J Stat Phys (2013) 153:763–800
DOI 10.1007/s10955-013-0865-y

# Occupancy Distributions Arising in Sampling from Gibbs-Poisson Abundance Models

**Thierry Huillet · Servet Martínez**

Received: 5 March 2013 / Accepted: 3 October 2013 / Published online: 16 October 2013
© Springer Science+Business Media New York 2013

**Abstract** Estimating the number $n$ of unseen species from a $k$-sample displaying only $p \leq k$ distinct sampled species has received attention for long. It requires a model of species abundance together with a sampling model. We start with a discrete model of iid stochastic species abundances, each with Gibbs-Poisson distribution. A $k$-sample drawn from the $n$-species abundances vector is the one obtained while conditioning it on summing to $k$. We discuss the sampling formulae (species occupancy distributions, frequency of frequencies) in this context. We then develop some aspects of the estimation of $n$ problem from the size $k$ of the sample and the observed value of $P_{n,k}$, the number of distinct sampled species.

It is shown that it always makes sense to study these occupancy problems from a Gibbs-Poisson abundance model in the context of a population with infinitely many species. From this extension, a parameter $\gamma$ naturally appears, which is a measure of richness or diversity of species. We rederive the sampling formulae for a population with infinitely many species, together with the distribution of the number $P_k$ of distinct sampled species. We investigate the estimation of $\gamma$ problem from the sample size $k$ and the observed value of $P_k$.

We then exhibit a large special class of Gibbs-Poisson distributions having the property that sampling from a discrete abundance model may equivalently be viewed as a sampling problem from a random partition of unity, now in the continuum. When $n$ is finite, this partition may be built upon normalizing $n$ infinitely divisible iid positive random variables by its partial sum. It is shown that the sampling process in the continuum should generically be biased on the total length appearing in the latter normalization. A construction with size-biased sampling from the ranked normalized jumps of a subordinator is also supplied, would the problem under study present infinitely many species. We illustrate our point of view with many examples, some of which being new ones.

T. Huillet (✉)
Laboratoire de Physique Théorique et Modélisation, Université de Cergy-Pontoise, CNRS UMR-8089, Site de Saint Martin, 2 avenue Adolphe-Chauvin, 95302 Cergy-Pontoise, France
e-mail: huillet@u-cergy.fr

S. Martínez
Depto. Ingenieria Matematica and Centro Modelamiento Matematico, Universidad de Chile, UMI 2071, Uchile-Cnrs, Casilla 170-3 Correo 3, Santiago, Chile
e-mail: smartine@dim.uchile.cl

**Keywords** Occupancy distributions · Sampling from Gibbs-Poisson distribution · Species abundance and frequencies · Biodiversity · Combinatorial probability · Subordinators

## 1 Introduction and Outline of Main Results

Estimating the number $n$ of unseen species from a $k$-sample displaying only $p \le k$ distinct sampled species has been a challenging problem since the mid-twentieth century, [21]. It requires a model of species abundance together with a sampling model [16], and the answer to the latter question is of course model-dependent. In this work, we start with a discrete model of independent and identically distributed (iid) stochastic species abundances $\boldsymbol{\xi}_n := (\xi_1, \ldots, \xi_n)$, based on compound Poisson distributions for $\xi \stackrel{d}{=} \xi_1$. We discuss the sampling formulae (species occupancy distributions, frequency of frequencies) in this discrete context. Typically, a $k$-sample drawn from the $n$-species abundances vector is the one obtained while conditioning this vector on summing to $k$ (the sample size). Sampling from iid compound Poisson abundance random variables (rvs) in this sense results in a Gibbs-Poisson sampling model from $\boldsymbol{\xi}_n$. It has to do with random allocation of balls into boxes, [37, 38]. Various combinatorial identities arising in this setup are discussed. A distribution for the number of distinct visited species $P_{n,k}$ in a $k$-sample from a population of size $n$ with compound Poisson abundance is derived. For this class of sampling problems, a 'temperature' type parameter $\theta > 0$ pops in naturally. It is a measure of how similar the box occupancy numbers look like statistically, after the sampling process: the smaller the values of $\theta$, the more likely it is that these occupancy numbers are disparate. When sampling from $\boldsymbol{\xi}_n$, we then discuss some aspects of the problem of the estimation of the number of species $n$ from the size $k$ of the sample and the number $P_{n,k}$ of distinct sampled species, assuming $\theta$ to be known. These results are supplied in Propositions 1 and 3.

It turns out that it always makes sense to study these occupancy problems from a Gibbs-Poisson abundance model in the context of a population with infinitely many species, provided $n$ goes to $\infty$ together with $\theta$ going to 0 while $n\theta \to \gamma > 0$. From this construction, $\gamma$ then appears as a measure of species richness or diversity. We rederive the sampling formulae (species occupancy distributions, frequency of frequencies) for a population with infinitely many species, together with the distribution of the number $P_k$ of distinct sampled species. We discuss the problem of the estimation of the diversity parameter $\gamma$ from the size $k$ of the sample and the number $P_k$.

One particular model in the compound Poisson class has been discussed at length in the literature: the sampling problem from a population with discrete negative binomial distribution abundance $\xi$, both when the population is made of a finite number of species and when there are infinitely many of them. For this particular model, when there are infinitely many species, the obtained sampling formulae are the ones of Ewens, [18]. It is also well-known that the Ewens sampling formulae may also be viewed as sampling from a random Dirichlet partition of the unity when the number of species is finite or as sampling from a random Poisson-Dirichlet partition of unity when there are infinitely many classes, [26]. This property is remarkable. By sampling from a partition of the continuum [0, 1], we mean that we draw independently $k$ uniform random variables on the unit interval, looking at the subintervals of the partition which are being hit in the process to form the occupancy distributions of classes.

In this work, we exhibit a large class of compound Poisson distributions sharing with the negative binomial distribution this property that sampling from a discrete abundance model may equivalently be viewed as a sampling problem from a random partition of unity in the continuum. When $n$ is finite, this partition may be built upon normalizing $n$ infinitely divisible iid non-negative random variables $\mathbf{Y}_n := (Y_1, Y_2, \dots, Y_n)$ by its partial sum. We exhibit the one-to-one correspondence between the laws of $\xi$ and $Y \stackrel{d}{=} Y_1$, assuming $\xi$ to be in the special class. It is however shown that the sampling process in the continuum should generically be biased on the total length appearing in the latter normalization. A construction with such biased sampling from the ranked normalized jumps of a subordinator is also supplied, would the problem under study present infinitely many species. The biasing factors account for the fact that the Gibbs-Poisson occupancy models are not in general sampling consistent as $k$ varies (are not EPPFs). A complete classification of EPPFs induced by the unbiased multinomial sampling from partition of unity can be found in [23, 25].

With this correspondence in mind, we discuss several examples, among which the Engen extended negative binomial model [15], the Berestycki-Pitman model [3] for the enumeration of forests of trees with generalized binomial generator, the polylog and the Mittag-Leffler models. When there are some reasons to suspect that the ranked species frequencies decay algebraically with the rank number, then the Engen model is well suited. Would one think of the ranked species frequencies as decaying exponentially with the rank number, then the Ewens model seems relevant. If the ranked species frequencies are believed to decay exponentially as some power of the rank number, then one should opt for the polylog model.

We end up giving a new example of $\xi$ sharing some common issues with the Engen's model (in particular the algebraic decay property of the ranked frequencies). For this precise model, we are able to give an exact estimator of the biodiversity parameter.

## 2 Sampling from Discrete Gibbs-Poisson Distributions

The sampling problem from a negative binomial abundance model and its Dirichlet counterpart in the continuum suggest to study the following general construction (see [3, 27, 28, 37] and [45] for similar recent interest).

### 2.1 Generating and Partition Function (see [10] and [45], Sect. 1)

With $\phi_\bullet := (\phi_m; m \geq 1)$ a sequence of non-negative real numbers with $\phi_1 > 0$, let

$$\phi(x) := \sum_{m \geq 1} \frac{\phi_m}{m!} x^m \tag{1}$$

be a formal power series in $x$. Assume that $x_0 := \sup(x > 0 : \phi(x) < \infty) \in (0, +\infty]$ is its convergence radius. Then $\phi(x)$ defines a convergent series on $|x| < x_0$ and it is absolutely monotone on $(0, x_0)$ in the sense that $\phi^{(n)}(x) \geq 0$ for all $n \geq 0$ and $x \in (0, x_0)$. We call it the local exponential generating function.

Let $\theta > 0$ and consider the exponential 'partition' generating function

$$Z_\theta(x) = e^{\theta \phi(x)}. \tag{2}$$

This function also defines a convergent series on $|x| < x_0$ with $Z_\theta(0) = 1$. Further, with $\sigma_k(\theta) = k![x^k] Z_\theta(x)$ (where $[x^k] f(x)$ is the $x^k$-coefficient in the series expansion of the function $f(x)$):

$$Z_\theta(x) = 1 + \sum_{k \geq 1} \frac{x^k}{k!} \sigma_k(\theta).$$

Since $\partial_x Z_\theta(x) = \theta\phi'(x)Z_\theta(x)$, we get the recurrence:

$$\sigma_{k+1}(\theta) = \theta \sum_{l=0}^{k} \binom{k}{l} \phi_{k-l+1}\sigma_l(\theta), \quad k \geq 0, \sigma_0(\theta) \equiv 1. \tag{3}$$

Similarly, since $\partial_\theta Z_\theta(x) =: Z'_\theta(x) = \phi(x)Z_\theta(x)$, we find:

$$\sigma'_k(\theta) = \sum_{l=0}^{k-1} \binom{k}{l} \phi_{k-l}\sigma_l(\theta), \quad k \geq 1, \sigma_0(\theta) = 1. \tag{4}$$

Then, clearly,

$$\sigma_k(\theta) = \sum_{l=1}^{k} B_{k,l}(\phi_\bullet)\theta^l, \tag{5}$$

with:

$$B_{k,l}(\phi_\bullet) = \frac{k!}{l!}\left[x^k\right]\phi(x)^l = \frac{k!}{l!} \sum_{\mathbf{m}_l : |\mathbf{m}_l| = k} \prod_{j=1}^{l} \frac{\phi_{m_j}}{m_j!} \geq 0.$$

In the latter sum, summation runs over $\mathbf{m}_l := (m_1, \ldots, m_l) \in \mathbb{N}^l$, with $|\mathbf{m}_l| := \sum_{j=1}^{l} m_j = k$ and $\mathbb{N} := \{1, 2, \ldots\}$; there are $\binom{k-1}{l-1}$ terms in such sums. So $\sigma_k(\theta)$ is a degree-$k$ Bell polynomial in $\theta$ whose $\theta^l$ coefficient is $B_{k,l}(\phi_\bullet)$ which is known as the Bell exponential polynomial in the variables $\phi_\bullet$ (see [10]). On $\theta > 0$, the function $\sigma_k(\theta)$ is convex and log-concave, for all $k$. As a polynomial with non-negative coefficients of degree $k$, $\sigma_k(\theta)$ has no strictly positive real root and at most $k$ real non-positive roots (including 0), counting roots with their multiplicity.

*Remarks* (Bell polynomials and convolutions)

(i) Define $(\phi * \phi)_m := \sum_{l=1}^{m-1} \binom{m}{l} \phi_l \phi_{m-l}$, $m \geq 2$, as the binomial self-convolution sequence of $\phi_m$. Define $\phi_m^{*p}$ as the $m$th term, $m \geq p$, of the sequence $\phi^{*p} := \phi * \cdots * \phi$, $p$ times; then the following convolution identity is well-known to hold:

$$B_{k,p}(\phi_\bullet) = \phi_k^{*p}/p!.$$

(ii) Because $Z_{\theta+\theta'}(x) = Z_\theta(x)Z_{\theta'}(x)$, the polynomials $\sigma_k(\theta)$ satisfy

$$\sigma_k\left(\theta + \theta'\right) = \sum_{l=0}^{k} \binom{k}{l} \sigma_l(\theta)\sigma_{k-l}\left(\theta'\right) \quad \text{for all } \theta, \theta' > 0, \tag{6}$$

and so they form a so-called binomial convolution sequence of polynomials.

If $p \geq 1$ is an integer, with $\sigma(1)_k^{*p} := (\sigma(1)^{*p})_k$, $\mathbf{k}_p := (k_1, \ldots, k_p)$ in $\mathbb{N}_0^p$, $|\mathbf{k}_p| := k_1 + \cdots + k_p$ and $\mathbb{N}_0 := \{0, 1, 2, \ldots\}$

$$\sigma_k(p) = \sigma(1)_k^{*p} = \sum_{\mathbf{k}_p \in \mathbb{N}_0^p : |\mathbf{k}_p| = k} \binom{k}{k_1 \ldots k_p} \prod_{q=1}^{p} \sigma_{k_q}(1).$$

We clearly have

$$\sigma_k(p) = \sum_{q=1}^{p} \binom{p}{q} \sum_{\mathbf{k}_q \in \mathbb{N}^q : |\mathbf{k}_p| = k} \binom{k}{k_1 \ldots k_q} \prod_{r=1}^{q} \sigma_{k_r}(1).$$

In other words,

$$\sigma_k(p) = \sum_{q=1}^{k} \binom{p}{q} \sum_{\mathbf{k}_q \in \mathbb{N}^q : |\mathbf{k}_p| = k} \binom{k}{k_1 \ldots k_q} \prod_{r=1}^{q} \sigma_{k_r}(1), \tag{7}$$

where $\binom{p}{q} = 0$ if $q > p$. This expression extends to non-integral arguments $\theta > 0$ of $\sigma_k(\cdot)$ as

$$\sigma_k(\theta) =: \sigma(1)_k^{*\theta} = \sum_{q=1}^{k} \binom{\theta}{q} \sum_{\mathbf{k}_q \in \mathbb{N}^q :\, |\mathbf{k}_p| = k} \binom{k}{k_1 \ldots k_q} \prod_{r=1}^{q} \sigma_{k_r}(1), \tag{8}$$

where $\binom{\theta}{q} =: \{\theta\}_q / q!$ with $\{\theta\}_q := \Gamma(\theta+1)/\Gamma(\theta-q+1) = \theta(\theta-1)..(\theta-q+1)$, the usual extension of $\binom{p}{q}$ for the expansion of $(1+x)^\theta$. From (8), it is clear again that $\sigma_k(\theta)$ is a degree-$k$ polynomial in $\theta$ with no constant term. This expression should be used instead of (5) whenever the values at $\theta = 1$ of $\sigma_k(\cdot)$ are available in the first place, instead of the $\phi_\bullet$.

(iii) Putting the expression of $\sigma_k(\theta)$ in (5) into the recurrence equation (3) which $(\sigma_k(\theta);\, k \geq 1)$ satisfies gives

$$l \cdot B_{k,l}(\phi_\bullet) = \sum_{j=l-1}^{k-1} \binom{k}{j} \phi_{k-j} B_{j,l-1}(\phi_\bullet). \tag{9}$$

Recalling the boundary conditions

$$B_{k,0}(\phi_\bullet) = B_{0,l}(\phi_\bullet) = 0, \quad k, l \geq 1 \quad \text{and} \quad B_{0,0}(\phi_\bullet) := 1,$$

we get

$$B_{k,1}(\phi_\bullet) = \phi_k \quad \text{and} \quad B_{k,k}(\phi_\bullet) = \phi_1^k. \tag{10}$$

(iv) While performing the substitution $\theta \to 1/\theta$, $\sigma_k(\theta)$ should be mapped into the new polynomial with respect to $1/\theta$

$$\sigma_k(1/\theta) = \theta^{-(k+1)} \sum_{l=1}^{k} B_{k,k-l+1}(\phi_\bullet)\theta^l,$$

involving the 'reversed' Bell sequence $B_{k,k-l+1}(\phi_\bullet)$.

### 2.2 Discrete Compound Poisson Distributions Arising from $Z_\theta(x)$

Let now $\xi \in \mathbb{N}_0$ be a discrete random variable whose probability generating (pgf) is given by:

$$\Phi(u) := \mathbf{E}\big[u^\xi\big] = \frac{Z_\theta(xu)}{Z_\theta(x)}, \quad |u| \leq 1.$$

Since

$$\mathbf{E}\big[u^{\xi}\big] = e^{-\theta\phi(x)(1-\frac{\phi(xu)}{\phi(x)})}, \tag{11}$$

$\xi$ is in the compound Poisson (CP) class, as a Poisson sum of iid jumps, hence infinitely divisible. The jumps' height law is given by its pgf $\mathbf{E}[u^{\delta}] = \frac{\phi(xu)}{\phi(x)}$, where $\delta \in \mathbb{N}$ is one of these jumps. Note that both $\mathbf{E}[\delta] = x\frac{\phi'(x)}{\phi(x)}$ and $\mathbf{E}[\xi] = \theta\phi(x)\mathbf{E}[\delta] = \theta x\phi'(x)$ are finite when $|x| < x_0$. Clearly

$$\mathbf{P}(\delta = m) = \frac{\phi_m x^m}{\phi(x)\cdot m!}, \quad m \geq 1 \quad \text{and}$$

$$\mathbf{P}(\xi = k) = \frac{\sigma_k(\theta)x^k}{Z_\theta(x)\cdot k!}, \quad k \geq 0.$$

With $y$ defined by $x =: e^{-y}$, $y$ is indeed the Legendre conjugate of $\mu := \mathbf{E}(\xi)$. So the parameter $x$ in (11) can serve to adjust the mean $\mu$ of $\xi$. The random variable $\xi$ will be used in the sequel as the abundance of some species in a population with $n$ species. Due to its compound Poisson structure, it is tacitly assumed that species abundance is modelled as a Poisson sum of iid 'clusters' each with random size distributed like $\delta \geq 1$.

Consider now a sequence $\boldsymbol{\xi} := (\xi_1, \ldots, \xi_n, \ldots)$ of iid compound Poisson random variables, each on $\mathbb{N}_0$. Let $\zeta_n := \sum_{m=1}^{n} \xi_m$ denote their partial sum. Then, because $\xi$ is in the compound-Poisson class due to $Z_\theta(x)^n = Z_{n\theta}(x)$

$$\mathbf{P}(\zeta_n = k) = \frac{\sigma_k(n\theta)x^k}{Z_{n\theta}(x)\cdot k!}, \quad k \geq 0.$$

This is also a compound Poisson distribution with corresponding partition function $Z_{n\theta}(x)$.

*Remark* One could think of starting with $\phi(x) := \phi_0 + \sum_{m\geq 1}\frac{\phi_m}{m!}x^m$ with $\phi_0 \geq 0$ but because we shall deal with CP distributions whose pgfs are given by (11), $\phi_0$ plays no role in our problem.

### 2.3 Sampling from Infinitely Divisible CP Distributions

Define a random allocation scheme of $k$ distinguishable particles or balls into $n$ distinguishable boxes by

$$\mathbf{K}_{n,k} := \big(K_{n,k}(1), \ldots, K_{n,k}(n)\big) \stackrel{d}{=} (\xi_1, \ldots, \xi_n \mid \zeta_n = k),$$

so that $K_{n,k}(m)$ counts the number of particles in box $m$, $m = 1, \ldots, n$ in a $k$-sample. Defining $\mathbf{K}_{n,k}$ from $n$ iid $\xi$'s conditioned on summing to $k$, we get the generalized allocation scheme defined by Kolchin, (see [37]). When the $\xi$'s are in addition CP distributed, we call this model sampling from Gibbs-Poisson (GP) distributions.

*Remark* Since $\mathbf{E}[\xi] = \Phi'(1) = \theta x\phi'(x)$, $\theta > 0$ and $x \in (0, x_0)$, we could adjust the mean $\mu$ of $\xi$ so that $\mathbf{E}[\xi] = \mu$. Then we would have the relation $\mu/\theta = x\phi'(x)$ (Legendre conjugation of $x$ and $\mu$) from which, by Lagrange inversion formula, an expression of $x$ as a function of $\mu/\theta$ would follow. However, as we shall see, the actual value of the mean $\mu$ does not really matter after the sampling process.

