Bernoulli **12**(1), 2006, 35–53

On sampling distributions for coalescent processes with simultaneous multiple collisions

M. MÖHLE

Mathematisches Institut, Eberhard Karls Universität Tübingen, Auf der Morgenstelle 10, 72076 Tübingen, Germany. E-mail: martin.moehle@uni-tuebingen.de

Recursions for a class of sampling distributions of allele configurations are derived for the situation where the genealogy of the underlying population is modelled by a coalescent process with simultaneous multiple collisions of ancestral lineages. These recursions describe a new family of partition structures in terms of the composition probability function, parametrized by the infinitesimal rates of the coalescent process. For the Kingman coalescent process with only binary mergers of ancestral lines, the recursion reduces to that known for the classical Ewens sampling distribution. We solve the recursion for the star-shaped coalescent. The asymptotic behaviour of the number K_n of alleles (types) for large sample size n is studied, in particular for the star-shaped coalescent and the Bolthausen–Sznitman coalescent.

Keywords: Bolthausen–Sznitman coalescent; composition probability function; Ewens sampling formula; mutation rate; neutral infinite alleles model; partition structure; sampling distribution; simultaneous multiple collisions; star-shaped tree

1. Introduction

In population genetics the ancestry of a sample of n genes, taken from a large population, is often modelled by a continuous-time stochastic process known as the n-coalescent (Kingman 1982a; 1982b; 1982c). Mutations are superimposed on the genealogical tree of the sample as follows. Conditional on the tree, mutations occur independently of the tree at the points of a homogeneous Poisson process with rate $r = \theta/2 > 0$ acting along each branch of the tree. Usually, the infinitely-many-alleles model is assumed, that is, each mutation leads to a new type (allele) never seen before in the population. The special choice $r = \theta/2$ for the rate of the Poisson process has historical reasons. With this choice the right-hand sides in the distributions (1) and (2) below have a quite simple form.

Assume that there are $k \in \{1, \ldots, n\}$ types in the sample. Ewens' celebrated sampling distribution states that, if you label these k types randomly, that is, in an exchangeable order, you will see a specific non-ordered allele configuration $\mathbf{n} = (n_1, \ldots, n_k)$, that is, n_j genes of (randomly labelled) type j, $j \in \{1, \ldots, k\}$, with probability

1350–7265 © 2006 ISI/BS

$$p(\mathbf{n}) = \frac{\theta^k \, n!}{[\theta]_n \, k! \, n_1 \cdots n_k}, \tag{1}$$

where $[\theta]_n := \theta(\theta+1)\cdots(\theta+n-1)$. Obviously, (1) defines an exchangeable distribution on the set $S_n := \{\mathbf{n} = (n_1, \ldots, n_k) \in \mathbb{N}^k \mid 1 \leqslant k \leqslant n, n_1 + \cdots + n_k = n\}$ of all non-ordered allele configurations, as

$$\sum_{\mathbf{n}\in S_n} p(\mathbf{n}) = \frac{n!}{[\theta]_n} \sum_{k=1}^{n} \frac{\theta^k}{k!} \sum_{\substack{n_1,\ldots,n_k\in\mathbb{N}\\ n_1+\cdots+n_k=n}} \frac{1}{n_1 \cdots n_k} = \frac{1}{[\theta]_n} \sum_{k=1}^{n} \theta^k s(n, k) = 1,$$

where the $s(n, k)$ denote the absolute Stirling numbers of the first kind. Note that $|S_n| = \sum_{k=1}^{n} \binom{n-1}{k-1} = 2^{n-1}$ and that the function p is symmetric, that is, $p(\mathbf{n})$ does not depend on the order of the entries $n_1, \ldots, n_k$. In the terminology of Gnedin and Pitman (2005, Section 2), the function p in (1) is a special example of a *composition probability function* (CPF), which is associated with the composition structure obtained by putting the components of a partition structure in exchangeable random order. The corresponding *exchangeable partition probability function* (EPPF) introduced by Pitman (1995; 2002) is the function p in (1) multiplied by $k!\, n_1! \cdots n_k!/n!$. For more information on exchangeable random partitions, see Pitman (2002, Section 2).

For $l \in \mathbb{N}$ let $a_l := \#\{1 \leqslant j \leqslant k \mid n_j = l\}$ denote the number of n_js which are equal to l. Note that $\sum_i a_i = k$ and that $\sum_i i a_i = n$. Multiplication of the right-hand side in (1) by the number $k!/(a_1! \cdots a_n!)$ of reorderings of $(n_1, \ldots, n_k)$ yields the Ewens sampling formula in the classical form

$$q(\mathbf{a}) = q(a_1, a_2, \ldots) = \frac{n!}{[\theta]_n} \prod_{i=1}^{n} \left(\frac{\theta}{i}\right)^{a_i} \frac{1}{a_i!}. \tag{2}$$

