Theoretical Population Biology 122 (2018) 5–11

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Simulating the component counts of combinatorial structures

Richard Arratia [a], A.D. Barbour [b], W.J. Ewens [c], Simon Tavaré [d,*]

[a] Department of Mathematics, University of Southern California, Los Angeles, CA 90089, USA
[b] Institut für Mathematik, Universität Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland
[c] Department of Biology, University of Pennsylvania, 433 S University Ave, Philadelphia, PA 19104, USA
[d] DAMTP, University of Cambridge, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, UK

ARTICLE INFO

Article history:
Available online 9 February 2018

Keywords:
Chinese restaurant process
Ewens sampling formula
Feller coupling
Random mappings
Random permutations
Spaghetti loop distribution

ABSTRACT

This article describes and compares methods for simulating the component counts of random logarithmic combinatorial structures such as permutations and mappings. We exploit the Feller coupling for simulating permutations to provide a very fast method for simulating logarithmic assemblies more generally. For logarithmic multisets and selections, this approach is replaced by an acceptance/rejection method based on a particular conditioning relationship that represents the distribution of the combinatorial structure as that of independent random variables conditioned on a weighted sum. We show how to improve its acceptance rate. We illustrate the method by estimating the probability that a random mapping has no repeated component sizes, and establish the asymptotic distribution of the difference between the number of components and the number of distinct component sizes for a very general class of logarithmic structures.

© 2018 Published by Elsevier Inc.

1. Introduction

Paul Joyce had a long-standing interest in the structure of the Ewens Sampling Formula (Joyce and Tavaré, 1987; Tavaré et al., 1989; Joyce, 1995), denoted by ESF in what follows, its asymptotics (Joyce et al., 2002), and in likelihood and simulation-based methods for inference in population genetics (Nordborg et al., 2001; Joyce et al., 2012). Our contribution to this memorial volume also exploits simulation, asymptotics and the Ewens Sampling Formula, to study the component counting structure of a broad class of combinatorial objects. We hope you like it, Paul!

We begin with a recreational motivation. Peter Winkler, in his book Mathematical Mind-Benders (Winkler, 2003), posed the following question[1] :

> *Spaghetti loops.* The 100 ends of 50 strands of cooked spaghetti are paired at random and tied together. How many pasta loops should you expect to result from this process, on average?

In our view, this question involves the case $n = 50, \theta = 1/2$ of the **E**nds of **S**paghetti **F**ormula, more popularly known as the ESF; we will ask more advanced questions, such as:

> What is the chance that all the loops have different lengths? (Either with exactly 50 strands, or in the limit, as the number of strands tends to infinity.)

To get to the connection between spaghetti loops and the ESF, we begin with a brief description of the relationship between the cycle structure of random permutations and the ESF. It is convenient to do this by describing two methods for simulating random permutations of length n by exploiting a sequence of independent random variables $B_1, B_2, \ldots$ with distribution given by

$$\mathbb{P}_\theta(B_i = j) = \begin{cases} \dfrac{\theta}{\theta + i - 1}, & j = i, \\[2ex] \dfrac{1}{\theta + i - 1}, & j = 1, 2, \ldots, i - 1, \end{cases} \tag{1}$$

where the parameter $\theta \in (0, \infty)$. The first method, the Chinese Restaurant Process, simulates a biased permutation of $[n] = \{1, 2, \ldots, n\}$ using $B_1, B_2, \ldots, B_n$, while the second, the Feller Coupling, achieves the same end by using the reverse order $B_n, B_{n-1}, \ldots, B_1$; see Arratia et al. (1992).

1.1. The Chinese Restaurant Process

This generates the cycles of a permutation as follows. The integer 1 starts a cycle. The integer 2 is placed to the right of 1, in the same cycle, with probability $1/(\theta + 1)$, or begins a new cycle

* Corresponding author.
E-mail address: st321@cam.ac.uk (S. Tavaré).

[1] Winkler noted that this is equivalent to the 'blades of grass' game described on page 198 of Gardner (Gardner, 1971).

https://doi.org/10.1016/j.tpb.2018.02.002
0040-5809/© 2018 Published by Elsevier Inc.

with probability $\theta/(\theta+1)$. Suppose that the first $n-1$ integers have been assigned to cycles. Then integer n starts a new cycle with probability $\theta/(\theta+n-1)$, or is placed to the right of integer j with probability $\mathbb{P}_\theta(B_n = j) = 1/(\theta+n-1), j = 1, 2, \ldots, n-1$. It follows that for any permutation π of $[n]$ having k cycles,

$$\mathbb{P}_\theta(\pi) = \frac{\theta^k}{\theta_{(n)}}, \tag{2}$$

where $\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$. The cycles generated in this way are ordered, in that the first contains the integer 1, the second cycle the smallest integer not in the first cycle, and so on. Furthermore, if we define independent Bernoulli random variables $\xi_i = 1$ if $B_i = i$ and $= 0$ if $B_i < i$, then

$$\mathbb{P}_\theta(\xi_i = 1) = \frac{\theta}{\theta+i-1},\ i = 1, 2, \ldots$$

and the number of cycles in π is given by $K_n = \xi_1 + \cdots + \xi_n$.

1.2. The Feller Coupling

We start with 1 in the first cycle. If $B_n = n$, so that $\xi_n = 1$, we finish that cycle, and start the next cycle with the smallest available integer. If $B_n < n$, then B_n indicates which of the remaining $n-1$ integers is used next, and this is placed to the right of 1 in the same cycle. Continuing in this way also produces a permutation with cycles ordered by their smallest integer. If $B_i = i$, and so $\xi_i = 1$, the current cycle is finished, and the next starts with the smallest available integer. When $B_i < i$, B_i indicates which of the remaining $i-1$ integers is placed at the end of the growing cycle.

