Combinatorics, Probability and Computing (1999) 8, 193–208. Printed in the United Kingdom
© 1999 Cambridge University Press

On Poisson–Dirichlet Limits for Random Decomposable Combinatorial Structures

RICHARD ARRATIA[1]†, A.D. BARBOUR[2]‡
and SIMON TAVARÉ[1]†

[1] Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113, USA
(e-mail: rarratia@math.usc.edu stavare@gnome.usc.edu)

[2] Abteilung für Angewandte Mathematik, Universität Zürich, Winterthurerstrasse 190, CH-8057, Zürich, Switzerland
(e-mail: adb@amath.unizh.ch)

Received 20 June 1997; revised 16 March 1998

We prove a joint local limit law for the distribution of the r largest components of decomposable logarithmic combinatorial structures, including assemblies, multisets and selections. Our method is entirely probabilistic, and requires only weak conditions that may readily be verified in practice.

1. Introduction

The proportion of integers in the largest, second largest, … cycles of a random permutation of n objects have, asymptotically as $n \to \infty$, the Poisson–Dirichlet distribution PD(θ), with parameter $\theta = 1$. The result for the largest cycle appears in Goncharov [8], that for the kth largest in Shepp and Lloyd [17], and the joint distributional result in Kingman [12] and Vershik and Shmidt [19]. Related limit laws are now known for a variety of decomposable combinatorial structures. For example, the relative sizes of the largest, second largest, … components of a random mapping have asymptotically the PD(θ) law with $\theta = 1/2$. Convergence for the marginal distributions appears in Kolchin [14, 15], and for the joint distribution in Aldous [1]. An analogous result holds for the ordered degree sequence of the factors of a polynomial over $GF(q)$; in this case $\theta = 1$, just as for permutations [3].

† Supported in part by NSF grant DMS 90-05833 and DMS 96-26412.
‡ Supported in part by Schweizerischer NF Projekt Nr 20-43453.95.

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press

More recently, Hansen [9] provided a unified approach to the asymptotics of the order statistics of the component sizes of multisets and assemblies, which are families of decomposable combinatorial structures. She establishes weak convergence to the PD(θ) distribution, and shows how to identify the appropriate value of θ. Her arguments rely in part on complex analysis, and involve conditions on generating functions which are not always easy to verify.

In this paper we establish a local limit theorem for the joint distribution of the large components, refining previous results, under very weak conditions. Our method uses probabilistic, as opposed to complex analytic, arguments, and the conditions become correspondingly transparent. We consider randomly chosen decomposable combinatorial structures of total size n whose component counts $(C_1(n), \ldots, C_n(n))$, where $C_i(n)$ denotes the number of components of size i, have joint distribution determined by the conditioning relation

$$\text{(CR)} \qquad \mathscr{L}(C_1(n), \ldots, C_n(n)) = \mathscr{L}(Z_1, \ldots, Z_n \mid T_n = n), \tag{1.1}$$

where $(Z_i, i \geqslant 1)$ are independent random variables over $\mathbb{Z}_+$, and $T_n = \sum_{i=1}^{n} iZ_i$. Such structures include assemblies (for which the Z_i are Poisson-distributed), multisets (for which the Z_i are negative binomially distributed), and selections (for which the Z_i are binomially distributed): see [5]. We require only that the Z_i satisfy the logarithmic condition LC:

$$\text{(LC)} \qquad \lim_{i \to \infty} i\mathbb{P}(Z_i = 1) = \theta = \lim_{i \to \infty} i\mathbb{E}Z_i \tag{1.2}$$

for some $\theta > 0$, together with the additional tail condition

$$\mathbb{P}(Z_i \geqslant 2) = O(i^{-2}), \tag{1.3}$$

conditions weaker than those in [9]. In fact, (1.3) is implied by (1.2) for the examples we consider in Section 5. More general structures can also be analysed, under the mild extra condition given in (5.14), since the approach also ties in with arguments using Stein's method, discussed in detail in [4].

The simplest example of such a construct is that in which the Z_i are Poisson-distributed with means *exactly* θ/i. In this case, the $C_i(n)$ have as joint distribution the Ewens Sampling Formula $\mathrm{ESF}_n(\theta)$ given in (2.1). In Section 2, we collect known facts about PD(θ), $\mathrm{ESF}_n(\theta)$, and the limit distribution P_θ of $n^{-1}T_n$ under $\mathrm{ESF}_n(\theta)$. The local limit approximation to the joint distribution of the sizes of the large cycles by PD(θ), when the $C_i(n)$ are distributed according to $\mathrm{ESF}_n(\theta)$, is then established in Section 3. The argument used is readily generalized in Section 4 to arbitrary combinatorial structures that satisfy (LC), and for which T_n/n admits the local limit approximation (LLA) given in (4.6). In Section 5 it is shown that (LLA) holds for assemblies, multisets and selections satisfying (LC).

2. The Ewens Sampling Formula

Let $\mathscr{S}_n$ denote the set of permutations of $\{1, 2, \ldots, n\}$. We write $\pi \in \mathscr{S}_n$ as an ordered product of cycles. The integer 1 starts the first cycle, followed by the image of 1, the image of that point and so on. The smallest integer not in the first cycle begins the second cycle, followed by its images. In this way, π is decomposed into an ordered product of cycles.

