A SIZE-BIASED INTRODUCTION TO KINGMAN'S THEORY OF RANDOM PARTITIONS

NGOC M TRAN

Abstract. Size-biased permutation is motivated by applications in species sampling. In the 1960s, biologists in population genetics were interested in inferring the distribution of alleles in a population through sampling. Size-biased permutation models the outcome of successive sampling, where one samples without replacement from the population and records the abundance of newly discovered species in the order that they appear. To account for the occurrence of new types of alleles through mutation and migration, biologists considered random abundance sequences and did not assume an upper limit to the number of possible types. This leads to the study of size-biased permutation of an infinite, summable sequence of i.i.d. random variables, in other words, jumps of a subordinator. We will head towards major results of this theory, starting with the case of finitely many terms.

These are lecture notes for the Vietnam 2016 Spring School on Combinatorial Stochastic Processes. The lecture notes are largely based on Jim's textbook 'Combinatorial Stochastic Processes', Bertoin's textbook 'Random fragmentation and coagulation processes', and on my paper with Jim. This is part of an ongoing effort to update Jim's textbook and the status of the open problems in that text.

1. Introduction

1.1. **Different ways to think about partitions.** In general, for a sequence $x = (x(1), x(2), \ldots)$, write $x^{\downarrow}$ to denote the same sequence presented in decreasing order. Let $\Delta = \{x = (x(1), x(2), \ldots) : x(i) \geq 0, \sum_i x(i) \leq 1\}$ and $\Delta^{\downarrow} = \{x^{\downarrow} : x \in \Delta\}$ be closed infinite simplices, the later contains sequences with non-increasing terms. Denote their boundaries by $\Delta_1 = \{x \in \Delta : \sum_i x(i) = 1\}$ and $\Delta_1^{\downarrow} = \{x \in \Delta^{\downarrow}, \sum_i x(i) = 1\}$ respectively. Any finite sequence can be associated with an element of Δ_1 after being normalized by its sum and extended with zeros.

Example 1.1 (Mass partitions)**.** Consider a mass T split into countably many smaller masses $s(1), s(2), \ldots \geq 0$. If $\sum_i s(i) = T$, we say that the partition is conservative (ie: total mass is conserved). If $\sum_i s(i) < T$, we say that the partition is dissipative. One can imagine that when the mass splits, a sizeable chunk of it becomes dust (isolated infinitesimal particles) and dissipate into the air. The normalized sequence $(s(1)/T, \ldots)$ is in Δ. Thus, Δ is also called the space of mass partitions.

Example 1.2 (Interval partitions)**.** Represent a mass T as the interval $[0, T]$, each point on the interval is an infinitesimal particle making up the mass. Given $x \in \Delta$, represent $T \cdot x(1), T \cdot x(2), \ldots$ as lengths of disjoint open intervals on $[0, T]$. This is

Date: March 1, 2016.

called an interval partition representation of $T \cdot x$. A point $u \in [0, T]$ that does not belong to the union of such intervals represents a dust.

Example 1.3 (Partition of n)**.** For an integer n, a partition of n is an unordered set $\{n_1, \ldots, n_k\}$ of integers summing up to n: $\sum_i n_i = n$. The normalized sequence $(n_1/n, \ldots, n_k/n)$ is a mass partition. In this case, one can think of the mass as making up of n indivisible units, each of mass 1.

Example 1.4 (Random mass partition)**.** Here is a simple way to get a random mass partition. Fix an integer n. Let F be a distribution on $(0, \infty)$, mean $\mu < \infty$. Let $X(1), \ldots, X(n)$ be independent and identically distributed (i.i.d) random variables with distribution F. Define $T_n = \sum_{i=1}^n X(i)$. Then $(X(1)/T, \ldots, X(n)/T)$ is a random mass partition with n parts.

Example 1.5 (Stick-breaking)**.** Here is a simple way to get a random interval partition. Start with a stick of length 1. Choose a point on the stick according to some distribution F_1 supported on $[0, 1]$, 'break' the stick into two pieces, discard the left-hand piece, and rescale the remaining half to have length 1. Repeating this procedure with distribution F_2, and so on. This gives a random mass partition $X = (X(1), X(2), \ldots)$, where

$$X(1) = W_k \prod_{i=1}^{k-1} \overline{W}_i,$$

where W_i's are independent, and W_i distributed as F_i.

1.2. **Orderings of a partition.** Let $x = (x(1), \ldots, x(n)) \in \Delta$ be a mass partition with n parts, $t = \sum_i x(i)$. There are three natural ways to order the elements of x.

Decreasing order: $x^{\downarrow}$.

Exchangeable random order: let σ be a uniformly distributed random permutation of $[n]$. The mass partition x presented in exchangeable random order is the random mass partition $X = (x(\sigma_1), \ldots, x(\sigma_n))$.

Size-biased order: here $X = (x(\sigma_1), \ldots, x(\sigma_n))$, where σ is the random permutation with $\mathbb{P}(\sigma_1 = i) = \dfrac{x(i)}{t}$, and for k distinct indices $i_1, \ldots, i_k$,

$$\mathbb{P}(\sigma_k = i_k | \sigma_1 = i_1, \ldots, \sigma_{k-1} = i_{k-1}) = \frac{x(i_k)}{t - (x(i_1) + \ldots + x(i_{k-1}))}. \tag{1}$$

An index i with bigger 'size' $x(i)$ tends to appear earlier in the permutation, hence the name size-biased. Call σ the size-biased order, and call the random sequence X the size-biased permutation of x.

The size-biased order is important because it is the order in which new elements appear in a sampling without replacement scheme. For this reason, it is sometimes called the *order of appearance*.

Example 1.6 (Size-biased order and sampling without replacement)**.** Let $(n_1, \ldots, n_k)$ be a vector of integers, n_i interpreted as the number of balls of color i (or the number of animals of species i). Let $n = \sum_{i=1}^k n_k$ be the total number of balls. Sample without replacement from the set of n balls. Let σ denote the order of colors that appear. This is a size-biased order.

Example 1.7 (Kingman's paintbox and size-biased permutation)**.** *Kingman's paintbox* [5] is a useful way to describe and extend size-biased permutations. For $x \in \Delta$, let s_k be the sum of the first k terms. Note that x defines a partition $\varphi(x)$ of the unit interval $[0,1]$, consisting of *components* which are intervals of the form $[s_k, s_{k+1})$ for $k = 1, 2, \ldots$, and the interval $[s_\infty, 1]$, which we call the *zero component*. Sample points $\xi_1, \xi_2, \ldots$ one by one from the uniform distribution on $[0,1]$. Each time a sample point discovers a new component that is not in $[s_\infty, 1]$, write down its size. If the sample point discovers a new point of $[s_\infty, 1]$, write 0. Let $X^* = (X^*(1), X^*(2), \ldots)$ be the random sequence of sizes. Since the probability of discovery of a particular (non-zero) component is proportional to its length, the non-zero terms in X^* form the size-biased permutation of the non-zero terms in x as defined by (1). In the paintbox terminology, the components correspond to different colors used to paint the balls with labels $1, 2, \ldots$. Two balls i, j have the same paint color if and only if ξ_i and ξ_j fall in the same component. The size-biased permutation X^* records the size of the newly discovered components, or paint colors. The zero component represents a continuum of distinct paint colors, each of which can be represented at most once.

Example 1.8 (Kingman's paintbox and random partition of n)**.** Kingman's paintbox also gives a random partition of n (or more precisely, a random partition of $[n]$). Consider the previous setup. Say that $i \sim j$ iff ξ_i and ξ_j fall on the same interval (ie: if the two balls have the same color). Then for each $n \in \mathbb{N}$, we get a random partition of the n balls by colors.

We end this section with some questions on the objects we have introduced so far. We will answer these in the next couple of lectures.

(1) Take the size-biased permutation X^* of a random partition X (eg: from i.i.d). What is the distribution of X^*?
(2) When does a random mass partition has the stick-breaking form for independent stick lengths? When does a size-biased permutation has the stick-breaking form for independent stick lengths?

2. Size-biased permutation of a finite i.i.d sequence

The size-biased permutation of a random sequence X is defined conditioned on the sequence's values. We now focus on the size-biased permutation of an i.i.d sequence $X_n = (X_n(1), \ldots, X_n(n))$ with finite length n. We will use square brackets $(X_n[1], \ldots, X_n[n])$ to denote the size-biased permutation, or X_n^*, to avoid having to list out the terms.

Assume that F has density ν_1. Let $T_{n-k} = X_n[k+1] + \ldots + X_n[n]$ denote the sum of the last $n-k$ terms in an i.i.d. size-biased permutation of length n. For $1 \le k \le n$, let ν_k be the density of S_k, the sum of k i.i.d. random variables with distribution F.

