Combinatorics, Probability and Computing (2007) 16, 435–444. © 2006 Cambridge University Press
doi:10.1017/S0963548306008212 First published online 3 November 2006 Printed in the United Kingdom

On a Class of Non-Regenerative Sampling Distributions

MARTIN MÖHLE

Mathematical Institute, University of Tübingen, Auf der Morgenstelle 10, 72076 Tübingen, Germany
(e-mail: martin.moehle@uni-tuebingen.de)

Received 26 April 2006; *revised 13 September 2006*

We show that the sampling formula induced from a Λ-coalescent process with multiple collisions is regenerative if and only if the measure Λ is either concentrated in 0 (Kingman case) or concentrated in 1 (star-shaped case). The Ewens sampling formula is the only sampling formula in this class which also belongs to Pitman's two-parameter family of sampling distributions.

1. Introduction

We start by recalling some basic properties of regenerative composition structures going back to Gnedin [4, 5] and Gnedin and Pitman [6, 7]. A *composition* of an integer $n \in \mathbb{N} := \{1, 2, \dots\}$ is a vector $\mathbf{n} = (n_1, \dots, n_k)$ of positive integers with sum $\sum_{j=1}^{k} n_j = n$. The component n_j is called the jth part of the composition. A *random composition* of n is a random variable C_n with values in the set of all 2^{n-1} compositions of n. A *composition structure* is a sequence $(C_n)_{n=1,2,\dots}$ of random compositions of n, which is consistent, *i.e.*, if n balls are distributed into an ordered series of boxes according to C_n, then C_{n-1} is obtained by disregarding one of the balls picked uniformly at random, and deleting an empty box in case one is created. A composition structure $(C_n)_n$ is called *regenerative* if, for all $1 \leqslant m < n$, given the first part F_n of C_n is m, the remaining composition of $n - m$ is distributed like C_{n-m}.

For a random composition C_n and for a composition $\mathbf{n} = (n_1, \dots, n_k)$ of n write $c(\mathbf{n}) := P(C_n = \mathbf{n})$ for convenience. The function c is called the *composition probability function*. The corresponding *symmetric composition probability function* c^* is defined via

$$c^*(\mathbf{n}) := \frac{1}{k!} \sum_{\pi} c(n_{\pi 1}, \dots, n_{\pi k}), \tag{1.1}$$

where the sum extends over all permutations π of $\{1, \dots, k\}$. Gnedin and Pitman [7, Proposition 3.1] showed that a composition structure $(C_n)_n$ is regenerative if and only if,

https://doi.org/10.1017/S0963548306008212 Published online by Cambridge University Press

for all n and all compositions $\mathbf{n} = (n_1, \ldots, n_k)$ of n, we have the product formula

$$c(\mathbf{n}) = \prod_{j=1}^{k} q(N_j, n_j), \tag{1.2}$$

where $N_j := n_j + \cdots + n_k$, $j \in \{1, \ldots, k\}$, and $q(n,m) := P(F_n = m)$, $1 \leqslant m \leqslant n$, *i.e.*, $q(n,.)$ is the distribution of the first part F_n of C_n. The matrix $(q(n,m))_{n \geqslant 1, 1 \leqslant m \leqslant n}$ is called the *decrement matrix*. For more information on regenerative composition structures, in particular on their asymptotics for large n, we refer to Gnedin, Pitman and Yor [8, 9]. We present two simple, but important examples. For $x \in \mathbb{R}$ and for a non-negative integer n we use the notation $[x]_0 := 1$ and $[x]_n := x(x+1)\cdots(x+n-1)$ for the rising factorial.

Example (Ewens composition structure). Fix $\theta \geqslant 0$. It is well known that

$$c(\mathbf{n}) := \frac{\theta^{k-1} n!}{[\theta+1]_{n-1} N_1 \cdots N_k} \tag{1.3}$$

defines a regenerative composition structure with decrement matrix

$$q(n,m) = \binom{n}{m} \frac{[\theta]_{n-m} m!}{[\theta+1]_{n-1} n} \tag{1.4}$$

and symmetric composition probability function

$$c^*(\mathbf{n}) = \frac{\theta^{k-1} n!}{[\theta+1]_{n-1} k! n_1 \cdots n_k}. \tag{1.5}$$

For more information on the Ewens composition structure and its relation to the celebrated Ewens sampling formula (ESF) we refer to Donnelly and Joyce [2], Ewens and Tavaré [3], Gnedin [4, Section 2] and Gnedin and Pitman [7, Section 4, Example 2].