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press

We consider random permutations with distribution determined by

$$\mathbb{P}(\pi) = \frac{\theta^{|\pi|}}{\theta_{(n)}}, \pi \in \mathscr{S}_n,$$

where $|\pi|$ denotes the number of cycles in π, $\theta \in (0, \infty)$, and

$$\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1).$$

Let $C^{(n)} \equiv (C_1(n), C_2(n), \ldots, C_n(n))$ be the counts of cycles of sizes $1, 2, \ldots, n$ in such a θ-biased random permutation of size n. The distribution of $C^{(n)}$ is given by the Ewens Sampling Formula [7] $\mathrm{ESF}_n(\theta)$: for any $a \in \mathbb{Z}_+^n$,

$$\mathbb{P}(C^{(n)} = a) = \mathbb{1}\left(\sum_{i=1}^n i a_i = n\right) \frac{n!}{\theta_{(n)}} \prod_{j=1}^n \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!}, \tag{2.1}$$

$\mathbb{1}(A)$ denoting the indicator of A. Let $Z_1, Z_2, \ldots$ be independent Poisson random variables with means $\mathbb{E}Z_i = \theta/i, i = 1, 2, \ldots, n$, and let $Z[1,n] = (Z_1, \ldots, Z_n)$. It is well known [21, 2] that

$$\mathbb{P}(C^{(n)} = a) = \mathbb{P}(Z[1,n] = a | T_n = n), \tag{2.2}$$

where

$$T_n = Z_1 + 2Z_2 + \cdots + nZ_n,$$

so we have a combinatorial structure satisfying (1.1), with Z_i having a Poisson distribution with mean θ/i, for which the conditions (LC) and (1.3) are clearly satisfied. Understanding the asymptotic behaviour of (2.2) requires knowledge of the asymptotic behaviour of T_n, which we now review.

2.1. The limit T of T_n/n

The density of T_n satisfies the recursion

$$k\mathbb{P}(T_n = k) = \theta \sum_{j=1}^n \mathbb{P}(T_n = k-j),\ k = 1, 2, \ldots. \tag{2.3}$$

It follows from this that

$$k\mathbb{P}(T_n = k) = (k-1+\theta)\mathbb{P}(T_n = k-1),\ k = 1, 2, \ldots, n,$$

so that

$$\mathbb{P}(T_n = k) = \frac{\theta_{(k)}}{k!}\mathbb{P}(T_n = 0) = \exp(-\theta h(n))\frac{\theta_{(k)}}{k!},\quad k \leqslant n,$$

where $h(n) = \sum_{j=1}^n 1/j$. Hence

$$\lim_{n\to\infty} n\mathbb{P}(T_n = k) = \frac{e^{-\gamma\theta}x^{\theta-1}}{\Gamma(\theta)} \quad \text{if } k \leqslant n,\ k/n \to x \in (0,1]. \tag{2.4}$$

Using (2.3) and (2.4), we conclude that

$$\lim_{n\to\infty} \mathbb{P}(T_n/n \leqslant x) = \frac{x^\theta e^{-\gamma\theta}}{\Gamma(\theta+1)} \tag{2.5}$$

if $x \in (0,1]$. In fact, $\lim_{n\to\infty} \mathbb{P}(T_n/n \leqslant x)$ exists for all $x > 0$.

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press

Theorem 2.1. *As $n \to \infty$, the random variable T_n/n converges in distribution to a random variable T whose distribution P_θ has Laplace transform given by*

$$\mathbb{E}e^{-sT} = \exp\left(-\int_0^1 (1-e^{-sx})\frac{\theta}{x}dx\right). \tag{2.6}$$

Proof. Let μ_n be the measure that puts mass n^{-1} at points $in^{-1}, i = 1,2,\ldots,n$ and note that μ_n converges weakly to Lebesgue measure. The Laplace transform of the random variable T_n/n is

$$\begin{aligned}\mathbb{E}e^{-sT_n/n} &= \exp\left(-\sum_{i=1}^{n}\frac{\theta}{i}(1-e^{-si/n})\right)\\ &= \exp\left(-\int_0^1 (1-e^{-sx})\frac{\theta}{x}\mu_n(dx)\right)\\ &\to \exp\left(-\int_0^1 (1-e^{-sx})\frac{\theta}{x}dx\right),\end{aligned}$$

the last step following by dominated convergence. □

It follows from (2.5) that the density g_θ of T satisfies

$$g_\theta(x) = \frac{e^{-\gamma\theta}x^{\theta-1}}{\Gamma(\theta)},\ 0 \leqslant x \leqslant 1,$$

so that

$$g_\theta(1) = \frac{e^{-\gamma\theta}}{\Gamma(\theta)}. \tag{2.7}$$

An expression for $g_\theta(x)$ for $x > 1$ is given in [20]; it satisfies the integral equation

$$xg_\theta(x) = \theta\int_{x-1}^{x} g_\theta(u)du, \quad x > 0, \tag{2.8}$$

with $g_\theta(x) = 0$ if $x < 0$.

2.2. The Poisson–Dirichlet distribution

The Poisson–Dirichlet distribution, denoted by PD(θ), was defined by Kingman [11] to be the distribution of the normalized points $\sigma_1 > \sigma_2 > \cdots$ of a Poisson process with intensity $\theta e^{-x}/x, x > 0$:

$$\mathrm{PD}(\theta) = \mathscr{L}((\sigma_1/\sigma, \sigma_2/\sigma, \ldots)), \tag{2.9}$$

where $\sigma = \sigma_1 + \sigma_2 + \cdots$. Other representations of the Poisson–Dirichlet distribution may be found in [13, 4], for example.

The density $f_\theta^{(r)}$ of the first r coordinates of PD(θ) was found by Watterson [22] in the form

$$f_\theta^{(r)}(x_1,\ldots,x_r) = \frac{e^{\gamma\theta}\theta^r\Gamma(\theta)x_r^{\theta-1}}{x_1x_2\cdots x_r}g_\theta\left(\frac{1-x_1-\cdots-x_r}{x_r}\right), \tag{2.10}$$

for $r \geqslant 1$, and $x_1,\ldots,x_r$ satisfying $0 < x_r < \cdots < x_1 < 1$ and $0 < x_1 + \cdots + x_r < 1$.

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press

The Poisson–Dirichlet distribution with parameter θ arises as the limit law of the renormalized sizes $L_1(n), L_2(n), \ldots$ of the largest, second largest, $\ldots$ cycles of a θ-biased permutation.

Theorem 2.2 ([12]). *As $n \to \infty$,*

$$n^{-1}(L_1(n), L_2(n), \ldots) \Rightarrow (L_1, L_2, \ldots),$$

where $(L_1, L_2, \ldots)$ has the $PD(\theta)$ distribution.

3. A local limit law for large cycles under $\mathrm{ESF}_n(\theta)$

3.1. Point probabilities for T_n

Theorem 2.1 extends the convergence in (2.5) from $x \in (0, 1]$ to all $x > 0$. This suggests that the same extension may also be feasible for (2.4), as is shown in the following lemma.

