THE POISSON–DIRICHLET DISTRIBUTION AND ITS RELATIVES REVISITED

LARS HOLST

Department of Mathematics, Royal Institute of Technology
SE–100 44 Stockholm, Sweden

E-mail: lholst@math.kth.se

December 17, 2001

Abstract

The Poisson-Dirichlet distribution and its marginals are studied, in particular the largest component, that is Dickman's distribution. Size-biased sampling and the GEM distribution are considered. Ewens sampling formula and random permutations, generated by the Chinese restaurant process, are also investigated. The used methods are elementary and based on properties of the finite-dimensional Dirichlet distribution.

Keywords: Chinese restaurant process; Dickman's function; Ewens sampling formula; GEM distribution; Hoppe's urn; random permutations; residual allocation models; size-biased sampling

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 60G57
SECONDARY 60C05, 60K99

Running title: The Poisson–Dirichlet distribution revisited

1 Introduction

Random discrete probability distributions occur in many situations in pure and applied probability. One such is the *Poisson–Dirichlet distribution* introduced by Kingman (1975). This distribution or special cases of it has proved to be useful in a variety of interesting applications in combinatorics, analytic number theory, Bayesian statistics, population genetics and ecology; see Arratia, Barbour and Tavaré (2001),

Donnelly and Grimmett (1993), Kingman (1980), Pitman and Yor (1997), Tenenbaum (1995), and the references therein.

Following Kingman (1993, Chapter 9) we introduce the Poisson–Dirichlet distribution as follows. Let Π be a Poisson point process in the first quadrant of the real (t, x)-plane with intensity measure $e^{-x}/x\, dtdx$. Then, the number of points of Π in the set

$$A_{\theta,z} := \{(t, x) : 0 \le t \le \theta, z < x\}$$

is, for each $z > 0$, Poisson distributed with mean $\theta E_1(z)$, where

$$E_1(z) := \int_z^\infty \frac{e^{-x}}{x}\, dx = \int_1^\infty \frac{e^{-zx}}{x}\, dx$$

denotes the exponential integral function. With probability one

$$\Pi \cap A_{\theta,0} = \{(T_j, X_{[j]}) : j = 1, 2, \dots\},$$

where $X_{[1]} > X_{[2]} > \cdots > 0$ are points of a non-homogenous Poisson process on the positive real half line with intensity measure $\theta e^{-x}/x\, dx$, and independent of these points the T's are independent random variables uniform on $(0, \theta)$. The density of $X_{[j]}$ is

$$f_{X_{[j]}}(x) = \frac{\theta e^{-x}}{x} \frac{(\theta E_1(x))^{j-1}}{(j-1)!}\, e^{-\theta E_1(x)}, \quad x > 0.$$

The random variable $S_\theta = X_{[1]} + X_{[2]} + \cdots$ has the Laplace transform

$$E(e^{-zS_\theta}) = \exp\big(-\int_0^\infty (1 - e^{-zx})\, \frac{\theta e^{-x}}{x}\, dx\big) = (1+z)^{-\theta}, \quad |z| < 1,$$

implying that S_θ is $\Gamma(\theta)$ distributed with density $x^{\theta-1}e^{-x}I(x > 0)/\Gamma(\theta)$. The stochastic process $\{S_\theta : \theta > 0\}$, sometimes called *Moran's subordinator*, has independent stationary gamma distributed increments. Define the *Poisson–Dirichlet distribution,* abbreviated $PD(\theta)$, on the (infinite-dimensional) simplex

$$\Delta := \{(x_1, x_2, \dots) : x_i \ge 0,\ x_1 + x_2 + \cdots = 1\}$$

by the random vector

$$(V_1, V_2, \dots) := (X_{[1]}/S_\theta, X_{[2]}/S_\theta, \dots).$$

Here $V_1 > V_2 > \cdots > 0$, $V_1 + V_2 + \cdots = 1$, are the ordered normalized (almost surely different) jumps of Moran's subordinator up to time θ.

Kingman (1993, Section 9.6) says that "the distribution PD(θ) is rather less than user-friendly"; few explicit results for the marginal distributions are given there.

Some such results are obtained for example in Watterson (1976) and Griffiths (1988). The main purpose of this paper is to derive and present properties of the Poisson–Dirichlet and related distributions in a way, which we think is more transparent than previously. Marginal distributions and moments of the V's are obtained in Section 2. Certain aspects of the distribution of $1/V_1$ are studied in Section 3. In Section 4 we review size-biased sampling and the GEM distribution. Ewens sampling formula is considered in Section 5. Finally, random permutations generated by the Chinese restaurant process, and Hoppe's urn are investigated in Section 6.

2 Marginal distributions of the Poisson–Dirichlet

Let $\{S_\theta : \theta > 0\}$ be Moran's subordinator. For $\theta > 0$ fixed and $\alpha = \theta/n$ consider the independent $\Gamma(\alpha)$ distributed increments

$$Y_j = S_{j\alpha} - S_{(j-1)\alpha}, \quad j = 1, 2, \ldots, n,$$

and their (almost surely different) order statistic $Y_{[1]} > Y_{[2]} > \cdots > Y_{[n]}$. With X's as in the Introduction we have

$$(Y_{[1]}, Y_{[2]}, \ldots, Y_{[n]}, 0, 0, \ldots) \Rightarrow (X_{[1]}, X_{[2]}, \ldots), \quad n \to \infty.$$

An elementary calculation by change of variables shows, that the random variable $S_\theta = Y_1 + \cdots + Y_n$ is independent of the random vector

$$(Z_1, \ldots, Z_n) := (Y_1/S_\theta, \ldots, Y_n/S_\theta),$$

and that the vector has a symmetric Dirichlet distribution, $\mathrm{D}(\alpha, \ldots, \alpha)$, with density

$$\frac{\Gamma(\theta)}{\Gamma(\alpha)^n}\, x_1^{\alpha-1} \cdots x_n^{\alpha-1},$$

relative to the $(n-1)$-dimensional Lebesgue measure on the simplex

$$\Delta_n := \{(x_1, \ldots, x_n) : x_i \geq 0,\ x_1 + \cdots + x_n = 1\}.$$

Hence, for the ordered normalized increments $Z_{[1]} > Z_{[2]} > \cdots > Z_{[n]}$, and the ordered normalized jumps $V_1, V_2, \ldots$, we have

$$(Z_{[1]}, Z_{[2]}, \ldots, Z_{[n]}, 0, 0, \ldots) \Rightarrow (V_1, V_2, \ldots), \quad n \to \infty,$$

and also that the V's are independent of S_θ.