We shall write $gamma(a, \lambda)$ for a Gamma distribution whose density at x is $\lambda^a x^{a-1} e^{-\lambda x}/\Gamma(a)$ for $x > 0$, and $beta(a, b)$ for the Beta distribution whose density at x is $\dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$ for $x \in (0,1)$.

2.1. Joint distribution.

We first derive joint distribution of the first k terms $X_n[1], \ldots, X_n[k]$.

Proposition 2.1 (Barouch-Kaufman [1])**.** *We have*

$$\begin{aligned}
&\mathbb{P}(X_n[1] \in dx_1, \ldots, X_n[k] \in dx_k) \\
&= \frac{n!}{(n-k)!} \left(\prod_{j=1}^{k} x_j \, \nu_1(x_j) \, dx_j \right) \int_0^\infty \nu_{n-k}(s) \prod_{j=1}^{k} (x_j + \ldots + x_k + s)^{-1} \, ds \qquad (2) \\
&= \frac{n!}{(n-k)!} \left(\prod_{j=1}^{k} x_j \, \nu_1(x_j) \, dx_j \right) \mathbb{E} \left(\prod_{j=1}^{k} \frac{1}{x_j + \ldots + x_k + S_{n-k}} \right). \qquad (3)
\end{aligned}$$

Proof. Let σ denote the random permutation on n letters defined by size-biased permutation as in (1). Then there are $\frac{n!}{(n-k)!}$ distinct possible values for $(\sigma_1, \ldots, \sigma_k)$. By exchangeability of the underlying i.i.d. random variables $X_n(1), \ldots, X_n(n)$, it is sufficient to consider $\sigma_1 = 1, \ldots, \sigma_k = k$. Note that

$$\mathbb{P}\left((X_n(1), \ldots, X_n(k)) \in dx_1 \ldots dx_k, \sum_{j=k+1}^{n} X_n(j) \in ds \right) = \nu_{n-k}(s) \, ds \prod_{j=1}^{k} \nu_1(x_j) \, dx_j.$$

Thus, restricted to $\sigma_1 = 1, \ldots, \sigma_k = k$, the probability of observing $(X_n[1], \ldots, X_n[k]) \in dx_1 \ldots dx_k$ and $T_{n-k} \in ds$ is precisely

$$\frac{x_1}{x_1 + \ldots + x_k + s} \frac{x_2}{x_2 + \ldots + x_k + s} \cdots \frac{x_k}{x_k + s} \, \nu_{n-k}(s) \left(\prod_{j=1}^{k} \nu_1(x_j) \, dx_j \right) ds.$$

By summing over $\frac{n!}{(n-k)!}$ possible values for $(\sigma_1, \ldots, \sigma_k)$, and integrating out the sum T_{n-k}, we arrive at (2). Equation (3) follows by rewriting. □

Note that $X_n[k] = T_{n-k+1} - T_{n-k}$ for $k = 1, \ldots, n-1$. Thus we can rewrite (2) in terms of the joint law of $(T_n, T_{n-1}, \ldots, T_{n-k})$:

$$\mathbb{P}(T_n \in dt_0, \ldots, T_{n-k} \in dt_k) = \frac{n!}{(n-k)!} \left(\prod_{i=0}^{k-1} \frac{t_i - t_{i+1}}{t_i} \nu_1(t_i - t_{i+1}) \right) \nu_{n-k}(t_k) \, dt_0 \ldots dt_k. \qquad (4)$$

Rearranging (4) yields the following result, which appeared as an exercise in [2, §2.3].

Corollary 2.2 (Chaumont-Yor [2])**.** *The sequence $(T_n, T_{n-1}, \ldots, T_1)$ is an inhomogeneous Markov chain with transition probability*

$$\mathbb{P}(T_{n-k} \in ds | T_{n-k+1} = t) = (n-k+1) \frac{t-s}{t} \nu_1(t-s) \frac{\nu_{n-k}(s)}{\nu_{n-k+1}(t)} \, ds, \qquad (5)$$

for $k = 1, \ldots, n-1$. Together with $T_n \stackrel{d}{=} S_n$, equation (5) specifies the joint law of $(X_n[1], \ldots, X_n[n])$, and vice versa.

2.2. **Stick-breaking representation.** For $k \geq 1$, conditioned on $T_{n-k+1} = t$, $X_n[k]$ is distributed as the first size-biased pick out of $n-k+1$ i.i.d. random variables conditioned to have sum $S_{n-k+1} = t$. This provides a recursive way to generate a finite i.i.d. size-biased permutation: first generate T_n (which is distributed as S_n). Conditioned on the value of T_n, generate T_{n-1}, let $X_n[1]$ be the difference. Now conditioned on the value of T_{n-1}, generate T_{n-2} via (5), let $X_n[2]$ be the difference, and so on.

Let us explore this recursion from a different angle by considering the ratio $W_{n,k} := \dfrac{X_n[k]}{T_{n-k+1}}$ and its complement, $\overline{W}_{n,k} = 1 - W_{n,k} = \dfrac{T_{n-k}}{T_{n-k+1}}$. For $k \geq 2$, note that

$$\frac{X_n[k]}{T_n} = \frac{X_n[k]}{T_{n-k+1}} \frac{T_{n-k+1}}{T_{n-k+2}} \cdots \frac{T_{n-1}}{T_n} = W_{n,k} \prod_{i=1}^{k-1} \overline{W}_{n,i}. \tag{6}$$

The variables $\overline{W}_{n,i}$ can be interpreted as residual fractions in a *stick-breaking* scheme: start with a stick of length 1. Choose a point on the stick according to distribution $W_{n,1}$, 'break' the stick into two pieces, discard the piece of length $W_{n,1}$ and rescale the remaining half to have length 1. Repeating this procedure k times, and (6) is the fraction broken off at step k relative to the original stick length.

Together with $T_n \stackrel{d}{=} S_n$, one could use (6) to compute the marginal distribution for $X_n[k]$ in terms of the ratios $\overline{W}_{n,i}$. In general the $W_{n,i}$ are not necessarily independent, and their joint distributions need to be worked out from (5).

Lukacs [6] proved that if X, Y are non-degenerate, positive independent random variables, then $X + Y$ is independent of $\frac{X}{X+Y}$ if and only if $X \sim gamma(a, \lambda)$, $Y \sim gamma(b, \lambda)$ for some parameters a, b, λ. In this case, $\frac{X}{X+Y} \sim beta(a, b)$, and $X + Y \sim gamma(a + b, \lambda)$. This leads to the following.

Proposition 2.3 (Patil-Taillie [7])**.** *Consider the stick-breaking representation in (6) of the size-biased permutation of an i.i.d sequence with distribution F. The random variables T_n and the $W_{n,1}, \ldots, W_{n,n-1}$ in (6) are mutually independent if and only if F is $gamma(a, \lambda)$ for some $a, \lambda > 0$. In this case,*

$$\begin{aligned} X_n[1] &= \gamma_0 \beta_1 \\ X_n[2] &= \gamma_0 \bar{\beta}_1 \beta_2 \\ &\cdots \\ X_n[n-1] &= \gamma_0 \bar{\beta}_1 \bar{\beta}_2 \ldots \bar{\beta}_{n-2} \beta_{n-1} \\ X_n[n] &= \gamma_0 \bar{\beta}_1 \bar{\beta}_2 \ldots \bar{\beta}_{n-1} \end{aligned}$$

where γ_0 has distribution $gamma(an, \lambda)$, β_k has distribution $beta(a + 1, (n - k)a)$, $\bar{\beta}_k = 1 - \beta_k$ for $1 \leq k \leq n - 1$, and the random variables $\gamma_0, \beta_1, \ldots, \beta_{n-1}$ are independent.