Example (hook composition structure). A composition $\mathbf{n}$ is called a *hook* if it has at most one part larger than 1. Gnedin and Pitman [6, Section 6] showed that all regenerative hook composition structures are of the following form. Fix a constant $r \in [0, \infty]$. For $n \in \mathbb{N} \setminus \{1\}$ define

$$q(n,1) := \frac{nr}{1+nr}, \qquad q(n,n) := \frac{1}{1+nr} \tag{1.6}$$

and $q(n,m) := 0$ for $1 < m < n$. In agreement with (1.2) define, for $1 \leqslant k < n$,

$$\begin{aligned} c(\underbrace{1, \ldots, 1}_{(k-1)\text{-times}}, n-k+1) &:= q(n,1)\, q(n-1,1) \cdots q(n-k+2,1)\, q(n-k+1, n-k+1) \\ &= \frac{1}{1+(n-k+1)r} \prod_{i=n-k+2}^{n} \frac{ir}{1+ir} = \frac{n! r^{k-1}}{(n-k+1)!} \prod_{i=n-k+1}^{n} \frac{1}{1+ir}. \end{aligned} \tag{1.7}$$

For all other compositions $\mathbf{n}$ define $c(\mathbf{n}) := 0$. For $n \in \mathbb{N}$ let C_n be a random composition with composition probability function c. It is obvious that $(C_n)_n$ is a composition structure, which by construction is regenerative with decrement matrix q. The symmetric composition

https://doi.org/10.1017/S0963548306008212 Published online by Cambridge University Press

probability function c^* satisfies

$$c^*(\mathbf{n}) = \begin{cases} \prod_{i=2}^{n} q(i,1) & \text{if } b(\mathbf{n}) = 0, \\ \frac{q(n-k+1,n-k+1)}{k} \prod_{i=n-k+2}^{n} q(i,1) & \text{if } b(\mathbf{n}) = 1, \\ 0 & \text{if } b(\mathbf{n}) \geqslant 2, \end{cases} \tag{1.8}$$

where $b(\mathbf{n}) := \#\{j \mid n_j > 1\}$ denotes the number of $j \in \{1,\dots,k\}$ with $n_j > 1$. For more examples on regenerative composition structures see Gnedin [4].

The aim of this article is not to extend the theory on regenerative composition structures or to add a new regenerative structure to the literature. We rather focus on an important class of sampling distributions related to coalescent processes. Surprisingly, all these sampling distributions turn out to be non-regenerative (in the sense of Definition 1), except for some well-known boundary cases which are related to the Ewens composition structure and the hook composition structure. In Section 2 this class of sampling distributions is introduced in terms of a recursion going back to Möhle [13]. The main result (Theorem 3.1) is presented in Section 3.

2. Sampling formulas induced from coalescent processes

Let $\mathscr{E}$ denote the set of all equivalence relations on $\mathbb{N}$. For $n \in \mathbb{N}$ let $\varrho_n : \mathscr{E} \to \mathscr{E}_n$, $\varrho_n(\xi) := \{(i,j) \mid 1 \leqslant i,j \leqslant n, (i,j) \in \xi\}$, denote the natural restriction to the (finite) set $\mathscr{E}_n$ of all equivalence relations on $\{1,\dots,n\}$. The set $\mathscr{E}$ is equipped with the metric $d(\xi,\eta) := \inf\{2^{-n} \mid n \in \mathbb{N}, \varrho_n(\xi) = \varrho_n(\eta)\}$, $\xi,\eta \in \mathscr{E}$, and the associated Borel-σ-algebra generated by the open subsets of $\mathscr{E}$.

Let Λ be a finite measure on $[0,1]$. The standard Λ-coalescent, independently studied by Pitman [15] and Sagitov [16], is a Markovian process $R = (R_t)_{t \geqslant 0}$ with state space $\mathscr{E}$ and initial state $R_0 = \Delta := \{(i,i) \mid i \in \mathbb{N}\}$, the diagonal relation having only blocks of size 1 (singletons). The probabilistic structure of R depends on the measure Λ as follows. For each $n \in \mathbb{N}$, the restricted process $(\varrho_n R_t)_{t \geqslant 0}$ is Markovian with state space $\mathscr{E}_n$ and infinitesimal rates

$$q_{\xi\eta} = \begin{cases} \int_{[0,1]} \frac{1-(1-x)^{b-1}(1-x+bx)}{x^2} \Lambda(dx) & \text{if } \xi = \eta, \\ \int_{[0,1]} x^{b-a-1}(1-x)^{a-1} \Lambda(dx) & \text{if } \xi \prec \eta, \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

