Combinatorics, Probability and Computing (1997) 6, 185–195
Copyright © 1997 Cambridge University Press

From GEM back to Dirichlet via Hoppe's Urn

ULRICH MARTIN HIRTH†

Mathematics Department, Royal Holloway and Bedford New College, University of London, Egham Hill, Surrey TW20 0EX, UK

Received 20 February 1995; revised 19 September 1995

In generalisation of the beta law obtained under the GEM/Poisson–Dirichlet distribution in Hirth [12] we undertake here an analogous construction which results in the Dirichlet law. Our proof makes use of Hoppe's Pólya-like urn model in population genetics.

Foreword

The central result of this paper, the Dirichlet law of Theorem 2 (Section 2), and hence also Theorem 1 (Section 1), has already been shown by Donnelly and Tavaré [9, §3, pp. 387–388], though using a completely different approach. The proof in Donnelly and Tavaré [9] is based on the construction of the Poisson–Dirichlet distribution by means of an inhomogeneous Poisson process and uses Poisson process arguments.

By contrast, our proof is based on the rather elementary construction of the GEM distribution by means of a sequence of independent, identically distributed random variables. Furthermore, we use Hoppe's urn which is more closely related to the GEM distribution than to the Poisson–Dirichlet measure, and we feel that our sub-urn argument in §3 may be interesting in its own right, as an idea for future developments in related fields.

1. Introduction: GEM/Poisson–Dirichlet and beta laws

Let Z_k, $k \in \mathbb{N} := \{1, 2, \ldots\}$ be independent random variables, identically distributed each with the beta $(1, \theta)$ law, with $\theta > 0$ remaining fixed throughout Sections 1–3. Set

$$X_1 = Z_1, \qquad X_k = Z_k \prod_{j=1}^{k-1} (1 - Z_j), \qquad k \in \mathbb{N}. \tag{1.1}$$

† Now at Mathematisch-Geographische Fakultät, Katholische Universität Eichstätt, Ostenstr. 26-28, D–85071 Eichstätt, Germany. E-mail: ulrich.hirth@eo_nwfs-1.ku_eichstaett.de

https://doi.org/10.1017/S0963548396002842 Published online by Cambridge University Press

The distribution of the random vector $X = (X_1, X_2, \ldots)$ is called the *GEM* (Griffiths-Engen-McCloskey) law with parameter θ. Closely related to the GEM law is the *Poisson–Dirichlet* measure (with parameter θ), which is obtained by applying the non-increasing order statistic to X. They are both essentially the same distribution, since the GEM law can be recovered from the Poisson–Dirichlet measure by *size-biasing* (see [5, 7, 8]). The GEM law is usually easier to handle. Moreover, we will be interested in the distribution of functions of X which do not depend upon the labelling of the components, and for this reason it will be irrelevant which one of the two laws we take (see, for example, Hirth [12]). The GEM/Poisson–Dirichlet law has proved significant as a limit measure in various contexts, including neutral allele models in population genetics (see [9–11, 13–15, 18–20, 26, 27, 32]), random permutations (see [16, 28, 31]), random mappings (see [6]) and the prime factorisation of a random integer (see [7, 22, 25, 30]). It was surprisingly the last of these which gave the impetus to the present research line. Namely in number theory there is an arc-sine law (=beta $(1/2, 1/2)$ law) by Deshouillers, Dress and Tenenbaum [4] (see also Tenenbaum [29, II.6.2, p.233]) for the mean distribution of the divisors of a random integer, and the author of this paper has recognised that something similar can be done in the Donnelly-Grimmett [7] framework, equally resulting in the arc-sine law, and more generally, resulting in the beta $(\theta/2, \theta/2)$ law (see Hirth [12]).

In this paper we generalise this to the Dirichlet law, which we will prove by induction. The start of this induction is obtained by the methods of Hirth [12, §3], but slightly more generally it is the beta $(\alpha\theta, (1-\alpha)\theta)$ law with $0 < \alpha < 1$. For this reason we include both statement and proof for it.

Let B_k, $k \in \mathbb{N}$ be independent random variables, identically distributed each

$$P[B_k = 1] \;=\; 1 - P[B_k = 0] \;= \alpha,$$

and independent of X.

Theorem 1.1. *The random variable*

$$S_0 := \sum_{k=1}^{\infty} B_k X_k \tag{1.2}$$

has the distribution beta $(\alpha\theta, (1-\alpha)\theta)$.