The observations above are the main tools used below. First we prove, cf. Griffiths (1988, Theorem 2):

Proposition 2.1 *For $x > 0$:*

$$P(V_1 \le x) = 1 + \sum_{j=1}^{[1/x]} \frac{(-\theta)^j}{j!} \int_x^1 \cdots \int_x^1 \frac{(1 - y_1 - \cdots - y_j)_+^{\theta-1}}{y_1 \cdots y_j}\, dy_1 \cdots dy_j.$$

Proof. For $0 < x < 1$ inclusion-exclusion, symmetry and properties of the Dirichlet distribution give

$$\begin{aligned} P(Z_{[1]} \le x) &= 1 - P(Z_1 > x \cup \cdots \cup Z_n > x) \\ &= 1 + \sum_{j=1}^{[1/x]} (-1)^j \binom{n}{j} P(Z_1 > x \cap \cdots \cap Z_j > x) \\ &= 1 + \sum_{j=1}^{[1/x]} (-1)^j \binom{n}{j} \int_x^1 \cdots \int_x^1 \frac{\Gamma(\theta)(1 - y_1 - \cdots - y_j)_+^{\theta - j\alpha - 1}}{\Gamma(\alpha)^j \Gamma(\theta - j\alpha) y_1^{1-\alpha} \cdots y_j^{1-\alpha}}\, dy_1 \cdots dy_j. \end{aligned}$$

Thus, for $n \to \infty$ and $\alpha = \theta/n \to 0$ we have

$$\binom{n}{j} \alpha^j \to \theta^j / j!, \quad \alpha\Gamma(\alpha) = \Gamma(\alpha + 1) \to 1, \quad P(Z_{[1]} \le x) \to P(V_1 \le x),$$

and therefore

$$P(Z_{[1]} \le x) \to 1 + \sum_{j=1}^{[1/x]} \frac{(-\theta)^j}{j!} \int_x^1 \cdots \int_x^1 \frac{(1 - y_1 - \cdots - y_j)_+^{\theta-1}}{y_1 \cdots y_j}\, dy_1 \cdots dy_j,$$

proving the assertion. □

Proposition 2.2 *For $k = 1, 2, \ldots$:*

$$E(V_1^k) = \frac{E(X_{[1]}^k)}{E(S_\theta^k)} = \frac{\int_0^\infty y^{k-1} e^{-y} e^{-\theta E_1(y)}\, dy}{(\theta + 1) \cdots (\theta + k - 1)}.$$

Proof. The independence between S_θ and V_1 $(= X_{[1]}/S_\theta)$ implies

$$E(X_{[1]}^k) = E(V_1^k S_\theta^k) = E(V_1^k) E(S_\theta^k).$$

As S_θ is $\Gamma(\theta)$ distributed and $X_{[1]}$ has the density

$$f_{X_{[1]}}(y) = \frac{\theta e^{-y}}{y}\, e^{-\theta E_1(y)}, \quad y > 0,$$

the assertion follows. □

Using similar arguments formulas for higher and mixed moments of the V's can readily be obtained, cf. Griffiths (1979). The following result is essentially given in Watterson (1976, Theorem).

Proposition 2.3 *The joint density of* $(V_1, \ldots, V_r)$ *satisfies*

$$f_{V_1,\ldots,V_r}(z_1, \ldots, z_r) = \frac{\theta^r (1 - z_1 - \cdots - z_r)^{\theta-1}}{z_1 \cdots z_r} P\Big(V_1 \le \frac{z_r}{1 - z_1 - \cdots - z_r}\Big),$$

for $z_1 > z_2 > \cdots > z_r > 0$ *and* $z_1 + z_2 + \cdots + z_r < 1$*; it is 0 elsewhere.*

Proof. Let $Z_{[1]} > Z_{[2]} > \cdots > Z_{[n]}$ be the order statistic of $\mathrm{D}(\alpha, \ldots, \alpha)$. Integrating over the set

$$B = \{(x_{r+1}, \ldots, x_n) : 0 \le x_i < z_r,\ x_{r+1} + \cdots + x_n = 1 - z_1 - \cdots - z_r\},$$

we see that the density of $(Z_{[1]}, \ldots, Z_{[r]})$ can be written

$$\begin{aligned}\int_B \frac{n(n-1)\ldots(n-r+1)\Gamma(\theta)}{\Gamma(\alpha)^n} z_1^{\alpha-1} \cdots z_r^{\alpha-1} x_{r+1}^{\alpha-1} \cdots x_n^{\alpha-1}\, dx_{r+1} \cdots dx_{n-1}&\\ = \frac{n(n-1)\cdots(n-r+1)\alpha^r\Gamma(\theta)}{\Gamma(\alpha+1)^r\Gamma(\theta-r\alpha)} z_1^{\alpha-1} \cdots z_r^{\alpha-1}&\\ \times \int_B \frac{\Gamma(\theta-r\alpha)}{\Gamma(\alpha)^{n-r}} x_{r+1}^{\alpha-1} \cdots x_n^{\alpha-1}\, dx_{r+1} \cdots dx_{n-1}.&\end{aligned}$$

