

STATISTICS OF BOSE SAMPLES FROM DIRICHLET PROPORTIONS

Thierry HUILLET

Laboratoire de Physique Théorique et Modélisation,
CNRS-UMR 8089 et Université de Cergy-Pontoise,
2, Avenue Adolphe Chauvin, 95302, Cergy-Pontoise, France

October 18, 2006

Abstract

To fix the background and notations, we shall first briefly revisit some aspects of the following Ewens-like randomized occupancy problem: assume distinguishable particles are to be placed at random into the cells of the unit interval which was previously broken into random pieces according to the (Poisson-)Dirichlet partitioning model. Particles being distinguishable, the statistical structure of the problem can be understood within the Maxwell-Boltzmann setup.

In this note, we shall address the following sampling problem of a different nature: assume now that indistinguishable particles are to be placed at random within the cells with (Poisson-)Dirichlet distributed sizes. Then the statistical formalism to be used is the one of Bose-Einstein. We show that in the grand canonical ensemble, the Bose sampling procedure from (Poisson-)Dirichlet proportions is, to a large extent, amenable to exact analytic calculations. This concerns for example the full Bose occupancy distributions, the distribution of the number of distinct occupied fragments, the number of cells with a prescribed amount of particles. Using a grand canonical approach, a phase transition phenomenon is shown to take place provided the disorder parameter of the (Poisson-)Dirichlet partition is large enough; we describe this phase transition in some details.

Keywords and Phrases: random discrete distribution, Dirichlet, sampling, Ewens, urns, Maxwell-Boltzmann and Bose-Einstein statistics, disordered systems, phase transition.

1 Introduction

Sampling from random Dirichlet and Poisson-Dirichlet partitions has for long been a subject of recurrent interest (see Tavaré and Ewens (1997) and

references therein for historical background and applications to various fields). In one model, a number k of distinguishable particles (balls) are sequentially and uniformly thrown on the interval which has been previously partitioned at random into n pieces (fragments, bins or cells) according to the Dirichlet law with “disorder” parameter $\theta > 0$. Since particles are distinguishable, the statistical structure of the problem can nicely be understood within the Maxwell-Boltzmann setup. Many interesting questions can be (and have been) developed within this randomized occupancy framework, for instance and to cite only a few:

- What is the joint cells occupancy distribution? - What is the state of cell occupancies if sequential sampling process is stopped when some cell has received $c > 1$ particles for the first time ? (the randomized Banach match box problem if $n = 2$).
- What is the sample size (particle number) till the first visit to smallest fragment? - What is the sample size till some fragment has been visited twice for the first time? (the birthday problem).
- What is the sample size till all fragments have been visited at least once (r times)? (the coupon collector problem).
- What is the sample size between consecutive visits to distinct fragments? - What is the number of distinct fragments visited by the k -sample? - The laws of succession and Pólya urn scheme....

These problems were also naturally investigated within the extended framework of the Poisson-Dirichlet partitioning model. This model may be viewed by taking an appropriate Kingman weak limit ($n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$) of the Dirichlet partitioning model after reordering the random fragment sizes. For instance, the occupancy distribution in this context yields the celebrated Ewens Sampling Formula.

The purpose of this note is to address the following sampling problem: assume now that k indistinguishable particles are to be placed at random into the cells with Dirichlet distributed random sizes. Then the statistical formalism to be used is the one of Bose-Einstein. Since the image of a sequential throw is lost (in a way, particles are now thrown all at once), the questions relative to stopping times become meaningless, to some extent. However the questions relative to the statistics of occupancies pertain. We shall essentially be concerned here by this distributional problem and its specificities. It turns out that the canonical Bose occupancy distributions (particle number k is fixed) are difficult to handle analytically. However, we will show that in the grand canonical ensemble (where particle number is appropriately randomized), the Bose sampling procedure from Dirichlet and Poisson-Dirichlet proportions is, to a large extent, amenable to exact analytic calculations. We shall also show that when the Dirichlet disorder parameter θ is large enough (namely when $\theta > 1/(n-1)$), a phase transition occurs which is reminiscent of a Bose-Einstein like condensation phenomenon in a different random energy levels occupancy context. In the Poisson-Dirichlet situation, a similar phase transition occurs at the condition that $\gamma > 1$. No such critical phenomena arise in the classical Maxwell-Boltzmann formulation of the sampling problem.

2 Preliminaries on Ewens-sampling from Dirichlet populations

To fix the ideas, notations and analogies, we start recalling some ingredients of the classical occupancies statistics when particles are distinguishable (Maxwell-Boltzmann) before turning in the next section to the main purpose of this work: the statistical properties of Bose samples from Dirichlet populations (when particles to be placed are indistinguishable).

2.1 Dirichlet partition of the interval

Consider the following random partition into n fragments (cells or states) of the unit interval. Let $\theta > 0$ be some ‘disorder’ parameter and assume that the random fragment sizes $\mathbf{S}_n := (S_1, \dots, S_n)$ (with $\sum_{m=1}^n S_m = 1$) are distributed according to the (exchangeable) Dirichlet density function on the simplex, that is to say

$$(2.1) \quad f_{S_1, \dots, S_n}(s_1, \dots, s_n) = \frac{\Gamma(n\theta)}{\Gamma(\theta)^n} \prod_{m=1}^n s_m^{\theta-1} \cdot \delta(\sum_{m=1}^n s_m - 1).$$

Alternatively, with $(\theta)_q := \Gamma(\theta + q) / \Gamma(\theta)$ and $\Gamma(\cdot)$ the Euler-gamma function, using well-known properties of Dirichlet integrals, the law of \mathbf{S}_n can also be characterized by its joint moment function

$$(2.2) \quad \mathbf{E} \left(\prod_{m=1}^n S_m^{q_m} \right) = \frac{1}{(n\theta)_{\sum_{m=1}^n q_m}} \prod_{m=1}^n (\theta)_{q_m} \quad \text{with } q_m > -\theta.$$

If this is so, we shall say $\mathbf{S}_n \stackrel{d}{\sim} D_n(\theta)$. There are two ways to generate Dirichlet partitioning:

- (normalizing) Firstly, \mathbf{S}_n can be obtained while considering the independent and identically distributed (iid) random vector $\mathbf{X}_n := (X \stackrel{d}{=} X_1, \dots, X_n)$, satisfying $X \stackrel{d}{\sim} \text{gamma}(\theta)$ and by letting $S_m = X_m / (X_1 + \dots + X_n)$, $m = 1, \dots, n$.
- (conditioning) Secondly, with $x > 0$, consider the partitioning of the interval $[0, x]$ obtained while conditioning as follows:

$$\mathbf{S}_n(x) := (X_1, \dots, X_n \mid X_1 + \dots + X_n = x)$$

where \mathbf{X}_n is as above. Then $\mathbf{S}_n := \mathbf{S}_n(1)$ has Dirichlet distribution and the following important scaling property holds: $\mathbf{S}_n(x) \stackrel{d}{=} x\mathbf{S}_n(1)$.

If $\mathbf{S}_n \stackrel{d}{\sim} D_n(\theta)$, $S_m \stackrel{d}{=} S_n$, $m = 1, \dots, n$, independently of m and the individual fragments sizes are all identically distributed. Their common density on the interval $(0, 1)$ is a $\text{beta}(\theta, (n-1)\theta)$ density, with $\mathbf{E}(S_n) = 1/n$ and $\sigma^2(S_n) =$

$\frac{n-1}{n^2(n\theta+1)}$. In particular, $\phi(q) := \mathbb{E}(S_n^q) = (\theta)_q / (n\theta)_q$ is the moment function of typical fragment size S_n . Further,

$$nS_n \xrightarrow{d} \Gamma_{\theta,\theta} \stackrel{d}{\sim} \text{gamma}(\theta, \theta), \text{ with density } f_{\Gamma_{\theta,\theta}}(t) = \frac{\theta^\theta}{\Gamma(\theta)} t^{\theta-1} e^{-\theta t}, t > 0.$$

For each $m_1 \neq m_2 \in [n]$, as conventional wisdom suggests, (S_{m_1}, S_{m_2}) are negatively correlated with $Cov(S_{m_1}, S_{m_2}) = -\frac{\sigma^2(S_n)}{n-1} = -\frac{1}{n^2(n\theta+1)}$. When $\theta = 1$, the partition model Eqs.(2.1, 2.2) corresponds to the standard uniform random partitioning model of the interval. When $\theta \uparrow \infty$, $\mathbf{S}_n = (1/n, \dots, 1/n)$, the deterministic uniform partition. Consider next the sequence $\mathbf{S}_{(n)} := (S_{(m)}; m = 1, \dots, n)$ obtained while ranking the spacings vector \mathbf{S}_n according to descending sizes, hence with $S_{(1)} > \dots > S_{(m)} > \dots > S_{(n)}$. The $S_{(m)}$ s distribution can hardly be derived in closed form. However, one can prove that, as $n \uparrow \infty$

$$n^{(1+\theta)/\theta} S_{(n)} \xrightarrow{d} W_\theta \text{ and } n\theta \left(S_{(1)} - \frac{1}{n\theta} \log \left(n \log^{\theta-1} n \right) \right) \xrightarrow{d} G_\theta$$

where W_θ is a Weibull random variable, G_θ a Gumbel random variable such that $\mathbb{P}(W_\theta > t) = \exp\left(-\frac{t^\theta}{s_\theta}\right)$, $t > 0$ and $\mathbb{P}(G_\theta \leq t) = \exp\left(-\frac{1}{s_\theta} \exp(-t)\right)$, $t \in \mathbb{R}$, $s_\theta := \frac{\Gamma(1+\theta)}{\theta^\theta} > 0$ a scale parameter. Note that $s_\theta > 1$ if $\theta \in (0, 1)$, $s_{\theta=1} = 1$ and $s_\theta < 1$ if $\theta > 1$ and $s_\theta \rightarrow_{\theta \downarrow 0} 1$. In the random division of the interval as in Eq. (2.1), although all fragments are identically distributed with sizes of order n^{-1} , the smallest fragment size grows like $n^{-(\theta+1)/\theta}$ while the one of the largest is of order $\frac{1}{n\theta} \log \left(n \log^{\theta-1} n \right)$. The smaller disorder θ is, the larger (smaller) the largest (smallest) fragment size is: hence, the smaller θ is, the more the values of the S_m s are, with high probability, disparate. When θ is small, the size of the largest fragment $S_{(1)}$ tends to dominate the other ones. On the contrary, large values of θ correspond to situations in which the range of fragment sizes is lower: the fragment sizes look more homogeneous and, in the limit $\theta \uparrow \infty$, distribution Eq. (2.1) concentrates on its centre $(\frac{1}{n}, \dots, \frac{1}{n})$. For large disorder θ , the diversity of the partition is small.