Combinatorial proofs of (2) can be found in Griffiths and Lessard (2005). It is straightforward to verify that the sampling probabilities (1) satisfy the recursion $p(1) = 1$ and

$$p(\mathbf{n}) = \frac{\theta}{\theta+n-1} \sum_{\substack{j=1\\ n_j=1}}^{k} \frac{1}{k} p(\tilde{\mathbf{n}}_j) + \frac{n-1}{\theta+n-1} \sum_{\substack{j=1\\ n_j>1}}^{k} \frac{n_j-1}{n-1} p(\mathbf{n}-\mathbf{e}_j) \tag{3}$$

for $\mathbf{n} = (n_1, \ldots, n_k)$ with $n := \sum_{j=1}^{k} n_j \geqslant 2$, where $\tilde{\mathbf{n}}_j := (n_1, \ldots, n_{j-1}, n_{j+1}, \ldots, n_k)$ and $\mathbf{e}_j$ denotes the jth unit vector in $\mathbb{R}^k$. Similar recursions for models with a finite number of types have been used by De Iorio and Griffiths (2004, Section 3). The recursion (3) can be either verified directly from (1), or deduced by looking at the first event which happens backwards in time. As we look back into the past at the history of the genes in the sample, we will eventually see either a mutation or a coalescence. Under the Kingman coalescent, the time W_n back to the first mutation is exponentially distributed with parameter nr. The time T_n back to the first coalescence is independent of W_n and exponentially distributed with parameter $g_n := n(n-1)/2$. Therefore, the first event backwards in time is a mutation with probability $P(W_n < T_n) = nr/(g_n + nr) = \theta/(\theta + n - 1)$, and a coalescence with the complementary probability $P(W_n > T_n) = (n-1)/(\theta + n - 1)$. These two probabilities are the fractions in front of the sums on the right-hand side in (3). Given that the first event backwards in time is a mutation, one of the (randomly labelled) types which appear only once in the sample is produced by this mutation, which explains the first sum

$$\frac{a_1}{k}\,p(\tilde{\mathbf{n}}) = \sum_{\substack{j=1\\ n_j=1}}^{k} \frac{p(\tilde{\mathbf{n}}_j)}{k}$$

on the right-hand side in (3). Here $\tilde{\mathbf{n}}$ is derived from $\mathbf{n}$ by removing one of its entries which are equal to 1. Note that $p(\tilde{\mathbf{n}}_j)$ does not depend on j as long as $n_j = 1$. If the first event backwards in time is a coalescence, then two genes merge into a singleton. Such a merger occurs among genes of type j, $j \in \{1, \ldots, k\}$, with probability $(n_j - 1)/(n - 1)$, provided the constraint $n_j > 1$ is satisfied, which explains the second sum on the right-hand side in (3).

Arguments of this type are often helpful in the context of coalescent theory to derive useful recursions (Griffiths and Tavaré 1996, Section 3). For example, similar arguments show that the recursion in terms of the probabilities $q(\mathbf{a})$ in (2) is given by $q(1, 0, 0, \ldots) = 1$ and

$$q(\mathbf{a}) = \frac{\theta}{\theta + n - 1}\, q(\mathbf{a} - \mathbf{e}_1) + \frac{n-1}{\theta + n - 1} \sum_{i=1}^{n-1} \frac{i(a_i - 1)}{n-1} q(\mathbf{a} + \mathbf{e}_i - \mathrm{e}_{i+1}) \tag{4}$$

for $n = \sum_i i a_i \geqslant 2$, with $q(\mathbf{a}) = 0$ if at least one of the entries $a_1, a_2, \ldots$ is negative. Note that in (4) the unit vectors are in $\mathbb{R}^\infty$. Of course, the recursion for q can be also deduced directly from (2). It also follows from the recursion for p via multiplication of (3) by $k!/(a_1! \cdots a_n!)$.

It is known (Pitman 1999; Sagitov 1999; Möhle and Sagitov 2001) that, in addition to the Kingman coalescent, a richer class of coalescent processes plays an important role in population genetics. While the Kingman coalescent has only binary mergers of ancestral lineages, these more general coalescent processes allow for multiple and even simultaneous multiple mergers of ancestral lineages. Figure 1 shows a multiple collision and a simultaneous multiple collision for a sample of size $n = 8$. A simultaneous multiple collision appears (by definition) if at least two multiple collisions happen at exactly the same time.

In this paper recursions for sampling formulae of allele configurations are presented when the underlying genealogical tree is a coalescent with simultaneous multiple collisions. We

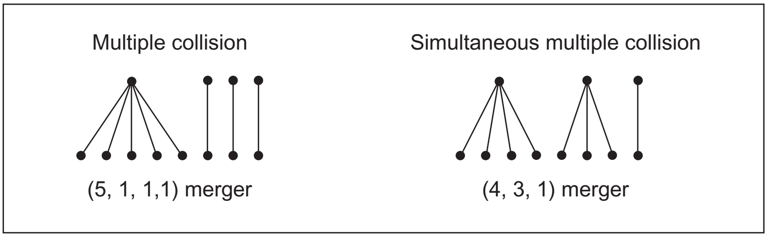

Figure 1. Collisions for a sample of size 8.

start in Section 2 with a brief summary on coalescent processes with simultaneous multiple collisions. In Section 3 sampling recursions are studied when the coalescent allows only for multiple collisions of ancestral lineages. In Section 4 the star-shaped coalescent is studied in detail. Extensions to the general case with simultaneous collisions are presented in Section 5. The paper concludes in Section 6 with examples and numerical studies, in particular for the Bolthausen–Sznitman coalescent.