Proof. We use the method of moments and the recursion idea of Lloyd and Williams [24]. Let Z and B denote independent copies of Z_k and B_k, respectively. Consider the moments $(l \in \mathbb{N})$:

$$\begin{aligned} E[S_0^l] &= E\left[\left(BZ + (1-Z)\sum_{k=1}^{\infty} B_{k+1}X_k\right)^l\right] \\ &= (1-\alpha)\cdot E[(1-Z)^l]\cdot E[S_0^l] + \alpha \cdot E[(Z + (1-Z)S_0)^l]. \end{aligned}$$

https://doi.org/10.1017/S0963548396002842 Published online by Cambridge University Press

Simplifying this, we get

$$\begin{aligned} E[S_0^l] &= \frac{\alpha}{1-(1-\alpha)E[(1-Z)^l]} \cdot E[(Z+(1-Z)S_0)^l] \\ &= \frac{\alpha}{1-(1-\alpha)E[(1-Z)^l]} \cdot \sum_{j=0}^{l} \binom{l}{j} \cdot E[Z^{l-j} \cdot (1-Z)^j] \cdot E[S_0^j] \\ &= \frac{\alpha}{1-(1-\alpha)\theta/(\theta+l)} \cdot \sum_{j=0}^{l} \binom{l}{j} \frac{\theta \cdot \Gamma(\theta+j) \cdot \Gamma(l-j+1)}{\Gamma(\theta+l+1)} \cdot E[S_0^j] \\ &= \frac{\alpha\theta}{l+\alpha\theta} \cdot \frac{l!}{\Gamma(\theta+l)} \cdot \sum_{j=0}^{l} \frac{\Gamma(\theta+j)}{j!} \cdot E[S_0^j]. \end{aligned}$$

Cancelling the $E[S_0^l]$ term, we arrive at

$$E[S_0^l] = \frac{\alpha\theta \cdot (l-1)!}{\Gamma(\theta+l)} \cdot \sum_{j=0}^{l-1} \frac{\Gamma(\theta+j)}{j!} \cdot E[S_0^j],$$

and we can solve this recursive equation explicitly:

$$E[S_0^{l+1}] - \frac{l}{\theta+l} \cdot E[S_0^l] = \frac{\alpha\theta}{\theta+l} \cdot E[S_0^l]$$

and

$$E[S_0^{l+1}] = \frac{l+\alpha\theta}{l+\theta} \cdot E[S_0^l].$$

This formula holds for $l = 0$ as well ($E[S_0^0] = 1$). So we get by induction

$$\begin{aligned} E[S_0^l] &= \frac{\Gamma(l+\alpha\theta) \cdot \Gamma(\theta)}{\Gamma(\alpha\theta) \cdot \Gamma(l+\theta)} \\ &= \frac{\Gamma(l+\alpha\theta) \cdot \Gamma((1-\alpha)\theta)}{\Gamma(l+\theta)} \cdot \frac{\Gamma(\theta)}{\Gamma(\alpha\theta) \cdot \Gamma((1-\alpha)\theta)}, \end{aligned}$$

the moments of the beta $(\alpha\theta, (1-\alpha)\theta)$ law. □

2. Statement of the new result: Dirichlet law

Speaking in terms of population genetics, what we have done in Section 1 is to divide the infinitely many neutral alleles into two subsets, each allele independently going with probability α into the first and with probability $1-\alpha$ into the second, and considering the distribution of the proportion of the population in the first subset.

Generalising this, we will next partition the alleles into $d+1$ subsets, $d \in \mathbb{N}$. Fix real numbers α_m, $m \in \{0, \ldots, d\}$ with $0 < \alpha_m < 1$ and $\sum_{m=0}^{d} \alpha_m = 1$. Think of α_m as the probability that an allele is grouped into the m'th subset, independently for each allele.