Although \mathbf{S}_n has a degenerate weak limit when $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$, this limit is worth being considered (see Kingman (1975), (1978) and (1993)). Indeed, in this $*$ -limit, $\mathbf{S}_{(n)} \rightarrow_* \mathbf{S}_{(\infty)} \stackrel{d}{\sim} PD(\gamma)$ which is the Poisson-Dirichlet distribution with parameter γ ; see Kingman (1993) and Tavaré-Ewens (1997). We shall also call γ the ‘disorder’ parameter since PD partitions with large γ approach the ‘uniform distribution’ on the infinite-dimensional simplex whereas for small values of γ , the largest fragment is the dominant one.

2.2 Maxwell-Boltzmann approach to sampling problems from Dirichlet partition

Before discussing the specific statistical features of Bose samples drawn from

Dirichlet populations, we first revisit the Ewens approach to sampling formulae which is akin to a Maxwell-Boltzmann sampling procedure.

• **Sampled fragment size when sample size is 1:**

Assume first sample size is $k = 1$ and suppose the sampled tagged fragment is the one hit by a uniform random throw of a particle on the interval. Under our hypothesis, this particle will launch on fragment number m with conditional (given \mathbf{S}_n) probability S_m . The corresponding fragment size attached to this single particle, say \mathcal{S}_n , has conditional law given by

$$\mathbb{P}_{\mathbf{S}_n}(\mathcal{S}_n = S_m) = S_m, \quad m = 1, \dots, n.$$

(Here and throughout, the subscript \mathbf{S}_n in $\mathbb{P}_{\mathbf{S}_n}$ (or $\mathbb{E}_{\mathbf{S}_n}$) will denote conditional probability (or expectation) given \mathbf{S}_n). Let $\mathbb{E}_{\mathbf{S}_n}(\mathcal{S}_n^q)$ be its moment function and put $\phi_{\mathbf{S}_n}(q) := \sum_{m=1}^n S_m^q$. Then $\mathbb{E}_{\mathbf{S}_n}(\mathcal{S}_n^q) = \phi_{\mathbf{S}_n}(q+1)/\phi_{\mathbf{S}_n}(1)$. Averaging over \mathbf{S}_n ,

$$\mathbb{E}(\mathcal{S}_n^q) := \mathbb{E}\mathbb{E}_{\mathbf{S}_n}(\mathcal{S}_n^q) = n\mathbb{E}(S_n^{q+1}) =: n\phi(q+1)$$

characterizes the distribution of the fragment size of this single particle. In particular,

$$\mathbb{E}(\mathcal{S}_n) = n\phi(2) = \frac{\theta + 1}{n\theta + 1} > \frac{1}{n}.$$

In this size-biased picking procedure $\mathbf{S}_n \rightarrow \mathcal{S}_n$, states with large size are clearly favored. One therefore expects (and this indeed true) that \mathcal{S}_n is stochastically larger than the typical fragment size S_n from \mathbf{S}_n .

• **Maxwell-Boltzmann-Ewens sampling (sample size is $k > 1$):**

The full Maxwell-Boltzmann sampling version of the randomized occupancy problem proceeds as follows: let (U_1, \dots, U_k) be k iid uniform throws on $[0, 1]$ partitioned by \mathbf{S}_n . Let $(B_{n,k}(1), \dots, B_{n,k}(n))$ be an integral-valued random vector which counts the number of visits of particles thus thrown to the different fragments in a k -sample in the following sense: if M_l is the random state label which the l -th trial hits, then $B_{n,k}(m) := \sum_{l=1}^k \mathbf{I}(M_l = m)$, $m = 1, \dots, n$ (where $\mathbf{I}(A)$ stands for the set indicator of the event A). As stated above $\mathbb{P}_{\mathbf{S}_n}(M_l = m) = S_m$ and state m is chosen proportionally to its size S_m . With $(b_1, \dots, b_n) \in \mathbb{N}_0^n$, (where $\mathbb{N}_0 := \{0, 1, 2, \dots\}$) satisfying $\sum_{m=1}^n b_m = k$, $(B_{n,k}(m) = b_m; m = 1, \dots, n)$ clearly follows the conditional multinomial distribution with randomized probabilities \mathbf{S}_n :

$$(2.3) \quad \mathbb{P}_{\mathbf{S}_n}(B_{n,k}(m) = b_m; m = 1, \dots, n) = \frac{k!}{\prod_{m=1}^n b_m!} \prod_{m=1}^n S_m^{b_m}.$$

Proceeding in this way to fill up sequentially the states \mathbf{S}_n , particles are clearly assumed distinguishable. Indeed, in the above expression of the probability, the multinomial factor $\frac{k!}{\prod_{m=1}^n b_m!}$ represents the number of ways to distribute

k labelled particles into n distinguishable boxes with respective occupancies (b_1, \dots, b_n) .

We note from Eq. (2.3) that, given \mathbf{S}_n ,

$$(B_{n,k}(m); m = 1, \dots, n) \stackrel{d}{=} (\xi_1, \dots, \xi_n \mid \zeta_n = k)$$

where (ξ_1, \dots, ξ_n) are mutually independent on \mathbb{N}_0^n with sum $\zeta_n := \sum_1^n \xi_m$ and

$$\mathbb{P}_{\mathbf{S}_n}(\xi_m = b_m) = \frac{S_m^{b_m} e^{-S_m}}{b_m!}, \quad b_m \in \mathbb{N}_0,$$

which are Poisson distributions with random means $S_m \stackrel{d}{\sim} \text{beta}(\theta, (n-1)\theta)$, for each $m = 1, \dots, n$.

Averaging over \mathbf{S}_n gives the Maxwell-Boltzmann distribution

$$\begin{aligned} \mathbb{P}(B_{n,k}(m) = b_m; m = 1, \dots, n) &= \mathbb{E} \mathbb{P}_{\mathbf{S}_n}(B_{n,k}(m) = b_m; m = 1, \dots, n) \\ &= \frac{k!}{\prod_{m=1}^n b_m!} \frac{1}{(n\theta)_k} \prod_{m=1}^n (\theta)_{b_m}, \end{aligned}$$

also known in the statistical context as the Dirichlet multinomial distribution.

Examples: When $\theta \uparrow \infty$, the partition \mathbf{S}_n reduces to $(\frac{1}{n}, \dots, \frac{1}{n})$ which is not random. It follows from Eq. (2.3) that

$$\mathbb{P}(B_{n,k}(m) = b_m; m = 1, \dots, n) = \frac{k!}{\prod_{m=1}^n b_m!} n^{-k}.$$

Unless otherwise specified, we shall assume in the sequel that $\theta < \infty$ which means that sampling really is from Dirichlet probabilities which are indeed random.

When $\theta = 1$, the partition \mathbf{S}_n reduces to the random uniform partition. It follows from Eq. (2.3) that

$$\mathbb{P}(B_{n,k}(m) = b_m; m = 1, \dots, n) = \frac{k!}{(n)_k} = \frac{1}{\binom{n+k-1}{k}},$$

the uniform distribution on the set $\{b_m \in \mathbb{N}_0, m = 1, \dots, n : \sum_1^n b_m = k\}$. \diamond

We also recall the following almost sure convergence which follows from conditional strong law of large numbers:

Lemma 1 *It holds that*

$$(2.4) \quad (B_{n,k}(m); m = 1, \dots, n) / k \rightarrow \mathbf{S}_n \text{ as } k \uparrow \infty,$$

in distribution and almost surely.

In a Maxwell-Boltzmann approach to the sampling problem from Dirichlet proportions, the proportions of sampled fragments when sample size is large is balanced (no concentration phenomenon within a specific fragment) and the Dirichlet partition is recovered in the limit.

• **Sampling distribution as a random allocation scheme:**

Let $(\eta_m)_{m \geq 1}$ be an iid sequence of negative-binomial (or Pölya) distributed random variables on \mathbb{N}_0 with mean 1 and distribution

$$(2.5) \quad \mathbb{P}(\eta_1 = b_1) = \frac{(\theta)_{b_1}}{b_1!} x^{b_1} \bar{x}^\theta, \quad b_1 = 0, 1, \dots$$

with $x = (1 + \theta)^{-1}$, $\bar{x} := 1 - x$. The generating function of η_1 is

$$(2.6) \quad \mathbb{E}(u_1^{\eta_1}) = \left(\frac{1 - xu_1}{\bar{x}} \right)^{-\theta}, \quad 0 \leq u_1 < 1/x.$$

The random variable η_1 has mean 1 and variance $(1 + \frac{1}{\theta})$, exceeding 1 for finite θ (it is over-dispersing compared to a mean 1 Poisson distribution). Its distribution can be obtained while randomizing the intensity a Poisson distribution by a gamma(θ, θ) – distributed independent random variable (with mean 1 and variance $1/\theta$); it is a gamma-Poisson mixture. Let $\mu_n := \sum_{m=1}^n \eta_m$, $n \geq 1$, be the partial sum sequence of $(\eta_m)_{m \geq 1}$ with $\mu_0 := 0$. Then, one can check that

$$\mathbb{P}(B_{n,k}(m) = b_m; m = 1, \dots, n) = \mathbb{P}(\eta_1 = b_1, \dots, \eta_n = b_n \mid \mu_n = k).$$

The unconditional multinomial-Dirichlet distribution is in the class of random allocation schemes as the ones obtained by conditioning a random walk by its terminal value (see Kolchin (1986), Johnson and Kotz (1977) for instance).