2. Coalescent processes with simultaneous multiple collisions

Basically, there are two approaches in the literature to coalescent processes with simultaneous multiple collisions, the 'consistent rates approach' of Pitman (1999) and Schweinsberg (2000b), and the 'weak limit approach' of Möhle and Sagitov (2001). The following method essentially coincides with the 'consistent rates approach'. However, we modify some parts of Pitman and Schweinsberg's approach for the sake of simplicity. In the following it is assumed that there exist functions $\phi_j : \mathbb{N}^j \to \mathbb{R}$, $j \in \mathbb{N}$, with the following properties:

(i) For each $j \in \mathbb{N}$, the function ϕ_j is symmetric with respect to the j coordinates, that is, $\phi_j(k_{\pi 1}, \ldots, k_{\pi j}) = \phi_j(k_1, \ldots, k_j)$ for all $k_1, \ldots, k_j \in \mathbb{N}$ and each permutation π of the indices $\{1, \ldots, j\}$.
(ii) $\phi_j(k_1, \ldots, k_j) \geqslant 0$ for all $j, k_1, \ldots, k_j \in \mathbb{N}$, with $k_1 + \ldots + k_j > j$.
(iii) $\sum_{i=1}^{j} \frac{1}{i!} \sum_{k_1,\ldots,k_i \in \mathbb{N}; k_1+\ldots+k_i=j} \frac{\phi_i(k_1, \ldots, k_i)}{k_1! \cdots k_i!} = 0$ for all $j \in \mathbb{N}$.

The existence of such functions ϕ_j is obvious. Note that $\phi_1(1) = 0$ and that property (iii) determines $\phi_j(1, \ldots, 1)$. For $n \in \mathbb{N}$, let $\mathcal{E}_n$ denote the set of all equivalence relations on $\{1, \ldots, n\}$. For $\xi \in \mathcal{E}_n$, let $|\xi|$ denote the number of blocks (equivalence classes) of ξ.

Definition 2.1. *Fix* $n \in \mathbb{N}$. *A time-continuous Markovian chain* $(R_t^{(n)})_{t \geqslant 0}$ *on a probability space* $(\Omega, \mathcal{F}, P)$ *with state space* $\mathcal{E}_n$ *is called n-coalescent with rate functions* $\phi_1, \ldots, \phi_n$, *if* $R_0^{(n)} = \{(i, i) \,|\, 1 \leqslant i \leqslant n\}$ *and if the infinitesimal rates* $q_{\xi\eta} := \lim_{h \searrow 0} h^{-1}(P(R_{t+h}^{(n)} = \eta \,|\, R_t^{(n)} = \xi) - \delta_{\xi\eta})$, $\xi, \eta \in \mathcal{E}_n$, *are given by*

$$q_{\xi\eta} = \begin{cases} \phi_a(b_1, \ldots, b_a), & \textit{if } \xi \subseteq \eta, \\ 0, & \textit{otherwise}, \end{cases} \tag{5}$$

where $a = |\eta|$ *and* $b_1, \ldots, b_a \in \mathbb{N}$ *are the group sizes of merging equivalence classes of* ξ.

Remark. Note that the rates (5) do not depend on n. As the functions ϕ_j are symmetric, $\phi_a(b_1, \ldots, b_a)$ does not depend on the order of the group sizes of merging classes of ξ. Thus the rates (5) are well defined. Assumption (iii) ensures that $\sum_{\eta \in \mathcal{E}_n} q_{\xi\eta} = 0$ for all $\xi \in \mathcal{E}_n$. Thus $Q := (q_{\xi\eta})_{\xi,\eta \in \mathcal{E}_n}$ is a generator matrix. The existence of a Markovian chain with state space $\mathcal{E}_n$ and rates (5) is obvious, as on finite state spaces Markovian chains can be constructed for arbitrary generators.

Definition 2.2. *A family of functions* $\phi_j : \mathbb{N}^j \to \mathbb{R}$, $j \in \mathbb{N}$, *is called natural or consistent, if*

$$\phi_j(k_1, \ldots, k_j) = \phi_{j+1}(k_1, \ldots, k_j, 1) + \sum_{i=1}^{j} \phi_j(k_1, \ldots, k_{i-1}, k_i + 1, k_{i+1}, \ldots, k_j) \tag{6}$$

for all $j, k_1, \ldots, k_j \in \mathbb{N}$.

Remark. Condition (6) coincides with the consistency relation for an EPPF described in Pitman (1995, Proposition 10). Note that (6) and (ii) imply that the functions ϕ_j, $j \in \mathbb{N}$, are monotone in the sense that

$$\phi_j(k_1, \ldots, k_j) \leqslant \phi_l(m_1, \ldots, m_l) \tag{7}$$

whenever $j \geqslant l \geqslant 1$ and $k_1, \ldots, k_j, m_1, \ldots, m_l \in \mathbb{N}$ with $k_1 \geqslant m_1, \ldots, k_l \geqslant m_l$ and $m_1 + \ldots + m_l > l$. We verify (7) by induction on the difference $d := j - l \in \mathbb{N}_0$. The consistency property (6) and (ii) ensure that $\phi_l(m_1, \ldots, m_l) \geqslant \phi_l(m_1, \ldots, m_{i-1}, m_i + 1, m_{i+1}, \ldots, m_l)$ for $i \in \{1, \ldots, l\}$. Iteratively it follows that (7) holds for $j = l$, that is, for $d = 0$. Again using (6) and (7) for $d = 0$, we see that $\phi_l(m_1, \ldots, m_l) \geqslant \phi_{l+1}(m_1, \ldots, m_l, 1) \geqslant \phi_{l+1}(k_1, \ldots, k_{l+1})$, which shows that (7) is valid for $j = l + 1$, that is, for $d = 1$. Finally, applying (7) with $d = 1$ exactly $j - l$ times yields $\phi_l(m_1, \ldots, m_l) \geqslant \phi_{l+1}(k_1, \ldots, k_{l+1}) \geqslant \phi_{l+2}(k_1, \ldots, k_{l+2}) \geqslant \cdots \geqslant \phi_j(k_1, \ldots, k_j)$.