We will find it convenient to express this formally in terms of random variables as follows: Let $\beta_{k,m}$ be such that for each $k \in \mathbb{N}$, $m \in \{0, \ldots, d\}$

$$P[\beta_{k,m} = 1] = 1 - P[\beta_{k,m} = 0] = \alpha_m \tag{2.1}$$

https://doi.org/10.1017/S0963548396002842 Published online by Cambridge University Press

and

$$P\left[\sum_{m=0}^{d}\beta_{k,m}=1\right]\quad=1; \tag{2.2}$$

and let the collection $\beta_k := (\beta_{k,0},\ldots,\beta_{k,d})$, $k\in\mathbb{N}$ be a sequence of independent random vectors, additionally independent of the GEM vector X. Think of $\beta_{k,m}$ as being 1 or 0 according to whether the k'th oldest allele is grouped into the m'th subset or not. We are interested in the distribution of the induced partition of the population:

$$S=(S_0,\ldots,S_d) \tag{2.3}$$

with

$$S_m=\sum_{k=1}^{\infty}\beta_{k,m}X_k, \tag{2.4}$$

$m\in\{0,\ldots,d\}$. Our double-use of the notation S_0 both for (1.2) and (2.3)–(2.4) is intentional for the later proof by induction, where we will set $\alpha=\alpha_0$. The central result of this paper is

Theorem 2.1. *S has the Dirichlet distribution with parameters $(\alpha_0\theta,\ldots,\alpha_d\theta)$.*

The Dirichlet distribution with parameters $(a_0,\ldots,a_d)$, $d\in\mathbb{N}$, $a_m>0$ for all $m\in\{0,\ldots,d\}$, is concentrated on the d-dimensional simplex

$$\Delta_d := \left\{(s_0,\ldots,s_d)\in[0,1]^{d+1} : \sum_{m=0}^{d}s_m=1\right\}, \tag{2.5}$$

and its density with respect to the d-dimensional Lebesgue measure is the mapping $\Delta_d\to\mathbb{R}$ given by

$$(s_0,\ldots,s_d)\to\frac{\Gamma(a_0+\ldots+a_d)}{\Gamma(a_0)\cdots\Gamma(a_d)}\;s_0^{a_0-1}\cdots s_d^{a_d-1}. \tag{2.6}$$

This is clearly a generalisation of the $d=1$ case, the beta (a_0,a_1) distribution. There is a nice – and for our purposes especially suitable – characterisation of the Dirichlet distribution by recursion on d (see Johnson and Kotz [17, ch. 40, §5, p. 234], Connor and Mosimann [2], and for further background Darroch and Ratcliff [3]). For S to have the asserted Dirichlet distribution, it is necessary and sufficient that

$$S_0 \text{ has the beta } (\alpha_0\theta,(1-\alpha_0)\theta) \text{ distribution}, \tag{2.7}$$

$$S' := \left(\frac{S_1}{1-S_0},\ldots,\frac{S_d}{1-S_0}\right) \text{ has the Dirichlet } (\alpha_1\theta,\ldots,\alpha_d\theta) \text{ distribution}, \tag{2.8}$$

and

$$\{S_0,S'\} \text{ is stochastically independent.} \tag{2.9}$$

We have already proved (2.7) with Theorem 1 (Section 1). In Section 3 we will prove Theorem 2 via (2.8) and (2.9) by induction on d, where in the case $d=1$ Theorem 2, of course, reduces to Theorem 1. We will thereby bring into play Hoppe's Pólya-like urn model from population genetics.

https://doi.org/10.1017/S0963548396002842 Published online by Cambridge University Press

The Dirichlet distribution occurs in population genetics as well as the GEM/Poisson–Dirichlet law. In fact, Kingman [19] (see also his recent book [21, ch. 9]) used the Dirichlet law to *define* the Poisson–Dirichlet measure, as follows: take the symmetric Dirichlet $(a,\dots,a)$ law with $d+1$ parameters, $a>0$, and let $(d+1)\cdot a$, the sum of the parameters (which in our case of Theorem 2 equals θ), converge to the limit $\theta>0$ as $d\to\infty$, $a\to 0$. Then, after applying the order statistic, the ordered Dirichlet laws converge weakly to the Poisson–Dirichlet measure with parameter θ. From this point of view, the novelty of our Theorem 2 consists in providing a way from the limiting Poisson–Dirichlet law back to the Dirichlet law.

