3

Sequential constructions of random partitions

This chapter introduces a basic sequential construction of random partitions, motivated at first by consideration of uniform random permutations of $[n]$ which are consistent in a certain sense as n varies. This leads to consideration of a particular two-parameter family of exchangeable random partition structures, which can be characterized in various ways, and which is closely related to gamma and stable subordinators.

3.1. The Chinese restaurant process This process defines a sequence of random permutations σ_n of the set $[n] := \{1, \dots, n\}$ such that the random partitions Π_n generated by cycles of σ_n are consistent as n varies. The most general exchangeable random partition of positive integers can be obtained this way.

3.2. The two-parameter model This section treats a particularly tractable family of random partitions of $\mathbb{N}$, parameterized by a pair of real numbers (α, θ) subject to appropriate constraints. The distribution $\mathbb{P}_{\alpha,\theta}$ of an (α, θ) partition is characterized by the product form of its partition probabilities, and by the induced distribution of its frequencies in size-biased random order. This distribution is known in the literature as GEM(α, θ), after Griffiths-Engen-McCloskey, while the corresponding distribution of ranked frequencies is known as the Poisson-Dirichlet distribution PD(α, θ). These distributions and associated random partitions arise in numerous contexts, such as population genetics, number theory, Bayesian nonparametric statistics, and the theory of excursions of Brownian motion and Bessel processes.

3.3. Asymptotics This section treats various asymptotic properties of partitions of $\mathbb{N}$, with special emphasis on the two-parameter family, whose asymptotic properties are radically different according to whether α is positive, negative, or zero. In particular, the number K_n of blocks of Π_n is bounded if $\alpha < 0$, grows like $\theta \log n$ if $\alpha = 0 < \theta$, and grows like a random multiple of n^α if $0 < \alpha < 1$, where the distribution of the multiplier depends on θ. For fixed $\alpha \in (0, 1)$ the probability laws of (α, θ) partitions turn out to be mutually absolutely continuous as θ varies with $\theta > -\alpha$, and the Radon-Nikodym density is described.

3.4. A branching process construction of the two-parameter model
This section offers a construction of the two-parameter model in terms of a branching process in continuous time.

3.1. The Chinese restaurant process

Consistent random permutations Consider a sequence of random permutations $(\sigma_n, n = 1, 2, \cdots)$ such that

(i) σ_n is a uniformly distributed random permutation of $[n]$ for each n;

(ii) for each n, if σ_n is written as a product of cycles, then σ_{n-1} is derived from σ_n by deletion of element n from its cycle.

For example, using standard cycle notation for permutations,

if $\sigma_5 = (134)(25)$ then $\sigma_4 = (134)(2)$;
if $\sigma_5 = (134)(2)(5)$ then $\sigma_4 = (134)(2)$.

It is easily seen that these requirements determine a unique distribution for the sequence (σ_n), which can be described as follows.

An initially empty restaurant has an unlimited number of circular tables numbered $1, 2, \ldots$, each capable of seating an unlimited number of customers. Customers numbered $1, 2, \cdots$ arrive one by one and are seated at the tables according to the following:

Simple random seating plan Person 1 sits at table 1. For $n \geq 1$ suppose that n customers have already entered the restaurant, and are seated in some arrangement, with at least one customer at each of the tables j for $1 \leq j \leq k$ say, where k is the number of tables occupied by the first n customers to arrive. Let customer $n + 1$ choose with equal probability to sit at any of the following $n + 1$ places: to the left of customer j for some $1 \leq j \leq n$, or alone at table $k + 1$. Define $\sigma_n : [n] \to [n]$ as follows. If after n customers have entered the restaurant, customers i and j are seated at the same table, with i to the left of j, then $\sigma_n(i) = j$, and if customer i is seated alone at some table then $\sigma_n(i) = i$. The sequence (σ_n) then has features (i) and (ii) above by a simple induction.

Many asymptotic properties of uniform random permutations can be read immediately from this construction. For instance, the number of occupied tables after n customers have been seated is

$$K_n = \#\{ \text{ cycles of } \sigma_n\} = Z_1 + \cdots + Z_n \tag{3.1}$$

where the Z_j is the indicator of the event that the jth customer is seated at a new table. By construction, the Z_j are independent Bernoulli$(1/j)$ variables, hence,

$$\frac{K_n}{\log n} \to 1 \text{ almost surely}, \qquad \frac{K_n - \log n}{(\log n)^{1/2}} \xrightarrow{d} B_1 \tag{3.2}$$

where B_1 is a standard Gaussian variable. This and other results about random permutations now recalled are well known.

Let Π_n be the partition of $[n]$ generated by the cycles of σ_n. Then Π_n is an exchangeable random partition of $[n]$, and the Π_n are consistent as n varies. Thus

the sequence $\Pi_\infty := (\Pi_n)$ is an exchangeable random partition of $\mathbb{N}$. Let X_n be the indicator of the event that the $(n+1)$th customer sits at table 1. Then the sequence $(X_n)_{n\geq 1}$ is an exchangeable sequence which evolves by the dynamics of Pólya's urn scheme (2.15) with $a = b = 1$. Hence $S_n := X_1 + \cdots + X_n$ has uniform distribution on $\{0, 1, \ldots, n\}$. Equivalently, the size $S_n + 1$ of the cycle of σ_{n+1} containing 1 has uniform distribution on $\{1, \ldots, n+1\}$. The asymptotic frequency of the class of Π_∞ containing 1 is the almost sure limit of S_n/n, which evidently has uniform distribution on $[0, 1]$.

The limit frequencies Let $(N_{n,1}, \ldots, N_{n,K_n})$ denote the sizes of blocks of Π_n, in order of least elements. In terms of the restaurant construction, $N_{n,i}$ is the number of customers seated at table i after n customers have been seated. From above, $N_{n,1}$ has uniform distribution on $[n]$. Similarly, given $N_{n,1} = n_1 < n$, $N_{n,2}$ has uniform distribution on $[n - n_1]$. And so on. Asymptotic behavior of this *discrete uniform stick-breaking scheme* is quite obvious: as $n \to \infty$, the relative frequencies $(N_{n,i}/n, i \geq 1)$ of the sizes of cycles of σ_n, which are in a size-biased random order, converge in distribution to the *continuous uniform stick-breaking sequence*

$$(\tilde{P}_1, \tilde{P}_2, \ldots) = (U_1,\ \overline{U}_1 U_2,\ \overline{U}_1 \overline{U}_2 U_3, \ldots)$$

where the U_i are independent uniform$[0, 1]$ variables, and $\bar{U} := 1 - U$. By an obvious combinatorial argument, the corresponding infinite exchangeable partition probability function (EPPF), which gives for each n the probability that Π_n equals any *particular* partition of $[n]$ with n_i elements in the ith cycle, for some arbitrary ordering of cycles, is

$$p_{0,1}(n_1, \ldots, n_k) := \frac{1}{n!} \prod_{i=1}^{k} (n_i - 1)!. \tag{3.3}$$

Compare with (2.19) to see that this *continuous uniform stick-breaking sequence* $(\tilde{P}_1, \tilde{P}_2, \ldots)$ has the same distribution as a size-biased permutation of the jumps of the Dirichlet process with exchangeable increments

$$(\Gamma_u/\Gamma_1, 0 \leq u \leq 1)$$

where $(\Gamma_u, u \geq 0)$ is a gamma process. Since the limiting ranked frequencies $P_i^\downarrow$ are recovered from the $(\tilde{P}_j)$ by ranking, it follows that if Γ_1 is a standard exponential variable independent of the limiting ranked frequencies $P_i^\downarrow$ defined by the Chinese restaurant construction of random permutations, then

$$\Gamma_1 P_1^\downarrow > \Gamma_1 P_2^\downarrow > \Gamma_1 P_3^\downarrow > \cdots > 0$$

are the ranked points of a Poisson point process whose intensity measure $x^{-1}e^{-x}dx$ on $(0, \infty)$ is the Lévy measure of the gamma process. This allows calculation of moments of the $P_i^\downarrow$. For instance

$$\mathbb{E}\#\{i : \Gamma_1 P_i^\downarrow > y\} = E_1(y) := \int_y^\infty x^{-1} e^{-x} dx.$$

So as $n \to \infty$ the asymptotic mean fraction of elements in the longest cycle of a uniform random permutation of $[n]$ is

$$\mathbb{E}(P_1^{\downarrow}) = \mathbb{E}(\Gamma_1)\mathbb{E}(P_1^{\downarrow}) = \mathbb{E}(\Gamma_1 P_1^{\downarrow}) = \int_0^{\infty} (1 - e^{-E_1(x)})dx.$$

This technique of *random scaling* to simplify the probabilistic structure of random partitions has many other applications. See for instance [85, 372, 374, 24]. The distribution of $(P_1^{\downarrow}, P_2^{\downarrow}, \ldots)$, constructed here from random permutations using the Chinese restaurant process, is known as the *Poisson-Dirichlet distribution* with parameter 1. Some references: Shepp-Lloyd [400], Vershik-Shmidt [422, 423], Flajolet-Odlyzko [156], Arratia-Barbour-Tavaré [27].