where $a := |\eta|$ and $b := |\xi|$ are the number of blocks of the equivalence relations $\eta \in \mathscr{E}_n$ and $\xi \in \mathscr{E}_n$, and $\xi \prec \eta$ means (by definition) that exactly $b - a + 1$ equivalence classes of ξ merge together to form one class of η, while all the other $a - 1$ classes of ξ remain unchanged. If $\Lambda = \delta_0$ is the Dirac measure concentrated in 0, then the Λ-coalescent R is Kingman's coalescent (Kingman [11, 12]), which allows only for binary mergers of ancestral lines. If $\Lambda = U$ is uniformly distributed on $[0,1]$, then R is the Bolthausen–Sznitman coalescent (Bolthausen and Sznitman [1]). If Λ is concentrated in 1, then the process R is called a star-shaped coalescent. From (2.1) it follows that the block counting process $D = (D_t)_{t \geqslant 0}$ is a death process with infinitesimal rates

$$g_{nk} = \binom{n}{k-1} \int_{[0,1]} x^{n-k-1}(1-x)^{k-1} \Lambda(dx) \tag{2.2}$$

https://doi.org/10.1017/S0963548306008212 Published online by Cambridge University Press

($n, k \in \mathbb{N}$ with $k < n$) and total rates

$$g_n = \sum_{k=1}^{n-1} g_{nk} = \int_{[0,1]} \frac{1-(1-x)^{n-1}(1-x+nx)}{x^2}\,\Lambda(dx), \quad n \in \mathbb{N}. \tag{2.3}$$

The differences $h_n := g_n - g_{n-1}$, $n \geqslant 2$, therefore satisfy

$$h_n = (n-1)\int_{[0,1]} (1-x)^{n-2}\,\Lambda(dx), \quad n \geqslant 2. \tag{2.4}$$

Note that (2.2) and (2.4) imply that the rates g_{nk}, $1 \leqslant k < n$, are determined by the total rates $g_1, g_2, \ldots$ via

$$g_{nk} = \binom{n}{k-1} \sum_{j=k+1}^{n} \frac{(-1)^{k+1-j}}{j-1} \binom{n-k-1}{n-j} h_j, \quad 1 \leqslant k < n.$$

In particular, $g_{n,n-1} = nh_n/2$, $n \geqslant 2$, a formula which will be useful later.

In population genetics the ancestry of n genes is often modelled by the restricted coalescent process $(\varrho_n R_t)_{t \geqslant 0}$. By definition, two genes i and j have the same ancestor at time t in the past if and only if $(i, j) \in \varrho_n R_t$. Mutations are superimposed on the genealogical tree $(\varrho_n R_t)_{t \geqslant 0}$ as follows. Conditional on the tree, mutations occur independently of the tree at the points of a homogeneous Poisson process with rate $r > 0$ acting along each branch of the tree. Usually, the infinitely-many-alleles model is assumed, *i.e.*, each mutation leads to a new type (allele) never seen before.

Assume there are $k \in \{1, \ldots, n\}$ types among the n genes and assume that you label these k types randomly, *i.e.*, in an exchangeable order. Let $p(\mathbf{n})$ be the probability to see a specific allele configuration $\mathbf{n} = (n_1, \ldots, n_k)$, *i.e.*, to see n_j genes of (randomly labelled) type j, $j \in \{1, \ldots, k\}$. By definition, the probability $p(\mathbf{n})$ is symmetric with respect to the k coordinates $n_1, \ldots, n_k$. Möhle [13, Theorem 3.1] showed that the sampling probabilities $p(\mathbf{n})$ satisfy the recursion $p(1) = 1$ and

$$p(\mathbf{n}) = \frac{nr}{g_n + nr} \sum_{\substack{j=1 \\ n_j = 1}}^{k} \frac{1}{k}\, p(\tilde{\mathbf{n}}_j) + \sum_{i=1}^{n-1} \frac{g_{n,n-i}}{g_n + nr} \sum_{\substack{j=1 \\ n_j > i}}^{k} \frac{n_j - i}{n-i}\, p(\mathbf{n} - i\mathbf{e}_j) \tag{2.5}$$

for any composition $\mathbf{n} = (n_1, \ldots, n_k)$ of n with $n = \sum_{j=1}^{n} n_j \geqslant 2$, where $\mathbf{e}_j = \mathbf{e}_j^{(k)}$ denotes the jth unit vector in $\mathbb{R}^k$, and $\tilde{\mathbf{n}}_j := (n_1, \ldots, n_{j-1}, n_{j+1}, \ldots, n_k)$.