By change of variables, $y_j = x_j/(1 - z_1 - \cdots - z_r)$, the density becomes

$$\begin{aligned}\frac{n(n-1)\cdots(n-r+1)\alpha^r\Gamma(\theta)}{\Gamma(\alpha+1)^r\Gamma(\theta-r\alpha)} z_1^{\alpha-1} \cdots z_r^{\alpha-1} (1 - z_1 - \cdots - z_r)^{\theta-r\alpha-1}&\\ \times P(\max_{r+1\le j\le n} Z'_j \le \frac{z_r}{1 - z_1 - \cdots - z_r}),&\end{aligned}$$

where $(Z'_{r+1}, \ldots, Z'_n)$ is $\mathrm{D}(\alpha, \ldots, \alpha)$. Hence, as $n \to \infty$,

$$f_{Z_{[1]},\ldots,Z_{[r]}}(z_1, \ldots, z_r) \to \frac{\theta^r (1 - z_1 - \cdots - z_r)^{\theta-1}}{z_1 \cdots z_r} P(V_1 \le \frac{z_r}{1 - z_1 - \cdots - z_r}).$$

As $(Z_{[1]}, \ldots, Z_{[r]}) \Rightarrow (V_1, \ldots, V_r)$, the assertion follows. □

3 Dickman's function

Dickman (1930) found the limiting distribution of the largest prime factor in a large integer. Surprisingly, the distribution is the same as that of V_1 for $\theta = 1$. Dickman's results have later been extended in different forms; see Donnelly and Grimmett (1993) and the references therein. We call $\rho_\theta(x) = P(1/V_1 > x)$ *Dickman's function* following Tenenbaum (1995, Chapter III.5), where the case $\theta = 1$ is studied.

Remark. Karl Dickman, born 1862, worked as an actuary in the end of the $19th$ and the beginning of the $20th$ centuary. Proabably, he studied mathematics in the 1880's at Stockholm University, where the legendary Mittag–Leffler was professor. However, Dickman's 1930 paper seems to be his only publication in mathematics. The paper is a remarkable achievement of an old man.

Proposition 3.1 *Dickman's function* $\rho_\theta(x) = P(1/V_1 > x)$ *is continuous for* $x \geq 0$ *and satisfies*

$$\rho_\theta(x) = 1, \quad for \quad 0 \leq x \leq 1,$$

$$x^\theta \rho'_\theta(x) + \theta(x-1)^{\theta-1}\rho_\theta(x-1) = 0, \quad for \quad x > 1,$$

$$x^\theta \rho_\theta(x) = \int_{x-1}^{x} \theta y^{\theta-1}\rho_\theta(y)\,dy, \quad for \quad x \geq 1,$$

$$\rho_\theta(x) = 1 + \sum_{j=1}^{[x]} \frac{(-\theta)^j}{j!} \int_{1/x}^{1} \cdots \int_{1/x}^{1} \frac{(1-y_1-\cdots-y_j)_+^{\theta-1}}{y_1\cdots y_j}\,dy_1\cdots dy_j.$$

Proof. From Proposition 2.3 we have

$$f_{V_1}(z) = \frac{\theta(1-z)^{\theta-1}}{z}\,P(V_1 \leq \frac{z}{1-z}), \quad 0 < z < 1.$$

Hence for $x > 1$

$$\rho'_\theta(x) = -f_{V_1}(1/x)/x^2 = -\frac{\theta(x-1)^{\theta-1}}{x^\theta}\,\rho_\theta(x-1).$$

Therefore for $x > 1$

$$(x^\theta\rho_\theta(x))' = \theta x^{\theta-1}\rho_\theta(x) + x^\theta\rho'_\theta(x) = \theta x^{\theta-1}\rho_\theta(x) - \theta(x-1)^{\theta-1}\rho_\theta(x-1).$$

As $\rho_\theta(y) = 1$ for $0 \leq y \leq 1$, it follows that

$$x^\theta \rho_\theta(x) = \int_{x-1}^{x} \theta y^{\theta-1}\rho_\theta(y)\,dy, \quad x \geq 1.$$

The explicit formula for ρ_θ is obtained in Proposition 2.1. □

Proposition 3.2 *Let* $V_1, U_1, U_2, U_3, \ldots$ *be independent random variables such that* $P(V_1 \leq v) = \rho_\theta(1/v)$ *and the U's be B(1,θ) (density* $\theta(1-u)^{\theta-1}I(0<u<1)$*). Then*

$$V_1, \quad \max\big(U_1, (1-U_1)V_1\big), \quad \max\big(U_1, (1-U_1)U_2, (1-U_1)(1-U_2)U_3, \ldots\big),$$

have the same distribution.

Proof. By Proposition 3.1 we get for $0 < v \leq 1$

$$P(\max(U_1, (1-U_1)V_1) \leq v)$$

$$= \int_0^v \theta(1-u)^{\theta-1} P(V_1 \leq v/(1-u))\, du = \int_0^v \theta(1-u)^{\theta-1} \rho_\theta((1-u)/v)\, du$$

$$= v^\theta \int_{1/v-1}^{1/v} \theta y^{\theta-1} \rho_\theta(y)\, dy = v^\theta v^{-\theta} \rho_\theta(1/v) = P(V_1 \leq v).$$

From this it follows that the random variables

$$V_1, \quad \max(U_1, (1-U_1)V_1), \quad \max(U_1, (1-U_1)U_2, (1-U_1)(1-U_2)V_1), \quad \ldots$$

have the same distribution. As almost surely

$$(1-U_1)(1-U_2)\cdots(1-U_n) \to 0, \quad n \to \infty,$$

the assertion follows. □

The explicit formula for $\rho_\theta(x)$ in Proposition 3.1 is not useful for calculations. However, the differential-difference equation can be solved recursively. For $1 < x < 2$ we get by integration by parts

$$\rho_\theta(x) = 1 - \int_1^x \theta(1-t)^{\theta-1} t^{-\theta}\, dt = 1 - \theta \sum_{j=0}^{\infty} (1-1/x)^{\theta+j}/(\theta+j),$$

and in particular $\rho_1(x) = 1 - \log x$. Griffiths (1988, Theorem 1) derives a recursive algorithm, useful for numerical computation of $\rho_\theta(x)$ ($= h(x)$ in Griffiths' notation), for the function

$$g_\theta(x) = e^{-\gamma\theta} x^{\theta-1} \rho_\theta(x)/\Gamma(\theta),$$

where γ is Euler's constant, using the following result in Watterson (1976).