Example. Fix $l \in \mathbb{N}$. Define $\phi_j(k_1, \ldots, k_j) := (l)_j / l^k$ for $j, k_1, \ldots, k_j \in \mathbb{N}$ with $k := k_1 + \ldots + k_j > j$ and $\phi_j(1, \ldots, 1) := -\sum_{i=1}^{j-1} \phi_i(2, 1, \ldots, 1) = -\sum_{i=1}^{j-1} (l)_i / \mathrm{l}^{i+1}$, where $(l)_j := l(l-1)\cdots(l-j+1)$. It is straightforward to verify that this family of functions ϕ_j is natural. We will return to this example later.

Remarks.

1. Assume that the values $\phi_j(k_1, \ldots, k_j)$ are known only for all $k_1, \ldots, k_j \geqslant 2$. Then (6) determines the values $\phi_j(k_1, \ldots, k_j)$ for all $k_1, \ldots, k_j \in \mathbb{N}$ satisfying $k_1 + \ldots + k_j > j$. Moreover, as $\phi_1(1) = 0$, all the values $\phi_j(1, \ldots, 1)$ with $j \geqslant 2$ can be derived from (6). Thus, the values $\phi_j(k_1, \ldots, k_j)$ for $k_1, \ldots, k_j \geqslant 2$ and (6) determine the functions ϕ_j, $j \in \mathbb{N}$, completely.
2. For $m, n \in \mathbb{N}$ with $m \leqslant n$, let $\varrho_{nm} : \mathcal{E}_n \to \mathcal{E}_m$ denote the natural projection. Assume that the family of functions $\{\phi_j\}_{j \in \mathbb{N}}$ is natural. Applying Lemma 3.2.1 in Rosenblatt (1959), it follows that if $(R_t^{(n)})_{t \geqslant 0}$ is a n-coalescent with rate functions $\phi_1, \ldots, \phi_n$, then $(\varrho_{nm} R_t^{(n)})$ is a m-coalescent with rate functions $\phi_1, \ldots, \phi_m$. This property is called 'natural coupling'.

Let $\mathcal{E}$ denote the set of all equivalence relations on $\mathbb{N}$. Each $\xi \in \mathcal{E}$ can be considered as an element $x = (x_{ij})_{i,j \in \mathbb{N}}$ of $X := \prod_{(i,j) \in \mathbb{N}^2} \{0, 1\}$ via the identification $(i, j) \in \xi \Leftrightarrow x_{ij} = 1$. The product topology on X induces the topology of $\mathcal{E}$ (subspace topology). It is well known that $\mathcal{E}$, together with this topology, is Hausdorff, compact, separable and complete. We consider $\mathcal{E}$ with this topology and with the Borel-σ-algebra $\mathcal{B}(\mathcal{E})$ generated from this topology. For $n \in \mathbb{N}$ the natural projection $\varrho_n : \mathcal{E} \to \mathcal{E}_n$, defined via $\varrho_n(\xi) := \{(i, j) \in \xi \,|\, 1 \leqslant i, j \leqslant n\}$, is continuous and hence $\mathcal{B}(\mathcal{E})$-measurable.

Definition 2.3. *A time-continuous Markovian process $R = (R_t)_{t \geqslant 0}$ on a probability space $(\Omega, \mathcal{F}, P)$ with state space $\mathcal{E}$ is called a coalescent with rate functions ϕ_j, $j \in \mathbb{N}$, if for each $n \in \mathbb{N}$ the process $(\varrho_n R_t)_{t \geqslant 0}$ is a n-coalescent with rate functions $\phi_1, \ldots, \phi_n$ in the sense of Definition 2.1.*

Kolmogoroff's extension theorem ensures that such a process R exists if and only if the family of functions $\{\phi_j\}_{j \in \mathbb{N}}$ is natural. The consistency needed to apply Kolmogoroff's extension theorem follows from the natural coupling property, which is satisfied if and only if the family $\{\phi_j\}_{j \in \mathbb{N}}$ is natural. We will henceforth assume that the family $\{\phi_j\}_{j \in \mathbb{N}}$ is natural.

We claim that, for each $j \in \mathbb{N}$, there exists a measure Λ_j on the simplex $\Delta_j := \{(x_1, \ldots, x_j) \in [0, 1]^j \,|\, x_1 + \ldots + x_j \leqslant 1\}$, uniquely determined via its moments

$$\int_{\Delta_j} x_1^{k_1 - 2} \cdots x_j^{k_j - 2} \Lambda_j(\mathrm{d}x_1, \ldots, \mathrm{d}x_j) = \phi_j(k_1, \ldots, k_j), \qquad k_1, \ldots, k_j \geqslant 2. \tag{8}$$