Taking now into account the conditioning on the sample size in the definition of $\mathbf{K}_{n,k}$'s law, with $\mathbf{k}_n := (k_1, \dots, k_n) \in \mathbb{N}_0^n$ a vector of non-negative integers obeying $|\mathbf{k}_n| := \sum_{m=1}^n k_m = k$

$$\mathbf{P}(\mathbf{K}_{n,k} = \mathbf{k}_n) = \frac{\mathbf{P}(\xi_1 = k_1, \dots, \xi_n = k_n)}{\mathbf{P}(\zeta_n = k)} = \frac{1}{\sigma_k(n\theta)} \binom{k}{k_1 \dots k_n} \prod_{m=1}^n \sigma_{k_m}(\theta), \tag{12}$$

this (Maxwell-Boltzmann) joint law being independent of $x$ and so of the mean $\mu$ of the $\xi$'s. In other words, the joint probability generating function of $\mathbf{K}_{n,k}$ reads ($|u_m| \le 1; m = 1, \dots, n$):

$$\mathbf{E}\left[\prod_{m=1}^n u_m^{K_{n,k}(m)}\right] = \frac{1}{\sigma_k(n\theta)} \sum_{\mathbf{k}_n \in \mathbb{N}_0^n:\, |\mathbf{k}_n| = k} \binom{k}{k_1 \dots k_n} \prod_{m=1}^n \sigma_{k_m}(\theta) u_m^{k_m}. \tag{13}$$

From (12), $w_{k_m}(\theta) := \sigma_{k_m}(\theta)/k_m!$ is seen to be the Boltzmann weight of box $m$ with $e_{k_m}(\theta) := -\log(\sigma_{k_m}(\theta)/k_m!)$ being the energy required to put $k_m$ balls into box number $m$. More precisely, for our random allocation GP model of particles (13) and from (5), the price to pay for having the $l$th particle, $l \in \{1, \dots, k_m\}$, in box $m$ simply is $l$ and this event is assigned the weight $B_{k_m,l}(\phi_\bullet)/k_m!$. From this, one may view $\theta$ as a box temperature parameter which, under our assumptions, is here common to all boxes (or species). Due to $\sigma_{k_m}(\theta)$ being a polynomial in $\theta$ with positive coefficients, the energy $e_{k_m}(\theta)$ is indeed a decreasing function of $\theta$ and one may therefore interpret $\theta$ as some temperature[1] of the boxes (maybe through the monotone transformation $\theta \leftrightarrow e^{-1/T}$). Note that when $\theta$ approaches 0, the energy $e_{k_m}(\theta) \sim -\log\theta$ tends to $+\infty$: because the price to pay to put any number of particles into a box is extremely high, the optimal strategy is to put them all into a single box. One therefore expects that, as $\theta$ gets very small, the vector $\mathbf{K}_{n,k}$ gets very skewed (most balls into a single box), that is, completely opposite to the balanced multinomial $(k; \frac{1}{n}, \dots, \frac{1}{n})$ situation

$$\mathbf{P}(\mathbf{K}_{n,k} = \mathbf{k}_n) = \frac{k!}{\prod_{m=1}^n k_m!} n^{-k}, \quad |\mathbf{k}_n| = k,$$

which is obtained for $\theta \to \infty$, as a result of $\sigma_{k_m}(\theta) \sim (\phi_1\theta)^{k_m}$. As a conclusion, smaller the values of $\theta$, the more likely it is that the occupancy numbers $K_{n,k}(m)$ are disparate.

From (12), the random vector-count $\mathbf{K}_{n,k}$ has exchangeable distribution (invariance under any permutation of the boxes numbers). But obviously, in the ordered version $\mathbf{K}_{(n),k}$ of the box occupancies $\mathbf{K}_{n,k}$, say with $K_{(n),k}(1) \ge \dots \ge K_{(n),k}(n)$, the boxes are not equally filled and so $\mathbf{K}_{(n),k}$ is not exchangeable.

– Let us now compute the distribution of one of its typical components, say $K_{n,k}(1)$. With $l \in \{0, \dots, k\}$, we get

$$\begin{aligned}\mathbf{P}\big(K_{n,k}(1) = l\big) &= \mathbf{P}(\xi_1 = l)\frac{[u^{k-l}]\Phi(u)^{n-1}}{[u^k]\Phi(u)^n} \\ &= \frac{\sigma_l(\theta)x^l}{l!}\frac{[u^{k-l}]Z_\theta(xu)^{n-1}}{[u^k]Z_\theta(xu)^n} = \binom{k}{l}\frac{\sigma_l(\theta)\sigma_{k-l}((n-1)\theta)}{\sigma_k(n\theta)}.\end{aligned}$$

Note that $\sum_{l=0}^k \mathbf{P}(K_{n,k}(1) = l) = 1$, as required, in view of (6) with $\theta' = (n-1)\theta$.

[1] In statistical contexts, this temperature parameter is also called the concentration parameter.

– Proceeding similarly, with $l \in \{0, \dots, k\}$, we would obtain the law of the partial sums $K_{n,k}(1) + \cdots + K_{n,k}(m)$, $m < n$, as

$$\mathbf{P}\big(K_{n,k}(1) + \cdots + K_{n,k}(m) = l\big) = \binom{k}{l} \frac{\sigma_l(m\theta)\sigma_{k-l}((n-m)\theta)}{\sigma_k(n\theta)}.$$

As required also, $\sum_{l=0}^{k} \mathbf{P}(K_{n,k}(1) + \cdots + K_{n,k}(m) = l) = 1$, as a result of $\sigma_k(\theta)$ being a convolution sequence of polynomials, from (6).

– Finally, define $\{k\}_l := k(k-1)\cdots(k-l+1)$ with $\{k\}_0 := 1$ and let us now consider the falling factorial moments of $\mathbf{K}_{n,k}$.

Fix $\mathbf{l}_n := (l_1, \dots, l_n) \in \mathbb{N}_0^n$ summing to $l \le k$. We have

$$\mathbf{E}\left[\prod_{m=1}^{n} \big\{K_{n,k}(m)\big\}_{l_m}\right] = \prod_{m=1}^{n} l_m! \frac{[v^k]\prod_{m=1}^{n}[v_m^{l_m}]Z_\theta(xv(v_m+1))}{[v^k]Z_{n\theta}(xv)}.$$

Since $l_m![v_m^{l_m}]Z_\theta(xv(v_m+1)) = \sum_{k_m \ge l_m} \frac{\sigma_{k_m}(\theta)\cdot(xv)^{k_m}}{(k_m - l_m)!}$, with $\mathbf{k}_n$ summing to $|\mathbf{k}_n| = k$, with $\mathbf{k}_n \ge \mathbf{l}_n$ meaning $k_1 \ge l_1, \dots, k_n \ge l_n$, we get

$$\mathbf{E}\left[\prod_{m=1}^{n} \big\{K_{n,k}(m)\big\}_{l_m}\right] = \frac{\sum_{\mathbf{k}_n \ge \mathbf{l}_n} \prod_{m=1}^{n} \sigma_{k_m}(\theta)/(k_m - l_m)!}{\sigma_k(n\theta)/k!}. \tag{14}$$

These combinatorial quantities arise in the following resampling problem:

*Subsampling Without Replacement from* $\mathbf{K}_{n,n}$ Suppose $K_{n,n}(m)$, $m = 1, \dots, n$ are the random box occupancies of some sample with size exactly equal to the number $n$ of boxes, generated by some compound-Poisson vector $\boldsymbol{\xi}_n := (\xi_1, \dots, \xi_n)$. So there are at most $n$ boxes filled by a singleton as a result of $\sum_{m=1}^{n} K_{n,n}(m) = n$. Let $p \le k \le n$. In connection with the theory of compound-Poisson coalescent processes, [31], we are interested in the event that after a random $k$-subsampling without replacement from $\mathbf{K}_{n,n}$, balls are reassigned at random into boxes so as to end up in a new occupancy $\mathbf{K}'_{n,k} := (K'_{n,k}(q); q = 1, \dots, p)$ where only a fixed number $p$ of the random number $\Pi_{n,k}$ of filled boxes (labeled in arbitrary order) are occupied. So $\mathbf{K}'_{n,k}$ obeys $\sum_{q=1}^{p} K'_{n,k}(q) = k$ and $K'_{n,k}(q) \ge 1$. Then, with $(k_1, \dots, k_p) \in \mathbb{N}^p$ summing to $k$, the sampling without replacement strategy yields:

$$\begin{aligned}\mathbf{P}\big(K'_{n,k}(1) = k_1, .., K'_{n,k}(p) = k_p; \Pi_{n,k} = p\big) &= \binom{n}{p}\binom{k}{k_1..k_p}\frac{\mathbf{E}(\prod_{q=1}^{p}\{K_{n,n}(q)\}_{k_q})}{\{n\}_k} \\ &= \frac{\binom{n}{p}}{\binom{n}{k}}\mathbf{E}\prod_{q=1}^{p}\binom{K_{n,n}(q)}{k_q}.\end{aligned}$$

Summing over $(k_1, \dots, k_p) \in \mathbb{N}^p$

$$\mathbf{P}(\Pi_{n,k} = p) = \frac{\binom{n}{p}}{\{n\}_k} \sum_{\mathbf{k}_p \in \mathbb{N}^p : |\mathbf{k}_p| = k} \binom{k}{k_1 \dots k_p} \mathbf{E}\left(\prod_{q=1}^{p} \big\{K_{n,n}(q)\big\}_{k_q}\right)$$

is the probability that in a $k$-subsampling without replacement from $\mathbf{K}_{n,n}$ exactly $p \le k \le n$ boxes will be filled. Using (14), with $\mathbf{k}_p := (k_1, \dots, k_p) \in \mathbb{N}^p$ a vector of positive integers satisfying $|\mathbf{k}_p| := \sum_{q=1}^{p} k_q = k$, we have

$$\mathbf{E}\left[\prod_{q=1}^{p}\{K_{n,n}(q)\}_{k_q}\right] = \frac{\sum_{\mathbf{l}_p \in \mathbb{N}_0^p} \prod_{q=1}^{p} \sigma_{k_q+l_q}(\theta)/l_q!}{\sigma_n(n\theta)/n!},$$

and the full expression of the probabilities $\mathbf{P}(\Pi_{n,k} = p)$ can be obtained in terms of the original weights $w_k(\theta) = \sigma_k(\theta)/k!$.

These questions arise in the discrete theory of compound-Poisson coalescent processes. Suppose $\mathbf{K}_{n,n}$ is the random exchangeable reproduction law of some Markov branching process preserving the total number $n$ of individuals over the subsequent generations, [31]. That is, independently in each generation, individual number $m$ produces $K_{n,n}(m)$ offspring, $m = 1, \ldots, n$ and $\sum_{m=1}^{n} K_{n,n}(m) = n$.

We first wish to count, forward in time, the number of descendants of any size-$m$ subsample of the full population with size $n$, defining thereby a discrete-time Markov chain. Clearly, the $(m, l)$ entry of the transition matrix of this Markov process on the state-space $\{0, \ldots, n\}$ is

$$\mathbf{P}\big(K_{n,n}(1) + \cdots + K_{n,n}(m) = l\big) = \binom{n}{l}\frac{\sigma_l(m\theta)\sigma_{n-l}((n-m)\theta)}{\sigma_n(n\theta)}, \quad m, l \in \{0, \ldots, n\},$$

looking at the descent of all size-$m$ subsamples. For this Markov chain, clearly, the states $\{0, n\}$ are both absorbing.

Looking now at this branching process backward in time, individuals are seen to merge, giving rise to some ancestral coalescent process where individuals are identified if they share a common ancestor one generation backward in time. The process stops when a single individual is present (at time to their most recent common ancestor).

The quantity $\mathbf{P}(K'_{n,k}(1) = k_1, \ldots, K'_{n,k}(p) = k_p; \Pi_{n,k} = p)$ is then the probability that a one-step back $(k_1, \ldots, k_p)$ to $p$ merger for a subsample of size $k$ occurs in this ancestral process. The lower-triangular stochastic matrix $\mathcal{Q}_{k,p}^{(n)} := \mathbf{P}(\Pi_{n,k} = p)$ is the transition matrix of this pure death coalescent Markov process on $\{0, \ldots, n\}$, with both states $\{0, 1\}$ absorbing. The forward and backward Markov processes are easily seen to be duals in the sense and for the duality kernel defined in [39].

*Number of Filled Boxes in* $\mathbf{K}_{n,k}$ With $\mathbf{I}(\cdot)$ denoting the indicator function, let now $P_{n,k} := \sum_{m=1}^{n} \mathbf{I}(K_{n,k}(m) > 0)$ count the number of non empty boxes in the sampling process from $\xi_n$. With $1 \le p \le n \wedge k$, the probability that there are only $P_{n,k} = p \in [n]$ visited boxes in the sampling process, the $n - p$ remaining ones remaining empty, is easily obtained as follows: In the event $P_{n,k} = p$, for any fixed subset $(m_1, \ldots, m_p)$ of $p$ different box numbers and each $\mathbf{k}_p = (k_1, \ldots, k_p)$ in $\mathbb{N}^p$ summing to $k$, we have from (12)

$$\mathbf{P}\big(\big(K_{n,k}(m_q) = k_q, q = 1 \ldots, p\big); P_{n,k} = p\big) = \frac{k!}{\sigma_k(n\theta)}\prod_{q=1}^{p}\frac{\sigma_{k_q}(\theta)}{k_q!}.$$

The above probability is independent of the $\binom{n}{p}$ different subsets $(m_1, \ldots, m_p)$. Denote by $\{L_1, .., L_p\}$ the random subset of indexes of the occupied $p$ boxes in the event $P_{n,k} = p$. From the above argument, we get

$$\mathbf{P}\big(\big\{K_{n,k}(L_q), q = 1, \ldots, p\big\} = \{k_q, q = 1, \ldots, p\}; P_{n,k} = p\big) = \binom{n}{p}\frac{k!}{\sigma_k(n\theta)}\prod_{q=1}^{p}\frac{\sigma_{k_q}(\theta)}{k_q!},$$

where $\{K_{n,k}(L_q), q=1,\ldots,p\}=\{k_q, q=1,\ldots,p\}$ is an equality of multisets (the multisets are needed to keep in mind the repetitions that could exist in $k_q$, $q=1,..,p$). Letting $\widehat{K}_{n,k}(q):=K_{n,k}(L_q)$, $q=1,..,p$, the last equality will simply be written as

$$\mathbf{P}\big(\widehat{K}_{n,k}(1)=k_1,\ldots,\widehat{K}_{n,k}(p)=k_p;\,P_{n,k}=p\big)=\binom{n}{p}\frac{k!}{\sigma_k(n\theta)}\prod_{q=1}^{p}\frac{\sigma_{k_q}(\theta)}{k_q!}. \tag{15}$$

This is the probability that there are $p\in[n]$ non-empty boxes labeled in arbitrary way and that $(k_1,\ldots,k_p)$ are their respective occupancies. Note that

$$\mathcal{P}_{k,p}^{(n)}:=\mathbf{P}(P_{n,k}=p)=\binom{n}{p}\frac{k!}{\sigma_k(n\theta)}\sum_{\mathbf{k}_p\in\mathbb{N}^p:\,|\mathbf{k}_p|=k}\prod_{q=1}^{p}\frac{\sigma_{k_q}(\theta)}{k_q!}$$

is the probability that in a $k$-sample from $n$ species with abundance $\boldsymbol{\xi}_n$, the exact number of distinct observed species is $p$. In particular, $\mathcal{P}_{k,1}^{(n)}:=n\frac{\sigma_k(\theta)}{\sigma_k(n\theta)}$ is the probability that in this $k$-sample, only one species is discovered (whichever it is).

The lower-triangular stochastic matrix with $(k,p)$ entries $\mathcal{P}_{k,p}^{(n)}:=\mathbf{P}(P_{n,k}=p)$ is the transition matrix of some other pure death Markov process on $\{0,\ldots,n\}$ which does not coincide in general with the coalescent transition matrix $\mathcal{Q}_{k,p}^{(n)}$ defined in the latter paragraph (in fact both transition matrices match iff $\xi$ is negative binomial distributed, see [32]).

The expression (15) turns out to be the canonical Gibbs distribution on finite size-$n$ partitions of $k$ into $p$ distinct clusters (the filled boxes), derived from the weight sequence $\phi_\bullet$. In this language, the normalizing quantity $\sigma_k(n\theta)/k!$ is called the canonical Gibbs partition function.

Now, from (15), with $\{n\}_p:=n!/(n-p)!$

$$\mathbf{P}(P_{n,k}=p)=\frac{\{n\}_p}{\sigma_k(n\theta)}B_{k,p}\big(\sigma_\bullet(\theta)\big),\qquad p\in\{1,\ldots,n\wedge k\}, \tag{16}$$

where

$$B_{k,p}\big(\sigma_\bullet(\theta)\big):=\frac{k!}{p!}\sum_{\mathbf{k}_p\in\mathbb{N}^p:\,|\mathbf{k}_p|=k}\prod_{q=1}^{p}\frac{\sigma_{k_q}(\theta)}{k_q!}=\frac{k!}{p!}\big[x^k\big]\big(Z_\theta(x)-1\big)^p \tag{17}$$

is now a Bell polynomial in the polynomial variables $\sigma_\bullet(\theta):=(\sigma_1(\theta),\sigma_2(\theta),\ldots)$.

Conditioning the canonical Gibbs distribution on the number of filled cells being equal to $p$ yields the corresponding micro-canonical distribution as

$$\begin{aligned}&\mathbf{P}\big(\widehat{K}_{n,k}(1)=k_1,\ldots,\widehat{K}_{n,k}(p)=k_p\mid P_{n,k}=p\big)\\&\quad=\frac{k!}{p!}\frac{1}{B_{k,p}(\sigma_\bullet(\theta))}\prod_{q=1}^{p}\frac{\sigma_{k_q}(\theta)}{k_q!}.\end{aligned}$$

The new normalizing constant $B_{k,p}(\sigma_\bullet(\theta))/k!$ may be called the microcanonical partition function.

The microcanonical distribution is independent of $n$. So, for all models studied here, $P_{n,k}$ is a sufficient statistic in the estimation of $n$ problem from occupancy data (assuming $\theta$ known).

Let us now give some additional details on the distribution of $P_{n,k}$.

**Proposition 1**

(a) *Assume $k \geq n$. The probability generating function of $P_{n,k}$ is given by*

$$\mathbf{E}\big(u^{P_{n,k}}\big) = \sum_{p=0}^{n-1} \binom{n}{p} u^{n-p}(1-u)^p \frac{\sigma_k((n-p)\theta)}{\sigma_k(n\theta)}, \tag{18}$$

*with*:

$$\mathbf{P}(P_{n,k} = p) = \binom{n}{p} \sum_{q=1}^{p} (-1)^{p-q} \binom{p}{q} \frac{\sigma_k(q\theta)}{\sigma_k(n\theta)}, \quad p \in \{1, \ldots, n\}. \tag{19}$$

*In addition*,

$$\mathbf{E}(P_{n,k}) = n\left(1 - \frac{\sigma_k((n-1)\theta)}{\sigma_k(n\theta)}\right)$$

$$\mathrm{Var}(P_{n,k}) = n\left(\frac{\sigma_k((n-1)\theta)}{\sigma_k(n\theta)} + (n-1)\frac{\sigma_k((n-2)\theta)}{\sigma_k(n\theta)} - n\left(\frac{\sigma_k((n-1)\theta)}{\sigma_k(n\theta)}\right)^2\right).$$

(b) *If $k < n$*, (18) *and* (19) *still hold, but now with a modified support for $P_{n,k}$'s law*:

$$\mathbf{P}(P_{n,k} = p) = \binom{n}{p} \sum_{q=1}^{p} (-1)^{p-q} \binom{p}{q} \frac{\sigma_k(q\theta)}{\sigma_k(n\theta)}, \quad p \in \{1, \ldots, k\}. \tag{20}$$

*Proof*

(a) This follows from $B_{k,p}(\sigma_\bullet(\theta)) = \frac{k!}{p!}[x^k](Z_\theta(x) - 1)^p$. Indeed, from (16)

$$\begin{aligned}
\mathbf{E}\big(u^{P_{n,k}}\big) &= \sum_{p=0}^{n} u^p \{n\}_p \frac{B_{k,p}(\sigma_\bullet(\theta))}{\sigma_k(n\theta)} = \frac{k!}{\sigma_k(n\theta)} \sum_{p=0}^{n} \binom{n}{p} \big[x^k\big]\big(u\big(Z_\theta(x) - 1\big)\big)^p \\
&= \frac{k!}{\sigma_k(n\theta)} \big[x^k\big]\big(1 - u + uZ_\theta(x)\big)^n \\
&= \frac{k!}{\sigma_k(n\theta)} \sum_{p=0}^{n} \binom{n}{p} u^{n-p}(1-u)^p \big[x^k\big] Z_\theta(x)^{n-p} \\
&= \sum_{p=0}^{n-1} \binom{n}{p} u^{n-p}(1-u)^p \frac{\sigma_k((n-p)\theta)}{\sigma_k(n\theta)}.
\end{aligned}$$

The alternating sum expression of $\mathbf{P}(P_{n,k} = p)$ follows from extracting $[u^p]\mathbf{E}(u^{P_{n,k}})$ and the mean and variance of $P_{n,k}$ from the evaluations of the first and second derivatives of $\mathbf{E}(u^{P_{n,k}})$ with respect to $u$ at $u = 1$.

