



ASYMPTOTICS OF THE ALLELE FREQUENCY SPECTRUM AND THE NUMBER OF ALLELES

ROSS A. MALLER ,* *The Australian National University*

SODABEH SHEMEHSAVAR,** *Murdoch University and University of Tehran*

Abstract

We derive large-sample and other limiting distributions of components of the allele frequency spectrum vector, \mathbf{M}_n , joint with the number of alleles, K_n , from a sample of n genes. Models analysed include those constructed from gamma and α -stable subordinators by Kingman (thus including the Ewens model), the two-parameter extension by Pitman and Yor, and a two-parameter version constructed by omitting large jumps from an α -stable subordinator. In each case the limiting distribution of a finite number of components of \mathbf{M}_n is derived, joint with K_n . New results include that in the Poisson–Dirichlet case, \mathbf{M}_n and K_n are asymptotically independent after centering and norming for K_n , and it is notable, especially for statistical applications, that in other cases the limiting distribution of a finite number of components of \mathbf{M}_n , after centering and an unusual $n^{\alpha/2}$ norming, conditional on that of K_n , is normal.

Keywords: Allele frequency spectrum; generalised Poisson–Dirichlet laws; Ewens sampling formula; Kingman’s Poisson–Dirichlet distributions; Pitman sampling formula; gene and species distributions

2020 Mathematics Subject Classification: Primary 60G51; 60G52; 60G55

Secondary 60G57; 62E20; 62P10

1. Introduction

In population genetics, the allele frequency spectrum (AFS) records the numbers of types of alleles represented a designated number of times in a sample of genes. Its distribution was derived under an infinite alleles model of mutation by Ewens [10]. An intimate connection of this distribution, now known as the Ewens sampling formula (ESF), with Kingman’s Poisson–Dirichlet distribution [31], was subsequently exploited in a large range of important applications. The AFS plays an important role for example in the formulation of Kingman’s coalescent [32]; for background, see Berestycki *et al.* [4] and Basdevant and Goldschmidt [3].

Kingman [31] constructed the Poisson–Dirichlet distribution $\mathbf{PD}(\theta)$, for $\theta > 0$, as a random distribution on the infinite unit simplex $\nabla_\infty := \{x_i \geq 0, i = 1, 2, \dots, \sum_{i \geq 1} x_i = 1\}$, by ranking and renormalising the jumps of a driftless gamma subordinator up to a specified time. Another of Kingman’s distributions arises when a driftless α -stable subordinator with

Received 1 November 2023; accepted 11 August 2024.

* Postal address: Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, ACT, 0200, Australia. Email address: Ross.Maller@anu.edu.au

** Postal address: College of Science, Technology, Engineering and Mathematics, and Centre for Healthy Ageing, Health Future Institute, Murdoch University, and School of Mathematics, Statistics & Computer Sciences, University of Tehran. Email address: Soudabeh.Shemehsavar@murdoch.edu.au

© The Author(s), 2024. Published by Cambridge University Press on behalf of Applied Probability Trust.

parameter $\alpha \in (0, 1)$ is used instead of the gamma subordinator. Later again, an encompassing two-parameter Poisson–Dirichlet distribution $\mathbf{PD}(\alpha, \theta)$ was constructed by Pitman and Yor [43]. This specialises to $\mathbf{PD}(\theta)$ when $\alpha \downarrow 0$ and to the second of Kingman’s examples, denoted herein as $\mathbf{PD}(\alpha, 0)$, when $\theta = 0$.

These distributions and the methodologies associated with them have subsequently had a huge impact in many applications areas, especially in population genetics, but also in the excursion theory of stochastic processes, the theory of random partitions, random graphs and networks, probabilistic number theory, machine learning, Bayesian statistics, and others. They have also given rise to a number of generalisations and a large literature analysing the various versions. Among these, generalising $\mathbf{PD}(\alpha, 0)$, is the $\mathbf{PD}_\alpha^{(r)}$ class of Ipsen and Maller [23], defined for $\alpha \in (0, 1)$ and $r > 0$.

In this paper we derive large-sample ($n \rightarrow \infty$) and other limiting distributions of the AFS, joint with the number of alleles, for each of these classes. The AFS in a sample of size $n \in \mathbb{N} := \{1, 2, \dots\}$ is the vector $\mathbf{M}_n = (M_{1n}, M_{2n}, \dots, M_{nn})$, where M_{jn} is the number of allele types represented j times, and $K_n = \sum_{j=1}^n M_{jn}$ is the total number of alleles in the sample. (Vectors and matrices are denoted in boldface, with a superscript ‘ \top ’ for transpose.) K_n is a deterministic function of \mathbf{M}_n , but analysing \mathbf{M}_n and K_n jointly, as we do, leads to important new insights and useful practical results.

2. Notation, models analysed, and overview

Here we set out some notation to be used throughout. Recall that the sample size is $n \in \mathbb{N}$. The sample AFS is the vector \mathbf{M}_n with its dependence on parameters α, θ , or r specified as required, with components $(M_{1n}, M_{2n}, \dots, M_{nn})$ indicated correspondingly. Each \mathbf{M}_n takes values in the set

$$A_{kn} := \left\{ \mathbf{m} = (m_1, \dots, m_n) : m_j \geq 0, \sum_{j=1}^n jm_j = n, \sum_{j=1}^n m_j = k \right\}, \tag{1}$$

and each K_n takes values $k \in \mathbb{N}_n := \{1, 2, \dots, n\}$, $n \in \mathbb{N}$.

Specialising our general notation, when the model under consideration depends on parameters α, θ , or r , these are distinguished in the particular \mathbf{M}_n and K_n analysed; thus, for $\mathbf{PD}(\alpha, \theta)$ we have $(\mathbf{M}_n(\alpha, \theta), K_n(\alpha, \theta))$ (with $(\mathbf{M}_n(\alpha, 0), K_n(\alpha, 0))$ abbreviated to $(\mathbf{M}_n(\alpha), K_n(\alpha))$), for $\mathbf{PD}(\theta)$ we have $(\mathbf{M}_n(\theta), K_n(\theta))$, and similarly for $\mathbf{PD}_\alpha^{(r)}$. Explicit formulae for the distributions of (\mathbf{M}_n, K_n) are available for each of these models, as we now set out in detail.

Distribution of $(\mathbf{M}_n(\alpha, \theta), K_n(\alpha, \theta))$ for $\mathbf{PD}(\alpha, \theta)$. Pitman’s sampling formula ([40, Prop. 9], [43, Sect. A.2, p. 896]) gives, for $\theta > 0, 0 < \alpha < 1$,

$$\begin{aligned} & \mathbf{P}(\mathbf{M}_n(\alpha, \theta) = \mathbf{m}, K_n(\alpha, \theta) = k) \\ &= \frac{n!}{\alpha} \frac{\Gamma(\theta/\alpha + k) \Gamma(\theta + 1)}{\Gamma(\theta/\alpha + 1) \Gamma(n + \theta)} \left(\frac{\alpha}{\Gamma(1 - \alpha)} \right)^k \times \prod_{j=1}^n \frac{1}{m_j!} \left(\frac{\Gamma(j - \alpha)}{j!} \right)^{m_j}. \end{aligned} \tag{2}$$

The corresponding formula for $\mathbf{PD}(\alpha, 0)$ is given just by setting $\theta = 0$:

$$\mathbf{P}(\mathbf{M}_n(\alpha) = \mathbf{m}, K_n(\alpha) = k) = \frac{n(k - 1)!}{\alpha} \left(\frac{\alpha}{\Gamma(1 - \alpha)} \right)^k \times \prod_{j=1}^n \frac{1}{m_j!} \left(\frac{\Gamma(j - \alpha)}{j!} \right)^{m_j}. \tag{3}$$

Distribution of $(\mathbf{M}_n(\theta), K_n(\theta))$ for $\mathbf{PD}(\theta)$. The Ewens sampling formula ([10], [42, p. 46]), for $\theta > 0$, is

$$\mathbf{P}(\mathbf{M}_n(\theta) = \mathbf{m}, K_n(\theta) = k) = \frac{n! \Gamma(\theta) \theta^k}{\Gamma(n + \theta)} \prod_{j=1}^n \frac{1}{m_j!} \left(\frac{1}{j}\right)^{m_j}. \tag{4}$$

Distribution of $(\mathbf{M}_n(\alpha, r), K_n(\alpha, r))$ for $\mathbf{PD}_\alpha^{(r)}$. Equation (2.1) of [26] gives the following formula: for $r > 0, 0 < \alpha < 1$,

$$\mathbf{P}(\mathbf{M}_n(\alpha, r) = \mathbf{m}, K_n(\alpha, r) = k) = n \int_0^\infty \frac{\Gamma(r+k) \lambda^{\alpha k}}{\Gamma(r) \Psi(\lambda)^{r+k}} \prod_{j=1}^n \frac{1}{m_j!} (F_j(\lambda))^{m_j} \frac{d\lambda}{\lambda}, \tag{5}$$

where

$$\Psi(\lambda) = 1 + \alpha \int_0^1 (1 - e^{-\lambda z}) z^{-\alpha-1} dz \tag{6}$$

and

$$F_j(\lambda) = \frac{\alpha}{j!} \int_0^\lambda e^{-z} z^{j-\alpha-1} dz, \quad j \in \mathbb{N}_n, \lambda > 0. \tag{7}$$

In each of (2)–(5), $k \in \mathbb{N}_n, n \in \mathbb{N}$, and \mathbf{m} takes values in the set A_{kn} in (1).

The limiting distribution of K_n is already known for each case, and for components of \mathbf{M}_n separately for $\mathbf{PD}(\alpha, 0)$ and $\mathbf{PD}(\theta)$, but the joint approach adds new information. For example, the limiting distributions of $\mathbf{M}_n(\theta)$ and $K_n(\theta)$ are known separately for $\mathbf{PD}(\theta)$ (the ESF), but we show that they are asymptotically independent. For $\mathbf{PD}(\alpha, \theta)$ and $\mathbf{PD}_\alpha^{(r)}$ we obtain the limiting covariance matrix of a finite number of components of \mathbf{M}_n after centering and normalising, and show that the conditional limiting distribution of these, given K_n , is normal – a useful fact for statistical applications.

Here we list some additional notation and preliminaries. The multinomial distribution arises in a natural way when considering the distribution of \mathbf{M}_n . We use the following generic notation for a multinomial vector with its length and associated occupancy probabilities specified by context. If $J \geq 0$ and $n > J$ are integers, $\mathbf{p} = (p_{J+1}, \dots, p_n)$ is a vector with positive entries such that $p_{J+1} + \dots + p_n = 1$, and m_{J+1}, \dots, m_n are non-negative integers with $m_{J+1} + \dots + m_n = b \leq n$, we set

$$\mathbf{P}(\mathbf{Mult}(J, b, n, \mathbf{p}) = (m_{J+1}, \dots, m_n)) = b! \prod_{j=J+1}^n \frac{p_j^{m_j}}{m_j!}. \tag{8}$$

We recall a useful representation from [26, p. 374]. Denote the components of the multinomial $\mathbf{Mult}(J, b, n, \mathbf{p})$ by (M_{J+1}, \dots, M_n) . This vector has moment generating function (MGF)

$$E \left(\exp \left(\sum_{j=J+1}^n v_j M_j \right) \right) = \left(\sum_{j=J+1}^n p_j e^{v_j} \right)^b,$$

where $v_j > 0, J + 1 \leq j \leq n$. Choosing $v_j = \nu j, J + 1 \leq j \leq n$, where $\nu > 0$, gives

$$E \left(\exp \left(\nu \sum_{j=J+1}^n j M_j \right) \right) = \left(\sum_{j=J+1}^n p_j e^{\nu j} \right)^b. \tag{9}$$

Introduce independent and identically distributed (i.i.d.) random variables X_i with $P(X_i = j) = p_j, J + 1 \leq j \leq n$. Then the right-hand side of (9) is the MGF of $\sum_{i=1}^b X_i$.

A related covariance matrix that occurs in our analyses is the $J \times J$ matrix \mathbf{Q}_J with diagonal elements $q_i(1 - q_i)$ and off-diagonal elements $-q_i q_j, 1 \leq i \neq j \leq J$, where

$$q_j = \frac{\alpha \Gamma(j - \alpha)}{j! \Gamma(1 - \alpha)}, \quad j \in \mathbb{N}_J, J \in \mathbb{N}, J < n. \tag{10}$$

We develop a systematic method of proof in which the novelty is to combine detailed asymptotics of the combinatorial parts of (2)–(5) together with a local limit theorem arising from the representation (9) applied to the distributions of (\mathbf{M}_n, K_n) in the various cases. In two instances key results concerning subordinators due to Covo [8] and Hensley [22] are also used.