For $l \in \mathbb{N}$ let $a_l := \#\{j \in \{1, \ldots, k\} \mid n_j = l\}$ denote the number of indices $j \in \{1, \ldots, k\}$ with $n_j = l$. In terms of the probabilities $q(\mathbf{a}) := k!/(n_1! \cdots n_k!)p(\mathbf{n})$, $\mathbf{a} = (a_1, a_2, \ldots)$, the recursion (2.5) becomes $q(\mathbf{e}_1) = 1$ and

$$(g_n + nr)q(\mathbf{a}) = nrq(\mathbf{a} - \mathbf{e}_1) + \sum_{i=1}^{n-1} \frac{g_{ni}}{i} \sum_{j=1}^{i} j(a_j + 1)q(\mathbf{a} + \mathbf{e}_j - \mathbf{e}_{n-i+j}), \tag{2.6}$$

for $n = \sum_i i a_i > 1$, with the convention $q(\mathbf{a}) := 0$ if at least one of the entries $a_1, a_2, \ldots$ is negative. Here $\mathbf{e}_j$ denotes the jth unit vector in $\mathbb{R}^\infty$. In terms of the probability generating functions $f_n(s^{(n)}) := \sum_{\mathbf{a}} q(\mathbf{a}) s_1^{a_1} \cdots s_n^{a_n}$, $n \in \mathbb{N}$, $s^{(n)} := (s_1, \ldots, s_n) \in \mathbb{R}^n$, where the

https://doi.org/10.1017/S0963548306008212 Published online by Cambridge University Press

sum extends over all $\mathbf{a} = (a_1, a_2, \dots)$ with $\sum_i i a_i = n$, the recursion (2.6) is equivalent to

$$(g_n + nr) f_n(s^{(n)}) = nr s_1 f_{n-1}(s^{(n-1)}) + \sum_{i=1}^{n-1} \frac{g_{ni}}{i} \sum_{j=1}^{i} j s_{n-i+j} \frac{\partial}{\partial s_j} f_i(s^{(i)}), \quad n > 1,$$

i.e., $f_1(s_1) = s_1$, $f_2(s_1, s_2) = (2rs_1^2 + g_2 s_2)/(g_2 + 2r)$ and so on.

Solutions of the recursion (2.5) or, equivalently, (2.6) are only known for special cases. For the Kingman coalescent ($\Lambda = \delta_0$), the solution of the recursion (2.5) is the symmetric composition probability function $p = c^*$ in (1.5) of the Ewens composition structure with parameter $\theta = 2r$. For the star-shaped tree ($\Lambda = \delta_1$), the solution p (see Möhle [13, Section 4]) is the symmetric composition probability function c^* in (1.8) of the hook composition structure. Thus, for both of these cases, there exists a regenerative composition structure $(C_n)_n$ such that the corresponding symmetric composition probability function c^* coincides with p. The following definition is hence natural.

Definition 1. Let p be any real-valued function on the set of all compositions. Call p *regenerative* if there exists a regenerative composition structure $(C_n)_n$ such that the associated symmetric composition probability function c^* (defined via (1.1)) coincides with p, *i.e.*, $c^*(\mathbf{n}) = p(\mathbf{n})$ for all compositions $\mathbf{n}$.

Remark. The definition implies that only symmetric functions p can be regenerative.

By the above discussion it is tempting to conjecture, that all the sampling formulas based on the recursion (2.5) are regenerative in the sense of Definition 1. We verify in the following section the contrary (Theorem 3.1), which explains the complexity of the sampling probabilities p, and as well the fact that not much is known about p, except for the two boundary cases when the measure Λ is either concentrated in 0 or concentrated in 1.

3. Properties of sampling distributions induced from coalescents

For arbitrary finite measure Λ, the recursion (2.5) seems to be difficult to solve for all compositions $\mathbf{n}$. However, for certain compositions $\mathbf{n} = (n_1, \dots, n_k)$ of n, namely for those satisfying $k = n$ or $k = n - 1$, solutions can be derived easily. Define $\mathbf{v}^{(n)} := \sum_{i=1}^{n} \mathbf{e}_i^{(n)} = (1, \dots, 1) \in \mathbb{R}^n$ for convenience. For $k = n$, the recursion (2.5) reduces to $p(\mathbf{v}^{(n)}) = nr\, p(\mathbf{v}^{(n-1)})/(g_n + nr)$, $n \geqslant 2$, with solution

$$p(\underbrace{1, \dots, 1}_{n\text{-times}}) = p(\mathbf{v}^{(n)}) = \prod_{i=2}^{n} \frac{ir}{g_i + ir}. \tag{3.1}$$