Proposition 3.3 *The function g_θ is a probability density on the positive real half line with Laplace transform*

$$\int_0^\infty e^{-zx} g_\theta(x)\, dx = \exp\big(-\theta \int_0^1 \frac{1-e^{-zu}}{u}\, du\big).$$

Proof. As S_θ is $\Gamma(\theta)$ distributed and independent of V_1 we get

$$e^{-\theta E_1(z)} = P(X_{[1]} \leq z) = P(S_\theta V_1 \leq z) = \int_0^\infty P(V_1 \leq z/s) s^{\theta-1} e^{-s}/\Gamma(\theta)\, ds$$

$$= z^\theta \int_0^\infty e^{-zx} P(V_1 \le 1/x) x^{\theta-1}/\Gamma(\theta)\, dx = z^\theta e^{\gamma\theta} \int_0^\infty e^{-zx} g_\theta(x)\, dx.$$

Thus

$$\int_0^\infty e^{-zx} g_\theta(x) dx = z^{-\theta} e^{-\gamma\theta} e^{-\theta E_1(z)}$$

$$= \exp\big(-\theta(\int_1^\infty \frac{e^{-zu}}{u}\, du + \log z + \gamma)\big) = \exp\big(-\theta \int_0^1 \frac{1-e^{-zu}}{u}\, du\big),$$

using a well-known identity for the exponential integral. □

Note that g_θ is the density of an infinitely divisible probability distribution with the Lévy–Khinchine measure $\theta I(0 < u < 1)/u\, du$, whose Laplace transform is an entire analytic function. Inverting the transform we get for $x > 0$ and any real a:

$$\rho_\theta(x) = \Gamma(\theta) e^{\gamma\theta} x^{1-\theta} \frac{1}{2\pi i} \int_{a-i\infty}^{a+i\infty} e^{zx+H(z)}\, dz,$$

where

$$H(z) = -\theta \int_0^1 (1 - e^{-zu})/u\, du.$$

The saddle-point method can be used to study $\rho_\theta(x)$ for large x. Formally

$$g_\theta(x) = \frac{1}{2\pi i} \int_{a-i\infty}^{a+i\infty} e^{zx+H(z)}\, dz \approx \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{(a+it)x + H(a) + itH'(a) - t^2 H''(a)/2}\, dt.$$

Choosing a such that $x + H'(a) = 0$, we get for large x

$$g_\theta(x) \approx \frac{e^{ax+H(a)}}{\sqrt{2\pi H''(a)}}.$$

With $b = -a$ the equation $x + H'(a) = 0$ becomes $(e^b - 1)/b = x/\theta$, and for $x \to \infty$ we have $b = \log(x/\theta) + \log\log(x/\theta) + \cdots$.

These formal calculations can be justified, see Tenenbaum (1995, pp. 373–376) for the case $\theta = 1$. By small modifications of Tenenbaum's proof we obtain for the general case:

Proposition 3.4 *For Dickman's function*

$$\rho_\theta(x) = \Gamma(\theta) e^{\gamma\theta} x^{1-\theta} \frac{e^{-xb + \theta \int_0^b (e^y - 1)/y\, dy}}{\sqrt{2\pi(x - x/b + \theta/b)}} (1 + O(\frac{1}{x})), \quad x \to \infty,$$

where $(e^b - 1)/b = x/\theta$.

From this we get the following tail behaviour of $1/V_1$:

$$-\log P(1/V_1 > x) = -\log \rho_\theta(x) \asymp x \log x, \quad x \to \infty.$$

4 The GEM distribution and size-biased sampling

The content of this section can to a great extent be found in Kingman (1993, Section 9.6); see also Pitman (1996).

Let $P_* = (P_1, P_2, \dots)$ be a probability vector and ν_1 a random variable such that

$$P(\nu_1 = j \,|\, P_*) = P_j, \quad j = 1, 2, \dots \ .$$

Given P_* and ν_1 define ν_2 so that

$$P(\nu_2 = k \,|\, P_*, \nu_1) = P_k/(1 - P_{\nu_1}), \quad k = 1, 2, \dots, \ k \neq \nu_1.$$

Define similarly ν_3

$$P(\nu_3 = \ell \,|\, P_*, \nu_1, \nu_2) = P_\ell/(1 - P_{\nu_1} - P_{\nu_2}), \quad \ell = 1, 2, \dots, \ \ell \neq \nu_1, \nu_2,$$

and analogously $\nu_4, \nu_5, \dots$. The random probability vector

$$P_*^\pi = (P_{\nu_1}, P_{\nu_2}, P_{\nu_3}, \dots)$$

is a *size-biased permutation* of P_*. Note that any permutation of the original probability vector P_* has the same size-biased permutation P_*^π.