The paper is organised as follows. We start in the next section with $\mathbf{PD}_\alpha^{(r)}$ by analysing the distribution of $(\mathbf{M}_n(\alpha, r), K_n(\alpha, r))$ as $n \rightarrow \infty$. This is the most difficult of the cases we deal with, but having established the method, the analogous results for $\mathbf{PD}(\alpha, \theta)$, $\mathbf{PD}(\alpha, 0)$, and $\mathbf{PD}(\theta)$ follow with only necessary changes in Sections 4 and 5. Section 6 gives other related limiting results; for example, the limit as $r \downarrow 0$ of $\mathbf{PD}_\alpha^{(r)}$ is $\mathbf{PD}(\alpha, 0)$. Further discussion and more references are in the concluding Section 7.

3. Limiting distribution of $(\mathbf{M}_n(\alpha, r), K_n(\alpha, r))$ as $n \rightarrow \infty$

In the $\mathbf{PD}_\alpha^{(r)}$ model, $\mathbf{M}_n(\alpha, r)$ and $K_n(\alpha, r)$ depend on the parameters $\alpha \in (0, 1)$ and $r > 0$, and the sample size $n \in \mathbb{N}$. In this section α and r are kept fixed for the large-sample analysis ($n \rightarrow \infty$). The limiting distribution of $K_n(\alpha, r)$ was derived in [26]. Here we extend that analysis to get the joint limiting distribution of a finite number of components of $\mathbf{M}_n(\alpha, r)$, self-normalised by $K_n(\alpha, r)$, together with $K_n(\alpha, r)$. A surprising and novel aspect is the $n^{\alpha/2}$ norming needed for the self-normalised frequency spectrum (after appropriate centering).

Introduce for each $\lambda > 0$ a subordinator $(Y_t(\lambda))_{t>0}$ having Lévy measure

$$\Pi_\lambda(dy) := \frac{\alpha y^{-\alpha-1} dy}{\Gamma(1 - \alpha)} (\mathbf{1}_{\{0 < y < \lambda \leq 1\}} + \mathbf{1}_{\{0 < y < 1 < \lambda\}}). \tag{11}$$

As shown in [26], each $Y_t(\lambda), t > 0, \lambda > 0$, has a continuous bounded density which we denote by $f_{Y_t(\lambda)}(y), y > 0$. Let $J \geq 1$ be a fixed integer, and define q_j and \mathbf{Q}_J as in (10). Let $\mathbf{a} = (a_1, a_2, \dots, a_J) \in \mathbb{R}^J, c > 0$, and recall that for the components of $\mathbf{M}_n(\alpha, r)$ we write $(M_{jn}(\alpha, r))_{1 \leq j \leq n}$. Throughout, when $\mathbf{m} \in A_{kn}$, let $m_+ = \sum_{j=1}^J m_j$ and $m_{++} = \sum_{j=1}^J j m_j$. Then we have the following.

Theorem 1. *For the $\mathbf{PD}_\alpha^{(r)}$ model, we have*

$$\begin{aligned} & \lim_{n \rightarrow \infty} \mathbf{P} \left(\frac{M_{jn}(\alpha, r)}{K_n(\alpha, r)} \leq q_j + \frac{a_j}{n^{\alpha/2}}, 1 \leq j \leq J, \frac{K_n(\alpha, r)}{n^\alpha} \leq c \right) \\ &= \frac{1}{\Gamma(r) \Gamma^r(1 - \alpha)} \int_{\mathbf{y} \in \mathbb{R}^J, \mathbf{y} \leq \mathbf{a}} \int_{0 < x \leq c} \int_{\lambda > 0} \frac{x^{r+J/2-1} e^{-\frac{x}{2} \mathbf{y}^\top \mathbf{Q}_J^{-1} \mathbf{y}}}{\sqrt{(2\pi)^J \det(\mathbf{Q}_J)}} \\ & \times e^{-x(\lambda^{-\alpha} \vee 1) / \Gamma(1 - \alpha)} f_{Y_x(\lambda)}(1) \frac{d\lambda}{\lambda^{\alpha r + 1}} dx d\mathbf{y}. \end{aligned} \tag{12}$$

Corollary 1. Let $(\tilde{\mathbf{M}}(\alpha, r), K(\alpha, r))$ denote a vector having the distribution on the right-hand side of (12). Then the distribution of $\tilde{\mathbf{M}}(\alpha, r)$, conditional on $K(\alpha, r) = x > 0$, is $N(\mathbf{0}, \mathbf{Q}_J/x)$, that is, with density

$$\frac{x^{J/2} e^{-\frac{x}{2}\mathbf{y}^\top \mathbf{Q}_J^{-1} \mathbf{y}}}{\sqrt{(2\pi)^J \det(\mathbf{Q}_J)}}, \quad \mathbf{y} \in \mathbb{R}^J. \tag{13}$$

Proof of Theorem 1. Take $n, J \in \mathbb{N}, n > J, \mathbf{m} = (m_1, \dots, m_n) \in A_{kn}$ and $k \in \mathbb{N}_n$. From (5) we can write

$$\begin{aligned} & \mathbf{P}(M_{jn}(\alpha, r) = m_j, 1 \leq j \leq J, K_n(\alpha, r) = k) \\ &= n \int_0^\infty \frac{\Gamma(r+k)\lambda^{\alpha k}}{\Gamma(r)\Psi(\lambda)^{r+k}} \prod_{j=1}^J \frac{1}{m_j!} (F_j(\lambda))^{m_j} \times \sum_{\mathbf{m}^{(J)} \in A_{kn}^{(J)}} \prod_{j=J+1}^n \frac{1}{m_j!} (F_j(\lambda))^{m_j} \frac{d\lambda}{\lambda}, \end{aligned} \tag{14}$$

with $\mathbf{m}^{(J)} = (m_{J+1}, \dots, m_n)$ and

$$A_{kn}^{(J)} = \left\{ m_j \geq 0, J < j \leq n: \sum_{j=J+1}^n j m_j = n - m_{++}, \sum_{j=J+1}^n m_j = k - m_+ \right\}. \tag{15}$$

(Recall $m_+ = \sum_{j=1}^J m_j$ and $m_{++} = \sum_{j=1}^J j m_j$.) For each $\lambda > 0$, let

$$\mathbf{p}_n^{(J)}(\lambda) = (p_{jn}^{(J)}(\lambda))_{J+1 \leq j \leq n} = \left(\frac{F_j(\lambda)}{\sum_{\ell=J+1}^n F_\ell(\lambda)} \right)_{J+1 \leq j \leq n}.$$

In the notation of (8), let $\mathbf{Mult}(J, k - m_+, n, \mathbf{p}_n^{(J)}(\lambda))$ be a multinomial vector with

$$\mathbf{P}(\mathbf{Mult}(J, k - m_+, n, \mathbf{p}_n^{(J)}(\lambda)) = (m_{J+1}, \dots, m_n)) = (k - m_+)! \prod_{j=J+1}^n \frac{(p_{jn}^{(J)}(\lambda))^{m_j}}{m_j!}, \tag{16}$$

where $m_j \geq 0, J + 1 \leq j \leq n$, and $\sum_{j=J+1}^n m_j = k - m_+$. We can represent the summation over $\mathbf{m}^{(J)} \in A_{kn}^{(J)}$ of the right-hand side of (16) using (9), the probability

$$\mathbf{P}\left(\sum_{i=1}^{k-m_+} X_{in}^{(J)}(\lambda) = n - m_{++} \right), \tag{17}$$

and a method detailed in [26, p. 375]. In (17), $(X_{in}^{(J)}(\lambda))_{1 \leq i \leq k-m_+}$ are i.i.d. with

$$\mathbf{P}(X_{1n}^{(J)}(\lambda) = j) = p_{jn}^{(J)}(\lambda), \quad J + 1 \leq j \leq n.$$

For brevity let $k' = k - m_+$ and $n' = n - m_{++}$, where convenient in what follows. Then

$$\begin{aligned} & \sum_{\mathbf{m}^{(J)} \in A_{kn}^{(J)}} \prod_{j=J+1}^n \frac{1}{m_j!} (F_j(\lambda))^{m_j} \\ &= \frac{1}{k'} \sum_{\mathbf{m}^{(J)} \in A_{kn}^{(J)}} k'! \prod_{j=J+1}^n \frac{1}{m_j!} (p_{jn}^{(J)}(\lambda))^{m_j} \left(\sum_{\ell=J+1}^n F_\ell(\lambda) \right)^{m_j} \\ &= \frac{1}{k'} \left(\sum_{\ell=J+1}^n F_\ell(\lambda) \right)^{k'} \mathbf{P} \left(\sum_{i=1}^{k'} X_{in}^{(J)}(\lambda) = n' \right). \end{aligned}$$

So now we can represent (14) as

$$\begin{aligned} & \mathbf{P}(M_{jn}(\alpha, r) = m_j, 1 \leq j \leq J, K_n(\alpha, r) = k) \\ &= n \int_0^\infty \frac{\Gamma(r+k)\lambda^{\alpha k}}{\Gamma(r)\Psi(\lambda)^{r+k}} \prod_{j=1}^J \frac{1}{m_j!} (F_j(\lambda))^{m_j} \\ & \quad \times \frac{1}{k'} \left(\sum_{\ell=J+1}^n F_\ell(\lambda) \right)^{k'} \mathbf{P} \left(\sum_{i=1}^{k'} X_{in}^{(J)}(\lambda) = n' \right) \frac{d\lambda}{\lambda}. \end{aligned} \tag{18}$$

In this we change variable from λ to λn , and let

$$q_{jn} = \frac{F_j(\lambda n)}{\sum_{\ell=1}^n F_\ell(\lambda n)}, \quad 1 \leq j \leq J, \quad \text{and} \quad q_{+n} = \sum_{j=1}^J q_{jn}. \tag{19}$$

The q_{jn} and q_{+n} depend on λ , but this is omitted in the notation. From (19),

$$\frac{\sum_{\ell=J+1}^n F_\ell(\lambda n)}{\sum_{\ell=1}^n F_\ell(\lambda n)} = 1 - \sum_{j=1}^J q_{jn} = 1 - q_{+n},$$

so we can rewrite (18) as

$$\begin{aligned} & \mathbf{P}(M_{jn}(\alpha, r) = m_j, 1 \leq j \leq J, K_n(\alpha, r) = k) \\ &= \int_0^\infty \frac{\Gamma(r+k)(1-q_{+n})^{k'}}{k'! \prod_{j=1}^J m_j!} \prod_{j=1}^J q_{jn}^{m_j} \times \frac{(\lambda n)^{\alpha k}}{\Psi(\lambda n)^k} \left(\sum_{\ell=1}^n F_\ell(\lambda n) \right)^k \\ & \quad \times \frac{1}{\Gamma(r)\Psi(\lambda n)^r} \times n \mathbf{P} \left(\sum_{i=1}^{k'} X_{in}^{(J)}(\lambda n) = n' \right) \frac{d\lambda}{\lambda}. \end{aligned} \tag{20}$$

Introducing continuous variables $z > 0$ and $u_j \geq 0, 1 \leq j \leq J$, we can write

$$\begin{aligned} & \mathbf{P}\left(\frac{M_{jn}(\alpha, r)}{K_n(\alpha, r)} \leq q_j + \frac{a_j}{n^{\alpha/2}}, 1 \leq j \leq J, \frac{K_n(\alpha, r)}{n^\alpha} \leq c\right) \\ &= \sum_{1 \leq j \leq J, 1 \leq k \leq cn^\alpha} \mathbf{1}\{m_j \leq (q_j + a_j/n^{\alpha/2})k\} \\ & \quad \times \mathbf{P}(M_{jn}(\alpha, r) = m_j, K_n(\alpha, r) = k) \\ &= \int \mathbf{P}(M_{jn}(\alpha, r) = \lfloor u_j \rfloor, K_n(\alpha, r) = \lfloor z \rfloor) \prod_{j=1}^J du_j dz, \end{aligned} \tag{21}$$

where the integration is over the range

$$\{u_j \leq (q_j + a_j/n^{\alpha/2})z, 1 \leq j \leq J, 0 < z \leq cn^\alpha\}.$$

In this integral make the change of variables $z = n^\alpha x, u_j = (q_j + y_j/n^{\alpha/2})z$, so $dz = n^\alpha dx, du_j = dy_j x n^{\alpha/2}$, to write (21) as

$$\begin{aligned} & \mathbf{P}\left(\frac{M_{jn}(\alpha, r)}{K_n(\alpha, r)} \leq q_j + \frac{a_j}{n^{\alpha/2}}, 1 \leq j \leq J, \frac{K_n(\alpha, r)}{n^\alpha} \leq c\right) \\ &= n^{\alpha(1+J/2)} \\ & \quad \times \int_{\mathbf{y} \leq \mathbf{a}, 0 < x \leq c} x^J \mathbf{P}(M_{jn}(\alpha, r) = \lfloor (q_j + y_j/n^{\alpha/2})x n^\alpha \rfloor, 1 \leq j \leq J, K_n(\alpha, r) = \lfloor x n^\alpha \rfloor) dy dx. \end{aligned} \tag{22}$$

Substitute from (20) to write the last expression as

$$\int_{\lambda > 0} \int_{\mathbf{y} \leq \mathbf{a}, 0 < x \leq c} f_n(\mathbf{y}, x, \lambda) dy dx d\lambda =: \int_{\lambda > 0} I_n(\mathbf{a}, c, \lambda) d\lambda, \tag{23}$$

where

$$I_n(\mathbf{a}, c, \lambda) = \int_{\mathbf{y} \leq \mathbf{a}, 0 < x \leq c} f_n(\mathbf{y}, x, \lambda) dy dx$$

and $f_n(\mathbf{y}, x, \lambda)$ denotes the probability in (22), together with a factor $n^{\alpha(1+J/2)}x^J$, and with m_j replaced by $(q_j + y_j/n^{\alpha/2})k$ and k replaced by $\lfloor x n^\alpha \rfloor$ in (20). The limiting behaviours of the four factors in (20) which contribute to $f_n(\mathbf{y}, x, \lambda)$ are set out in the next lemma, whose proof is given in the supplementary material to this paper. The matrix \mathbf{Q}_J is defined using (10), its determinant is $\det(\mathbf{Q}_J)$ and a tilde means the ratio of the connected quantities tends to 1.

