_{j+1})}{(i_{j+1} - (j + 1)\alpha)(s_{j+1})} \right) \frac{\Gamma(i_k + n - \alpha k)}{\Gamma(i_k - \alpha k)} \frac{V_{n+i_k, k}}{V_{i_k, k}}.
\]

hence

\[
\mathbb{E}\left( \prod_{j=1}^k X_j^{n_j} | i_1 \equiv 1, i_2, \ldots, i_k \right) P(i_2, \ldots, i_k)
\]

\[
= (i_k - \alpha k)(n_k) V_{n+i_k, k} \left( \prod_{j=1}^{k-1} \frac{(1 - \alpha)n_{j+1} \Gamma(i_{j+1} - j\alpha - 1 + S_{j+1}) \Gamma(i_{j+1} - (j + 1)\alpha)}{\Gamma(i_{j+1} - \alpha(j + 1) + S_{j+1}) \Gamma(i_j - \alpha j)} \right)
\]

\[
= \prod_{j=1}^k (1 - \alpha)(n_j) \left( \prod_{j=1}^{k-1} \frac{(i_j - \alpha j + S_j)(i_{j+1} - i_j - 1)}{V_{i_k + n_k, k} \prod_{j=1}^{k-1} (i_j - \alpha j + S_j)(i_{j+1} - i_j - 1)} \right),
\]

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:

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

(74)

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 (74) reads
\[ E(\alpha,\nu) \left( \prod_{j=1}^{k} X_j^{\nu_j} \right) = \prod_{j=1}^{k} (1 - \alpha)_{(\nu_j)} \sum_{i=k}^{\infty} \lambda_i V_{n+i,k}. \] (75)

For Pitman’s two-parameter family, this becomes
\[ E(\alpha,\theta) \left( \prod_{j=1}^{k} X_j^{\nu_j} \right) = \prod_{j=1}^{k} (1 - \alpha)_{(\nu_j)} \sum_{i=k}^{\infty} \lambda_i \frac{\prod_{j=1}^{k} (\theta + \alpha(j - 1))}{\theta_{(n+i)}}. \] (76)

The two-parameter GEM distribution implies that
\[ E \left( \prod_{j=1}^{k} X_j^{n_j} \right) = E \left( \prod_{j=1}^{k} B_j^{n_j}(1 - B_j)^{n - S_j} \right) = \prod_{j=1}^{k} \frac{(1 - \alpha)_{(\nu_j)}(\theta + j\alpha)_{(n - S_j)}}{(\theta - j\alpha)_{(n - S_{j-1})}} = \frac{1}{\theta_{(n)}} \prod_{j=1}^{k} (1 - \alpha)_{(\nu_j)}(\theta + \alpha(j - 1)). \] (77)

Therefore, from (76) we derive the identity:
\[ \prod_{j=1}^{k} \frac{1}{(\theta + \alpha(j - 1) + n - S_{j-1})} = \sum_{i=k}^{\infty} \lambda_i \frac{\theta_{(n)}}{\theta_{(n+i)}}, \quad \theta > 0. \] (78)

For $\theta > 0$ replace $\theta$ by $\theta - n$ in (78), and denote $n^*_j = 1 - \alpha + n_j$. We now find an expansion of the left-hand side of (78) in terms of products of the type $\prod_{i=1}^{k} (n^*_j)_{(m_j)}$, for $m_1, \ldots, m_k \geq 0$. The left-hand side of (78) is now
\[ \prod_{j=1}^{k} \frac{1}{(\theta + \alpha(j - 1) + n - S^*_j)} = \prod_{j=1}^{k} \frac{1}{(\theta - j\alpha + n^-)} \]
\[ = \prod_{j=1}^{k} \int_{0}^{1} t_{j-1}^{\theta - j - 1} dt_{j-1} = \int \left( \prod_{j=1}^{k-1} t_j \right)^{\theta - 1} (t_1 \cdots t_{k-1})^{\sum_{j=1}^{k-1} n^*_j} \cdots (t_{k-1})^{\sum_{j=1}^{k-1} n^*_j} \prod_{j=1}^{k-1} t_j^{\theta} dt_0 \cdots dt_{k-1}, \] (79)

where $S^*_0 = 0$ and $S^*_j = \sum_{i=1}^{j} n^*_i$, $j = 1, \ldots, k - 1$. Make the change of variable
\[ u_j = 1 - \prod_{i=j}^{k-1} t_i, \quad j = 0, \ldots, k - 1. \]