Now consider $\mathbf{v}_i^{(k)} := \mathbf{v}^{(k)} + \mathbf{e}_i^{(k)}$, *i.e.*, the ith component of $\mathbf{v}_i^{(k)}$ is equal to 2 and all other components of $\mathbf{v}_i^{(k)}$ are equal to 1. Note that the sum of all components of $\mathbf{v}_i^{(k)}$ is equal to $n := k + 1$. The recursion (2.5) for the composition $\mathbf{n} := \mathbf{v}_1^{(k)}$ reads

$$p(\mathbf{v}_1^{(k)}) = \frac{nr}{g_n + nr} \frac{k-1}{k} p(\mathbf{v}_1^{(k-1)}) + \frac{g_{n,n-1}}{g_n + nr} \frac{1}{n-1} p(\mathbf{v}^{(k)}).$$

https://doi.org/10.1017/S0963548306008212 Published online by Cambridge University Press

From $k = n-1$, $g_{n,n-1} = nh_n/2$ and (3.1) it follows that

$$p(\mathbf{v}_1^{(n-1)}) = \frac{nr}{g_n + nr}\frac{n-2}{n-1}p(\mathbf{v}_1^{(n-2)}) + \frac{n!h_n r^{n-2}}{2(n-1)\prod_{i=2}^{n}(g_i + ir)}.$$

Now verify by induction on n that

$$p(2, \underbrace{1, \ldots, 1}_{(n-2)\text{-times}}) = p(\mathbf{v}_1^{(n-1)}) = \frac{n!g_n r^{n-2}}{2(n-1)\prod_{i=2}^{n}(g_i + ir)}, \qquad n \geqslant 2. \tag{3.2}$$

Thus, based on the solution (3.1) of the recursion (2.5) for $k = n$, we have found the solution (3.2) of the recursion (2.5) for $k = n-1$. Now proceed as follows to derive further solutions. For $i, j \in \{1, \ldots, k\}$ define $\mathbf{v}_{ij}^{(k)} := \mathbf{v}^{(k)} + \mathbf{e}_i^{(k)} + \mathbf{e}_j^{(k)} \in \mathbb{N}^k$. For the composition $\mathbf{n} := \mathbf{v}_{12}^{(k)}$ of $n = k+2$, the recursion (2.5) reads

$$p(\mathbf{v}_{12}^{(k)}) = \frac{nr}{g_n + nr}\frac{k-2}{k}p(\mathbf{v}_{12}^{(k-1)}) + \frac{g_{n,n-1}}{g_n + nr}\frac{2}{n-1}p(\mathbf{v}_1^{(k)}).$$

From $k = n-2$, $g_{n,n-1} = nh_n/2$ and (3.2) it follows that

$$p(\mathbf{v}_{12}^{(n-2)}) = \frac{nr}{g_n + nr}\frac{n-4}{n-2}p(\mathbf{v}_{12}^{(n-3)}) + \frac{n!g_{n-1}h_n r^{n-3}}{2(n-1)(n-2)\prod_{i=2}^{n}(g_i + ir)}. \tag{3.3}$$

Now verify by induction on n that

$$\begin{aligned} p(\mathbf{v}_{12}^{(n-2)}) &= p(2, 2, \underbrace{1, \ldots, 1}_{(n-4)\text{-times}}) \\ &= \frac{n!r^{n-3}}{2(n-2)(n-3)\prod_{i=2}^{n}(g_i + ir)}\sum_{j=3}^{n-1}\frac{(j-2)g_j h_{j+1}}{j}, \quad n \geqslant 4. \end{aligned}$$

The complexity of such formulas for $p(\mathbf{n})$ increases rapidly with the number of components of $\mathbf{n} = (n_1, \ldots, n_k)$ being larger than 1. We now present our main result.

Theorem 3.1. *Let Λ be a finite measure on $[0,1]$ and let $p(\mathbf{n})$ denote the sampling probabilities induced from a Λ-coalescent with mutation rate $r > 0$, i.e., $p(\mathbf{n})$ is defined via the recursion* (2.5). *The following two conditions are equivalent.*

(i) *The sampling distribution p is regenerative in the sense of Definition 1.*
(ii) *$\Lambda((0,1]) = 0$ or $\Lambda([0,1)) = 0$, i.e., the measure Λ is either concentrated in 0 or concentrated in 1.*

Proof. **(ii) $\Rightarrow$ (i)** Without loss of generality, assume $\Lambda \neq 0$. If Λ is concentrated in 0, *i.e.*, $\Lambda = c\delta_0$ for some $0 < c < \infty$, then it is straightforward to verify that the solution p of the recursion (2.5) coincides with the symmetric composition probability function c^* in (1.5) of the Ewens composition structure with parameter $\theta = 2r/c$. If Λ is concentrated in 1, *i.e.*, $\Lambda = c\delta_1$ for some $0 < c < \infty$, then the solution p of the recursion (2.5) coincides with the symmetric composition probability function c^* in (1.8) of a hook composition structure (with r replaced by r/c). Both composition structures are regenerative.

https://doi.org/10.1017/S0963548306008212 Published online by Cambridge University Press