• **The number of distinct visited fragments:**

Let now $P_{n,k} := \sum_{m=1}^n \mathbf{I}(B_{n,k}(m) > 0)$ count the number of distinct fragments which have been visited in the k -sampling process. With $1 \leq m_1 < \dots < m_p \leq n$ a subset of p labels from $\{1, \dots, n\}$, with $b_q \in \mathbb{N} := \{1, 2, \dots\}$, $q = 1, \dots, p$, we clearly have

$$(2.7) \quad \begin{aligned} \mathbb{P}_{\mathbf{S}_n}(M_1, \dots, M_k \in \{m_1, \dots, m_p\}; B_{n,k}(m_1) = b_1, \dots, B_{n,k}(m_p) = b_p; P_{n,k} = p) \\ = \frac{k!}{\prod_{q=1}^p b_q!} \prod_{q=1}^p S_{m_q}^{b_q}. \end{aligned}$$

Define next $B_{n,k}(q) > 0$, $q = 1, \dots, p$ to be the numbers of type- q fragments where the $P_{n,k} = p$ fragments observed were labelled in an arbitrary way (independently of the sampling mechanism). Averaging the last formula over \mathbf{S}_n , summing over the $\binom{n}{p}$ sequences of hit labels and making use of its exchangeability, we easily obtain (see Huillet (2005), for details)

Theorem 2 (i) With $b_q \in \mathbb{N} : \sum_{q=1}^p b_q = k$, we have

$$(2.8) \quad \begin{aligned} \mathbb{P}(B_{n,k}(1) = b_1, \dots, B_{n,k}(p) = b_p; P_{n,k} = p) \\ = \binom{n}{p} \frac{k!}{\prod_{q=1}^p b_q!} \frac{1}{(n\theta)_k} \prod_{q=1}^p (\theta)_{b_q}. \end{aligned}$$

(ii) With $(\theta)_\bullet := (\theta)_1, (\theta)_2, \dots$ and

$$\mathfrak{B}_{k,p}((\theta)_\bullet) := \frac{k!}{p!} \sum_{b_q \in \mathbb{N} : \sum_{q=1}^p b_q = k} \prod_{q=1}^p \frac{(\theta)_{b_q}}{b_q!}$$

Bell polynomials in the indeterminates $(\theta)_\bullet$, it holds that,

$$(2.9) \quad \mathbb{P}(P_{n,k} = p) = \frac{n!}{(n-p)!} \frac{1}{(n\theta)_k} \mathfrak{B}_{k,p}((\theta)_\bullet)$$

where $p = 1, \dots, n \wedge k$.

Concerning the distribution of $P_{n,k}$, we also have the conditional transition probabilities

$$\begin{aligned} \mathbb{P}(P_{n,k+1} = p+1 \mid P_{n,k} = p) &= \frac{(n-p)\theta}{n\theta+k} \\ \mathbb{P}(P_{n,k+1} = p \mid P_{n,k} = p) &= \frac{\sum_{r=1}^p (\theta + b_r)}{n\theta+k} = \frac{p\theta+k}{n\theta+k}. \end{aligned}$$

Therefore, the following recurrence holds

$$\mathbb{P}(P_{n,k+1} = p) = \frac{(n-p+1)\theta}{n\theta+k} \mathbb{P}(P_{n,k} = p-1) + \frac{p\theta+k}{n\theta+k} \mathbb{P}(P_{n,k} = p).$$

As a simple application of the inclusion-exclusion principle, we shall finally recall a straightforward representation of the probability $\mathbb{P}(P_{n,k} = p)$ under the form of an alternate sum (see for example Keener et al (1987), pages 1471-1472). This is an explicit expression of this probability in contrast with Eq. (2.9) which, as just shown, is recursive.

Corollary 3 (i) With $\langle \theta \rangle_{n,k;m} := \frac{((n-m)\theta)_k}{(n\theta)_k}$, $m = 0, \dots, n-1$, the generating function of $P_{n,k}$ reads

$$(2.10) \quad \mathbb{E}(u^{P_{n,k}}) = \sum_{m=0}^{n-1} \binom{n}{m} u^{n-m} (1-u)^m \langle \theta \rangle_{n,k;m}.$$

In particular, the mean and variance are given by

$$(2.11) \quad \mathbb{E}(P_{n,k}) = n \left(1 - \langle \theta \rangle_{n,k;1} \right) = n \left(1 - \frac{((n-1)\theta)_k}{(n\theta)_k} \right),$$

$$\sigma^2(P_{n,k}) = n \left(\langle \theta \rangle_{n,k;1} + (n-1) \langle \theta \rangle_{n,k;2} - n \langle \theta \rangle_{n,k;1}^2 \right).$$

(ii)

$$(2.12) \quad \mathbb{P}(P_{n,k} = p) = \sum_{q=1}^p (-1)^{p-q} \binom{n}{p} \binom{p}{q} \langle \theta \rangle_{n,k;n-q}.$$

Remark: When $\theta = 1$, one can check that $\mathbb{E}(P_{n,k}) = (nk)/(n+k)$, which is half the geometric average of n and k . \diamond

• **Kingman limit:**

With $s_{k,p} := \mathfrak{B}_{k,p}((\bullet - 1)!)$ the absolute value of the first kind Stirling numbers, taking the $*$ -limit $n \uparrow \infty$, $\theta \downarrow 0$, $n\theta = \gamma > 0$, using $\mathfrak{B}_{k,p}((\theta)_\bullet) \sim_* \theta^p \mathfrak{B}_{k,p}((\bullet - 1)!)$, we easily get

$$(2.13) \quad \mathbb{P}(P_{n,k} = p) \rightarrow_* \mathbb{P}_*(P_k = p) = \frac{\gamma^p s_{k,p}}{(\gamma)_k}, \quad p = 1, \dots, k$$

and

$$(2.14) \quad \mathbb{P}_*(B_k(1) = b_1, \dots, B_k(p) = b_p \mid P_k = p) \rightarrow_* \frac{k!}{p!} \frac{1}{s_{k,p} \prod_{q=1}^p b_q}.$$

We note that the law of P_k in this case is in the class of exponential families. Further, the generating function of P_k takes the simple form

$$(2.15) \quad \mathbb{E}_*[u^{P_k}] = \frac{(\gamma u)_k}{(\gamma)_k}.$$

In particular, the mean and variance are given by

$$\begin{aligned} \mathbb{E}_*(P_k) &= \sum_{l=0}^{k-1} \frac{\gamma}{\gamma + l}, \\ \sigma_*^2(P_k) &= \sum_{l=0}^{k-1} \frac{\gamma l}{(\gamma + l)^2}. \end{aligned}$$

In this context, we recall the important result of Korwar and Hollander (1973)

$$(2.16) \quad \frac{P_k}{\log k} \rightarrow \gamma, \quad k \uparrow \infty, \text{ almost surely.}$$

• **The second Ewens formula for Dirichlet populations:**

Let now $A_{n,k}(i)$, $i \in \{0, \dots, k\}$ count the number of fragments in the k -sample with i representatives, that is

$$(2.17) \quad A_{n,k}(i) = \# \{m \in \{1, \dots, n\} : B_{n,k}(m) = i\} = \sum_{m=1}^n \mathbf{I}(B_{n,k}(m) = i).$$

Then $\sum_{i=0}^k A_{n,k}(i) = n$, $\sum_{i=1}^k A_{n,k}(i) = p$ is the number of fragments visited by the k -sample and $A_{n,k}(0)$ the number of unvisited ones. Note that $\sum_{i=1}^k i A_{n,k}(i) = k$ is the sample size.

The vector $(A_{n,k}(1), \dots, A_{n,k}(k))$ is called the fragments vector count or the species vector count in biology, see Ewens (1990). In Sibuya (1993), it is called the size-index vector and in Good (1968), the frequency of frequencies.

In this case (see Huillet (2005) for computational details), with $\{n\}_p := n(n-1) \dots (n-p+1)$ the order p falling factorial of n , we have

Theorem 4 For any $a_i \geq 0$, $i = 1, \dots, k$ satisfying $\sum_{i=1}^k i a_i = k$ and $\sum_{i=1}^k a_i = p$, we have

$$(2.18) \quad \begin{aligned} \mathbb{P}(A_{n,k}(1) = a_1, \dots, A_{n,k}(k) = a_k; P_{n,k} = p) \\ = \{n\}_p \frac{k!}{\prod_{i=1}^k i!^{a_i} a_i!} \frac{1}{(n\theta)_k} \prod_{i=1}^k (\theta)_i^{a_i}. \end{aligned}$$

Considering the Kingman limit $n \uparrow \infty$, $\theta \downarrow 0$ while $n\theta = \gamma > 0$, using $(\theta)_i \sim_{\theta \downarrow 0} \theta(i-1)!$ and $\{n\}_p \sim_{n \uparrow \infty} n^p$, we recover the celebrated Ewens Sampling Formula (1972):

Corollary 5 In the Kingman limit, the probability displayed in (2.18) converges to

$$(2.19) \quad \mathbb{P}_*(A_k(1) = a_1, \dots, A_k(k) = a_k; P_k = p) = \frac{k! \gamma^p}{(\gamma)_k \prod_{i=1}^k i^{a_i} a_i!}.$$

• **Ewens grand-canonical sampling formula:**

Although Poissonization of sample size in occupancy problems was addressed in Cesaroli (1983), this point has not been discussed in the specific random context of Dirichlet partitioning, to the best of the author knowledge. As it will prove essential when considering Bose samples, we shall briefly introduce this topic.

The problem here is to randomize sample size k . Let $z > 0$ be some ‘‘activity’’ parameter. Let $K_{n,z}$ be the random sample size and assume it has Poisson distribution with mean $\kappa := z > 0$.

Multiplying the probability displayed in Eq. (2.3) by $\frac{z^k}{k!} e^{-z}$, with $b_m \in \mathbb{N}_0$, $m = 1, \dots, n$, we get

$$\mathbb{P}_{\mathbf{S}_n}(B_{n,z}(m) = b_m; m = 1, \dots, n) = \prod_{m=1}^n \frac{e^{-z S_m} (z S_m)^{b_m}}{b_m!}$$

where the random occupancies B are now indexed by z instead of k . In this formulation, the annoying restriction that $\sum_1^m b_m = k$ has been lifted, which is the usual trick used in the grand-canonical ensemble of equilibrium statistical mechanics. Given \mathbf{S}_n , the grand canonical distribution of $B_{n,z}(m); m = 1, \dots, n$ turns out to be the product of n independent Poisson random variables with intensities zS_m , $m = 1, \dots, n$. Averaging the last formula over \mathbf{S}_n and making use of its exchangeability, we get

$$\mathbb{P}(B_{n,z}(m) = b_m; m = 1, \dots, n) = \frac{e^{-z}}{(n\theta)_{b_1+\dots+b_n}} \prod_{m=1}^n \frac{z^{b_m} (\theta)_{b_m}}{b_m!}.$$

The unconditional distribution of $B_{n,z}(m) = b_m; m = 1, \dots, n$ is still exchangeable but independence is lost.