In the following we verify the existence of the measures Λ_j by applying Kingman's representation theorem for exchangeable random equivalence relations and a lemma known from the 'weak limit approach' of Möhle and Sagitov (2001), which results in a simplification compared to the proof of Schweinsberg (2000b). Let $N \geqslant 2$. The N-coalescent $(\varrho_N R_t)_{t \geqslant 0}$ jumps at its first jump time $T_N := \inf\{t > 0 |\, |\varrho_N R_t| < N\}$ to a state $\xi \in \mathcal{E}_N$ with probability $P(\varrho_N R_{T_N} = \xi) = \phi_j(k_1, \ldots, k_j)/g_N$, where $j := |\xi| < N$, $k_1, \ldots, k_j$ are the sizes of the equivalence classes of ξ and $g_N := -\phi_N(1, \ldots, 1)$ denotes the total rate of the N-coalescent. From (5), a backward induction on k yields

$$P(\varrho_k R_{T_N} = \xi) = \frac{\phi_j(k_1, \ldots, k_j)}{g_N} \tag{9}$$

for all $k \in \{1, \ldots, N\}$ and all $\xi \in \mathcal{E}_k$ with $j := |\xi| < k$, where $k_1, \ldots, k_j$ are the sizes of the equivalence classes of ξ. There is another important representation of (9). Let $B := |\varrho_N R_{T_N}|$ denote the number of equivalence classes of $\varrho_N R_{T_N}$ and let $\lambda_1 \geqslant \ldots \geqslant \lambda_B$ be the sizes of these B equivalence classes. Define $\lambda_i := 0$ for $i \in \{B + 1, \ldots, N\}$. Note that $\lambda_i = \lambda_i(N)$ depends on N. As R_{T_N} is an exchangeable random equivalence relation, it follows (Kingman 1982c, equation (3.13)) that

$$P(\varrho_k R_{T_N} = \xi) = \sum_{\substack{r_1, \ldots, r_j = 1 \\ \text{distinct}}}^{N} \frac{\mathrm{E}((\lambda_{r_1})_{k_1} \cdots (\lambda_{r_j})_{k_j})}{(N)_k} = \frac{(N)_j}{(N)_k} \mathrm{E}((\nu_1)_{k_1} \cdots (\nu_j)_{k_j}), \tag{10}$$

where $(\nu_1, \ldots, \nu_N)$ is a random permutation of $(\lambda_1, \ldots, \lambda_N)$. The random variables $\nu_1, \ldots, \nu_N$ can be viewed as the family sizes of an exchangeable discrete population model with fixed population size N introduced by Cannings (1974; 1975). The right-hand sides in (9) and (10) coincide, in particular for $j = 1$ and $k = k_1 = 2$, that is, $c_N :=$

$\mathrm{E}((\nu_1)_2)/(N-1) = \phi_1(2)/g_N$. From $B < N$ we conclude that $P(\nu_1 > 1) > 0$ and hence $c_N > 0$. Moreover, (9) and (10) yield

$$\lim_{N\to\infty} \frac{\mathrm{E}((\nu_1)_{k_1}\cdots(\nu_j)_{k_j})}{N^{k_1+\ldots+k_j-j}c_N} = \frac{\phi_j(k_1, \ldots, k_j)}{\phi_1(2)}$$

for all $j, k_1, \ldots, k_j \in \mathbb{N}$ with $k_1 + \ldots + k_j > j$. Finally, Lemma 3.1. of Möhle and Sagitov (2001) shows that, for each $j \in \mathbb{N}$, there exists a measure Λ_j on the simplex Δ_j, uniquely determined via its moments (8). The measure Λ_j is finite ($\Lambda_j(\Delta_j) = \phi_j(2, \ldots, 2) < \infty$) and symmetric, as ϕ_j is assumed to be symmetric. Note that (7) implies that $\phi_j(2, \ldots, 2) \geqslant \phi_{j+1}(2, \ldots, 2)$, that is, the sequence $(\Lambda_j(\Delta_j))_{j\in\mathbb{N}}$ of total masses decreases. As a natural family $\{\phi_j\}_{j\in\mathbb{N}}$ of rate functions is completely determined via the measures $\Lambda_1, \Lambda_2, \ldots,$ a coalescent process R, as introduced in Definition 2.3, is also called a $(\Lambda_1, \Lambda_2, \ldots)$-coalescent.

Example. For the rate functions presented in the example after Definition 2.2, the corresponding coalescent process allows for up to $l \in \mathbb{N}$ multiple collisions of ancestral lineages simultaneously. The measure Λ_j assigns its total mass $\Lambda_j(\Delta_j) := (l)_j/l^{2j}$ to the single point $(1/l, \ldots, 1/l) \in \mathbb{R}^j$, being the zero measure for $j > l$.

Obviously, a $(b_1, \ldots, b_a)$ merger, as described in Figure 1, occurs with rate

$$\frac{b!\,\phi_a(b_1, \ldots, b_a)}{a!\,b_1!\cdots b_a!},$$

where $b := b_1 + \ldots + b_a$, as exactly $b!/(a!b_1!\cdots b_a!)$ equivalence relations in $\mathcal{E}_b$ correspond to a given $(b_1, \ldots, b_a)$ merger. Summing over all $b_1, \ldots, b_a \in \mathbb{N}$ with $b_1 + \ldots + b_a = b$, it follows that

$$g_{ba} := \frac{b!}{a!} \sum_{\substack{b_1,\ldots,b_a\in\mathbb{N}\\ b_1+\ldots+b_a=b}} \frac{\phi_a(b_1, \ldots, b_a)}{b_1!\cdots b_a!} \tag{11}$$

are the rates of the corresponding death process $D = (D_t)_{t\geqslant 0}$, where $D_t := |R_t|$ counts the number of blocks (equivalence classes) of R_t. Finally, we define the total rates

$$g_b := \sum_{a=1}^{b-1} g_{ba}, \qquad b \in \mathbb{N}. \tag{12}$$

Schweinsberg (2000a; 2000b) provides information about the death process D, especially the question of whether it ‘comes down from infinity’ at time $t = 0+$. From (11) it is obvious that the consistency condition (6) puts certain constraints on the rates of the death process. In the following section it is shown that, if the underlying coalescent process allows only for multiple collisions of ancestral lineages, these constraints can be easily expressed directly in terms of the rates (11).