Proposition 4.1 *Let $Z_* = (Z_1, \dots, Z_n)$ be $D(\alpha, \dots, \alpha)$ distributed, $\theta = n\alpha$, and define $\nu_1, \nu_2, \dots, \nu_n$ as above with $P_* = Z_*$. Then*

$$T_1 = Z_{\nu_1}, \quad T_2 = Z_{\nu_2}/(1 - Z_{\nu_1}), \quad T_3 = Z_{\nu_3}/(1 - Z_{\nu_1} - Z_{\nu_2}), \quad \dots \quad ,$$

$$T_{n-1} = Z_{\nu_{n-1}}/(1 - Z_{\nu_1} - \cdots - Z_{\nu_{n-2}}),$$

are independent Beta distributed random variables such that T_j is $B(1 + \alpha, \theta - j\alpha)$ with density

$$f_{T_j}(x) = \frac{\Gamma(\theta - (j-1)\alpha + 1)}{\Gamma(1 + \alpha)\Gamma(\theta - j\alpha)} \, x^\alpha (1 - x)^{\theta - j\alpha - 1}, \quad 0 < x < 1,$$

for $j = 1, \dots, n-1$, and a size-biased permutation of Z_* has the representation

$$Z_*^\pi = \big(T_1, (1 - T_1)T_2, (1 - T_1)(1 - T_2)T_3, \dots, (1 - T_1)(1 - T_2) \cdots (1 - T_{n-1})T_n\big),$$

where $T_n \equiv 1$.

Proof. A straightforward calculation shows that $(Z_{\nu_1}, Z_1, \dots, Z_{\nu_1-1}, Z_{\nu_1+1}, \dots, Z_n)$ is $\mathrm{D}(1+\alpha, \alpha, \dots, \alpha)$ distributed. Well-known properties of the Dirichlet distribution or a direct calculation give that $T_1 = Z_{\nu_1}$ is $\mathrm{B}(1+\alpha, \theta-\alpha)$ distributed and independent of the vector $(Z_1, \dots, Z_{\nu_1-1}, Z_{\nu_1+1}, \dots, Z_n)/(1-T_1)$, which has a $\mathrm{D}(\alpha, \dots, \alpha)$ distribution.

The same argument can be used again on the last vector giving $T_2 = Z_{\nu_2}/(1-T_1)$ with a $\mathrm{B}(1+\alpha, \theta-2\alpha)$ distribution, etc. □

Next we generalize Proposition 3.2.

Proposition 4.2 *Let $V_* = (V_1, V_2, \dots)$ be PD(θ) distributed and V_*^π be a size-biased permutation of it. Then*

$$V_*^\pi = \big(U_1, (1-U_1)U_2, (1-U_1)(1-U_2)U_3, \dots\big),$$

where $U_1, U_2, \dots$ are independent identically distributed $B(1, \theta)$ random variables.

Proof. Let $Z_* = (Z_1, \dots, Z_n)$ be $\mathrm{D}(\alpha, \dots, \alpha)$ and $\alpha = \theta/n$. We have as $n \to \infty$

$$(Z_{[1]}, \dots, Z_{[n]}, 0, 0, \dots) \Rightarrow (V_1, V_2, \dots).$$

From the previous proposition we get

$$(Z_*^\pi, 0, 0, \dots) \Rightarrow \big(U_1, (1-U_1)U_2, (1-U_1)(1-U_2)U_3, \dots\big).$$

That this limiting distribution is the distribution of the size-biased permutation V_*^π is intuitively obvious; a rigerous proof can be found in Donnelly and Joyce (1989, Theorem 3). □

The distribution of V_*^π is the *GEM(θ) distribution* called after Griffiths, Engen and McCloskey, cf. Johnson, Kotz and Balakrishnan (1997, p. 237) and Pitman and Yor (1997, p. 858). The representation of the GEM distribution using independent identically distributed random variables makes it more 'user-friendly' than the Poisson–Dirichlet. Note that for any permutation-invariant function f the random variables $f(V_*^\pi)$ and $f(V_*)$ have the same distribution (provided they are well-defined), which can be used to simplify certain calculations.

Define for $0 \le \beta < 1$ and $\theta > -\beta$ the residual allocation model, cf. Pitman (1996),

$$V_{*\beta}^\pi := \big(U_{1\beta}, (1-U_{1\beta})U_{2\beta}, (1-U_{1\beta})(1-U_{2\beta})U_{3\beta}, \dots\big),$$

where the U's are independent and $U_{j\beta}$ is $\mathrm{B}(1-\beta, \theta+j\beta)$ for $j = 1, 2, \dots$. The distribution of the size-ordered permutation $V_{*\beta}$ of $V_{*\beta}^\pi$ defines a *two-parameter Poisson-Dirichlet distribution*, abbreviated *PD(β, θ)*; see Pitman and Yor (1997). Of course $\mathrm{PD}(0, \theta)$ is the same as $\mathrm{PD}(\theta)$. Also the general case $\mathrm{PD}(\beta, \theta)$ has many interesting properties and applications as shown by Pitman and Yor.

5 Ewens sampling formula

Let the probability vector $P_* = (P_1, P_2, \dots)$ be the subinterval-lengths of a division of the unit interval. Draw m points at random in the unit interval by the uniform distribution and let A_j be the number of subintervals containing exactly j of the points for $j = 1, 2, \dots$. Clearly

$$A_1 + 2A_2 + 3A_3 + \cdots = m.$$

Proposition 5.1 *Let $Z_* = (Z_1, \dots, Z_n)$ be $D(\alpha, \dots, \alpha)$ distributed, $\theta = n\alpha$, and define $A_1, A_2, \dots$ as above with $P_* = Z_*$. Then for non-negative integers $a_1, a_2, \dots$ with $a_1 + 2a_2 + 3a_3 + \cdots = m$ and $a_1 + a_2 + a_3 + \cdots = k$ we have*

$$P(A_j = a_j, j = 1, 2, \dots) = \frac{m!}{\prod_j a_j!\, j!^{a_j}} \frac{n(n-1)\cdots(n-k+1)}{\theta(\theta+1)\cdots(\theta+m-1)} \prod_j \Big(\frac{\Gamma(j+\alpha)}{\Gamma(\alpha)}\Big)^{a_j}.$$