3. Proof of Theorem 2 via the Hoppe urn

Let $d\geq 2$. We plan to show (2.8) and (2.9) under the induction hypothesis that Theorem 2 holds for $d-1$ replacing d. What we are intending to achieve becomes evident when writing the parameter vector of the Dirichlet distribution in (2.8) in the form

$$\left(\frac{\alpha_1}{1-\alpha_0}\cdot(1-\alpha_0)\theta,\dots,\frac{\alpha_d}{1-\alpha_0}\cdot(1-\alpha_0)\theta\right). \tag{3.1}$$

To make our idea clear, set

$$\alpha'_m=\frac{\alpha_{m+1}}{1-\alpha_0} \tag{3.2}$$

for $m\in\{0,\dots,d-1\}$ and

$$\theta'=(1-\alpha_0)\theta. \tag{3.3}$$

We have $\sum_{m=0}^{d-1}\alpha'_m=1$, $0<\alpha'_m<1$ for $m\in\{0,\dots,d-1\}$ and $\theta'>0$. We have in mind to show that the distribution of S' in (2.8) is equal to the distribution of S with $d-1$ replacing d, the numbers α'_m taking over the role of the probabilities α_m and the quantity θ' replacing the parameter θ of the GEM distribution. Then, by the induction hypothesis, we are through provided that we can also show (2.9).

Each time when an allele is grouped into the zero'th subset (with probability α_0), think of this allele being sorted out and concentrate on the partition of the remaining alleles into $(d-1)+1=d$ subsets, a problem we have already solved by the induction hypothesis, thereby planning that the remaining alleles are distributed, after appropriate rescaling, according to the GEM/Poisson–Dirichlet law with parameter θ'.

We find that this plan is made precise and realised most easily by means of Hoppe's Pólya-like urn model from population genetics (see Hoppe [13, 15]).

The urn $\mathscr{U}$ starts with one black ball of mass θ only. This ball is drawn on the first move and returned together with one additional ball of unit mass 1 and any non-black colour. From now on, on each move one ball is drawn with probability proportional to its mass. If the drawn ball is non-black, then it is returned together with one additional ball of the same colour and unit mass 1. If the drawn ball is black, then it is returned together with one additional ball of a new, previously unobserved colour and unit mass 1.

We label the colours as usual by natural numbers, starting with 1 for the colour of the first non-black ball and proceeding in order of their appearance in the urn. We adopt the

https://doi.org/10.1017/S0963548396002842 Published online by Cambridge University Press

language of the population geneticists, saying that we label the alleles (=colours) in age order.

Let $X_k(n)$ be the normalised number of non-black balls in $\mathscr{U}$ labelled k, $k \in \mathbb{N}$ (i.e. divided by n, the total number of non-black balls in $\mathscr{U}$ after the n'th draw and return). We have almost sure convergence of $X_k(n)$ to a limit random variable X_k for each $k \in \mathbb{N}$, as $n \to \infty$, and this $X := (X_1, X_2, \ldots)$ has the GEM distribution with parameter θ – our double-use of the notation $X = (X_1, X_2, \ldots)$ for both (1.1) and for the limit of the random variables $X_k(n)$ is intentional and cannot lead to any confusion.

Within this urn model $\mathscr{U}$ we can control each allele by means of the black mass θ whose draw generates this allele. For our purpose of proof by induction, we partition the black mass θ into one black ball of mass $\alpha_0\theta$ plus a second black ball of mass $(1-\alpha_0)\theta$. To simplify notation and to link our arguments with Theorem 1, we write α for α_0. Conditional on the event of black being drawn, it will be the first black ball with probability α and the second with probability $1-\alpha$, independently for each draw of black. The allele generated by such a draw of black mass will then be put into one of two sub-urns, $\mathscr{U}_0$ and $\mathscr{U}_1$, according to whether it is initiated by the first or the second black ball, respectively. As already indicated, we will then perform a further partitioning of alleles within $\mathscr{U}_1$.

Coming next to the formal elaboration of this idea, consider the urn $\mathscr{U}$ as consisting of the two sub-urns $\mathscr{U}_0$ and $\mathscr{U}_1$ as above. Initially, $\mathscr{U}_0$ contains only one black ball of mass $\alpha\theta$, and $\mathscr{U}_1$ contains only one black ball of mass $(1-\alpha)\theta$, where $\theta > 0$ and $0 < \alpha < 1$. To understand the following arguments, it is helpful to regard $\mathscr{U}$ as containing the balls of both $\mathscr{U}_0$ and $\mathscr{U}_1$, and as containing initially only one black ball of mass $\alpha\theta + (1-\alpha)\theta = \theta$. When regarded in this manner, $\mathscr{U}$ will develop according to the rules of Hoppe's Pólya-like urn process with black mass parameter θ, and in particular, we will label the colours by age order within the whole of $\mathscr{U}$ as usual. Each time the black mass θ is drawn and thus a new colour generated, we consider the black ball to be drawn from the sub-urn $\mathscr{U}_0$ with (conditional) probability α, and from $\mathscr{U}_1$ with probability $1-\alpha$. The first ball of the new colour will then be placed into $\mathscr{U}_0$ or $\mathscr{U}_1$ according to from where the generating black ball has been drawn. It is immediately clear by our construction that $\mathscr{U}_0$ and $\mathscr{U}_1$ develop each according to the rules of a Hoppe process with black mass parameters $\alpha\theta$ and $(1-\alpha)\theta$, respectively, up to an obvious (random) relabelling of alleles and a (random) transformation of time – provided only that both $\mathscr{U}_0$ and $\mathscr{U}_1$ are served infinitely often, but this is almost surely the case, because we know that the black mass θ is drawn almost surely infinitely often: the limiting GEM distribution (see Hoppe [15]) almost surely assigns positive frequency to each of the infinitely many alleles.