Multiplying now the probability displayed in Eq. (2.7) by $\frac{z^k}{k!} e^{-z}$, with $b_q \in \mathbb{N}$, $q = 1, \dots, p$, we get

$$\begin{aligned} \mathbb{P}_{\mathbf{S}_n}(M_1, \dots, M_k \in \{m_1, \dots, m_p\}; B_{n,z}(m_1) = b_1, \dots, B_{n,z}(m_p) = b_p; P_{n,z} = p) \\ = e^{-z} \prod_{q=1}^p \frac{(zS_{m_q})^{b_q}}{b_q!}. \end{aligned}$$

Summing over $b_q \in \mathbb{N}$

$$\begin{aligned} \mathbb{P}_{\mathbf{S}_n}(M_1, \dots, M_k \in \{m_1, \dots, m_p\}; P_{n,z} = p) \\ = e^{-z} \prod_{q=1}^p (e^{zS_{m_q}} - 1). \end{aligned}$$

Averaging the last formula over \mathbf{S}_n and making use of its exchangeability, we get the unconditional grand-canonical probability for the number $P_{n,z}$ of distinct visited fragments

$$\mathbb{P}(P_{n,z} = p) = \binom{n}{p} e^{-z} \mathbb{E} \left[\prod_{q=1}^p (e^{zS_q} - 1) \right].$$

Here $p \in \{0, 1, \dots, n\}$ with the convention that $\mathbb{P}(P_{n,z} = 0) = e^{-z}$ which, as required, is the probability that there is no particle in the system: the event $K_{n,z} = 0$. Clearly, for each $p \in \{1, \dots, n\}$, one can check that

$$\mathbb{P}(P_{n,z} = p) = \sum_{k \geq p} \frac{z^k e^{-z}}{k!} \mathbb{P}(P_{n,k} = p)$$

where $\mathbb{P}(P_{n,k} = p)$ is the canonical distribution given that the sample size is k , displayed above in Eq. (2.9) or Eq. (2.12).

3 Bose samples from Dirichlet populations

We now come to the announced Bose-Einstein version of the sampling process from Dirichlet proportions. In this problem, particles are assumed to be indistinguishable.

3.1 The statistical structure of the Bose model

Let there now be k indistinguishable particles to place at random on the states \mathbf{S}_n . Conditionally on \mathbf{S}_n (quenched disorder), let $\mathcal{B}_{n,k}(m)$ denote the occupancy of state m with Bose equilibrium collective law given by

$$(3.1) \quad \mathbb{P}_{\mathbf{S}_n}(\mathcal{B}_{n,k}(m) = b_m; m = 1, \dots, n) = \frac{1}{Z_{k,\mathbf{S}_n}(\theta)} \prod_{m=1}^n S_m^{b_m}.$$

With $[z^k] f(z)$ the coefficient of z^k in the power-series expansion of $f(z)$ the normalizing partition function term reads

$$(3.2) \quad Z_{k,\mathbf{S}_n}(\theta) = \sum_{b'_1 + \dots + b'_n = k} \prod_{m=1}^n S_m^{b'_m} = [z^k] \prod_{m=1}^n (1 - z S_m)^{-1}.$$

The distribution thus defined favors configurations with minimal (interaction free) “energy”: $H_{n,k}(\mathbf{S}_n) := -\sum_{m=1}^n b_m \log S_m$.

In Eq. (3.1), $b_m \in \mathbb{N}_0$, $m = 1, \dots, n$, with no restriction but $b_1 + \dots + b_n = k$. Imposing the additional condition that $b_m \in \{0, 1\}$, $m = 1, \dots, n$ (the Pauli exclusion principle), would lead to a Fermi-Dirac occupancy problem which we shall not further develop specifically.

Example: When $\theta \uparrow \infty$, the limiting partition \mathbf{S}_n reduces to $(\frac{1}{n}, \dots, \frac{1}{n})$ which is not random. It follows from the above equation that

$$\mathbb{P}(\mathcal{B}_{n,k}(m) = b_m; m = 1, \dots, n) = \frac{1}{\binom{n+k-1}{k}},$$

the uniform distribution on the set $\{b_m \in \mathbb{N}_0, m = 1, \dots, n : \sum_1^n b_m = k\}$. This distribution is known as the Bose-Einstein distribution (see Feller (1971) and Holst (1985)). Curiously, it coincides with the Maxwell-Boltzmann sampling formula from the random uniform partition \mathbf{S}_n (the Dirichlet partition obtained when $\theta = 1$). \diamond

Thanks to the representation of \mathbf{S}_n in terms of ratios of iid gamma distributed random variables \mathbf{X}_n , this is also

$$\mathbb{P}_{\mathbf{S}_n}(\mathcal{B}_{n,k}(m) = b_m; m = 1, \dots, n) = \mathbb{P}_{\mathbf{X}_n}(\mathcal{B}_{n,k}(m) = b_m; m = 1, \dots, n)$$

where

$$\begin{aligned}\mathbb{P}_{\mathbf{X}_n}(\mathcal{B}_{n,k}(m) = b_m; m = 1, \dots, n) &= \frac{1}{Z_{k, \mathbf{X}_n}(\theta)} \prod_{m=1}^n X_m^{b_m}, \\ Z_{k, \mathbf{X}_n}(\theta) &= \sum_{b'_1 + \dots + b'_n = k} \prod_{m=1}^n X_m^{b'_m}.\end{aligned}$$

Given there are k particles, averaging over disorder \mathbf{S}_n (or \mathbf{X}_n), the Bose unconditional occupancy probability now is

$$\begin{aligned}\mathbb{P}(\mathcal{B}_{n,k}(m) = b_m; m = 1, \dots, n) &= \mathbb{E}\mathbb{P}_{\mathbf{S}_n}(\mathcal{B}_{n,k}(m) = b_m; m = 1, \dots, n) \\ &= \mathbb{E}\mathbb{P}_{\mathbf{X}_n}(\mathcal{B}_{n,k}(m) = b_m; m = 1, \dots, n).\end{aligned}$$

As a symmetric function of the b_m s, this distribution is exchangeable. In particular, $\mathbb{E}(\mathcal{B}_{n,k}(m)) = k/n$, $m = 1, \dots, n$. Even though, by using this ‘ratio trick’, the average to perform can be over the simpler sequence of iid random variables \mathbf{X}_n (rather than over \mathbf{S}_n on the simplex), these canonical occupancy distributions conditioned on sample size being equal to k remain clearly hard to evaluate in practice.

• **One-dimensional distribution:**

We here briefly give the occupancy law of any cell. With $b_1 \in \{0, \dots, k\}$ and $\mathbf{X}_{n \setminus 1} := (X_2, \dots, X_n)$

$$\begin{aligned}\mathbb{P}_{\mathbf{X}_n}(\mathcal{B}_{n,k}(1) = b_1) &= \frac{X_1^{b_1}}{Z_{k, \mathbf{X}_n}(\theta)} \sum_{b'_2 + \dots + b'_n = k - b_1} \prod_{m=2}^n X_m^{b'_m}, \\ &= \frac{X_1^{b_1} Z_{k-b_1, \mathbf{X}_{n \setminus 1}}(\theta)}{Z_{k, \mathbf{X}_n}(\theta)}\end{aligned}$$

and

$$\mathbb{P}(\mathcal{B}_{n,k}(1) = b_1) = \mathbb{E} \left[\frac{X_1^{b_1} Z_{k-b_1, \mathbf{X}_{n \setminus 1}}(\theta)}{Z_{k, \mathbf{X}_n}(\theta)} \right]$$

is the one-dimensional marginal of $(\mathcal{B}_{n,k}(m); m = 1, \dots, n)$.

• **Random allocation scheme representation of Bose distribution:**

We first observe from Eqs. (3.1, 3.2) that, given \mathbf{S}_n :

$$(\mathcal{B}_{n,k}(m) = b_m; m = 1, \dots, n) \stackrel{d}{=} (\xi_1, \dots, \xi_n \mid \zeta_n = k)$$

where (ξ_1, \dots, ξ_n) are mutually independent on \mathbb{N}_0^n with sum $\zeta_n := \sum_{m=1}^n \xi_m$ and

$$\mathbb{P}_{\mathbf{S}_n}(\xi_m = b_m) = S_m^{b_m} (1 - S_m), \quad b_m \in \mathbb{N}_0,$$

geometric distributions with random success probabilities $S_m \stackrel{d}{\sim} \text{beta}(\theta, (n-1)\theta)$, for each $m = 1, \dots, n$. Such a representation of the occupancies is called a random allocation scheme property in Kolchin (1986).

• **A concentration phenomenon:**

Let us now show that, when the number of fragments is fixed, the proportions of particles tend to concentrate on ground state (which is the fragment with largest size) when the number of particles increases. This result should be compared with the one displayed in Lemma 1 when sampling uses a Maxwell-Boltzmann procedure.

Proposition 6 Let $\mathbf{S}_n \stackrel{d}{\sim} D_n(\theta)$ with $0 < \theta < \infty$. For each $m \in \{1, \dots, n\}$, let $\mathbf{S}_{n \setminus m} := (S_1, \dots, S_{m-1}, S_{m+1}, \dots, S_n)$. With n fixed, as the number of particles grows, conditionally given \mathbf{S}_n , we have

$$(3.3) \quad \left(\frac{\mathcal{B}_{n,k}(m)}{k}; m = 1, \dots, n \right) \xrightarrow{k \uparrow \infty} (P_{m,n} := \mathbf{I}(S_m > \mathbf{S}_{n \setminus m}); m = 1, \dots, n)$$

in distribution.

Proof: Let us first consider the ordered version $\mathbf{S}_{(n)}$ of the energy sequence \mathbf{S}_n , namely: $\mathbf{S}_{(n)} := (S_{(1)}, \dots, S_{(n)})$ with $S_{(1)} > \dots > S_{(n)}$. Developing the product partition function

$$\prod_{m=1}^n (1 - zS_m)^{-1} = \prod_{m=1}^n (1 - zS_{(m)})^{-1}$$

appearing in Eq. (3.2) into a sum of n rational fractions, extracting its coefficient of z^k , we easily get (after obvious identification of the coefficients)

$$Z_{k, \mathbf{S}_n}(\theta) = Z_{k, \mathbf{S}_{(n)}}(\theta) = \sum_{m=1}^n C_{(m)} S_{(m)}^k \text{ where } C_{(m)} := \prod_{l \neq m} \left(1 - \frac{S_{(l)}}{S_{(m)}} \right)^{-1}.$$

Isolate the ground state term and factorize $S_{(1)}$. Then

$$Z_{k, \mathbf{S}_{(n)}}(\theta) = S_{(1)}^k \left(C_{(1)} + \sum_{m=2}^n C_{(m)} \tilde{S}_{(m)}^k \right)$$

where $\tilde{S}_{(m)} := S_{(m)}/S_{(1)}$, $m = 1, \dots, n$. With $b_1 + \dots + b_n = k$, we want to compute the law of the occupancies $\mathcal{B}_{(n),k}(m)$ of $S_{(m)}$ which is

$$\mathbb{P}_{\mathbf{S}_{(n)}}(\mathcal{B}_{(n),k}(m) = b_m; m = 1, \dots, n) = \frac{1}{Z_{k, \mathbf{S}_{(n)}}(\theta)} \prod_{m=1}^n S_{(m)}^{b_m}.$$