Proof. It is sufficient to consider the first stick-break. Note that

$$\mathbb{P}(X_n[1]/T_n \in du, T_n \in dt) = n\, u\, \mathbb{P}\left(\frac{X_n(1)}{X_n(1) + (X_n(2) + \ldots + X_n(n))} \in du, T_n \in dt \right). \tag{7}$$

Suppose $F = gamma(a, \lambda)$. Since $X_n(1) \stackrel{d}{=} gamma(a, \lambda)$, $S_{n-1} = X_n(2) + \ldots + X_n(n) \stackrel{d}{=} gamma(a(n-1), \lambda)$, independent of $X_n(1)$, the ratio $\frac{X_n(1)}{X_n(1)+S_{n-1}}$ has distribution $beta(a, a(n-1))$ and is independent of T_n. Thus

$$\begin{aligned} \mathbb{P}(X_n[1]/T_n \in du) &= n\, u\, \frac{\Gamma(a + a(n-1))}{\Gamma(a)\Gamma(a(n-1))} u^{a-1} (1-u)^{a(n-1)-1} \\ &= \frac{\Gamma(a + 1 + a(n-1))}{\Gamma(a+1)\Gamma(a(n-1))} u^{a} (1-u)^{a(n-1)-1}. \end{aligned}$$

In other words, $X_n[1]/T_n \stackrel{d}{=} beta(a, a(n-1))$. This proves the if direction. Now suppose $W_{n,1} = X_n[1]/T_n$ is independent of T_n. Reverse the argument, this implies that $X_n(1)$ is independent of $X_n(1)/T_n$. Apply Lukacs' theorem for $X = X_n(1)$, $Y = X_n(2) + \ldots + X_n(n)$, we see that F must be the gamma distribution. □

3. Size-biased permutation of an infinite i.i.d sequence

3.1. **Subordinator as limit of i.i.d.** We now want to send $n \to \infty$, and derive the analogue of the above results on the infinite simplex Δ. First, we need an infinite sequence of i.i.d random variables $X = (X(1), X(2), \ldots)$ that is a.s. summable $T := \sum_i X(i) < \infty$. This condition is necessary since we will divide by T to obtain a distribution on Δ.

Let $((X_n), n \geq 1)$ be an i.i.d. positive triangular array, that is, $X_n = (X_n(1), \ldots, X_n(n))$, where $X_n(i), i = 1, \ldots, n$ are i.i.d. and a.s. positive. Write T_n for $\sum_{i=1}^n X_n(i)$. A classical result in probability states that $T_n \stackrel{d}{\to} T$ if and only if $T = \tilde{T}(1)$ for some *Lévy process* $\tilde{T}$, which in this case is a *subordinator*.

Definition 3.1. A *Lévy process* $\tilde{T}$ in $\mathbb{R}$ is a stochastic process with right-continuous left-limits paths, stationary independent increments, and $\tilde{T}(0) = 0$. A *subordinator* $\tilde{T}$ is a Lévy process, with real, finite, non-negative increments.

Positive increments of an subordinator are called its jumps. Here is one way to cook up a subordinator with infinitely many jumps but the sum of all jumps is finite. Consider a measure Λ on $(0, \infty)$ such that

$$\int_0^\infty (1 \wedge x)\Lambda(dx) < \infty, \tag{8}$$

and

$$\Lambda((0, \infty)) = \infty. \tag{9}$$

Let $\mathcal{X}$ be a Poisson point process on $(0, \infty)$ with intensity measure Λ. Equation (8) tells us that this point process has a.s. finitely many points above 1, so it is possible to list them in decreasing order. Let $X^\downarrow = (X^\downarrow(1), X^\downarrow(2), \ldots)$ denote the sequence of points of $\mathcal{P}$ in decreasing order. Equation (9) tells us that $\mathcal{P}$ has a.s. infinitely many points, so $X^\downarrow$ is an infinite sequence.

Now let us list $X^\downarrow$ in exchangeable random order. Introduce an independent sequence $U(1), U(2), \ldots$ of i.i.d uniform random variables on $[0, 1]$. 'Order' the pairs $(X^\downarrow(i), U(i))$ in increasing U-coordinate, and let this be the list of (jump-size, jump-time) description of our subordinator.[1] That is, define the process $\tilde{T} : [0, 1] \to \mathbb{R}_{\geq 0}$ by

$$(\tilde{T})(t) = \mathsf{d}\, t + \sum_i X(i)\mathbf{1}_{\{U(i) \leq t\}}$$

for $0 \leq t \leq 1$, and $\mathsf{d} \geq 0$ a fixed number called the drift coefficient. One can check that $\tilde{T}$ is indeed a subordinator restricted to $[0, 1]$. Furthermore, all subordinators satisfying $\tilde{T}(1) < \infty$ a.s. is of this form. The measure Λ is called the Lévy measure of $\tilde{T}$.

Finally, we need a classical result on convergence of i.i.d. positive triangular arrays to subordinators (see [4, §15]).

[1] We can't really speak of the smallest $U(i)$ since there are infinitely many of them. However, the expression for $\tilde{T}$ below still makes sense.

Theorem 3.1. *Let $(X(n), n \geq 1)$ be an i.i.d. positive triangular array, $T_n = \sum_{i=1}^{n} X_n(i)$. Then $T_n \xrightarrow{d} T$ for some random variable T, $T < \infty$ a.s. if and only if $T = \tilde{T}(1)$ for some subordinator $\tilde{T}$ whose Lévy measure Λ satisfies (8).*

We would also want the following result on convergence of densities.

Theorem 3.2. *Consider the setup of Theorem 3.1. Assume $T_n \xrightarrow{d} T$, $T < \infty$ a.s. Let μ_n be the measure of $X_n(i)$. If μ_n, Λ have densities ρ_n, ρ, respectively, then we have pointwise convergence for all $x > 0$*

$$n\rho_n(x) \to \rho(x).$$

Consider a subordinator with Lévy measure Λ, drift $\mathtt{d} = 0$. Let $\tilde{T}_0$ be the subordinator at time 1. Assume $\Lambda(1, \infty) < \infty$, $\Lambda(0, \infty) = \infty$, $\int_0^1 x\Lambda(dx) < \infty$, and $\Lambda(dx) = \rho(x)\, dx$ for some density ρ. Note that $\tilde{T}_0 < \infty$ a.s, and it has a density determined by ρ via its Laplace transform, which we denote ν. Let $\tilde{T}_k$ denote the remaining sum after removing the first k terms of the size-biased permutation of the sequence $X^{\downarrow}$ of ranked jumps.

Proposition 3.3 ([8])**.** *The sequence $(\tilde{T}_0, \tilde{T}_1, \ldots)$ is a Markov chain with stationary transition probabilities*

$$\mathbb{P}(\tilde{T}_1 \in dt_1 | \tilde{T}_0 = t) = \frac{t - t_1}{t} \cdot \rho(t - t_1)\, \frac{\nu(t_1)}{\nu(t)}\, dt_1.$$

Proof. Note the similarity to (5). Starting with (4) and send $n \to \infty$, for any finite k, we have $\nu_{n-k} \to \nu$ pointwise, and by Theorem 3.1, $(n-k)\nu_1 \to \rho$ pointwise over $\mathbb{R}$, since there is no drift term. Thus the analogue of (4) in the limit is

$$\mathbb{P}(\tilde{T}_0 \in dt_0, \ldots, \tilde{T}_k \in dt_k) = \left(\prod_{i=0}^{k-1} \frac{t_i - t_{i+1}}{t_i} \rho(t_i - t_{i+1}) \right) \nu(t_k)\, dt_0 \ldots dt_k. \tag{10}$$

Rearranging gives the transition probability in Proposition 3.3. □

4. Random mass partition with independent stick-breaks. Stable subordinators and Poisson-Dirichlet distributions.

The stick-breaking representation in (6) in the limit as $n \to \infty$ takes the form

$$\frac{X[k]}{\tilde{T}_0} = W_k \prod_{i=1}^{k-1} \overline{W}_i, \tag{11}$$

where $X[k]$ is the kth size-biased pick from the jumps of the subordinator $\tilde{T}$, and $W_i = \frac{X[i]}{\tilde{T}_{i-1}}$, $\overline{W}_i = 1 - W_i = \frac{\tilde{T}_i}{\tilde{T}_{i-1}}$.

Let us for a moment forget about the subordinator and consider (11). As long as the W_i's are random, we have a random mass partition $(P_1, P_2, \ldots)$, with

$$P_i = W_1 \cdots W_{i-1} \overline{W}_i \tag{12}$$

for $i = 1, 2, \ldots$. Let us look for 'nice' random partitions. Specifically, we want the W_i's are independent, and P given by (12) is the size-biased permutation of some mass partition Q. But if $P \stackrel{d}{=} Q^*$, then $P^* \stackrel{d}{=} (Q^*)^* \stackrel{d}{=} Q^* \stackrel{d}{=} P$. So this is equivalent to saying that the distribution of P is invariant under size-biased permutation. Pitman [9] proved a complete characterization.

Theorem 4.1 ([9]). *Let $P \in \Delta_1$, $P_1 < 1$, and $P_n = W_1 \cdots W_{n-1}\overline{W}_n$ for independent W_i. Then $P = P^*$ if and only if one of the four following conditions holds.*

(1) *$P_n \geq 0$ a.s. for all n, in which case the distribution of W_n is*
$$beta(1-\alpha, \theta + n\alpha)$$
for every $n = 1, 2, \ldots$, for some $0 \leq \alpha < 1$, $\theta > -\alpha$.