Generalization The Chinese restaurant construction is easily generalized to allow construction of a sequence of random permutations σ_n of $[n]$ such that the associated sequence of random partitions $\Pi_\infty := (\Pi_n)$ is the most general possible exchangeable random partition of integers, as discussed in Section 2.2. Recall that the corresponding exchangeable partition probability function (EPPF) $p(n_1, \ldots, n_k)$ gives for each $(n_1, \ldots, n_k)$ the the probability that Π_n equals any specific partition of $[n]$ into sets of sizes $(n_1, \ldots, n_k)$. In terms of the Chinese restaurant, the permutation σ_n is thought of as a configuration of n customers seated at K_n tables, where K_n is the number of cycles of σ_n. For present purposes, we only care about the random partition Π_n induced by the cycles of σ_n. So for $1 \le i \le K_n$ the statement "customer $n+1$ is placed at occupied table i" means Π_{n+1} is the partition of $[n+1]$ whose restriction to $[n]$ is Π_n, with $n+1$ belonging to the ith class of Π_n. Similarly "customer $n+1$ is placed at a new table" means Π_{n+1} is the partition of $[n+1]$ whose restriction to $[n]$ is Π_n, with $\{n+1\}$ a singleton block. Given an infinite EPPF $p(n_1, \ldots, n_k)$, a corresponding exchangeable random partition of $\mathbb{N}$ (Π_n) can thus be constructed as follows.

Random seating plan for an exchangeable partition The first customer is seated at the first table, that is $\Pi_1 = \{1\}$. For $n \ge 1$, given the partition Π_n, regarded as a placement of the first n customers at tables of the Chinese restaurant, with k occupied tables, the next customer $n+1$ is

- placed at occupied table j with probability $p(\ldots, n_j + 1, \ldots)/p(n_1, \ldots, n_k)$
- placed at new table with probability $p(n_1, \ldots, n_k, 1)/p(n_1, \ldots, n_k)$

In particular, it is clear that a simple product form for the EPPF will correspond to a simple prescription of these conditional probabilities. But before discussing specific examples, it is worth making some more general observations. Any sequential seating plan for the Chinese restaurant, corresponding to a *prediction rule* for the conditional distribution of Π_{n+1} given Π_n for each n, whereby $n+1$ is either assigned to one of the existing blocks of Π_n or declared to be a singleton block of Π_{n+1}, can be used to construct a random partition $\Pi_\infty := (\Pi_n)$ of the positive integers. Most seating plans will fail to produce a Π_∞ that is exchangeable. But it is instructive to experiment with simple plans

to see which ones do generate exchangeable partitions. According to Kingman's theory of exchangeable random partitions described in Section 2.2, a necessary condition for Π_∞ to be exchangeable is that for each i there exists an almost sure limiting frequency $\tilde{P}_i$ of customers seated at table i. More formally, this is the limit frequency of the ith block of Π_∞ when blocks are put in order of appearance. The simplest way to achieve this is to consider the following:

Random seating plan for a partially exchangeable partition Let $(P_i, i = 1, 2, \ldots)$ be an arbitrary sequence of random variables with $P_i \geq 0$ and $\sum_i P_i \leq 1$. Given the entire sequence $(P_i, i = 1, 2, \ldots)$ let the first customer be seated at the first table, and for $n \geq 1$, given the partition Π_n, regarded as a placement of the first n customers at tables of the Chinese restaurant, with k occupied tables, let the next customer $n + 1$ be

- placed at occupied table j with probability P_j
- placed at new table with probability $1 - \sum_{i=1}^{k} P_i$

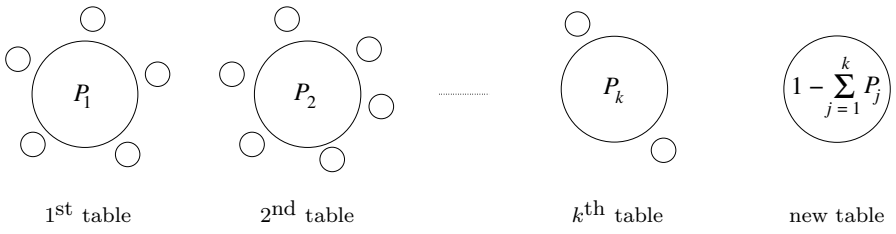

Figure 3.1: Chinese Restaurant Process with random seating plan.

By construction and the law of large numbers, for each i the limiting frequency of customers seated at table i exists and equals P_i. Moreover, by conditioning on the entire sequence P_i, the probability that Π_n equals any specific partition of $[n]$ into sets of sizes $(n_1, \ldots, n_k)$, in order of least elements, is given by the formula

$$p(n_1, \cdots, n_k) = \mathbb{E}\left[\prod_{i=1}^{k} P_i^{n_i - 1} \prod_{i=1}^{k-1} \left(1 - \sum_{j=1}^{i} P_j\right)\right] \tag{3.4}$$

Such a random partition of $[n]$ is called *partially exchangeable* [347]. These considerations lead to the following variation of Kingman's representation:

Theorem 3.1. [347] *Let (P_i) be a sequence of random variables with $P_i \geq 0$ and $\sum_i P_i \leq 1$, and let $p(n_1, \ldots, n_k)$ be defined by in formula* (3.4).
(i) *There exists an exchangeable random partition Π_∞ of $\mathbb{N}$ whose block frequencies in order of appearance $(\tilde{P}_i)$ are distributed like (P_i) if and only if the function $p(n_1, \ldots, n_k)$ is a symmetric function of $(n_1, \ldots, n_k)$ for each k.*
(ii) *If Π_∞ is such an exchangeable random partition of $\mathbb{N}$ with block frequencies $(\tilde{P}_i)$, then the EPPF of Π_∞ is $p(n_1, \ldots, n_k)$ defined by* (3.4) *for $P_i = \tilde{P}_i$, and*

the conditional law of Π_∞ *given* $(\tilde{P}_i)$ *is governed by the random seating plan for a partially exchangeable partition, described above.*

Proof. The "if" part of (i) is read from the preceding argument. See [347] for the "only if" part of (i). Granted that, part (ii) follows easily. □

Exercises

3.1.1. Let $\Pi_\infty := (\Pi_n)$ be an infinite exchangeable (or partially exchangeable) random partition, with $\tilde{N}_{n,i}$ the number of elements of $[n]$ in the ith class of Π_∞ to appear, and $\tilde{P}_i := \lim_n \tilde{N}_{n,i}/n$. The conditional distribution of $\tilde{N}_{n,1} - 1$ given $\tilde{P}_1$ is binomial$(n-1, \tilde{P}_1)$, hence the distribution of $\tilde{N}_{n,1}$ is determined by that of $\tilde{P}_1$ via

$$\mathbb{P}(\tilde{N}_{n,1} = j) = \binom{n-1}{j-1} \mathbb{E}\left[\tilde{P}_1^{j-1}(1-\tilde{P}_1)^{n-j}\right] \qquad (1 \le j \le n).$$

Use a similar description of the law of $(\tilde{N}_{n,1}, \ldots, \tilde{N}_{n,k})$ given $(\tilde{P}_1, \ldots, \tilde{P}_k)$ to show that for each $n, k \ge 1$ the law of $(\tilde{N}_{n,1}, \ldots, \tilde{N}_{n,k})$ is determined by that of $(\tilde{P}_1, \ldots, \tilde{P}_k)$.