Lemma 1. *With the substitutions $k = \lfloor x n^\alpha \rfloor, m_j = \lfloor (q_j + y_j/n^{\alpha/2})k \rfloor, k' = k - m_+$, we have the following limiting behaviours as $n \rightarrow \infty$:*

$$\frac{\Gamma(r+k)(1-q_{+n})^{k'}}{k'! \prod_{j=1}^J m_j!} \prod_{j=1}^J q_{jn}^{m_j} \sim \frac{(x n^\alpha)^{r-J/2-1} e^{-\frac{x}{2} \mathbf{y}^\top \mathbf{Q}_J^{-1} \mathbf{y}}}{\sqrt{(2\pi)^J \det(\mathbf{Q}_J)}}, \tag{24}$$

$$\lim_{n \rightarrow \infty} \frac{(\lambda n)^{\alpha k}}{\Psi(\lambda n)^k} \left(\sum_{\ell=1}^n F_\ell(\lambda n) \right)^k = e^{-x(\lambda^{-\alpha} \vee 1)/\Gamma(1-\alpha)}, \tag{25}$$

$$\frac{1}{\Gamma(r)\Psi(\lambda n)^r} \sim \frac{1}{\Gamma(r)(\lambda n)^{r\alpha}\Gamma^r(1-\alpha)}, \tag{26}$$

$$\lim_{n \rightarrow \infty} n\mathbf{P}\left(\sum_{i=1}^{k-m_+} X_{in}^{(J)}(\lambda n) = n - m_{++}\right) = f_{Y_x(\lambda)}(1). \tag{27}$$

Multiplying together the right-hand sides of (24)–(27), and keeping in mind the factor of $n^{\alpha(1+J/2)}$ from (22), which exactly matches the factors of n in the right-hand side of (24) and (26), then substituting in (20), shows the existence of the limit

$$\begin{aligned} \lim_{n \rightarrow \infty} f_n(\mathbf{y}, x, \lambda) &=: f(\mathbf{y}, x, \lambda) \\ &= \frac{1}{\Gamma(r)\Gamma^r(1-\alpha)} \frac{x^{r+J/2-1} e^{-\frac{x}{2}\mathbf{y}^\top \mathbf{Q}_J^{-1}\mathbf{y}}}{\sqrt{(2\pi)^J \det(\mathbf{Q}_J)}} e^{-x(\lambda^{-\alpha}\vee 1)/\Gamma(1-\alpha)} \frac{f_{Y_x(\lambda)}(1)}{\lambda}. \end{aligned} \tag{28}$$

This is the integrand in (12). Let

$$I(\lambda) := \int_{\mathbf{y} \in \mathbb{R}^J, x > 0} f(\mathbf{y}, x, \lambda) \, d\mathbf{y} \, dx.$$

Then $\int_{\lambda > 0} I(\lambda) \, d\lambda$ is the integral in (12) taken over $(\mathbf{y}, x) \in \mathbb{R}^J \times (0, \infty)$. The integral over \mathbf{y} of the right-hand side of (28) equals

$$\frac{x^{r-1}}{\Gamma(r)\Gamma^r(1-\alpha)} e^{-x(\lambda^{-\alpha}\vee 1)/\Gamma(1-\alpha)} \frac{f_{Y_x(\lambda)}(1)}{\lambda^{\alpha r+1}},$$

and when integrated over $\lambda > 0$ this equals the limiting density of $K_n(\alpha, r)$ in equation (2.8) of [26]. Thus $\int_{\lambda > 0} I(\lambda) \, d\lambda = 1$.

Now argue as follows. Let E_n denote the event in the probability on the left-hand side of (22). Then, by (23), $\mathbf{P}(E_n) = \int_{\lambda > 0} I_n(\mathbf{a}, c, \lambda) \, d\lambda$, and Fatou’s lemma gives

$$\liminf_{n \rightarrow \infty} \mathbf{P}(E_n) = \liminf_{n \rightarrow \infty} \int_{\lambda > 0} I_n(\mathbf{a}, c, \lambda) \, d\lambda \geq \int_{\lambda > 0} I(\mathbf{a}, c, \lambda) \, d\lambda, \tag{29}$$

where

$$I(\mathbf{a}, c, \lambda) = \liminf_{n \rightarrow \infty} I_n(\mathbf{a}, c, \lambda) \geq \int_{\mathbf{y} \leq \mathbf{a}, 0 < x \leq c} f(\mathbf{y}, x, \lambda) \, d\mathbf{y} \, dx,$$

again by Fatou’s lemma. Let E_n^c denote the complement of E_n and set $I^c(\mathbf{a}, c, \lambda) = I(\lambda) - I(\mathbf{a}, c, \lambda)$ and $I_n^c(\mathbf{a}, c, \lambda) = I(\lambda) - I_n(\mathbf{a}, c, \lambda)$. Then

$$\begin{aligned} \liminf_{n \rightarrow \infty} \mathbf{P}(E_n^c) &= \liminf_{n \rightarrow \infty} (1 - \mathbf{P}(E_n)) \\ &= \liminf_{n \rightarrow \infty} \int_{\lambda > 0} (I(\lambda) - I_n(\mathbf{a}, c, \lambda)) \, d\lambda \\ &= \liminf_{n \rightarrow \infty} \int_{\lambda > 0} I_n^c(\mathbf{a}, c, \lambda) \, d\lambda \\ &\geq \int_{\lambda > 0} \liminf_{n \rightarrow \infty} I_n^c(\mathbf{a}, c, \lambda) \, d\lambda. \end{aligned}$$

Now

$$\begin{aligned} \liminf_{n \rightarrow \infty} I_n^c(\mathbf{a}, c, \lambda) &= \liminf_{n \rightarrow \infty} \int_{\{\mathbf{y} \leq \mathbf{a}, 0 < x \leq c\}^c} f_n(\mathbf{y}, x, \lambda) \, d\mathbf{y} \, dx \\ &\geq \int_{\{\mathbf{y} \leq \mathbf{a}, 0 < x \leq c\}^c} f(\mathbf{y}, x, \lambda) \, d\mathbf{y} \, dx \\ &= \left(\int_{\{\mathbf{y} \in \mathbb{R}^J, x > 0\}} - \int_{\{\mathbf{y} \leq \mathbf{a}, 0 < x \leq c\}} \right) f(\mathbf{y}, x, \lambda) \, d\mathbf{y} \, dx \\ &\geq I(\lambda) - I(\mathbf{a}, c, \lambda), \end{aligned}$$

and hence

$$\int_{\lambda > 0} \liminf_{n \rightarrow \infty} I_n^c(\mathbf{a}, c, \lambda) \, d\lambda \geq \int_{\lambda > 0} (I(\lambda) - I(\mathbf{a}, c, \lambda)) \, d\lambda = 1 - \int_{\lambda > 0} I(\mathbf{a}, c, \lambda) \, d\lambda.$$

It follows that

$$\limsup_{n \rightarrow \infty} \mathbf{P}(E_n) = 1 - \liminf_{n \rightarrow \infty} \mathbf{P}(E_n^c) \leq \int_{\lambda > 0} I(\mathbf{a}, c, \lambda) \, d\lambda, \tag{30}$$

and together with (29) this proves

$$\lim_{n \rightarrow \infty} \mathbf{P}(E_n) = \int_{\lambda > 0} I(\mathbf{a}, c, \lambda) \, d\lambda,$$

i.e. (12). □

Proof of Corollary 1. $(\tilde{\mathbf{M}}(\alpha, r), K(\alpha, r))$ has density equal to

$$\int_{\lambda > 0} f(\mathbf{y}, x, \lambda) \, d\lambda,$$

where $f(\mathbf{y}, x, \lambda)$ is defined in (28). Integrating out \mathbf{y} from this integral gives

$$f_{K(\alpha, r)}(x) := \frac{x^{r-1}}{\Gamma(r)\Gamma^r(1-\alpha)} \int_{\lambda > 0} e^{-x(\lambda^{-\alpha} \vee 1)/\Gamma(1-\alpha)} f_{Y_x(\lambda)}(1) \frac{d\lambda}{\lambda^{\alpha r+1}}, \quad x > 0,$$

for the density of $K(\alpha, r)$ at $x > 0$, agreeing with equation (2.8) of [26]. Then dividing

$$\int_{\lambda > 0} f(\mathbf{y}, x, \lambda) \, d\lambda$$

by $f_{K(\alpha, r)}(x)$ gives (13). □

4. The Pitman sampling formula

The next theorem deals with Pitman’s formula for the AFS from $\mathbf{PD}(\alpha, \theta)$. Again we see an $n^{\alpha/2}$ norming for the frequency spectrum, as also occurred in Theorem 1. The marginal limiting distribution obtained for $K_n(\alpha, \theta)/n^\alpha$ in Theorem 2 agrees with that in equation (3.27)

of [42, p. 68]. Let $c \geq 0, J \in \mathbb{N}, \mathbf{a} = (a_1, a_2, \dots, a_J) \in \mathbb{R}^J$, define \mathbf{Q}_J as in (10) and write $(M_{jn}(\alpha, \theta))_{1 \leq j \leq n}$ for the components of $\mathbf{M}_n(\alpha, \theta)$.

Theorem 2. For the PD(α, θ) model we have the asymptotic property

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbf{P} \left(\frac{M_{jn}(\alpha, \theta)}{K_n(\alpha, \theta)} \leq q_j + \frac{a_j}{n^{\alpha/2}}, 1 \leq j \leq J, \frac{K_n(\alpha, \theta)}{n^\alpha} \leq c \right) \\ = \frac{\Gamma(\theta + 1)}{\Gamma(\theta/\alpha + 1)} \int_{\mathbf{y} \in \mathbb{R}^J, \mathbf{y} \leq \mathbf{a}} \int_{0 < x \leq c} \frac{x^{J/2 + \theta/\alpha - 1} e^{-\frac{x}{2} \mathbf{y}^\top \mathbf{Q}_J^{-1} \mathbf{y}}}{\alpha \sqrt{(2\pi)^J \det(\mathbf{Q}_J)}} f_{Y_{x(1)}(1)} dx d\mathbf{y}. \end{aligned} \tag{31}$$

Integrating out \mathbf{y} gives, for the limiting density of $K_n(\alpha, \theta)/n^\alpha$,

$$\frac{\Gamma(\theta + 1)}{\alpha \Gamma(\theta/\alpha + 1)} x^{\theta/\alpha - 1} e^{-x/\Gamma(1-\alpha)} f_{Y_x(1)}(1), \quad x > 0.$$

When $\theta = 0$ this can alternatively be written as a Mittag-Leffler density, $f_{L_\alpha}(x), x > 0$.