(b) Follows from similar considerations. Indeed, in principle, we should start with $\mathbf{E}(u^{P_{n,k}}) = \sum_{p=0}^{k} u^p \{n\}_p \frac{B_{k,p}(\sigma_\bullet(\theta))}{\sigma_k(n\theta)}$ where the $p$-sum now stops at $p = k = k \wedge n$. But the upper bound of this $p$-sum can be extended to $n$ because $B_{k,p}(\sigma_\bullet(\theta)) = 0$ if $p > k$. □

In the particular case discussed below when $\sigma_k(\theta) = \theta(\theta+1)\cdots(\theta+k-1)$ (Ewens-Dirichlet model), these results can be found in [33].

In (16), the new combinatorial coefficients $B_{k,p}(\sigma_\bullet(\theta))$ come into the game. They are given by

**Corollary 2** *With $S_{l,p}$ the second kind Stirling numbers,*

$$B_{k,p}\big(\sigma_\bullet(\theta)\big) = \sum_{l=p}^{k} B_{k,l}(\phi_\bullet) S_{l,p}\theta^l = \theta^p \cdot \sum_{l=0}^{k-p} B_{k,p+l}(\phi_\bullet) S_{l+p,p}\theta^l,$$

*showing that $B_{k,p}(\sigma_\bullet(\theta))$ is itself a polynomial in $\theta$ with larger (smaller) degree $k$ (respectively $p$).*

*Proof* From (16) and (19), we have[2]

$$B_{k,p}\big(\sigma_\bullet(\theta)\big) = \frac{1}{p!}\sum_{q=1}^{p}(-1)^{p-q}\binom{p}{q}\sigma_k(q\theta).$$

Recalling $\sigma_k(\theta) = \sum_{l=1}^{k}\theta^l B_{k,l}(\phi_\bullet)$ and observing $S_{l,p} = \sum_{q=1}^{p}(-1)^{p-q}\binom{p}{q}q^l$ gives the result after reversing the sums. This result actually is in accordance with the Faa di Bruno formula (see [10]) giving the Taylor coefficients of the composition function $g$ of the two analytic functions $g(x) := e_{\lambda,\theta} \circ \phi(x)$ where $e_{\lambda,\theta}(x) := e^{\lambda(e^{\theta x}-1)}$ as

$$S_k(\lambda) = \sum_{l=1}^{k} e_l(\theta,\lambda) B_{k,l}(\phi_\bullet),$$

with $e_l(\theta,\lambda) = \theta^l \sum_{p=1}^{l}\lambda^p S_{l,p}$ the $l$th Taylor coefficient of $e_{\lambda,\theta}(x)$. Clearly indeed,

$$g(x) = e^{\lambda(Z_\theta(x)-1)} = 1 + \sum_{k\geq 1}\frac{x^k}{k!}S_k(\lambda) =: 1 + \sum_{k\geq 1}\frac{x^k}{k!}\left(\sum_{p=1}^{k}\lambda^p B_{k,p}\big(\sigma_\bullet(\theta)\big)\right)$$

and the $\lambda^p$-coefficient of $S_k(\lambda)$ is exactly $\sum_{l=p}^{k} B_{k,l}(\phi_\bullet) S_{l,p}\theta^l$. $\square$

### 2.4 The Estimation of $n$ Problem

Let us now discuss the important question of estimating the unknown number of species $n$ based on the data $k$ and $P$ (assuming $\theta$ is known), recalling $\mathbf{P}(P_{n,k} = P)$ is a sufficient statistic in this estimation problem. Our forthcoming statement holds for a class of $\phi$ which is such that the degree-$k$ polynomial $\sigma_k(\theta) \in ZR_-$ (has only real non-positive zeroes). We recall that $\sigma_k(\theta) \in ZR_-$ iff the matrix $M$ with entries $M_{i,j} = B_{k,i-j}(\phi_\bullet)$, $i, j = 0, \ldots, k$, with $B_{k,l}(\phi_\bullet) = 0$ if $l \notin \{l : B_{k,l}(\phi_\bullet) > 0\}$ is totally positive of order $k$ (with $l = 1, \ldots, k$, each $l \times l$ minor of $M$ has a nonnegative determinant), [49]. Therefore, there is no simple way to check whether or not $\sigma_k(\theta) \in ZR_-$.

We also recall here, [49], that if and only if the matrix $M = M_{i,j}$ would be such that all its $2 \times 2$ minors have a nonnegative determinant, then the sequence $B_{k,l}(\phi_\bullet)$, $l = 1, \ldots, k$ (with no internal zeros) is $l$-log-concave (the $l$-sequence $B_{k,l}(\phi_\bullet)$ is a Pòlya frequency sequence of order 2). If this is the case, we shall say $\sigma_k(\theta) \in PF_2$.

[2] This identity was derived in a different way in [53].

**Proposition 3** *Suppose $\sigma_k(\theta) \in ZR_-$. Then the log-likelihood* $\log \mathbf{P}(P_{n,k} = P)$ *attains its maximum in $n$ at least once and at most twice in which latter case, the two values are adjacent integers. This leads to the maximum likelihood estimator $\widehat{n}$ of $n$ characterized by*:

$$\widehat{n} = \sup\left\{n > 0 : \frac{\mathbf{P}(P_{n,k} = P)}{\mathbf{P}(P_{n-1,k} = P)} > 1\right\}.$$

*When the set of integers* $\{n > 0 : \frac{\mathbf{P}(P_{n,k}=P)}{\mathbf{P}(P_{n-1,k}=P)} > 1\}$ *is empty*, $\widehat{n} = P$.

*When this is not the case and for large $n$, an approximation of the estimator $\widehat{n}$ of $n$ is given by the implicit equation*

$$P = \widehat{n}\left(1 - \frac{\sigma_k((\widehat{n}-1)\theta)}{\sigma_k(\widehat{n}\theta)}\right).$$

*Proof* We extend (16) to $n$ a real variable, so we can differentiate $\log \mathbf{P}(P_{n,k} = P)$ with respect to $n > P$. In this domain, we have $\partial_n \log\{n\}_P = \sum_{q=0}^{P-1} \frac{1}{n-q}$, and so we get

$$\partial_n \log \mathbf{P}(P_{n,k} = P) = \sum_{q=0}^{P-1} \frac{1}{n-q} - \partial_n \log \sigma_k(n\theta).$$

Suppose the polynomial $\sigma_k(\theta) \in ZR_-$ has zeroes $-r_{l,k}$ where: $0 = r_{1,k} \le \cdots \le r_{k,k}$. Then $\sigma_k(n\theta) = \prod_{l=1}^{k}(n\theta + r_{l,k})$ and $\partial_n \log \sigma_k(n\theta) = \sum_{l=1}^{k}(n + r_{l,k}/\theta)^{-1}$, together with $\partial_n^2 \log \sigma_k(n\theta) = -\sum_{l=1}^{k}(n + r_{l,k}/\theta)^{-2} < 0$.

If $\sum_{q=0}^{P-1} \frac{1}{n-q} - \sum_{l=1}^{k}(n + r_{l,k}/\theta)^{-1} \overset{(*)}{=} 0$ , then

$$\partial_n^2 \log \mathbf{P}(P_{n,k} = P) = -\sum_{q=0}^{P-1} \frac{1}{(n-q)^2} + \sum_{l=1}^{k}(n + r_{l,k}/\theta)^{-2} < 0,$$

showing that the likelihood is log-concave around the critical points. Hence, if $\widehat{n}$ solves $(*)$ it is a local maximum and there is no local minimum. The maximum likelihood estimator of real $n$ is thus unique.

Coming back to $n$ integer, we deduce that the maximum likelihood estimator of $n$ is the integer $\sup\{n > 0 : \frac{\mathbf{P}(P_{n,k}=P)}{\mathbf{P}(P_{n-1,k}=P)} > 1\}$. When $n$ is large, it may thus be approximated by $\frac{\mathbf{P}(P_{\widehat{n},k}=P)}{\mathbf{P}(P_{\widehat{n}-1,k}=P)} = 1$, leading to

$$\frac{\{\widehat{n}\}_P \sigma_k((\widehat{n}-1)\theta)}{\{\widehat{n}-1\}_P \sigma_k(\widehat{n}\theta)} = 1 \quad \text{or} \quad P = \widehat{n}\left(1 - \frac{\sigma_k((\widehat{n}-1)\theta)}{\sigma_k(\widehat{n}\theta)}\right).$$ □

*An Alternative Estimator* Let us now come to an alternative estimator of $n$ (see [33] for a similar approach in the particular context of the Dirichlet model given by $\phi(x) = -\alpha \log(1-x)$). Suppose that for all $\theta > 0$ and $k \ge 1$, $B_{k,p}(\sigma_\bullet(\theta))$ is a log-concave $p$-sequence (equivalently, each degree-$k$ $\lambda$-polynomial $S_k(\lambda) \in PF_2$). Then, by Darroch Theorem [12], $B_{k,p}(\sigma_\bullet(\theta))$ is $p$-unimodal or bimodal at two consecutive $p$. Because the $p$-sequence $\{n\}_p$ is also log-concave, $\{n\}_p B_{k,p}(\sigma_\bullet(\theta))$ is itself $p$-log-concave. For each $n$ therefore, there is a unique $\widetilde{p}$ defined as $\widetilde{p} = \sup\{p > 0 : \frac{\mathbf{P}(P_{n,k}=p)}{\mathbf{P}(P_{n,k}=p-1)} > 1\}$. Inverting the map $n \to \widetilde{p}(n)$, given $p = P$, there exists a unique $\widetilde{n}$, approximately characterized by

$\frac{\mathbf{P}(P_{\widetilde{n},k}=P-1)}{\mathbf{P}(P_{\widetilde{n},k}=P)}=1$, which can serve as an alternative estimator of $n$ given the data $(k,P)$. From (16), it is thus given by

$$\widetilde{n}=P+\frac{B_{k,P-1}(\sigma_\bullet(\theta))}{B_{k,P}(\sigma_\bullet(\theta))}.$$

If $k\geq n$, taking the expectation with respect to $P_{n,k}$, from (19), we have

$$\begin{aligned}\mathbf{E}(\widetilde{n}) &= \mathbf{E}(P_{n,k})+\sum_{p=1}^{n}\frac{B_{k,p-1}(\sigma_\bullet(\theta))}{B_{k,p}(\sigma_\bullet(\theta))}\frac{\{n\}_p}{\sigma_k(n\theta)}B_{k,p}\big(\sigma_\bullet(\theta)\big)\\ &= \mathbf{E}(P_{n,k})+\sum_{p=2}^{n}\frac{\{n\}_p B_{k,p-1}(\phi_\bullet)}{\sigma_k(n\theta)}=\mathbf{E}(P_{n,k})+\sum_{p=2}^{n}\big(n-(p-1)\big)\frac{\{n\}_{p-1}B_{k,p-1}(\phi_\bullet)}{\sigma_k(n\theta)}\\ &= \mathbf{E}(P_{n,k})+n\left(1-\frac{\{n\}_n B_{k,n}(\phi_\bullet)}{\sigma_k(n\theta)}\right)-\left(\mathbf{E}(P_{n,k})-n\frac{\{n\}_n B_{k,n}(\phi_\bullet)}{\sigma_k(n\theta)}\right)=n.\end{aligned}$$

So, when $k\geq n$, $\widetilde{n}$ is an unbiased estimator of $n$. The Fisher information of $n$ is

$$I(n)=-\mathbf{E}\big(\partial_n^2\log\mathbf{P}(P_{n,k}=P)\big)=\mathbf{E}\left(\sum_{q=0}^{P-1}\frac{1}{(n-q)^2}\right)-\sum_{l=1}^{k}(n+r_{l,k}/\theta)^{-2}>0,$$

giving the Cramér-Rao bound for the variance: $\mathrm{Var}(\widetilde{n})\geq I(n)^{-1}$.

### 2.5 Frequency of Frequencies

This suggests to look at the frequency of frequencies distribution problem. For $i=0,\dots,k$, let now

$$A_{n,k}(i)=\sum_{m=1}^{n}\mathbf{I}\big(K_{n,k}(m)=i\big) \tag{21}$$

count the number of boxes visited $i$ times by the $k$-sample, with $A_{n,k}(0)=n-P_{n,k}$, the number of empty boxes.

Let $(a_1,a_2,\dots)$ be non-negative integers satisfying $\sum_{i\geq1}a_i=p$ and $\sum_{i\geq1}ia_i=k$. It follows from (12) that

$$\mathbf{P}\big(A_{n,k}(1)=a_1,A_{n,k}(2)=a_2,\dots\big)=\frac{\{n\}_p\cdot k!}{\sigma_k(n\theta)}\prod_{i\geq1}\left\{\left(\frac{\sigma_i(\theta)}{i!}\right)^{a_i}\frac{1}{a_i!}\right\}. \tag{22}$$

Taking $A_{n,k}(0)$ into account, let $(a_0,a_1,\dots,a_k)$ be non-negative integers satisfying $\sum_{i=0}^{k}a_i=n$ and $\sum_{i=1}^{k}ia_i=k$. Then

$$\mathbf{P}\big(A_{n,k}(0)=a_0,A_{n,k}(1)=a_1,\dots,A_{n,k}(k)=a_k\big)=\frac{n!\cdot k!}{\sigma_k(n\theta)}\prod_{i=0}^{k}\left\{\left(\frac{\sigma_i(\theta)}{i!}\right)^{a_i}\frac{1}{a_i!}\right\}.$$

Note from this that, with $\sum_{i=1}^{k}ia_i=k$ and $\sum_1^k a_i\leq n$, the normalization condition gives the identity

$$\sum_{a_1,\dots,a_k}\frac{k!}{(n-\sum_1^k a_i)!}\prod_{i=1}^{k}\left\{\left(\frac{\sigma_i(\theta)}{i!}\right)^{a_i}\frac{1}{a_i!}\right\}=\frac{\sigma_k(n\theta)}{n!}. \tag{23}$$

From this, we get (see also [45] Sect. 1.5):

**Proposition 4** *If $p = n - a_0$, the joint distribution of $(A_{n,k}(1), \dots, A_{n,k}(k))$ and $P_{n,k}$ reads*

$$\mathbf{P}\big(A_{n,k}(1) = a_1, \dots, A_{n,k}(k) = a_k;\, P_{n,k} = p\big) = \frac{\{n\}_p \cdot k!}{\sigma_k(n\theta)} \prod_{i=1}^{k} \left\{ \left( \frac{\sigma_i(\theta)}{i!} \right)^{a_i} \frac{1}{a_i!} \right\}. \tag{24}$$

Let us compute the falling factorial moments of $A_{n,k}(i)$, $i = 1, \dots, k$.

**Proposition 5** *Let $r_i$, $i = 1, \dots, k$ be non-negative integers satisfying $\sum_1^k r_i = r \le n$ and $\sum_1^k i r_i = \kappa \le k$. We have*

$$\mathbf{E}\left[ \prod_{i=1}^{k} \{A_{n,k}(i)\}_{r_i} \right] = \{n\}_r \{k\}_\kappa \frac{\sigma_{k-\kappa}((n-r)\theta)}{\sigma_k(n\theta)} \prod_{i=1}^{k} \left( \frac{\sigma_i(\theta)}{i!} \right)^{r_i}. \tag{25}$$

*Proof*

$$\begin{aligned}
&\mathbf{E}\left[ \prod_{i=1}^{k} \{A_{n,k}(i)\}_{r_i} \right] \\
&= \frac{n! \cdot k!}{\sigma_k(n\theta)} \sum_{a_1,\dots,a_k} \frac{1}{(n - \sum_1^k a_i)!} \prod_{i=1}^{k} \left\{ \left( \frac{\sigma_i(\theta)}{i!} \right)^{a_i} \frac{1}{(a_i - r_i)!} \right\} \\
&= \frac{n! \cdot k!}{\sigma_k(n\theta)} \prod_{i=1}^{k} \left( \frac{\sigma_i(\theta)}{i!} \right)^{r_i} \sum_{a_1,\dots,a_k} \frac{1}{(n - \sum_1^k a_i)!} \prod_{i=1}^{k} \left\{ \left( \frac{\sigma_i(\theta)}{i!} \right)^{a_i - r_i} \frac{1}{(a_i - r_i)!} \right\}.
\end{aligned}$$

The normalization condition (23) gives:

$$\sum_{a_1,\dots,a_k} \frac{1}{(n - \sum_1^k a_i)!} \prod_{i=1}^{k} \left\{ \left( \frac{\sigma_i(\theta)}{i!} \right)^{a_i - r_i} \frac{1}{(a_i - r_i)!} \right\} = \frac{\sigma_{k-\kappa}((n-r)\theta)}{(n-r)! \cdot (k-\kappa)!}.$$

Finally, we get

$$\mathbf{E}\left[ \prod_{i=1}^{k} \{A_{n,k}(i)\}_{r_i} \right] = \{n\}_r \{k\}_\kappa \frac{\sigma_{k-\kappa}((n-r)\theta)}{\sigma_k(n\theta)} \prod_{i=1}^{k} \left( \frac{\sigma_i(\theta)}{i!} \right)^{r_i}. \qquad \square$$

In particular, if all $r_i = 0$, except for one $i$ for which $r_i = 1$ ($r = 1$, $\kappa = i$), then

$$\mathbf{E}\big[A_{n,k}(i)\big] = n\{k\}_i \frac{\sigma_{k-i}((n-1)\theta)}{\sigma_k(n\theta)} \frac{\sigma_i(\theta)}{i!} = n\mathbf{P}\big(K_{n,k}(1) = i\big). \tag{26}$$

This shows that the expected number of cells visited $i$ times is $n$ times the probability that there are $i$ visits to (say) cell one. In fact, we have the more general statement (see also [45] Sect. 1.5):

**Corollary 6** *If $r_i = \#\{m \in \{1, \dots, n\} : k_m = i\}$, then*

$$\mathbf{E}\left[\prod_{i=1}^{k}\left\{A_{n,k}(i)\right\}_{r_i}\right] = n!\mathbf{P}\big(K_{n,k}(1) = k_1, \dots, K_{n,k}(n) = k_n\big),$$

*so that the joint falling factorial moments of the $A$'s can directly be obtained in terms of the joint distribution of the $K$'s.*

*Proof* With the $r_i$ as stated, using a sampling without replacement argument

$$\mathbf{P}\big(K_{n,k}(1) = k_1, \dots, K_{n,k}(n) = k_n \mid A_{n,k}(1), \dots, A_{n,k}(k)\big)$$
$$= \frac{1}{n!}\prod_{i=1}^{k}\left\{A_{n,k}(i)\right\}_{r_i}.$$

Averaging over the $A$'s gives the announced result. □

2.6 The $*$-Limit of Sampling Distributions (the Infinitely Many Species Abundance Model)

Theoretical biologists work in a framework of a population with infinitely many species, with the more frequent one occurring with abundance $\xi_{(1)}$, second more frequent with abundance $\xi_{(2)}, \dots$ with $\xi_{(1)} \geq \xi_{(2)} \geq \cdots$. Sampling from $(\xi_{(1)}, \xi_{(2)}, \dots)$ turns out to be a challenging problem. This requires the introduction of a model with infinitely many species (not only $n$) with ordered abundance $\xi_{(m)}$, $m \geq 1$. For such abundance models, a $k$-sample will represent the met individuals of various species when sampling from a population with infinitely many species, [8]. One can think of obtaining such models while considering the limit $n \to \infty$ and $\theta \to 0$ in the finite model with $n$ species. Indeed, as we saw, small values of the temperature $\theta > 0$ was an indication on how disparate the abundance numbers $\boldsymbol{\xi}_n$ were. Then, although (as a result of $\mathbf{P}(\xi_1 = 0) = \sigma_0(\theta)/Z_\theta(x) \underset{\theta \to 0}{\to} 1$) the $(\xi_m)_{m=1}^{n}$ are all small in the limit, there is some hope that sampling from the ranked $\xi_{(m)}$'s would have a non-degenerate limit as $n \to \infty$, $\theta \to 0$ while $n\theta \to \gamma > 0$. We call such a limit the $*$-limit.