Notes and comments

Basic references on random permutations are Feller [148] and Goncharov [177] from the 1940's. There is a nice bijection between the structure of records and cycles. For this and more see papers by Ignatov [207, 206], Rényi, Goldie [175], Stam [403, 405]. The fact that the cycle structure of uniform random permutations is consistent as n varies was pointed out by Greenwood [179]. Lester Dubins and I devised the Chinese Restaurant Process in the early 1980's as a way of constructing consistent random permutations and random partitions. The notion first appears in print in [14, (11.19)]. See also Joyce and Tavaré [224], and Arratia, Barbour and Tavaré [27] for many further results and references. The Chinese Restaurant Process and associated computations with random partitions have found applications in Bayesian statistics [109, 287, 210, 212], and in the theory of representations of the infinite symmetric group [243].

3.2. The two-parameter model

The EPPF's calculated in (2.17) and (2.19) suggest the following seating plan for the Chinese restaurant construction of a random partition of $\mathbb{N}$, say $\Pi_\infty := (\Pi_n)$, starting from $\Pi_1 := \{1\}$.

(α, θ) **seating plan** [347] Given at stage n there are k occupied tables, with n_i customers at the ith table, let the next customer be
- placed at occupied table i with probability $(n_i - \alpha)/(n + \theta)$,
- placed at new table with probability $(\theta + k\alpha)/(n + \theta)$.

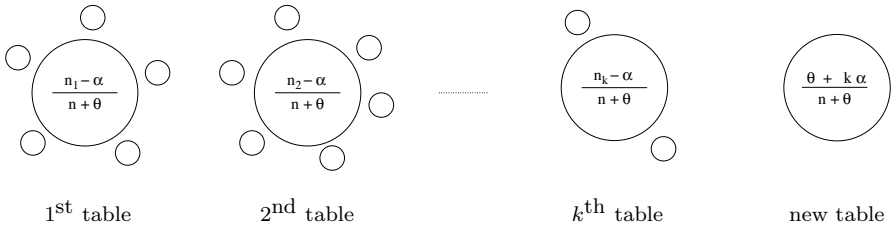

Figure 3.2: Chinese Restaurant Process with (α, θ) seating plan.

To satisfy the rules of probability it is necessary to suppose that

$$\begin{array}{ll} \bullet \text{ either} & \alpha = -\kappa < 0 \text{ and } \theta = m\kappa \text{ for some } m = 1, 2, \ldots \\ \bullet \text{ or} & 0 \le \alpha \le 1 \text{ and } \theta > -\alpha. \end{array} \tag{3.5}$$

Case ($\alpha = -\kappa < 0$ and $\theta = m\kappa$, for some $m = 1, 2, \ldots$) Compare the (α, θ) seating plan with Exercise 2.2.5 to see that in this case Π_∞ is distributed as if by sampling from a symmetric Dirichlet distribution with m parameters equal to κ. This can also be seen by comparison of (2.17) and (3.6) below.

Case ($\alpha = 0$ and $\theta > 0$) This is the weak limit of the previous case as $\kappa \to 0$ and $m\kappa \to \theta$. By consideration of this weak limit, or by the Blackwell-MacQueen urn scheme (2.18), such a Π_∞ is distributed as if by sampling from a Dirichlet process with parameter θ.

Case ($\alpha = 0$ and $\theta = 1$) This instance of the previous case corresponds to Π_∞ generated by the cycles of a consistent sequence of uniform random permutations, as in the previous section.

Case ($0 < \alpha < 1$ and $\theta > -\alpha$) This case turns out to be related to the stable subordinator of index α, as will be explained in detail in Section 4.2.

Theorem 3.2. [347] *For each pair of parameters (α, θ) subject to the constraints above, the Chinese restaurant with the (α, θ) seating plan generates an exchangeable random partition Π_∞ of $\mathbb{N}$. The corresponding EPPF is*

$$p_{\alpha,\theta}(n_1, \ldots, n_k) = \frac{(\theta + \alpha)_{k-1\uparrow\alpha} \prod_{i=1}^{k} (1 - \alpha)_{n_i - 1\uparrow 1}}{(\theta + 1)_{n-1\uparrow 1}} \tag{3.6}$$

where

$$(x)_{n\uparrow\alpha} := \prod_{i=0}^{n-1}(x + i\alpha). \tag{3.7}$$

The corresponding limit frequencies of classes, in size-biased order of least elements, can be represented as

$$(\tilde{P}_1, \tilde{P}_2, \ldots) = (W_1,\ \overline{W}_1 W_2,\ \overline{W}_1\overline{W}_2 W_3,\ \ \ldots) \tag{3.8}$$

where the W_i are independent, W_i has beta$(1-\alpha, \theta+i\alpha)$ distribution, and $\overline{W}_i := 1 - W_i$.

Proof. By construction, the probability that Π_n equals a specific partition of $[n]$ is found to depend only on the sizes $(n_1, \ldots, n_k)$ of the blocks of the partition, as indicated by $p_{\alpha,\theta}(n_1, \ldots, n_k)$. Since this function is symmetric in $(n_1, \ldots, n_k)$, each Π_n is exchangeable, and by construction the sequence (Π_n) is consistent. So $\Pi := (\Pi_n)$ is an exchangeable random partition of $\mathbb{N}$. The joint law of the W_i can be identified either using formula (3.4), or by repeated application of the beta-binomial relation described around (2.15) . □

Definition 3.3. (GEM and PD distributions) For (α, θ) subject to the constraints (3.5), call the distribution of size-biased frequencies $(\tilde{P}_j)$, defined by the residual allocation model (3.8), the *Griffiths-Engen-McCloskey distribution with parameters* (α, θ), abbreviated GEM(α, θ). Call the corresponding distribution on $\mathcal{P}^{\downarrow}_{[0,1]}$ of ranked frequencies $(P_i^{\downarrow})$ of an (α, θ) partition, obtained by ranking $(\tilde{P}_j)$ with GEM(α, θ) distribution, the *Poisson-Dirichlet distribution with parameters* (α, θ), abbreviated PD(α, θ).

Explicit but complicated formulae are known for the joint density of the first j coordinates of a PD(α, θ) distributed sequence [371], but these formulae are of somewhat limited use.

Characterizations of the two-parameter scheme. The closure of the two-parameter family of models consists of the original two-parameter family subject to the constraints on (α, θ) discussed above, plus the following models:
• the degenerate case with Π_n the partition of singletons for all n; this arises for $\alpha = 1$ and as the weak limit of (α, θ) partitions as $\theta \to \infty$ for any fixed $\alpha \in [0, 1)$.
• for each $m = 1, 2, \ldots$ the coupon collectors partition (2.16) defined by m frequencies identically equal to $1/m$; this is the weak limit of $(-\kappa, m\kappa)$ partitions as $\kappa \to \infty$ for fixed m.
• for each $0 \le p \le 1$ the mixture with weights p and $1-p$ of the one block partition and the partition into singletons. As observed by Kerov [240, (1.10)], this limit is obtained as $\alpha \to 1$ and $\theta \to -1$ with $(1-\alpha)/(1+\theta) \to p$. The cases $p = 0$ and $p = 1$ arise also as indicated just above.

Theorem 3.4. [240, 349] *Suppose that an exchangeable random partition* Π_∞ *of* $\mathbb{N}$ *has block frequencies* $\tilde{P}_j$ *(in order of least elements) such that* $0 < \tilde{P}_1 < 1$ *almost surely, and either*
(i) *The restriction* Π_n *of* Π_∞ *to* $[n]$ *is a Gibbs*$_{[n]}(v_\bullet, w_\bullet)$ *partition, meaning its EPPF is of the product form* (1.48), *for some pair of non-negative sequences* $v_\bullet$ *and* $w_\bullet$, *or*
(ii) *the frequencies* $\tilde{P}_j$ *are of the product form* (3.8) *for some independent random variables* W_i.
Then the distribution of Π_∞ *is either that determined by* (α, θ) *model for some* (α, θ), *or that of a coupon collectors partition, for some* $m = 2, 3, \ldots$.