Proof of Theorem 2. Start from Pitman’s sampling formula (2) and proceed as in (14) to get, with $1 \leq J < n$,

$$\begin{aligned} \mathbf{P}(M_{jn}(\alpha, \theta) = m_j, 1 \leq j \leq J, K_n(\alpha, \theta) = k) \\ = \frac{n! \Gamma(\theta/\alpha + k) \Gamma(\theta + 1)}{\alpha \Gamma(\theta/\alpha + 1) \Gamma(n + \theta)} \left(\frac{\alpha}{\Gamma(1 - \alpha)} \right)^k \prod_{j=1}^J \frac{1}{m_j!} \left(\frac{\Gamma(j - \alpha)}{j!} \right)^{m_j} \\ \times \sum_{\mathbf{m}^{(J)} \in A_{kn}^{(J)}} \prod_{j=J+1}^n \frac{1}{m_j!} \left(\frac{\Gamma(j - \alpha)}{j!} \right)^{m_j}, \end{aligned} \tag{32}$$

where $\mathbf{m}^{(J)}$ and $A_{kn}^{(J)}$ are as in (15). Let

$$\mathbf{p}_n^{(J)} = (p_{jn}^{(J)})_{J+1 \leq j \leq n} = \left(\frac{\Gamma(j - \alpha)/j!}{\sum_{\ell=J+1}^n \Gamma(\ell - \alpha)/\ell!} \right)_{J+1 \leq j \leq n},$$

and in the notation of (8) let $\mathbf{Mult}(J, k - m_+, n, \mathbf{p}_n^{(J)}(\lambda))$ be a multinomial with

$$\mathbf{P}(\mathbf{Mult}(J, k - m_+, n, \mathbf{p}_n^{(J)}) = (m_{J+1}, \dots, m_n)) = (k - m_+)! \prod_{j=J+1}^n \frac{(p_{jn}^{(J)})^{m_j}}{m_j!}, \tag{33}$$

where $m_j \geq 0, J + 1 \leq j \leq n, \sum_{j=J+1}^n m_j = k - m_+$, and recall that

$$m_+ = \sum_{j=1}^J m_j \quad \text{and} \quad m_{++} = \sum_{j=1}^J j m_j.$$

As in the previous proof we can represent the summation over $\mathbf{m}^{(J)}$ of the left-hand side of (33) using (9) and the probability

$$\mathbf{P} \left(\sum_{i=1}^{k-m_+} X_{in}^{(J)} = n - m_{++} \right), \tag{34}$$

where $(X_{in}^{(J)})_{1 \leq i \leq k-m_+}$ are i.i.d. with

$$\mathbf{P}(X_{1n}^{(J)} = j) = p_{jn}^{(J)}, \quad J + 1 \leq j \leq n.$$

Again let $k' = k - m_+$ and $n' = n - m_{++}$ where convenient. Then we obtain

$$\begin{aligned} & \sum_{\mathbf{m}^{(J)} \in A_{kn}^{(J)}} \prod_{j=J+1}^n \frac{1}{m_j!} \left(\frac{\Gamma(j-\alpha)}{j!} \right)^{m_j} \\ &= \frac{1}{k'!} \sum_{\mathbf{m}^{(J)} \in A_{kn}^{(J)}} k'! \prod_{j=J+1}^n \frac{(p_{jn}^{(J)})^{m_j}}{m_j!} \left(\sum_{\ell=J+1}^n \frac{\Gamma(\ell-\alpha)}{\ell!} \right)^{m_j} \\ &= \frac{1}{k'!} \left(\sum_{\ell=J+1}^n \frac{\Gamma(\ell-\alpha)}{\ell!} \right)^{k'} \mathbf{P} \left(\sum_{i=1}^{k'} X_{in}^{(J)} = n' \right). \end{aligned} \tag{35}$$

To economise on notation we let

$$q_{jn} = \frac{\Gamma(j-\alpha)/j!}{\sum_{\ell=1}^n \Gamma(\ell-\alpha)/\ell!}, \quad 1 \leq j \leq J, \quad \text{so that} \quad \frac{\sum_{\ell=J+1}^n \Gamma(\ell-\alpha)/\ell!}{\sum_{\ell=1}^n \Gamma(\ell-\alpha)/\ell!} = 1 - q_{+n}, \tag{36}$$

still with $q_{+n} = \sum_{j=1}^J q_{jn}$. The q_{jn} in (36) play an exactly analogous role to those in (19). We can write, with the same change of variables as in (22),

$$\begin{aligned} & \mathbf{P} \left(\frac{M_{jn}(\alpha, \theta)}{K_n(\alpha, \theta)} \leq q_j + \frac{a_j}{n^{\alpha/2}}, \quad 1 \leq j \leq J, \quad \frac{K_n(\alpha, \theta)}{n^\alpha} \leq c \right) \\ &= n^{\alpha(1+J/2)} \\ & \quad \times \int_{\mathbf{y} \in \mathbb{R}^J, \mathbf{y} \leq \mathbf{a}} \int_{0 < x \leq c} x^J \mathbf{P}(\mathbf{M}_{jn}(\alpha, \theta) = \lfloor (q_j + y_j/n^{\alpha/2})xn^\alpha \rfloor, \quad 1 \leq j \leq J, \quad K_n(\alpha, \theta) \\ &= \lfloor xn^\alpha \rfloor) \, dy \, dx, \end{aligned}$$

where $1 \leq J < n$, and then, using (32), (34), and (35), we get

$$\begin{aligned} & \mathbf{P}(M_{jn}(\alpha, \theta) = m_j, \quad 1 \leq j \leq J, \quad K_n(\alpha, \theta) = k) \\ &= \left(\frac{\alpha}{\Gamma(1-\alpha)} \sum_{\ell=1}^n \frac{\Gamma(\ell-\alpha)}{\ell!} \right)^k \times \frac{\Gamma(n)\Gamma(\theta/\alpha+k)\Gamma(\theta+1)}{\alpha k! \Gamma(\theta/\alpha+1)\Gamma(n+\theta)} \\ & \quad \times (1 - q_{+n})^{k'} \prod_{j=1}^J \frac{q_{jn}^{m_j}}{m_j!} \times n \mathbf{P} \left(\sum_{i=1}^{k'} X_{in}^{(J)} = n' \right). \end{aligned} \tag{37}$$

The following counterpart of Lemma 1 is proved in the supplementary material.

Lemma 2. *With the substitutions $k = \lfloor xn^\alpha \rfloor$, $m_j = \lfloor q_j + y_j/n^{\alpha/2} \rfloor k$, $k' = k - m_+$, we have the following limiting behaviours as $n \rightarrow \infty$:*

$$\begin{aligned} \frac{(k-1)!(1-q_{+n})^{k'}}{k'! \prod_{j=1}^J m_j!} \prod_{j=1}^J q_{jn}^{m_j} &\sim \frac{(xn^\alpha)^{-J/2-1} e^{-\frac{x}{2} \mathbf{y}^\top \mathbf{Q}_J^{-1} \mathbf{y}}}{\sqrt{(2\pi)^J \det(\mathbf{Q}_J)}}, \\ \lim_{n \rightarrow \infty} \left(\frac{\alpha}{\Gamma(1-\alpha)} \sum_{\ell=1}^n \frac{\Gamma(\ell-\alpha)}{\ell!} \right)^{\lfloor xn^\alpha \rfloor} &= e^{-x/\Gamma(1-\alpha)}, \\ \frac{\Gamma(n)\Gamma(\theta/\alpha+k)\Gamma(\theta+1)}{\alpha k'! \Gamma(\theta/\alpha+1)\Gamma(n+\theta)} &\sim \frac{x^{\theta/\alpha}(k-1)!\Gamma(\theta+1)}{\alpha k'! \Gamma(\theta/\alpha+1)}, \\ \lim_{n \rightarrow \infty} n \mathbf{P} \left(\sum_{i=1}^{k'} X_{in}^{(J)} = n' \right) &= f_{Y_x(1)}(1). \end{aligned}$$

Substituting these estimates in (37), we arrive at (31). □

The $\mathbf{PD}(\alpha, \theta)$ model reduces exactly to the $\mathbf{PD}(\alpha, 0)$ model when θ is set equal to 0, and the same is true of the asymptotic result for it. There is an interesting connection between the probability density of a Stable(α) subordinator and that of a Mittag–Leffler random variable, which we exploit. Write the probability density function (PDF) of a Stable(α) subordinator $(S_x(\alpha))_{x \geq 0}$ (using variable $x \geq 0$ for the time parameter) having Laplace transform $e^{-x\tau^\alpha}$ and Lévy density $\alpha z^{-\alpha-1} \mathbf{1}_{\{z>0\}} / \Gamma(1-\alpha)$, as

$$f_{S_x(\alpha)}(s) = \frac{1}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^{k+1}}{k!} \frac{\Gamma(\alpha k + 1)}{s^{\alpha k + 1}} x^k \sin(\pi \alpha k), \tag{38}$$

and the PDF of a Mittag–Leffler random variable L_α with parameter α as

$$f_{L_\alpha}(s) = \frac{1}{\pi \alpha} \sum_{k=0}^{\infty} \frac{(-1)^{k+1}}{k!} \Gamma(\alpha k + 1) s^{k-1} \sin(\pi \alpha k) \tag{39}$$

(Pitman [42, pp. 10, 11]). Then observe the useful relation

$$\frac{1}{\alpha x} f_{S_x(\alpha)}(1) = f_{L_\alpha}(x), \quad x > 0. \tag{40}$$

Let $\mathbf{a} = (a_1, a_2, \dots, a_J) \in \mathbb{R}^J$, $c > 0$, and write $(M_{jn}(\alpha))_{1 \leq j \leq n}$ for the components of $\mathbf{M}_n(\alpha)$. We prove the following theorem as a corollary to Theorem 2.

Theorem 3. *For the $\mathbf{PD}(\alpha, 0)$ model, we have*

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbf{P} \left(\frac{M_{jn}(\alpha)}{K_n(\alpha)} \leq q_j + \frac{a_j}{n^{\alpha/2}}, \quad 1 \leq j \leq J, \quad \frac{K_n(\alpha)}{n^\alpha} \leq c \right) \\ = \int_{\mathbf{y} \in \mathbb{R}^J, \mathbf{y} \leq \mathbf{a}} \int_{0 < x \leq c} \frac{x^{J/2-1} e^{-\frac{x}{2} \mathbf{y}^\top \mathbf{Q}_J^{-1} \mathbf{y}}}{\alpha \sqrt{(2\pi)^J \det(\mathbf{Q}_J)}} e^{-x/\Gamma(1-\alpha)} f_{Y_x(1)}(1) \, dx \, d\mathbf{y}. \end{aligned} \tag{41}$$

Integrating out \mathbf{y} gives, for the limiting density of $K_n(\alpha)/n^\alpha$,

$$\frac{1}{\alpha x} e^{-x/\Gamma(1-\alpha)} f_{Y_x(1)}(1), \quad x > 0.$$

This can alternatively be written as a Mittag–Leffler density, $f_{L_\alpha}(x)$, $x > 0$.

Let $(\tilde{\mathbf{M}}(\alpha), K(\alpha))$ have the distribution on the right-hand side of (41). The distribution of $\tilde{\mathbf{M}}(\alpha)$, conditional on $K(\alpha) = x > 0$, is $N(\mathbf{0}, \mathbf{Q}_\mathbf{I}/x)$, as in (13).

Remarks. (i) Pitman [42, Thm 3.8, p. 68] gives the almost sure convergence of $K_n(\alpha)/n^\alpha$ to a Mittag–Leffler random variable and the corresponding almost sure convergence of each M_{jn}/n^α . Equation (41) also shows that the M_{jn} are of order n^α , but we express (41) as we do because $K_n(\alpha)$ is a natural normaliser of $\mathbf{M}_n(\alpha)$, producing a vector whose components add to 1. Similarly in Theorems 1 and 2.

(ii) Following the fidi convergence proved in Theorems 1, 2, and 3, it is natural to ask for functional versions. We leave consideration of this to another time.

Proof of Theorem 3. Equation (31) reduces to (41) when $\theta = 0$, so we need only show how the density $f_{Y_x(1)}(1)$ in (41) is related to the density of the Mittag–Leffler distribution. Here a result of Covo [8], which relates the Lévy density of a subordinator to that of a subordinator with the same but truncated Lévy measure, plays a key role.

The Lévy density of $Y_x(1)$ is obtained from (11) with $\lambda = 1$, and is the same as that of $S_x(\alpha)$ truncated at 1. Using Corollary 2.1 of Covo [8] (in his formula set $x = 1$, $s = \lambda$, $t = x$, and take $\Lambda(\lambda) = \lambda^{-\alpha}/\Gamma(1 - \alpha)$), we have

$$f_{Y_x(\lambda)}(1) = e^{x\lambda^{-\alpha}/\Gamma(1-\alpha)} \left(f_{S_x(\alpha)}(1) + \sum_{\kappa=1}^{\lceil 1/\lambda \rceil - 1} (-x)^\kappa A_{\lambda;\kappa}^{(1)}(1, x) \right) \tag{42}$$

(with $\sum_1^0 \equiv 0$), where the $A_{\lambda;\kappa}^{(1)}$ are certain functions defined by Covo. When $\lambda = 1$ these functions disappear from the formula, and we simply have

$$f_{Y_x(1)}(1) = e^{x/\Gamma(1-\alpha)} f_{S_x(\alpha)}(1).$$

Using this together with (40) to replace $f_{Y_x(1)}(1)$ in (41), we obtain a representation of the limiting density of $K_n(\alpha)/n^\alpha$ in terms of a Mittag–Leffler density, in agreement with Theorem 3.8 of [42].