For any permutation π of $[n]$, it is immediate that (2) holds, but the cycles have been constructed using the B_i in the order $B_n, B_{n-1}, \ldots, B_1$. Note also that in the Feller Coupling the cycles are completed sequentially, unlike in the Chinese Restaurant Process.

The lengths of the ordered cycles are precisely the spacings between the 1s in the sequence $1, \xi_n, \xi_{n-1}, \ldots, \xi_1$, so that the number of cycles in a permutation is $\xi_n + \cdots + \xi_1$.

We can now see the connection with the spaghetti problem. Starting with $n = 50$ cooked pieces, we had 100 ends; pretending these are labelled 1 to 100, the random choices begin with end 1 making a 99-way choice to determine which end to join; finishing a loop at this first step corresponds to the event $\xi_{50} = 50$, having probability 1/99. At subsequent steps, if the last step did not complete a loop, then continue to work with the developing strand. In this way, the lengths of the loops formed, in order, are the spacings between ones, reading the sequence $\xi_1\xi_2\cdots\xi_{50}1$ from right to left. To determine the value of θ, we see that

$$\mathbb{P}_\theta(\xi_i = 1) = \frac{1}{2i-1} = \frac{1/2}{1/2+i-1},$$

so we have identified $\theta = 1/2$.

1.3. The Ewens Sampling Formula

For most purposes, interest focuses on the distribution of the cycle counts of a permutation π. We denote by $C_j(n)$ the number of cycles of size j in a permutation of size n, so that

$$C_1(n) + 2C_2(n) + \cdots + nC_n(n) = n.$$

It follows from (2) that the joint distribution of the cycle counts is given by the Ewens Sampling Formula (ESF) (Ewens, 1972):

$$\mathbb{P}_\theta(C_j(n) = c_j, j = 1, 2, \ldots, n) = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{c_j} \frac{1}{c_j!}, \tag{3}$$

for non-negative integers $c_1, \ldots, c_n$ satisfying $c_1+2c_2+\cdots+nc_n = n$. Thus the joint distribution of the counts of spaghetti loops of length $1, 2, \ldots, n$ is given by the ESF with parameter $\theta = 1/2$.

We remark that the ESF with parameter θ arises from the ESF with parameter $\theta = 1$ by biasing the uniform case by θ^{K_n}, where $K_n = C_1(n) + C_2(n) + \cdots + C_n(n) = \xi_1 + \cdots + \xi_n$.

1.4. The conditioning relation

Watterson (Watterson, 1974) showed that the distribution $\mathcal{L}(C_1(n), \ldots, C_n(n))$ in (3) can be realized in the form

$$\mathcal{L}(C_1(n), \ldots, C_n(n)) = \mathcal{L}(Z_1, \ldots, Z_n | T_{0n} = n), \tag{4}$$

where

$$Z_1, Z_2, \ldots \text{ are independent Poisson random variables} \tag{5}$$

satisfying

$$\mathbb{E}(Z_j) = \frac{\theta}{j}, \tag{6}$$

and

$$T_{0n} = Z_1 + 2Z_2 + \cdots + nZ_n. \tag{7}$$

The relationship in (4) suggests a third way to simulate samples from the ESF with parameter θ, namely a rejection method that simulates independent $Z_1, \ldots, Z_n$ and accepts $(Z_1, \ldots, Z_n)$ as a realization of $(C_1(n), \ldots, C_n(n))$ if $T_{0n} = n$; otherwise, reject the simulation, and repeat. The acceptance rate is then $\mathbb{P}(T_{0n} = n)$. If the acceptance rate is small, this third approach can be inefficient; we return to some of its properties in a more general setting later.

2. What is the chance that all the cycle lengths are distinct?

We return to the question raised in the introduction, namely what is the probability that a θ-biased permutation has all its cycle lengths distinct? For the spaghetti loop problem with $\theta = 1/2$, simulation of one million permutations via the Feller coupling with $n = 50$ yielded an estimate of 0.8377.

For a uniform ($\theta = 1$) random permutation of n objects, the probability q_n that it has distinct cycle lengths was shown by analytical means in Greene and Knuth (1982) to satisfy the asymptotic formula

$$q_n \sim e^{-\gamma}\left(1 + \frac{1}{n}\right), \text{ as } n \to \infty, \tag{8}$$

where $\gamma \approx 0.577216$ is Euler's constant. The analogous result is also known when the random permutation is distributed according to the ESF with parameter $\theta > 0$. To ease the notation, in what follows we suppress the parameter θ in $\mathbb{P}_\theta(\cdot)$ when there is no cause for confusion.

In Arratia and Tavaré (1992) it is shown that the asymptotic distribution of the difference D_n between the number of cycles and the number of distinct cycle lengths for a permutation of size n satisfies

$$D_n = \sum_{j=1}^{n} (C_j(n) - 1)_+ \Rightarrow D = \sum_{j\geq 1} (Z_j - 1)_+, \tag{9}$$

where the Z_j satisfy (5) and (6) and $(x)_+ = \max(0, x)$.