Lemma 3.1. *Suppose that $m/n \to y \in (0, \infty)$ as $n \to \infty$. Then*

$$\lim_{n\to\infty} n\mathbb{P}(T_n = m) = g_\theta(y). \tag{3.1}$$

Proof. Equation (2.3) gives

$$m\mathbb{P}(T_n = m) = \theta\mathbb{P}(m - n \leqslant T_n < m).$$

Multiplying by n/m and using (2.8) and the fact that $T_n/n \Rightarrow T$, which has a continuous distribution function, shows that

$$\begin{aligned} \lim_{n\to\infty} n\mathbb{P}(T_n = m) &= \frac{\theta}{y}\mathbb{P}(y - 1 \leqslant T \leqslant y) \\ &= g_\theta(y), \end{aligned}$$

completing the proof. □

The next result uses elementary arguments to derive bounds for the point probabilities $\mathbb{P}(T_{bn} = m)$, where

$$T_{bn} = \sum_{j=b+1}^{n} jZ_j, \ 0 \leqslant b < n.$$

Lemma 3.2. *Write $\bar{\theta} = \min(1, \theta)$. Then*

$$\max_{k\geqslant 0} \mathbb{P}(T_{bn} = k) \leqslant e^{-\bar{\theta}(h(n)-h(b))}. \tag{3.2}$$

Proof. First consider the case $0 < \theta \leqslant 1$. We use the fact ([5]) that the point probabilities for T_{bn} satisfy

$$m\mathbb{P}(T_{bn} = m) = \theta\mathbb{P}(m - n \leqslant T_{bn} < m - b). \tag{3.3}$$

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press

Hence, for $m \geqslant 1$,

$$\mathbb{P}(T_{bn} = m) \leqslant \frac{1}{m} \sum_{j=0}^{m-1} \mathbb{P}(T_{bn} = j).$$

Thus $\mathbb{P}(T_{bn} = m)$ is at most the average of the previous m values, and so, by induction, $\max_{k \geqslant 0} \mathbb{P}(T_{bn} = k) \leqslant \mathbb{P}(T_{bn} = 0) = e^{-\theta(h(n)-h(b))}$.

For the case $\theta > 1$, let $\tilde{Z}_j, j \geqslant 1$ be independent Poisson random variables with $\mathbb{E}\tilde{Z}_j = 1/j$, and define $\tilde{T}_{bn} = \sum_{j=b+1}^{n} j\tilde{Z}_j$. Define $T'_{bn} = \sum_{j=b+1}^{n} jZ'_j$, where the Z'_j are independent Poisson random variables with mean $(\theta - 1)/j$, independent of the $\tilde{Z}_j$. Then we can write

$$T_{bn} = \tilde{T}_{bn} + T'_{bn},$$

with independent summands. It follows that

$$\begin{aligned} \mathbb{P}(T_{bn} = m) &= \sum_{j=0}^{m} \mathbb{P}(\tilde{T}_{bn} = j)\mathbb{P}(T'_{bn} = m - j) \\ &\leqslant \max_{0 \leqslant j \leqslant m} \mathbb{P}(\tilde{T}_{bn} = j) \\ &\leqslant e^{-(h(n)-h(b))}, \end{aligned}$$

the last step following from the case proved earlier. □

3.2. The local limit theorem

In this section we derive a joint local limit law for the distribution of the r largest cycle lengths $L_1(n), \ldots, L_r(n)$ under $\mathrm{ESF}_n(\theta)$.

Theorem 3.3. *For $r \geqslant 1$, suppose that $0 < x_r < x_{r-1} < \cdots < x_1 < 1$ satisfy $0 < x_1 + \cdots + x_r < 1$. Then*

$$\lim_{n \to \infty} n^r \mathbb{P}(L_i(n) = \lfloor nx_i \rfloor, 1 \leqslant i \leqslant r) = f_\theta^{(r)}(x_1, \ldots, x_r), \tag{3.4}$$

where the density $f_\theta^{(r)}$ is given in (2.10).

Proof. First assume that integers $m_1, m_2, \ldots, m_r$ satisfy the conditions

$$1 \leqslant m_r < m_{r-1} < \cdots < m_1 < n, \ m \equiv m_1 + \cdots + m_r \leqslant n,$$

and let $A_n(C^{(n)}) = A_n(C^{(n)}; m_1, m_2, \ldots, m_{r-1}, m_r)$ denote the event

$$\begin{aligned} \Big\{ & C_n(n) = 0, \ldots, C_{m_1+1}(n) = 0, C_{m_1}(n) = 1, C_{m_1-1}(n) = 0, \ldots, \\ & C_{m_2+1}(n) = 0, C_{m_2}(n) = 1, C_{m_2-1}(n) = 0, \ldots, C_{m_{r-1}+1}(n) = 0, \\ & C_{m_{r-1}}(n) = 1, C_{m_{r-1}-1}(n) = 0, \ldots, C_{m_r+1}(n) = 0 \Big\}. \end{aligned}$$

Then

$$\mathbb{P}(L_1(n) = m_1, \ldots, L_r(n) = m_r) = \mathbb{P}(A_n(C^{(n)}), C_{m_r}(n) \geqslant 1).$$

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press

This last probability can be written

$$\mathbb{P}(A_n(C^{(n)}), C_{m_r}(n) = 1) + \sum_{l \geqslant 2} \mathbb{P}(A_n(C^{(n)}), C_{m_r}(n) = l).$$

The first term is, using (2.2),

$$\mathbb{P}(A_n(Z), Z_{m_r} = 1 | T_n = n) = \mathbb{P}(A_n(Z))\mathbb{P}(Z_{m_r} = 1)\frac{\mathbb{P}(T_{m_r - 1} = n - m)}{\mathbb{P}(T_n = n)}, \tag{3.5}$$

which reduces to

$$\frac{\mathbb{P}(T_{m_r - 1} = n - m)}{\mathbb{P}(T_n = n)} \frac{\theta^r e^{-\theta(h(n) - h(m_r - 1))}}{m_1 \cdots m_r} \tag{3.6}$$