(2) *For some integer constant m, $P_n \geq 0$ a.s. for all $1 \leq n \leq m$, and $P_n = 0$ a.s. otherwise. Then either*
   (a) *For some $\alpha > 0$, W_n has distribution $beta(1+\alpha, m\alpha - n\alpha)$ for $n = 1, \ldots, m$;*
   or
   (b) *$W_n = 1/(m-n+1)$ a.s., that is, $P_n = 1/m$ a.s. for $n = 1, \ldots, m$;*
   or
   (c) *$m = 2$, and the distribution F on $(0,1)$ defined by $F(dw) = \bar{w}\mathbb{P}(W_1 \in dw)/\mathbb{E}(\bar{W}_1)$ is symmetric about $1/2$.*

The Patil-Taillie case of Proposition 2.3 is case 2(a). Case 2(b) is the limit of 2(a) as $\alpha \to \infty$. Case 2(c) is the special situation where the random mass partition only has two parts.

In case (1), such a distribution P is known as the $GEM(\alpha, \theta)$ distribution. The abbreviation GEM was introduced by Ewens, which stands for Griffiths-Engen-McCloskey. If P is $GEM(\alpha, \theta)$, then $P^{\downarrow}$ is called a Poisson-Dirichlet distribution with parameters (α, θ), denoted $PD(\alpha, \theta)$ [8]. This is an important family of ranked random mass partitions, with applications in fragmentation and coalescence, Bayesian statistics, and machine learning. See [10] and references therein.

We now prove one direction of this theorem by deriving $PD(\alpha, \theta)$ as ranked jumps of a subordinator.

Definition 4.1. The subordinator $\tilde{T}$ with Lévy measure
$$\Lambda(dx) = \theta x^{-1} e^{-cx},\, dx$$
for $x > 0$ is called a gamma subordinator with parameter (θ, c).

The parameter $c > 0$ is called the scaling parameter. Indeed, if $\tilde{T}$ a gamma subordinator with parameter $(\theta, 1)$, then $c \cdot \tilde{T}$ is a gamma subordinator with parameter (θ, c). Thus, the parameter c plays no role in the random mass partition defined by a gamma subordinator.

Proposition 4.2 ($PD(0, \theta)$). *Let $\tilde{T}$ be a gamma(θ, c) subordinator. For $X^{\downarrow}$ its sequence of ranked jumps, the random mass partition $X^{\downarrow}/\tilde{T}(1)$ has distribution $PD(0, \theta)$.*

Definition 4.2. A stable process is a real-valued Lévy process $(Y(s), s \geq 0)$ with initial value $Y(0) = 0$ that has the self-similar property
$$Y(s) \stackrel{d}{=} s^{1/\alpha} Y(1)$$
for all $s \geq 0$. The parameter α is the exponent of the process.

Stable processes are important in probability. A stable subordinator is a subordinator that is also a stable process.

Lemma 4.3. *Consider a subordinator $\tilde{T}$ with Lévy measure Λ with density*

$$\rho(x) = cx^{-(1+\alpha)}, \tag{13}$$

for some constant $c > 0$, $x > 0$. Then $\tilde{T}$ satisfies (8) if and only if $\alpha \in (0, 1)$.

Proof. Plug (13) in to (8) and integrate. □

Proposition 4.4 ($PD(\alpha, 0)$)**.** *For $\alpha \in (0, 1)$, let $\tilde{T}$ be a stable(α) subordinator with density (13). For $X^{\downarrow}$ its sequence of ranked jumps, the random mass partition $X^{\downarrow}/\tilde{T}(1)$ has distribution $PD(\alpha, 0)$.*

To obtain $PD(\alpha, \theta)$, we will do a change of measure from $PD(\alpha, 0)$. First we need a technical lemma.

Lemma 4.5. *For $\alpha \in (0, 1)$, let $\tilde{T}$ be a stable(α) subordinator with density (13). Then $\tilde{T}(1) > 0$ a.s. and $\mathbb{E}(\tilde{T}(1)^{-\theta}) < \infty$ for all $\theta > -\alpha$.*

Proof. For $\theta > 0$, start with the identity

$$x^{-\theta} = \frac{1}{\Gamma(\theta)} \int_0^\infty e^{-xt} t^{\theta-1}\, dt.$$

Apply Fubini-Tonelli's theorem for nonnegative functions

$$\mathbb{E}(\tilde{T}(1)^{-\theta}) = \frac{1}{\Gamma(\theta)} \int_0^\infty \mathbb{E} e^{-\tilde{T}(1)t} t^{\theta-1}\, dt$$

Since $\tilde{T}$ is α-stable,

$$\mathbb{E} e^{-\tilde{T}(1)t} = e^{-ct^\alpha}.$$

(One can also use the Lévy-Khitchine formula for subordinators). Plug in and simplify, we get

$$\mathbb{E}(\tilde{T}(1)^{-\theta}) = \frac{\Gamma(\theta/\alpha)}{\alpha c^{\theta/\alpha} \Gamma(\theta)}.$$

For $\theta < 0$, express the Γ function as an integral, and use analytic continuity to extend the above computation. Note that $\Gamma(\theta/\alpha) < \infty$ for $\theta > -\alpha$. □

Proposition 4.6 ($PD(\alpha, \theta)$)**.** *For $\alpha \in (0, 1)$, let $\tilde{T}$ be a stable(α) subordinator with density (13). Let $X^{\downarrow}$ be its sequence of ranked jumps. Let P_α denote the distribution of $X^{\downarrow}$. For $\theta > -\alpha$, define the probability measure $P_{\alpha,\theta}$ to be absolutely continuous with respect to P_α and has density*

$$P_{\alpha,\theta} = \frac{\tilde{T}(1)^{-\theta}}{\mathbb{E}\tilde{T}(1)^{-\theta}} P_\alpha.$$

Then under $P_{\alpha,\theta}$, the random mass partition $X^{\downarrow}/\tilde{T}(1)$ has distribution $PD(\alpha, \theta)$.

Proof of Propositions 4.2, 4.4 and 4.6. From (10), for general ρ, the joint density of $\tilde{T}_1$ and W_1 is

$$f_{W_1,\tilde{T}_1}(w_1, t_1) = f_{\tilde{T}_0,\tilde{T}_1}(t_1 \bar{w_1}^{-1}, t_1) = w_1 \rho(\frac{t_1 w_1}{\bar{w_1}}) t_1 \bar{w_1}^{-1} \nu(t_1),$$

where $\bar{w_1} = 1 - w_1$.

For Propositions 4.2 and 4.4, plug in the corresponding formulas for ρ and simplify. For instance, consider the α-stable subordinator of Proposition 4.4. Plug in (13) for ρ gives

$$f_{W_1,\tilde{T}_1}(w_1, t_1) = c w_1^{-\alpha} \bar{w_1}^{\alpha-1} t_1^{-\alpha} \nu(t_1). \tag{14}$$

Since c is a constant independent of w_1 and t_1, we conclude that W_1 and T_1 are independent. In particular, W_1 has $beta(1-\alpha,\alpha)$ distribution. Repeatedly apply (10) as above show that the W_i's are independent and have the desired distributions.

For Proposition 4.6, note that $\tilde{T}(1) = \tilde{T}_0 = t_1 \bar{w_1}^{-1}$. So the density $f_{W_1,\tilde{T}_1}(w_1,t_1)$ under $\mathbb{P}_{\alpha,\theta}$ is

$$C f_{W_1,\tilde{T}_1}(w_1,t_1)(\frac{t_1}{\bar{w_1}})^{-\theta}$$

for some constant C depending on α and θ, and $f_{W_1,\tilde{T}_1}(w_1,t_1)$ is given in (14). Simplify to get

$$C c w_1^{-\alpha} \bar{w_1}^{\alpha+\theta-1} t_1^{-(\alpha+\theta)} \nu(t_1).$$

So under $\mathbb{P}_{\alpha,\theta}$, W_1 and $\tilde{T}_1$ are independent, W_1 has $beta(1-\alpha,\alpha+\theta)$ distribution, and $\tilde{T}_1$ is distributed like $\tilde{T}_0$ under $\mathbb{P}_{\alpha,\theta+\alpha}$. Recurse this computation gives the desired result for the distribution of the W_i's.

We have shown that the size-biased permutation of the jumps of the subordinators defined in Propositions 4.2, 4.4 and 4.6 follow the $GEM(0,\theta)$, $GEM(\alpha,0)$ and $GEM(\alpha,\theta)$ distributions, respectively. Thus, the ranked jumps follow the Poisson-Dirichlet distributions with corresponding parameters. □

5. Exchangeable random partitions. Kingman's correspondence.

5.1. **Partition of $[n]$.** Let $n \in \mathbb{N} \cup \{\infty\}$. Define $[n] = \{1,2,\ldots,n\}$, with the convention that $[\infty] = \mathbb{N}$. For a set A, let $|A|$ denote its cardinality (number of elements).