It turns out that for the class of Gibbs-Poisson allocation models considered in this Section, the $*$-limit always makes sense. This illustrates that limiting models should come down from some finitary counterpart, [22]. We first verify our claim intuitively (see also [45], Sect. 1.5). Observing indeed that

$$\sigma_k(\theta) \sim_{\theta \downarrow 0} \theta B_{k,1}(\phi_\bullet) = \theta\phi_k \quad \text{and} \quad B_{k,p}\big(\sigma_\bullet(\theta)\big) \sim_{\theta \downarrow 0} \theta^p B_{k,p}(\phi_\bullet)$$

and recalling $\{n\}_p \sim_{n \to \infty} n^p$, we easily get:

**Proposition 7** *From (15), with $(k_1, \dots, k_p) \in \mathbb{N}^p$ summing to $k$ and $p \leq k$*

$$\mathbf{P}\big(\widehat{K}_{n,k}(1) = k_1, \dots, \widehat{K}_{n,k}(p) = k_p;\, P_{n,k} = p\big)$$
$$\to_* \mathbf{P}^*\big(\widehat{K}_k(1) = k_1, \dots, \widehat{K}_k(p) = k_p;\, P_k = p\big) = \frac{k!}{p!}\frac{\gamma^p}{\sigma_k(\gamma)}\prod_{q=1}^{p}\frac{\phi_{k_q}}{k_q!} \tag{27}$$

*and, from (16), (17)*

$$\mathbf{P}(P_{n,k}=p)\to_* \mathbf{P}^*(P_k=p)=\frac{\gamma^p}{\sigma_k(\gamma)}B_{k,p}(\phi_\bullet). \tag{28}$$

*Equivalently, the limiting probability generating function of $P_k$ also reads*

$$\mathbf{E}^*\big(u^{P_k}\big)=\frac{\sigma_k(\gamma u)}{\sigma_k(\gamma)}, \tag{29}$$

*with mean $\mathbf{E}^*(P_k)=\gamma\frac{\sigma_k'(\gamma)}{\sigma_k(\gamma)}$. From this,*

$$\mathbf{P}^*\big(\widehat{K}_k(1)=k_1,\dots,\widehat{K}_k(p)=k_p\mid P_k=p\big)=\frac{k!}{p!}\frac{1}{B_{k,p}(\phi_\bullet)}\prod_{q=1}^{p}\frac{\phi_{k_q}}{k_q!} \tag{30}$$

*which is independent of $\gamma$.*

*Further, from (22), with $(a_1,a_2,\dots)$ satisfying $\sum_{i\ge1} i a_i=k$ and $\sum_{i\ge1}a_i=p$*

$$\begin{aligned}&\mathbf{P}\big(A_{n,k}(1)=a_1,\,A_{n,k}(2)=a_2,\dots\big)\\&\quad\to_* \mathbf{P}^*\big(A_k(1)=a_1,\,A_k(2)=a_2,\dots\big)=\frac{\gamma^p k!}{\sigma_k(\gamma)}\prod_{i=1}^{k}\frac{(\phi_i/i!)^{a_i}}{a_i!}.\end{aligned} \tag{31}$$

*Equivalently, from (24)*

$$\begin{aligned}&\mathbf{P}\big(A_{n,k}(1)=a_1,\dots,A_{n,k}(k)=a_k;\,P_{n,k}=p\big)\\&\quad\to_* \mathbf{P}^*\big(A_k(1)=a_1,\dots,A_k(k)=a_k;\,P_k=p\big)=\frac{\gamma^p k!}{\sigma_k(\gamma)}\prod_{i=1}^{k}\frac{(\phi_i/i!)^{a_i}}{a_i!}\end{aligned} \tag{32}$$

*and*

$$\mathbf{P}^*\big(A_k(1)=a_1,\dots,A_k(k)=a_k\mid P_k=p\big)=\frac{k!}{B_{k,p}(\phi_\bullet)}\prod_{i=1}^{k}\frac{(\phi_i/i!)^{a_i}}{a_i!}, \tag{33}$$

*which is also independent of $\gamma$.*

Equations (27) or (32) are the canonical Gibbs distributions on partitions of $k$ into $p$ distinct clusters, derived from the weight sequence $\phi_\bullet$. In this context, the normalizing quantity $\sigma_k(\gamma)/k!$ is called the canonical Gibbs partition polynomial.[3] Conditioning the canonical Gibbs distribution on the number of filled boxes being equal to $p$ yields the corresponding micro-canonical distributions (30) or (33). The new normalizing constant $B_{k,p}(\phi_\bullet)/k!$ is called the microcanonical partition function.

Let us finally compute the falling factorial moments of $A_k(i)$, $i=1,\dots,k$.

[3]The occupancy distribution (32) also appears in Ecology in a species abundance model occurring in the Hubbell's unified neutral theory of biodiversity. In this context, $\gamma$ is the fundamental biodiversity number, [29].

**Proposition 8** *Let $r_i$, $i=1,\dots,k$ be non-negative integers satisfying $\sum_1^k r_i = r$ and $\sum_1^k i r_i = \kappa \le k$. We have*

$$\mathbf{E}^*\left[\prod_{i=1}^{k}\{A_k(i)\}_{r_i}\right] = \gamma^r \{k\}_\kappa \frac{\sigma_{k-\kappa}(\gamma)}{\sigma_k(\gamma)}\prod_{i=1}^{k}\left(\frac{\phi_i}{i!}\right)^{r_i}. \tag{34}$$

*Proof* This follows straightforwardly from Proposition 5 while taking the $*$-limit and using $\sigma_i(\theta) \sim \theta\phi_i$ for small $\theta$. This formula is a generalization of the Watterson expression [52] obtained in the special Ewens context when $\phi(x) = -\log(1-x)$, with $\phi_i = (i-1)!$ and $\sigma_k(\gamma) = \Gamma(\gamma+k)/\Gamma(\gamma) =: (\gamma)_k$; see Sect. 3 for a special account on this model. From (34), we easily get a closed-form expression for the mean $\mathbf{E}^*(A_k(i))$, $i \le k$, the variance $\mathrm{Var}^{*2}(A_k(i))$, for all $i$ with $2i \le k$ and the covariance $\mathrm{Cov}^*(A_k(i_1), A_k(i_2))$ for all $i_1 \ne i_2$, $i_1 + i_2 \le k$. □

We observed that (30) or (33) were independent of $\gamma$, meaning that $P_k$ is a sufficient statistic in the estimation of $\gamma$ problem. Let us now briefly investigate this problem.

### 2.7 The Estimation of $\gamma$ Problem

We wish now to discuss the question of estimating $\gamma$ from the data $k$ and $P$. From (28)

$$\partial_\gamma \log \mathbf{P}^*(P_k = P) = P/\gamma - \partial_\gamma \log\sigma_k(\gamma).$$

Suppose the polynomial $\sigma_k(\gamma) \in ZR_-$ with zeroes $-r_{l,k}$ where: $0 = r_{1,k} \le \cdots \le r_{k,k}$. Then $\sigma_k(\gamma) = \prod_{l=1}^k (\gamma + r_{l,k})$ and $\partial_\gamma \log\sigma_k(\gamma) = \sum_{l=1}^k (\gamma + r_{l,k})^{-1}$, together with $\partial_\gamma^2 \log\sigma_k(\gamma) = -\sum_{l=1}^k (\gamma + r_{l,k})^{-2} < 0$ ($\gamma \to \sigma_k(\gamma)$ is log-concave).

If $P/\gamma - \sum_{l=1}^k (\gamma + r_{l,k})^{-1} \overset{(*)}{=} 0$ , then

$$\partial_\gamma^2 \log \mathbf{P}^*(P_k = P) = -P/\gamma^2 + \sum_{l=1}^{k}(\gamma + r_{l,k})^{-2} < 0,$$

showing that $\widehat{\gamma}$ solving $(*)$ is a local maximum and that $\log \mathbf{P}^*(P_k = P)$ has no local minima. So $\widehat{\gamma}$ is the maximum likelihood estimator of $\gamma$. Even though $\sigma_k(\gamma)$ $(1/\sigma_k(\gamma))$ is a log-concave (respectively log-convex) function of $\gamma$, the log-likelihood is a log-concave function of $\gamma$ leading to the existence of $\widehat{\gamma}$. To summarize, there exists a maximum likelihood estimator $\widehat{\gamma}$ of $\gamma$ which is characterized by the implicit equation:

$$P = \widehat{\gamma}\frac{\sigma_k'(\widehat{\gamma})}{\sigma_k(\widehat{\gamma})}.$$

Let us now come to another estimator of $\gamma$. If $\sigma_k(\gamma) \in ZR_-$, then by Newton's inequality ([24], p. 52)

$$B_{k,p}(\phi_\bullet)^2 \ge B_{k,p-1}(\phi_\bullet)B_{k,p+1}(\phi_\bullet)\left(1+\frac{1}{p}\right)\left(1+\frac{1}{k-p}\right) > B_{k,p-1}(\phi_\bullet)B_{k,p+1}(\phi_\bullet).$$

So $B_{k,p}(\phi_\bullet)$ is $p$-log-concave and by Darroch Theorem, $B_{k,p}(\phi_\bullet)$ is $p$-unimodal or bimodal at two consecutive $p$, with mode (maybe up to one unit) equal to $\sigma_k'(1)/\sigma_k(1)$. Because the

$p$-sequence $\gamma^p$ is also log-concave (and log-convex), $\gamma^p B_{k,p}(\phi_\bullet)$ is itself $p$-log-concave and therefore there exists a unique $\widetilde{\gamma}$ such that $\frac{\mathbf{P}^*(P_k=P)}{\mathbf{P}^*(P_k=P-1)}=1$. It is thus defined by

$$\frac{\gamma^p B_{k,P}(\phi_\bullet)}{\gamma^{p-1} B_{k,P-1}(\phi_\bullet)}=1, \quad \text{or} \quad \widetilde{\gamma}=\frac{B_{k,P-1}(\phi_\bullet)}{B_{k,P}(\phi_\bullet)}.$$

This $\widetilde{\gamma}$ is an alternative explicit estimator of $\gamma$ based on the data $k$ and $P$.

Taking the expectation with respect to $P_k$, we have

$$\begin{aligned}\mathbf{E}^*(\widetilde{\gamma}) &= \sum_{p=1}^{k} \frac{B_{k,p-1}(\phi_\bullet)}{B_{k,p}(\phi_\bullet)} \frac{\gamma^p}{\sigma_k(\gamma)} B_{k,p}(\phi_\bullet) = \gamma \sum_{p=2}^{k} B_{k,p-1}(\phi_\bullet) \frac{\gamma^{p-1}}{\sigma_k(\gamma)} \\ &= \gamma \sum_{p=1}^{k-1} B_{k,p}(\phi_\bullet) \frac{\gamma^p}{\sigma_k(\gamma)} = \gamma \left(1-\frac{(\phi_1\gamma)^k}{\sigma_k(\gamma)}\right) < \gamma.\end{aligned}$$

This shows that $\widetilde{\gamma}$ is not an unbiased estimator of $\gamma$.

*Remark* The estimator $\widetilde{\gamma}$ only requires that the sequence $B_{k,p}(\phi_\bullet)$ be $p$-log-concave and, although sufficient, it is therefore not necessary that $\sigma_k(\gamma) \in ZR_-$; the sequence $B_{k,p}(\phi_\bullet)$ only needs to be a Pòlya frequency sequence of order 2 (so $\sigma_k(\gamma) \in PF_2$) for $\widetilde{\gamma}$ to be well-defined. In this spirit, we draw the attention on a result in [2], stating that if the non-null roots of $\sigma_k(\gamma)$ all lie in the angular cone $\phi \in (2\pi/3, 4\pi/3)$ of the complex plane, then $\sigma_k(\gamma)$ has $p$-log-concave coefficients. See [47] for a bulk of work pertaining to the diversity estimation parameter for Gibbs partitions.

## 3 Sampling from Dirichlet Partition: A Special Case

We now briefly investigate one particular model of species abundance $\boldsymbol{\xi}_n$.

- **Sampling from a negative binomial sample.**

Assume $\phi(x) = -\log(1-x)$, with $\phi_m = (m-1)!$ and let $Z_\theta(x) = (1-x)^{-\theta}$. Thus, with $(\theta)_k := \theta(\theta+1)\cdots(\theta+k-1)$ denoting the (rising factorial) Pochhammer symbol, $\sigma_k(\theta) = (\theta)_k$ and $\xi$ is a negative binomial random variable with parameters $\theta$ and $1-x$. Note that $\sigma_k(\theta) \in ZR_-$. From (11), the jumps' height $\delta$ of $\xi$ is seen to obey a logarithmic series distribution.

When sampling from this discrete species-abundance model $\boldsymbol{\xi}_n = (\xi_1, \ldots, \xi_n)$, for instance (12) takes the particular form:

$$\mathbf{P}(\mathbf{K}_{n,k} = \mathbf{k}_n) = \frac{\mathbf{P}(\xi_1 = k_1, \ldots, \xi_n = k_n)}{\mathbf{P}(\zeta_n = k)} = \frac{k!}{(n\theta)_k} \prod_{m=1}^{n} \frac{(\theta)_{k_m}}{k_m!}. \tag{35}$$

Substituting $(\theta)_k$ to $\sigma_k(\theta)$ in (15) gives its particular expression.

Because $\sigma_{k+1}(\theta) = (k+\theta)\sigma_k(\theta)$, it follows from (3) and (4) that with $S_k(\lambda) = k![x^k]e^{\lambda((1-x)^{-\theta}-1)}$, $S_{k+1}(\lambda) = (\theta\lambda + k)S_k(\lambda) + \theta\lambda S_k'(\lambda)$. Thus, the Bell coefficients $B_{k,p}(\sigma_\bullet(\theta)) = B_{k,p}((\theta)_\bullet) = [\lambda^p]S_k(\lambda)$, appearing in (16), obey a simple 3-term recurrence [14, 30]

$$B_{k+1,p}\big((\theta)_\bullet\big) = \theta B_{k,p-1}\big((\theta)_\bullet\big) + (p\theta + k) B_{k,p}\big((\theta)_\bullet\big),$$

which should be considered with the boundary conditions

$$B_{k,0}\big((\theta)_\bullet\big) = B_{0,p}\big((\theta)_\bullet\big) = 0,$$

except for $B_{0,0}((\theta)_\bullet) := 1$. This observation is important because it follows from (16), that, there exist transition probabilities

$$\mathbf{P}(P_{n,k+1} = p+1 \mid P_{n,k} = p) = \frac{(n-p)\theta}{n\theta + k} \quad \text{and}$$

$$\mathbf{P}(P_{n,k+1} = p \mid P_{n,k} = p) = \frac{\sum_{r=1}^{p}(\theta + k_r)}{n\theta + k} = \frac{p\theta + k}{n\theta + k}$$

such that,

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

The first transition probability gives the probability of the event that a new species is discovered given $p < n$ of them were discovered from a previous sample of size $k \geq p$ (the so-called law of succession, [17, 19]) in a population with $n$ species. Note that $P_{n,k}$ is a Markov chain in $k$.

Considering the sampling formulae in the $*$-limit, the expressions (30) and (33) with $\phi_i = (i-1)!$ and $B_{k,p}(\phi_\bullet) = s_{k,p}$ (the absolute first kind Stirling numbers) are the Ewens sampling formulae [18]. Due to $\sigma_{k+1}(\theta) = (k+\theta)\sigma_k(\theta)$, the Bell coefficients $B_{k,p}(\phi_\bullet) = B_{k,p}((\bullet - 1)!)$ also obey a 3-term recurrence

$$B_{k+1,p}\big((\bullet - 1)!\big) = B_{k,p-1}\big((\bullet - 1)!\big) + kB_{k,p}\big((\bullet - 1)!\big).$$

- **Sampling from a symmetric Dirichlet prior.**

It turns out that this sampling formula can be obtained while following a different path for the sampling procedure:

Consider indeed the following random partition into $n$ fragments of the unit interval. Let $\theta > 0$ be some parameter and assume that the random fragments sizes $\mathbf{S}_n(\theta) := (S_{1,\theta}, \dots, S_{n,\theta})$ (with $\sum_{m=1}^{n} S_{m,\theta} = 1$) are distributed according to the (exchangeable) Dirichlet $D_n(\theta)$ density function on the $n$-simplex, that is to say

$$f_{S_{1,\theta},\dots,S_{n,\theta}}(s_1, \dots, s_n) = \frac{\Gamma(n\theta)}{\Gamma(\theta)^n}\prod_{m=1}^{n} s_m^{\theta - 1} \cdot \delta_{(\sum_{m=1}^{n} s_m = 1)}. \tag{36}$$

Alternatively, with $(\theta)_q := \Gamma(\theta + q)/\Gamma(\theta)$, the law of $\mathbf{S}_n(\theta)$ is characterized by its joint moment function

$$\mathbf{E}\left(\prod_{m=1}^{n} S_{m,\theta}^{q_m}\right) = \frac{1}{(n\theta)_{\sum_{m=1}^{n} q_m}}\prod_{m=1}^{n}(\theta)_{q_m}. \tag{37}$$

We shall put $\mathbf{S}_n(\theta) \overset{d}{\sim} D_n(\theta)$ if $\mathbf{S}_n(\theta)$ is Dirichlet distributed with parameter $\theta$. $\mathbf{S}_n(\theta)$ can be obtained while considering $(Y_\theta \overset{d}{=} Y_{1,\theta}, \dots, Y_{n,\theta})$, an iid random vector with $Y_\theta \overset{d}{\sim} \text{gamma}(\theta)$ and letting $S_{m,\theta} = Y_{m,\theta}/(Y_{1,\theta} + \cdots + Y_{n,\theta})$, $m = 1, \dots, n$ (normalizing the $Y_{m,\theta}$'s by their sum). $\mathbf{S}_n(\theta)$ accounts now for a $n$-species frequency (proportion) model, but now in the continuum. We now come to the sampling procedure from $\mathbf{S}_n(\theta)$.

Let $(U_1, \ldots, U_k)$ be $k$ iid uniform throws on the unit interval partitioned according to $\mathbf{S}_n(\theta)$. Let

$$\mathbf{K}_{n,k} := \big(K_{n,k}(1), \ldots, K_{n,k}(n)\big)$$

be an integral-valued random vector which counts the number of visits to the different fragments of $\mathbf{S}_n(\theta)$ in this $k$-sample. Hence, if $M_l$ is the random fragment label in which the $l$th trial $U_l$ falls, $K_{n,k}(m) := \sum_{l=1}^{k} \mathbf{I}(M_l = m)$, $m = 1, \ldots, n$.