Define

$$G_k = j \iff \text{the } k\text{'th draw of black mass is from sub-urn } \mathscr{U}_j,\ j \in \{0,1\},\ k \in \mathbb{N}.$$

We have

$$\alpha = P[G_k = 0] = 1 - P[G_k = 1]$$

for all $k \in \mathbb{N}$, and the sequence of random variables $(G_k)_{k\in\mathbb{N}}$ is independent. We have $G_k = j$ if and only if all the balls of colour k are in sub-urn $\mathscr{U}_j$. As already indicated,

https://doi.org/10.1017/S0963548396002842 Published online by Cambridge University Press

we consider $\mathscr{U}_0$ and $\mathscr{U}_1$ as Hoppe urn processes (via the obvious relabelling of alleles and time-rescaling) which are as such stochastically independent of the sequence of random variables $(G_k)_{k\in\mathbb{N}}$.

So as not to get confused by relabelling alleles within $\mathscr{U}_j$, $j \in \{0,1\}$, we speak unambigously of the l'th oldest allele within $\mathscr{U}_j$, $l \in \mathbb{N}$.

Let $X_{\mathscr{U},k}(n)$ denote the normalised number of balls representing the k'th oldest allele within $\mathscr{U}$ at time n (i.e. divided by n, the total number of non-black unit balls in the whole of the urn $\mathscr{U}$ at time n). Theorems 1 and 2 (plus the Corollary) of Hoppe [15] state that $X_{\mathscr{U},k}(n)$ converges almost surely $(n \to \infty)$ to X_k with $X = (X_1, X_2, \ldots)$ being distributed as GEM with parameter θ.

For $j \in \{0,1\}$ and n the same time as above, let $X_{j,l}(n)$ be the normalised number of balls representing the l'th oldest allele within $\mathscr{U}_j$ at time n (i.e. now divided by the total number of non-black unit balls in the sub-urn $\mathscr{U}_j$ at time n). The sequence $X_{j,l}(n)$, $n \in \mathbb{N}$ converges almost surely to $X_{j,l}$ with $(X_{0,1}, X_{0,2}, \ldots)$ being distributed as GEM with parameter $\alpha\theta$ and $(X_{1,1}, X_{1,2}, \ldots)$ being distributed as GEM with parameter $(1-\alpha)\theta$. Define and calculate

$$\begin{aligned} S_0(n) :=& \frac{\text{number of non-black balls in } \mathscr{U}_0 \text{ at time } n}{\text{number of non-black balls in } \mathscr{U} \text{ at time } n} \\ =& \frac{1}{n} \cdot (\text{number of non-black balls in } \mathscr{U}_0 \text{ at time } n) \\ =& \frac{1}{n} \cdot \sum_{k=1}^{\infty} (1 - G_k) \cdot n \cdot X_{\mathscr{U},k}(n) \\ \to& \sum_{k=1}^{\infty} (1 - G_k) \cdot X_k \\ =:& S_0, \end{aligned}$$

where the convergence step (almost surely as $n \to \infty$) is based on knowledge about the topological structure of the infinite-dimensional simplex Δ (e.g. see Hirth [12, §2, Lemma 1]) and the triple-use of the notation S_0, which we introduced by (1.2) and (2.3)–(2.4), is intentional. We already know by Theorem 1 that S_0 has the beta $(\alpha\theta, (1-\alpha)\theta)$ distribution. Since the limit behaviour of the sub-urns is independent of the times when $\mathscr{U}_0$ and $\mathscr{U}_1$ are served, S_0 is stochastically independent of $((X_{0,1}, X_{0,2}, X_{0,3}, \ldots), (X_{1,1}, X_{1,2}, X_{1,3}, \ldots))$.