The conditional distribution of $\tilde{\mathbf{M}}(\alpha)$ given $K(\alpha)$ follows easily. □

5. The Ewens sampling formula

In the next theorem, dealing with $\mathbf{PD}(\theta)$, the limiting behaviours of $\mathbf{M}_n(\theta)$ and $K_n(\theta)$ as (independent) Poissons and normal are well known separately ([2, p. 96], [42, pp. 68, 69]), but the asymptotic independence of $\mathbf{M}_n(\theta)$ and $K_n(\theta)$ seems not to have been previously noted, and the way the joint limit arises from the methodology of the previous sections is also interesting. A result of Hensley [22] plays a key role. Let $\mathbf{m} = (m_1, \dots, m_n) \in A_{kn}$, $c \in \mathbb{R}$, and write $(M_{jn}(\theta))_{1 \leq j \leq n}$ for the components of $\mathbf{M}_n(\theta)$.

Theorem 4. For the $\mathbf{PD}(\theta)$ model we have

$$\begin{aligned} & \lim_{n \rightarrow \infty} \mathbf{P}\left(M_{jn}(\theta) = m_j, 1 \leq j \leq J, \frac{K_n(\theta) - \theta \log n}{\sqrt{\theta \log n}} \leq c\right) \\ &= \prod_{j=1}^J \frac{1}{m_j!} \left(\frac{\theta}{j}\right)^{m_j} e^{-\theta/j} \times \frac{1}{\sqrt{2\pi}} \int_{x \leq c} e^{-\frac{1}{2}x^2} dx. \end{aligned} \tag{43}$$

Proof of Theorem 4. Starting from Ewens' sampling formula (4), with $n, J \in \mathbb{N}, n > J, k \in \mathbb{N}_n$, and $m_+ = \sum_{j=1}^J m_j$, we get

$$\begin{aligned} & \mathbf{P}(M_{jn}(\theta) = m_j, 1 \leq j \leq J, K_n(\theta) = k) \\ &= \frac{n! \Gamma(\theta) \theta^{k-m_+}}{\Gamma(n+\theta)} \prod_{j=1}^J \frac{1}{m_j!} \left(\frac{\theta}{j}\right)^{m_j} \sum_{\mathbf{m}^{(J)} \in A_{kn}^{(J)}} \prod_{j=J+1}^n \frac{1}{m_j!} \left(\frac{1}{j}\right)^{m_j}, \end{aligned} \tag{44}$$

where $\mathbf{m}^{(J)}$ and $A_{kn}^{(J)}$ are as in (15). Let vector $\mathbf{p}_n^{(J)}$ have components

$$p_{jn} = \frac{1}{j} / \sum_{\ell=J+1}^n \frac{1}{\ell}, \quad J+1 \leq j \leq n, \tag{45}$$

and let $\mathbf{Mult}(J, k - m_+, n, \mathbf{p}_n^{(J)})$ be multinomial with distribution

$$\mathbf{P}(\mathbf{Mult}(J, k - m_+, n, \mathbf{p}_n^{(J)}) = (m_{J+1}, \dots, m_n)) = (k - m_+)! \prod_{j=J+1}^n \frac{(p_{jn})^{m_j}}{m_j!},$$

for $m_j \geq 0$, with $\sum_{j=J+1}^n m_j = k - m_+$. Thus, arguing as in (35), we find

$$\sum_{\mathbf{m}^{(J)} \in A_{kn}^{(J)}} \prod_{j=J+1}^n \frac{1}{m_j!} \left(\frac{1}{j}\right)^{m_j} = \frac{1}{k'!} \left(\sum_{\ell=J+1}^n \frac{1}{\ell}\right)^{k'} \mathbf{P}\left(\sum_{i=1}^{k'} X_{in}^{(J)} = n'\right),$$

where $k' = k - m_+, n' = n - m_{++}$ and $(X_{in}^{(J)})_{1 \leq i \leq k'}$ are i.i.d. with

$$\mathbf{P}(X_{in}^{(J)} = j) = p_{jn}, \quad J+1 \leq j \leq n. \tag{46}$$

So we can write, from (44),

$$\begin{aligned} & \mathbf{P}(M_{jn}(\theta) = m_j, 1 \leq j \leq J, K_n(\theta) = k) \\ &= \frac{\Gamma(n) \Gamma(\theta) \theta^{k'}}{\Gamma(n+\theta) k'!} \left(\sum_{\ell=J+1}^n \frac{1}{\ell}\right)^{k'} \prod_{j=1}^J \frac{1}{m_j!} \left(\frac{\theta}{j}\right)^{m_j} n \mathbf{P}\left(\sum_{i=1}^{k'} X_{in}^{(J)} = n'\right). \end{aligned} \tag{47}$$

Then, for $c \in \mathbb{R}$,

$$\begin{aligned} & \mathbf{P}\left(M_{jn}(\theta) = m_j, 1 \leq j \leq J, \frac{K_n(\theta) - \theta \log n}{\sqrt{\theta \log n}} \leq c\right) \\ &= \int_{z \leq \theta \log n + c \sqrt{\theta \log n}} \mathbf{P}(M_{jn}(\theta) = m_j, 1 \leq j \leq J, K_n(\theta) = \lfloor z \rfloor) dz \\ &= \sqrt{\theta \log n} \int_{x \leq c} \mathbf{P}(M_{jn}(\theta) = m_j, 1 \leq j \leq J, K_n(\theta) = \lfloor \theta \log n + x \sqrt{\theta \log n} \rfloor) dx. \end{aligned} \tag{48}$$

Let $k_n(x) = \lfloor \theta \log n + x\sqrt{\theta \log n} \rfloor$ and calculate

$$\begin{aligned} \sqrt{\theta \log n} \mathbf{P}(M_{jn}(\theta) = m_j, 1 \leq j \leq J, K_n(\theta) = k_n(x)) \\ = \sqrt{\theta \log n} \times \text{right-hand side of (47) with } k = k_n(x). \end{aligned}$$

We let $n \rightarrow \infty$ in this. Consider the various factors. First,

$$\frac{\Gamma(n)\Gamma(\theta)}{\Gamma(n + \theta)} \sim \frac{\Gamma(\theta)}{n^\theta}. \tag{49}$$

Next, using a standard asymptotic for the harmonic series,

$$\begin{aligned} \left(\sum_{\ell=J+1}^n \frac{1}{\ell} \right)^{k-m_+} &= \left(\sum_{\ell=1}^n \frac{1}{\ell} - \sum_{j=1}^J \frac{1}{j} \right)^{k-m_+} \\ &= \left(\log n + \gamma - \sum_{j=1}^J \frac{1}{j} + O(1/n) \right)^{k-m_+} \\ &= (\log n)^{k-m_+} \left(1 + \frac{\gamma - \sum_{j=1}^J 1/j + O(1/n)}{\log n} \right)^{k-m_+} \\ &\sim (\log n)^{k-m_+} e^{\theta\gamma - \theta \sum_{j=1}^J 1/j}, \quad \text{as } k = k_n(x) \rightarrow \infty. \end{aligned} \tag{50}$$

Here $\gamma = 0.577 \dots$ is Euler’s constant and we recall $k_n(x) = \lfloor \theta \log n + x\sqrt{\theta \log n} \rfloor$.

Substitute (49) and (50) in (47), remembering the factor of $\sqrt{\theta \log n}$ from (48), to get (47) asymptotic to

$$\frac{\sqrt{\theta \log n}}{n^\theta} \frac{(\theta \log n)^{k'}}{k'!} \times e^{\theta\gamma} \Gamma(\theta) \times \prod_{j=1}^J \frac{1}{m_j!} \left(\frac{\theta}{j} \right)^{m_j} e^{-\theta/j} \times n \mathbf{P} \left(\sum_{i=1}^{k'} X_{in}^{(j)} = n' \right). \tag{51}$$

Using Stirling’s formula and the relations $k' = k - m_+ = k - \sum_{j=1}^J m_j$ and $k = k_n(x) = \lfloor \theta \log n + x\sqrt{\theta \log n} \rfloor$, we find that

$$\begin{aligned} \frac{\sqrt{\theta \log n}}{n^\theta} \frac{(\theta \log n)^{k'}}{k'!} &\sim \frac{\sqrt{\theta \log n}}{n^\theta} \frac{(\theta \log n)^{k'}}{\sqrt{2\pi k'} (k')^{k'} e^{-k'}} \\ &\sim \frac{e^{k'}}{\sqrt{2\pi n^\theta}} \left(\frac{\theta \log n}{k'} \right)^{k'} \\ &= \frac{e^{k'}}{\sqrt{2\pi n^\theta}} \left(1 - \frac{k' - \theta \log n}{k'} \right)^{k'} \\ &= \frac{e^{k'}}{\sqrt{2\pi n^\theta}} \exp \left(k' \log \left(1 - \frac{k' - \theta \log n}{k'} \right) \right). \end{aligned} \tag{52}$$

Note that

$$\frac{k' - \theta \log n}{k'} = \frac{[\theta \log n + x\sqrt{\theta \log n}] - m_+ - \theta \log n}{k'} = \frac{x\sqrt{\theta \log n} + O(1)}{k'}$$

is $O(1/\sqrt{\log n}) \rightarrow 0$ as $n \rightarrow \infty$, so we can use the expansion $\log(1 - z) = -z - z^2/2 - \dots$ for small z to get, for the right-hand side of (52),

$$\begin{aligned} & \frac{e^{k'}}{\sqrt{2\pi n^\theta}} \exp\left(k' \left(-\frac{k' - \theta \log n}{k'} - \frac{1}{2} \left(\frac{k' - \theta \log n}{k'}\right)^2\right)\right) + O\left(\frac{1}{\log n}\right)^{3/2} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x\sqrt{\theta \log n} + O(1))^2}{\theta \log n + x\sqrt{\theta \log n}}\right) + O\left(\frac{1}{\log n}\right)^{1/2} \\ &\rightarrow \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \end{aligned}$$

Thus (51) is asymptotic to

$$\prod_{j=1}^J \frac{1}{m_j!} \left(\frac{\theta}{j}\right)^{m_j} e^{-\theta/j} \times \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \times e^{\theta\gamma} \Gamma(\theta) \times n\mathbf{P}\left(\sum_{i=1}^{k'} X_{in}^{(j)} = n'\right). \tag{53}$$

It remains to deal with the $X_{in}^{(j)}$ term in (53). We need a version of a local limit theorem for a sum of i.i.d. random variables; this kind of usage has arisen before in various analyses of Poisson–Dirichlet processes [2, 42]. Use the Fourier inversion formula for the mass function of a discrete random variable (Gnedenko and Kolmogorov [15, p. 233]) to write

$$\begin{aligned} n\mathbf{P}\left(\sum_{i=1}^{k'} X_{in}^{(j)} = n'\right) &= \frac{n}{2\pi} \int_{-\pi}^{\pi} e^{-n'iv} (\phi_n(v))^{k'} dv \\ &= \frac{n}{2\pi n'} \int_{-n'\pi}^{n'\pi} e^{-iv} (\mathbf{E}(e^{ivX_{1n}/n'}))^{k'} dv, \end{aligned} \tag{54}$$

where $\phi_n(v) = \mathbf{E}(e^{ivX_{1n}^{(j)}})$, $v \in \mathbb{R}$. By (45) and (46),

$$\mathbf{E}(e^{ivX_{1n}^{(j)}/n'}) = \sum_{j=J+1}^n \frac{e^{ivj/n'}}{j} / \sum_{\ell=J+1}^n \frac{1}{\ell} = 1 + \sum_{j=J+1}^n \frac{1}{j} (e^{ivj/n'} - 1) / \sum_{\ell=J+1}^n \frac{1}{\ell},$$

in which

$$\begin{aligned} \frac{\sum_{j=J+1}^n (e^{ivj/n'} - 1)/j}{\sum_{\ell=J+1}^n 1/\ell} &= \frac{(\sum_{j=1}^n - \sum_{j=1}^J)(e^{ivj/n'} - 1)/j}{(\sum_{\ell=1}^n - \sum_{\ell=1}^J)1/\ell} \\ &= \frac{\sum_{j=1}^n (e^{ivj/n'} - 1)/j}{\log n + O(1)} + O\left(\frac{1}{n \log n}\right). \end{aligned}$$