With $|\mathbf{k}_n| = k$ and $\mathbf{k}_n := (k_1, \ldots, k_n) \in \mathbb{N}_0^n$ the non-negative occupancy vector, as sampling in terms of uniforms $U_l$ is equivalent to the multinomial, $\mathbf{K}_{n,k}$ follows the conditional multinomial distribution:

$$\mathbf{P}\big(\mathbf{K}_{n,k} = \mathbf{k}_n \mid \mathbf{S}_n(\theta)\big) = \frac{k!}{\prod_{m=1}^{n} k_m!} \prod_{m=1}^{n} S_{m,\theta}^{k_m}. \tag{38}$$

Averaging over $\mathbf{S}_n(\theta)$, we find

$$\mathbf{P}(\mathbf{K}_{n,k} = \mathbf{k}_n) = \mathbf{EP}\big(\mathbf{K}_{n,k} = \mathbf{k}_n \mid \mathbf{S}_n(\theta)\big) = \frac{k!}{(n\theta)_k} \prod_{m=1}^{n} \frac{(\theta)_{k_m}}{k_m!}, \tag{39}$$

which is the Dirichlet-multinomial distribution, with $\mathbf{E}(K_{n,k}(m)) = k/n$. We shall put $\mathbf{K}_{n,k} \overset{d}{\sim} D_{n,k}(\theta)$.

The sampling from $\mathbf{S}_n(\theta) \overset{d}{\sim} D_n(\theta)$ formula (39) coincides with the one (35) obtained while sampling from a discrete species abundance model $\boldsymbol{\xi}_n$ with negative binomial distributions. The $*$-limit of this Dirichlet model is known to lead to the Ewens sampling formulae which are particular incarnation of (30) and (33) with $\phi_i = (i-1)!$ and $B_{k,p}(\phi_\bullet) = s_{k,p}$. See [35] and [36].

It is worthwhile exploring if this remarkable property (or maybe a weaker one) propagates to sampling from other discrete species abundance model.

## 4 Sampling Problems from a Special CP Class

We shall now exhibit a sub-class of CP models whose statistical properties are very similar to the ones developed in the latter Section for the Dirichlet model.

### 4.1 Sampling from a Special CP Class

Let us first define the class of $\phi$ we will be interested in.

*The Special Class $\mathcal{S}$* We first recall that a function $h(x)$ defined on some interval $x \in (-\infty, x_0)$ is absolutely monotone on some open interval $I \subseteq (-\infty, x_0)$ if it is $C^\infty$ with $h^{(n)}(x) \geq 0$ for all $n \geq 0$ and $x \in I$.

We shall consider the following special class model

**Definition 1** Suppose that $\phi(x)$ (with $\phi_1 > 0$ and $\phi_m \geq 0$, $m \geq 2$) as from (1), is defined (finite) on the unbounded half-domain $x \in (-\infty, x_0)$ with $0 < x_0 \leq \infty$ and that $\phi'(x)$ is absolutely monotone for all $x \in (-\infty, x_0)$. If this is the case, we shall put $\phi \in \mathcal{S}$. If $\phi \in \mathcal{S}$, $Z_\theta(x) = \exp(\theta\phi(x))$ is also defined on $x \in (-\infty, x_0)$ and absolutely monotone there.

- Examples of $\phi \in \mathcal{S}$ are $x$, $e^x - 1$ (Bell), $-\log(1-x)$, $(1-x)^{-\alpha} - 1$, $\alpha > 0$ and $1-(1-x)^\alpha$, $\alpha \in (0,1)$.
- Examples of $\phi \notin \mathcal{S}$ are polynomials with positive coefficients $\sum_{l=1}^d c_l x^l$ ($d \geq 2$), $xe^x$, $\sinh(x)$, $\cosh(x) - 1$ and $\tan(x)$. Although the latter $\phi$'s can be expanded as in (1) and all have non-negative Taylor coefficients $\phi_m$ ($\phi_1 > 0$), the corresponding $\phi'(x)$ are not absolutely monotone on $(-\infty, x_0)$ although they are of course on $(0, x_0)$.

**Remarks and properties:**

– If $\phi \in \mathcal{S}$, so does clearly $\widetilde{\phi}(x) := a\phi(bx)$ for all $a, b > 0$. We can check that: $B_{k,p}(\widetilde{\phi}_\bullet) = a^p b^k B_{k,p}(\phi_\bullet)$.
– If $\phi^1, \phi^2 \in \mathcal{S}$, then $\phi^1 + \phi^2 \in \mathcal{S}$ and the composition $\phi^1 \circ \phi^2 \in \mathcal{S}$. This allows to produce a lot of new examples of $\phi$'s in $\mathcal{S}$ from the ones already introduced. For instance because $\phi^1 = (1-x)^{-\alpha} - 1$ and $\phi^2 = 1-(1-x)^\alpha$ both belong to $\mathcal{S}$, would $\alpha \in (0,1)$, $\phi^1 + \phi^2 = 2\sinh(-\alpha \log(1-x))$ belongs to $\mathcal{S}$, together with $\phi^1 \circ \phi^2 = (1-x)^{-\alpha^2} - 1$ and $\phi^2 \circ \phi^1 = 1-(2-(1-x)^{-\alpha})^\alpha$.
– If $\phi^1, \phi^2 \in \mathcal{S}$, the product $\phi := \phi^1 \cdot \phi^2 \notin S$ (in the first place because $\phi_1 = 0$). The Taylor coefficients $\phi_m$ of $\phi$ are

$$\phi_m = \sum_{l=1}^{m-1} \binom{m}{l} \phi_l^1 \phi_{m-l}^2 = \left(\phi^1 * \phi^2\right)_m, \quad m \geq 2$$

and the $\phi_m$ do not necessarily form a log-convex sequence, even though $\phi_m^1$, $\phi_m^2$, $m \geq 1$, would be log-convex themselves. This is not in contradiction with the Davenport and Pòlya theorem [13] stating that the binomial convolution of two log-convex sequences is log-convex because the $\phi^1, \phi^2$ sequences here have no constant terms: $\phi_0^1 = \phi_0^2 = 0$ (resulting in $\phi_1 = 0$). The reason why, when $\phi(x) \in \mathcal{S}$, log-convexity of the sequences $(\phi_m)_{m\geq 1}$ pops in is (see [4] and [48]):

**Proposition 9** *When $\phi \in \mathcal{S}$, the function $h(x) := \phi'(-x)$ is completely monotone on the domain $x \in (-x_0, \infty)$, meaning it is $C^\infty$ with $(-1)^n h^{(n)}(x) \geq 0$ for all $n \geq 0$ and $x \in (-x_0, \infty)$. So (from Bernstein theorem [5]), $h(x)$ is the Laplace-Stieltjes transform (LST) of some finite non-negative measure $\mu$ on $[0, +\infty)$: $h(x) = \int_0^\infty e^{-xt} \mu(dt)$. We have*

$$h(x) = \sum_{m \geq 0} \frac{\phi_{m+1}}{m!} (-x)^m$$

*and so $\phi_{m+1}$ is the $m$th moment of $\mu$, with finite total mass $\phi_1$. By the Cauchy-Schwarz inequality, for all $m \geq 2$, $\phi_{m+1}\phi_{m-1} \geq \phi_m^2$, showing that when $\phi \in \mathcal{S}$, $(\phi_m)_{m\geq 1}$ is a log-convex sequence. Upon shifting, $(\phi_m)_{m\geq 1}$ is the moment sequence of some non-negative measure $\pi(dt) := t^{-1}\mu(dt)$.*

Let us now consider $Z_\theta(-x) = e^{\theta\phi(-x)} =: e^{-\theta\psi(x)}$, with

$$\psi(x) := -\phi(-x), \quad x > -x_0.$$

**Proposition 10** *When $\phi \in \mathcal{S}$, it holds that $\psi'(x) = h(x) = \int_0^\infty e^{-xt} t\pi(dt)$ is completely monotone, so $Z_\theta(-x) = e^{-\theta\psi(x)}$ is the LST of some infinitely divisible random variable (or subordinator process) $Y_\theta$ on $[0, +\infty)$, whose integral moments are all finite. The coefficients*

$(\phi_m)_{m\geq 1}$ *are the cumulants of* $Y_\theta$. *The function* $\psi$ *is the Laplace exponent of* $Y_\theta$ *with* $\psi(x) = cx + \int_0^\infty (1-e^{-xt})\pi(dt)$ *for some* $c \geq 0$ *and some positive Lévy measure* $\pi(dt)$ *on* $(0,\infty)$, *integrating* $1 \wedge t$ [50]. *Therefore*, *when* $\phi \in \mathcal{S}$,

$$Z_\theta(-x) = \mathbf{E}\big(e^{-xY_\theta}\big) = e^{-\theta\psi(x)} = 1 + \sum_{k\geq 1} \frac{(-x)^k}{k!}\sigma_k(\theta),$$

*with* $(\sigma_k(\theta),\ k \geq 0)$ *being the Stieltjes moment sequence of* $Y_\theta$: $\sigma_k(\theta) = \mathbf{E}(Y_\theta^k)$. *Thus*, *when* $\phi \in \mathcal{S}$, *for all* $\theta > 0$, $(\sigma_k(\theta))_{k\geq 0}$ *forms a k-log-convex sequence and for all* $k \geq 1$, *all* $\theta > 0$: $\sigma_{k+1}(\theta)\sigma_{k-1}(\theta) \geq \sigma_k(\theta)^2$.

*Since* $\mathbf{E}(e^{-x\overline{Y}_{n,\theta}}) = e^{-n\theta\psi(x)}$, $\sigma_k(n\theta)$ *is also the kth moment of the sum* $\overline{Y}_{n,\theta} := Y_{1,\theta} + \cdots + Y_{n,\theta}$ *of n iid terms* $Y_{m,\theta} := Y_{m\theta} - Y_{(m-1)\theta}$. *So*, $\sigma_k(n\theta) = \mathbf{E}(\overline{Y}_{n,\theta}^k) = \mathbf{E}(Y_{n\theta}^k)$.

Note finally that taking $Z_\theta(x) = Z_\theta^1(x)Z_\theta^2(x)$ where $Z_\theta^i(x) = e^{\theta\phi_i(x)}$ for two $\phi_i$ in $\mathcal{S}$, with $\sigma_k^i(\theta)$ defined by $Z_\theta^i(x) = 1 + \sum_{k\geq 1}\frac{x^k}{k!}\sigma_k^i(\theta)$, two $k$-log-convex sequences, the sequence $\sigma_k(\theta)$ defined by $Z_\theta(x) = 1 + \sum_{k\geq 1}\frac{x^k}{k!}\sigma_k(\theta)$ obeys

$$\sigma_k(\theta) = \sum_{l=0}^{k}\binom{k}{l}\sigma_l^1(\theta)\sigma_{k-l}^2(\theta) = \big(\sigma^1(\theta) * \sigma^2(\theta)\big)_k, \quad k \geq 0,$$

and is $k$-log-convex by Davenport and Pòlya theorem, as a binomial convolution of two log-convex sequences.

*Sampling from* $\boldsymbol{\xi}_n$ *when* $\phi \in \mathcal{S}$ Assume $\phi \in \mathcal{S}$ and consider the sampling problem from $\boldsymbol{\xi}_n$, where $\xi$ is constructed as in Sect. 2 from $\phi$, but now for $\phi \in \mathcal{S}$. Note that in this case

$$\mathbf{E}\big(u^{\xi}\big) = e^{\theta[\phi(xu)-\phi(x)]} = e^{-\theta[\psi(-xu)-\psi(-x)]}.$$

In a general sampling problem from $\boldsymbol{\xi}_n$, the joint probability generating function of $\mathbf{K}_{n,k}$ was given by (13). From (12) and making use of $\phi \in \mathcal{S}$, from Proposition 10, we have

$$\mathbf{P}(\mathbf{K}_{n,k} = \mathbf{k}_n) = \frac{k!}{\sigma_k(n\theta)}\prod_{m=1}^{n}\frac{\sigma_{k_m}(\theta)}{k_m!} = \binom{k}{k_1\ldots k_n}\frac{\prod_{m=1}^n \mathbf{E}(Y_{m,\theta}^{k_m})}{\mathbf{E}(\overline{Y}_{n,\theta}^k)}, \tag{40}$$

*Remark* Because (40) does not depend on the common mean of the $Y_{m,\theta}$'s, we can as well define the reduced (iid) random variables with mean 1: $X_{m,\theta} := Y_{m,\theta}/(\theta\phi_1)$, $m = 1,\ldots,n$ and $\overline{X}_{n,\theta} := \sum_{m=1}^n X_{m,\theta}$. Then, with $S_{m,\theta} := X_{m,\theta}/\overline{X}_{n,\theta}$, $m = 1,\ldots,n$ defining a random partition $\mathbf{S}_n(\theta) = (S_{1,\theta},\ldots,S_{n,\theta})$ of unity into $n$ exchangeable (mean $1/n$) parts

$$\begin{aligned}\mathbf{P}(\mathbf{K}_{n,k} = \mathbf{k}_n) &= \binom{k}{k_1\ldots k_n}\frac{\prod_{m=1}^n \mathbf{E}(X_{m,\theta}^{k_m})}{\mathbf{E}(\overline{X}_{n,\theta}^k)} \\ &= \binom{k}{k_1\ldots k_n}\frac{\prod_{m=1}^n \mathbf{E}(\overline{X}_{n,\theta}^{k_m}S_{m,\theta}^{k_m})}{\mathbf{E}(\overline{X}_{n,\theta}^k)} = \binom{k}{k_1\ldots k_n}\frac{\mathbf{E}(\overline{X}_{n,\theta}^k\prod_{m=1}^n S_{m,\theta}^{k_m})}{\mathbf{E}(\overline{X}_{n,\theta}^k)}\end{aligned} \tag{41}$$

as well. The latter expression is identified to an occupancy distribution arising from sampling from the random partition of unity $\mathbf{S}_n(\theta)$ but now biased by the total length $\overline{X}_{n,\theta}$. In the occupancy distribution (41) indeed, realizations of $(X_{m,\theta})_{m=1}^{n}$ giving rise to large values of the sum $\overline{X}_{n,\theta}$ are favored, compared to the "unbiased" multinomial one, say $\mathbf{Q}(\mathbf{K}_{n,k}=\mathbf{k}_n):=\binom{k}{k_1\ldots k_n}\prod_{m=1}^{n}\mathbf{E}(S_{m,\theta}^{k_m})$, based on the same $\mathbf{S}_n(\theta)$.

Would $\overline{X}_{n,\theta}$ be independent of $S_{m,\theta}=X_{m,\theta}/\overline{X}_{n,\theta}$, $m=1,\ldots,n$, (the only possible way, by Lukacs' criterion, to have this is when $\mathbf{S}_n(\theta)$ has Dirichlet$(\theta)$ distribution, [26]), the latter expression boils down to the usual sampling one

$$\mathbf{P}(\mathbf{K}_{n,k}=\mathbf{k}_n)=\binom{k}{k_1\ldots k_n}\mathbf{E}\left(\prod_{m=1}^{n}S_{m,\theta}^{k_m}\right)=\mathbf{Q}(\mathbf{K}_{n,k}=\mathbf{k}_n).$$

Alternatively, from (41), the joint pgf of $\mathbf{K}_{n,k}$ also reads

$$\mathbf{E}\left[\prod_{m=1}^{n}u_m^{K_{n,k}(m)}\right]=\frac{\mathbf{E}[(\sum_{m=1}^{n}u_mX_{m,\theta})^k]}{\mathbf{E}(\overline{X}_{n,\theta}^k)}=\frac{\mathbf{E}[\overline{X}_{n,\theta}^k(\sum_{m=1}^{n}u_mS_{m,\theta})^k]}{\mathbf{E}(\overline{X}_{n,\theta}^k)}.$$

Its computation is thus amenable to the normalized $k$th moment of the weighted sum $\sum_1^n u_mX_{m,\theta}$ of iid mean 1 infinitely divisible random variables with LST $\mathbf{E}(e^{-xX_\theta})=e^{\theta\phi(-x/(\theta\phi_1))}=e^{-\theta\psi(x/(\theta\phi_1))}$ and moments $\mathbf{E}(X_\theta^k)=\sigma_k(\theta)/(\theta\phi_1)^k$, $k\geq 1$. Unlike $(Y_\theta;\theta\geq 0)$, the process $(X_\theta;\theta\geq 0)$ is not a Lévy process.

Note also that with $\mathbf{k}_p:=(k_1,\ldots,k_p)\in\mathbb{N}^p$ obeying $|\mathbf{k}_p|=k$

$$\begin{aligned}&\mathbf{P}\big(\widehat{K}_{n,k}(1)=k_1,\ldots,\widehat{K}_{n,k}(p)=k_p;\,P_{n,k}=p\big)\\&=\binom{n}{p}\binom{k}{k_1\ldots k_p}\frac{\prod_{q=1}^{p}\mathbf{E}(\overline{X}_{n,\theta}^{k_q}S_{q,\theta}^{k_q})}{\mathbf{E}(\overline{X}_{n,\theta}^k)}=\binom{n}{p}\binom{k}{k_1\ldots k_p}\frac{\mathbf{E}(\overline{X}_{n,\theta}^k\prod_{q=1}^{p}S_{q,\theta}^{k_q})}{\mathbf{E}(\overline{X}_{n,\theta}^k)}\end{aligned}$$

is the joint probability that there are $p\in[n]$ non-empty boxes and that $(k_1,\ldots,k_p)$ are the respective occupancies of the $p$ filled boxes, labeled in arbitrary order. Again

$$\begin{aligned}\mathcal{P}_{k,p}^{(n)}&:=\binom{n}{p}\sum_{\mathbf{k}_p\in\mathbb{N}^p:|\mathbf{k}_p|=k}\binom{k}{k_1\ldots k_p}\frac{\prod_{q=1}^{p}\mathbf{E}(\overline{X}_{n,\theta}^{k_q}S_{q,\theta}^{k_q})}{\mathbf{E}(\overline{X}_{n,\theta}^k)}\\&=\binom{n}{p}\sum_{\mathbf{k}_p\in\mathbb{N}^p:|\mathbf{k}_p|=k}\binom{k}{k_1\ldots k_p}\frac{\mathbf{E}(\overline{X}_{n,\theta}^k\prod_{q=1}^{p}S_{q,\theta}^{k_q})}{\mathbf{E}(\overline{X}_{n,\theta}^k)}\end{aligned}$$

is the probability that in a $k$-sample from $n$ species with abundance $\boldsymbol{\xi}_n$ in the special class $\mathcal{S}$, the exact number of distinct visited species is $p$.

To summarize, we conclude

**Proposition 11** *When $\phi\in\mathcal{S}$ and when the discrete species abundance model $\boldsymbol{\xi}_n$ is built on $\phi$, its occupancy distribution (12) can alternatively be given the interpretation of an occupancy distribution (41) arising from sampling from the random partition of unity $\mathbf{S}_n(\theta)$ but biased by the total length $\overline{X}_{n,\theta}$ appearing in the normalization of $S_{m,\theta}:=X_{m,\theta}/\overline{X}_{n,\theta}$. The positive random variable $X_\theta\overset{d}{=}X_{1,\theta}$ is infinitely divisible. The correspondence between $\xi$ and (mean 1) $X_\theta$ is:*

$$\mathbf{E}\big[u^{\xi}\big]=e^{-\theta\phi(x)(1-\frac{\phi(xu)}{\phi(x)})}\quad\text{and}\quad\mathbf{E}\big(e^{-xX_\theta}\big)=e^{\theta\phi(-x/(\theta\phi_1))}=e^{-\theta\psi(x/(\theta\phi_1))}.$$

Note finally that $\psi(x/(\theta\phi_1))$ being the Laplace exponent of $X_\theta$:

$$\mathbf{E}\big(e^{-xX_\theta}\big) = e^{-\theta\int_0^\infty(1-e^{-xt})\pi_\theta(dt)},$$

where the Lévy measure $\pi_\theta(dt)$ integrates $1 \wedge t$. The measure $t\pi_\theta(dt)$ is a finite positive measure with all finite $m$-moments: $\int_0^\infty t^m t\pi_\theta(dt) = \phi_{m+1}/(\theta\phi_1)^{m+1}$, $m \geq 0$. So $((\theta\phi_1)^{-m}\phi_m)_{m\geq 1}$ is the moment sequence of $\pi_\theta(dt)$.

With $S_{1,\theta} := X_{1,\theta}/\overline{X}_{n,\theta}$, define finally $\mu_k := \mathbf{E}[S_{1,\theta}^k]$, $k \geq 1$, the sequence of the moments of $S_{1,\theta}$.