Set

$$L(k) := \sum_{m=1}^{k} G_m,$$

i.e. if the allele originally labelled k (k'th oldest within the whole of $\mathscr{U}$) is in sub-urn $\mathscr{U}_1$, then it is $(L(k))$'th oldest within $\mathscr{U}_1$; and if it is in $\mathscr{U}_0$, then it is $(k - L(k))$'th oldest there. At one stage near the end of the proof we will need the generalised inverse of the random mapping L. We will denote it by L^{-1}, and for $l \in \mathbb{N}$ we will understand by $L^{-1}(l)$ the original $\mathscr{U}$-label of the l'th oldest allele in sub-urn $\mathscr{U}_1$. Defining this formally, $L^{-1}(l)$ is the smallest $k \in \mathbb{N}$ with the property $L(k) = l$. We now exploit the identity

$$X_{\mathscr{U},k}(n) = S_0(n) \cdot X_{0,k-L(k)}(n) \cdot (1 - G_k) + (1 - S_0(n)) \cdot X_{1,L(k)}(n) \cdot G_k,$$

https://doi.org/10.1017/S0963548396002842 Published online by Cambridge University Press

taking the almost sure limit as $n \to \infty$:

$$X_k = S_0 \cdot X_{0,k-L(k)} \cdot (1 - G_k) + (1 - S_0) \cdot X_{1,L(k)} \cdot G_k.$$

Let $\delta_{k,m}$, $k \in \mathbb{N}$, $m \in \{1, \ldots, d\}$, be $\{0,1\}$-valued random variables, independent of everything else, with

$$\sum_{m=1}^{d} \delta_{k,m} = 1,$$

$$P[\delta_{k,m} = 1] \quad = \frac{\alpha_m}{1 - \alpha},$$

$$0 < \alpha_m < 1, \ \sum_{m=0}^{d} \alpha_m = 1, \ \alpha_0 \equiv \alpha, \ d \geq 2,$$

and the sequence of random vectors $((\delta_{k,1}, \ldots, \delta_{k,d}))_{k \in \mathbb{N}}$ independent. Set

$$\beta_{k,0} = 1 - G_k$$

and

$$\beta_{k,m} = G_k \cdot \delta_{k,m}$$

for $m \in \{1, \ldots, d\}$. Then we have

$$\sum_{m=0}^{d} \beta_{k,m} = 1, \quad P[\beta_{k,m} = 1] \ = \alpha_m$$

for $m \in \{0, 1, \ldots, d\}$, and the sequence of random vectors $((\beta_{k,0}, \ldots, \beta_{k,d}))_{k \in \mathbb{N}}$ is independent in itself and independent of the urn process $\mathscr{U}$, in particular independent of the limiting GEM vector X of $\mathscr{U}$. Again, the coincidence of the notation $\beta_{k,m}$ with (2.1)–(2.2) is intentional. We are interested in

$$\begin{aligned} S &= (S_0, \ldots, S_d) \\ &= \left(\sum_{k=1}^{\infty} \beta_{k,0} X_k, \ldots, \sum_{k=1}^{\infty} \beta_{k,d} X_k \right) \end{aligned}$$

via

$$\begin{aligned} S' : &= \left(\frac{S_1}{1 - S_0}, \ldots, \frac{S_d}{1 - S_0} \right) \\ &= \frac{1}{1 - S_0} \cdot \left(\sum_{k=1}^{\infty} G_k \cdot \delta_{k,1} \cdot X_k, \ldots, \sum_{k=1}^{\infty} G_k \cdot \delta_{k,d} \cdot X_k \right), \end{aligned}$$

where

$$\begin{aligned} \frac{1}{1 - S_0} \cdot \sum_{k=1}^{\infty} G_k \cdot \delta_{k,m} \cdot X_k &= \sum_{k=1}^{\infty} G_k \cdot \delta_{k,m} \cdot X_{1,L(k)} \\ &= \sum_{l=1}^{\infty} \delta_{L^{-1}(k),m} \cdot X_{1,l} \end{aligned} \tag{3.4}$$

https://doi.org/10.1017/S0963548396002842 Published online by Cambridge University Press

has the same distribution as

$$\sum_{l=1}^{\infty} \delta_{l,m} \cdot X_{1,l}$$

whose joint ($m \in \{1, \ldots, d\}$) distribution is, by induction hypothesis,

$$\begin{aligned} &\text{Dirichlet } \left(\frac{\alpha_1}{1-\alpha}(1-\alpha)\theta, \ldots, \frac{\alpha_d}{1-\alpha}(1-\alpha)\theta\right) \\ &= \text{Dirichlet } (\alpha_1\theta, \ldots, \alpha_d\theta); \end{aligned}$$

and S' appears by (3.4) to be independent of S_0. This observation terminates the proof of Theorem 2. □