The numerator here is

$$\begin{aligned} \sum_{j=1}^n \frac{1}{j} (e^{ivj/n'} - 1) &= \sum_{j=1}^n \frac{iv}{n'} \int_0^1 e^{ivzj/n'} dz \\ &= \frac{iv}{n'} \int_0^1 \sum_{j=0}^n e^{ivzj/n'} dz - \frac{iv}{n'} \\ &= \frac{iv}{n'} \int_0^1 \frac{e^{ivz(n+1)/n'} - 1}{e^{ivz/n'} - 1} dz + O\left(\frac{1}{n}\right) \\ &= iv \int_0^1 \frac{e^{ivz(n+1)/n'} - 1}{ivz + O(1/n)} dz + O\left(\frac{1}{n}\right) \\ &= \int_0^1 (e^{ivz} - 1) \frac{dz}{z} + O\left(\frac{1}{n}\right), \end{aligned}$$

and consequently

$$\begin{aligned} &\left(1 + \sum_{j=J+1}^n \frac{1}{j} (e^{ivj/n'} - 1) / \sum_{\ell=J+1}^n \frac{1}{\ell}\right)^{k_n(x) - m_+} \\ &= \left(1 + \frac{\int_0^1 (e^{ivz} - 1) dz/z}{\log n + O(1)}\right)^{\lfloor \theta \log n + x\sqrt{\theta \log n - m_+} \rfloor}. \end{aligned}$$

The last expression has limit $\exp(\theta \int_0^1 (e^{ivz} - 1) dz/z)$ so it follows that

$$\lim_{n \rightarrow \infty} (\mathbf{E} e^{ivX_{in}^{(J)}/n'})^{k'} = e^{\theta \int_0^1 (e^{ivz} - 1) dz/z} = \widehat{w}_\theta(v),$$

in the notation of Hensley [22] (his equations (2.6) and (2.12) with $\alpha = \theta$ and $K(\alpha) = 1$). Hensley’s function $\widehat{w}_\theta(v)$ is the characteristic function of a density $w_\theta(u)$, so taking the limit in (54) gives

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iv\widehat{w}_\theta(v)} dv = w_\theta(1).$$

(Justification for taking the limit through the integral in (54) can be given by using the kind of arguments in (29)–(30).) By [22, eq. (2.1)], $w_\theta(1) = e^{-\theta\gamma} / \Gamma(\theta)$. We conclude that

$$\lim_{n \rightarrow \infty} n\mathbf{P}\left(\sum_{i=1}^{k'} X_{in}^{(J)} = n'\right) = e^{-\theta\gamma} / \Gamma(\theta).$$

Substituting this in (53) gives (43); in fact we get density convergence in (43), which is stronger. □

6. Other limits

In this section we derive other limits of the processes studied above. Section 6.1 analyses $\mathbf{PD}_\alpha^{(r)}$ as $r \downarrow 0$ or $r \rightarrow \infty$. Section 6.2 summarises the results in a convergence diagram for $\mathbf{M}_n(\alpha, r)$, showing that the convergences $n \rightarrow \infty$ and $r \downarrow 0$ commute.

6.1. Limiting behaviour of $\mathbf{PD}_\alpha^{(r)}$ as $r \downarrow 0$ or $r \rightarrow \infty$

Kingman’s $\mathbf{PD}(\alpha, 0)$ distribution arises by taking the ordered jumps, $(\Delta S_1^{(i)})_{i \geq 1}$, up to time 1, of a driftless α -stable subordinator $S = (S_t)_{t > 0}$, normalised by their sum S_1 , as a random distribution on the infinite simplex ∇_∞ . A natural generalisation is to delete the r largest jumps ($r \geq 1$ an integer) up to time 1 of S , and consider the vector whose components are the remaining jumps $(\Delta S_1^{(i)})_{i \geq r+1}$ normalised by their sum. Again we obtain a random distribution on ∇_∞ , now with an extra parameter, r . This is the $\mathbf{PD}_\alpha^{(r)}$ distribution of Ipsen and Maller [23]. Some limiting and other properties of it were studied in [24], where it was extended to all $r > 0$, not just integer r , and in Chegini and Zarepour [6]. Zhang and Dassios [46] also consider an α -stable subordinator with largest jumps removed for simulating functionals of a Pitman–Yor process.

By its construction, $\mathbf{PD}_\alpha^{(r)}$ reduces to $\mathbf{PD}(\alpha, 0)$ when r is set equal to 0, but Theorem 5 shows a kind of continuity property: as $r \downarrow 0$ the distribution of $(\mathbf{M}_n(\alpha, r), K_n(\alpha, r))$ from $\mathbf{PD}_\alpha^{(r)}$ converges to the corresponding quantity from $\mathbf{PD}(\alpha, 0)$; a useful result for statistical analysis, where we might fit the $\mathbf{PD}_\alpha^{(r)}$ model with general α and r to data, and test $H_0: r = 0$, i.e. whether r is needed in the model to describe the data. The parameter r models ‘over-dispersion’ in the data [6].

In this subsection we keep n and α fixed and let $r \downarrow 0$ or $r \rightarrow \infty$. Part (i) of the next theorem is an analogue of the Pitman and Yor result [43, p. 880] that, for each $\theta > 0$, the limit of $\mathbf{PD}(\alpha, \theta)$ as $\alpha \downarrow 0$ is $\mathbf{PD}(\theta)$.

Theorem 5. *We have the following limiting behaviour for $\mathbf{PD}_\alpha^{(r)}$.*

- (i) *As $r \downarrow 0$, the limit of $\mathbf{P}(\mathbf{M}_n(\alpha, r) = \mathbf{m}, K_n(\alpha, r) = k)$ is the probability in (3), i.e. the distribution of $(\mathbf{M}_n(\alpha), K_n(\alpha))$.*
- (ii) *In the limit as $r \rightarrow \infty$, a sample from $\mathbf{PD}_\alpha^{(r)}$ is of n alleles each with 1 representative, i.e. the singleton distribution with mass $\mathbf{1}_{(\mathbf{m}, k) = ((n, 0, \dots, 0_n), n)}$.*

Proof of Theorem 5. For this proof it is convenient to work from a formula for $\mathbf{PD}_\alpha^{(r)}$ derived from equations (2.1) and (3.7) of [26]. For $\mathbf{m} \in A_{kn}$ (see (1)) and $k \in \mathbb{N}_n$,

$$\mathbf{P}(\mathbf{M}_n(\alpha, r) = \mathbf{m}, K_n(\alpha, r) = k) = n \int_0^\infty \ell_n(\lambda)^k \mathbf{P}(\mathbf{Mult}(0, k, n, \mathbf{p}_n(\lambda)) = \mathbf{m}) \mathbf{P}\left(\text{Negbin}\left(r, \frac{1}{\Psi(\lambda)}\right) = k\right) \frac{d\lambda}{\lambda}, \tag{55}$$

where the multinomial notation is as in (8) with

$$\mathbf{p}_n(\lambda) = (p_{jn}(\lambda))_{j \in \mathbb{N}_n} = \left(\frac{F_j(\lambda)}{\sum_{\ell=1}^n F_\ell(\lambda)} \right)_{j \in \mathbb{N}_n}, \tag{56}$$

and $F_j(\lambda)$ is as in (7). The function $\ell_n(\lambda)$ is defined in [26, eq. (3.6)] as

$$\ell_n(\lambda) := \frac{\Psi_n(\lambda) - 1}{\Psi(\lambda) - 1} = \frac{\int_0^\lambda \sum_{j=1}^n (z^j/j!) z^{-\alpha-1} e^{-z} dz}{\int_0^\lambda z^{-\alpha-1} (1 - e^{-z}) dz} \leq 1, \tag{57}$$

where $\Psi(\lambda)$ is as in (6). Finally, in (55), $\text{Negbin}(r, 1/\Psi(\lambda))$ is a negative binomial random variable with parameter $r > 0$ and success probability $1/\Psi(\lambda)$, thus

$$\mathbf{P}\left(\text{Negbin}\left(r, \frac{1}{\Psi(\lambda)}\right) = k\right) = \frac{\Gamma(r+k)}{\Gamma(r)k!} \left(1 - \frac{1}{\Psi(\lambda)}\right)^k \left(\frac{1}{\Psi(\lambda)}\right)^r, \quad k \in \mathbb{N}_n. \tag{58}$$

(i) Observe, by (6), for all $\lambda > 0$,

$$1 \vee \lambda^\alpha \int_0^\lambda e^{-z} z^{-\alpha} dz \leq \Psi(\lambda) = 1 + \alpha \lambda^\alpha \int_0^\lambda (1 - e^{-z}) z^{-\alpha-1} dz \leq 1 + \lambda^\alpha \Gamma(1 - \alpha), \tag{59}$$

so (recalling that $\Psi(\lambda) > 1$)

$$\mathbf{P}\left(\text{Negbin}\left(r, \frac{1}{\Psi(\lambda)}\right) = k\right) = \frac{\Gamma(r+k) (\Psi(\lambda) - 1)^k}{\Gamma(r)k! \Psi(\lambda)^{k+r}} \leq \frac{\Gamma(r+k)\Gamma^k(1 - \alpha)}{\Gamma(r)k!} \lambda^{k\alpha}. \tag{60}$$

Fix $A > 0$. Using (60) and $\ell_n(\lambda) \leq 1$, the component of the integral over $(0, A)$ in (55) does not exceed

$$n \int_0^A \frac{\Gamma(r+k)\Gamma^k(1 - \alpha)}{\Gamma(r)k!} \lambda^{k\alpha-1} d\lambda = \frac{n\Gamma(r+k)\Gamma^k(1 - \alpha)}{\Gamma(r)k!} \times \frac{A^{k\alpha}}{k\alpha},$$

and this tends to 0 as $r \downarrow 0$ since then $\Gamma(r) \rightarrow \infty$.

For the component of the integral over (A, ∞) in (55), change variable to $\lambda = \lambda/r$ and write it as

$$n \int_{Ar}^\infty \ell_n(\lambda/r)^k \mathbf{P}(\text{Mult}(0, k, n, \mathbf{p}_n(\lambda/r)) = \mathbf{m}) \mathbf{P}\left(\text{Negbin}\left(r, \frac{1}{\Psi(\lambda/r)}\right) = k\right) \frac{d\lambda}{\lambda}. \tag{61}$$

From (7), (56), and (57),

$$\lim_{\lambda \rightarrow \infty} \ell_n(\lambda) =: \ell_n^\infty = \frac{\alpha}{\Gamma(1 - \alpha)} \sum_{j=1}^n \frac{\Gamma(j - \alpha)}{j!}$$

and

$$\lim_{\lambda \rightarrow \infty} p_j(\lambda) =: p_j(\infty) = \left(\frac{F_j(\infty)}{\sum_{\ell=1}^n F_\ell(\infty)} \right) = \frac{\Gamma(j - \alpha)}{j!} / \sum_{\ell=1}^n \frac{\Gamma(\ell - \alpha)}{\ell!},$$

so that, as $\lambda \rightarrow \infty$,

$$(\ell_n(\lambda))^k \mathbf{P}(\text{Mult}(0, k, n, \mathbf{p}_n(\lambda)) = \mathbf{m}) \rightarrow k! \left(\frac{\alpha}{\Gamma(1 - \alpha)} \right)^k \prod_{j=1}^n \frac{1}{m_j!} \left(\frac{\Gamma(j - \alpha)}{j!} \right)^{m_j}. \tag{62}$$

Given $\varepsilon > 0$, choose $A = A(\varepsilon)$ large enough for the left-hand side of (62) to be within ε of the right-hand side when $\lambda > A$. Note that, by (59), when $\lambda/r > A$,

$$\Psi(\lambda/r) \geq (\lambda/r)^\alpha \int_0^{\lambda/r} e^{-z} z^{-\alpha} dz \geq (\lambda/r)^\alpha \int_0^A e^{-z} z^{-\alpha} dz =: c_{\alpha,A} (\lambda/r)^\alpha.$$

So, using (58), an upper bound for the integral in (61) is, for $r < 1$,

$$\begin{aligned} & \frac{n}{c_{\alpha,A}^r} \int_{Ar}^\infty \left(k! \left(\frac{\alpha}{\Gamma(1 - \alpha)} \right)^k \prod_{j=1}^n \frac{1}{m_j!} \left(\frac{\Gamma(j - \alpha)}{j!} \right)^{m_j} + \varepsilon \right) \frac{\Gamma(r+k) r^{\alpha r} d\lambda}{\Gamma(r)k! \lambda^{\alpha r+1}} \\ &= \frac{n}{c_{\alpha,A}^r} \left(k! \left(\frac{\alpha}{\Gamma(1 - \alpha)} \right)^k \prod_{j=1}^n \frac{1}{m_j!} \left(\frac{\Gamma(j - \alpha)}{j!} \right)^{m_j} + \varepsilon \right) \frac{\Gamma(r+k)}{\alpha r \Gamma(r)k!} A^{-\alpha r}. \end{aligned} \tag{63}$$

Let $r \downarrow 0$ and note that $c_{\alpha,A}^r \rightarrow 1$, $A^{-\alpha r} \rightarrow 1$, $r\Gamma(r) = \Gamma(r+1) \rightarrow 1$ and $\Gamma(r+k) \rightarrow \Gamma(k) = (k-1)!$. Then, letting $\varepsilon \downarrow 0$, we obtain an upper bound for the limit of the integral in (55) equal to the distribution in (3).