*Examples* Examples of admissible $\phi \in \mathcal{S}$ were $-\log(1-x)$, $(1-x)^{-\alpha} - 1$, $\alpha > 0$ and $1-(1-x)^\alpha$, $\alpha \in (0,1)$.

The LST $\mathbf{E}(e^{-xX_\theta})$ of $X_\theta$ in each case is $(1+x/\theta)^{-\theta}$, $\exp[-\theta(1-(1+\frac{x}{\alpha\theta})^{-\alpha})]$ and $\exp[-\theta((1+\frac{x}{\alpha\theta})^\alpha - 1)]$ corresponding respectively to a Gamma$(\theta,\theta)$ distribution, a compound Poisson sum of iid gamma$(\alpha,\alpha\theta)$ random variables and an exponentially damped stable$(\theta,\alpha)$. For this last case, let $\Sigma > 0$ be a stable$(\theta,\alpha)$ random variable i.e. with LST $\mathbf{E}(e^{-x\Sigma}) := \exp[-\theta x^\alpha]$, $x \geq 0$. Let $f_\Sigma$ be its density. Define a random variable $Y_\theta$ with damped density $f_{Y_\theta}(t) = \frac{1}{\mathbf{E}(e^{-\Sigma})}e^{-t}f_\Sigma(t)$, $t > 0$. Its LST is $\mathbf{E}(e^{-xY_\theta}) = \mathbf{E}(e^{-(x+1)\Sigma})/\mathbf{E}(e^{-\Sigma}) = \exp -\theta[(1+x)^\alpha - 1]$. Upon scaling $Y_\theta$, $X_\theta := Y_\theta/(\theta\alpha)$ is mean 1. In the sampling context, the last example was recently considered in [15, 16, 27, 28]. They were named the generalized inverse Gaussian or Engen models.

*Remark* In the degenerate case, $\phi(x) = x$, $X_\theta$ is purely atomic with $X_\theta \overset{d}{\sim} \delta_1$. The LST of $X_\theta$ can be obtained from the one of the first gamma$(\theta,\theta)$ example: $\mathbf{E}(e^{-xX_\theta}) = (1+x/\theta)^{-\theta}$ as $\theta \to \infty$. In this very particular (admissible) case, $\mathbf{S}_n = (1/n,\ldots,1/n)$ is the uniform deterministic partition of unity (the Maxwell-Boltzmann case).

### 4.2 The $*$-Limit

We now come back to the $*$-limit.

Let $\phi \in \mathcal{S}$. With $\gamma > 0$, let $(Y_\gamma)_{\gamma\geq 0}$ be a subordinator with $Y_0 = 0$ and LST

$$\mathbf{E}\big(e^{-xY_\gamma}\big) = e^{-\gamma\psi(x)}, \quad \psi(x) = -\phi(-x).$$

Under our assumptions on $\phi$, $\mathbf{E}(Y_\gamma) = \gamma\phi_1 < \infty$. Then the Laplace exponent $\psi$ reads

$$\psi(x) = \int_0^\infty \big(1 - e^{-xt}\big)\pi(dt), \tag{42}$$

for some positive Lévy measure $\pi$ on $(0,\infty)$, integrating $1 \wedge t$, [6]. Let $\overline{\pi}(t) := \int_t^\infty \pi(ds)$ be the tail function of $\pi$ and assume $\overline{\pi}(t) \to \infty$ as $t \to 0$.[4] Then

$$Y_\gamma = \sum_{k\geq 1} \overline{\pi}^{-1}(\Gamma_k/\gamma) \tag{43}$$

[4] If $\overline{\pi}$ has a finite limit, the random partition of unity defined in (46) is finite with a random Poisson number of pieces (see Example (iii) below). The corresponding subordinator has an atom at point $\gamma = 0$ with positive probability. This case deserves a special treatment.

where $(\Gamma_k)_{k\geq 1}$ are the points of a standard Poisson Point Process (PPP) on $(0,\infty)$ with intensity 1. The random variables

$$\Delta_{(k)}(\gamma) := \overline{\pi}^{-1}(\Gamma_k/\gamma)$$

with $\Delta_{(1)}(\gamma) \geq \Delta_{(2)}(\gamma) \geq \cdots$ constitute the ranked jumps' heights of the subordinator $Y_\gamma$ (they are countably many, with 0 as a limit point). They form a PPP on the half-line with intensity $\gamma\pi(dt)$, and the law of $\Delta_{(k)}(\gamma)$ can easily be computed to be [6]

$$\mathbf{P}\big(\Delta_{(k)}(\gamma) \in dt\big) = \frac{\gamma^k \overline{\pi}(t)^{k-1}}{(k-1)!} e^{-\gamma\overline{\pi}(t)} \pi(dt). \tag{44}$$

By Campbell formula (see [36, 40]), for all measurable function $g$ for which $\int_0^\infty (1 - e^{-xg(t)})\pi(dt) < \infty$, we have

$$\mathbf{E}\bigg(\exp\bigg\{-x\sum_{k\geq 1} g\big(\overline{\pi}^{-1}(\Gamma_k/\gamma)\big)\bigg\}\bigg) = \exp\bigg\{-\gamma\int_0^\infty \big(1 - e^{-xg(t)}\big)\pi(dt)\bigg\}.$$

Putting $g(t) = t$, $\mathbf{E}(e^{-xY_\gamma}) = e^{-\gamma\psi(x)}$, showing that (43) holds in law.

From the above construction, when $\pi$ has infinite mass, we can define a random distribution on the infinite-dimensional 1-simplex by normalizing the ranked jumps' heights of $Y_\gamma$ by itself. Consider again $Y_\gamma$ and, with $\theta := \gamma/n$, define $Y_{m,\theta} := Y_{m\theta} - Y_{(m-1)\theta}$, $m = 1,\ldots,n$ which are mutually independent. Then, $\overline{Y}_{n,\theta} := \sum_{m=1}^n Y_{m,\theta} = Y_{n\theta} = Y_\gamma$. If we rank the $Y_{m,\theta}$'s, with $Y_{(1),\theta} \geq \cdots \geq Y_{(n),\theta}$,[5] then, [34], as $n\to\infty$, $\theta \to 0$, $n\theta = \gamma$

$$(Y_{(1),\theta},\ldots,Y_{(n),\theta},0,0,\ldots) \overset{d}{\underset{*}{\to}} \big(\Delta_{(1)}(\gamma),\Delta_{(2)}(\gamma),\ldots\big). \tag{45}$$

Normalizing,

$$\begin{aligned}&(Y_{(1),\theta}/Y_\gamma,\ldots,Y_{(n),\theta}/Y_\gamma,0,0,\ldots)\\ &\quad\overset{d}{\underset{*}{\to}} \big(\Delta_{(1)}(\gamma)/Y_\gamma,\Delta_{(2)}(\gamma)/Y_\gamma,\ldots\big) =: \mathbf{S}_\infty(\gamma) := (S_{(1),\gamma},S_{(2),\gamma},\ldots),\end{aligned} \tag{46}$$

with $\mathbf{S}_\infty(\gamma)$ defining a random partition of unity with infinitely many (ordered) pieces.

If $t>0$ is some (small) cutoff or threshold value, let $N_+(t) := \sum_{k\geq 1}\mathbf{I}(\Delta_{(k)}(\gamma) > t)$ count the numbers of atoms of the partition of $Y_\gamma$ exceeding $t$. By Campbell formula

$$\begin{aligned}\mathbf{E}\big(\exp\{-xN_+(t)\}\big) &= \exp\bigg\{-\gamma\int_0^\infty \big(1 - e^{-x\mathbf{I}(s>t)}\big)\pi(ds)\bigg\}\\ &= \exp\big\{-\gamma\overline{\pi}(t)\big(1-e^{-x}\big)\big\}\end{aligned} \tag{47}$$

is the full LST of $N_+(t)$. This shows that $N_+(t)$ is Poisson distributed with mean $\gamma\overline{\pi}(t)$. Recalling $\overline{\pi}(t) \underset{t\to 0}{\to} \infty$, the law of large numbers gives

$$N_+(t)/\overline{\pi}(t) \overset{a.s.}{\to} \gamma, \quad \text{as } t\to 0. \tag{48}$$

[5]If $Y_\theta$ has a density ($\pi$ has no atom), these inequalities are strict.

The fact that $N_+(t)$ is Poisson may be also checked as follows. We have $N_+(t) = \inf(k \geq 1 : \Delta_{(k)}(\gamma) \leq t) - 1$ and $\mathbf{P}(N_+(t) \geq k) = \mathbf{P}(\Delta_{(k)}(\gamma) > t) = \mathbf{P}(\Gamma_k \leq \gamma\overline{\pi}(t)) = e^{-\gamma\overline{\pi}(t)} \sum_{l\geq k} \frac{[\gamma\overline{\pi}(t)]^l}{l!}$. So $N_+(t)$ is Poisson with mean $\gamma\overline{\pi}(t)$.

Because also, by the strong law of large numbers, $\Gamma_k/k \to 1$ a.s. as $k \to \infty$, recalling $\Gamma_k = \gamma\overline{\pi}(Y_\gamma S_{(k),\gamma})$, we get

$$\gamma\overline{\pi}(Y_\gamma S_{(k),\gamma})/k \to 1 \ a.s. \quad \text{as } k \to \infty.$$

From the behavior of $\overline{\pi}(t)$ near $t = 0$, the decay rate of $S_{(k),\gamma}$ to 0 as $k \to \infty$ follows.

Sampling from $S_{m,\theta} := Y_{m,\theta}/Y_\gamma$, $m = 1, \ldots, n$. Define as in (40) a biased sampling procedure for which ($|\mathbf{k}_n| = k$)

$$\mathbf{P}(\mathbf{K}_{n,k} = \mathbf{k}_n) = \binom{k}{k_1 \ldots k_n} \frac{\mathbf{E}(Y_\gamma^k \prod_{m=1}^n S_{m,\theta}^{k_m})}{\mathbf{E}(Y_\gamma^k)}. \tag{49}$$

Recall that this biased procedure is not the standard sampling one from a $k$ uniform throw on $S_{m,\theta}$, $m = 1, \ldots, n$, obtained while counting the number of uniform hits within each $S_{m,\theta}$. Indeed, would the latter sampling model hold, instead of (49), one would rather expect the strict multinomial occupancy distribution

$$\mathbf{Q}(\mathbf{K}_{n,k} = \mathbf{k}_n) = \binom{k}{k_1 \ldots k_n} \mathbf{E}\left(\prod_{m=1}^n (Y_{m,\theta}/Y_\gamma)^{k_m}\right),$$

and in general, we have $\mathbf{Q}(\mathbf{K}_{n,k} = \mathbf{k}_n) \neq \mathbf{P}(\mathbf{K}_{n,k} = \mathbf{k}_n)$. According to (49), the joint pgf of $\mathbf{K}_{n,k}$ is

$$\begin{aligned}
\mathbf{E}\left(\prod_{m=1}^n u_m^{K_{n,k}(m)}\right) &= \frac{1}{\mathbf{E}(Y_\gamma^k)} \sum_{\mathbf{k}_n \in \mathbb{N}_0^n :\, |\mathbf{k}_n| = k} \binom{k}{k_1 \ldots k_n} \prod_{m=1}^n u_m^{k_m} \mathbf{E}\left(\prod_{m=1}^n Y_{m,\theta}^{k_m}\right) \\
&= \frac{\mathbf{E}[(\sum_{m=1}^n u_m Y_{m,\theta})^k]}{\mathbf{E}(Y_\gamma^k)},
\end{aligned} \tag{50}$$

which is akin to (40).

Biased sampling from $\mathbf{S}_\infty(\gamma) = (S_{(1),\gamma}, S_{(2),\gamma}, \ldots)$ can also be defined whenever the sampling process amounts to draw $k$ points at random in the unit interval partitioned according to $\mathbf{S}_\infty(\gamma)$, counting the number of points in each subintervals and when biasing some functional $f(S_{(1),\gamma}, S_{(2),\gamma}, \ldots)$ under concern to produce $\mathbf{E}^*(Y_\gamma^k f(S_{(1),\gamma}, S_{(2),\gamma}, \ldots))/\mathbf{E}^*(Y_\gamma^k)$ when averaging over $\mathbf{S}_\infty(\gamma)$.

From these considerations, we can state the following results:

**Proposition 12** *Let $\gamma = n\theta$. When $\phi \in \mathcal{S}$, with $(\sigma_k(\theta), k \geq 0)$ the Stieltjes moment sequence of some infinitely divisible subordinator $Y_\gamma$ with Laplace exponent $\psi(x) = -\phi(-x)$, the occupancy distributions* (12), (15) *and* (24) *are biased sampling multinomial distributions from $S_{m,\theta} := Y_{m,\theta}/Y_\gamma$, $m = 1, \ldots, n$ as defined by* (49).

**Corollary 13** *When $\phi \in \mathcal{S}$ and $\pi$ has infinite mass ($\phi(x) \underset{x\to-\infty}{\to} -\infty$), the occupancy distributions* (27), (31) *and* (32) *are biased sampling multinomial distributions from $\mathbf{S}_\infty(\gamma) = (S_{(1),\gamma}, S_{(2),\gamma}, \ldots)$ defined in* (46) *from the subordinator $Y_\gamma$ with Laplace exponent $\psi(x) = -\phi(-x)$.*

*Proof* The proof follows from the previous Proposition, the fact that (27) and (32) were obtained as weak $*$-limits of (15) and (24), from (46) and from exchangeability of the $K_{n,k}(m)$'s. $\square$

Let us now illustrate Corollary 13. For instance, when $\phi \in \mathcal{S}$, from (27),

$$\begin{aligned}
&\mathbf{P}^*\big(\widehat{K}_k(1)=k_1,\ldots,\widehat{K}_k(p)=k_p;\, P_k=p\big) \\
&\quad = \frac{k!}{p!}\frac{\gamma^p}{\sigma_k(\gamma)}\prod_{q=1}^{p}\frac{\phi_{k_q}}{k_q!} = \frac{\mathbf{E}^*(Y_\gamma^k\sum_{1\le m_1<\cdots<m_p}\prod_{q=1}^{p}S_{(m_q),\gamma}^{k_q})}{\mathbf{E}^*(Y_\gamma^k)}
\end{aligned}$$

is the probability that there are $p$ observed species, labeled in arbitrary way, in the $k$-sample, each visited $k_q$ times, and that they were obtained after biased sampling from $S_{(m_1),\gamma}>\cdots> S_{(m_p),\gamma}$ for any ordered sequence $1\le m_1<\cdots<m_p$.

In particular, the probability that, in a biased sampling procedure from $\mathbf{S}_\infty(\gamma)$, all elements of the $k$-sample are of the same species (whichever species it can be) is thus

$$\mathbf{P}^*\big(\widehat{K}_k(1)=k;\, P_k=1\big) = \frac{\mathbf{E}^*(Y_\gamma^k\sum_{m\ge1}S_{(m),\gamma}^k)}{\mathbf{E}^*(Y_\gamma^k)} = \gamma\frac{\phi_k}{\sigma_k(\gamma)} = \mathbf{E}^*\big(A_k(k)\big). \tag{51}$$

The latter identity also follows from (31) with $a_1=\cdots=a_{k-1}=0$, $a_k=1$ and $p=1$ (only one species visited $k$ times).

We observe that, as $\gamma\to0$ (or $\mathbf{E}^*(Y_\gamma)\to0$ as well), due to $\sigma_k(\gamma)\sim\gamma\phi_k$, this probability tends to 1, showing that $\gamma$ itself may be viewed as some temperature parameter for the population with infinitely many species: the smaller $\gamma$, the larger the probability is that any $k$-sample visits a single one species (among which the one with largest frequency $S_{(1),\gamma}$).

Similarly, the probability that all elements of the $k$-sample reveal only two species (whichever species they can be) is

$$\sum_{l=1}^{k-1}\frac{\mathbf{E}^*(Y_\gamma^k\sum_{1\le m_1<m_2}S_{(m_1),\gamma}^{l}S_{(m_2),\gamma}^{k-l})}{\mathbf{E}^*(Y_\gamma^k)} = \frac{1}{2}\frac{\gamma^2 k!}{\sigma_k(\gamma)}\sum_{l=1}^{k-1}\frac{\phi_l}{l!}\frac{\phi_{k-l}}{(k-l)!} = \frac{\gamma^2}{\sigma_k(\gamma)}B_{k,2}(\phi_\bullet).$$

This identity follows from (31) with $a_l=1, a_{k-l}=1, a_j=0$ if $j\ne\{l,k-l\}$ and $p=2$ (only two species visited, one $l$ times and the other one $k-l$ times), summing on $l=1,\ldots,k-1$ and from $\phi_k^{*2}=2B_{k,2}(\phi_\bullet)$. More generally, if $p\le k$, $\frac{\gamma^p}{\sigma_k(\gamma)}B_{k,p}(\phi_\bullet)$ is the probability that all elements of the $k$-sample reveal $p$ distinct species (consistently with (28)), $\frac{(\gamma\phi_1)^k}{\sigma_k(\gamma)}$ the probability that all species in the $k$-sample are of distinct types. When $\gamma$ is small this latter probability is polynomially small $\sim\gamma^{k-1}$.

Finally, the probability that only one species is visited by the $k$-sample and that it is the $m$th more abundant one is

$$\begin{aligned}
\frac{\mathbf{E}^*(Y_\gamma^k S_{(m),\gamma}^k)}{\mathbf{E}^*(Y_\gamma^k)} &= \frac{\mathbf{E}^*(\Delta_{(m)}(\gamma)^k)}{\mathbf{E}^*(Y_\gamma^k)} = \frac{1}{(m-1)!}\frac{\int_0^\infty e^{-x}x^{m-1}\overline{\pi}^{-1}(x/\gamma)^k dx}{\sigma_k(\gamma)} \\
&= \frac{\gamma}{\sigma_k(\gamma)}\frac{1}{(m-1)!}\int_0^\infty t^k\big(\gamma\overline{\pi}(t)\big)^{m-1}e^{-\gamma\overline{\pi}(t)}\pi(dt),
\end{aligned} \tag{52}$$

consistently with (44). Summing (52) over $m\ge1$, we recover from (51), that $\phi_k=\frac{1}{\gamma}\int_0^\infty\overline{\pi}^{-1}(x/\gamma)^k dx=\int_0^\infty t^k\pi(dt)$ is the $k$th moment of the Lévy measure $\pi$. In particular, the probability that only one species is visited by the $k$-sample *and* that it is the more abundant one is (compare with (51))

$$\frac{\mathbf{E}^*(Y_\gamma^k S_{(1),\gamma}^k)}{\mathbf{E}^*(Y_\gamma^k)} = \frac{\gamma}{\sigma_k(\gamma)} \int_0^\infty t^k e^{-\gamma\overline{\pi}(t)} \pi(dt)$$
$$= \frac{\gamma\phi_k}{\sigma_k(\gamma)} \left[1 - \frac{1}{\phi_k} \int_0^\infty t^k \left(1 - e^{-\gamma\overline{\pi}(t)}\right)\pi(dt)\right].$$

When $\gamma$ gets very small, this probability approaches 1 from below, up to an $\mathcal{O}(\gamma)$ residual term: again, $S_{(1),\gamma}$ dominates the other smaller $S_{(m),\gamma}$ and for small values of the biodiversity parameter $\gamma$ therefore, the species frequencies $S_{(m),\gamma}$; $m \geq 1$ turn out to be very disparate.

Similarly, from (32), when $\phi \in \mathcal{S}$

$$\mathbf{P}^*\left(A_k(1) = a_1, \ldots, A_k(k) = a_k;\, P_k = p\right)$$
$$= \frac{\gamma^p k!}{\sigma_k(\gamma)} \prod_{i=1}^k \frac{(\phi_i / i!)^{a_i}}{a_i!} = \frac{k!}{\prod_{i\geq 1}(i!^{a_i} a_i!)} \frac{\mathbf{E}^*(Y_\gamma^k \sum \prod_{i\geq 1} \prod_{j=1}^{a_i} S_{(m_{i,j}),\gamma}^i)}{\mathbf{E}^*(Y_\gamma^k)}$$

where in the latter numerator, the unindexed sum runs over all distinct $(m_{i,j})$, $i = 1, \ldots, k$; $j = 1, \ldots, a_i$ with $(a_1, a_2, \ldots)$ satisfying $\sum_{i\geq 1} i a_i = k$ and $\sum_{i\geq 1} a_i = p$.