(i) ⇒ (ii) Assume that $\Lambda([0,1)) > 0$. Then, by (2.4), $h_n > 0$ for $n \geqslant 2$. The assumption (i) ensures that there exists a regenerative composition structure $(C_n)_n$ with $c^*(\mathbf{n}) = p(\mathbf{n})$ for all compositions $\mathbf{n}$. As $(C_n)_n$ is regenerative, we have $c(\mathbf{v}^{(n)}) = q(n,1)\, c(\mathbf{v}^{(n-1)})$, and hence

$$q(n,1) = \frac{c(\mathbf{v}^{(n)})}{c(\mathbf{v}^{(n-1)})} = \frac{c^*(\mathbf{v}^{(n)})}{c^*(\mathbf{v}^{(n-1)})} = \frac{p(\mathbf{v}^{(n)})}{p(\mathbf{v}^{(n-1)})} = \frac{nr}{g_n + nr}, \qquad n \in \mathbb{N}, \tag{3.4}$$

where the last equality follows from (3.1). In order to compute $q(n,2)$ consider the composition $\mathbf{v}_1^{(k)}$ of $n = k+1$ defined earlier. We have

$$\begin{aligned} k\, p(\mathbf{v}_1^{(k)}) &= \sum_{i=1}^{k} c(\mathbf{v}_i^{(k)}) = c(\mathbf{v}_1^{(k)}) + \sum_{i=2}^{k} c(\mathbf{v}_i^{(k)}) \\ &= q(n,2)\, c(\mathbf{v}^{(k-1)}) + q(n,1) \sum_{i=1}^{k-1} c(\mathbf{v}_i^{(k-1)}) \\ &= q(n,2)\, p(\mathbf{v}^{(k-1)}) + q(n,1)\,(k-1)\, p(\mathbf{v}_1^{(k-1)}). \end{aligned}$$

Using $k = n-1$, (3.2) and (3.4) yield

$$\begin{aligned} q(n,2)\, p(\mathbf{v}^{(n-2)}) &= (n-1)\, p(\mathbf{v}_1^{(n-1)}) - q(n,1)\,(n-2)\, p(\mathbf{v}_1^{(n-2)}) \\ &= \frac{n!\, g_n r^{n-2}}{2 \prod_{i=2}^{n} (g_i + ir)} - \frac{nr}{g_n + nr} \frac{(n-1)!\, g_{n-1} r^{n-3}}{2 \prod_{i=2}^{n-1} (g_i + ir)} \\ &= \frac{n!\, r^{n-2} h_n}{2 \prod_{i=2}^{n} (g_i + ir)}. \end{aligned}$$

Dividing both sides by

$$p(\mathbf{v}^{(n-2)}) = \prod_{i=2}^{n-2} \frac{ir}{g_i + ir} = \frac{(n-2)!\, r^{n-3}}{\prod_{i=2}^{n-2} (g_i + ir)}$$

yields

$$q(n,2) = \frac{\binom{n}{2} r h_n}{(g_n + nr)(g_{n-1} + (n-1)r)}, \qquad n \geqslant 2. \tag{3.5}$$

For the composition $\mathbf{v}_{ij}^{(k)} := \mathbf{v}^{(k)} + \mathbf{e}_i^{(k)} + \mathbf{e}_j^{(k)} \in \mathbb{N}^k$ of the integer $n = k+2$ we have

$$\begin{aligned} \binom{k}{2} p(\mathbf{v}_{12}^{(k)}) &= \sum_{1 \leqslant i < j \leqslant k} c(\mathbf{v}_{ij}^{(k)}) = \sum_{j=2}^{k} c(\mathbf{v}_{1j}^{(k)}) + \sum_{2 \leqslant i < j \leqslant k} c(\mathbf{v}_{ij}^{(k)}) \\ &= q(n,2) \sum_{i=1}^{k-1} c(\mathbf{v}_i^{(k-1)}) + q(n,1) \sum_{1 \leqslant i < j \leqslant k-1} c(\mathbf{v}_{ij}^{(k-1)}) \\ &= q(n,2)\,(k-1)\, p(\mathbf{v}_1^{(k-1)}) + q(n,1) \binom{k-1}{2} p(\mathbf{v}_{12}^{(k-1)}). \end{aligned}$$

From $k = n-2$ we conclude that

$$p(\mathbf{v}_{12}^{(n-2)}) = q(n,2)\, \frac{2}{n-2}\, p(\mathbf{v}_1^{(n-3)}) + q(n,1)\, \frac{n-4}{n-2}\, p(\mathbf{v}_{12}^{(n-3)}).$$

https://doi.org/10.1017/S0963548306008212 Published online by Cambridge University Press