Definition 5.1. A partition π_n of $[n]$ into k blocks is an unordered collection of non-empty disjoint sets $\{A_1,\ldots,A_k\}$, whose union is $[n]$. The set of unordered block sizes $\{|A_1|,\ldots,|A_k|\}$ is a partition of n, which we denote $|\pi_n|$.

The canonical way to order a partition of $[n]$ is by an increasing order on the least element of each block. This is called order by appearance.

A partition of $[n]$ induces a partition of $[m]$ for all $m \le n$ by restricting to the first m elements.

Definition 5.2. A sequence $(\pi_{[n]}, n \ge 1)$ of partitions of $[n]$ is called consistent or compatible if the restriction of $\pi_{[n]}$ to $[m]$ equals $\pi_{[m]}$ for all $m \le n$.

Lemma 5.1. *A sequence $(\pi_{[n]}, n \ge 1)$ of partitions of $[n]$ is consistent if and only if $\pi_{[n]}$ equals the restriction of π to $[n]$ for some partition π of $\mathbb{N}$.*

Proof. Suppose π is a partition of $\mathbb{N}$. Then clearly its restrictions to $[n]$ form a consistent sequence. Conversely, suppose we have a consistent sequence. Label the blocks by order of appearance. Let $\pi_{[n]}(i)$ be the block of $\pi_{[n]}$ containing in i. For each fixed $i \in \mathbb{N}$, $(\pi_{[n]}(i), n \ge 1)$ is a non-decreasing sequence of sets. Define

$$\pi(i) = \bigcup_{n\in\mathbb{N}} \pi_{[n]}(i).$$

Then $(\pi(i), i \in \mathbb{N})$ is a partition of $\mathbb{N}$, call it π. Clearly π restricted to $[n]$ equals $\pi_{[n]}$. □

5.2. **Exchangeable partitions.** In many applications we want to forget about the labels of the individual elements of $[n]$, treating them as 'exchangeable' objects. Effectively we want to forget about the partition of $[n]$, and concentrate on partitions of n.

A permutation is a bijection $\sigma : [n] \to [n]$. For $n = \infty$, a permutation of $\mathbb{N}$ is a bijection $\sigma : \mathbb{N} \to \mathbb{N}$ that leaves all but finitely many elements of $\mathbb{N}$ fixed. The group of permutations of $[n]$ naturally acts on a partition of $[n]$ by permuting the labels of the elements. That is,

$$\sigma \cdot \{A_1, \ldots, A_k\} = \{\sigma \cdot A_1, \ldots, \sigma \cdot A_k\},$$

where if $A_1 = \{i, j, k, \ldots\}$, then

$$\sigma \cdot A_1 = \{\sigma(i), \sigma(j), \sigma(k), \ldots\}.$$

Example 5.1. Example: order by appearance, action of symmetric group. Exchangeable. Size-biased permutation, sampling without replacement and order by appearance.

Definition 5.3 (Exchangeable random partitions)**.** Let $n \in \mathbb{N} \cup \{\infty\}$. A random partition Π of $[n]$ is called exchangeable if for every permutation σ of $[n]$, $\sigma \cdot \Pi \stackrel{d}{=} \Pi$.

It follows from this definition that a random partition Π of $\mathbb{N}$ is exchangeable if and only if its restrictions to $[n]$ are exchangeable for all $n \in \mathbb{N}$.

Example 5.2 (Random partitions from Kingman's paintbox)**.** Fix $p = (p_1, p_2 \ldots) \in \Delta$. Represent it as an interval partition of $[0, 1]$, viewed as paint buckets. Use Kingman's paintbox to color balls numbered $i = 1, 2, \ldots$. Say that $i \sim j$ if i and j have the same color. This defines a random partition Π of $[\infty]$. Call this the paintbox based on p.

Lemma 5.2. *Fix $p = (p_1, p_2 \ldots) \in \Delta$. The paintbox based on p is an exchangeable random partition.*

Proof. Let $(U_i, i \geq 1)$ be the sequence of i.i.d uniform random variables used to construct the paint boxes. By definition, $\sigma \cdot \Pi$ has the same distribution as the paintbox constructed with the sequence $(U_{\sigma_i}, i \geq 1)$. But the U_i's are i.i.d, so $(U_{\sigma_i}, i \geq 1) \stackrel{d}{=} (U_i, i \geq 1)$. Thus, $\sigma \cdot \Pi \stackrel{d}{=} \Pi$. So Π is exchangeable. □

Example 5.3 (Finite paintbox)**.** For $k \in \mathbb{N}$, let $\pi = (n_1, \ldots, n_k)$ be a partition of n into k parts. Represent them as k disjoint intervals, each containg $n_1, \ldots, n_k$ integer points on $[1, n]$. Use a discrete version of Kingman's paintbox to paint n balls as follows: for each $i \in [n]$, pick an integer on $[1, n]$ uniformly at random out of the remaining integers, and color it according to the interval that it falls on. Then remove this integer from $[1, n]$. This gives a random partition Π_n of $[n]$. Call this the finite paintbox based on π.

Lemma 5.3. *Fix $\pi = (n_1, \ldots, n_k)$. The finite paintbox based on π is exchangeable.*

The following says that exchangeable finite partitions arise as mixture of finite paintboxes. It setups a bijection between exchangeable finite partitions and random mass partitions supported on indivisible units. It is a finite version of the Kingman's correspondence.

Proposition 5.4 (Finite Kingman's Correspondence)**.** *Let Π_n be an exchangeable partition of $[n]$. Then the distribution of Π_n is a mixture of finite paintboxes. That is, let $\pi_n^\downarrow$ be the corresponding partition of n arranged in decreasing block sizes. Then the distribution of Π_n conditioned on $\pi_n^\downarrow = \pi$ equals the distribution of the finite paintbox based on π.*

Proof. Exercise □

Theorem 5.5 (Kingman's correspondence)**.** *Let Π be an exchangeable random partition of $\mathbb{N}$. Then the law of Π is a mixture of paintboxes. That is,*

$$\mathbb{P}(\Pi \in \cdot) = \int_\Delta \mathbb{P}(|\Pi|^\downarrow \in dp)\rho_p(\cdot),$$

where ρ_p stands for the law of the paintbox based on p.

Proof. For a given partition π of $\mathbb{N}$, let $b_\pi : \mathbb{N} \to \mathbb{N}$ be the least element selector, where $b_\pi(i)$ is the smallest element of the block that contains i. Let $U_1, \dots$ be a sequence of i.i.d uniform, independent of Π. Conditioned on values of Π, define

$$X(i) = U_{b_\Pi(i)}.$$

Note that X defines Π a.s. via $i \sim j$ iff $X(i) = X(j)$.

We claim that $X = (X(1), \dots)$ is exchangeable. Indeed, for a permutation σ of $\mathbb{N}$,

$$\sigma(X)(i) = X(\sigma_i) = U_{b_\Pi(\sigma_i)} = U'_{b_{\sigma(\Pi)}(i)},$$

where $U'_j = U_{b_\Pi(i)}$. Since $\Pi \stackrel{d}{=} \sigma(\Pi)$ and independent of U, and U_i's are i.i.d., we conclude that

$$\sigma(X) = (X(\sigma_1), \dots) \stackrel{d}{=} X = (X(1), \dots).$$

So X is exchangeable.

By de Finetti's theorem, there exists some random probability measure μ on $[0,1]$ such that $X(i)$'s are i.i.d conditioned on μ. Now, conditioned on μ, let F be its distribution function. Introduce an independent sequence $V_1, \dots$ of i.i.d uniform random variables on $[0,1]$. Then conditioned on μ, $(F^{-1}(V_1), \dots) \stackrel{d}{=} X$. Recover Π a.s. from X. Then i and j belongs to the same block of Π if and only if V_i and V_j belong to the same interval in the domain of F^{-1}. So conditioned on μ, Π is distributed as a paintbox based on the intervals in the domain of F^{-1}. So Π is a mixture of paintboxes (with mixture law μ) as required. □

Let us now consider some consequences of Kingman's correspondence.

Definition 5.4. Let (π_n) be a sequence of partitions. Write π_n as $(N_{n,1}, N_{n,2}, \dots)$ in the order of least element. For each $i = 1, 2, \dots$, the limit

$$\lim_{n\to\infty} \frac{N_{n,i}}{n},$$

if it exists, is called the asymptotic frequency of block i. If all blocks of (π_n) have asymptotic frequency, say that (π_n) has asymptotic frequencies.