Since $b_1 = k - (b_2 + \dots + b_n)$, using the expression of $Z_{k, \mathbf{S}_{(n)}}(\theta)$, the occupancy distribution of all states but ground state reads

$$\begin{aligned} \mathbb{P}_{\mathbf{S}_{(n)}}(\mathcal{B}_{(n),k}(m) = b_m; m = 2, \dots, n) &= \frac{\prod_{m=2}^n \tilde{S}_{(m)}^{b_m}}{C_{(1)} \left(1 + \sum_{m=2}^n \frac{C_{(m)}}{C_{(1)}} \tilde{S}_{(m)}^k\right)} \\ &= \frac{\prod_{m=2}^n \tilde{S}_{(m)}^{b_m} (1 - \tilde{S}_{(m)})}{1 + \sum_{m=2}^n \frac{C_{(m)}}{C_{(1)}} \tilde{S}_{(m)}^k}, \text{ using } C_{(1)} = \prod_{m \neq 1} (1 - \tilde{S}_{(m)})^{-1}. \end{aligned}$$

Developing the denominator in power series, we finally obtain

$$(3.4) \quad \mathbb{P}_{\mathbf{S}_{(n)}}(\mathcal{B}_{(n),k}(m) = b_m; m = 2, \dots, n) = (1 - \varepsilon(k)) \prod_{m=2}^n \left\{ \tilde{S}_{(m)}^{b_m} (1 - \tilde{S}_{(m)}) \right\}.$$

Since $\tilde{S}_{(n)} < \dots < \tilde{S}_{(3)} < \tilde{S}_{(2)} < 1$, the corrective term $\varepsilon(k) := \tilde{S}_{(2)}^k C_{(2)}/C_{(1)} < 0$ is dominant to the second order. It goes to 0 exponentially fast with k becoming large. When k is large, a good approximation of occupancies of all ordered states but ground state therefore is a product of geometrically distributed random variables with normalized success probabilities $\tilde{S}_{(m)}$.

Suppose $b_m = \lfloor kx_m \rfloor$ for some fixed $x_m \in (0, 1]$; $m = 2, \dots, n$. In this case, the probability displayed in Eq. (3.4) goes to 0 when k goes to ∞ : in other words, the probabilities of $\mathcal{B}_{(n),k}(m)/k$; $m = 2, \dots, n$ all concentrate at 0 and therefore all the probability mass goes to ground state ($m = 1$). This is the content of statement displayed in Eq. (3.3) where by the event $S_m > \mathbf{S}_{n \setminus m}$ it is meant that S_m is larger than all entries constituting the random vector $\mathbf{S}_{n \setminus m}$. Note that almost surely $\sum_{m=1}^n P_{m,n} = 1$ and that for each m , $\mathbb{E}(P_{m,n}) = 1/n$ and $\sigma^2(P_{m,n}) = (1 - 1/n)/n$. \square

• **Gibbs randomization of sample size (variable particle number):**

As it can be guessed from above, the canonical conditional distributions given sample size is k are difficult to evaluate in general (except for $k = 1$). To circumvent this drawback, we shall again assume that the number of particles is variable and so randomize sample size. In this way, we shall obtain a tractable grand-canonical version of Bose sample from Dirichlet proportions.

Let $\alpha > 0$ stand for fugacity and let $z = e^{-\alpha} \in (0, z_c := 1)$ be the activity parameter. Assume the number of particles $\mathcal{K}_{n,z}$ is now random with law given by the Gibbs model

$$\mathbb{P}_{\mathbf{S}_n}(\mathcal{K}_{n,z} = k) = \frac{z^k Z_{k, \mathbf{S}_n}(\theta)}{Z_{z, \mathbf{S}_n}(\theta)}$$

where the grand canonical partition now reads

$$Z_{z, \mathbf{S}_n}(\theta) = \sum_{k \geq 0} z^k Z_{k, \mathbf{S}_n}(\theta) = \prod_{m=1}^n \frac{1}{1 - zS_m}.$$

Alternatively, with $u \in [0, 1]$, we clearly have

$$\begin{aligned} \mathbb{E}_{\mathbf{S}_n}(u^{\mathcal{K}_{n,z}}) &= \frac{Z_{uz, \mathbf{S}_n}(\theta)}{Z_{z, \mathbf{S}_n}(\theta)} = \prod_{m=1}^n \frac{1 - zS_m}{1 - uzS_m} \\ \mathbb{E}(u^{\mathcal{K}_{n,z}}) &= \mathbb{E}\left(\prod_{m=1}^n \frac{1 - zS_m}{1 - uzS_m}\right). \end{aligned}$$

Under this form, this shows that, given \mathbf{S}_n , $\mathcal{K}_{n,z}$ is the sum of n independent geometric random variables with respective success probabilities zS_m , $m = 1, \dots, n$.

To each value $z \in (0, 1)$ there is a unique corresponding value of $\kappa := \mathbb{E}(\mathcal{K}_{n,z})$ through:

$$\begin{aligned} \mathbb{E}_{\mathbf{S}_n}(\mathcal{K}_{n,z}) &= -\partial_\alpha \log Z_{e^{-\alpha}, \mathbf{S}_n}(\theta) = \sum_{m=1}^n \frac{zS_m}{1 - zS_m} \\ \mathbb{E}(\mathcal{K}_{n,z}) &= : \kappa = \mathbb{E}\mathbb{E}_{\mathbf{S}_n}(\mathcal{K}_{n,z}) = n\mathbb{E}\left(\frac{zS_n}{1 - zS_n}\right). \end{aligned}$$

As will be checked below, κ is an increasing function of $z \in (0, 1)$, possibly diverging when $z \uparrow z_c$, depending on the range of the variables θ and n parameterizing the Dirichlet model (see below where condition $(n-1)\theta \leq 1$ versus $(n-1)\theta > 1$ appears that separates two phases depending on whether $\kappa \uparrow \infty$ or not when $z \uparrow 1^-$).

Indexing now cell occupancies by z rather than k , with $b_m \in \mathbb{N}_0$; $m = 1, \dots, n$, the joint occupancies probability takes the product form

$$\mathbb{P}_{\mathbf{S}_n}(\mathcal{B}_{n,z}(m) = b_m; m = 1, \dots, n) = \prod_{m=1}^n \left\{ (zS_m)^{b_m} (1 - zS_m) \right\}$$

where each $\mathcal{B}_{n,z}(m)$ now follows a geometric distribution with success probability zS_m .

In other words, given \mathbf{S}_n , the grand canonical distribution of $\mathcal{B}_{n,z}(m)$; $m = 1, \dots, n$ now turns out to be the product of n independent geometric random variables with success probabilities zS_m , $m = 1, \dots, n$.

- **The Bose sample grand-canonical distribution:**

First we note that $\mathbb{P}_{\mathbf{S}_n}(\mathcal{B}_{n,z}(m) \geq b_m; m = 1, \dots, n) = \prod_{m=1}^n (zS_m)^{b_m}$. Averaging over \mathbf{S}_n , with $k := \sum_1^n b_m$,

$$\mathbb{P}(\mathcal{B}_{n,z}(m) \geq b_m; m = 1, \dots, n) = \frac{1}{(n\theta)_k} \prod_{m=1}^n \{z^{b_m} (\theta)_{b_m}\}.$$

In particular,

$$\mathbb{P}(\mathcal{B}_{n,z}(m) \geq 1; m = 1, \dots, n) = \frac{(\theta z)^n}{(n\theta)_n}$$

is the probability that all fragments have been visited at least once in a Bose sample (the coupon collector problem, see Feller (1971)). In other related applications (reminiscent of the Banach match box problem), with $c \in \mathbb{N}$, some cell capacity parameter, one can find useful to estimate the probability $\mathbb{P}(\mathcal{B}_{n,z}(m) \leq c; m = 1, \dots, n)$ to have less than c particles in all cells. To compute this quantity, we first need to estimate $\mathbb{P}(\mathcal{B}_{n,z}(m) = b_m; m = 1, \dots, n)$ and then sum over each $b_m \in \{0, \dots, c\}$. Averaging over \mathbf{S}_n , this unconditional occupancy probability is

Theorem 7 With $k := \sum_{m=1}^n b_m$,

$$\begin{aligned} \mathbb{P}(\mathcal{B}_{n,z}(m) = b_m; m = 1, \dots, n) &:= \mathbb{E}\mathbb{P}_{\mathbf{S}_n}(\mathcal{B}_{n,z}(m) = b_m; m = 1, \dots, n) \\ &= \frac{z^k}{(n\theta)_k} \prod_{m=1}^n (\theta)_{b_m} \sum_{q=0}^n \frac{(-z)^q}{(n\theta + k)_q} \binom{n}{q} \prod_{r=1}^q (\theta + b_r). \end{aligned}$$

Proof: Using exchangeability of \mathbf{S}_n

$$\begin{aligned} \mathbb{P}(\mathcal{B}_{n,z}(m) = b_m; m = 1, \dots, n) &= \mathbb{E} \left[\prod_{m=1}^n \{ (zS_m)^{b_m} (1 - zS_m) \} \right] \\ &= \sum_{q=0}^n (-1)^q \sum_{1 \leq m_1 < \dots < m_q \leq n} \mathbb{E} \left(\prod_{r=1}^q (zS_{m_r})^{b_{m_r}+1} \prod_{m \neq \{m_1, \dots, m_q\}} (zS_m)^{b_m} \right) \\ &= z^k \sum_{q=0}^n (-z)^q \binom{n}{q} \mathbb{E} \left(\prod_{r=1}^q S_r^{b_r+1} \prod_{r=q+1}^n S_r^{b_r} \right) \\ &= z^k \sum_{q=0}^n \frac{(-z)^q}{(n\theta)_{k+q}} \binom{n}{q} \prod_{r=1}^q (\theta)_{b_r+1} \prod_{r=q+1}^n (\theta)_{b_r} \\ &= \frac{z^k}{(n\theta)_k} \prod_{m=1}^n (\theta)_{b_m} \sum_{q=0}^n \frac{(-z)^q}{(n\theta + k)_q} \binom{n}{q} \prod_{r=1}^q (\theta + b_r). \end{aligned}$$

We used $(\theta)_{k+q} = (\theta)_k (\theta + k)_q$ and $(\theta)_0 = 1$.