Proof. Let $Y_1, \dots, Y_n$ be independent $\Gamma(\alpha)$ random variables with sum S_θ. Then $Z_* = (Y_1, \dots, Y_n)/S_\theta$ is $\mathrm{D}(\alpha, \dots, \alpha)$ distributed and independent of S_θ. The number of ways of drawing k different Y's, allocating m objects into k classes of which a_j contain exactly j objects, and not ordering the classes is

$$n(n-1)\cdots(n-k+1)\, \frac{m!}{\prod_j j!^{a_j}}\, \frac{1}{\prod_j a_j!}\,.$$

Thus by symmetry and independence we get

$$\begin{aligned}
&P(A_j = a_j, j = 1, 2, \dots) \\
&= n(n-1)\cdots(n-k+1)\, \frac{m!}{\prod_j j!^{a_j}\, a_j!}\, E\Big(\frac{Y_1}{S_\theta} \cdots \frac{Y_{a_1}}{S_\theta} \frac{Y_{a_1+1}^2}{S_\theta^2} \cdots \frac{Y_{a_1+a_2}^2}{S_\theta^2} \frac{Y_{a_1+a_2+1}^3}{S_\theta^3} \cdots\Big) \\
&= \frac{m!}{\prod_j a_j!\, j!^{a_j}}\, \frac{n(n-1)\cdots(n-k+1)}{E(S_\theta^{a_1+2a_2+\cdots})}\, E(Y_1)\cdots E(Y_{a_1})\, E(Y_{a_1+1}^2)\cdots \\
&= \frac{m!}{\prod_j a_j!\, j!^{a_j}}\, \frac{n(n-1)\cdots(n-k+1)}{\theta(\theta+1)\cdots(\theta+m-1)} \prod_j \Big(\frac{\Gamma(j+\alpha)}{\Gamma(\alpha)}\Big)^{a_j},
\end{aligned}$$

which proves the assertion. □

Proposition 5.2 *Let $V_* = (V_1, V_2, \dots)$ be PD(θ) distributed, and define $A_1, A_2, \dots$ as above with $P_* = V_*$. Then we have for non-negative integers $a_1, a_2, \dots$ with $a_1 + 2a_2 + 3a_3 + \cdots = m$ and $a_1 + a_2 + a_3 + \cdots = k$*

$$P(A_j = a_j, j = 1, 2, \dots) = \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+m-1)}\, \frac{m!}{\prod_{j=1}^m a_j!\, j^{a_j}}.$$

Proof. Letting $n \to \infty$, $\alpha = \theta/n \to 0$ and using $\alpha\Gamma(\alpha) = \Gamma(\alpha+1) \to 1$ we get

$$\frac{m!}{\prod_j a_j!\, j!^{a_j}} \frac{n(n-1)\cdots(n-k+1)\alpha^k}{\theta(\theta+1)\cdots(\theta+m-1)} \prod_j \Big(\frac{\Gamma(j+\alpha)}{\Gamma(\alpha)}\Big)^{a_j}$$

$$\to \frac{m!}{\prod_j a_j!\, j^{a_j}} \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+m-1)}\,.$$

As we have convergence of the ordered Z_* to V_*, that is

$$(Z_{[1]}, Z_{[2]}, \dots, Z_{[n]}, 0, 0, \dots) \Rightarrow (V_1, V_2, \dots), \quad n \to \infty,$$

the assertion follows from Proposition 5.1 and symmetry. □

The distribution of $(A_1, A_2, \dots)$ for $P_* = V_*$ in Proposition 5.2 is the famed *Ewens sampling formula*, abbreviated $ESF(\theta)$. It has been established for many different models; see Arratia, Barbour and Tavaré (2001), Johnson, Kotz and Balakrishnan (1997, Chapter 41), and the references therein. In the next section we will see how it comes up in connection with cycle-lengths of random permutations. We have the following representation of ESF(θ); this is the starting-point of the thorough investigations by Arratia et al.

Proposition 5.3 *Let $T_1, T_2, \dots$ be independent Poisson random variables with $E(T_j) = \theta/j$. Then for any integer $m \geq 0$ the conditional distribution of $(T_1, T_2, \dots)$ given $T_1 + 2T_2 + \cdots = m$ is ESF(θ).*

Proof. For non-negative integers $a_1, a_2, \dots$ with $a_1 + 2a_2 + \cdots = m$ we have

$$P(T_j = a_j, j = 1, 2, \dots \,|\, T_1 + 2T_2 + \cdots = m)$$

$$= P(T_j = a_j, j = 1, 2, \dots m)/P(T_1 + 2T_2 + \cdots + mT_m = m)$$

$$= \prod_{j=1}^{m} \frac{(\theta/j)^{a_j}\, e^{-\theta/j}}{a_j!} \,/P(T_1 + 2T_2 + \cdots + mT_m = m).$$

Hence, the assertion is proved if

$$P(T_1 + 2T_2 + \cdots + mT_m = m) = \frac{\theta(\theta+1)\cdots(\theta+m-1)}{m!} e^{-\theta\sum_1^m 1/j}\,.$$

This follows from the following calculation with generating functions:

$$E(s^{T_1+2T_2+\cdots+mT_m}) = \exp\big(\sum_{j=1}^{m}(s^j - 1)\theta/j\big)$$

$$= \exp\big(-\sum_1^m \theta/j - \theta\log(1-s) - \theta\sum_{m+1}^\infty s^j/j\big) = e^{-\theta\sum_1^m 1/j}(1-s)^{-\theta}e^{-\theta\sum_{m+1}^\infty s^j/j}$$

$$= e^{-\theta\sum_1^m 1/j}\left(\sum_{\ell=0}^m \binom{-\theta}{\ell}(-s)^\ell + \sum_{\ell=m+1}^\infty b_\ell s^\ell\right).$$

□

6 Random permutations

Every permutation of the numbers $1, 2, \ldots, m$ can be broken down into cycles. The study of such cycles has a long history going back at least to Cauchy. In the seminal paper by Shepp and Lloyd (1966) asymptotics of cycle-lengths in random permutations (uniform distribution) are investigated; see also the references therein. More general combinatorial structures and other distributions than the uniform for such objects are thoroughly investigated in the coming book by Arratia, Barbour and Tavaré (2001), which also contains an extensive list of references. Below we will study cycle-lengths in random permutations weighted by the number of cycles. Our treatment is influenced by Arratia et al.