Fix $p \in \Delta$. Suppose $\sum_i p(i) = 1$. Let Π be a paintbox based on p, Π_n be its restriction on $[n]$. Write Π_n as $(N_{n,1}, N_{n,2}, \dots)$ in the order of least element and as

$(N_{n,1}^{\downarrow}, N_{n,2}^{\downarrow}, \ldots)$ in the order of decreasing block sizes. Then as $n \to \infty$, by law of large numbers, for each i,

$$\lim_{n\to\infty} \frac{N_{n,i}^{\downarrow}}{n} \to p^{\downarrow}(i) \tag{15}$$

and

$$\lim_{n\to\infty} \frac{N_{n,i}}{n} \stackrel{a.s.}{\to} P^*(i), \tag{16}$$

where P^* is the size-biased permutation of p. If $\sum_i p(i) < 1$, then we still have (15). For each i, either (16) holds, or

$$\lim_{n\to\infty} \frac{N_{n,i}}{n} \stackrel{a.s.}{\to} 0.$$

The later case occurs if and only if $\Pi(i)$ is a singleton. The set of singletons $\Pi(0) := \{i \in \mathbb{N} : \Pi(i) = \{i\}\}$ is a random set with mass $p(0) := 1 - \sum_i p(i)$.

That is, paintboxes (and hence mixture of paintboxes) have asymptotic frequencies. So by Kingman's theorem,

Corollary 5.6. *All exchangeable random partitions of $\mathbb{N}$ have asymptotic frequencies.*

5.3. **EPPF.**

Lemma 5.7. *Let $n \in \mathbb{N}$. A random partition Π_n of $[n]$ is exchangeable if and only if*

$$\mathbb{P}(\Pi_n = \{A_1, \ldots, A_k\}) = \mathsf{p}(|A_1|, \ldots, |A_k|)$$

for some probability mass function p taking values on finite (but not fixed) partitions of n, and p is symmetric in its argument.

The function p of Lemma 5.7 is called the exchangeable partition probability function (EPPF) of Π_n. Since p is symmetric, it is customary to list its arguments in decreasing sizes. Let $\mathcal{C}_n$ denote the set of partitions of n ordered in decreasing sizes (these are called compositions of n). So p maps $\mathcal{C}_n$ to $[0, 1]$.

Since Π_n is exchangeable, it induces a sequence of consistent exchangeable random partitions by restrictions to $m < n$. Thus, one can regard p as a map from $\bigcup_{m=1}^{n} \mathcal{C}_m \to [0, 1]$. Consistency implies that p satisfies the following addition rule: For each composition $(n_1, \ldots, n_k)$ of $m \text{ ¡ } n$,

$$\mathsf{p}(n_1, \ldots, n_k) = \mathsf{p}(n_1, \ldots, n_k, 1) + \sum_{j=1}^{k} \mathsf{p}(n_1, \ldots, n_{j-1}, n_j + 1, n_{j+1}, \ldots, n_k). \tag{17}$$

In addition,

$$\mathsf{p}(1) = 1. \tag{18}$$

Conversely, if $\mathsf{p} : \bigcup_{m=1}^{n} \mathcal{C}_m \to [0, 1]$ satisfies (17) and (18), then by Lemma 5.7, it is an EPPF of some exchangeable partition Π_n of $[n]$.

Definition 5.5 (Infinite EPPF)**.** An infinite EPPF is a function $\mathsf{p} : \bigcup_{m=1}^{\infty} \mathcal{C}_m \to [0, 1]$ that satisfies (17) and (18).

By Lemma 5.7 and definition of infinite exchangeable partitions, each infinite EPPF specifies the law of an exchangeable partition Π of $\mathbb{N}$.

5.4. **Kingman's correspondence in terms of EPPF.** Our goal is to cast Kingman's result in terms of the EPPF. This leads us to Theorem 5.8 of Pitman.

Fix $p \in \Delta$. Consider the paintbox Π based on p. What is its EPPF p? The answer is given by the Chinese Restaurant Process, introduced by Dubins and Pitman and later generalized by Pitman [10, §3].

Definition 5.6 (Partially Exchangeable Chinese Restaurant Process)**.** Start with an initially empty restaurant with an unlimited number of tables numbered $1, 2, \ldots$, each capable of seating an unlimited number of customers. Customers numbered $1, 2, \ldots$ arrive one by one and choose their seats according to the following rules.

(1) Customer 1 sits at table 1
(2) Given that the first n customers have sat at k tables, the $(n+1)$-st customer will:
   - Join table i with probability $p(i)$
   - Join a new table with probability $1 - \sum_{j=1}^{k} P(j)$.

Identify the seating configuration of n customers with a partition of $[n]$, where $i \sim j$ if and only if customers i and j sit at the same table. This process defines a consistent sequence (Π_n) of partitions of n with asymptotic block frequencies p. For blocks A_i listed in order of appearance, the law of Π_n is specified by

$$\mathbb{P}(\Pi_n = \{A_1, \ldots, A_k\}) = \mathsf{p}(n_1, \ldots, n_k) = \prod_{i=1}^{k} p(i)^{n_i - 1} \prod_{i=1}^{k-1} \left(1 - \sum_{j=1}^{i} p(j)\right), \tag{19}$$

where $n_i = |A_i|$. For general $p \in \Delta$, p defined in (19) is not symmetric in its argument. Thus, (Π_n) in general is not exchangeable.

Now let P be a random mass partition on Δ. Condition on the value of P, draw a random partition (Π_n) of $[n]$ by the Chinese Restaurant Process above. This gives a consistent sequence of partitions (Π_n), with asymptotic block frequencies P, and law

(20)

$$\mathbb{P}(\Pi_n = \{A_1, \ldots, A_k\}) = \mathsf{p}(n_1, \ldots, n_k) = \mathbb{E}\left[\prod_{i=1}^{k} P(i)^{n_i - 1} \prod_{i=1}^{k-1} \left(1 - \sum_{j=1}^{i} P(j)\right)\right].$$

The key theorem of this section states that this sequence (Π_n) is exchangeable if and only if p in (20) is symmetric in its argument. Roughly, this says that exchangeable partitions of $\mathbb{N}$ are in bijection with 'exchangeable mixtures' of Chinese Restaurants.

Theorem 5.8 (Pitman EPPF theorem)**.** *Let P be a random mass partition on Δ. Define $\mathsf{p} : \bigcup_{m=1}^{\infty} \mathcal{C}_m \to [0, 1]$*

$$\mathsf{p}(n_1, \ldots, n_k) = \mathbb{E}\left[\prod_{i=1}^{k} P(i)^{n_i - 1} \prod_{i=1}^{k-1} \left(1 - \sum_{j=1}^{i} P(j)\right)\right]. \tag{21}$$

There exists an exchangeable partition Π_∞ of $\mathbb{N}$ whose asymptotic block frequencies in order of appearance $\tilde{P}$ has the same distribution as P if and only if p is a symmetric function in its arguments. In this case, the EPPF of Π_∞ is p defined for $P(i) = \tilde{P}(i)$.

5.5. **EPPF for independent stick-break family.** Let P be a random mass partition on Δ invariant under size-biased permutation, such that its stick-breaking representation (12) consists of independent W_i's. For each fixed k, by Theorem 5.8, one can compute the EPPF p using (21) and condition on $P(k)$, take the expectation, condition on $P(k-1)$, take the expectation, and so on. The independence property simplifies the formula.

Recall from Theorem **??** that there are only a few families satisfying the independent stick-break criterion. The more interesting one is the $GEM(\alpha, \theta)$ family. The EPPF for $P \sim GEM(\alpha, \theta)$ is the Pitman sampling formula

$$\mathsf{p}_{\alpha,\theta}(n_1, \ldots, n_k) = \frac{(\theta+\alpha)_{k-1\uparrow\alpha}}{(\theta+1)_{n-1\uparrow 1}} \prod_{i=1}^{k} (1-\alpha)_{n_i-1\uparrow 1}, \tag{22}$$

where for integer m and real numbers x,a,

$$(x)_{m\uparrow a} = \prod_{i=0}^{m-1} (x+ia).$$

For example, for $\alpha = 0$, $\theta > 0$, this is the Ewens sampling formula

$$\mathsf{p}_{0,\theta}(n_1, \ldots, n_k) = \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{i=1}^{k} (n_i-1)!. \tag{23}$$

The seating plan for the Chinese Restaurant Process has the following simple description [11].

- Customer 1 sits at table 1.
- Given that n customers have sat at k tables, with n_i customers at table i, customer $(n+1)$ will
  - Sit at table i with probability $\frac{n_i-\alpha}{n+\theta}$
  - Sit at a new table with probability $\frac{\theta+k\alpha}{n+\theta}$.

6. Various applications of the GEM process

The Chinese Restaurant description of the GEM process gives a simple way to check if some random partition is GEM. We give some examples.

6.1. **A branching process description of the GEM distribution.** There is an interesting interpretation of the $GEM(\alpha, \theta)$ distribution in terms of a branching process in continuous time. In other lecture notes in this school we have seen Galton-Watson process, which is a branching process in discrete time.