As a consequence, the probability that all cycle lengths of an n-permutation are distinct, $\mathbb{P}(D_n = 0)$, satisfies

$$\begin{aligned}
\mathbb{P}(D_n = 0) \to \mathbb{P}(D = 0) &= \mathbb{P}\Big(\bigcap_{j\geq 1}\{Z_j \leq 1\}\Big) \\
&= \prod_{j\geq 1} e^{-\theta/j}(1 + \theta/j) \\
&= e^{-\gamma\theta} \lim_{n\to\infty} n^{-\theta} \frac{(\theta+1)\cdots(\theta+n)}{n!} \\
&= e^{-\gamma\theta}/\Gamma(\theta+1).
\end{aligned} \tag{10}$$

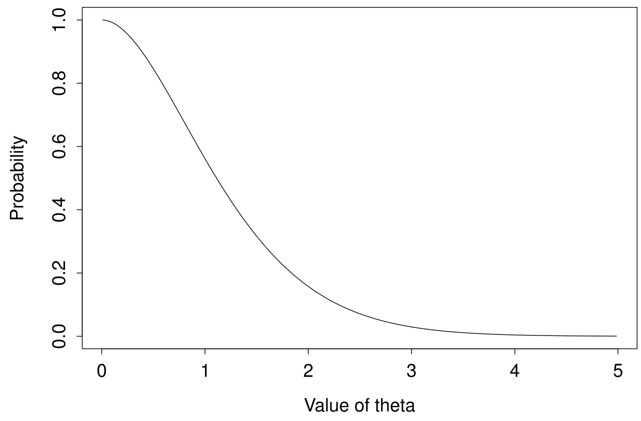

Fig. 1. Plot of $\mathbb{P}(D = 0)$ as a function of θ, from (10).

Fig. 1 plots $\mathbb{P}(D = 0)$ as a function of θ. For the spaghetti loop problem, $\theta = 1/2$ and we have

$$\mathbb{P}(D = 0) = e^{-\gamma/2}/\Gamma(3/2) = 2e^{-\gamma/2}/\sqrt{\pi} \approx 0.84550,$$

thus linking π, e and γ in a single formula. For $\theta = 1$, $\mathbb{P}(D = 0) = e^{-\gamma} \approx 0.561460$ as anticipated in (8); thus about 56% of large random permutations have no repeated cycle lengths.

The asymptotic analysis above leaves open how good the approximations actually are for any given value of n. In what follows, we investigate how simulation can be used to show when the asymptotics are adequate, and extend the discussion beyond the ESF, to some more general combinatorial structures.

3. Logarithmic assemblies

The conditioning relation (4), for independent random variables $Z_1, Z_2, \ldots$ taking values in $\{0, 1, 2, \ldots\}$, defines the distribution of the component counts of a broad class of decomposable random structures. For a more detailed overview of these structures, see Chapter 2 of Arratia et al. (2003) for example. Among the examples are the ESF with parameter θ, as observed by Watterson. The ESF itself is a member of the larger family of assemblies, for which, for some $x \in (0, \infty)$ and $m_j \in \mathbb{R}_+, j \geq 1$, the Z_j are Poisson distributed with

$$\mathbb{E}(Z_j) = m_j x^j/j!. \tag{11}$$

Note that, if x is varied, and $Z_j{}^{(x)}$ is used to denote the corresponding random variables, then the probability

$$\begin{aligned}&\mathbb{P}((Z_1{}^{(x)}, \ldots, Z_n{}^{(x)}) = (c_1, \ldots, c_n))\\ &= x^n \mathbb{P}((Z_1{}^{(1)}, \ldots, Z_n{}^{(1)}) = (c_1, \ldots, c_n))/\psi_n(x),\end{aligned}$$

whenever $\sum_{j=1}^n jc_j = n$, where

$$\psi_n(x) := \exp\left\{\sum_{j=1}^n m_j(x^j - 1)/j!\right\}$$

is the same for all choices of $c_1, \ldots, c_n$. Hence it follows that

$$\mathbb{P}(T_{0n}{}^{(x)} = n) = x^n \mathbb{P}(T_{0n}{}^{(1)} = n)/\psi_n(x) \tag{12}$$

also, so that the distribution $\mathcal{L}(C_1(n), \ldots, C_n(n))$ given by (4) is the same for all $x > 0$. Thus the choice of x may even be allowed to depend on n.

Here we focus primarily on the *logarithmic* structures, those that satisfy

$$j\,\mathbb{P}(Z_j = 1) \to \theta \quad \text{and} \quad j\mathbb{E}(Z_j) \to \theta, \quad \text{as } j \to \infty, \tag{13}$$

for some $\theta \in (0, \infty)$. For assemblies satisfying

$$\frac{m_j}{j!} \sim \frac{\theta y^j}{j}, \tag{14}$$

for some $y > 0$, $\theta > 0$, we can take $x = 1/y$ to express them in logarithmic form. An example is given by the random mapping. Letting $\text{Po}(\lambda)$ denote a Poisson distributed random variable with mean λ, we have

$$\begin{aligned}m_j &= (j-1)! \sum_{i=0}^{j-1} \frac{j^i}{i!}\\ &= (j-1)!\, e^j\, \mathbb{P}(\text{Po}(j) < j)\\ &\sim (j-1)!\, e^j/2,\end{aligned}$$

so that we can take $x = 1/e$, the same for all n, giving $\theta = 1/2$.

4. Simulating logarithmic assemblies

To estimate the chance that a random mapping of size n has no repeated component sizes, we resort to simulation once more. The obvious method is via the conditioning relation (4). However, given the speed of the Feller coupling for simulating from the ESF, it makes sense to ask whether it is possible to use simulations of the ESF for some value of θ to generate observations from any other logarithmic structure. This is, indeed, the case.