Plugging in the formulas (3.4) and (3.5) for $q(n,1)$ and $q(n,2)$, and the formula (3.2) for $p(\mathbf{v}_1^{(n-3)})$ yields

$$p(\mathbf{v}_{12}^{(n-2)}) = \frac{n!g_{n-2}h_n r^{n-3}}{2(n-2)(n-3)\prod_{i=2}^{n}(g_i+ir)} + \frac{nr}{g_n+nr}\frac{n-4}{n-2}p(\mathbf{v}_{12}^{(n-3)}).$$

Comparing this recursion with the standard recursion (3.3) yields

$$\frac{n!g_{n-2}h_n r^{n-3}}{2(n-2)(n-3)} = \frac{n!g_{n-1}h_n r^{n-3}}{2(n-1)(n-2)}, \qquad n \geqslant 4.$$

As $h_n > 0$ and $r > 0$ we conclude that $(n-3)g_{n-1} = (n-1)g_{n-2}$ for all $n \geqslant 4$. By induction on n it follows that $g_n = \binom{n}{2}g_2$ for all n, *i.e.*, up to the constant g_2, the total rates are those of the Kingman coalescent. Thus $\Lambda = g_2\delta_0$ is concentrated in zero, *i.e.*, $\Lambda((0,1]) = 0$. □

We finally present some results related to Pitman's two-parameter family of sampling distributions. For $\alpha, \theta \in \mathbb{R}$ such that either $0 \leqslant \alpha < 1$ and $\theta \geqslant -\alpha$ or $\alpha < 0$ and $\theta = -m\alpha$ for some fixed non-negative integer $m \in \mathbb{N}$, Pitman [14] studied the two-parameter sampling formula

$$p_{\alpha,\theta}(\mathbf{n}) = \frac{n![\theta+\alpha]_{k-1,\alpha}}{[\theta+1]_{n-1}k!}\prod_{j=1}^{k}\frac{[1-\alpha]_{n_j-1}}{n_j!},$$

$\mathbf{n} = (n_1,\dots,n_k) \in \mathbb{N}^k$ with $n := n_1 + \cdots + n_k$, where $[\theta]_{k,\alpha} := \prod_{i=0}^{k-1}(\theta + j\alpha)$. The case $\alpha < 0$ was also studied by Keener, Rothman and Starr [10] with $A = -\alpha$. The following proposition shows that the Ewens sampling formula ($\alpha = 0$) is the only case where Pitman's two-parameter sampling distribution coincides with a sampling distribution induced from a coalescent with multiple collisions.

Proposition 3.2. *Let Λ be a finite measure on $[0,1]$ and let $p(\mathbf{n})$ be the sampling probabilities given by the recursion* (2.5). *If $p_{\alpha,\theta}(\mathbf{n}) = p(\mathbf{n})$ for all $\mathbf{n} = (n_1,\dots,n_k) \in \mathbb{N}^k$, then $\alpha = 0$.* □

Remark. For the sub-range of parameters $0 \leqslant \alpha < 1$ and $\theta \geqslant 0$, Proposition 3.2 follows directly from the previous theorem, as in this case $p_{\alpha,\theta}$ is regenerative in the sense of Definition 1 with decrement matrix entries

$$q(n,n) = \frac{[1-\alpha]_{n-1}}{[\theta+1]_{n-1}} \quad \text{and} \quad q(n,m) = \binom{n}{m}\frac{[1-\alpha]_{m-1}}{[\theta+n-m]_m}\frac{(n-m)\alpha+m\theta}{n}$$

for $1 \leqslant m < n$. See Gnedin and Pitman [7, Section 8] for more details.

Proof. Define $a_n := p(\mathbf{v}^{(n)}) = \prod_{i=2}^{n}(ir/(g_i+ir))$ and $b_n := p_{\alpha,\theta}(\mathbf{v}^{(n)}) = [\theta+1]_{n-1,\alpha}/[\theta+1]_{n-1}$. By assumption, $a_n = b_n$ for all n, *i.e.*,

$$\frac{nr}{g_n+nr} = \frac{a_n}{a_{n-1}} = \frac{b_n}{b_{n-1}} = \frac{\theta+(n-1)\alpha}{\theta+(n-1)}$$

or equivalently

$$\frac{g_n}{nr} = \frac{(n-1)(1-\alpha)}{\theta+(n-1)\alpha}.$$

https://doi.org/10.1017/S0963548306008212 Published online by Cambridge University Press