3. Sampling recursions for coalescent processes with multiple collisions

A coalescent is called a coalescent with multiple collisions if the measures Λ_j, $j \geqslant 2$, are equal to zero. In other words, $\phi_j(k_1, \ldots, k_j) = 0$ whenever at least two of the indices $k_1, \ldots, k_j$ are greater than 1, which is to say that only singleton multiple mergers are allowed to occur with positive probability. A coalescent with multiple collisions is hence completely determined by the measure $\Lambda := \Lambda_1$, and is hence also called a Λ-coalescent. These coalescent processes have been studied independently by Pitman (1999) and Sagitov (1999). The infinitesimal rates $q_{\xi\eta}$ can be expressed in terms of the measure Λ as

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

where $a := |\eta|$, $b := |\xi|$ and $\xi \prec \eta$ means (by definition) that exactly $b_1 = b - a + 1$ equivalence classes of ξ merge together to form one equivalence class of η, while all the other $a - 1$ equivalence classes of ξ remain unchanged. If $\Lambda = \delta_0$ is the Dirac measure concentrated at 0, then the process R is Kingman's coalescent (Kingman 1982a; 1982b; 1982c; 2000), which allows only for binary mergers of ancestral lines. If $\Lambda = U$ is the uniform distribution on [0, 1], then the coalescent process R is the Bolthausen–Sznitman coalescent (Bolthausen and Sznitman 1998). From (5) it follows that the corresponding death process $D = (D_t)_{t \geqslant 0}$ has infinitesimal rates

$$g_{nk} = \binom{n}{k-1}\phi_k(n-k+1, 1, \ldots, 1) = \binom{n}{k-1}\int_{[0,1]} x^{n-k-1}(1-x)^{k-1}\,\Lambda(\mathrm{d}x) \qquad (13)$$

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

$$g_n = \sum_{k=1}^{n-1} g_{nk} = \int_{[0,1]} \frac{1-(1-x)^{n-1}(1-x+nx)}{x^2}\,\Lambda(\mathrm{d}x), \qquad n \in \mathbb{N}. \qquad (14)$$

For coalescent processes with only multiple collisions, the consistency condition (6) reduces to $\phi_k(n-k+1, 1, \ldots, 1) = \phi_{k+1}(n-k+1, 1, \ldots, 1) + \phi_k(n-k+2, 1, \ldots, 1)$, $1 \leqslant k < n$ and $\phi_n(1, \ldots, 1) = \phi_{n+1}(1, \ldots, 1) + n\phi_n(2, 1, \ldots, 1)$, $n \in \mathbb{N}$. Thus, from (13) we conclude that the consistency condition (6) is satisfied if and only if the rates of the block counting process satisfy the constraints

$$g_{nk} = \frac{k}{n+1}\,g_{n+1,k+1} + \frac{n-k+2}{n+1}\,g_{n+1,k}, \qquad 1 \leqslant k < n,\ n \in \mathbb{N}, \qquad (15)$$

and $g_n = g_{n+1} - 2g_{n+1,n}/(n+1)$, $n \in \mathbb{N}$. From these constraints it follows by induction on the difference $d := n - k \in \mathbb{N}$ that the rates g_{nk}, $1 \leqslant k < n$, are determined by the total rates $g_1, g_2, \ldots$ by

$$g_{nk} = \binom{n}{k-1} \sum_{j=k+1}^{n} \frac{(-1)^{k+1-j}}{j-1} \binom{n-k-1}{n-j} (g_j - g_{j-1}), \qquad 1 \leq k < n, \tag{16}$$

i.e. $g_{21} = g_2$, $g_{31} = \frac{3}{2}g_2 - \frac{1}{2}g_3$, $g_{32} = \frac{3}{2}(g_3 - g_2)$, $g_{41} = 2g_2 - \frac{4}{3}g_3 + \frac{1}{3}g_4$, $g_{42} = -2g_2 + \frac{10}{3}g_3 - \frac{4}{3}g_4$, $g_{43} = 2(g_4 - g_3)$ and so on. Also from (14) we conclude by induction on $k \in \{1, \ldots, n-1\}$ that the rates g_{nk}, $1 \leq k < n$, are determined by the moments $g_{n1} = \phi_1(n) = \int_{[0,1]} x^{n-2}\,\Lambda(\mathrm{d}x)$, $n \in \mathbb{N}\backslash\{1\}$, of the measure Λ via

$$g_{nk} = \binom{n}{k-1} (-1)^{n-k+1} \sum_{i=n-k+1}^{n} (-1)^i \binom{k-1}{n-i} g_{i1}, \qquad 1 \leq k < n. \tag{17}$$

Note that $g_{nn} = -\sum_{k=1}^{n-1} g_{nk}$ is then also determined by the moments of Λ. As already mentioned in Section 1, it is assumed that – independently of the underlying genealogical tree – mutations appear at the points of a Poisson process with rate $r = \theta/2$ along all branches of the tree. Each mutation leads to a new type. Such a coalescent process is then called a Λ-coalescent with mutation rate θ. If we take a sample of n genes, it is natural to ask for the probability $p(\mathbf{n})$ that we have sampled a specific non-ordered allele configuration $\mathbf{n} = (n_1, \ldots, n_k)$. We are now able to state our main result, a recursion for these probabilities $p(\mathbf{n})$.