In the rest of this section we mean by a *random permutation*, a permutation generated in the manner specified below by what is sometimes called the *Chinese restaurant process*. This is also closely connected with the so called *Hoppe's urn*.

Initially a list is empty and an urn contains one black ball of weight $\theta > 0$. Balls are successively drawn from the urn with probabilities proportional to weights. Each drawn ball is replaced into the urn together with a new ball of weight 1 and numbered by the *drawing number*. That number is also written in the list. At the first drawing 1 is written. If at drawing j the black ball is drawn, then j is written to the *left* of the list, else j is written *just to the right* of the number of the drawn ball in the list. In such a list the drawings $N_1 \equiv 1 < N_2 < N_3 < \cdots$ at which the black ball was drawn can be identified. Say, that after 8 drawings, the list is 5 8 3 6 4 1 2 7, then $N_1 \equiv 1$, $N_2 = 3$, $N_3 = 5$, and the urn contains balls numbered $1, 2, \ldots, 8$ and the black ball.

After m drawings let K be the number of times the black ball was obtained and let $N_1 \equiv 1 < N_2 < \cdots < N_K$ be the corresponding drawing numbers. Denote by $\mathcal{C}_1$ the part of the list from $N_1(\equiv 1)$ to the right; taking away $\mathcal{C}_1$, let $\mathcal{C}_2$ be the list from N_2 to the right, etc. The permutation with cycles $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$ defines our *random permutation*. The lengths of cycles are $C_1, C_2, \ldots, C_K$. Let A_j be the number of cycles of length $j = 1, 2, \ldots$, that is the cycle-length count. Obviously,

$$A_1 + A_2 + \cdots + A_m = K,\ C_1 + C_2 + \cdots + C_K = A_1 + 2A_2 + \cdots + mA_m = m.$$

In the above example we have: $m = 8$, $K = 3$, $\mathcal{C}_1 = (1\ 2\ 7)$, $\mathcal{C}_2 = (3\ 6\ 4)$, $\mathcal{C}_3 = (5\ 8)$, $C_1 = 3$, $C_2 = 3$, $C_3 = 2$, $A_1 = 0$, $A_2 = 1$, $A_3 = 2$.

The above drawing scheme is such that any in advance specified permutation of $1, 2, \ldots, m$ with k cycles has probability $\theta^k/\theta(\theta+1)\cdots(\theta+m-1)$. Thus, denoting the number of permutations with k cycles by $\left[\begin{smallmatrix} m \\ k \end{smallmatrix}\right]$,

$$\sum_{k=1}^{m} \left[\begin{matrix} m \\ k \end{matrix}\right] \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+m-1)} = \sum_{k=1}^{m} P(K=k) = 1.$$

Therefore

$$\theta(\theta+1)\cdots(\theta+m-1) = \sum_{k=1}^{m} \left[\begin{matrix} m \\ k \end{matrix}\right] \theta^k$$

is the generating function for the *cycle numbers* $\left[\begin{smallmatrix} m \\ k \end{smallmatrix}\right]$, usually called (sign-less) Stirling numbers of the first kind, cf. Knuth (1992).

Proposition 6.1 *Let K be the number of cycles in a random permutation of the numbers $1, 2, \ldots, m$. Then*

$$P(K=k) = P(I_1 + I_2 + \cdots + I_m = k) = \left[\begin{matrix} m \\ k \end{matrix}\right] \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+m-1)},$$

where $I_1, I_2, \ldots, I_m$ are independent Bernoulli random variables such that

$$P(I_j = 1) = 1 - P(I_j = 0) = \theta/(\theta+j-1).$$

Proof. The probability of getting the black ball in drawing j is $\theta/(\theta+j-1)$ independently of what has happened in the previous drawings. Let I_j be the indicator random variable for that event. These indicators are independet and K is the sum of them. □

Proposition 6.2 *Let in a random permutation the cycle-lengths (from right to left) be $C_1, C_2, \ldots$. Then*

$$P(C_1 = j) = \theta \frac{(m-1)(m-2)\cdots(m-j+1)}{(\theta+m-j)(\theta+m-j+1)\cdots(\theta+m-1)};$$

the conditional distribution of $(C_2, C_3, \ldots)$ given $C_1 = j$ is the same as the cycle-length distribution in a random permutation of $1, 2, \ldots, m-j$.