Applying the result of Lemma 3.1 and simplifying shows that

$$\lim_{n \to \infty} n^r \mathbb{P}(A_n(C^{(n)}), C_{m_r}(n) = 1) = f_\theta^{(r)}(x_1, \ldots, x_r).$$

It remains to show that $\sum_{l \geqslant 2} \mathbb{P}(A_n(C^{(n)}), C_{m_r}(n) = l) = o(n^{-r})$. But this probability is just

$$\begin{aligned} \mathbb{P}(A_n(Z)) \sum_{l \geqslant 2} \mathbb{P}(Z_{m_r} = l) \frac{\mathbb{P}(T_{m_r - 1} = n - m - (l - 1)m_r)}{\mathbb{P}(T_n = n)} \\ \leqslant \quad \mathbb{P}(A_n(Z)) \frac{e^{-\bar{\theta} h(m_r - 1)}}{\mathbb{P}(T_n = n)} \mathbb{P}(Z_{m_r} \geqslant 2), \end{aligned}$$

using Lemma 3.2. Since $\mathbb{P}(T_n = n) \sim n^{-1} g_\theta(1)$, $\mathbb{P}(A_n(Z)) \leqslant \theta^{r-1}/(m_1 \cdots m_{r-1})$, and $\mathbb{P}(Z_{m_r} \geqslant 2) \leqslant \theta^2/(2m_r^2)$, we see that this term is of order $O(n^{-r-1} \cdot n \cdot n^{-\bar{\theta}}) = O(n^{-r} n^{-\bar{\theta}}) = o(n^{-r})$, as required. □

Remarks.

(1) Theorem 2.2 follows from Theorem 3.3 using Scheffé's Theorem [16].

(2) It is crucial in the hypothesis $x_1 + \cdots + x_r < 1$ to have strict inequality. To see this, take $r = 1$ and $x_1 = 1$, and note that

$$n\mathbb{P}(L_1(n) = n) \sim \Gamma(\theta + 1) n^{1-\theta}.$$

4. Combinatorial structures

In this section, we show that a joint local limit law like that in Theorem 3.3 is true for a large class of decomposable combinatorial structures. $C_i(n)$ now denotes the number of components of size i, $i = 1, 2, \ldots, n$, and we consider structures that satisfy the relation (2.2) for independent random variables Z_i taking values in $\mathbb{Z}_+$. However, the Z_i no longer satisfy $Z_i \sim \mathrm{Po}(\theta/i)$; instead, we merely require the 'logarithmic condition' (LC), repeated here for convenience:

$$\lim_{i \to \infty} i\mathbb{P}(Z_i = 1) = \theta = \lim_{i \to \infty} i\mathbb{E}Z_i \tag{LC}$$

for some $\theta \in (0, \infty)$, and the tail condition (1.3). Note that (LC) implies that

$$\tilde{\theta} \equiv \sup_{i \geqslant 1} i\mathbb{E}Z_i < \infty. \tag{4.1}$$

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press

We begin with some preliminaries, the first of which requires no proof.

Lemma 4.1. *Suppose the Z_i satisfy* (LC). *Then as $i \to \infty$,*

$$\mathbb{P}(Z_i \geqslant 2) = o(i^{-1}), \tag{4.2}$$

and

$$\mathbb{P}(Z_i = 0) = 1 - \theta i^{-1} + o(i^{-1}). \tag{4.3}$$

We see from this that, for large i, the distribution of Z_i is indeed close to Poisson with mean θ/i.

Corollary 4.2. *Let Z_i^* be independent Poisson random variables with $\mathbb{E}Z_i^* = \theta/i$, $i \geqslant 1$. There is a sequence $\epsilon(i) \downarrow 0$ as $i \to \infty$ such that*

$$d_{TV}(Z_i, Z_i^*) \leqslant \epsilon(i) i^{-1}.$$

Proof.

$$\begin{aligned} 2d_{TV}(Z_i, Z_i^*) &= \sum_{j \geqslant 0} |\mathbb{P}(Z_i = j) - \mathbb{P}(Z_i^* = j)| \\ &\leqslant |\mathbb{P}(Z_i = 0) - \mathbb{P}(Z_i^* = 0)| + |\mathbb{P}(Z_i = 1) - \mathbb{P}(Z_i^* = 1)| \\ &\quad + \mathbb{P}(Z_i \geqslant 2) + \mathbb{P}(Z_i^* \geqslant 2). \end{aligned}$$

The result now follows from Lemma 4.1. □

In order to prove a local limit theorem for the r largest component sizes $L_1(n), \ldots, L_r(n)$, analogous to Theorem 3.3, we use the same recipe. The first ingredient is the counterpart of Theorem 2.1; an alternative proof may be found in [5].

Theorem 4.3. *For $i = 1, 2, \ldots$, let Z_i be independent random variables taking values in $\mathbb{Z}_+$ and satisfying* (LC). *Then, as $n \to \infty$,*

$$n^{-1} T_n \Rightarrow T. \tag{4.4}$$

Proof. Let Z_i^* be independent Poisson random variables with $\mathbb{E}Z_i^* = \theta/i$, and write $T_{bn}^* = \sum_{j=b+1}^n j Z_j^*$, $Z^*(b, n] = (Z_{b+1}^*, \ldots, Z_n^*)$. Corollary 4.2 shows that $d_{TV}(Z_i, Z_i^*) \leqslant \epsilon(i) i^{-1}$. Choose any sequence $b_n = o(n)$ such that $\epsilon(b_n) \log(n/b_n) \to 0$ as $n \to \infty$. Then we immediately find that

$$\begin{aligned} d_{TV}(T_{b_n n}, T_{b_n n}^*) &\leqslant d_{TV}(Z(b_n, n], Z^*(b_n, n]) \\ &\leqslant \sum_{j=b_n+1}^{n} \epsilon(j) j^{-1} \leqslant \epsilon(b_n) \log(n/b_n). \end{aligned} \tag{4.5}$$

Since $\mathbb{E}n^{-1} T_{b_n}^* = n^{-1} \theta b_n \to 0$, it follows that $n^{-1} T_{b_n}^* \Rightarrow 0$. Hence, since $n^{-1} T_n^* = n^{-1} T_{bn}^* + n^{-1} T_{b_n n}^*$, $n^{-1} T_{b_n n}^* \Rightarrow T$ by Theorem 2.1, and it then follows from (4.5) that $n^{-1} T_{b_n n} \Rightarrow T$.