It is exchangeable but not of product form. From this, we would get the law of $\mathcal{K}_{n,z}$ itself

$$\mathbb{P}(\mathcal{K}_{n,z} = k) = \sum_{b_1 + \dots + b_n = k} \mathbb{P}(\mathcal{B}_{n,z}(m) = b_m; m = 1, \dots, n). \quad \square$$

This result allows to extract some information of interest on the grand-canonical equilibrium law of individual cell occupancy. Indeed, with $b_1 \in \mathbb{N}_0$

$$\mathbb{P}(\mathcal{B}_{n,z}(1) = b_1) = \mathbb{E} \left[(zS_n)^{b_1} (1 - zS_n) \right] = z^{b_1} \left[\frac{(\theta)_{b_1}}{(n\theta)_{b_1}} - z \frac{(\theta)_{b_1+1}}{(n\theta)_{b_1+1}} \right]$$

is the one-dimensional expression of $\mathcal{B}_{n,z}(1)$ law. Therefore,

Corollary 8 (i) *The probability that in any cell there is more than b_1 particles is*

$$\mathbb{P}(\mathcal{B}_{n,z}(1) \geq b_1) = z^{b_1} \frac{(\theta)_{b_1}}{(n\theta)_{b_1}}, \quad b_1 \in \mathbb{N}_0.$$

(ii) *When $z < z_c := 1$, this one-dimensional probability decays exponentially.*

(iii) *At critical point $z = z_c$, $\mathcal{B}_{n,z}(1)$ has power law tails with exponent $(n-1)\theta > 0$: therefore, $\mathbb{E}(\mathcal{B}_{n,z}(1)) = \infty$ if and only $(n-1)\theta \in (0, 1]$ and $\sigma^2(\mathcal{B}_{n,z}(1)) < \infty$ if and only $(n-1)\theta > 2$.*

Proof: (iii) We have $\frac{(\theta)_{b_1}}{(n\theta)_{b_1}} = \frac{\Gamma(n\theta)}{\Gamma(\theta)} \frac{(b_1)_\theta}{(b_1)_{n\theta}}$ and when b_1 is large, using Stirling formula, $(b_1)_\theta \sim b_1^\theta$. This shows that $\frac{(\theta)_{b_1}}{(n\theta)_{b_1}} \sim \frac{\Gamma(n\theta)}{\Gamma(\theta)} b_1^{-(n-1)\theta}$ and $\mathcal{B}_{n,z}(1)$ has power law tails with exponent $(n-1)\theta > 0$. \square

• **The number of distinct visited fragments in a Bose sample:**

Let $\mathcal{P}_{n,z} := \sum_{m=1}^n \mathbf{I}(\mathcal{B}_{n,z}(m) > 0)$ be the number of distinct visited fragments in a Bose sample of the grand canonical ensemble. With $p \leq n$, assume that $\mathcal{P}_{n,z} = p$. Let $m_1 < \dots < m_p$ be a realization of the labels $M_q; q = 1, \dots, p$ of these p visited fragments. With $b_q \in \mathbb{N}$, $q = 1, \dots, p$, we have

$$\begin{aligned} & \mathbb{P}_{\mathbf{S}_n}(M_q = m_q; \mathcal{B}_{n,z}(m_q) = b_q; q = 1, \dots, p; \mathcal{P}_{n,z} = p) \\ &= \prod_{q=1}^p (zS_{m_q})^{b_q} \prod_{m=1}^n (1 - zS_m). \end{aligned}$$

Summing over $b_q \in \mathbb{N}$, $q = 1, \dots, p$

$$\mathbb{P}_{\mathbf{S}_n}(M_q = m_q; q = 1, \dots, p; \mathcal{P}_{n,z} = p) = \prod_{q=1}^p (zS_{m_q}) \prod_{m \in [n] \setminus \{m_1, \dots, m_p\}} (1 - zS_m)$$

showing that given \mathbf{S}_n , if the visited labels sequence is known, $\mathcal{P}_{n,z}$ is the sum of independent Bernoulli distributed random variables with success probability zS_{m_q} . Averaging over \mathbf{S}_n and using exchangeability of \mathbf{S}_n , with $p \in \{0, \dots, n\}$

$$\mathbb{P}(\mathcal{P}_{n,z} = p) = \binom{n}{p} \mathbb{E} \left[\prod_{q=1}^p (zS_q) \prod_{q=p+1}^n (1 - zS_q) \right].$$

Differentiating this expression with respect to z , we obtain

$$z \partial_z \mathbb{P}(\mathcal{P}_{n,z} = p) = p \mathbb{P}(\mathcal{P}_{n,z} = p) - (p+1) \mathbb{P}(\mathcal{P}_{n,z} = p+1)$$

so that if $\Phi_{n,z}(u) := \mathbb{E}(u^{\mathcal{P}_{n,z}})$ is the generating function of $\mathcal{P}_{n,z}$, it satisfies $z \partial_z \Phi_{n,z}(u) = -(1-u) \partial_u \Phi_{n,z}(u)$. In particular, $\partial_u \Phi_{n,z}(1) =: \mathbb{E}(\mathcal{P}_{n,z})$ satisfies $z \partial_z \mathbb{E}(\mathcal{P}_{n,z}) = \mathbb{E}(\mathcal{P}_{n,z})$ suggesting $\mathbb{E}(\mathcal{P}_{n,z}) \propto z$. In fact, as shown below

$$\mathbb{E}(\mathcal{P}_{n,z}) = z \in (0, 1),$$

independently of n : the Bose grand-canonical expected number of visited fragments is at most one.

Developing the second product in the expression of $\mathbb{P}(\mathcal{P}_{n,z} = p)$ and making use of Eq. (2.2), we get the alternate sum representation

$$\begin{aligned} \mathbb{P}(\mathcal{P}_{n,z} = p) &= (z\theta)^p \binom{n}{p} \sum_{q=0}^{n-p} \binom{n-p}{q} \frac{(-\theta z)^q}{(n\theta)_{p+q}} \\ &= \binom{n}{p} \sum_{r=p}^n (-1)^{r-p} \binom{n-p}{r-p} \frac{(\theta z)^r}{(n\theta)_r}. \end{aligned}$$

Summing over $p \in \{0, \dots, n\}$ and reversing the summation order, this gives

$$\begin{aligned} \mathbb{E}(\mathcal{P}_{n,z}) &= n! \sum_{p=1}^n \frac{1}{(p-1)!} \sum_{r=p}^n \frac{(-1)^{r-p}}{(r-p)! (n-r)!} \frac{(\theta z)^r}{(n\theta)_r} \\ &= n! \sum_{r=1}^n \frac{1}{(n-r)!} \frac{(\theta z)^r}{(n\theta)_r} \frac{1}{(r-1)!} \sum_{q=0}^{r-1} (-1)^q \binom{r-1}{q} \\ &= z \in (0, 1). \end{aligned}$$

More generally, proceeding similarly

$$\mathbb{E}(u^{\mathcal{P}_{n,z}}) = \sum_{r=0}^n \binom{n}{r} \frac{(-\theta z(1-u))^r}{(n\theta)_r}.$$

From this, the variance of $\mathcal{P}_{n,z}$ is $\sigma^2(\mathcal{P}_{n,z}) = z - z^2(\theta + 1)/(n\theta + 1) > 0$ and more generally,

$$\mathbb{E}(\{\mathcal{P}_{n,z}\}_r) = \{n\}_r \frac{(\theta z)^r}{(n\theta)_r}$$

are the falling factorial moments of $\mathcal{P}_{n,z}$. To summarize, we obtained

Theorem 9 *The law of $\mathcal{P}_{n,z}$ is characterized by any of the three equivalent properties*

(i) *For $p \in \{0, \dots, n\}$, with $z \in (0, 1)$*

$$\mathbb{P}(\mathcal{P}_{n,z} = p) = \binom{n}{p} \sum_{r=p}^n (-1)^{r-p} \binom{n-p}{r-p} \frac{(\theta z)^r}{(n\theta)_r}.$$

(ii) *With $u \in [0, 1]$, $\mathcal{P}_{n,z}$ has the generating function*

$$\mathbb{E}(u^{\mathcal{P}_{n,z}}) = \sum_{r=0}^n \binom{n}{r} \frac{(-\theta z(1-u))^r}{(n\theta)_r}.$$

(iii) *With $r \in \{0, \dots, n\}$, the falling factorial moments of $\mathcal{P}_{n,z}$ are*

$$\mathbb{E}(\{\mathcal{P}_{n,z}\}_r) = \{n\}_r \frac{(\theta z)^r}{(n\theta)_r}.$$

In particular: $\mathbb{E}(\mathcal{P}_{n,z}) = z$ and $\sigma^2(\mathcal{P}_{n,z}) = z - z^2(\theta + 1) / (n\theta + 1)$.

In the $*$ -Kingman limit, we clearly obtain

Corollary 10 *We have $\mathcal{P}_{n,z} \xrightarrow{d_*} \mathcal{P}_z$. Specifically,*

(i) *With $p \in \{0, \dots, n\}$*

$$\mathbb{P}(\mathcal{P}_{n,z} = p) \rightarrow_* \mathbb{P}_*(\mathcal{P}_z = p) = \sum_{r=p}^{\infty} \frac{(-1)^{r-p}}{r!} \binom{r}{p} \frac{(\gamma z)^r}{(\gamma)_r}$$

where $p \in \mathbb{N}_0$.

(ii)

$$\mathbb{E}(u^{\mathcal{P}_{n,z}}) \rightarrow_* \mathbb{E}_*(u^{\mathcal{P}_z}) = \sum_{r=0}^{\infty} \frac{(-\gamma z(1-u))^r}{r! \cdot (\gamma)_r}.$$

(iii) *From this, $\mathbb{E}_*(\mathcal{P}_z) = z$ and $\sigma_*^2(\mathcal{P}_z) = z - z^2 / (\gamma + 1) > 0$ and more generally, with $r \in \mathbb{N}_0$*

$$\mathbb{E}_*(\{\mathcal{P}_z\}_r) = \frac{(\gamma z)^r}{(\gamma)_r}$$

are the r -falling factorial moments of \mathcal{P}_z .