To see how this can be done, write

$$\lambda_j{}^{(x)} = \frac{m_j x^j}{(j-1)!},$$

and note that, from (12), for any $\theta > 0$,

$$\begin{aligned}&\mathbb{P}((C_1(n), \ldots, C_n(n)) = (c_1, \ldots, c_n))\\ &= \frac{\mathbb{P}(T_{0n}{}^{(x)} = 0)}{\mathbb{P}(T_{0n}{}^{(x)} = n)} \prod_{j=1}^n \left(\frac{\lambda_j{}^{(x)}}{j}\right)^{c_j} \frac{1}{c_j!}\\ &= \frac{\theta_{(n)}\, \mathbb{P}(T_{0n}{}^{(x)} = 0)}{n!\, \mathbb{P}(T_{0n}{}^{(x)} = n)} \left[\prod_{j=1}^n \left(\frac{\lambda_j{}^{(x)}}{\theta}\right)^{c_j}\right] \frac{n!}{\theta_{(n)}} \prod_{j=1}^n \left(\frac{\theta}{j}\right)^{c_j} \frac{1}{c_j!}\end{aligned} \tag{15}$$

for $c_1 + 2c_2 + \cdots + nc_n = n$.

Eq. (15) can be used in an acceptance/rejection algorithm, as follows. Suppose we can find x and θ such that

$$0 \leq \lambda_j{}^{(x)} \leq \theta, \ \text{for } j = 1, 2, \ldots, n.$$

The algorithm then simulates a sample $(c_1, \ldots, c_n)$ of cycle counts from ESF(θ) (using the Feller coupling for example), and accepts $(c_1, \ldots, c_n)$ as a realization from $\mathcal{L}(C_1(n), \ldots, C_n(n))$ for the assembly with probability

$$h(c_1, \ldots, c_n) = \prod_{j=1}^n \left(\frac{\lambda_j{}^{(x)}}{\theta}\right)^{c_j} \leq 1, \tag{16}$$

and otherwise rejects $(c_1, \ldots, c_n)$ and starts again. Notice that x may be chosen to be a function of n.

4.1. What is the chance that a random mapping has no repeated component lengths?

We can study the properties of $D_n := \sum_{j=1}^n (C_j(n) - 1)_+$ by exploiting this simulation method. For random mappings, we have seen that we may take $x = 1/e$, and then

$$\lambda_j = \mathbb{P}(\text{Po}(j) < j),$$

so that $\theta = 1/2$ works.

We can now generate the component counts of random mappings by simulating the ESF($\theta = 1/2$), and using the acceptance probability

$$h(c_1, \ldots, c_n) = \prod_{j=1}^{n} (2\mathbb{P}(\mathrm{Po}(j) < j))^{c_j}.$$

To estimate the chance that a random mapping of size $n = 50$ has no repeated component sizes, we simulated one million accepted values of the component counts, and estimated $\mathbb{P}(D_n = 0) = 0.888$, the proportion of runs that had no repeated sizes. Of course, the simulation provides estimates of the distribution of D_n, and we obtained estimates of $\mathbb{P}(D_n = 1) = 0.099$ and $\mathbb{P}(D_n = 2) = 0.012$, and $\mathbb{E}D_n \approx 0.1266$.

In the acceptance/rejection method for an assembly with rates $\lambda_j{}^{(x)}/j$, the acceptance rate is

$$\mathbb{E}\left(\prod_{j=1}^{n}\left(\frac{\lambda_j{}^{(x)}}{\theta}\right)^{C_j(n)}\right) \tag{17}$$

under the ESF(θ) distribution. For our random mapping example with $n = 50$, the simulation gave an estimated acceptance rate of 0.708.

For very large values of n, we can approximate (17) by

$$\mathbb{E}\left(\prod_{j=1}^{\infty}\left(\frac{\lambda_j{}^{(x)}}{\theta}\right)^{Z_j}\right) = \exp\left(-\theta\sum_{j\geq 1}\frac{1}{j}\left(1 - \frac{\lambda_j{}^{(x)}}{\theta}\right)\right), \tag{18}$$

since the Z_j are Poisson with mean θ/j for ESF(θ).

For random mappings, we took $x = 1/e, \theta = 1/2$ and (18) reduces to

$$\exp\left(-\frac{1}{2}\sum_{j\geq 1}\frac{1}{j}(1 - 2\mathbb{P}(\mathrm{Po}(j < j)))\right) = \frac{1}{\sqrt{2}} \approx 0.7071, \tag{19}$$

in good agreement with the simulated value.

We will prove later that, as $n \to \infty$,

$$D_n \Rightarrow D := \sum_{j\geq 1}(Z_j - 1)_+,$$

where Z_j are independent Poisson random variables with mean $\mathbb{E}Z_j = \mathbb{P}(\mathrm{Po}(j) < j)/j$. In particular, the limiting probability that a random mapping has distinct component lengths is given by

$$\mathbb{P}(D = 0) = \prod_{j=1}^{\infty} e^{-\rho_j/j}\left(1 + \frac{\rho_j}{j}\right) \approx 0.8959. \tag{20}$$

This may be compared to the simulated result for $n = 50$, namely 0.888. The probability in (20) may also be written as

$$\sqrt{2}\, e^{-\gamma/2} \lim_{n\to\infty} n^{-1/2} \prod_{j=1}^{n}\left(1 + \frac{\lambda_j}{j}\right). \tag{21}$$

It is also the case that

$$\mathbb{E}(D_n) \to \mathbb{E}(D) = \sum_{j\geq 1}\left(e^{-\lambda_j/j} - 1 + \frac{\lambda_j}{j}\right) \approx 0.1174. \tag{22}$$

5. Simulation via the conditioning relation

For logarithmic assemblies satisfying (14) we have described a very efficient method for simulating observations from the component counting process. For motivation for simulating more general logarithmic combinatorial structures, we note that we can also use the rejection method together with the conditioning relation (4): simulate values $c_1, c_2, \ldots, c_n$ from independent Poisson random variables $Z_1, Z_2, \ldots, Z_n$ with means given in (11), and accept $(c_1, c_2, \ldots, c_n)$ as an observation from $\mathcal{L}(C_1(n), \ldots, C(n))$ if $c_1 + 2c_2 + \cdots + nc_n = n$.