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press

Finally,

$$n^{-1}\mathbb{E}T_{b_n} = n^{-1}\sum_{j=1}^{b_n} j\mathbb{E}Z_j \leqslant \tilde{\theta}n^{-1}b_n \to 0,$$

so that $n^{-1}T_n \Rightarrow T$ also. □

The second ingredient is a bound on point probabilities, echoing Lemma 3.2.

Theorem 4.4. *As* $n \to \infty$, $\max_{k\geqslant 0}\mathbb{P}(T_n = k) \to 0$.

Proof. As in the proof of Theorem 4.3, let Z_i^* be independent Poisson random variables with $\mathbb{E}Z_i^* = \theta/i$, and choose any sequence $b_n = o(n)$ such that (4.5) holds. Since $T_n = T_{b_n} + T_{b_n n}$ and the two summands are independent,

$$\max_{k\geqslant 0}\mathbb{P}(T_n = k) \leqslant \max_{k\geqslant 0}\mathbb{P}(T_{b_n n} = k).$$

Now, from (4.5),

$$\mathbb{P}(T_{b_n n} = k) \leqslant \mathbb{P}(T^*_{b_n n} = k) + \epsilon(b_n)\log(n/b_n),$$

and by Lemma 3.2, defining $\bar{\theta} = \min(1, \theta)$,

$$\max_{k\geqslant 0}\mathbb{P}(T^*_{b_n n} = k) \leqslant e^{-\bar{\theta}(h(n)-h(b_n))} = O((b_n/n)^{\bar{\theta}}).$$

Hence $\max_{k\geqslant 0}\mathbb{P}(T_{b_n n} = k) \to 0$ as $n \to \infty$. □

The final ingredient is that T_n should satisfy a local limit approximation:

$$\text{(LLA)} \qquad n\mathbb{P}(T_n = m) \sim g_\theta(y), \quad \text{as } n \to \infty, m/n \to y \in (0, \infty). \tag{4.6}$$

The proofs of (LLA) are somewhat different for the various classes of combinatorial structure we consider, and a detailed treatment is given in the next section. Whenever (LC) and (LLA) hold, we have the following joint local limit law.

Theorem 4.5. *Suppose that a combinatorial structure* $C^{(n)}$ *has distribution given by* (CR)*, and satisfies* (LC)*,* (LLA)*, and the tail condition* (1.3)*. For* $r \geqslant 1$*, suppose that* $0 < x_r < x_{r-1} < \cdots < x_1 < 1$ *satisfy* $0 < x_1 + \cdots + x_r < 1$*. Then*

$$\lim_{n\to\infty} n^r\mathbb{P}(L_i(n) = \lfloor nx_i \rfloor, 1 \leqslant i \leqslant r) = f_\theta^{(r)}(x_1, \ldots, x_r),$$

where the density $f_\theta^{(r)}$ *is given in* (2.10)*; hence also, as* $n \to \infty$*,*

$$n^{-1}(L_1(n), L_2(n), \ldots) \Rightarrow \text{PD}(\theta).$$

Proof. The proof mimics that of Theorem 3.3 down to (3.5). Expression (3.6) is now replaced by

$$\frac{\mathbb{P}(T_{m_r-1} = n - m)}{\mathbb{P}(T_n = n)} \prod_{i=m_r}^{n}\mathbb{P}(Z_i = 0)\prod_{s=1}^{r}\frac{\mathbb{P}(Z_{m_s} = 1)}{\mathbb{P}(Z_{m_s} = 0)}.$$

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press

Under (LLA), the first term is asymptotic to $x_r^{-1} g_\theta((1-x_1-\cdots-x_r)/x_r)/g_\theta(1)$. Using (4.3), the first product is asymptotic to $x_r^{-\theta}$, while from (LC) the second product is asymptotic to $n^{-r}\theta^r x_1^{-1}\cdots x_r^{-1}$. Combining these terms and using (2.10) shows that

$$\lim_{n\to\infty} n^r \mathbb{P}(A_n(C^{(n)}), C_{m_r}(n) = 1) = f_\theta^{(r)}(x_1,\ldots,x_r).$$

To show that $\sum_{l\geqslant 2} \mathbb{P}(A_n(C^{(n)}), C_{m_r}(n) = l) = o(n^{-r})$, note that the left side is just

$$\begin{aligned}\mathbb{P}(A_n(Z)) \sum_{l\geqslant 2} \mathbb{P}(Z_{m_r} = l) \frac{\mathbb{P}(T_{m_r-1} = n - m - (l-1)m_r)}{\mathbb{P}(T_n = n)} \\ \leqslant \quad \frac{\mathbb{P}(A_n(Z))\mathbb{P}(Z_{m_r}\geqslant 2)}{\mathbb{P}(T_n = n)} \max_{k\geqslant 0} \mathbb{P}(T_{m_r-1} = k).\end{aligned}$$

Since

$$\mathbb{P}(A_n(Z)) \leqslant \mathbb{P}(Z_{m_1} = 1)\cdots\mathbb{P}(Z_{m_{r-1}} = 1) = O(n^{-(r-1)}),$$

and $\mathbb{P}(Z_{m_r}\geqslant 2) = O(n^{-2})$, we see from (LLA) that the first factor is of order n^{-r}, whereas the second term tends to 0 by Theorem 4.4. □

5. Verifying the local limit approximation

This section is devoted to establishing (LLA) for a wide variety of combinatorial models. Once done, Theorem 4.5 can then be applied.