## 5 Examples

Let us supply some Examples illustrating our results.

(i) Take the Fisher logarithmic series model $\phi(x) = -\log(1-x) \in \mathcal{S}$, resulting in $\xi$ obeying a negative binomial distribution with parameters $\theta > 0$ and $1 - x \in (0, 1)$, [21]. Here $\phi_\bullet = (\bullet - 1)!$. Then $Y_\gamma$ is a Moran subordinator with Lévy-measure: $\pi(dt) = t^{-1} e^{-t} dt$. The Laplace exponent of $Y_\gamma$ is $\psi(x) = \log(1 + x)$, in accordance with $\psi(x) = -\phi(-x)$. In that particular case, $(S_{(1),\gamma}, S_{(2),\gamma}, \ldots) \sim PD(0, \gamma)$, a Poisson-Dirichlet partition with parameter $\gamma$, [20, 26]. Because, due to well-known properties of Gamma-distributed random variables, $Y_\gamma$ is independent of $S_{m,\theta} = Y_{m,\theta}/Y_\gamma$, $m = 1, \ldots, n$, the biased sampling distributions from $(S_{1,\theta}, \ldots S_{n,\theta})$ corresponds to the usual multinomial one. In this well-known model for species frequency, $\sigma_k(\theta) = (\theta)_k$. So $\sigma_k(\theta) \in ZR_-$.

Because $\overline{\pi}(t) \sim -\log t$ as $t \to 0$, $N_+(t) := \#\{k : \Delta_{(k)}(\gamma) > t\}$ grows like $-\gamma \log t$ as $t \to 0$. Besides,

$$-\log S_{(k),\gamma} \sim k/\gamma \quad \text{as } k \to \infty$$

and the ordered frequencies decay exponentially fast with $k$: species with small frequency get exponentially rare.

Assuming $\theta$ known, the Maximum Likelihood Estimator (MLE) estimator of $n$ in the finitely many species model is given implicitly by $P = \widehat{n}(1 - \frac{\sigma_k((\widehat{n}-1)\theta)}{\sigma_k(\widehat{n}\theta)})$, so here

$$P = \widehat{n}\left(1 - \frac{((\widehat{n}-1)\theta)_k}{(\widehat{n}\theta)_k}\right).$$

When $\theta = 1$, this estimator is explicitly given by

$$\widehat{n} = \frac{(k-1)P}{k-P},$$

where, as conventional wisdom suggests, $\widehat{n}$ will be large when the difference between $1/P$ and $1/k$ is small (new species are being frequently discovered). The MLE estimator of $\gamma$ in the infinitely many species model is given implicitly by $P = \widehat{\gamma}\frac{\sigma_k'(\widehat{\gamma})}{\sigma_k(\widehat{\gamma})}$, [51], so here

$$P = \sum_{l=0}^{k-1} \frac{\widehat{\gamma}}{\widehat{\gamma}+l}.$$

The estimator $\widehat{\gamma}$ is biased but its bias decreases as $k$ grows. The alternative estimator $\widetilde{\gamma} = \frac{B_{k,P-1}(\phi_\bullet)}{B_{k,P}(\phi_\bullet)}$ with $B_{k,p}(\phi_\bullet) = s_{k,p}$ is also biased and can be computed using the recursion for third kind Stirling numbers

$$B_{k+1,p}\big((\bullet-1)!\big) = B_{k,p-1}\big((\bullet-1)!\big) + kB_{k,p}\big((\bullet-1)!\big).$$

(ii) The full two-parameters $PD(\alpha,\gamma)$ defined in [43] can be obtained while subordinating the damped $\alpha$-stable subordinator (see (iii) below) to an independent Moran one with parameter $\gamma/\alpha$. And considering the normalized ranked sizes of the subordinate jumps: here, independently of this partition of unity, $Y_\gamma$ again is gamma($\gamma$) distributed. As shown in [43], $PD(\alpha,\gamma)$ has many interesting properties, [20, 41]. This partition of unity leads to a generalized (unbiased) Ewens' sampling formula called Pitman's sampling formula, [42]. Connection of the two-parameters $PD(\alpha,\gamma)$ partition to Gibbs (EPPF) partitions and a complete classification of EPPFs induced by the unbiased multinomial sampling from partition of unity can be found in [25] and [23].

(iii) Take $\phi(x) = (1-x)^{-\alpha} - 1 \in \mathcal{S}$ where $\alpha > 0$. Here $\phi_\bullet = (\alpha)_\bullet$ resulting in $\xi$ being a Poisson sum of negative binomial increments $\delta$. The Lévy-measure corresponding to $Y_\gamma$ is the (mean $\alpha$) Gamma($\alpha$, 1) probability density: $\pi(dt) = 1/\Gamma(\alpha)\cdot t^{\alpha-1}e^{-t}dt$. The Laplace exponent of $Y_\gamma$ is $\psi(x) = 1-(1+x)^{-\alpha}$, in accordance with $\psi(x) = -\phi(-x)$. Because $\pi$ is integrable with mass 1, $Y_\gamma$ is a subordinator in the compound Poisson class (a Poisson($\gamma$) sum of iid positive jumps with Gamma($\alpha$, 1) density). For this reason,

$$Y_\gamma \overset{d}{=} \left[\sum_{k=1}^{P(\gamma)} \overline{\pi}^{-1}(U_{[k]})\right]\cdot \mathbf{I}\big(P(\gamma)\geq 1\big) + 0\cdot \mathbf{I}\big(P(\gamma)=0\big),$$

where $(U_{[k]}; k\geq 1)$ are the ranked $(U_{[1]} < \cdots < U_{[P_\gamma]})$ points of an iid uniform sequence $(U_k; k\geq 1)$ on $(0,1)$, independent of $P(\gamma)$ which is Poisson($\gamma$) distributed. Note that $Y_\gamma$ has an atom at $Y_\gamma = 0$ with positive probability and that, would $P(\gamma)\geq 1$, there are finitely many (Poissonian) terms in the Lévy decomposition of $Y_\gamma$. In this case, the random variables

$$\Delta_{(k)}(\gamma) := \overline{\pi}^{-1}(U_{[k]}); \quad k = 1,\ldots,P(\gamma)$$

with $\Delta_{(1)}(\gamma)\geq\cdots\geq\Delta_{(P(\gamma))}(\gamma)$ constitute the ranked (non-null) jumps' heights of the subordinator $Y_\gamma$. Considering $Y_\gamma$ on the event $P(\gamma)\geq 1$, with $\theta := \gamma/n$, the spacings $Y_{m,\theta}$ defined by $Y_{m,\theta} := Y_{m\theta} - Y_{(m-1)\theta}$, $m = 1,\ldots,n$ are non-negative and mutually independent; also $\overline{Y}_{n,\theta} := \sum_{m=1}^n Y_{m,\theta} = Y_{n\theta} - Y_0 = Y_\gamma > 0$. Normalizing the $Y_{m,\theta}$'s with $Y_\gamma$ defines a proper finite random partition of unity $S_m$ with a random number of non-zero parts and bias sampling (with $\pi$ finite with mass 1) is therefore to be understood from this partition. In the $*$-limit, its ranked (non-null) jumps' heights are the $\Delta_{(k)}(\gamma)$'s. Note that when $\pi$ is integrable with mass 1, the biodiversity parameter $\gamma$ takes on directly the interpretation of the expected number of species in the population.

Let us come back to our case study. We first recall that for $\phi_\bullet = (\alpha)_\bullet$

$$B_{k+1,p}(\phi_\bullet) = \alpha B_{k,p-1}(\phi_\bullet) + (k + p\alpha) B_{k,p}(\phi_\bullet).$$

When $\alpha = 1$, $B_{k,p}(\bullet!) = \binom{k-1}{p-1}\frac{k!}{p!}$ are the Lah numbers.

Recalling also $\mathbf{P}^*(P_k = p) = \frac{\gamma^p}{\sigma_k(\gamma)} B_{k,p}(\phi_\bullet)$, we get the recursion

$$\begin{aligned}\mathbf{P}^*(P_{k+1} = p) &= \frac{\gamma^p}{\sigma_{k+1}(\gamma)}\big(\alpha B_{k,p-1}(\phi_\bullet) + (k + p\alpha) B_{k,p}(\phi_\bullet)\big)\\ &= \frac{\sigma_k(\gamma)}{\sigma_{k+1}(\gamma)}\big(\alpha\gamma \mathbf{P}^*(P_k = p-1) + (k + p\alpha)\mathbf{P}^*(P_k = p)\big).\end{aligned}$$

This shows that the event $P_{k+1} = p$ only depends on the event $P_k = p - 1$ (respectively $P_k = p$), when a new species (respectively no new species) is being discovered as the sample size is increased by one unit. And not on further past events such as $P_l = p - 1$ for $p - 1 \le l < k$. The transition rates are $\lambda_{p,p+1} = \alpha\gamma\frac{\sigma_k(\gamma)}{\sigma_{k+1}(\gamma)}$ (independent of $p$ but dependent on $k$) and $\lambda_{p,p} = (k + p\alpha)\frac{\sigma_k(\gamma)}{\sigma_{k+1}(\gamma)}$. $\lambda_{p,p+1}$ is the rate at which a new species is being discovered given $p$ of them were previously discovered in a size-$k$ sample. This suggests an underlying sequential urn scheme, [7, 51].

The estimator $\widetilde{\gamma} = \frac{B_{k,P-1}(\phi_\bullet)}{B_{k,P}(\phi_\bullet)}$ of $\gamma$ can easily be evaluated numerically thanks to the three-term recurrence which $B_{k,p}(\phi_\bullet)$ fulfills. When $\alpha = 1$, it is

$$\widetilde{\gamma} = \frac{P(P-1)}{k - P + 1} = \frac{P}{k}\frac{1}{\frac{1}{P-1} - \frac{1}{k}}.$$

For the four following examples, an appeal to length-biased sampling distributions from $\mathbf{S}_\infty(\gamma)$ is required.

(iv) With $\alpha \in (0,1)$, take $\phi(x) = 1 - (1 - x)^\alpha \in \mathcal{S}$, resulting in $\xi$ being a Poisson sum of extended negative binomial increments $\delta$ (also called a Poisson-Pascal random variable). Here $\phi_1 = \alpha$, $\phi_m = \alpha(1-\alpha)_{m-1}$, $m \ge 1$ and the weight of large clusters is smaller than in Example (i) where $\phi_m = (m-1)!$. We therefore expect small clusters sizes to be enhanced. In this case, $Y_\gamma$ is a damped $\alpha$-stable subordinator with Lévy-measure: $\pi(dt) = \alpha/\Gamma(1-\alpha)\cdot t^{-(\alpha+1)}e^{-t}dt$. The Laplace exponent of $Y_\gamma$ is $\psi(x) = (1 + x)^\alpha - 1$, in accordance with $\psi(x) = -\phi(-x)$. The relevant subordinator is termed the generalized gamma (see [23, 44] and [25]).

Because $\overline{\pi}(t) \sim 1/\Gamma(1-\alpha)\cdot t^{-\alpha}$ as $t \to 0$, $N_+(t) := \#\{k : \Delta_{(k)}(\gamma) > t\}$ grows like $\gamma/\Gamma(1-\alpha)\cdot t^{-\alpha}$ as $t \to 0$. Besides,

$$S_{(k),\gamma} \sim \left(\frac{\gamma}{\Gamma(1-\alpha)}\right)^{1/\alpha} Y_\gamma^{-1} k^{-1/\alpha} \quad \text{as } k \to \infty$$

and the ordered frequencies only decay algebraically fast with $k$. Species with small frequency are long-tailed (there are many small size groups or rare species in the Engen model, compared to the Ewens model).

In this model, $\phi_\bullet = \alpha(1-\alpha)_{\bullet-1}$. Because $\phi_1 = \alpha$ and $\phi_{m+1} = \phi_m(m-\alpha)$, $m \geq 1$, it follows from (3), (4) that $\sigma_{k+1}(\theta) = (\theta\alpha + k)\sigma_k(\theta) - \theta\alpha\sigma_k'(\theta)$. Thus, the Bell coefficients $B_{k,p}(\phi_\bullet)$, appearing in (16), again obey a simple 3-term recurrence

$$B_{k+1,p}(\phi_\bullet) = \alpha B_{k,p-1}(\phi_\bullet) + (k - p\alpha)B_{k,p}(\phi_\bullet).$$

They constitute generalized Stirling numbers studied by [9]. It can be checked that $\sigma_k(\theta) \notin ZR_-$.

This model is amenable to similar conclusions as the ones from the previous example with recursion now given by

$$\mathbf{P}^*(P_{k+1} = p) = \frac{\sigma_k(\gamma)}{\sigma_{k+1}(\gamma)}\big(\alpha\gamma\mathbf{P}^*(P_k = p-1) + (k - p\alpha)\mathbf{P}^*(P_k = p)\big).$$

Equation (32) with $\phi_\bullet = \alpha(1-\alpha)_{\bullet-1}$ is the Engen's extended negative binomial sampling formula [27]. The particular case $\alpha = 1/2$ is studied in [28]. The microcanonical distribution (33) coincides when $\phi_\bullet = \alpha(1-\alpha)_{\bullet-1}$ with the one occurring in the Pitman sampling formula ([27], Remark 3).

(v) Let $\phi(x)$ solve the functional equation $\phi(x) = x\exp\phi(x)$. Then $\phi(x) = \sum_{m\geq 1}\frac{\phi_m}{m!}x^m$ with $\phi_m = m^{m-1}$ is the Cayley generating function appearing in the enumeration of rooted labeled trees with $m$ nodes. The convergence radius of this series is $x_0 = e^{-1}$ with $\phi(x_0) = 1$ and $\phi'(x_0) = \infty$. Clearly $\phi_m$ is log-convex, it is a Stieltjes moment sequence and $\phi \in \mathcal{S}$. The associated Laplace exponent $\psi(x) = -\phi(-x)$ is the Lambert function. Because $\psi(x) \sim \log x$ as $x \to \infty$, $\overline{\pi}(t) \sim -\log t$ as $t \to 0$ and $N_+(t) := \#\{k : \Delta_{(k)}(\gamma) > t\}$ grows like $-\gamma\log t$ as $t \to 0$. Besides, like in Example (i)

$$-\log S_{(k),\gamma} \sim k/\gamma \quad \text{as } k \to \infty.$$

The partition function $Z_\theta(x) = \exp\theta\phi(x)$ occurs in the enumeration of forests of Cayley trees. The Bell coefficients are $B_{k,p}(\phi_\bullet) = \binom{k-1}{p-1}k^{k-p}$ (number of forests with $k$ nodes and $p$ trees) in accordance with the global weights $\sigma_k(\theta) = \theta(k+\theta)^{k-1}$. So $\sigma_k(\theta) \in ZR_-$. Assuming $\theta$ known, the MLE estimator of $n$ in the finitely many species model is given implicitly by $P = \widehat{n}(1 - \frac{\sigma_k((\widehat{n}-1)\theta)}{\sigma_k(\widehat{n}\theta)})$, so here

$$P = \widehat{n} - (\widehat{n} - 1)\left(1 - \frac{\theta}{k + \widehat{n}\theta}\right)^{k-1}.$$

The MLE estimator of $\gamma$ in the infinitely many species model is given by $P = \widehat{\gamma}\frac{\sigma_k'(\widehat{\gamma})}{\sigma_k(\widehat{\gamma})}$, so here explicit

$$\widehat{\gamma} = \frac{k(P-1)}{k-P}.$$

The alternative (biased) estimator is $\widetilde{\gamma} = \frac{B_{k,P-1}(\phi_\bullet)}{B_{k,P}(\phi_\bullet)}$. Thus

$$\widetilde{\gamma} = \frac{k(P-1)}{k-P+1} = \frac{1}{\frac{1}{P-1} - \frac{1}{k}};$$

it is also explicit and very close to $\widehat{\gamma}$.

(vi) As a next example, let $\phi(x)$ solve the functional equation $\phi(x) = xg(\phi(x))$ where $g(x) = (1+bx)^a$ with either $b > 0$ and $a \geq 1$ or $a$ and $b$ both negative. $\phi(x)$ is the generating function appearing in the enumeration of rooted trees when the generating function $g$ of the offspring is either (generalized) binomial or negative binomial. Then $\phi_m = (m-1)!\binom{am}{m-1}b^{m-1}$ are non-negative numbers. We conjecture that $\phi \in \mathcal{S}$. It holds [31] that $x_0 = (ab)^{-1}(1-1/a)^{a-1}$ with $\phi(x_0) = 1/(b(a-1))$ and $\phi'(x_0) = \infty$. For this tree model first discussed in [3], the Lagrange inversion formula gives [1]

$$B_{k,p}(\phi_\bullet) = \binom{k-1}{p-1}\{ak\}_{k-p}b^{k-p},$$

where $\{a\}_l := a(a-1)\cdots(a-l+1)$. Recalling $\widetilde{\gamma} = \frac{B_{k,P-1}(\phi_\bullet)}{B_{k,P}(\phi_\bullet)}$, we get

$$\widetilde{\gamma} = \frac{b(P-1)}{k-P+1}\big((a-1)k + P\big) = \frac{b}{\frac{1}{P-1} - \frac{1}{k}}\left(a - 1 + \frac{P}{k}\right),$$

which is explicit. Again, would $1/k$ be close to $1/(P-1)$, then $\widetilde{\gamma}$ would be estimated to be large. Would $a \to \pm\infty$, $b \to \pm 0$ while $ab \to 1$, we recover the results just obtained for Cayley trees (consistently with $g(x) = (1+bx)^a \to e^x$). If $a = b = 1$, we recover Example (iii) with $\alpha = 1$. When $k$ is large, the minimum of $B_{k,p}^2(\phi_\bullet)/(B_{k,p-1}(\phi_\bullet)B_{k,p+1}(\phi_\bullet))$ is attained when $p = [\lambda k]$ for some $\lambda \in (0,1)$, with value

$$\frac{\lambda}{1-\lambda}\,\frac{(1-\lambda)k+1}{\lambda k - 1}\,\frac{(a-1+\lambda)k+1}{(a-1+\lambda)k} \underset{k\to\infty}{\to} 1$$

and the sequence $B_{k,p}(\phi_\bullet)$ is $p$-log-concave.

(vii) Let $\alpha > 0$ and let $\phi(x) = \sum_{m\geq 1} m^{-\alpha}x^m$ be the polylog function. The convergence radius of this series is $x_0 = 1$ with $\phi(x_0) < \infty$ iff $\alpha > 1$ and $\phi'(x_0) < \infty$ iff $\alpha > 2$. $\phi(x)$ is defined for $x < x_0$ and $\phi(x) \to -\infty$ as $x \to -\infty$. We have $\phi_m = m!m^{-\alpha}$ and $(\phi_m)_{m\geq 1}$ constitutes a log-convex sequence because for all $m \geq 2$,

$$\begin{aligned}\phi_{m+1}\phi_{m-1} &= (m+1)!(m-1)!\big(m^2-1\big)^{-\alpha}\\ &> (m+1)!(m-1)!m^{-2\alpha} > m!^2 m^{-2\alpha} = \phi_m^2.\end{aligned}$$

The sequence $\phi_m$ satisfies Carleman's condition $\sum_{m\geq 1}\phi_m^{-1/(2m)} = \infty$. Thus $\phi \in \mathcal{S}$ and $\psi(x) = -\phi(-x)$, $x > -1$, is the Laplace exponent of some polylog subordinator with Lévy measure $\pi$. Because $\phi(x) \sim -[\log(-x)]^\alpha/\Gamma(1+\alpha)$ as $x \to -\infty$, [11], $-\phi(-x) =: \psi(x) \to \infty$ as $x \to \infty$ and $\pi$ has infinite total mass. In this example, when $\alpha > 1$, the weight of large clusters $\phi_m$ is smaller than in Example (i) where $\phi_m = (m-1)!$. When $\alpha > 1$, we therefore expect small clusters sizes to be enhanced as in Example (iv), but to a lesser extent. Because indeed $\overline{\pi}(t) \sim [-\log t]^\alpha/\Gamma(1+\alpha)$ as $t \to 0$, $N_+(t) := \#\{k : \Delta_{(k)}(\gamma) > t\}$ grows like $\gamma[-\log t]^\alpha/\Gamma(1+\alpha)$ as $t \to 0$. Besides,

$$-\log S_{(k),\gamma} \sim \big(\Gamma(1+\alpha)/\gamma\big)^{1/\alpha}k^{1/\alpha} \quad \text{as } k \to \infty$$

and the ordered frequencies decay exponentially fast, but now with $k^{1/\alpha}$ (in a 'stretched exponential' Weibull way).