The acceptance rate of such an algorithm is $\mathbb{P}(T_{0n} = n)$. Theorem 4.13 in Arratia et al. (2003) establishes that

$$n\,\mathbb{P}(T_{0n} = n) \to \frac{e^{-\gamma\theta}}{\Gamma(\theta)}. \tag{23}$$

Hence the expected number s_n of simulated vectors per accepted vector satisfies

$$s_n/n \to \Gamma(\theta)\, e^{\gamma\theta}. \tag{24}$$

In view of the previous discussion, the acceptance rate could be improved by choosing $x = x(n)$ in (11) to maximize $\mathbb{P}(T_{0n} = n)$. For instance, for logarithmic assemblies with $x = 1/y$, we have $\mathbb{E}T_{0n} \sim n\theta$, and, if $\theta \neq 1$, this is far from the value n where we wish the probability of T_{0n} to be large. Appealing to (12) shows that, for an assembly, we should maximize $x^n/\psi_n(x)$. Now, for a logarithmic assembly, we have

$$\psi_n(x) = \exp\left\{\sum_{j=1}^{n} m_j(x^j - 1)/j!\right\} \sim \exp\left\{\theta\sum_{j=1}^{n} y^j(x^j - 1)/j\right\},$$

from (14). Taking $x := y^{-1}e^{-c/n}$, as in Arratia and Tavaré (1994), this gives

$$\begin{aligned} x^n/\psi_n(x) &\sim \Psi_n(y)e^{-c}\exp\left\{\theta\sum_{j=1}^{n} j^{-1}(1 - e^{-cj/n})\right\} \\ &\sim \Psi_n(y)\exp\left\{-c + \theta\int_0^1 x^{-1}(1 - e^{-cx})\,dx\right\}, \end{aligned} \tag{25}$$

where $\Psi_n(y) := y^{-n}\exp\left\{\theta\sum_{j=1}^{n}(y^j - 1)/j\right\}$ is asymptotically equivalent to $x^n/\psi_n(x)$ when $c = 0$, that is, when $x = 1/y$ fixed. Thus we should choose $c := c_1(\theta)$ to maximize $u(c) := -c + \theta\int_0^1 x^{-1}(1 - e^{-cx})\,dx$, which, at least asymptotically, improves the acceptance rate over the choice $x = 1/y$ by a factor of $e^{u(c_1(\theta))}$. Differentiating u with respect to c shows that $c_1(\theta)$ should be chosen to satisfy

$$0 = -1 + \theta\int_0^1 e^{-cx}\,dx, \quad \text{or} \quad \theta(1 - e^{-c}) = c. \tag{26}$$

It is straightforward to compute $c_1(\theta)$ numerically; see Fig. 2. If $\theta = 1$, we have $c_1 = 0$; if $\theta < 1$, the uncorrected mean $n\theta$ is too small, and we need to take $c_1 < 0$ to make the mean larger; conversely, if $\theta > 1$, we need to take $c_1 > 0$. In particular, if $\theta = 1/2$, as for random mappings, we have $c_1 = -1.25643$.

As shown above, the choice of $c = c_1$ improves the acceptance probability over the choice $x = 1/y$ by a factor asymptotic to $e^{u(c_1)}$. Applying (23) to obtain the asymptotics of the acceptance probability for $x = 1/y$ thus gives

$$n\mathbb{P}(T_{0n} = n) \to \frac{e^{-\gamma\theta}}{\Gamma(\theta)}\, e^{u(c_1)}. \tag{27}$$

It follows that, with this choice of c_1, the expected number $s_n(c_1)$, of simulated vectors per accepted vector satisfies

$$s_n(c_1)/n \to \Gamma(\theta)\, e^{\gamma\theta}\, e^{-u(c_1)}. \tag{28}$$

The quantity $e^{u(c_1)}$ is the asymptotic factor by which the number of simulations per accepted simulation is reduced when using the large deviation value of c, rather than the naive value $c = 0$. For example, when $\theta = 4$ this factor is 47.94. Values for $\theta \in (0, 5]$ are given in Fig. 3.

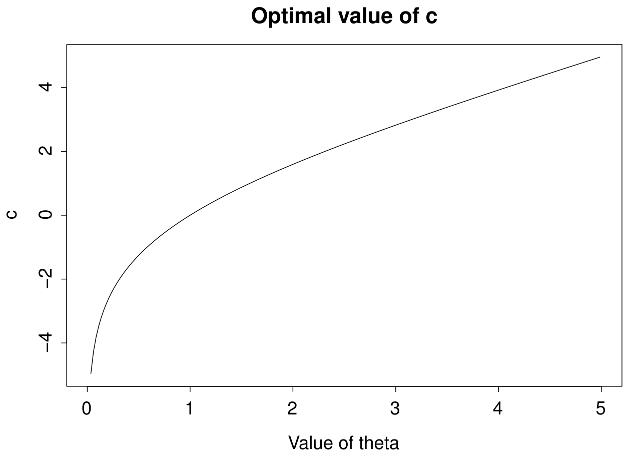

Fig. 2. Plot of the optimal value c_1, obtained from (26).