Then $0 < u_{k-1} < \ldots < u_0 < 1$. The absolute value of the Jacobian is
\[ \left| \frac{du}{dt} \right| = (t_1 \cdots t_{k-1}) \times (t_2 \cdots t_{k-1}) \times t_{k-1} \times 1 \]
\[ = \prod_{j=1}^{k-1} t_j^j. \]
Thus (79) transforms to
\[
\int_{0 < u_{k-1} < \ldots < u_0 < 1} (1 - u_0)^{\theta - 1} \prod_{j=1}^{k-1} (1 - u_j)^{-n_j^*} du_1 \cdots du_{k-1} du_0.
\]

Fix \(u_0\) and consider
\[
\int_{0 < u_{k-1} < \ldots < u_0} \prod_{j=1}^{k-1} (1 - u_j)^{m_j} du_1 \cdots du_{k-1}
\]
\[
= u_0^{k-1 + \sum_{i=1}^{k-1} m_i} \prod_{j=1}^{k-1} \frac{1}{m_{j-1} + 1} \cdot \frac{1}{m_{j-1} + m_{j-2} + 1} \cdot \frac{1}{m_{j-1} + \ldots + m_1 + 1}.
\]
The integral in (79) is thus
\[
\sum_{m_1, \ldots, m_{k-1} \geq 0} \frac{\lambda^{k-1} (1 - \alpha)(m_j)}{m_j!} \prod_{j=1}^{k-1} n_j^* \int_0^1 (1 - u_0)^{\theta - 1} u_0^{k-1 + \sum_{i=1}^{k-1} m_i} du_0,
\]
where
\[
c(m) = \prod_{j=1}^{k-1} \frac{1}{k - j + \sum_{i=j}^{k-1} m_i} \prod_{j=1}^{k-1} \frac{1}{m_j}.
\]
Now consider the right-hand side of (78), again with \(\theta\) 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 (80) to obtain a representation for the \(\lambda_i\)’s in (75):
\[
\frac{\lambda_{k+i}}{(k + i - 1)!} = \sum_{m \in \mathbb{N} \cup \{0\}} \frac{c(m)}{(1 - \alpha)(m_j)} \prod_{j=1}^{k-1} n_j^*,
\]
where \(\mathbb{N} = \mathbb{N} \cup \{0\}\). Recall that \(n_j^* = 1 - \alpha + n_j\) and consider the identity
\[
\frac{(1 - \alpha)(n_j)n_j^*}{(1 - \alpha)(m_j)} = (1 - \alpha + m_j)(n_j).
\]
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 (78) and (82) imply that
\[
\mathbb{E}_{(\alpha, \nu)} \left( \prod_{j=1}^{k} X_j^{n_j} \right) = \sum_{i=0}^{\infty} (k + i - 1)! \mathbb{E}(W_k^n | i_k = k + i) V_{k+i, k}
\]
\[
\times \sum_{m \in \mathbb{N} \cup \{0\}^k : |m| = i} c(m) \left[ \frac{(1 - \alpha)(n_k) \prod_{j=1}^{k-1} (1 - \alpha + m_j)(n_j)}{(k + i - \alpha k)(n)} \right].
\]
Now, for every \(m \in \mathbb{N}^{k-1}\) such that \(|m| = i\), define \(m'_j = m_j + 1\); then we can rewrite
\[
(k + i - 1)! c(m) V_{k+i, k} = \frac{V_{k+i, k}}{V_{k+i-1, k-1}} \left( \frac{(k + i - 1)!}{\prod_{j=1}^{k-1} m'_j!(k + i - 1 - \sum_{i=1}^{j-1} m'_j)} \right) V_{k+i-1, k-1} \prod_{j=1}^{k-1} (1 - \alpha)(m'_j-1)
\]
\[
= \frac{V_{k+i, k}}{V_{k+i-1, k-1}} \mu_{\alpha, \nu}(m').
\]
Thus the right-hand side of (83) becomes