For a lower bound, we can discard the component of the integral over $(0, A)$ in (55), then change variables as in (61). For $\lambda/r \geq A$, $\Psi(\lambda/r) \leq 1 + (\lambda/r)^\alpha \Gamma(1-\alpha) = (1 + o(1))(\lambda/r)^\alpha \Gamma(1-\alpha)$ as $r \downarrow 0$. So, from (58) and (59),

$$\mathbf{P}\left(\text{Negbin}\left(r, \frac{1}{\Psi(\lambda/r)}\right) = k\right) \geq \frac{\Gamma(r+k)}{\Gamma(r)k!} \left(1 - \frac{1}{\Psi(\lambda/r)}\right)^k \left(\frac{1}{(1 + o(1))(\lambda/r)^\alpha \Gamma(1-\alpha)}\right)^r.$$

We also have the limit in (62). Substituting these in (61), we get a lower bound of the same form as in the left-hand side of (63). Fatou’s lemma can then be used on the integral to get a lower bound equal to the upper bound derived earlier. Thus we prove part (i).

(ii) In the integral in (55), change variable to

$$\rho = \left(1 - \frac{1}{\Psi(\lambda)}\right) \frac{r}{1/\Psi(\lambda)} = r(\Psi(\lambda) - 1),$$

so that

$$\frac{d\lambda}{\lambda} = \frac{d\rho}{r\lambda\Psi'(\lambda)}, \quad \Psi(\lambda) = \frac{\rho}{r} + 1 \quad \text{and} \quad \lambda = \Psi^{\leftarrow}\left(\frac{\rho}{r} + 1\right),$$

where Ψ^{\leftarrow} is the inverse function to Ψ . Then $r \rightarrow \infty$ implies $\Psi(\lambda) \rightarrow 1$ thus $\lambda \rightarrow 0$. From (7), (56), (57), and L’Hôpital’s rule we see that

$$\lim_{\lambda \rightarrow 0} \ell_n(\lambda) = 1 \quad \text{and} \quad \lim_{\lambda \rightarrow 0} p_{jn}(\lambda) = p_{jn}(0) = \lim_{\lambda \rightarrow 0} \frac{1}{j! \sum_{\ell=1}^n \lambda^{\ell-j}/\ell!} = \mathbf{1}_{j=1}.$$

After changing variable in (55), we get

$$\begin{aligned} &\mathbf{P}(\mathbf{M}_n(\alpha, r) = \mathbf{m}, K_n(\alpha, r) = k) \\ &= n \int_0^\infty \ell_n\left(\Psi^{\leftarrow}\left(\frac{\rho}{r} + 1\right)\right)^k \mathbf{P}\left(\mathbf{Mult}(0, k, n, \mathbf{p}_n\left(\Psi^{\leftarrow}\left(\frac{\rho}{r} + 1\right)\right)) = \mathbf{m}\right) \\ &\quad \times \mathbf{P}\left(\text{Negbin}\left(r, \frac{r}{r+\rho}\right) = k\right) \frac{d\rho}{r\lambda\Psi'(\lambda)}. \end{aligned} \tag{64}$$

In this, by (6),

$$\Psi(\lambda) - 1 \sim \alpha\lambda/(1-\alpha) \quad \text{as } \lambda \rightarrow 0,$$

so $r\lambda \sim (1-\alpha)\rho/\alpha$, and

$$\Psi'(\lambda) = \alpha \int_0^1 e^{-\lambda z} z^{-\alpha} dz \rightarrow \alpha/(1-\alpha) \quad \text{as } \lambda \rightarrow 0.$$

We have $\ell_n(\cdot) \leq 1$, and by (60) the negative binomial probability is bounded above by a constant multiple of $r^k \lambda^k \asymp \rho^k$ for large r , and as $r \rightarrow \infty$ we have convergence of the negative binomial to the Poisson. So by dominated convergence the right-hand side of (64) converges to

$$\begin{aligned} &n \int_0^\infty \mathbf{P}(\mathbf{Mult}(0, k, n, \mathbf{p}_n(0)) = (1, 0, \dots, 0)) \frac{e^{-\rho} \rho^k}{k!} \frac{d\rho}{\rho} \\ &= \frac{n}{k} \mathbf{P}(\mathbf{Mult}(0, k, n, \mathbf{p}_n(0)) = (1, 0, \dots, 0)). \end{aligned}$$

Here $\text{Mult}(0, k, n, \mathbf{p}_n(0))$ is multinomial with $\mathbf{p}_n(0) = (p_{jn}(0))_{1 \leq j \leq n}$, and so the probability equals $\mathbf{1}_{(\mathbf{m}, k) = (n, 0, \dots, 0_n, n)}$ as required. \square

Theorem 6. For the $\text{PD}_\alpha^{(r)}$ model we have the property

$$\lim_{r \downarrow 0} (\text{right-hand side of (12)}) = \text{right-hand side of (41)}.$$

Proof of Theorem 6. From (12) it suffices to consider the integrand

$$\frac{x^{r+J/2-1}}{\Gamma(r)\Gamma^r(1-\alpha)} \frac{e^{-\frac{x}{2}\mathbf{y}^\top \mathbf{Q}_J^{-1}\mathbf{y}}}{\sqrt{(2\pi)^J \det(\mathbf{Q}_J)}} \int_{\lambda=0}^\infty e^{-x(\lambda^{-\alpha}\vee 1)/\Gamma(1-\alpha)} f_{Y_x(\lambda)}(1) \frac{d\lambda}{\lambda^{\alpha r+1}}, \tag{65}$$

and for this we look at \int_0^1 and \int_1^∞ separately. Recall that $f_{Y_x(\lambda)}(y)$ is uniformly bounded in λ by C , say. For the first part, make the transformation $y = \lambda^{-\alpha}$, so $\lambda = y^{-1/\alpha}$ and $d\lambda = -y^{-1/\alpha-1} dy/\alpha$. Then \int_0^1 is bounded by

$$\begin{aligned} & \frac{Cx^{r+J/2-1}}{\Gamma(r)\Gamma^r(1-\alpha)} \frac{e^{-\frac{x}{2}\mathbf{y}^\top \mathbf{Q}_J^{-1}\mathbf{y}}}{\sqrt{(2\pi)^J \det(\mathbf{Q}_J)}} \int_{\lambda=0}^1 e^{-x(\lambda^{-\alpha}\vee 1)/\Gamma(1-\alpha)} \frac{d\lambda}{\lambda^{\alpha r+1}} \\ &= \frac{Cx^{r+J/2-1}}{\alpha\Gamma(r)\Gamma^r(1-\alpha)} \frac{e^{-\frac{x}{2}\mathbf{y}^\top \mathbf{Q}_J^{-1}\mathbf{y}}}{\sqrt{(2\pi)^J \det(\mathbf{Q}_J)}} \int_{y=1}^\infty e^{-xy/\Gamma(1-\alpha)} y^{r-1} dy, \end{aligned}$$

which tends to 0 as $r \downarrow 0$. So we can neglect \int_0^1 . For \int_1^∞ the contribution is

$$\begin{aligned} & \frac{x^{r+J/2-1}}{\Gamma(r)\Gamma^r(1-\alpha)} \frac{e^{-\frac{x}{2}\mathbf{y}^\top \mathbf{Q}_J^{-1}\mathbf{y}}}{\sqrt{(2\pi)^J \det(\mathbf{Q}_J)}} \int_{\lambda=1}^\infty e^{-x/\Gamma(1-\alpha)} f_{Y_x(1)}(1) \frac{d\lambda}{\lambda^{\alpha r+1}} \\ &= \frac{x^{r+J/2-1}}{\alpha r\Gamma(r)\Gamma^r(1-\alpha)} \frac{e^{-\frac{x}{2}\mathbf{y}^\top \mathbf{Q}_J^{-1}\mathbf{y}}}{\sqrt{(2\pi)^J \det(\mathbf{Q}_J)}} e^{-x/\Gamma(1-\alpha)} f_{Y_x(1)}(1) \\ &\rightarrow \frac{x^{J/2-1}}{\alpha\sqrt{(2\pi)^J \det(\mathbf{Q}_J)}} e^{-x/\Gamma(1-\alpha)} f_{Y_x(1)}(1), \end{aligned}$$

using $\lim_{r \downarrow 0} r\Gamma(r) = \lim_{r \downarrow 0} \Gamma(r+1) = 1$. This equals the integrand on the right-hand side of (41). \square

Remarks. (i) We can simplify the limit distribution in (12) somewhat by using Corollary 2.1 of [8] as we did in the proof of Theorem 3. Once again in the integral in (65) look at \int_0^1 and \int_1^∞ separately. The component of the integral over $(1, \infty)$ can be explicitly evaluated as

$$\frac{1}{\Gamma(r)\Gamma^r(1-\alpha)} x^r f_{L_\alpha}^{(1)}(x).$$

The component over $(0,1)$ involves the functions $A_{\lambda;\kappa}^{(1)}$ defined by Covo (see (42)), which can be calculated using a recursion formula he gives; for an alternative approach, see [38].

(ii) Theorem 3.1 of Maller and Shemehsavar [35] shows that a sampling model based on the Dickman subordinator has the same limiting behaviour as the Ewens sampling formula. The Dickman function is well known to be closely associated with $\text{PD}(\theta)$ and some of its generalisations; see e.g. Arratia and Baxendale [1], Arratia *et al.* [2, pp. 14, 76], Handa [20], and [25], [26], and [35].

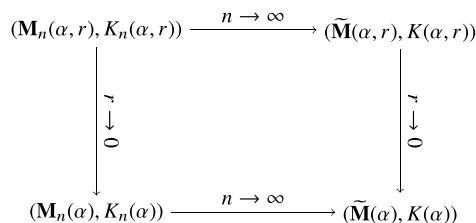


FIGURE 1. Convergence diagram for $\mathbf{PD}_\alpha^{(r)}$. Arrows denote convergence in distribution. Upper left to upper right is proved in Theorem 1; $(\tilde{\mathbf{M}}(\alpha, r), K(\alpha, r))$ has the distribution on the right-hand side of (12). Upper left to lower left is proved in Theorem 5(i). For upper right to lower right, the convergence is proved in Theorem 6. Lower left to lower right is proved in Theorem 3, with $(\tilde{\mathbf{M}}(\alpha), K(\alpha))$ denoting a random variable with the Normal-Mittag-Leffler mixture distribution on the right-hand side of (41). The diagram is schematic only: the random variables have to be normed and centred in certain ways before convergence takes place.

6.2. Convergence diagram for $\mathbf{PD}_\alpha^{(r)}$

Using Theorems 1 and 3 we can give the convergence diagram shown in Figure 1. We can also use Theorem 1 to complete the convergence diagram for $\mathbf{M}_n(\alpha, r)$ in [35]; the missing upper right entry in their Figure 2 is $(\tilde{\mathbf{M}}(\alpha, r), K(\alpha, r))$. Note that $r \rightarrow \infty$, $\alpha \downarrow 0$, such that $r\alpha \rightarrow a > 0$, in [35].

We also have the following useful results, which follow immediately from Theorems 1 and 3.

Corollary 2. We have $\sqrt{K(\alpha, r)}\tilde{\mathbf{M}}(\alpha, r) \stackrel{D}{=} N(\mathbf{0}, \mathbf{Q}_J)$, independent of $K(\alpha, r)$, and similarly $\sqrt{K(\alpha)}\tilde{\mathbf{M}}(\alpha) \stackrel{D}{=} N(\mathbf{0}, \mathbf{Q}_J)$, independent of $K(\alpha)$.

7. Discussion

The methods developed in Sections 3–6 offer a unified approach with nice interpretations to limit theorems of various sorts for Poisson–Dirichlet models, and we expect they can be applied in other situations too. For example, Grote and Speed [19] analyse a ‘general selection model’ based on the infinite alleles mutation scheme. Their formula (1.13) gives a version of the distribution of (\mathbf{M}_n, K_n) amenable to analysis by our methods. Ruggiero *et al.* [44, eqs (3.1), (3.2)] analyse a species sampling model based on normalised inverse-Gaussian diffusions as a generalisation of $\mathbf{PD}(\alpha, 0)$; an extra parameter β is introduced to the $\mathbf{PD}(\alpha, 0)$ model, somewhat analogous to our $\mathbf{PD}_\alpha^{(r)}$. See also Lijoi *et al.* [34] and Favaro and Feng [11].

Formula (5) can be compared with an analogous formula, equation (4.14) of Pitman [42, p. 81], in which the component $\prod_{i=1}^k \Psi^{(n_i)}(\lambda)$ can be replaced by $\prod_{j=1}^n (\Psi^{(j)}(\lambda))^{m_j}$ using the identity $m_j = \sum_{i=1}^k 1_{\{j=n_i\}}$. See also Pitman [41] and the Gibbs partitions in [42, pp. 25–26, eq. (1.52)]. Hansen [21] gave a general treatment of decomposable combinatorial structures having a Poisson–Dirichlet limit.