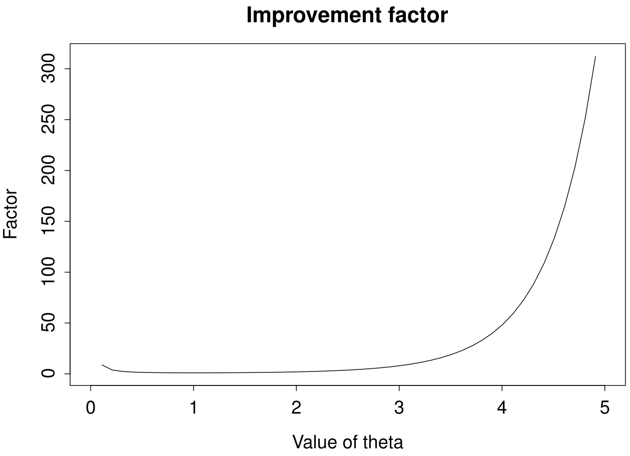

Fig. 3. Value of $e^{u(c_1)}$ for $\theta \in (0, 5]$.

6. Logarithmic multisets and selections

The discussion so far has concentrated on assemblies, for which the Z_j have Poisson distributions. There are two other big families of structures that are ubiquitous in classical combinatorics, and whose distributions can be obtained using the conditioning relation (4): multisets, and selections. Multisets have negative binomial distributions, with $Z_j \sim \mathrm{NB}(m_j, x^j)$, for $0 < x < 1$, and, as for assemblies, the distribution of component lengths obtained using the conditioning relation is the same for all choices of x. If $m_j \sim j^{-1}\theta y^j$ as $j \to \infty$ for some $\theta > 0$ and $y > 1$, we can take $x = 1/y$ to give a family $(Z_j, j \geq 1)$ satisfying the logarithmic condition (13). Selections have $Z_j \sim \mathrm{Bi}(m_j, x^j/(1+x^j))$, for $x > 0$, and a logarithmic representation is obtained using the choice $x = 1/y$ if $m_j \sim j^{-1}\theta y^j$ for some $y > 1$. Monic polynomials over the finite field GF(q) provide an example of a logarithmic multiset, and square free monic polynomials over GF(q) give rise to a logarithmic selection. More examples are to be found in Chapter 2 of Arratia et al. (2003). We shall work in the broader context of logarithmic combinatorial structures, whose distributions can be derived from the conditioning relation (4), for a fixed sequence of independent random variables $(Z_j, j \geq 1)$, that satisfy the logarithmic condition (13).

We note first that there seem to be no natural analogues for multisets and selections of the Feller coupling that proved so effective in simulating the ESF. But we can resort to the conditioning relation approach described in Section 5. For n very large, the rejection rate may become too large to make simulation a practicable option; in contrast, in such circumstances, asymptotic theory can be expected to give good results. Here, we give asymptotics sufficient to cover the behaviour of D_n, based on the material of Arratia et al. (2000) and Arratia et al. (2003), in a rather general context. We will also show that the improved acceptance rates for simulation obtained in Section 5 can be obtained in much greater generality.

In order to simplify the discussion, we make some further assumptions. We require that, for some $\theta > 0$,

$$\mathbb{P}(Z_j = 1) = \frac{\theta}{j}(1+\varepsilon_{j1}); \qquad \mathbb{P}(Z_j = r) = \frac{\theta}{j}\varepsilon_{jr}, \quad r \geq 2, \tag{29}$$

where, as $j \to \infty$,

$$|\varepsilon_{j1}| \leq \varepsilon(j)c_1; \qquad \varepsilon_{jr} \leq \varepsilon(j)c_r, \tag{30}$$

and

$$\lim_{j\to\infty} \varepsilon(j) = 0; \quad \sum_{r\geq 2} rc_r < \infty. \tag{31}$$

This is the Uniform Logarithmic Condition of Arratia et al. (2000), and is satisfied for all logarithmic assemblies, multisets and selections. It implies, in particular, that the convergence in (23) holds, by Corollary 2.8 of Arratia et al. (2000).

6.1. Convergence in distribution of D_n

Theorem. *Let D_n be the difference between the number of components and the number of distinct component lengths in a logarithmic combinatorial structure of size n satisfying the conditions (29), (30) and (31) for some $\theta > 0$. Suppose also that*

$$\sum_{j\geq 1} j^{-1}\varepsilon(j) < \infty. \tag{32}$$

Then, as $n \to \infty$,

$$D_n \Rightarrow D := \sum_{j\geq 1}(Z_j - 1)_+; \tag{33}$$

$$\mathbb{P}(D_n = 0) \to \mathbb{P}(D = 0) = \prod_{j=1}^{\infty}\mathbb{P}(Z_j \leq 1). \tag{34}$$