Proof. Assuming (i), the form of the EPPF is forced by elementary arguments using addition rules of an EPPF [240]. Assuming (ii), the form of the distribution of the W_i is forced by symmetry of the EPPF and the formula (3.4). See [349] for details. □

See also Zabell [441] for closely related characterizations by the simple form of the prediction rule for (Π_n) defined by the (α, θ) seating plan. Note in particular the following consequence of the previous theorem:

Corollary 3.5. McCloskey [302]. *An exchangeable random partition* Π_∞ *of* $\mathbb{N}$ *has block frequencies* $\tilde{P}_j$ *(in order of least elements) of the product form* (3.8) *for some sequence of independent and identically distributed random variables* W_i *with* $0 < W_i < 1$ *if and only if the common distribution of the* W_i *is beta*$(1, \theta)$ *for some* $\theta > 0$, *in which case* Π_∞ *is generated by the* $(0, \theta)$ *model.*

This result of McCloskey is easily transformed into another characterization of the $(0, \theta)$ model due to Kingman. The following formulation is adapted from Aldous [14, p. 89].

Corollary 3.6. *Let* Π_∞ *be an exchangeable random partition of* $\mathbb{N}$. *The distribution of* Π_∞ *is governed by the* $(0, \theta)$ *model iff for each pair of integers* i *and* j, *the probability that* i *and* j *belong to the same component of* Π_∞ *is* $1/(1+\theta)$, *and* Π_∞ *has the following further property: for each pair of non-empty disjoint finite sets of positive integers* A *and* B, *the event that* A *is a block of the restriction of* Π_∞ *to* $A \cup B$ *is independent of the restriction of* Π_∞ *to* B.

Proof. That the $(0, \theta)$ model satisfies the independence condition is evident from the form of its EPPF. Conversely, in terms of the general Chinese restaurant construction, the exchangeability of Π_∞ plus the independence condition means that the process of seating customers at tables $2, 3, \ldots$, watched only when customers are placed at one of these tables, can be regarded in an obvious way as a copy of the original process of seating customers at tables $1, 2, 3, \ldots$, and that this copy of the original process is independent of the sequence of times at which customers are seated at table 1. It follows that if the block frequencies $(\tilde{P}_j)$ of Π_∞ are represented in the product form (3.8), then the asymptotic frequency $\tilde{P}_1 = W_1$ of customers arriving at table 1 is independent of the sequence $(W_2, W_3, \ldots)$ governing the relative frequencies of arrivals at tables $2, 3, \ldots$, and

that $(W_2, W_3, \ldots) \stackrel{d}{=} (W_1, W_2, \ldots)$. So the W_i are i.i.d. and the conclusion follows from Corollary 3.5. □

Problem 3.7. *Suppose an exchangeable random partition Π_∞ has block frequencies $(\tilde{P}_i)$ such that $0 < \tilde{P}_i < 1$ and $\tilde{P}_1$ is independent of the sequence $(\tilde{P}_i/(1-\tilde{P}_1), i \geq 2)$. Is Π_∞ necessarily some (α, θ) partition?*

Exercises

3.2.1. (Deletion of Classes.) Given a random partition Π_∞ of $\mathbb{N}$ with infinitely many classes, for each $k = 0, 1, \cdots$ let $\Pi_\infty(k)$ be the partition of $\mathbb{N}$ derived from Π_∞ by deletion of the first k classes. That is, first let $\Pi'_\infty(k)$ be the restriction of Π_∞ to $H_k := \mathbb{N} - G_1 - \cdots - G_k$ where $G_1, \cdots G_k$ are the first k classes of Π_∞ in order of appearance, then derive $\Pi_\infty(k)$ on $\mathbb{N}$ from $\Pi'_\infty(k)$ by renumbering the points of H_k in increasing order. The following are equivalent:

(i) for each k, $\Pi_\infty(k)$ is independent of the frequencies $(\tilde{P}_1, \cdots, \tilde{P}_k)$ of the first k classes of Π_∞;

(ii) Π_∞ is an (α, θ)-partition for some $0 \leq \alpha < 1$ and $\theta > -\alpha$, in which case $\Pi_\infty(k)$ is an $(\alpha, \theta + k\alpha)$-partition.

3.2.2. (Urn scheme for a $(\frac{1}{2}, 0)$ partition) Let an urn initially contain two balls of different colors. Draw 1 is a simple draw from the urn with replacement. Thereafter, balls are drawn from the urn, with replacement of the ball drawn, and addition of two more balls as follows. If the ball drawn is of a color never drawn before, it is replaced together with two additional balls of two distinct new colors, different to the colors of balls already in the urn. Whereas if the ball drawn is of a color that has been drawn before, it is replaced together with two balls of its own color. Let Π_n be the partition of $[n]$ generated by the colors of the first n draws from the urn. Then $\Pi_\infty := (\Pi_n)$ is a $(\frac{1}{2}, 0)$ partition.

3.2.3. (Number of blocks) Let $\mathbb{P}_{\alpha,\theta}$ govern $\Pi_\infty = (\Pi_n)$ as an (α, θ) partition, for some (α, θ) subject to the constraints (3.5). Let K_n be the number of blocks of Π_n:

$$K_n := |\Pi_n| = \sum_{j=1}^{n} |\Pi_n|_j = \sum_{i=1}^{n} X_i$$

where $|\Pi_n|_j$ is the number of blocks of Π_n of size j, and X_i is the indicator of the event that i is the least element of some block of Π_∞ (customer i sits at an unoccupied table). Under $\mathbb{P}_{\alpha,\theta}$ the sequence $(K_n)_{n\geq 1}$ is a Markov chain, starting at $K_1 = 1$, with increments in $\{0, 1\}$, and inhomogeneous transition probabilities

$$\mathbb{P}_{\alpha,\theta}(K_{n+1} = k+1 \,|\, K_1, \ldots, K_n = k) = \frac{k\alpha + \theta}{n + \theta} \tag{3.9}$$

$$\mathbb{P}_{\alpha,\theta}(K_{n+1} = k \,|\, K_1, \ldots, K_n = k) = \frac{n - k\alpha}{n + \theta}. \tag{3.10}$$

The distribution of K_n is given by

$$\mathbb{P}_{\alpha,\theta}(K_n = k) = \frac{(\theta+\alpha)_{k-1\uparrow\alpha}}{(\theta+1)_{n-1\uparrow}} S_\alpha(n,k) \tag{3.11}$$

where

$$S_\alpha(n,k) := B_{n,k}((1-\alpha)_{\bullet-1\uparrow}) = S_{n,k}^{-1,-\alpha} \tag{3.12}$$

is a generalized Stirling number of the first kind, as in (1.20). The expected value of K_n is

$$\mathbb{E}_{\alpha,\theta}(K_n) = \sum_{i=1}^{n} \frac{(\theta+\alpha)_{i-1\uparrow}}{(\theta+1)_{i-1\uparrow}} = \begin{cases} \displaystyle\sum_{i=1}^{n} \frac{\theta}{\theta+i-1} & \text{if } \alpha = 0 \\ \displaystyle\frac{(\theta+\alpha)_{n\uparrow}}{\alpha(\theta+1)_{n-1\uparrow}} - \frac{\theta}{\alpha} & \text{if } \alpha \neq 0. \end{cases} \tag{3.13}$$

3.2.4. (Serban Nacu [318]) **(Independent indicators of new blocks)** Compare with Exercise 2.1.4 and Exercise 4.3.4 . 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 $X_i, 1 \le i \le n$ are independent if and only if Π_n is a $(0,\theta)$ partition of $[n]$ for some $\theta \in [0,\infty]$, with the obvious definition by continuity in the two endpoint cases.

3.2.5. (Equilibrium of a coagulation/fragmentation chain) [301, 110, 361, 169] Let $\mathcal{P}_1^\downarrow$ be the space of real partitions of 1. Define a Markov kernel Q on $\mathcal{P}_1^\downarrow$ as follows. For $p = (p_i) \in \mathcal{P}_1^\downarrow$, let I and J be independent and identically distributed according to p. If $I = J$ then replace p_I by two parts $p_I U$ and $p_I(1-U)$ where U is uniform$(0,1)$ independent of I, J, and rerank, but if $I \neq J$ then replace the two parts p_I and p_J by a single part $p_I + p_J$, and rerank.