(viii) As another example with $\phi \in \mathcal{S}$ but with $\pi$ integrable, consider the Mittag-Leffler function $\phi(x) = \sum_{m\geq 1} \frac{1}{\Gamma(1+m\alpha)} x^m$, where $\alpha \in (0,1)$. We have $\psi(x) := -\phi(-x) =: 1 - \varphi(x)$ where

$$\varphi(x) := \sum_{m\geq 0} \frac{1}{\Gamma(1+m\alpha)} (-x)^m.$$

$\varphi(x)$ is the Mittag-Leffler LST of the random variable $S_\alpha^{-\alpha}$ where $S_\alpha$ is an $\alpha$-stable random variable with LST $\mathbf{E}(e^{-xS_\alpha}) = e^{-x^\alpha}$, [46]. Here $\phi_\bullet = \frac{\Gamma(1+\bullet)}{\Gamma(1+\alpha\bullet)}$ and because of the latter link with the Mittag-Leffler LST, the $\phi_\bullet$ sequence is log-convex and $\phi \in \mathcal{S}$. For this model, the discrete abundance $\xi$ is thus a Poisson sum of discrete Mittag-Leffler increments $\delta$ with

$$\mathbf{P}(\delta = m) = \frac{1}{\Gamma(1+m\alpha)} \frac{x^m}{\phi(x)}, \quad m \geq 1.$$

In the bias sampling from a random partition point of view, the Lévy-measure corresponding to $Y_\gamma$ is $\pi(dt) = f_\alpha(t)dt$ where $f_\alpha(t)$ is the density of $S_\alpha^{-\alpha}$. The Laplace exponent of $Y_\gamma$ is $\psi(x) = -\phi(-x)$. Because $\pi$ is integrable with mass 1, $Y_\gamma$ is a subordinator in the compound Poisson class (a Poisson$(\gamma)$ sum of iid positive jumps with Mittag-Leffler density $f_\alpha(t)$). In the Mittag-Leffler case, the bias sampling is again from a finite random partition of unity, as in Example (iii). Note that as $\alpha \to 0$, $\phi(x) \sim (1-x)^{-1} - 1$ (which is a particular case of (iii)) whereas when $\alpha \to 1$, $\phi(x) \sim e^x - 1$ which is the Bell model, also in the $\mathcal{S}$ class.

(ix) Let $\phi(x)$ solve the functional equation $\phi(x) = xg(\phi(x))$ where $g(x) = 1 + x^2/2$. Then $\phi(x) = (1 - \sqrt{1-2x^2})/x$ is the generating function appearing in the enumeration of rooted binary labeled trees. Only the odd $\phi_m$'s are non-zero. The convergence radius of this series is $x_0 = 1/\sqrt{2}$ with $\phi(x_0) = \sqrt{2}$ and $\phi'(x_0) = \infty$. Clearly $\phi \notin \mathcal{S}$ because $\phi$ is only defined on $|x| \leq x_0$, so not absolutely monotone on $(-\infty, x_0)$.

## 6 A New Engen-Like Example

We end up giving a new example of $\xi$ sharing some common issues with the Engen's model.

**Preliminaries** Previously, let us start with a general fact. Let $\phi^\star(x)$ be some 'local' generating function with non-negative coefficients $\phi_m^\star$. Define $Z_1^\star(x) = \exp \phi^\star(x)$, together with $\sigma_k^\star(\theta)$, the Bell polynomials associated to $\phi^\star(x)$: $Z_1^\star(x)^\theta =: 1 + \sum_{k\geq 1} \frac{\sigma_k^\star(\theta)}{k!} x^k$. Define now the new generating functions

$$\phi(x) = xZ_1^\star(x) \quad \text{and} \quad Z_\theta(x) = \exp\bigl(\theta\phi(x)\bigr).$$

The Taylor coefficients of $\phi$ are: $\phi_m = m\sigma_{m-1}^\star(1)$. The Bell polynomials now associated to $\phi(x)$ are: $Z_\theta(x) = 1 + \sum_{k\geq 1} \frac{\sigma_k(\theta)}{k!} x^k$, with

$$\sigma_k(\theta) = \sum_{p=1}^{k} B_{k,p}\bigl(\bullet\sigma_{\bullet-1}^\star(1)\bigr)\theta^p.$$

Because $\sigma_k^\star(\theta)$ are binomial convolution polynomials, the following identity holds, [1]

$$B_{k,p}\bigl(\bullet\sigma_{\bullet-1}^\star(1)\bigr) = \binom{k}{p} \sigma_{k-p}^\star(p). \tag{53}$$

Three simple examples are:

- $\phi^\star(x) = \alpha x$, $\alpha > 0$. Then $\sigma_k^\star(\theta) = \alpha^k\theta^k$ leading to: $B_{k,p}(\bullet\alpha^{\bullet-1}) = \binom{k}{p}(\alpha p)^{k-p}$.
- $\phi^\star(x) = e^{\alpha x} - 1$, $\alpha > 0$. Then $\sigma_k^\star(\theta) = \alpha^k \sum_{p=1}^k S_{k,p}\theta^p$ (where $S_{k,p}$ are the second kind Stirling numbers), leading to: $B_{k,p}(\alpha^{\bullet-1}B_{\bullet-1}) = \binom{k}{p}\alpha^{k-p}\sum_{q=1}^{k-p} S_{k-p,q}p^q$ where $B_k = \sum_{p=1}^k S_{k,p}$ are the Bell numbers.
- $\phi^\star(x)$ solves $\phi^\star(x) = x\exp(\alpha\phi^\star(x))$, $\alpha > 0$. Then $\sigma_k^\star(\theta) = \sum_{p=1}^k B_{k,p}(\phi_\bullet^\star)\theta^p$ with $B_{k,p}(\phi_\bullet^\star) = \binom{k-1}{p-1}(\alpha k)^{k-p}$, leading to

$$\sigma_k^\star(\theta) = \theta(\theta + \alpha k)^{k-1}.$$

We conclude that, with $\phi_\bullet = \bullet(1 + \alpha(\bullet - 1))^{\bullet-2}$

$$B_{k,p}(\phi_\bullet) = \binom{k}{p} p\big(p + \alpha(k-p)\big)^{k-p-1}.$$

If $\alpha = 1$, $\phi_\bullet = \bullet(1 + \alpha(\bullet - 1))^{\bullet-2} = \bullet^{\bullet-1}$ and we recover $B_{k,p}(\bullet^{\bullet-1}) = \binom{k}{p}pk^{k-p-1} = \binom{k-1}{p-1}k^{k-p}$.

*Example* Let $\phi^\star(x) = -\alpha\log(1-x)$, $\alpha > 0$. Then $\sigma_k^\star(\theta) = (\alpha\theta)_k$. Looking at $\phi(x) = x\exp\phi^\star(x)$ and

$$Z_\theta(x) = \exp\big(\theta\phi(x)\big) = e^{\theta x(1-x)^{-\alpha}},$$

with $\phi_\bullet = \bullet(\alpha)_{\bullet-1}$, we get $\sigma_k(\theta) = \sum_{p=1}^k B_{k,p}(\phi_\bullet)\theta^p$ where

$$B_{k,p}\big(\bullet(\alpha)_{\bullet-1}\big) = \binom{k}{p}(\alpha p)_{k-p}. \tag{54}$$

**Proposition 14** *The new model* $\phi(x) = x(1-x)^{-\alpha} \in \mathcal{S}$ *iff* $\alpha \in [0, 1]$.

*Proof* First, the convergence radius of $\phi$ is $x_0 = 1$.

We have $\phi'(x) = (1-x)^{-(\alpha+1)}(1 - x(1-\alpha))$ and $\phi' > 0$ for all $x < x_0$ only if $\alpha \in [0, 1]$. Let then $\alpha \in [0, 1]$. Then $\phi^{(k)}(x) = (1-x)^{-(\alpha+k)}(a_k - xb_k)$ and suppose both $a_k$ and $b_k$ are positive with $a_k/b_k > 1$ in such a way that $\phi^{(k)} > 0$ for all $x < x_0$. Then

$$\phi^{(k+1)}(x) = (1-x)^{-(\alpha+k+1)}\big((\alpha+k)a_k - xb_k(\alpha+k-1)\big)$$

with $a_{k+1} = (\alpha+k)a_k$ and $b_{k+1} = b_k(\alpha+k-1)$. Both $a_{k+1}$ and $b_{k+1}$ are positive with $a_{k+1}/b_{k+1} > a_k/b_k > 1$. So $\phi^{(k+1)} > 0$ for all $x < x_0$. □

**Corollary 15** *When* $\alpha \in (0, 1)$, *in the infinitely many species context*, *sampling from a discrete abundance model* $\xi$ *built on* $\phi(x) = x(1-x)^{-\alpha}$ *interprets as bias sampling from a random partition of unity* $\mathbf{S}_\infty(\gamma)$ *with ordered frequencies decaying algebraically fast with* $k$. *The Laplace exponent associated to* $Y_\gamma$ *is* $\psi(x) = -\phi(-x) = x(1+x)^{-\alpha}$, $x > -1$. *The estimator* $\widetilde{\gamma}$ *of the biodiversity parameter* $\gamma$ *is explicitly given by*

$$\widetilde{\gamma} = \frac{P}{k-P+1}\frac{(\alpha(P-1))_{k-P+1}}{(\alpha P)_{k-P}}. \tag{55}$$

*Proof* Clearly $\psi(x) \sim x^{1-\alpha} \to \infty$ as $x \to \infty$ and the corresponding Lévy measure $\pi$ has infinite mass.

We have $\overline{\pi}(t) \sim t^{-(1-\alpha)} \to \infty$ as $t \to 0$ so that $N_+(t) := \#\{k : \Delta_{(k)}(\gamma) > t\}$ grows like $\gamma t^{-(1-\alpha)}$ as $t \to 0$ and

$$S_{(k),\gamma} \sim Y_\gamma^{-1}(k/\gamma)^{-1/(1-\alpha)} \quad \text{as } k \to \infty.$$

Like in the Engen model, the ordered frequencies decay algebraically fast with $k$.

The expression of $\widetilde{\gamma}$ in (55) follows from (54). □

When both $k$ and $P$ are large, together with $k-(1-\alpha)P$, using a simple asymptotic form for (54)

$$\widetilde{\gamma} \sim \frac{P(k-(1-\alpha)P)}{k-P+1}\left(1+\frac{\alpha+k-P}{\alpha(P-1)}\right)^{-\alpha}.$$

**Acknowledgements** T.H. acknowledges partial support from the ANR Modélisation Aléatoire en Écologie, Génétique et Évolution (ANR-Manège-09-BLAN-0215 project) and from the labex MME-DII (Modèles Mathématiques et Économiques de la Dynamique, de l' Incertitude et des Interactions). Part of this work was done while S.M. was visiting Professor at the University of Cergy-Pontoise. Both authors thank support from Basal CONICYT project PFB-03. The authors are indebted to their Referees and the Editor in Charge for suggesting improvements and correcting some mistakes appearing in a former draft.

## References

1. Abbas, M., Bouroubi, S.: On new identities for Bell's polynomials. Discrete Math. **293**(13), 5–10 (2005)
2. Bahls, P., Devitt-Ryder, R., Nguyen, T.: On the location of roots of logaritmically concave polynomials (2010). Preprint available at http://facstaff.unca.edu/pbahls/papers/BahlsDevittRyderNguyenV2.pdf
3. Berestycki, N., Pitman, J.: Gibbs distributions for random partitions generated by a fragmentation process. J. Stat. Phys. **127**(2), 381–418 (2007)
4. Berg, C., Christensen, J.P.R., Ressel, P.: Harmonic analysis on semigroups. In: Theory of positive definite and related functions. Graduate Texts in Mathematics, vol. 100. Springer, New York (1984)
5. Bernstein, S.: Sur les fonctions absolument monotones. Acta Math. **52**(1), 1–66 (1929)
6. Bertoin, J.: Lévy Processes. Cambridge University Press, Cambridge (1996)
7. Blackwell, D., MacQueen, J.B.: Ferguson distributions via Pólya urn schemes. Ann. Stat. **1**, 353–355 (1973)
8. Bunge, J., Fitzpatrick, M.: Estimating the number of species: A Review. J. Am. Stat. Assoc. **88**, 364–373 (1998)
9. Charalambides, Ch.A., Singh, J.: A review of the Stirling numbers, their generalizations and statistical applications. Commun. Stat. Theory Methods **17**(8) (1988)
10. Comtet, L.: Analyse Combinatoire, vols. 1 and 2. Presses Universitaires de France, Paris (1970)
11. Costin, O., Garoufalidis, S.: Resurgence of the fractional polylogarithms. Math. Res. Lett. **16**(5), 817–826 (2009)
12. Darroch, J.N.: On the distribution of the number of successes in independent trials. Ann. Math. Stat. **35**, 1317–1321 (1964)
13. Davenport, H., Pólya, G.: On the product of two power series. Can. J. Math. **1**, 1–5 (1949)
14. Donnelly, P.: Partition structures, Pòlya urns, the Ewens sampling formula and the age of alleles. Theor. Popul. Biol. **30**, 271–288 (1986)
15. Engen, S.: On species frequency models. Biometrika **61**, 263–270 (1974)
16. Engen, S.: Stochastic Abundance Models. Monographs on Applied Probability and Statistics. Chapman and Hall, London (1978)
17. Ewens, W.J.: Some remarks on the law of succession. In: Athens Conference on Applied Probability and Time Series Analysis (1995), Vol. I. Lecture Notes in Statistics, vol. 114, pp. 229–244. Springer, New York (1996)
18. Ewens, W.J.: The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3**, 87–112 (1972)

19. Ewens, W.J.: Population genetics theory—the past and the future. In: Lessard, S. (ed.) Mathematical and Statistical Developments of Evolutionary Theory. Kluwer, Dordrecht (1990)
20. Feng, S.: The Poisson-Dirichlet distribution and related topics. In: Models and Asymptotic Behaviors. Probability and Its Applications (New York). Springer, Heidelberg (2010)
21. Fisher, R.A., Corbet, A.S., Williams, C.B.: The relation between the number of species and the number of individuals in a random sample of an animal population. J. Anim. Ecol. **12**, 42–58 (1943)
22. Garibaldi, U., Scalas, E.: Finitary Probabilistic Methods in Econophysics. Cambridge University Press, Cambridge (2010)
23. Gnedin, A., Pitman, J.: Exchangeable Gibbs partitions and Stirling triangles. (English, Russian summary) Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) **325** (2005), Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. **12**, 83–102, 244–245; translation in J. Math. Sci. (N. Y.) **138**(3) (2006)
24. Hardy, G.H., Littlewood, J.E., Pólya, G.: Inequalities, 2nd edn. Cambridge University Press, Cambridge (1952)
25. Ho, M.W., James, L., Lau, J.W.: Gibbs Partitions (EPPF's) Derived From a Stable Subordinator are Fox H and Meijer G Transforms (2007). http://arxiv.org/abs/0708.0619
26. Holst, L.: The Poisson-Dirichlet distribution and its relatives revisited (2001). Available at http://www.math.kth.se/matstat/fofu/reports/PoiDir.pdf
27. Hoshino, N.: Engen's extended negative binomial model revisited. Ann. Inst. Stat. Math. **57**(2), 369–387 (2005)
28. Hoshino, N.: Random clustering based on the conditional inverse Gaussian-Poisson distribution. J. Jpn. Stat. Soc. **33**(1), 105–117 (2003)
29. Hubbell, S.P.: The neutral theory of biodiversity and biogeography and Stephen Jay Gould. Paleobiology **31**, 122–123 (2005)
30. Huillet, T.: Unordered and ordered sample from Dirichlet distribution. Ann. Inst. Stat. Math. **57**(3), 597–616 (2005)
31. Huillet, T., Möhle, M.: Asymptotics of symmetric compound Poisson population models. Submitted to Comb. Probab. Comput. Special issue dedicated to the memory of Philippe Flajolet, Preprint available at hal-00730734 (2012)
32. Huillet, T., Möhle, M.: Correction on 'Population genetics models with skewed fertilities: a forward and backward analysis'. Stoch. Models **28**(3), 527–532 (2012)
33. Keener, R., Rothman, E., Starr, N.: Distributions on partitions. Ann. Stat. **15**(4), 1466–1481 (1987)
34. Kingman, J.F.C.: Random discrete distributions. J. R. Stat. Soc. Ser. B **37**, 1–22 (1975)
35. Kingman, J.F.C.: Mathematics of Genetic Diversity. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 34. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1980)
36. Kingman, J.F.C.: Poisson processes. Clarendon Press, Oxford (1993)
37. Kolchin, V.F.: Random Mappings. Translation Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York (1986). Translated from the Russian. With a foreword by S.R.S. Varadhan
38. Kolchin, V.F.: Random Graphs. Encyclopedia of Mathematics and Its Applications, vol. 53. Cambridge University Press, Cambridge (1999)
39. Möhle, M.: The concept of duality and applications to Markov processes arising in neutral population genetics models. Bernoulli **5**(5), 761–777 (1999)
40. Neveu, J.: Processus ponctuels. In: École d' Été de Probabilités de Saint-Flour, VI 1976. Lecture Notes in Math., vol. 598, pp. 249–445. Springer, Berlin (1977)
41. Pitman, J.: Random discrete distributions invariant under size-biased permutation. Adv. Appl. Probab. **28**, 525–539 (1996)
42. Pitman, J.: Exchangeable and partially exchangeable random partitions. Probab. Theory Relat. Fields **102**, 145–158 (1995)
43. Pitman, J., Yor, M.: The two parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann. Probab. **25**, 855–900 (1997)
44. Pitman, J.: Poisson-Kingman partitions. In: Statistics and science: a Festschrift for Terry Speed, 134. IMS Lecture Notes Monogr. Ser., vol. 40. Inst. Math. Statist., Beachwood (2003)
45. Pitman, J.: Combinatorial stochastic processes, July 724, 2002. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour. Springer, Berlin (2002). With a foreword by Jean Picard. Lecture Notes in Mathematics, 1875
46. Pollard, H.: The completely monotonic character of the Mittag-Leffler function $Ea(-x)$. Bull. Am. Math. Soc. **54**, 1115–1116 (1948)
47. Prünster, I.: Bibliography: http://sites.carloalberto.org/pruenster/publications.html
48. Schilling, R.L., Song, R., Vondracek, Z.: Bernstein Functions. Theory and Applications. de Gruyter Studies in Mathematics, vol. 37. Walter de Gruyter & Co., Berlin (2010)

49. Schoenberg, I.J.: On the zeros of the generating functions of multiply positive sequences and functions. Ann. Math. (2) **62**, 447–471 (1955)
50. Steutel, F.W., van Harn, K.: Infinite Divisibility of Probability Distributions on the Real Line. Monographs and Textbooks in Pure and Applied Mathematics, vol. 259. Marcel Dekker, New York (2004)
51. Tavaré, S., Ewens, W.J.: Multivariate Ewens distribution. In: Johnson, N.L., Kotz, S., Balakrishnan, N. (eds.) Discrete Multivariate Distributions, pp. 232–246. Wiley, New York (1997), Chap. 41
52. Watterson, G.A.: The stationary distribution of the infinitely-many neutral alleles diffusion model. J. Appl. Probab. **13**(4), 639–651 (1976)
53. Yang, S.L.: Some identities involving the binomial sequences. Discrete Math. **308**, 51–58 (2008)