If, in addition, for some $C, \gamma > 0$ and for all j,

$$|\varepsilon_{j1}| \leq Cj^{-\gamma} \quad \text{and} \quad |\varepsilon_{j+1,1} - \varepsilon_{j1}| \leq Cj^{-1-\gamma}, \tag{35}$$

and also

$$\varepsilon_{j2} \leq Cj^{-1} \quad \text{and} \quad \varepsilon_{jr} \leq Cj^{-1-\gamma}r^{-2-\gamma},\ r \geq 3, \tag{36}$$

then, as $n \to \infty$,

$$\mathbb{E}(D_n) \to \mathbb{E}(D).$$

Proof. By Theorem 3.1 of Arratia et al. (2000), the total variation distance between the distribution of $\sum_{j\leq b}(C_j(n) - 1)_+$ and that of $\sum_{j\leq b}(Z_j - 1)_+$ tends to zero as $n \to \infty$ if $b = b_n$ satisfies $b_{n/n} \to 0$. Next, Theorem 3.2 of Arratia et al. (2000) shows that, for $b = b_n \to \infty$ and $b_{n/n} \to 0$, the total variation distance between the distributions of $\sum_{j=b+1}^{n}(C_j(n) - 1)_+$ and $\sum_{j=b+1}^{n}(C_j^*(n) - 1)_+$ tends to zero as $n \to \infty$, where $C_j^*(n), j = 1, \ldots, n$ are the cycle counts of the ESF with parameter θ. Finally, Lemma 14.2 of Arratia et al. (2003) shows that $\mathbb{P}(\sum_{b+1}^{n}(C_j^*(n) - 1)_+ > 0) = O(1/b)$ as

$b \to \infty$. Taking $b = b_n$ so that $b_n \to \infty$ and $b_n/n \to 0$, the first two limits are established.

The convergence of $\mathbb{E}(D_n)$ to $\mathbb{E}(D)$ under the extra assumptions requires considerably more work, and we omit the technical details here. □

6.2. *Choosing the optimal acceptance rate*

For simulation, the improved acceptance rate obtained by using random variables $(Z_j^{(x(n))}, j \geq 1)$ that are tailored to the choice of n also holds in greater generality. The appropriate choices of $Z_j^{(x)}$ are given by tilted versions of Z_j:

$$\mathbb{P}(Z_j^{(x)} = r) = \frac{x^{jr}\mathbb{P}(Z_j = r)}{\mathbb{E}(x^{jZ_j})}, \quad x > 0.$$

The acceptance probability obtained by using the $Z_j^{(x)}$ with $x = e^{-c/n}$ is thus equal to that using the original Z_j, multiplied by the factor

$$e^{-c} \Big/ \prod_{j=1}^{n} \mathbb{E}(e^{-cjZ_j/n}),$$

to be maximized with respect to c. Note that, for some $c < 0$, it may be the case that $\mathbb{E}(e^{-cjZ_j/n}) = \infty$ for some j. If this is so, then the factor is zero, and hence smaller than the value 1 obtained at $c = 0$, so that such c cannot be optimal. With this in mind, we define ϕ^* to be the supremum of those $\phi \geq 0$ such that

$$\lim_{j\to\infty} j\mathbb{E}\{Z_j e^{\phi Z_j}\mathbb{I}(Z_j \geq 3)\} = 0. \tag{37}$$

Theorem. *Under the conditions (29), (30) and (36), if $c > -\phi^*$, we have*

$$\prod_{j=1}^{n} \mathbb{E}(e^{-cjZ_j/n}) \sim \exp\left\{-\theta \int_0^1 \frac{(1 - e^{-cx})}{x}\, dx\right\}.$$

Hence, asymptotically, the optimal choice of c is still $c_1(\theta)$ satisfying (26), provided that $\theta \geq 1$, and the improvement is by a factor of $e^{u(c_1(\theta))}$, as for assemblies. If $\theta < 1$, then the asymptotically best choice is $c_1(\theta)$ if $c_1(\theta) > -\phi^*$, and again the improvement is as before.

Proof. We first note that

$$\begin{aligned}\mathbb{E}(e^{-cjZ_j/n}) &= 1 - \sum_{r\geq 1} \mathbb{P}(Z_j = r)(1 - e^{-cjr/n}) \\ &= 1 - x_j - y_j - w_j - z_j,\end{aligned}$$

where

$$x_j := \frac{\theta}{j}(1 - e^{-cj/n}); \quad y_j := \frac{\theta}{j}\varepsilon_{j1}(1 - e^{-cj/n});$$
$$w_j := \mathbb{P}(Z_j = 2)(1 - e^{-2cj/n}); \quad z_j := \sum_{r\geq 3} \mathbb{P}(Z_j = r)(1 - e^{-cjr/n}).$$

Take first the case $c > 0$. Since $0 < 1 - e^{-cjr/n} \leq cjr/n$, it follows from (36) that

$$0 \leq w_j \leq 2cjn^{-1}\mathbb{P}(Z_j = 2) = O(1/nj);$$
$$0 \leq z_j \leq cjn^{-1} \sum_{r\geq 3} r\mathbb{P}(Z_j = r) = O(1/nj^{1+\gamma}).$$

Then, similarly,

$$|y_j| \leq n^{-1}\theta c|\varepsilon_{j1}| \quad \text{and} \quad 0 \leq x_j \leq n^{-1}\theta c.$$

Hence, as $n \to \infty$,

$$\sum_{j=1}^{n} (|y_j| + w_j + z_j) = O(\log n/n);$$

$$\sum_{j=1}^{n} (x_j + |y_j| + w_j + z_j)^2 = O(n^{-1}),$$

so that

$$\begin{aligned}\sum_{j=1}^{n} \log(\mathbb{E}\{e^{-cjZ_j/n}\}) &= -\sum_{j=1}^{n} x_j + O(\log n/n) \\ &\sim -\theta \int_0^1 x^{-1}(1 - e^{-cx})\, dx,\end{aligned}$$

completing the proof for $c > 0$.