There are many other limiting results in the literature. Joyce *et al.* [29] prove a variety of Gaussian limit theorems as the mutation rate $\theta \rightarrow \infty$ in the $\mathbf{PD}(\theta)$ model; see also Griffiths [17] and Handa [20]. Feng [12, 13] gives large deviation results as $\theta \rightarrow \infty$. James [28] gives results like this related to $\mathbf{PD}(\alpha, \theta)$ in a Bayesian set-up. Dolera and Favaro [9] investigate α -diversity in $\mathbf{PD}(\alpha, \theta)$. Berestycki *et al.* [3] prove laws of large numbers for the AFS when the coalescent process is taken to be the Bolthausen–Sznitman coalescent. Möhle [37] introduces a

Mittag–Leffler process and shows that the block counting process of the Bolthausen–Sznitman n -coalescent, properly scaled, converges to it in the Skorokhod topology. See also Freund and Möhle [14]. Relevant to our Theorem 2, Koriyama, Matsuda, and Komaki [33] derive the asymptotic distribution of the maximum likelihood estimator of (α, θ) for the $\mathbf{PD}(\alpha, \theta)$ model and show that α_n is $n^{\alpha/2}$ consistent for α .

When $r \rightarrow \infty$, size-biased samples from $\mathbf{PD}_\alpha^{(r)}$ tend to those of $\mathbf{PD}(1 - \alpha)$, i.e. to $\mathbf{PD}(\theta)$ with $\theta = 1 - \alpha$; see [24]. Together with part (i) of Theorem 5, this suggests that $\mathbf{PD}_\alpha^{(r)}$ is intermediate in a sense between $\mathbf{PD}(\alpha, 0)$ and $\mathbf{PD}(\theta)$. Chegini and Zarepour [6] show that $\mathbf{PD}_\alpha^{(r)}$ can alternatively be obtained as the ordered normalised jumps of a negative binomial process, as defined in Gregoire [16]. When $r \rightarrow \infty$, $\alpha \rightarrow 0$ such that $r\alpha \rightarrow a > 0$, a limit involving the Dickman subordinator is obtained; see [35].

The AFS is alternatively known as the *site frequency spectrum* (SFS). It summarises the distribution of allele frequencies throughout the genome. According to Mas-Sandoval *et al.* [36, p. 2]:

The SFS is arguably one of the most important summary statistics of population genetic data . . . contain[ing] invaluable information on the demographic and adaptive processes that shaped the evolution of the population under investigation . . . For instance, an SFS showing an overrepresentation of rare alleles is an indication of an expanding population, while bottleneck events tend to deplete low-frequency variants.

The appearance of the normal as the limiting distribution of \mathbf{M}_n , after centering and normalising, conditional on K_n , for $\mathbf{PD}_\alpha^{(r)}$, $\mathbf{PD}(\alpha, 0)$ and $\mathbf{PD}(\alpha, \theta)$, is useful for statistical applications. In Ipsen *et al.* [27] we used a least-squares method to fit $\mathbf{PD}_\alpha^{(r)}$ to a variety of data sets, some quite small. Even so, we observed by simulations that the finite sample distributions of the estimates of α and r were quite closely approximated by the normal. The asymptotic normality derived in the present paper provides a rationale for using least-squares and helps explain the approximate normality of the parameter estimates. Similar ideas may be useful for $\mathbf{PD}(\alpha, \theta)$. Cereda and Corradi [5] give a methodology for fitting that model to forensic data. Keith *et al.* [30] give an example of testing differences between similar allelic distributions using the AFS. Functionals of \mathbf{M}_n such as $\sum_{j=1}^n (M_{jn} - EM_{jn})^2 / EM_{jn}$ are important in many aspects of population genetics; see e.g. the measures of sample diversity given by Watterson [45, Section 4].

Acknowledgements

We are grateful to the Editors and a referee for suggestions which led to simplifications and a significant improvement in the exposition.

Funding information

There are no funding bodies to thank relating to the creation of this article.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

Supplementary material

The supplementary material for this article can be found at <https://doi.org/10.1017/jpr.2024.84>.

References

- [1] ARRATIA, R. AND BAXENDALE, P. (2015). Bounded size bias coupling: a Gamma function bound, and universal Dickman-function behavior. *Prob. Theory Related Fields* **162**, 411–429.
- [2] ARRATIA, R., BARBOUR, A. AND TAVARÉ, S. (2003). *Logarithmic Combinatorial Structures: A Probabilistic Approach* (EMS Monographs in Mathematics). European Mathematical Society, Zurich.
- [3] BASDEVANT, A. AND GOLDSCHMIDT, C. (2008). Asymptotics of the allele frequency spectrum associated with the Bolthausen–Sznitman coalescent. *Electron. J. Prob.* **13**, 486–512.
- [4] BERESTYCKI, J., BERESTYCKI, N. AND SCHWEINSBERG, J. (2007). Beta-coalescents and continuous stable random trees. *Ann. Prob.* **35**, 1835–1887.
- [5] CEREDA, G. AND CORRADI, F. (2023). Learning the two parameters of the Poisson–Dirichlet distribution with a forensic application. *Scand. J. Statist.* **50**, 120–141.
- [6] CHEGINI, S. AND ZAREPOUR, M. (2023). Random discrete probability measures based on negative binomial process. Available at [arXiv:2307.00176](https://arxiv.org/abs/2307.00176).
- [7] COVO, S. (2009). On approximations of small jumps of subordinators with particular emphasis on a Dickman-type limit. *J. Appl. Prob.* **46**, 732–755.
- [8] COVO, S. (2009). One-dimensional distributions of subordinators with upper truncated Lévy measure, and applications. *Adv. Appl. Prob.* **41**, 367–392.
- [9] DOLERA, E. AND FAVARO, S. (2020). A Berry–Esseen theorem for Pitman’s α -diversity. *Ann. Appl. Prob.* **30**, 847–869.
- [10] EWENS, W. (1972). The sampling theory of selectively neutral alleles. *Theoret. Pop. Biol.* **3**, 87–112.
- [11] FAVARO, S. AND FENG, S. (2014). Asymptotics for the number of blocks in a conditional Ewens–Pitman sampling model. *Electron. J. Prob.* **19**, 21, 1–15.
- [12] FENG, S. (2007). Large deviations associated with Poisson–Dirichlet distribution and Ewens sampling formula. *Ann. Appl. Prob.* **17**, 1570–1595.
- [13] FENG, S. (2010). *The Poisson–Dirichlet Distribution and Related Topics: Models and Asymptotic Behaviours* (Probability and its Applications). Springer.
- [14] FREUND, F. AND MÖHLE, M. (2009). On the number of allelic types for samples taken from exchangeable coalescents with mutation. *Adv. Appl. Prob.* **41**, 1082–1101.
- [15] GNEDENKO, B. V. AND KOLMOGOROV, A. N. (1968). *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley.
- [16] GREGOIRE, G. (1984). Negative binomial distributions for point processes. *Stoch. Process. Appl.* **16**, 179–188.
- [17] GRIFFITHS, R. C. (1979). On the distribution of allele frequencies in a diffusion model. *Theoret. Pop. Biol.* **15**, 140–158.
- [18] GRIFFITHS, R. C. (2003). The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theoret. Pop. Biol.* **64**, 241–251.
- [19] GROTE, M. N. AND SPEED, T. P. (2002). Approximate Ewens formulae for symmetric overdominance selection. *Ann. Appl. Prob.* **12**, 637–663.
- [20] HANDA, K. (2009). The two-parameter Poisson–Dirichlet point process. *Bernoulli* **15**, 1082–1116.
- [21] HANSEN, J. (1994). Order statistics for decomposable combinatorial structures. *Random Structures Algorithms* **5I**, 517–533.
- [22] HENSLEY, D. (1982). The convolution powers of the Dickman function. *J. London Math. Soc.* **s2-33**, 395–406.
- [23] IPSEN, Y. F. AND MALLER, R. A. (2017). Negative binomial construction of random discrete distributions on the infinite simplex. *Theory Stoch. Process.* **22**, 34–46.
- [24] IPSEN, Y. F., MALLER, R. A. AND SHEMEHSAVAR, S. (2020). Limiting distributions of generalised Poisson–Dirichlet distributions based on negative binomial processes. *J. Theoret. Prob.* **33**, 1974–2000.
- [25] IPSEN, Y. F., MALLER, R. A. AND SHEMEHSAVAR, S. (2020). Size biased sampling from the Dickman subordinator. *Stoch. Process. Appl.* **130**, 6880–6900.
- [26] IPSEN, Y. F., MALLER, R. A. AND SHEMEHSAVAR, S. (2021). A generalised Dickman distribution and the number of species in a negative binomial process model. *Adv. Appl. Prob.* **53**, 370–399.
- [27] IPSEN, Y. F., SHEMEHSAVAR, S. AND MALLER, R. A. (2018). Species sampling models generated by negative binomial processes. Available at [arXiv:1904.13046](https://arxiv.org/abs/1904.13046).
- [28] JAMES, L. F. (2008). Large sample asymptotics for the two-parameter Poisson–Dirichlet process. *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, vol. 3, pp. 187–199. Institute of Mathematical Statistics.

- [29] JOYCE, P., KRONE, S. M. AND KURTZ, T. G. (2002). Gaussian limits associated with the Poisson–Dirichlet distribution and the Ewens sampling formula. *Ann. Appl. Prob.* **12**, 101–124.
- [30] KEITH, T. P., BROOKS, L. D., LEWONTIN, R. C., MARTINEZ-CRUZADO, J. C. AND RIGBY, D. L. (1985). Nearly identical allelic distributions of xanthine dehydrogenase in two populations of *Drosophila pseudoobscura*. *Mol. Biol. Evol.* **2**, 206–216.
- [31] KINGMAN, J. F. C. (1975). Random discrete distributions. *J. R. Statist. Soc. B* **37**, 1–22.
- [32] KINGMAN, J. F. C. (1982). The coalescent. *Stoch. Process. Appl.* **13**, 235–248.
- [33] KORIYAMA, T., MATSUDA, T. AND KOMAKI, F. (2023). Asymptotic analysis of parameter estimation for the Ewens–Pitman partition. Available at [arXiv:2207.01949v3](https://arxiv.org/abs/2207.01949v3).
- [34] LIJOI, A., MENA, R. H. AND PRUNSTER, I. (2005). Mixture modeling with normalized inverse-Gaussian priors. *J. Amer. Statist. Assoc.* **100**, 1278–1291.
- [35] MALLER, R. A. AND SHEMEHSAVAR, S. (2023). Generalized Poisson–Dirichlet distributions based on the Dickman subordinator. *Theory Prob. Appl.* **67**, 593–612.
- [36] MAS-SANDOVAL, A., POPE, N. S., NIELSEN, K. N., ALTINKAYA, I., FUMAGALLI, M. AND KORNELIUSSEN, T. S. (2022). Fast and accurate estimation of multidimensional site frequency spectra from low-coverage high-throughput sequencing data. *Gigascience* **11**, giac032.
- [37] MÖHLE, M. (2015). The Mittag–Leffler process and a scaling limit for the block counting process of the Bolthausen–Sznitman coalescent. *ALEA* **12**, 35–53.
- [38] PERMAN, M. (1993). Order statistics for jumps of normalised subordinators. *Stoch. Process. Appl.* **46**, 267–281.
- [39] PERMAN, M., PITMAN, J. AND YOR, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Prob. Theory Related Fields* **92**, 21–39.
- [40] PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Prob. Theory Related Fields* **102**, 145–158.
- [41] PITMAN, J. (1997). Partition structures derived from Brownian motion and stable subordinators. *Bernoulli* **3**, 79–96.
- [42] PITMAN, J. (2006). *Combinatorial Stochastic Processes*. Springer, Berlin.
- [43] PITMAN, J. AND YOR, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *Ann. Prob.* **25**, 855–900.
- [44] RUGGIERO, M., WALKER, S. G. AND FAVARO, S. (2013). Alpha-diversity processes and normalized inverse-Gaussian diffusions. *Ann. Appl. Prob.* **23**, 386–425.
- [45] WATTERSON, G. A. (1974). The sampling theory of selectively neutral alleles. *Adv. Appl. Prob.* **6**, 463–468.
- [46] ZHANG, J. AND DASSIOS, A. (2024). Truncated two-parameter Poisson–Dirichlet approximation for Pitman–Yor process hierarchical models. *Scand. J. Statist.* **51**, 590–611.
- [47] ZHOU, M., FAVARO, S. AND WALKER, S. G. (2017). Frequency of frequencies distributions and size-dependent exchangeable random partitions. *J. Amer. Statist. Assoc.* **112**, 1623–1635.