- (a) Show that PD$(0,1)$ is a Q-invariant measure.
- (b) [110] (hard). Show PD$(0,1)$ is the unique Q-invariant measure.
- (c) Modify the transition rule so that PD$(0,\theta)$ is an invariant measure.
- (d)(Open problem) Show that PD$(0,\theta)$ is the unique invariant measure for the modified rule.
- (d)(Open problem) Define some kind of coagulation/fragmentation kernel for which PD(α,θ) is an invariant measure.

3.2.6. The probabilities $q_{\alpha,\theta}(n,k) := P_{\alpha,\theta}(K_n = k)$ can be computed recursively from the forwards equations

$$q_{\alpha,\theta}(n+1,k) = \frac{n-k\alpha}{n+\theta} q_{\alpha,\theta}(n,k) + \frac{\theta+(k-1)\alpha}{n+\theta} q_{\alpha,\theta}(n,k-1),\ 1 \le k \le n. \tag{3.14}$$

and the boundary cases

$$q_{\alpha,\theta}(n,1) = \frac{(1-\alpha)_{n-1\uparrow}}{(\theta+1)_{n-1\uparrow}};\ q_{\alpha,\theta}(n,n) = \frac{(\theta+\alpha)_{n-1\uparrow\alpha}}{(\theta+1)_{n-1\uparrow\alpha}} \tag{3.15}$$

For instance, the distribution of K_3 is as shown in the following table:

k	1	2	3
$\mathbb{P}_{\alpha,\theta}(K_3 = k)$	$\frac{(1-\alpha)(2-\alpha)}{(\theta+1)(\theta+2)}$	$\frac{3(1-\alpha)(\theta+\alpha)}{(\theta+1)(\theta+2)}$	$\frac{(\theta+\alpha)(\theta+2\alpha)}{(\theta+1)(\theta+2)}$

3.2.7. Take $\theta = 0$ and use (3.11) to obtain a recursion for the $S_\alpha(n,k)$:

$$S_\alpha(n,1) = (1-\alpha)_{n-1\uparrow};\ S_\alpha(n,n) = 1 \tag{3.16}$$

$$S_\alpha(n+1,k) = (n-k\alpha)S_\alpha(n,k) + S_\alpha(n,k-1). \tag{3.17}$$

Toscano [415] used this recursion as his primary definition of these numbers, and obtained from it the formula

$$S_\alpha(n,k) = \frac{1}{k!}\Delta^k_{\alpha,x}(x)_{n\uparrow}\ \Big|_{x=0} \tag{3.18}$$

where $\Delta^k_{\alpha,x}$ is the kth iterate of the operator $\Delta_{\alpha,x}$, defined by $\Delta_{0,x} = \frac{d}{dx}$ (Jordan[222]) and for $\alpha \neq 0$,

$$(\Delta_{\alpha,x} f)(x) = \frac{f(x) - f(x-\alpha)}{\alpha}.$$

Check Toscano's formula

$$S_\alpha(n,k) = \frac{1}{\alpha^k k!}\sum_{j=1}^{k}(-1)^j\binom{k}{j}(-k\alpha)_{n\uparrow} \quad (\alpha \neq 0). \tag{3.19}$$

3.2.8. [132] Deduce the formula (3.13) for $\mathbb{E}_{\alpha,\theta}(K_n)$ by integration from the general formula (2.27) and the beta$(1-\alpha, \theta+\alpha)$ distribution of the frequency $\tilde{P}_1$ of the first block.

3.2.9. For real $p > 0$ let $[k]_p := \Gamma(k+p)/\Gamma(k)$ so that $[k]_p = (k)_{p\uparrow}$ for $p = 1, 2, \ldots$. For $0 < \alpha < 1$, and all real $p > 0$,

$$\mathbb{E}_{\alpha,0}[K_n]_p = \frac{\Gamma(p)[p\alpha]_n}{\Gamma(n)\alpha}. \tag{3.20}$$

3.2.10. Let $\Pi_\infty := (\Pi_n)$ be an exchangeable random partition of $\mathbb{N}$, with ranked frequencies denoted simply (P_j) instead of $(P_j^\downarrow)$. Let p be the EPPF of Π_∞, and let q be derived from p by (2.8).

- There is the formula

$$q(n_1, \ldots, n_k) = \mathbb{E}\left[\prod_{i=1}^{k}\left(\sum_{j=1}^{\infty} P_j^{n_i}\right)\right] \tag{3.21}$$

without further qualification if $\sum_j P_j = 1$ a.s., and with the qualification if $\mathbb{P}(\sum_j P_j < 1) > 0$ that $n_i \geq 2$ for all i.

- For each fixed $a > 0$, the distribution of (Π_n) on $\mathcal{P}_{\mathbb{N}}$, and that of (P_j) on $\mathcal{P}_1^{\downarrow}$, is uniquely determined by the values of $p(n_1, \ldots, n_k)$ for $n_i \geq a$ for all a. Similarly for q instead of p.
- (Π_n) is an (α, θ) partition, or equivalently (P_j) has $\text{PD}(\alpha, \theta)$ distribution, iff p satisfies the recursion

$$p(n_1 + 1, \ldots, n_k) = \frac{n_1 - \alpha}{n + \theta} p(n_1, \ldots, n_k). \tag{3.22}$$

  Note that p is subject also to the constraints of an EPPF, that is symmetry, the addition rule, and $P(1) = 1$. These constraints and (3.22) imply $p = p_{(\alpha,\theta)}$ as in (3.6).
- (Π_n) is an (α, θ) partition, or equivalently (P_j) has $\text{PD}(\alpha, \theta)$ distribution, iff q satisfies the recursion

$$q(n_1 + 1, \ldots, n_k) = \frac{n_1 - \alpha}{n + \theta} q(n_1, \ldots, n_k) + \sum_{s=2}^{n} q(n_1 + n_s, \ldots, n_k) \tag{3.23}$$

  where the number of arguments of $q(n_1 + 1, \ldots, n_k)$ and $q(n_1, \ldots, n_k)$ is k, and the number of arguments of $q(n_1 + n_s, \ldots, n_k)$ is $k - 1$, with n_s the missing argument. Note that q is subject also to the a priori constraints of symmetry, and $q(1, \ldots, 1) \equiv 1$. These constraints and (3.23) imply that $q = q_{(\alpha,\theta)}$ is given by formula (2.8) for $p = p_{(\alpha,\theta)}$ as in (3.6). There does not appear to be any simpler formula for $q_{(\alpha,\theta)}$.

In the case $\theta = 0$, the recursion (3.23) for $q = q_{(\alpha,0)}$ was derived by Talagrand [412, Proposition 1.2.2], using relations of Ghirlanda-Guerra [166] in the context of Derrida's random energy model [105] in the theory of spin glasses. The appearance of $\text{PD}(\alpha, 0)$ in that setting is explained in Exercise 4.2.1 . Once the parallel between (3.22) and (3.23) has been observed for $\theta = 0$, the result for general θ is easily guessed, and can be verified algebraically using (2.8). The identities (2.8) and (3.22) have a transparent probabilistic meaning, the latter in terms of the Chinese Restaurant Process. Can (3.23) too be understood without calculation in some setting? Does (3.23) or $\text{PD}(\alpha, \theta)$ have an interpretation in terms of spin glass theory for $\theta \neq 0$?

3.3. Asymptotics

The asymptotic properties of (α, θ) partitions of $[n]$ for large n depend on whether α is negative, 0, or positive. Recall the notations $K_n := |\Pi_n|$ for the number of blocks of Π_n, and $|\Pi_n|_j$ for the number of blocks of Π_n of size j. So

$$K_n := |\Pi_n| = \sum_{j=1}^{n} |\Pi_n|_j$$

Case $(\alpha < 0)$**.** Then $\theta = -m\alpha$ for some positive integer m, and $K_n = m$ for all sufficiently large n almost surely.