The recursion (2.5) implies that

$$p(2,2) = \frac{g_{43}}{g_4 + 4r}\frac{2}{3}p(2,1).$$

From $g_{43} = 2h_4 = 2(g_4 - g_3)$ we conclude that

$$\frac{4}{3}\frac{g_4 - g_3}{g_4 + 4r} = \frac{p(2,2)}{p(2,1)} = \frac{p_{\alpha,\theta}(2,2)}{p_{\alpha,\theta}(2,1)} = \frac{2(1-\alpha)}{\theta + 3},$$

where the last equality follows from Pitman's sampling formula. Thus

$$\frac{(1-\alpha)(g_4 + 4r)}{\theta + 3} = \frac{2}{3}(g_4 - g_3) = \frac{2}{3}g_4 - \frac{4r(1-\alpha)}{\theta + 2\alpha}.$$

Division by $4r$ leads to

$$\frac{1-\alpha}{\theta + 3}\frac{g_4 + 4r}{4r} = \frac{2}{3}\frac{g_4}{4r} - \frac{1-\alpha}{\theta + 2\alpha}.$$

Plugging in $(g_4 + 4r)/(4r) = (\theta + 3)/(\theta + 3\alpha)$ and $g_4/4r = 3(1-\alpha)/(\theta + 3\alpha)$, we see that

$$\frac{1-\alpha}{\theta + 3\alpha} = \frac{2}{3}\frac{3(1-\alpha)}{\theta + 3\alpha} - \frac{1-\alpha}{\theta + 2\alpha},$$

or equivalently

$$\frac{1}{\theta + 3\alpha} = \frac{2}{\theta + 3\alpha} - \frac{1}{\theta + 2\alpha},$$

which finally yields $\alpha = 0$. □

Final remark. We have studied properties of sampling formulas induced from coalescent processes with multiple collisions. There exists a wider class of sampling formulas (see Möhle [13]), where the underlying coalescent process allows for simultaneous multiple collisions of ancestral lineages. These coalescent processes are characterized in terms of a sequence of measures Λ_n on the simplex $\Delta_n := \{x = (x_1, \ldots, x_n) \in [0,1]^n \mid x_1 + \cdots + x_n \leqslant 1\}$, $n \in \mathbb{N}$. Similar methods as presented in this article should show that all these sampling distributions are non-regenerative in the sense of Definition 1, except for the case when Λ_j is the zero-measure for all $j \geqslant 2$ and Λ_1 is either concentrated in 0 or concentrated in 1.

References

[1] Bolthausen, E. and Sznitman, A.-S. (1998) On Ruelle's probability cascades and an abstract cavity method. *Comm. Math. Phys.* **197** 247–276.

[2] Donnelly, P. and Joyce, P. (1991) Consistent ordered sampling distributions: Characterization and convergence. *Adv. Appl. Probab.* **23** 229–258.

[3] Ewens, W. J. and Tavaré, S. (1995) The Ewens sampling formula. In *Multivariate Discrete Distributions* (N. S. Johnson *et al.*, eds), Wiley, New York.

[4] Gnedin, A. (2004) Three sampling formulas. *Combin. Probab. Comput.* **13** 185–193.

[5] Gnedin, A. (2004) The Bernoulli sieve. *Bernoulli* **10** 79–96.

[6] Gnedin, A. and Pitman, J. (2005) Regenerative composition structures. *Ann. Probab.* **33** 445–479.

[7] Gnedin, A. and Pitman, J. (2005) Regenerative partition structures. *Electron. J. Combin.* **11** # 12.

https://doi.org/10.1017/S0963548306008212 Published online by Cambridge University Press

[8] Gnedin, A., Pitman, J. and Yor, M. (2006) Asymptotic laws for regenerative compositions: Gamma subordinators and the like. *Probab. Theory Rel. Fields* **135** 576–602.
[9] Gnedin, A., Pitman, J. and Yor, M. (2006) Asymptotic laws for compositions derived from transformed subordinators. *Ann. Probab.* **34** 468–492.
[10] Keener, R., Rothman, E. and Starr, N. (1987) Distributions on partitions. *Ann. Stat.* **15** 1466–1481.
[11] Kingman, J. F. C. (1982) The coalescent. *Stoch. Process. Appl.* **13** 235–248.
[12] Kingman, J. F. C. (2000) Origins of the coalescent: 1974–1982. *Genetics* **156** 1461–1463.
[13] Möhle, M. (2006) On sampling distributions for coalescent processes with simultaneous multiple collisions. *Bernoulli* **12** 35–53.
[14] Pitman, J. (1995) Exchangeable and partially exchangeable random partitions. *Probab. Theory Rel. Fields* **102** 145–158.
[15] Pitman, J. (1999) Coalescents with multiple collisions. *Ann. Probab.* **27** 1870–1902.
[16] Sagitov, S. (1999) The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.* **36** 1116–1125.

https://doi.org/10.1017/S0963548306008212 Published online by Cambridge University Press