\[
\sum_{i=0}^{\infty} \frac{V_{k+i,k}}{V_{k+i-1,k-1}} \sum_{m' \in \mathbb{N}^{k-1}:|m'|=k+i-1} \mu_{\alpha,V}(m') \left[ \frac{(1-\alpha)(n_k) \prod_{j=1}^{k-1}(m_j'-\alpha)}{(k+i-\alpha)(n)} \mathbb{E}(W_e^{n} | i_k = k + i) \right] \tag{84}
\]

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

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

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

\[
\frac{Y_1(m')}{\sum_{i=1}^{k} Y_i(m')} \ldots \frac{Y_k(m')}{\sum_{i=1}^{k} Y_i(m')}
\]

is a Dirichlet distribution with parameters \((m_1'-\alpha, \ldots, m_{k-1}'-\alpha, 1-\alpha)\), therefore (84) completes the proof of part (i).

To prove part (ii), we only have to notice that, from (84) it must follow that

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

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

\[
\mathbb{P}(i_k = k + i) = \frac{V_{k+i,k}}{V_{k+i-1,k-1}} \sum_{m' \in \mathbb{N}^{k-1}:|m'|=k+i-1} \mu_{\alpha,V}(m'). \tag{85}
\]

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

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

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 (85). If we reconsider the term \( c(m) \) as in (81) (for \( m \in \mathbb{N}^{k-1}_0:|m| = i \)), we see that

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

is the coefficient of \( \zeta^i \) in

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

Thus

\[
\sum_{m \in \mathbb{N}^{k-1}_0:|m| = i} c(m) = \frac{\alpha^{-(k-1)}}{(k+i-1)!} \sum_{j=0}^{k-1} \frac{(-1)^{k+i+j-1}}{j!(k-1-j)!} (j\alpha)_{[k+i-1]}.
\] \tag{86}

Since

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

then part (ii) is proved by comparison of (80) with (85).
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.

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

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

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.

5. K. Bubeck and J. Wesołowski. The Dirichlet distribution and process through neutralities. *J. Theor. Probab.*, 20: 295–308, 2007.

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

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/0507501](http://arxiv.org/abs/math.PR/0507501)

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

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.

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

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

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

13. 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.

14. 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.

15. F. M. Hoppe. Pólya-like urns and the Ewens’ sampling formula. *J. Math. Biol.*, 20(1):91–94, 1984.

16. 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.

17. 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.

18. 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. Sistem. Kombin. i Algoritm. Metody. 1):181–218, 340, 1995.

19. 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. Sistem. Kombin. i Algoritm. Metody. 1):162–180, 339–340, 1995.

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

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

22. S. Nacu. Increments of random partitions, 2004. [http://arXiv.org/abs/math/0310091v2 [math.PR]](http://arXiv.org/abs/math/0310091v2 [math.PR])

23. 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.

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

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

26. 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.

27. 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.

28. 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.
29. S. Walker and P. Muliere. Beta-Stacy processes and a generalization of the Pólya-urn scheme. *Ann. Statist.*, 25(4):1762–1780, 1997.
30. 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.
31. G. Watterson. Lines of descent and the coalescent. *Theor. Popul. Biol.*, 26(1):77–92, 1984.
32. S. Zabell Predicting the unpredictable, *Synthese*, 90(2):205–232, 1992.