Proof. The number of permutations with $C_1 = j$ and with k cycles is

$$(m-1)(m-2)\cdots(m-j+1) \left[\begin{matrix} m-j \\ k-1 \end{matrix}\right];$$

note that $\mathcal{C}_1$ begins with a 1 and has $m-j$ numbers to the left with $k-1$ cycles. As permutations with the same number of cycles have the same probability (given above) we get:

$$P(C_1 = j, K = k) = (m-1)(m-2)\cdots(m-j+1) \begin{bmatrix} m-j \\ k-1 \end{bmatrix} \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+m-1)}.$$

Thus

$$P(C_1 = j) = \sum_{k=1}^{m} P(C_1 = j, K = k)$$

$$= \theta \frac{(m-1)(m-2)\cdots(m-j+1)}{\theta(\theta+1)\cdots(\theta+m-1)} \sum_k \begin{bmatrix} m-j \\ k-1 \end{bmatrix} \theta^{k-1}$$

$$= \theta \frac{(m-1)(m-2)\cdots(m-j+1)}{\theta(\theta+1)\cdots(\theta+m-1)} \theta(\theta+1)\cdots(\theta+m-j-1)\,,$$

proving the first assertion. The second assertion is obvious from the construction of the random permutation using the Chinese restaurant process. □

Proposition 6.3 *The cycle-length count* $(A_1, A_2, \dots)$ *in a random permutation is ESF(*θ*) distributed.*

Proof. The number of permutations of $1, 2, \dots, m$ with a_1 cycles of length 1, a_2 cycles of length 2, $\dots$, $a_1 + 2a_2 + \cdots = m$, is

$$\frac{m!}{\prod_j j!^{a_j}} \frac{\prod_j (j-1)!^{a_j}}{\prod_j a_j!} = \frac{m!}{\prod_j a_j! j^{a_j}};$$

the formula goes back to Cauchy. As each permutation with $k = a_1 + \cdots + a_m$ cycles has the probability $\theta^k/\theta(\theta+1)\cdots(\theta+m-1)$, the assertion follows. □

The next result gives the limiting distribution of the lengths of *long* cycles. This was first obtained for the case $\theta = 1$ by Shepp and Lloyd (1966). Thorough investigations on asymptotics of cycles and other combinatorial stuctures including results on rates of convergence are given in Arratia, Barbour and Tavaré (2001). Representations using independent random variables such as in Proposition 5.3 play an important rôle there. The limit behaviour for the number of cycles in a random permutation can be obtained using the representation in Proposition 6.1 of K as a sum of independent Bernoulli random variables, see Arratia et al.

Proposition 6.4 *In a random permutation the cycle-lengths converge as*

$$(C_1/m, C_2/m, \dots, C_K/m, 0, 0, \dots) \Rightarrow GEM(\theta), \quad m \to \infty,$$

and the size-ordered cycle-lengths as

$$(C_{[1]}/m, C_{[2]}/m, \dots, C_{[K]}/m, 0, 0, \dots) \Rightarrow PD(\theta), \quad m \to \infty.$$

Proof. Using the Gamma function and Proposition 6.2 we have

$$P(C_1 = j) = \frac{\theta}{m} \frac{\Gamma(m+1)}{\Gamma(m-j+1)} \frac{\Gamma(m-j+\theta)}{\Gamma(m+\theta)}.$$

For $0 < x < 1$, and $j, m \to \infty$ such that $j/m \to x$, Stirling's formula gives

$$m\, P(C_1 = j) \to \theta\, (1-x)^{\theta-1},$$

that is $C_1/m \Rightarrow B(1, \theta)$. The assertion follows by combining the second part of Proposition 6.2 with Proposition 4.2. Note that $(C_1, \dots, C_K)$ is a size-biased permutation of the ordered cycle-lengths. □

A similar urn scheme as in the Chinese restaurant process is *Hoppe's urn*, where the balls are coloured instead of numbered. The black ball has weight θ and other balls weight one. A drawn ball is replaced together with one of the same colour except the black, which is replaced together with one ball of a colour not already in the urn. The results in the propositions in this section describe the colour composition in the urn. For example, the ESF(θ) is the distribution of the composition after m drawings, and PD(θ) is the limit distribution as $m \to \infty$ of the random proportions arranged in decreasing order of the different colours in the urn.

References

[1] Arratia, R., Barbour, A.D. and Tavaré, S. (2001). *Logarithmic Combinatorial Structures: a Probabilistic Approach.* Book draft dated August 13, 2001, available on Internet.

[2] Dickman, K. (1930). On the frequency of numbers containing prime factors of a certain relative magnitude. *Arkiv för Matematik, Astronomi och Fysik* **22**, 1–14.

[3] Donnelly, P. and Grimmett, G. (1993). On the asymptotic distribution of large prime factors. *J. London Math. Soc.* **47**, 395–394.

[4] Donnelly, P. and Joyce, P. (1989). Continuity and weak convergence of ranked and size-biased permutations on the infinite simplex. *Stochastic Process. Appl.* **31**, 89–103.

[5] Griffiths, R.C. (1979). On the distribution of allele frequencies in a diffusion model. *Theoret. Popn. Biol.* **15**, 140–158.

[6] Griffiths, R.C. (1988). On the distribution of points in a Poisson process. *J. Appl. Prob.* **25**, 336–345.

[7] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions.* Wiley, New York.

[8] Kingman, J.F.C. (1975). Random discrete distributions. *J. R. Statist. Soc. B* **37**, 1–15.

[9] Kingman, J.F.C. (1980). *Mathematics of Genetic Diversity.* Society for Industrial and Applied Mathematics, Philadelphia, PA.

[10] Kingman, J.F.C. (1993). *Poisson Processes.* Oxford University Press.

[11] Knuth, D. (1992). Two notes on notations. *Amer. Math. Monthly.* **99**, 403–422.

[12] Pitman, J. (1996). Random discrete distributions invariant under size-biased permutation. *Adv. Appl. Prob.* **28**, 525–539.

[13] Pitman, J. and Yor, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *Ann. Prob.* **25**, 855–900.

[14] Shepp, L. and Lloyd, S. (1966). Ordered cycle lengths in a random permutation. *Amer. Math. Soc.* **121**, 340–357.

[15] Tenenbaum, G. (1995). *Introduction to Analytic and Probabilistic Number Theory.* Cambridge University Press.

[16] Watterson, G.A. (1976). The stationary distribution of the infinitely-many neutral alleles diffusion model. *J. Appl. Prob.* **13**, 639–651.