(iv) *When $\gamma \uparrow \infty$*

$$\mathcal{P}_z \xrightarrow{d} \text{Poisson}(z).$$

Proof: Points (i) to (iii) can easily be derived from the latter Theorem. Point (iv) is obtained by passing to the limit $\gamma \uparrow \infty$ on $\mathbb{E}_*(u^{\mathcal{P}_z})$ and, recalling $(\gamma)_r \sim \gamma^r$, $\mathbb{E}_*(u^{\mathcal{P}_z}) \rightarrow_{\gamma \uparrow \infty} \exp\{-z(1-u)\}$, the moment generating function

of a Poisson(z) random variable. \square

• **Statistics of the number of energy states with prescribed amount of particles (Bose-Ewens sampling formula):**

The understanding of the number of occupied (and therefore also of unoccupied) states is part of the broader problem of the number of states with prescribed amount of particles. Let therefore

$$\mathcal{A}_{n,z}(i) := \sum_{m=1}^n \mathbf{I}(\mathcal{B}_{n,z}(m) = i); i \geq 0$$

count the number of fragments with i particles. Clearly, $\mathcal{A}_{n,z}(0) = n - \mathcal{P}_{n,z}$ is the number of free fragments and $\sum_{i \geq 0} \mathcal{A}_{n,z}(i) = n$. The case $i = 1$ ($i = 2$) corresponds to singleton (doubleton) states. Recall first

$$\mathbb{P}(\mathcal{B}_{n,z}(m) = b_m; m = 1, \dots, n) = \mathbb{E} \prod_{m=1}^n \mathbb{P}_{\mathbf{S}_n}(\xi_{m,z} = b_m)$$

where, given \mathbf{S}_n , $(\xi_{1,z}, \dots, \xi_{n,z})$ is an independent sequence each with geometric distribution such that: $\mathbb{P}_{\mathbf{S}_n}(\xi_{m,z} \geq b_m) = (zS_m)^{b_m}$.

From this, using exchangeability of \mathbf{S}_n , after a simple rearrangement, we easily get:

Proposition 11 *For all sequences $(a_i \in \mathbb{N}_0; i \geq 0)$ satisfying $\sum_{i \geq 0} a_i = n$, the grand canonical Bose-Ewens sampling formula from Dirichlet proportions is:*

$$\mathbb{P}(\mathcal{A}_{n,z}(1) = a_1, \dots, \mathcal{A}_{n,z}(i) = a_i, \dots) = n! \cdot \mathbb{E} \left[\prod_{i=0}^{\sum_{l \geq 1} la_l} \frac{1}{a_i!} \prod_{j=a_{i-1}+1}^{\sum_{k=0}^i a_k} \mathbb{P}_{\mathbf{S}_n}(\xi_{j,z} = i) \right]$$

where $\mathbb{P}_{\mathbf{S}_n}(\xi_{j,z} = i) = (zS_j)^i (1 - zS_j)$ and $a_{-1} := 0$.

The exact distribution of $(\mathcal{A}_{n,z}(i); i \geq 0)$ could easily be obtained in closed form by further evaluating the Dirichlet integrals appearing in the right hand side of the latter formula. As it would lead to heavy combinatorial developments, we skip the details for the reader convenience.

We note however that the expected values of $(\mathcal{A}_{n,z}(i); i \geq 0)$ are quite easy to derive. Indeed, recalling $\mathcal{A}_{n,z}(i) = \sum_{m=1}^n \mathbf{I}(\mathcal{B}_{n,z}(m) = i)$, we get

$$\begin{aligned} \mathbb{E}(\mathcal{A}_{n,z}(i)) &= \sum_{m=1}^n \mathbb{P}(\mathcal{B}_{n,z}(m) = i) = \sum_{m=1}^n \mathbb{E} \mathbb{P}_{\mathbf{S}_n}(\xi_{m,z} = i) \\ &= n \mathbb{E} \mathbb{P}_{\mathbf{S}_n}(\xi_{n,z} = i) = n \mathbb{E} \left((zS_n)^i (1 - zS_n) \right) \\ &= n \left[z^i \frac{(\theta)_i}{(n\theta)_i} - z^{i+1} \frac{(\theta)_{i+1}}{(n\theta)_{i+1}} \right] = nz^i \frac{(\theta)_i}{(n\theta)_i} \left(1 - z \frac{\theta + i}{n\theta + i} \right). \end{aligned}$$

Thus, consistently,

$$\begin{aligned} \mathbb{E}(\mathcal{K}_{n,z}) &= \sum_{i \geq 1} i \mathbb{E}(\mathcal{A}_{n,z}(i)) = n \sum_{i \geq 1} iz^i \left[\frac{(\theta)_i}{(n\theta)_i} - z \frac{(\theta)_{i+1}}{(n\theta)_{i+1}} \right] \\ &= n \sum_{i \geq 1} z^i \frac{(\theta)_i}{(n\theta)_i} = n \mathbb{E} \left(\frac{zS_1}{1 - zS_1} \right) \end{aligned}$$

and

$$\mathbb{E}(\mathcal{P}_{n,z}) = \sum_{i \geq 1} \mathbb{E}(\mathcal{A}_{n,z}(i)) = z.$$

Note also that, taking the Kingman limit

$$\mathbb{E}(\mathcal{A}_{n,z}(i)) \rightarrow_* \mathbb{E}_*(\mathcal{A}_z(i)) = \gamma \frac{(i-1)!}{(\gamma)_i} z^i \left(1 - \frac{zi}{\gamma+i} \right).$$

• **Bose randomized occupancy and order statistics:**

In practice, it can be useful to consider the sampling process from $\mathbf{S}_{(n)}$ that is after having ordered the fragment sizes. Let then $\mathbf{S}_{(n)} := (S_{(1)}, \dots, S_{(n)})$ be the order statistics of \mathbf{S}_n , with $S_{(1)} > \dots > S_{(n)}$. The joint probability density of this vector now is

$$f_{\mathbf{S}_{(n)}}(s_1, \dots, s_n) = \frac{n! \Gamma(n\theta)}{\Gamma(\theta)^n} \prod_{m=1}^n s_m^{\theta-1} \cdot \mathbf{I}(s_1 > \dots > s_n) \cdot \delta_{(\sum_{m=1}^n s_{(m)} - 1)}.$$

Clearly, sampling under \mathbf{S}_n does not reduce to sampling under $\mathbf{S}_{(n)}$.

For instance, given $\mathbf{S}_{(n)}$, the joint probability of occupancies $\mathcal{B}_{(n),z}(m)$ of fragment m with size $S_{(m)}$; $m = 1, \dots, n$ now reads

$$\mathbb{P}_{\mathbf{S}_{(n)}}(\mathcal{B}_{(n),z}(m) = b_m; m = 1, \dots, n) = \prod_{m=1}^n \left\{ (zS_{(m)})^{b_m} (1 - zS_{(m)}) \right\}$$

Exchangeability is lost; averaging over $\mathbf{S}_{(n)}$ with distribution $f_{\mathbf{S}_{(n)}}$ on the simplex, the unconditional occupancy probability would be

$$\begin{aligned} \mathbb{P}(\mathcal{B}_{(n),z}(m) = b_m; m = 1, \dots, n) &:= \mathbb{E} \mathbb{P}_{\mathbf{S}_{(n)}}(\mathcal{B}_{(n),z}(m) = b_m; m = 1, \dots, n) = \\ &= \mathbb{E} \left[\prod_{m=1}^n (zS_{(m)})^{b_m} (1 - zS_{(m)}) \right]. \end{aligned}$$

In particular:

Proposition 12 *The grand canonical distribution of ground state occupancy is*

$$\mathbb{P}(\mathcal{B}_{(n),z}(1) \geq b_1) = z^{b_1} \left(1 - b_1 \int_0^\infty \mathbb{P}(S_{(1)} > s) s^{b_1-1} ds \right).$$

Proof: With $\mathbf{S}_{(n)\setminus 1} := (S_{(2)}, \dots, S_{(n)})$, let $\psi(b_1) := \mathbb{E}(S_{(1)}^{b_1})$ stand for the integral moments of $S_{(1)}$. With $b_1 \in \mathbb{N}_0$, the ground state $S_{(1)}$ occupancy is

$$\begin{aligned} \mathbb{P}(\mathcal{B}_{(n),z}(1) = b_1) &= \mathbb{E} \left[\frac{(zS_{(1)})^{b_1} Z_{z, \mathbf{S}_{(n)\setminus 1}}(z)}{Z_{z, \mathbf{S}_{(n)}}(z)} \right] \\ &= \mathbb{E} \left[(zS_{(1)})^{b_1} (1 - zS_{(1)}) \right] \\ &= z^{b_1} [\psi(b_1) - z\psi(b_1 + 1)]. \end{aligned}$$

Recalling (see Holst (2001), for instance) that, with $S_{(1)} \in (\frac{1}{n}, 1)$, the complementary distribution function of $S_{(1)}$ is

$$\mathbb{P}(S_{(1)} > s) = \sum_{m=1}^n \frac{(-1)^{m-1} \binom{n}{m} \Gamma(n\theta)}{\Gamma(\theta)^m \Gamma((n-m)\theta)} \int_s^1 \dots \int_s^1 \left(1 - \sum_{l=1}^m t_l \right)_+^{(n-m)\theta-1} \prod_{l=1}^m \frac{dt_l}{t_l^{1-\theta}},$$

we get

$$\psi(b_1) := \mathbb{E} \left[S_{(1)}^{b_1} \right] = 1 - b_1 \int_0^\infty \mathbb{P}(S_{(1)} > s) s^{b_1-1} ds.$$

Therefore

$$\mathbb{P}(\mathcal{B}_{(n),z}(1) \geq b_1) = z^{b_1} \left(1 - b_1 \int_0^\infty \mathbb{P}(S_{(1)} > s) s^{b_1-1} ds \right). \quad \square$$

Let $p \in \{1, \dots, n\}$ and $1 \leq m_1 < \dots < m_p \leq n$ be an increasing subsequence of $\{1, \dots, n\}$. Let M_1, \dots, M_p be the labels of visited fragments from $\mathbf{S}_{(n)}$ and $\mathcal{P}_{n,z}$ denote the number of such occupied states. With $b_q \in \mathbb{N}$, we clearly have

$$\begin{aligned} \mathbb{P}_{\mathbf{S}_{(n)}}(M_q = m_q, \mathcal{B}_{(n),z}(m_q) = b_q; q = 1, \dots, p; \mathcal{P}_{n,z} = p) = \\ \prod_{q=1}^p \left[(zS_{(m_q)})^{b_q} (1 - zS_{(m_q)}) \right] \prod_{q \neq 1, \dots, p} [1 - zS_{(m_q)}]. \end{aligned}$$

Averaging over $\mathbf{S}_{(n)}$ whose joint law is known, first gives the unconditional probability of the event $M_q = m_q, \mathcal{B}_{(n),z}(m_q) = b_q; q = 1, \dots, p; \mathcal{P}_{n,z} = p$.