Theorem 3.1. *Let Λ be a finite measure on* $[0, 1]$. *For the Λ-coalescent with mutation rate $\theta > 0$, the sampling probabilities $p(\mathbf{n})$ of non-ordered allele configurations $\mathbf{n} = (n_1, \ldots, n_k)$ satisfy the recursion $p(1) = 1$ and*

$$p(\mathbf{n}) = \frac{nr}{g_n + nr} \sum_{\substack{j=1 \\ n_j=1}}^{k} \frac{1}{k} p(\tilde{\mathbf{n}}_j) + \sum_{i=1}^{n-1} \frac{g_{n,n-i}}{g_n + nr} \sum_{\substack{j=1 \\ n_j>i}}^{k} \frac{n_j - i}{n - i} p(\mathbf{n} - i\mathbf{e}_j) \tag{18}$$

for $\mathbf{n} = (n_1, \ldots, n_k)$ with $n := \sum_{j=1}^{k} n_j \geq 2$, where the rates g_{nk} and g_n are given by (13) *and* (14), *$r = \theta/2$, $\tilde{\mathbf{n}}_j := (n_1, \ldots, n_{j-1}, n_{j+1}, \ldots, n_k)$ and $\mathbf{e}_j$ denotes the jth unit vector in $\mathbb{R}^k$.*

Proof. Fix the sample size n and look back into the past at the history of n sampled genes. The time W_n back to the first mutation is exponentially distributed with parameter nr. The time T_n back to the first coalescence is independent of W_n and exponentially distributed with parameter g_n. Thus, the first event backwards in time is a mutation with probability $P(W_n < T_n) = nr/(g_n + nr)$, and a coalescence with probability $P(T_n > W_n) = g_n/(g_n + nr)$. If the first event backwards in time is a mutation, then exactly the same argument as already used in the derivation of (3) leads to the first sum on the right-hand side in (18). If the first event backwards in time is a coalescence, then $i + 1$ genes will merge into a singleton, $i \in \{1, \ldots, n-1\}$, with probability $g_{n,n-i}/g_n$. Such a merger occurs among genes of type j, $j \in \{1, \ldots, k\}$, with probability $(n_j - i)/(n - i)$, provided the constraint $n_j > i$ is satisfied. Combining all these probabilities, the recursion follows similarly to the recursion (3) known for the Kingman coalescent. □

Remarks.

1. For the Kingman coalescent ($\Lambda = \delta_0$, i.e. $g_n = g_{n,n-1} = n(n-1)/2$ and $g_{ni} = 0$ for $i \in \{1, \ldots, n-2\}$), the recursion (18) reduces to the recursion (3) with solution (1). More generally, if $\Lambda = c\delta_0$ is concentrated at zero with total mass $c \in (0, \infty)$, then $g_n = g_{n,n-1} = n(n-1)c/2$ and the solution for $p(\mathbf{n})$ is given by (1) with θ replaced by θ/c.
2. For $k = 1$, the recursion (18) reduces to $p(n) = (g_n + nr)^{-1}\sum_{i=1}^{n-1} g_{ni}\, p(i)$, $n \geqslant 2$. An induction on n yields the solution

$$p(n) = \sum_{i=2}^{n} \sum_{1=k_1<\cdots<k_i=n} \prod_{j=2}^{i} \frac{g_{k_j,k_{j-1}}}{g_{k_j} + k_j r}, \qquad n \in \mathbb{N}\backslash\{1\}. \tag{19}$$

Note that $p(n)$ is the probability that n randomly sampled genes are identical by descent, or, in other words, of the same type, that is, there is no mutation in either lineage since their most recent common ancestor. The probability $p(n)$ is the $(n-1)$th moment of the so-called structural distribution of the random partition, which in Kingman's representation determines the expected number of frequencies in any interval. See Gnedin and Pitman (2005, Section 9) or Pitman (2002, Section 2.3) for more details. For the Kingman coalescent, it follows that $p(n) = (n-1)!/[\theta+1]_{n-1}$ is the $(n-1)$th moment of the beta$(1, \theta)$ distribution, in agreement with (1) for $k = 1$. For convenience, define $a_n := p(\mathbf{e})$ with $\mathbf{e} := (1, \ldots, 1) \in \mathbb{R}^n$. For $k = n$, the recursion (18) reduces to $a_n = nra_{n-1}/(g_n + nr)$, $n \geqslant 2$, with solution $a_n = \prod_{i=2}^{n}(ir/(g_i + ir))$, $n \in \mathbb{N}$. A solution for $k = n-1$ is presented later in (21).