4. Applications via weak convergence to the GEM/Poisson–Dirichlet law

Our result (Theorem 2) lends itself to immediate applications in the various contexts in which the GEM/Poisson–Dirichlet law occurs as limit distribution. The common feature, which we assume throughout this section, is convergence in distribution of some sequence of random vectors $X^{(n)} = (X_1^{(n)}, X_2^{(n)}, \ldots)$, $n \in \mathbb{N}$ with values (almost surely) in the infinite-dimensional simplex

$$\Delta := \left\{ x = (x_1, x_2, \ldots) \in [0,1]^{\infty} : \sum_{k=1}^{\infty} x_k = 1 \right\} \tag{4.1}$$

to the GEM random vector X in the sense of weak convergence of probability measures on Δ. Donnelly and Joyce [8] provide precise statements on the relationship between weak convergence to the GEM law and weak convergence to the Poisson–Dirichlet law. Via order statistic and size-biasing, both the GEM and the Poisson–Dirichlet laws are essentially one and the same; we call it GEM/Poisson–Dirichlet law. Most of the applications of this law are invariant under (random) permutations of the components, or speaking in terms of population genetics, invariant under relabelling of the infinitely many neutral alleles. Our Theorems 1 and 2 are no exception. We will not explain this in more detail here again, but refer instead to Hirth [12]. Here we only work with the GEM law, but Theorems 1 and 2 are equally valid under the Poisson–Dirichlet law, and what we are now going to mention works analogously under the assumption of a weak convergence to the Poisson–Dirichlet law.

By analogy to (2.3) and (2.4) we define

$$S^{(n)} := (S_0^{(n)}, \ldots, S_d^{(n)}), \ n \in \mathbb{N} \tag{4.2}$$

with

$$S_m^{(n)} := \sum_{k=1}^{\infty} \beta_{k,m} X_k^{(n)} \tag{4.3}$$

and $\beta_{k,m}$ as in §2, independent of the set of all the random variables $X_k^{(n)}$ as well as independent of X.

The result is of course the convergence in distribution of $S^{(n)}$ to S, which is by Theorem 2 weak convergence of probability measures on the d-dimensional simplex Δ_d to

https://doi.org/10.1017/S0963548396002842 Published online by Cambridge University Press

the Dirichlet distribution with parameters $(\alpha_0\theta, \ldots, \alpha_d\theta)$. We omit the proof here because it is essentially the same as in Hirth [12, §3]; the ingredients are knowledge of the topology of Δ, the continuous mapping theorem for weak convergence and Lebesgue's dominated convergence theorem.

Acknowledgements

The author wishes to thank his current PhD supervisor Professor N. H. Bingham for his excellent guidance, his parents and the College for financial support and the Department for its supportive research environment. The author's tuition fees are covered by an EPSRC 'EC-Fees Only' award.