For $-\phi^* < c < 0$, we have

$$\begin{aligned}0 \leq -z_j &\leq |c|jn^{-1} \sum_{r\geq 3} r\mathbb{P}(Z_j = r)e^{r|c|} \\ &\leq |c|jn^{-1}\mathbb{E}\{Z_j e^{|c|Z_j}\mathbb{I}(Z_j \geq 3)\},\end{aligned}$$

which, in view of (37), yields $\sum_{j=1}^{n} |z_j| = o(1)$ and $\sum_{j=1}^{n} z_j^2 = o(n^{-1})$ as $n \to \infty$. The remaining argument is as above. □

If $\phi^* > 0$, an analogous argument can be used to improve simulation for all values of $\theta < 1$, by first modifying the random variables Z_j. Let $Z_j(b) = Z_j$ for $1 \leq j \leq b$, and $Z_j(b) = Z_j\mathbb{I}(Z_j \leq 2)$ for $j > b$; the probability that the sequences differ is at most

$$\sum_{j>b} \mathbb{P}(Z_j \geq 3) = O(b^{-1-\gamma})$$

under (36). Thus, to simulate $(C_1(n), \ldots, C_n(n))$, we can use samples from $(Z_j(\delta n), j \geq 1)$, for any fixed $\delta > 0$, with an error probability of order $O(n^{-1-\gamma})$, which is of smaller order than the acceptance probability $\mathbb{P}(T_{0n} = n)$. Repeating the argument above by tilting the sequence $(Z_j(\delta n), j \geq 1)$, we only have $z_j > 0$ for $j \leq \delta n$, and hence the largest exponent in the moment generating function bound for $\sum_{j=1}^{n} z_j$ is $|c|\delta$. Thus, for each $c \leq -\phi^*$, by choosing $\delta(c)$ such that $|c|\delta(c) < \phi^*$, we find that the improvement factor is asymptotically given by $e^{u(c)}$ once more, with $c_1(\theta)$ the best choice of c. The simulations are then carried out with the $x = e^{-c_1(\theta)/n}$ tilted versions of the sequence $(Z_j(\delta n), j \geq 1)$, for any δ such that $|c_1(\theta)|\delta < \phi^*$.

7. Discussion

We conclude with a brief discussion of the efficiency of the three methods for simulating observations from the ESF. A reasonable way to assign a "cost" to simulation algorithms is to report the asymptotic growth, relative to n, of the number of calls to a random number generator. With this notion of cost, to get one ESF sample, the cost of the Feller coupling algorithm is $O(\log n)$, and the cost of the algorithm based on the Chinese Restaurant coupling is $O(n)$. For the algorithm based on the conditioning relation the cost of the straightforward algorithm is $O(n^2)$, a factor of n for the cost to propose the independent $(Z_1, \ldots, Z_n)$ and an additional factor of n for the expected number of trials needed to get one acceptance. The cost can be improved to $O(n \log n)$ by coding to propose $(Z_1, \ldots, Z_n)$ using only $O(\log n)$ calls to the random number generator; see Arratia and DeSalvo (2016), Section 5.1. For random mappings, the preferred method is acceptance/rejection relative to ESF($\theta = 1/2$), with an additional $O(1)$ cost factor relative to whichever ESF generator is used.

Acknowledgement

ADB acknowledges support from ARC grants DP150101459 and DP150103588.

References

Arratia, R., Barbour, A.D., Tavaré, S., 1992. Poisson process approximations for the Ewens Sampling Formula. Ann. Appl. Probab. 2, 519–535.

Arratia, R., Barbour, A.D., Tavaré, S., 2000. Limits of logarithmic combinatorial structures. Ann. Probab. 28, 1620–1644.

Arratia, R., Barbour, A.D., Tavaré, S., 2003. Logarithmic combinatorial structures: a probabilistic approach. In: EMS Monographs in Mathematics. European Mathematical Society.

Arratia, R., DeSalvo, S., 2016. Probabilistic divide-and-conquer: a new exact simulation method, with integer partitions as an example, probability. Comb. Comput. 25, 324–351.

Arratia, R., Tavaré, S., 1992. Limit theorems for combinatorial structures via discrete process approximations. Rand. Struct. Alg. 3, 321–345.

Arratia, R., Tavaré, S., 1994. Independent process approximations for random combinatorial structures. Adv. Math. 104, 90–154.

Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. Theor. Popul. Biol. 3, 87–112.

Gardner, M., 1971. The Sixth Book of Mathematical Puzzles and Diversions from Scientific American. Simon & Schuster.

Greene, D.H., Knuth, D.E., 1982. Mathematics for the Analysis of Algorithms, second ed.. Birkhaüser.

Joyce, P., 1995. Robustness of the Ewens Sampling Formula. J. Appl. Probab. 32, 609–622.

Joyce, P., Genz, A., Buzbas, E.O., 2012. Efficient simulation and likelihood methods for nonneutral multi-allele models. J. Comput. Biol. 19, 650–661.

Joyce, P., Krone, S., Kurtz, T., 2002. Gaussian limits associated with the Poisson–Dirichlet distribution and the Ewens Sampling Formula. Ann. Appl. Probab. 12, 101–124.

Joyce, P., Tavaré, S., 1987. Cycles permutations and the structure of the Yule process with immigration. Stochastic Process. Appl. 25, 89–103.

Nordborg, M., Donnelly, P.J., Joyce, P., 2001. Likelihood simulation methods and inference for a class of non-neutral population genetics models. Genetics 159, 853–867.

Tavaré, S., Ewens, W.J., Joyce, P., 1989. Is knowing the age-order of alleles in a sample useful in testing for selective neutrality? Genetics 122, 705–711.

Watterson, G.A., 1974. The sampling theory of selectively neutral alleles. Adv. Appl. Probab. 6, 463–488.

Winkler, P., 2003. Mathematical Mind-Benders. A K Peters/CRC Press..