5.1. Assemblies

Random assemblies are decomposable combinatorial structures for which the counts $C_j(n)$ of components of size j satisfy the conditioning relation (2.2) for Poisson-distributed Z_j with means

$$\mathbb{E}Z_j \equiv \lambda_j = \frac{m_j x^j}{j!}, \text{ for some } x > 0.$$

In these models, the integers m_j are prescribed in advance, and, for a satisfying $a_1 + 2a_2 + \cdots + na_n = n$, the probabilities

$$\begin{aligned}\mathbb{P}(C^{(n)} = a) &= \mathbb{P}(Z[1,n] = a | T_n = n) \\ &= \mathbb{P}(Z[1,n] = a)/\mathbb{P}(T_n = n) \\ &= \frac{\prod_{i=1}^n (m_i x^i/i!)^{a_i}/a_i!}{\sum_{\{d_1+2d_2+\cdots+nd_n=n\}} \prod_{i=1}^n (m_i x^i/i!)^{d_i}/d_i!}\end{aligned}$$

are the same for any arbitrary value of x. Hence, to be in the logarithmic class, it is enough that $m_j \sim \theta(j-1)!y^j$ for some $y > 0$, since we can take $x = y^{-1}$. Condition (LC) then reduces to the requirement that $j\lambda_j \to \theta$, in which case

$$\mathbb{P}(Z_j \geqslant 2) = [1 - e^{-\lambda_j}(1+\lambda_j)] \leqslant \frac{1}{2}\lambda_j^2 = O(j^{-2}),$$

so that the tail condition (1.3) is satisfied. Among the examples are permutations for which $m_j = (j-1)!, x = 1, \theta = 1$, and random mappings for which $m_j = (j-1)!\sum_{l=0}^{i-1} i^l/l!, x = e^{-1}, \theta = 1/2$. Many other examples are described in [5]. We note in passing that the

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press

following proofs make no use of the m_j being integers; an application in the more general setting appears in Section 5.4.

Most of the results depend on the analogue of equation (2.3) for the density of T_n. Using [5], it takes the form

$$k\mathbb{P}(T_n = k) = \sum_{j=1}^{n} \mathbb{P}(T_n = k - j)j\lambda_j,\ k = 0, 1, \ldots \tag{5.1}$$

with $j\lambda_j$ in the place of θ. Intuitively, this should make little difference for large n, because $j\lambda_j \to \theta$.

To verify (LLA), note that, according to equation (5.1),

$$\begin{aligned} k\mathbb{P}(T_n = k) &= \sum_{j=1}^{n} \mathbb{P}(T_n = k - j)j\lambda_j \\ &= \theta\mathbb{P}(k - n \leqslant T_n < k) + r_n(k), \end{aligned}$$

where

$$r_n(k) = \sum_{j=1}^{n} \mathbb{P}(T_n = k - j)(j\lambda_j - \theta).$$

The remainder of the proof follows just as for Lemma 3.1, but now using Theorem 4.3 instead of Theorem 2.1, if we can show that $|r_n(k)| \to 0$ as $n \to \infty$ when $k/n \to y > 0$. To do this, let $\epsilon > 0$ be arbitrary, and choose $j_0 = j_0(\epsilon)$ such that $|j\lambda_j - \theta| < \epsilon$ for all $j > j_0$. Then, for $n > j_0$,

$$\begin{aligned} |r_n(k)| &\leqslant \sum_{j=1}^{j_0} \mathbb{P}(T_n = k - j)|j\lambda_j - \theta| + \epsilon \sum_{j>j_0} \mathbb{P}(T_n = k - j) \\ &\leqslant \max_{j \geqslant 1} |j\lambda_j - \theta|\, \mathbb{P}(k - j_0 \leqslant T_n < k - 1) + \epsilon. \end{aligned}$$

Hence

$$\limsup_{n\to\infty} |r_n(k)| \leqslant \max_{j\geqslant 1} |j\lambda_j - \theta|\ \limsup_{n\to\infty}\ \sup_{x>0} \mathbb{P}(x - j_0/n \leqslant n^{-1}T_n < x) + \epsilon = \epsilon,$$

because T_n/n converges in distribution to T, which has continuous distribution function.

5.2. Multisets

For combinatorial multisets, the Z_i have negative binomial distributions $\mathrm{NB}(m_i, x^i)$, with

$$\mathbb{P}(Z_i = k) = \binom{m_i + k - 1}{k}(1 - x^i)^{m_i}x^{ik},\ k = 0, 1, \ldots,$$

for any $x \in (0, 1)$; once again, the integers m_i are prescribed in the structure, and the joint distribution of the component sizes is the same for any choice of x. We have

$$\mathbb{E}Z_i = \frac{m_i x^i}{1 - x^i},\quad \mathrm{Var}Z_i = \frac{m_i x^i}{(1 - x^i)^2};$$

in the logarithmic class are those structures for which

$$m_i \sim \frac{\theta y^i}{i},\quad \text{for some } y > 1,\ \theta \in (0, \infty),$$

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press

when we take $x = y^{-1}$, and record that then

$$\lim_{i\to\infty} i\mathbb{E}Z_i = \lim_{i\to\infty} im_i x^i = \theta. \tag{5.2}$$

Furthermore, the tail condition (1.3) is also satisfied, since

$$\mathbb{P}(Z_i \geqslant 2) = 1 - (1-x^i)^{m_i}[1 + m_i x^i] \leqslant \frac{m_i(m_i+1)x^{2i}}{2(1-x^i)^2} = O(i^{-2}).$$

We note once more that, in the proofs below, the m_i need not be integers: see the examples in Section 5.4.

The recursion analogous to (5.1) for the distribution of T_n is (see [5])

$$k\mathbb{P}(T_n = k) = \sum_{j=1}^{k} g_n(j)\mathbb{P}(T_n = k-j), \tag{5.3}$$

where

$$g_n(j) = x^j \sum_{l=1;\ l|j}^{n} lm_l. \tag{5.4}$$

This is already substantially more complicated than (5.1). However, we note that, for $j \leqslant n$,

$$g_n(j) = g(j) \equiv x^j \sum_{l=1;\ l|j}^{j} lm_l,$$

and that

$$\lim_{i\to\infty} g(i) = \theta. \tag{5.5}$$

On the other hand, for $j > n$ we have

$$\begin{aligned} g_n(j) &= x^j \sum_{l=1;l|j}^{n} lm_l \leqslant x^j \sum_{l=1}^{n} lm_l \\ &= x^{j-n}\sum_{l=1}^{n}(lm_l x^l)x^{n-l} \leqslant \tilde{\theta}x^{j-n}\sum_{l=0}^{n-1} x^l \leqslant \frac{\tilde{\theta}x^{j-n}}{1-x}, \end{aligned} \tag{5.6}$$

where

$$\tilde{\theta} = \sup_{j\geqslant 1} jm_j x^j < \infty$$

under assumption (5.2).