References

[1] Aldous, D. (1985) Exchangeability and related topics. In: *Ecole d'Eté de Probabilités de Saint-Flour XIII: Lecture Notes in Mathematics 1117* 1–198. Springer-Verlag.
[2] Connor, R. J. and Mosimann, J. E. (1969) Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Am. Statist. Assoc.* **64** 194–206.
[3] Darroch, J. N. and Ratcliff, D. (1971) A characterization of the Dirichlet distribution. *J. Am. Statist. Assoc.* **66** 641–643.
[4] Deshouillers J.-M., Dress, F. and Tenenbaum, G. (1979) Lois de répartition des diviseurs,1. *Acta Arith.* **23** 273–285.
[5] Donnelly, P. (1991) The heaps process, libraries, and size-biased permutations. *J. Appl. Prob.* **28** 321–335.
[6] Donnelly P. J., Ewens W. J. and Padmadisastra S. (1991) Functionals of random mappings: exact and asymptotic results. *Adv. Appl. Prob.* **23** 437–455.
[7] Donnelly, P. and Grimmett, G. (1993) On the asymptotic distribution of large prime factors. *J. London Math. Soc. (2)* **47** 395–404.
[8] Donnelly, P. and Joyce, P. (1989) Continuity and weak convergence of ranked and size-biased permutations on the infinite simplex. *Stoch. Processes Appl.* **31** 89–103.
[9] Donnelly, P. and Tavaré, S. (1987) The population genealogy of the infinitely-many neutral alleles model. *J. Math. Biol.* **25** 381–391.
[10] Engen, S. (1975) A note on the geometric series as a species frequency model. *Biometrika* **62** 697–699.
[11] Ewens, W.J. (1990) Population genetics theory – the past and the future. In: Lessard, S. (ed.), *Mathematical and statistical developments of evolutionary theory*. Kluwer Academic, pp. 177–227.
[12] Hirth, U. M. (to appear) Probabilistic number theory, GEM/Poisson–Dirichlet distribution and arc-sine law. *Combinatorics, Probability and Computing*.
[13] Hoppe, F. M. (1984) Pólya-like urns and the Ewens' sampling formula. *J. Math. Biol.* **20** 91–94.
[14] Hoppe, F.M. (1986) Size-biased filtering of Poisson–Dirichlet samples with an application to partition structures in genetics. *J. Appl. Prob.* **23** 1008–1012.
[15] Hoppe, F. M. (1987) The sampling theory of neutral alleles and an urn model in population genetics. *J. Math. Biol.* **25** 123–159.
[16] Ignatov, T. (1982) On a constant arising in the asymptotic theory of symmetric groups, and on Poisson–Dirichlet measures. *Th. Prob. Appl.* **27**(1) 136–147.
[17] Johnson, N. L. and Kotz, S. (1972) *Distributions in Statistics: Continuous multivariate distributions.* Wiley.
[18] Joyce, P. (1991) Estimating the frequency of the oldest allele: a Bayesian approach. *Adv. Appl. Prob.* **23** 456–475.

https://doi.org/10.1017/S0963548396002842 Published online by Cambridge University Press

[19] Kingman, J. F. C. (1975) Random discrete distributions. *J. Roy. Stat. Soc. B* **37**(1) 1–22.
[20] Kingman, J. F. C. (1977) The population structure associated with the Ewens sampling formula. *Theor. Popn. Biol.* **11** 274–283.
[21] Kingman, J. F. C. (1993) *Poisson Processes*. Clarendon Press.
[22] Knuth, D. E. and Trabb Pardo, L. (1976) Analysis of a simple factorization algorithm. *Theor. Comp. Sci.* **3** 321–348.
[23] Lloyd, C. J. (1989) Some further results on random splitting of an interval. *J. Appl. Prob.* **26** 466–476.
[24] Lloyd, C. J. and Williams, E. J. (1988) Recursive splitting of an interval when the proportions are identical and independent random variables. *Stoch. Processes Appl.* **27** 111–122.
[25] Lloyd, S. P. (1984) Ordered prime divisors of a random integer. *Ann. Prob.* **12** (4) 1205–1212.
[26] Möhle, M. (1988) *Der Coalescent*. Diplomarbeit, Fachbereich Mathematik, Johannes-Gutenberg-Universität Mainz.
[27] Notter, G. (1988) *Anwendungen eines Polya-Urnenmodells in der Populationsgenetik*. Diplomarbeit, Fachbereich Mathematik, Johannes-Gutenberg-Universität Mainz.
[28] Shepp, L. A. and Lloyd, S. P. (1966) Ordered cycle lengths in a random permutation. *Trans. Amer. Math. Soc.* **121** 340–357.
[29] Tenenbaum, G. (1990) Introduction à la théorie analytique et probabiliste des nombres. Université de Nancy 1, Institut Elie Cartan No. **13**.
[30] Vershik, A. M. (1987) The asymptotic distribution of factorizations of natural numbers into prime divisors. *Soviet Math. Dokl.* **34**(1) 57–61.
[31] Vershik, A. M. and Shmidt, A. A. (1977) Limit measures arising in the asymptotic theory of symmetric groups.I. *Th. Prob. Appl.* **22**(1) 70–85.
[32] Watterson, G. A. (1976) The stationary distribution of the infinitely-many neutral alleles diffusion model. *J. Appl. Prob.* **13** 639–651.

https://doi.org/10.1017/S0963548396002842 Published online by Cambridge University Press