In our case, we have a branching process in continuous time. Let $Z(t)$ be the population at time t. Each individual lives for a random time. Each individual has an independent exponential clock. When the clock rings, she produces a random number of offsprings.

Definition 6.1 ($GEM(\alpha, \theta)$ branching)**.** Fix $\alpha \in [0, 1]$, $\theta > 0$. Our population has two types, novel and clone. Each individual has a color, and has infinite lifetime. The reproduction rates are:

- A novel produces a novel at rate α, and independently produces a clone at rate $1-\alpha$.
- A clone produces a clone at rate 1.

Furthermore, novels migrate at rate θ, independent of the reproduction. The population starts with one novel at time $t = 0$. The coloring rules are:

- Each novel has a new unique color.
- Each clone has the same color as its parent.

Number the individuals $1, 2, \ldots$ in their order of appearance in the population. Let Π be the random partition of $\mathbb{N}$ generated by their colors, that is, $i \sim j$ if and only if i and j have the same color.

Proposition 6.1. *The random partition Π of $\mathbb{N}$ from the branching process in Defintion 6.1 is a $GEM(\alpha, \theta)$.*

Proof. The induced partition on $[n]$, Π_n, evolves according to a (α, θ) Chinese Restaurant Process. This uniquely determines the law of Π. We are done. □

With this description, one can use standard results from branching processes to derive properties of the $GEM(\alpha, \theta)$ process. For example,

Lemma 6.2. *Let Π be a $GEM(\alpha, 0)$ partition, Π_n be its restriction to $[n]$. Let K_n be the number of components of Pi_n. Then $\lim_{n\to\infty} \frac{K_n}{n^\alpha}$ exists a.s.. Denote this limit by S. Then the distribution of S is specified by the identity*

$$W^* = SW^\alpha,$$

where W is an exponential(1) independent of S, and $W^* \stackrel{d}{=}$ exponential(1).

Proof. Let N_t be the number of individuals at time t, N_t^* be the number of novel individuals at time t. let T_n be the first time where the n-th individual is born. Then

$$\frac{N^\alpha(T_n)}{e^{\alpha T_n}} \frac{K_n}{n^\alpha} = \frac{N^*(T_n)}{e^{\alpha T_n}}.$$

Note that K_n is independent of T_n, hence $\frac{N^\alpha(T_n)}{e^{\alpha T_n}} = \frac{n}{e^{\alpha T_n}}$ is independent of $\frac{K_n}{n^\alpha}$.

Now, $(N_t^*, t \geq 0)$ is a Yule process with rate α, that is, a pure birth process with transition rate $i\alpha$ from state i to state $i+1$. And $(N_t, t \geq 0)$ is a Yule process with rate 1. By known results for Yule processes,

$$\frac{N_t}{e^t} \stackrel{a.s.}{\to} W,$$

and

$$\frac{N_t^*}{e^{\alpha t}} \stackrel{a.s.}{\to} W^*,$$

where W and W^* are exponentially distributed with mean 1. Since $T_n \to \infty$ a.s. as $n \to \infty$, we have

$$\frac{N^\alpha(T_n)}{e^{\alpha T_n}} \stackrel{a.s.}{\to} W^\alpha,$$

and

$$\frac{N^*(T_n)}{e^{\alpha T_n}} \stackrel{a.s.}{\to} W.$$

So $\frac{K_n}{n^\alpha} \to S$ independent of W, as needed. □

By computing moments, one finds that

$$\mathbb{E}S^p = \frac{\Gamma(p+1)}{\Gamma(p\alpha+1)}.$$

In other words, S has the Mittag-Leffler distribution with parameter α.

Proposition 6.1 also reveals an interesting coupling between $GEM(0,\theta)$ and $GEM(\alpha,\theta)$ distribution for $\theta \geq 0$. If we ignore distinction between clone and novel, then this is just a birth process with rate 1, immigration at rate θ. So it is a $GEM(0,\theta)$ distribution, with $i \sim j$ if and only if i and j share a common ancestor. Conditioned on this partition, then the $GEM(\alpha,\theta)$ is a refinement, where each family is broken up by coloring rules according to a $GEM(\alpha,0)$.

Proposition 6.3. *Let $\alpha \in [0,1]$, $\theta \geq 0$. Break a stick of length 1 according to the $GEM(0,\theta)$ distribution. Then, break each of these stick further independently at random, according to the $GEM(\alpha,0)$ distribution. let Q be a size-biased random permutation of the lengths in this array. Then Q has the $GEM(\alpha,\theta)$ distribution.*

6.2. **Kingman coalescent with mutation.** As an application of our theory so far, we derive some properties of Kingman coalescent.

Proposition 6.4. *Let T be the line of descent tree from Kingman's coalescent on $[n]$. On each branch of the tree, introduce an independent Poisson point process with rate $\theta/2$ per unit length of marks, called mutations. Say that $i \sim j$ if and only if the unique path in T connecting i and j does not have mutations. This defines a random partition Π_n. Show that Π_n is a $GEM(0,\theta)$.*

Proof. We will use a coupling with the CRP for $GEM(0,\theta)$. Condition on Π_{n-1} and the tree T_{n-1}, consider element n. By Kingman's coalescent, for each $i = 1,\ldots,n-1$, there is an independent exponential clock for the pair (i,n), and the first clock that rings is the point on the tree T_{n-1} that n will join. Let S be the time that the first clock rings. Note that $S \sim exponential(n-1)$. For a some table j, the event that $\{n\text{joins } table j\}$ is the event that

Clock (n,i) is min for some i in table j AND *there is no mutation on $[0,2S]$ for an independent $PPP(\theta/2)$* .

Note that these two events are independent. The probability of the first event is $n_j/(n-1)$. The probability of the second event is the probability that $exponential(\theta) > exponential(n-1)$ for two independent exponentials. This happens with probability $\frac{n-1}{n-1+\theta}$. Thus, the probability of n joining table j is $\frac{n_j}{n-1+\theta}$. This gives the coupling required to the CRP for $GEM(0,\theta)$. □

6.3. **Random permutations.** Let σ_n be a random permutation of $[n]$. Write σ_n as a product of cycles. This defines a random partition of $[n]$, call it Π_n. There is an abundance of connections between random permutations and the GEM family. Here are some basic results.

Proposition 6.5. *Let $(\sigma_n, n \geq 1)$ be a sequence of random permutations of $[n]$ with the following properties*

- *σ_n is a uniform random permutation of $[n]$*
- *Conditioned on σ_n, σ_{n+1} is distributed as σ_n with the element $n+1$ inserted in one of its cycles.*

Let Π_n be the corresponding sequence of random partitions of $[n]$. Then Π_n is distributed as the restriction to $[n]$ of a $GEM(0,1)$.

Corollary 6.6. *Let K_n be the number of cycles in a uniform random permutation of $[n]$. Then*

$$K_n = \sum_{i=1}^{n} Bernoulli(\frac{1}{i})$$

for independent Bernoulli's. In particular,

$$\lim_{n\to\infty} \frac{K_n - \log n}{\sqrt{\log n}} \xrightarrow{d} N(0,1),$$

where $N(0,1)$ *is the standard Gaussian.*

7. Exercises

Exercise 7.1 (Size-biased pick)**.** Let $X_n = (X_n(1), \ldots, X_n(n))$ be i.i.d with distribution F supported on $[0, \infty)$, mean $\mu < \infty$. Let $X_n[1]$ be the first size-biased pick from X_n. Show that as $n \to \infty$,

$$\lim_{n\to\infty} \mathbb{P}(X_n[1] \in dx) = \frac{xF(dx)}{\mu}.$$

This distribution is called the size-biased distribution of F.

Exercise 7.2. Let F be a distribution supported on $[0, \infty)$ with mean $\mu < \infty$. Let G be its size-biased distribution, that is, $G(dx) = \frac{xF(dx)}{\mu}$. Give an example where G does not determine F.

Exercise 7.3 (Open problem, CSP 2.3.5)**.** Let P be a random mass partition, $P^*(1)$ be the first size-biased pick from P. Call the distribution of $P^*(1)$ the structural distribution of P. What is a necessary and sufficient condition for a distribution F on $(0, 1]$ to be a structural distribution?

Exercise 7.4 (Open problem, CSP 3.7)**.** Suppose that P is a random mass partition, $P = P^*$. Suppose that $P(1)$ is independent of the sequence $(\frac{P(i)}{1-P(1)}, i \geq 2)$. Does this necessary imply that P is a $GEM(\alpha, \theta)$?