Case ($\alpha = 0$)**.** Immediately from the prediction rule, for a $(0, \theta)$ partition, the X_i are independent Bernoulli($\theta/(\theta + i - 1)$) variables. Hence [263]

$$\lim_{n\to\infty} \frac{K_n}{\log n} = \theta, \quad \text{a.s. } \mathbb{P}_{0,\theta} \text{ for every } \theta > 0. \tag{3.24}$$

Moreover, it follows easily from Lindeberg's theorem that the $\mathbb{P}_{0,\theta}$ distribution of $(K_n - \theta \log n)/\sqrt{\theta \log n}$ converges to standard normal as $n \to \infty$. By consideration of the Ewens sampling formula (2.20), for each fixed k

$$\{(|\Pi_n|_j, j \geq 1); \mathbb{P}_{0,\theta}\} \xrightarrow{d} (Z_{\theta,j}, j \geq 1) \tag{3.25}$$

meaning that under $\mathbb{P}_{0,\theta}$ which governs Π_∞ as a $(0, \theta)$ partition, the finite dimensional distributions of the counts $(|\Pi_n|_j, j \geq 1)$ converge without normalization to those of $(Z_{\theta,j}, j \geq 1)$, where the $Z_{\theta,j}$ are independent Poisson variables with parameters θ/j. See [27] for various generalizations and refinements of these results.

Case ($0 < \alpha < 1$)**.** Now K_n is a sum of dependent indicators X_i. It is easily seen from (3.13) and Stirling's formula that

$$\mathbb{E}_{\alpha,\theta} K_n \sim \frac{\Gamma(\theta + 1)}{\alpha \Gamma(\theta + \alpha)}\, n^\alpha$$

which indicates the right normalization for a limit law.

Theorem 3.8. *For $0 < \alpha < 1$, $\theta > -\alpha$, under $\mathbb{P}_{\alpha,\theta}$ as $n \to \infty$,*

$$K_n/n^\alpha \to S_\alpha \text{ almost surely} \tag{3.26}$$

and in pth mean for every $p > 0$, for a strictly positive random variable S_α, with continuous density

$$\frac{d}{ds}\mathbb{P}_{\alpha,\theta}(S_\alpha \in ds) = g_{\alpha,\theta}(s) := \frac{\Gamma(\theta + 1)}{\Gamma(\frac{\theta}{\alpha} + 1)} s^{\frac{\theta}{\alpha}} g_\alpha(s) \qquad (s > 0) \tag{3.27}$$

where $g_\alpha = g_{\alpha,0}$ is the Mittag-Leffler density (0.43) *of the $\mathbb{P}_{\alpha,0}$ distribution of S_α, whose pth moment is $\Gamma(p + 1)/\Gamma(p\alpha + 1)$.*

Proof. Fix $\alpha \in (0, 1)$. Let $\mathcal{F}_n$ be the field of events generated by Π_n. The formula (3.6) for the EPPF of Π_n under $\mathbb{P}_{\alpha,\theta}$ gives the likelihood ratio

$$M_{\alpha,\theta,n} := \left.\frac{d\mathbb{P}_{\alpha,\theta}}{d\mathbb{P}_{\alpha,0}}\right|_{\mathcal{F}_n} = \frac{f_{\alpha,\theta}(K_n)}{f_{1,\theta}(n)} \tag{3.28}$$

where for $\theta > -\alpha$

$$f_{\alpha,\theta}(k) := \frac{(\theta + \alpha)_{k-1\uparrow\alpha}}{(\alpha)_{k-1\uparrow\alpha}} = \frac{\Gamma(\frac{\theta}{\alpha} + k)}{\Gamma(\frac{\theta}{\alpha} + 1)\Gamma(k)} \sim \frac{k^{\theta/\alpha}}{\Gamma(\frac{\theta}{\alpha} + 1)} \text{ as } k \to \infty. \tag{3.29}$$

Thus, for each $\theta > -\alpha$,

$$(M_{\alpha,\theta,n}, \mathcal{F}_n; n = 1, 2, \ldots) \text{ is a positive } \mathbb{P}_{\alpha,0}\text{-martingale.}$$

By the martingale convergence theorem $M_{\alpha,\theta,n}$ has a limit $M_{\alpha,\theta}$ almost surely $(\mathbb{P}_{\alpha,0})$. Theorem 3.2 shows that Π_∞ has infinitely many blocks with strictly positive frequencies, and hence $K_n \to \infty$ almost surely $(\mathbb{P}_{\alpha,0})$ so (3.29) gives

$$M_{\alpha,\theta,n} \sim \frac{\Gamma(\theta+1)}{\Gamma(\frac{\theta}{\alpha}+1)} \left(\frac{K_n}{n^\alpha}\right)^{\theta/\alpha} \quad \text{almost surely}(\mathbb{P}_{\alpha,0}) \tag{3.30}$$

Moreover the ratio of the two sides in (3.30) is bounded away from 0 and ∞. Using (3.20), it follows that for each $\theta > -\alpha$, the martingale $M_{\alpha,\theta,n}$ is bounded in $L^p(\mathbb{P}_{\alpha,0})$, hence convergent in $L^p(\mathbb{P}_{\alpha,0})$ to $M_{\alpha,\theta}$ for every $p > 1$. Hence

$$\mathbb{E}_{\alpha,0} M_{\alpha,\theta} = 1. \tag{3.31}$$

But also by (3.30),

$$\frac{\Gamma(\theta+1)}{\Gamma(\frac{\theta}{\alpha}+1)} \left(\frac{K_n}{n^\alpha}\right)^{\theta/\alpha} \to M_{\alpha,\theta} \;=\; \frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)} S_\alpha^{\theta/\alpha} \tag{3.32}$$

$\mathbb{P}_{\alpha,0}$ almost surely and in L^p, where $S_\alpha := M_{\alpha,\alpha}/\Gamma(\alpha+1)$. Now (3.31) and (3.32) yield the moments of the $\mathbb{P}_{\alpha,0}$ distribution of S. Since these are the moments (0.42) of the Mittag-Leffler distribution, the conclusions of the theorem in case $\theta = 0$ are evident. The corresponding results for $\theta > 0$ follow immediately from the results for $\theta = 0$, due to the following corollary of the above argument. □

Corollary 3.9. *Let $\mathbb{P}_{\alpha,\theta}$ denote the distribution on $\mathcal{P}_\mathbb{N}$ of an (α,θ)-partition $\Pi_\infty := (\Pi_n)$. For each $0 < \alpha < 1$, $\theta > -\alpha$, the laws $\mathbb{P}_{\alpha,\theta}$ and $\mathbb{P}_{\alpha,0}$ are mutually absolutely continuous, with density*

$$\frac{d\mathbb{P}_{\alpha,\theta}}{d\mathbb{P}_{\alpha,0}} = \frac{\Gamma(\theta+1)}{\Gamma(\frac{\theta}{\alpha}+1)} S_\alpha^{\frac{\theta}{\alpha}} \tag{3.33}$$

where S_α is the almost sure limit of $|\Pi_n|/n^\alpha$ under $\mathbb{P}_{\alpha,\theta}$ for every $\theta > -\alpha$.

Proof. This is read from the previous argument, by martingale theory. □

In view of Corollary 3.9, the limit random variable

$$S_\alpha := \lim_n |\Pi_n|/n^\alpha \tag{3.34}$$

plays a key role in describing asymptotic properties of an (α,θ) partition Π_∞.

Definition 3.10. Say that Π_∞, an exchangeable partition of $\mathbb{N}$ *has α-diversity* S_α if the limit (3.34) exists and is strictly positive and finite almost surely.

This limit random variable S_α can be characterized in a number of different ways, by virtue of the following lemma. According to Theorem 3.8 and Corollary 3.9, the conditions of the Lemma apply to an (α, θ) partition Π_∞, for each $\alpha \in (0,1)$, and each $\theta > -\alpha$.

Write $A_i \sim B_i$ if $A_i/B_i \to 1$ almost surely as $i \to \infty$.