Next, summing the above conditional probability over $b_q; q = 1, \dots, p$, we get

$$\mathbb{P}_{\mathbf{S}_{(n)}}(M_q = m_q; q = 1, \dots, p; \mathcal{P}_{n,z} = p) = \prod_{q \neq \{1, \dots, p\}} (1 - zS_{(m_q)}) \prod_{q=1}^p (zS_{(m_q)}).$$

Averaging over $\mathbf{S}_{(n)}$, using the joint law of $\mathbf{S}_{(n)}$, we get the unconditional probability $\mathbb{P}(M_q = m_q; q = 1, \dots, p; \mathcal{P}_{n,z} = p)$ that only $p \leq n$ of the ordered cells with labels m_q are occupied. In particular

$$\mathbb{E} \left\{ \prod_{q=2}^n [1 - zS_{(q)}] (zS_{(1)}) \right\}$$

is the unconditional probability that only ground state $S_{(1)}$ is occupied.

3.2 Evidence of a phase transition in some cases

It remains to interpret z in more details to get a deeper insight into the phase transition question. Recall that

$$\mathbb{E}_{\mathbf{S}_n} (\mathcal{B}_{n,z}(m)) = \frac{zS_m}{1 - zS_m} = \sum_{i \geq 1} z^i S_m^i$$

and so, using exchangeability of \mathbf{S}_n

$$\begin{aligned} \mathbb{E}_{\mathbf{S}_n} (\mathcal{K}_{n,z}) &= \sum_{m=1}^n \frac{zS_m}{1 - zS_m} \\ \kappa &: = \mathbb{E} \mathbb{E}_{\mathbf{S}_n} (\mathcal{K}_{n,z}) = n \mathbb{E} \left(\frac{zS_n}{1 - zS_n} \right). \end{aligned}$$

Recalling $S_n \stackrel{d}{\sim} \text{beta}(\theta, (n-1)\theta)$, we get the following expression for the expected number of particles in the sample

$$\kappa = n \sum_{i \geq 1} z^i \mathbb{E} (S_n^i) = n \sum_{i \geq 1} z^i \frac{(\theta)_i}{(n\theta)_i}.$$

This is also the implicit state equation

$$\kappa = H_{n,\theta}(z)$$

where $H_{n,\theta}(z)$ has the power series representation

$$H_{n,\theta}(z) = n \sum_{i \geq 1} z^i \frac{(\theta)_i}{(n\theta)_i}.$$

Usually, it is of interest to try to deduce z from (θ, κ) which are the physical quantities which are known in practice.

The convergence radius of the series $H_{n,\theta}(z)$ clearly is $z_c = 1$. Two cases arise: either $H_{n,\theta}(1) = \infty$ or $H_{n,\theta}(1) < \infty$.

The condition $H_{n,\theta}(1) < \infty$ is fulfilled if and only if $\theta > \frac{1}{n-1}$; indeed, $\frac{(\theta)_i}{(n\theta)_i} = \frac{\Gamma(n\theta)}{\Gamma(\theta)} \frac{(i)_\theta}{(i)_{n\theta}}$ and, when i is large, by Stirling formula, $(i)_\theta \sim i^\theta$. This shows that $\frac{(\theta)_i}{(n\theta)_i} \sim \frac{\Gamma(n\theta)}{\Gamma(\theta)} i^{-(n-1)\theta}$ which is the term of a summable series if and only if $\theta > \frac{1}{n-1}$: A phase transition phenomenon pops in when the disorder parameter θ is large enough.

When $\theta > \frac{1}{n-1}$, the function $\theta \rightarrow H_{n,\theta}(1)$ is monotone decreasing and maps $\theta \in \left(\frac{1}{n-1}, \infty\right)$ onto $\left(\infty, \frac{n}{n-1}\right)$; assuming $\kappa > \frac{n}{n-1}$, there is therefore a unique $\theta_c > \frac{1}{n-1}$ defined by $\kappa = H_{n,\theta_c}(1)$. For $\theta < \theta_c$, by Bürmann-Lagrange inversion formula

$$z = 1 + \sum_{l \geq 1} \frac{\kappa^l}{l} h_l(\theta) \quad \text{with } h_l(\theta) := [z^{l-1}] \left(\frac{H_{n,\theta}(z)}{z} \right)^{-l}.$$

Remark: In this approach, the free parameter is the average number of particles κ and the Lagrange inversion formula only holds for $\theta < \theta_c$ (disorder is small enough). If θ is the free parameter, Lagrange inversion formula only is valid for $\kappa < \kappa_c := H_{n,\theta}(1)$ (the expected number of particles is small enough). In the (θ, κ) plane, the critical line $\kappa_c =: \kappa_c(\theta)$ separates a weak disorder phase $\kappa < \kappa_c$ from a strong disorder phase $\kappa > \kappa_c$.

When $\theta \leq \frac{1}{n-1}$ (disorder is small enough), the largest fragment is dominant: particles largely tend to accumulate within this fragment. The same holds true when $\theta > \frac{1}{n-1}$ provided $\kappa < \kappa_c$. However, if $\theta > \frac{1}{n-1}$ (disorder strong enough) and $\kappa > \kappa_c$ (the expected number of particles is large enough), we conjecture that κ_c particles in average will accumulate in the largest fragment while the rest $(\kappa - \kappa_c)$ is scattered on all the other fragments. In the strong disorder phase, particles tend to spread on all cells. This property seems to result from the conjunction of two effects: the randomness of the sampling probabilities within bins and the indistinguishability of balls to be packed. \diamond

One can summarize shortly the results as follows:

Proposition 13 *We have:*

- (i) *If $0 < \theta \leq 1/(n-1)$, there is a unique weakly disordered phase.*
- (ii) *If $\infty > \theta > 1/(n-1)$, a phase transition occurs: the critical line separating the strong disorder ($\kappa > \kappa_c > n/(n-1)$) from weak disorder ($\kappa < \kappa_c$) phases has equation $\kappa = \kappa_c(\theta) = n\mathbb{E}\left(\frac{S_n}{1-S_n}\right)$ which is also the convergent series*

$$\kappa = \kappa_c(\theta) = n \sum_{i \geq 1} \frac{(\theta)_i}{(n\theta)_i}.$$

Recalling that $S_n \stackrel{d}{\sim} \text{beta}(\theta, (n-1)\theta)$, the phase transition criterion $\theta > 1/(n-1)$ indicates that the probability mass of fragment sizes in a neighborhood of $s = 1$ has to be small enough (in the sense that the density of S_n should vanish at point 1).

Passing to the Kingman limit $n \uparrow \infty$, $\theta \downarrow 0$, while $n\theta = \gamma > 0$, assuming $z < 1$, one can check that

$$\mathcal{K}_{n,z} \xrightarrow{d} \mathcal{K}_z.$$

Indeed, with $\mathbf{S}_{(\infty)}$ the Poisson-Dirichlet weak $*$ -limit of $\mathbf{S}_{(n)}$

$$\mathbb{E}_*(u^{\mathcal{K}_z}) = \mathbb{E} \left(\prod_{m=1}^{\infty} \frac{1 - zS_{(m)}}{1 - uzS_{(m)}} \right),$$

showing that, given $\mathbf{S}_{(\infty)}$, \mathcal{K}_z is the sum of independent geometric random variables with respective success probabilities $zS_{(m)}$, $m = 1, \dots, \infty$.

In particular,

$$\kappa := \mathbb{E}(\mathcal{K}_{n,z}) \rightarrow_* \kappa_* := \mathbb{E}_*(\mathcal{K}_z) = \Gamma(\gamma + 1) \sum_{i \geq 1} \frac{z^i}{(i)_\gamma}.$$

At $z_c = 1$, this series is convergent if and only if $\gamma > 1$. Proceeding similarly as for the finite Dirichlet case, we have

Corollary 14 *Consider Bose samples from Poisson-Dirichlet partitioning. Then:*

(i) *If $0 < \gamma \leq 1$, there is no phase transition and the only available phase is the weakly disordered one.*

(ii) *If $\infty > \gamma > 1$, there is a phase transition: the critical line separating the strong disorder ($\kappa > \kappa_c > 1$) from weak disorder ($\kappa < \kappa_c$) phases has equation $\kappa = \kappa_c(\gamma)$ which is the convergent series*

$$\kappa = \kappa_c(\gamma) = \gamma \sum_{i \geq 1} \frac{(i-1)!}{(\gamma)_i} = \Gamma(\gamma + 1) \sum_{i \geq 1} \frac{1}{(i)_\gamma}.$$

References

- [1] Cesaroli, M. (1983). Poisson randomization in occupancy problems. *Journal of Mathematical Analysis and Applications*, **94**, 150-165.
- [2] Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, **3**, 87-112.
- [3] Ewens, W.J. (1990). Population genetics theory - the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory*, Edt. S. Lessard, Kluwer, Dordrecht.

- [4] Feller, W. (1971). *An introduction to probability theory and its applications*, **1** and **2**. John Wiley and Sons, Second Edition, New York.
- [5] Good, I. J. (1968). *The estimation of probabilities. An essay on modern Bayesian methods*. MIT Research Monograph, No **30** The M.I.T. Press, Cambridge, Mass.
- [6] Holst, L. (1985). On discrete spacings and the Bose-Einstein distribution. *Contributions to Probability and Statistics. Essays in honour of Gunnar Blom*. Ed. by Jan Lanke and Georg Lindgren, Lund, 169–177.
- [7] Holst, L. (2001). The Poisson-Dirichlet distribution and its relatives revisited. Preprint of the *Royal Institute of Technology*, Stockholm, Sweden.
- [8] Huillet, T. (2005). Sampling formulae arising from random Dirichlet populations. *Communications in Statistics - Theory and Methods*, **34**, No 5, 1019-1040.
- [9] Johnson, N. L., Kotz, S. (1977). *Urn models and their application. An approach to modern discrete probability theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York-London-Sydney, xiii+402 pp.
- [10] Kolchin, V. F. (1986). *Random mappings*. Translated from the Russian. With a foreword by S. R. S. Varadhan. Translation Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York.
- [11] Keener, R., Rothman, E., Starr, N. (1987). Distributions on partitions. *The Annals of Statistics*, **15**, No 4, 1466–1481.
- [12] Kingman, J.F.C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society. Series B*, **37**, 1–22.
- [13] Kingman, J.F.C. (1978). Random partitions in population genetics. *Proceedings of the Royal Society. London. Series A*, **361**, No 1704, 1–20.
- [14] Kingman, J.F.C. (1993). *Poisson processes*. Clarendon Press, Oxford.
- [15] Korwar, R. M., Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *Annals of Probability*, **1**, 705–711.
- [16] Sibuya, M. (1993). A random clustering process. *Ann. Inst. Statist. Math.* **45**, No 3, 459–465.
- [17] Tavaré, S., Ewens, W.J. (1997). Multivariate Ewens distribution. Chapter **41** in *Discrete Multivariate Distributions*, Edts N.L. Johnson, S. Kotz and N. Balakrishnan, pages 232-246, (Wiley, New York).