Applying Theorem 4.4 when $k > n$ and using (5.6) shows that

$$\begin{aligned} \sum_{i=n+1}^{k} g_n(i)\mathbb{P}(T_n = k-i) &\leqslant \sum_{i=n+1}^{k} \frac{\tilde{\theta}x^{i-n}}{1-x}\mathbb{P}(T_n = k-i) \\ &= \frac{\tilde{\theta}}{1-x}\max_{l\geqslant 0}\mathbb{P}(T_n = l)\sum_{l=0}^{k-n-1} x^{k-n-l} \\ &\leqslant \max_{l\geqslant 0}\mathbb{P}(T_n = l)\frac{\tilde{\theta}x}{(1-x)^2} \\ &= o(1), \end{aligned} \tag{5.7}$$

uniformly in $k > n$.

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press

This can be exploited to verify (LLA) as follows. The bound (5.7) shows that, for $k = 0, 1, \ldots,$

$$k\mathbb{P}(T_n = k) = \sum_{i=1}^{n} g(i)\mathbb{P}(T_n = k - i) + o(1), \tag{5.8}$$

uniformly in $k \geqslant 0$. The method of proof in the previous section, together with (5.5), then shows that

$$k\mathbb{P}(T_n = k) = \theta\mathbb{P}(k - n \leqslant T_n < k) + r_n(k),$$

where $r_n(k) \to 0$ as $n \to \infty$, uniformly in k. The result follows from Theorem 4.3.

5.3. Selections

The next case we consider is the case of combinatorial selections, for which the Z_j are binomially distributed with

$$\mathbb{P}(Z_i = k) = \binom{m_i}{k} \left(\frac{x^i}{1 + x^i}\right)^k \left(\frac{1}{1 + x^i}\right)^{m_i - k}, \quad k = 0, 1, \ldots, m_i,$$

for any $0 < x < 1$. Once more, the assumption that

$$m_i \sim \frac{\theta y^i}{i}$$

is necessary. In this case, we take $x = y^{-1} \in (0, 1)$, and (LC) and the tail condition hold.

To verify (LLA), the method of the previous section can be used, but this time based on the recurrence (see [5]) in (5.3), where

$$g_n(j) = x^j \sum_{l=1;\ l|j}^{n} (-1)^{j/l-1} l m_l. \tag{5.9}$$

The steps that lead to (5.6) and (5.7) follow immediately, with appropriate modification for the alternating nature of the $g_n(j)$.

5.4. Biased combinatorial structures

The preceding results are applicable to combinatorial structures that are not chosen uniformly from the set of possible structures of weight n, but rather with probability proportional to $\kappa^{\#\text{components}}$, for some $\kappa > 0$. According to the results in Section 8 of [5], these models also satisfy the identity (2.2); for assemblies the Z_j are Poisson with mean $\kappa m_j x^j / j!$, for multisets they are negative binomial with parameters m_j and κx^j, and for selections, binomial with parameters m_j and $\kappa x^j/(1 + \kappa x^j)$. It follows that if the uniform structure satisfies the conditions in (LC), then so do the biased structures, with θ replaced by $\kappa\theta$. Theorem 4.5 then follows from the earlier results from this section. For further examples of biasing, see [5].

5.5. General combinatorial structures

Now suppose the Z_i are arbitrary $\mathbb{Z}_+$-valued random variables, with means $\mathbb{E}Z_i$ satisfying (LC). In combinatorial settings we are aware of (for example, [10]), Z_j can be decomposed

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press

into the sum of m_j i.i.d. random variables $Y_{j1}, \ldots, Y_{jm_j}$, each with p.g.f. $\phi_j(s)$ and means

$$\mathbb{E}Y_{j1} = y_j,$$

with the y_j eventually decreasing, and such that

$$j\mathbb{E}Z_j = jm_j y_j \to \theta \in (0, \infty).$$

Theorems 4.3 and 4.4 continue to hold, and

$$\begin{aligned} \mathbb{P}(Z_j \geqslant 2) &\leqslant \mathbb{E}\left\{ \sum_{i=1}^{m_j} \mathbb{1}(Y_{j_i} \geqslant 2) + \sum_{1 \leqslant i < l \leqslant m_j} \mathbb{1}(Y_{j_i} \geqslant 1)\mathbb{1}(Y_{j_l} \geqslant 1) \right\} \\ &\leqslant m_j \mathbb{P}(Y_{j_1} \geqslant 2) + \frac{1}{2} m_j^2 y_j^2 \end{aligned}$$

is of order $O(j^{-2})$ under (LC) if also

$$m_j \mathbb{P}(Y_{j_1} \geqslant 2) = O(j^{-2});$$

this is typically the case. However, to get further we need a recursion for the point probabilities $\mathbb{P}(T_n = k)$. Since

$$\mathbb{E}s^{T_n} = \prod_{j=1}^{n} \left(\phi_j(s^j)\right)^{m_j},$$

logarithmic differentiation leads to

$$k\mathbb{P}(T_n = k) = \sum_{l=1}^{k} g_n(l)\mathbb{P}(T_n = k - l),$$

where

$$g_n(l) = \sum_{j=1}^{n} jm_j [s^{l-j}] \frac{\phi_j'(s^j)}{\phi_j(s^j)}, \tag{5.10}$$

$[x^l]f(x)$ denoting the coefficient of x^l in $f(x)$. The following example shows that this recursion need not be easy to use.