Lemma 3.11. *Fix $\alpha \in (0,1)$, An exchangeable random partition Π_∞ has α-diversity S_α, defined as an almost sure limit* (3.34)*, which is strictly positive and finite, if and only if*

$$P_i^\downarrow \sim Z i^{-1/\alpha} \text{ as } i \to \infty \tag{3.35}$$

for some random variable Z with $0 < Z < \infty$. In that case S_α and Z determine each other by

$$Z^{-\alpha} = \Gamma(1-\alpha) S_\alpha$$

and the following conditions also hold:

$$(1 - \textstyle\sum_{i=1}^k \tilde{P}_i) \sim \alpha\Gamma(1-\alpha)^{1/\alpha} Z k^{1-1/\alpha} \text{ as } k \to \infty \tag{3.36}$$

where $\tilde{P}_i$ is the frequency of the ith block of Π_∞ in order of appearance;

$$|\Pi_n|_j \sim p_\alpha(j) S_\alpha n^\alpha \text{ for each } j = 1, 2, \ldots \tag{3.37}$$

where $|\Pi_n|_j$ is the number of blocks of Π_n of size j, and $(p_\alpha(j), j = 1, 2, \ldots)$ is the discrete probability distribution defined by

$$p_\alpha(j) = (-1)^{j-1} \binom{\alpha}{j} = \frac{\alpha(1-\alpha)_{j-1\uparrow}}{j!} \tag{3.38}$$

and

$$|\Pi_n|_j / |\Pi_n| \to p_\alpha(j) \text{ for every } j = 1, 2, \ldots \text{ a.s. as } n \to \infty. \tag{3.39}$$

Sketch of proof. By Kingman's representation, it suffices to establish the Lemma for Π_∞ with deterministic frequencies $P_i^\downarrow$. Most of the claims in this case can be read from the works of Karlin [232] and Rouault [392], results in the theory of regular variation [66], and large deviation estimates for sums of bounded independent random variables obtained by Poissonization [158]. □

The discrete probability distribution (3.38) arises in other ways related to the positive stable law of index α. See the exercises below, and [351, 355] for further references.

The ranked frequencies

Theorem 3.12. Case $(\alpha = 0)$ [153]. *A random sequence $(P_i^\downarrow)$ has* PD$(0, \theta)$ *distribution iff for Γ_θ a gamma(θ) variable independent of $(P_i^\downarrow)$, the sequence $(\Gamma_\theta P_i^\downarrow)$ is the ranked sequence of points of a Poisson process on $(0, \infty)$ with intensity $\theta x^{-1} e^{-x} dx$.*

Proof. This follows from previous discussion.

Theorem 3.13. Case $(0 < \alpha < 1)$.
(i) [341] *A random sequence* $(P_i^{\downarrow})$ *with* $\sum_i P_i^{\downarrow} = 1$ *has* $\mathrm{PD}(\alpha, 0)$ *distribution iff the limit*

$$S_\alpha := \lim_{i\to\infty} i\Gamma(1-\alpha)(P_i^{\downarrow})^\alpha \tag{3.40}$$

exists almost surely, and the sequence $(S_\alpha^{-1/\alpha} P_i^{\downarrow})$ *is the ranked sequence of points of a Poisson process on* $(0,\infty)$ *with intensity* $\alpha\Gamma(1-\alpha)^{-1}x^{-\alpha-1}dx$.
(ii) [351] *For* $\theta > -\alpha$, *and* $(P_i^{\downarrow})$ *the* $\mathrm{PD}(\alpha,\theta)$ *distributed sequence of ranked frequencies of an* (α,θ)*-partition* Π_∞, *the limit* S_α *defined by* (3.40) *exists and equals almost surely the* α*-diversity of* Π_∞, *that is*

$$S_\alpha = lim_{n\to\infty} |\Pi_n|/n^\alpha. \tag{3.41}$$

(iii) [341] *For* $\theta > -\alpha$, *the* $\mathrm{PD}(\alpha,\theta)$ *distribution is absolutely continuous with respect to* $\mathrm{PD}(\alpha,0)$, *with density*

$$\frac{d\,\mathrm{PD}(\alpha,\theta)}{d\,\mathrm{PD}(\alpha,0)} = \frac{\Gamma(\theta+1)}{\Gamma(\frac{\theta}{\alpha}+1)} S_\alpha^{\frac{\theta}{\alpha}}$$

for S_α *as in* (3.40).

Proof. Part (i) follows from results of [341] which are reviewed in Section 4.1. If a sequence of ranked frequencies admits the limit (3.40) almost surely in $(0,\infty)$, then it can be evaluated as in (3.41) using the associated random partition Π_∞. This was shown by Karlin [232] for Π_∞ with deterministic frequencies, and the general result follows by conditioning on the frequencies. This gives (ii), and (iii) is just a translation of Corollary 3.9 via Kingman's correspondence, using (ii). □

In particular, parts (i) and (ii) imply that if $S := S_\alpha$ is the α-diversity of an $(\alpha,0)$ partition Π_∞, then $S^{-1/\alpha}$ has the stable(α) law whose Lévy density is $\alpha\Gamma(1-\alpha)^{-1}x^{-\alpha-1}dx$. This can also be deduced from Theorem 3.8, since we know from (0.43) that a random variable S has Mittag-Leffler(α) law iff $S^{-1/\alpha}$ has this stable(α) law. It must also be possible to establish the Poisson character of the random set of points $\{S^{-1/\alpha}P_i^{\downarrow}\} = \{S^{-1/\alpha}\tilde{P}_i\}$ by some direct computation based on the prediction rule for an $(\alpha,0)$ partition, but I do not know how to do this.

Exercises

3.3.1. [351] **(Poisson subordination)** Fix $\alpha \in (0,1)$, and let Z be the closure of the range of a stable subordinator of index α. Let N be a homogeneous Poisson point process on $\mathbb{R}_{>0}$ and let X_i be the number of points of N in the ith interval component of the complement of Z that contains at least one point

of N. Then the X_i are independent and identically distributed with distribution $(p_\alpha(j), j = 1, 2, \ldots)$ as in (3.38). Generalize to a drift-free subordinator that is not stable.

3.3.2. If $\mathbb{P}_\alpha$ governs independent $X_1, X_2, \ldots$ with distribution (3.38), as in the previous exercise, then

$$\mathbb{E}_\alpha(z^{X_i}) = 1 - (1-z)^\alpha. \tag{3.42}$$

Let $S_k := X_1 + \cdots + X_k$. Then

$$\mathbb{P}_\alpha(S_k = n) = [z^n](1 - (1-z)^\alpha)^k \tag{3.43}$$

so the generalized Stirling number $S_\alpha(n,k)$ in (3.11), (3.12), (3.18), (3.19), acquires another probabilistic meaning as

$$S_\alpha(n,k) = \frac{n!}{k!}\alpha^{-k}\mathbb{P}_\alpha(S_k = n) \tag{3.44}$$

and the distribution of K_n for an (α, θ) partition is represented by the formula

$$\mathbb{P}_{\alpha,\theta}(K_n = k) = \frac{(\frac{\theta}{\alpha}+1)_{k-1\uparrow}}{\alpha(\theta+1)_{n-1\uparrow}}\,\frac{n!}{k!}\mathbb{P}_\alpha(S_k = n). \tag{3.45}$$

3.3.3. (A local limit theorem) [355] In the setting of Theorem 3.8, establish the local limit theorem

$$\mathbb{P}_{\alpha,\theta}(K_n = k) \sim g_{\alpha,\theta}(s)n^{-\alpha} \text{ as } n \to \infty \text{ with } k \sim sn^\alpha. \tag{3.46}$$

Deduce from (3.11) and (3.46) an asymptotic formula for $S_\alpha(n,k)$ as $n \to \infty$ with $k \sim sn^\alpha$.

3.3.4. For $0 < \alpha < 1$, as $n \to \infty$, for each $p > 0$

$$\mathbb{E}_{\alpha,\theta}(K_n^p) \sim n^{\alpha p}\frac{\Gamma(\frac{\theta}{\alpha}+p+1)\Gamma(\theta+1)}{\Gamma(\theta+p\alpha+1)\Gamma(\frac{\theta}{\alpha}+1)}. \tag{3.47}$$

Notes and comments

Lemma 3.11 is from unpublished work done jointly with Ben Hansen. There is much interest in power law behaviour, such as described by Lemma 3.11, in the literature of physical processes of fragmentation and coagulation. See [306] and papers cited there.

3.4. A branching process construction of the two-parameter model

This section offers an interpretation of the (α, θ) model for $0 \le \alpha \le 1, \theta > -\alpha$, in terms of a branching process in continuous time, which generalizes the model

of Tavaré [414] in case $\theta = 0$. This brings out some interesting features of (α, θ) partitions which are hidden from other points of view.

Fix $0 \le \alpha \le 1$. Consider a population of individuals of two types, *novel* and *clone*. Each individual is assigned a color, and has infinite lifetime. Starting from a single novel individual at time $t = 0$, of some first color, suppose that each individual produces offspring throughout its infinite lifetime as follows:

- Novel individuals produce novel offspring according to a Poisson process with rate α, and independently produce clone offspring according to a Poisson process with rate $1 - \alpha$.
- Clones produce clone offspring according to a Poisson process with rate 1.