Exercise 7.5. Kingman's paintbox gives one method to extend size-biased permutation from proper mass partitions (ie: mass partitions $p \in \Delta$ such that $\sum_i p(i) = 1$) to include improper mass partitions (ie: mass partitions $p \in \Delta$ such that $\sum_i p(i) < 1$). Consider the following naive extension of size-biased permutation: for $p \in \Delta$, perform Kingman's paintbox but do not write down a zero whenever the sample point falls outside of the paint buckets. Denote the resulting random permutation of p by $\hat{P}$. Let P^* denote the usual (Kingman's) size-biased permutation of p.

Construct a sequence $(p_n) \in \Delta$ such that p_n approaches $p \in \Delta$ component-wise, but $\hat{P}_n$ does not converge in distribution to $\hat{P}$.

Exercise 7.6. Prove Proposition 5.4.

Exercise 7.7. Give an example of a sequence of consistent partitions (π_n) of $[n]$ that does not have asymptotic frequencies.

Exercise 7.8 (Chinese Restaurant and Polya's urn)**.** Consider the Polya's urn process: start with w white balls and b black balls. At each step, remove a random ball from the urn and replace with two balls of the same color. Repeat this process n times, thus adding in total n balls to the urn. Let W_n denote the number of white balls newly added.

(1) Let X_i be the indicator that the i-th ball is black. Show that the X_i's are exchangeable.
(2) Show that for $x \in \{0, 1, \ldots, n\}$,
$$\mathbb{P}(W_n = x) = \binom{n}{x} \frac{(w + x - 1)!(b + n - x - 1)!(w + b - 1)!}{(w - 1)!(b - 1)!(w + b + n - 1)!}$$
(3) Show that $\lim_{n\to\infty} W_n/n$ has the $beta(w, b)$ distribution.
(4) Let P be a random mass partition such that $P(i)$ has the stick-breaking form (12) for W_i i.i.d $beta(a, b)$. Let p be defined as (20). Give a Chinese

Restaurant construction for p using Polya's urn scheme. (Note that unless $a = 1$, this partition is not exchangeable).

Exercise 7.9. Let X_i be the indicator of the event that i is the least element of some block of an exchangeable random partition Π_n of $[n]$. Show that the joint law $(X_i, 1 \leq i \leq n)$ determines the law of Π_n.

Exercise 7.10. Show that the X_i's of the previous exercise are independent if and only if Π_n is the partition obtained by running the Chinese Restaurant Process with parameters $(0, \theta)$, for $\theta \in [0, \infty]$, with obvious definitions for the limiting cases by continuity.

Exercise 7.11 (Open problem, CSP 2.1.5)**.** Let P be a distribution on binary strings of length n. Give necessary and sufficient conditions for P to be the law of block indicators of an exchangeable random partition as defined Problem 7.9.

Exercise 7.12 (Open problem, based on [3])**.** Fix For $\alpha \in [0, 1)$, $\theta > -\alpha$. Let $Frag_\alpha : \Delta^\downarrow \to \Delta^\downarrow$ be a random operator as follows: for a $p \in \Delta^\downarrow$, let $p^*(1)$ be a size-biased pick from p. Let $\eta = (\eta_1, \eta_2, \ldots) \sim PD(\alpha, 1 - \alpha)$ be independent from $p^*(1)$. Replace $p^*(1)$ in p by the sequence $p^*(1) \cdot \eta$, and return the resulting vector in decreasing order as $Frag_\alpha(p)$. In other words, this operator fragments a size-biased mass into smaller masses.

Now define the random operator $Coag_{\alpha,\theta} : \Delta^\downarrow \to \Delta^\downarrow$ as follows. For $p \in \Delta^\downarrow$, pick a $B \sim Beta(\frac{1-\alpha}{\alpha}, \frac{\theta+\alpha}{\alpha})$. If $\alpha = 0$, choose $B = \frac{1}{\theta+1}$. Conditoned on B, select each $p(i)$ with probability B independently at random. Remove the selected blocks and replace it with their sum, and return the resulting vector in decreasing order as $Coag_{\alpha,\theta}(p)$. In other words, this operator coagulates a random fraction of masses into one.

Dong, Goldschmidt and Martin [3] showed that $Frag_\alpha$ and $Coag_\alpha$ are inverses of each other, in the following sense.

Theorem 7.1. *Suppose X, Y are random ordered mass partitions. Then $Y \sim Frag_\alpha(X)$ for $X \sim PD(\alpha, \theta)$ if and only if $X \sim Coag_{\alpha,\theta}(Y)$ for $Y \sim PD(\alpha, \theta + 1)$.*

They gave a branching process interpretation of $Frag_\alpha$ as follows. Consider the $GEM(\alpha, \theta)$ branching process. Assume further that

- When a parent is killed, each child (first generation clone) becomes a novel and has a unique color different from all others in the population.
- Each clone actually generates novel individuals at rate α and independently generates clones at rate $1 - \alpha$, but this difference (in type as well as color) among its offspring is invisible until the clone becomes a novel due to its parents being killed.

Now kill the first individual, then the second, then the third, and so on. Starting with a $GEM(\alpha, \theta)$ population, it is clear that this process of killing is the $Frag_\alpha$ operator above.

(1) Is there a natural branching process description for the coagulation operator?
(2) Run the branching process with fragmentation and coagulation start with time $t > 0$. (Instead of waiting to obtain a $GEM(\alpha, \theta)$ population first before introducing killing or merging). What are the possible equilibria?

8. Supplementary materials

8.1. **Poisson point processes.** Let E be a Polish space, Λ a sigma-finite measure on E. For a random measure M on E, Borel measureable set E, let $\mathsf{M}(B)$ denotes its measure.

Definition 8.1. Say that M is a Poisson point process with intensity measure Λ (or Poisson measure with intensity Λ) if for every collection of k disjoint Borel measurable sets $B_1, \ldots, B_k$ with $\Lambda(B_i) < \infty$,

$$\mathbb{P}(\mathsf{M}(B_i) = n_i, i = 1, \ldots, k) = \prod_{i=1}^{k} \frac{[\Lambda(B_i)]^{n_i}}{n_i!} e^{-\Lambda(B_i)}.$$

A Poisson point process M is a sum of Dirac point masses

$$\mathsf{M} = \sum_{i \in I} \delta_{a_i}.$$

The random points $a_i \in E$ are called atoms of M. If $\Lambda(E) = \infty$, then one can take $I = \mathbb{N}$. If $\Lambda(E) < \infty$, then I is a finite set a.s.

The key property is the independence property: that for disjoint sets B_i with $\Lambda(B_i) < \infty$, $\mathsf{M}(B_i)$ are independent. The Poisson distribution follows from this property after eliminating trivial cases.

8.2. **de Finetti's theorem.** Say that the sequence of random variables $X = (X_1, X_2, \ldots)$ is exchangeable if the distribution of X is invariant under permutations of a finite subset of its terms.

Define the exchangeable sigma-algebra $\mathcal{E}$, generated by all events unchanged under permutations of a finite subset of the sequence X.

Theorem 8.1 (de Finetti's theorem)**.** *Suppose $X_1, X_2, \ldots$ are exchangeable. Conditioned on $\mathcal{E}$, $X_1, X_2, \ldots$ are i.i.d.*

Proof. See Durrett, theorem 4.6.5. □

References

[1] E. Barouch and G.M Kaufman. Probabilistic modelling of oil and gas discovery. In *Energy: mathematics and models*, pages 133–150. Philadelphia: SIAM, 1976.

[2] L. Chaumont and M. Yor. *Exercises in Probability: A Guided Tour from Measure Theory to Random Processes, via Conditioning*. Cambridge University Press, 2003.

[3] Rui Dong, Christina Goldschmidt, James B Martin, et al. Coagulation–fragmentation duality, poisson–dirichlet distributions and random recursive trees. *The Annals of Applied Probability*, 16(4):1733–1750, 2006.

[4] O. Kallenberg. *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.

[5] J. F. C. Kingman. Random partitions in population genetics. *Proc. R. Soc. Lond. A.*, 361:1–20, 1978.

[6] E. Lukacs. A characterization of the Gamma distribution. *Ann. Math. Statist.*, 26:319–324, 1955.

[7] G. P. Patil and C. Taillie. Diversity as a concept and its implications for random communities. *Bull. Int. Stat. Inst.*, XLVII:497 – 515, 1977.

[8] M. Perman, J. Pitman, and M. Yor. Size-biased sampling of poisson point processes and excursions. *Probability Theory and Related Fields*, 92:21–39, 1992. 10.1007/BF01205234.

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

[10] J. Pitman. Combinatorial Stochastic Processes - Saint-Flour Summer School of Probabilities XXXII - 2002. In *Combinatorial Stochastic Processes*, volume 1875 of *Lecture Notes in Mathematics*, pages 1+. Springer-Verlag Berlin, 2006.
[11] Jim Pitman. Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2):145–158, 1995.

University of Bonn, Germany and University of Texas at Austin