Example 5.1. Suppose that a combinatorial structure is conditioned to have at most one component of each size [18]. If the original structure $\tilde{C}^{(n)}$ satisfies a conditioning relation like (2.2), then

$$\begin{aligned} \mathbb{P}(\tilde{C}^{(n)} = a | \tilde{C}^{(n)} \leqslant \mathbf{1}) &= \mathbb{P}(\tilde{Z}[1, n] = a | \tilde{Z}[1, n] \leqslant \mathbf{1}, \tilde{T}_n = n) \\ &= \mathbb{P}(Z[1, n] = a | T_n = n), \end{aligned}$$

where $\mathbf{1} = (1, \ldots, 1)$ and $Z = (Z_1, \ldots)$ is a vector of independent Bernoulli random variables satisfying

$$\mathbb{P}(Z_j = a) = \mathbb{P}(\tilde{Z}_j = a | \tilde{Z}_j \leqslant 1), \ a = 0, 1. \tag{5.11}$$

If the original $\tilde{Z}_j$ satisfy the (LC), then so too do the Z_j; that is,

$$\pi_j \equiv \mathbb{P}(Z_j = 1) \sim \frac{\theta}{j}, \ \theta \in (0, \infty), \tag{5.12}$$

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press

and then $j\mathbb{E}Z_j = j\mathbb{P}(Z_j = 1) \to \theta$ automatically, and the tail condition is trivially satisfied. The point probabilities $\mathbb{P}(T_n = k)$ satisfy an equation of the form (5.3). Adapting (5.10) to the present setting with $m_j = 1$, $y_j = \pi_j$ and $\phi_j(s) = 1 - \pi_j + \pi_j s$ leads, after some simplification, to the fact that

$$g_n(l) = -\sum_{j=1, j|l}^{n} (-1)^{l/j} j h_j^{l/j}, \ h_j \equiv \frac{\pi_j}{1-\pi_j}. \tag{5.13}$$

It seems difficult to make progress with this approach in general, although in special cases verification of (LLA) should be possible. In [4], we have developed an alternative approach based on Stein's method for compound Poisson approximation (*cf.* [6]). This leads to recursions for point probabilities that are easier to handle. If, for example, the Z_i satisfy (LC), together with the mild additional condition that

$$\sum_{i \geqslant 1} i\mathbb{E}(Z_i \mathbb{1}[Z_i \geqslant r]) < \infty, \quad \text{for some } r, \tag{5.14}$$

then (LLA) follows. Note that (5.14) clearly holds with $r = 2$ for Example 5.1. The alternative approach also provides bounds on the accuracy of the approximations in this paper.

Acknowledgement

We thank Dr Jennie Hansen for helpful comments on an earlier draft.

References

[1] Aldous, D. J. (1985) *Exchangeability and Related Topics*, Vol. 1117 of *Lecture Notes in Mathematics*, Springer, New York, pp. 1–198.
[2] Arratia, R., Barbour, A. D. and Tavaré, S. (1992) Poisson process approximations for the Ewens Sampling Formula. *Ann. Appl. Probab.* **2** 519–535.
[3] Arratia, R., Barbour, A. D. and Tavaré, S. (1993) On random polynomials over a finite field. *Math. Proc. Camb. Phil. Soc.* **114** 347–368.
[4] Arratia, R., Barbour, A. D. and Tavaré, S. (1998) *Logarithmic Combinatorial Structures.* In preparation.
[5] Arratia, R. and Tavaré, S. (1994) Independent process approximations for random combinatorial structures. *Adv. Math.* **104** 90–154.
[6] Barbour, A. D., Chen, L. H. Y. and Loh, W.-L. (1992) Compound Poisson approximation for nonnegative random variables via Stein's method. *Ann. Probab.* **20** 1843–1866.
[7] Ewens, W. J. (1972) The sampling theory of selectively neutral alleles. *Theoret. Pop. Biol.* **3** 87–112.
[8] Goncharov, V. L. (1944) Some facts from combinatorics. *Izvestia Akad. Nauk. SSSR, Ser. Mat.* **8** 3–48. See also: On the field of combinatory analysis. *Trans. Amer. Math. Soc.* **19** 1–46.
[9] Hansen, J. C. (1994) Order statistics for decomposable combinatorial structures. *Random Struct. Alg.* **5** 517–533.
[10] Hansen, J. C. and Schmutz, E. (1993) How random is the characteristic polynomial of a random matrix? *Math. Proc. Camb. Phil. Soc.* **114** 507–515.
[11] Kingman, J. F. C. (1975) Random discrete distributions. *J. Royal Statist. Soc.* **37** 1–22.

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press

[12] Kingman, J. F. C. (1977) The population structure associated with the Ewens sampling formula. *Theoret. Pop. Biol* **11** 274–283.
[13] Kingman, J. F. C. (1993) *Poisson Processes*, Oxford University Press, Oxford.
[14] Kolchin, V. F. (1976) A problem of the allocation of particles in cells and random mappings. *Theor. Probab. Appl.* **21** 48–63.
[15] Kolchin, V. F. (1986) *Random Mappings*, Optimization Software, Inc., New York.
[16] Scheffé, H. (1947) A useful convergence theorem for probability distributions. *Ann. Math. Stat.* **18** 434–438.
[17] Shepp, L. A. and Lloyd, S. P. (1966) Ordered cycle lengths in a random permutation. *Trans. Amer. Math. Soc.* **121** 340–357.
[18] Stark, D. (1994) Total variation distance for independent process approximations of random combinatorial objects. PhD thesis, University of Southern California.
[19] Vershik, A. M. and Shmidt, A. A. (1977) Limit measures arising in the theory of groups I. *Theor. Probab. Appl.* **22** 79–85.
[20] Vervaat, W. (1972) *Success Epochs in Bernoulli Trials with Applications in Number Theory*. Vol. 42 of *Mathematical Center Tracts*, Mathematisch Centrum, Amsterdam.
[21] Watterson, G. A. (1974) The sampling theory of selectively neutral alleles. *Adv. Appl. Probab.* **6** 463–488.
[22] Watterson, G. A. (1976) The stationary distribution of the infinitely-many-alleles diffusion model. *J. Appl. Probab.* **13** 639–651.

https://doi.org/10.1017/S0963548399003788 Published online by Cambridge University Press