Each novel individual to appear is assigned a new color, distinct from the colors of all individuals in the current population. Each clone has the same color as its parent. Let

$$N_t := \text{ number of all individuals at time } t$$

$$N_t^* := \text{ number of novel individuals at time } t.$$

Thus $N_0^* = N_0 = 1$, and $1 \le N_t^* \le N_t$ for all $t \ge 0$. The process $(N_t^*, t \ge 0)$ is a *Yule process* with rate α, that is a pure birth process with transition rate $i\alpha$ from state i to state $i + 1$. Similarly, $(N_t, t \ge 0)$ is a Yule process with rate 1. Think of the individuals as colored balls occupying boxes labelled by $\mathbb{N} := \{1, 2, \ldots\}$. So the nth individual to be born into the population is placed in box n. The colors of individuals then induce a random partition of $\mathbb{N}$. Each novel individual appears in the first of an infinite subset of boxes containing individuals of the same color.

Proposition 3.14. *The random partition of* Π *of* $\mathbb{N}$*, generated by the colors of successive individuals born into the population described above, is an* $(\alpha, 0)$ *partition. The number of blocks in the induced random partition of* $[n]$ *is the value of* N_t^* *at every time* t *such that* $N_t = n$*. For each* $t > 0$*, the conditional distribution of* N_t^* *given* $N_t = n$ *is identical to the distribution of* K_n*, the number of blocks of the partition of* $[n]$*, for an* $(\alpha, 0)$ *partition.*

Proof. Let Π_n be the partition of $[n]$ induced by Π. It follows easily from the description of the various birth rates that $(\Pi_n, n = 1, 2, \ldots)$ is a Markov chain with transition probabilities described by the (α, θ) urn scheme, independent of the process $(N_t, t \ge 0)$. □

According to a standard result for the Yule process

$$e^{-t} N_t \overset{a.s.}{\to} W$$

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

where W and W^* are both exponentially distributed with mean 1. Combined with Proposition 3.14 this implies

Corollary 3.15. $W^* = SW^\alpha$ *where* $S := \lim_{n\to\infty} K_n/n^\alpha$ *is independent of* W*.*

A formula for the moments of S follows immediately, confirming the result of Theorem 3.8 that S has Mittag-Leffler distribution with parameter α.

To present a continuous time variation of the residual allocation model, let $N_t^{(k)} =$ be the number of individuals of the kth color to appear that are present in the population at time t. From the previous analysis, as $t \to \infty$

$$\left(e^{-t}N_t, \frac{N_t^{(1)}}{N_t}, \frac{N_t^{(2)}}{N_t}, \frac{N_t^{(3)}}{N_t}, \ldots\right) \stackrel{a.s.}{\to} (W,\ X_1,\ \overline{X}_1X_2,\ \overline{X}_1\overline{X}_2X_3, \ldots) \tag{3.48}$$

where W, $X_1, X_2, \ldots$ are independent, W has exp(1) distribution, and X_i (denoted W_i in (3.8)) has beta$(1-\alpha, i\alpha)$ distribution. Equivalently,

$$e^{-t}(N_t, N_t^{(1)}, N_t^{(2)}, \ldots) \stackrel{a.s.}{\to} (W, WX_1, W\overline{X}_1X_2, \ldots). \tag{3.49}$$

In particular, the limit law of $e^{-t}(N_t^{(1)}, N_t - N_t^1)$ is that of WX_1 and $W\overline{X}_1$, which are independent gamma$(1-\alpha)$ and gamma(α) respectively. The subsequent terms have more complicated joint distributions.

Case $0 \le \alpha \le 1$, $\theta > -\alpha$. Define a population process with two types of individuals, exactly as in the case $\theta = 0$ treated as above, but with the following modification of the rules for the offspring process of the first novel individual only. This first individual produces novel offspring at rate $\alpha + \theta$ (instead of α as before) and clone offspring at rate $1 - \alpha$ (exactly as before). Both clone and novel offspring of the first individual reproduce just as before. And the rules for coloring are just as before. It is easily checked that the transition rules when the partition is extended from n individuals to $n + 1$ individuals are exactly those of the (α, θ) prediction rule. So the random partition Π of $\mathbb{N}$ induced by this population process is an (α, θ) partition.

Case $0 \le \alpha \le 1$, $\theta \ge 0$. This can be described more simply by a slight modification of the rules for the above scheme. The modified scheme is then a generalization of the process described by Tavaré [414] in case $\alpha = 0$, $\theta > 0$. Instead of letting the first individual produce novel offspring at rate $\alpha+\theta$, let the first individual produce novel offspring at rate α, and let an independent Poisson migration process at rate θ bring further novel individuals into the population. Otherwise the process runs as before. Now the first novel individual follows the same rules as all other novel individuals.

If the distinction between novel and clone individuals is ignored, we just have a Yule process with immigration, where all individuals produce offspring at rate 1, and there is immigration at rate θ. If we keep track of the type of individuals, since each immigrant is novel by definition, it is clear that the partition generated by all the colors is a refinement of the partition whose classes are the progeny of the first individual, the progeny of the first immigrant, the progeny of the second immigrant, and so on. Each of these classes is created by a Yule process with rate 1, whose individuals are partitioned by coloring exactly as before in case $\theta = 0$. This structure reveals the following result:

Proposition 3.16. *Let $0 \le \alpha \le 1, \theta \ge 0$. Let a stick of length 1 be broken into lengths $P_n = \overline{X}_1 \dots \overline{X}_{n-1} X_n$ according to the GEM $(0, \theta)$ distribution, as in (3.8). Then, let each of these stick be broken further, independently of each other, according to the GEM $(\alpha, 0)$, to create a countable array of sticks of lengths*

$$P_1 X_{11}, P_1 \overline{X}_{11} X_{12}, P_1 \overline{X}_{11} \overline{X}_{12} X_{13}, \dots$$
$$P_2 X_{21}, P_2 \overline{X}_{21} X_{22}, P_2 \overline{X}_{21} \overline{X}_{22} X_{23}, \dots$$

where $X_1, X_2, \dots, X_{11}, X_{12}, \dots, X_{21}, X_{22}, \dots$ are independent, with $X_j \sim \text{beta}(1, \theta)$ for all j, and $X_{ij} \sim \text{beta}(1 - \alpha, j\alpha)$ for all i and j. Let $Q_1, Q_2, \dots$ be a size-biased random permutation of the lengths in this array. Then the Q_n are distributed according to GEM(α, θ), that is:

$$Q_n = \overline{Y}_1 \overline{Y}_2 \dots \overline{Y}_{n-1} Y_n$$

where the Y_j are independent beta$(1 - \alpha, \theta + j\alpha)$.

By arguing as in Hoppe [202], Proposition 3.16 can be restated as follows:

Proposition 3.17. *Let $0 < \alpha <, \theta \ge 0$. Let $\{A_i\}$ be a $(0, \theta)$ random partition of $[n]$. Given $\{A_i\}$, with say k blocks, let $\{A_{ij}\}$, $j = 1, \dots, k$ be independent $(\alpha, 0)$ random partitions of A_i. Then $\{A_{ij}\}$ is an (α, θ) random partition of $[n]$.*

In view of Theorem 3.2, either of these propositions follow easily from the other. A direct calculation shows that the result for finite partitions reduces to the following variant of formula (1.16) for the generating function of numbers of cycles in a random permutation of $[n]$:

$$\sum_{j=1}^{n} \theta^j \sum_{\{C_i, 1 \le i \le j\}} \Pi_{i=1}^{j} (|C_i| - 1)! \alpha^{|C_i| - 1} = \theta(\theta + \alpha) \dots (\theta + (n-1)\alpha) \quad (3.50)$$

where the second sum is over all partitions $\{C_i, 1 \le i \le j\}$ of $[n]$ into j parts, and $|C_i|$ is the number of elements of C_i. See also [368, (67)], [371, Proposition 22] for further discussion, and (5.26) for a more refined result.

Notes and comments

This section is based on an unpublished supplement to the technical report [346], written in November 1992. See also Feng and Hoppe [152] for a similar approach, with reference to an earlier model of Karlin. See Dong, Goldschmidt and Martin [113] for some recent developments.