In general, for $1 < k < n-1$, explicit solutions for $p(\mathbf{n})$ seem to be difficult to find. In principle, the recursion (18) can be solved for any fixed sample size n, but for large n the results become rather complicated. For example, from $p(2) = g_{21}/(g_2 + 2r)$ and $p(1, 1) = 2r/(g_2 + 2r)$ we conclude that, for $n = 3$,

$$p(3) = \frac{g_{32}g_{21}}{(g_3 + 3r)(g_2 + 2r)} + \frac{g_{31}}{g_3 + 3r},$$

$$p(2, 1) = p(1, 2) = \frac{r(g_{32} + \frac{3}{2}g_{21})}{(g_3 + 3r)(g_2 + 2r)}$$

and

$$p(1, 1, 1) = \frac{6r^2}{(g_3 + 3r)(g_2 + 2r)}.$$

For $n = 4$, we derive

$$p(4) = \frac{1}{g_4 + 4r}\left(\frac{g_{43}g_{32}g_{21}}{(g_3 + 3r)(g_2 + 2r)} + \frac{g_{42}g_{21}}{g_2 + 2r} + \frac{g_{43}g_{31}}{g_3 + 3r} + g_{41}\right),$$

$$p(3,\ 1) = p(1,\ 3)$$

$$= \frac{r}{g_4 + 4r}\left(\frac{2g_{32}g_{21} + \frac{2}{3}g_{43}g_{32} + g_{43}g_{21}}{(g_3 + 3r)(g_2 + 2r)} + \frac{2g_{31}}{g_3 + 3r} + \frac{g_{42}}{g_2 + 2r}\right),$$

$$p(2,\ 2) = \frac{r(\frac{2}{3}g_{43}g_{32} + g_{43}g_{21})}{(g_4 + 4r)(g_3 + 3r)(g_2 + 2r)},$$

$$p(2,\ 1,\ 1) = p(1,\ 2,\ 1) = p(1,\ 1,\ 2) = \frac{r^2(\frac{8}{3}g_{32} + 4g_{21} + 2g_{43})}{(g_4 + 4r)(g_3 + 3r)(g_2 + 2r)}$$

and

$$p(1,\ 1,\ 1,\ 1) = \frac{24r^3}{(g_4 + 4r)(g_3 + 3r)(g_2 + 2r)}.$$

If these formulae for the CPF p are converted to the corresponding formulae for the EPPF, then (see also the remark after Definition 2.2) the consistency relation (6) must hold for the EPPF. In other words,

$$p(n_1,\ \ldots,\ n_k) = \frac{k+1}{n+1}p(n_1,\ \ldots,\ n_k,\ 1) + \sum_{j=1}^{k}\frac{n_j + 1}{n+1}p(n_1,\ \ldots,\ n_{j-1},\ n_j + 1,\ n_{j+1},\ \ldots,\ n_k), \tag{20}$$

which provides an additional significant check on computations. Choosing $n_1 = \cdots = n_k = 1$ yields $a_n = a_{n+1} + 2n/(n+1)p(2,\ 1,\ \ldots,\ 1)$. Therefore,

$$p\left(2,\ \underbrace{1,\ldots,1}_{(n-2)\text{ times}}\right) = \frac{n(a_{n-1} - a_n)}{2(n-1)} = \frac{n}{2(n-1)}\left(\prod_{i=2}^{n-1}\frac{ir}{g_i + ir} - \prod_{i=2}^{n}\frac{ir}{g_i + ir}\right)$$

$$= \frac{n!\,g_n r^{n-2}}{2(n-1)\prod_{i=2}^{n}(g_i + ir)}, \qquad n \geqslant 2, \tag{21}$$

that is, we have found the solution of the recursion (18) for the case $k = n - 1$

In the examples displayed above for $n \in \{2,\ 3,\ 4\}$, the coefficients involved in the rational expressions are all positive. This is so in general, as all the expressions involved in (18) are positive. The recursion (18) provides a powerful method to compute $p(\mathbf{n})$ numerically in reasonable time. As the probabilities $p(\mathbf{n})$ do not depend on the order of the entries $n_1,\ \ldots,\ n_k$, only the $p(\mathbf{n})$ with $n_1 \geqslant \ldots \geqslant n_k$ have to be computed. For $\mathbf{n} = (n_1,\ \ldots,\ n_k)$ with $n_1 \geqslant \ldots \geqslant n_k$, (18) can be rewritten as

$$p(\mathbf{n}) = \frac{nr}{g_n + nr}\frac{a_1}{k}\, p(\tilde{\mathbf{n}}) + \sum_{i=1}^{n-1} \frac{g_{n,n-i}}{g_n + nr} \sum_{l=i+1}^{n} \frac{a_l(l-i)}{n-i}\, p(\tilde{\mathbf{n}}(i,\, l)). \tag{22}$$

Here, for $l \in \{1, \ldots, n\}$, $a_l := \#\{1 \leqslant j \leqslant k \mid n_j = l\}$ denotes the number of indices $1 \leqslant j \leqslant k$ with $n_j = l$, $\tilde{\mathbf{n}} := (n_1, \ldots, n_{k-1})$ and $\tilde{\mathbf{n}}(i,\, l)$ is obtained from $\mathbf{n} = (n_1, \ldots, n_k)$ by replacing one entry n_j with $n_j = l$ by $l - i$ and then sorting the entries in ascending order.

In comparison to (18), the recursion (22) might be more useful for practical purposes. In terms of the probabilities $q(\mathbf{a}) = k!/(a_1! \cdots a_n!)p(\mathbf{n})$, $\mathbf{a} = (a_1, a_2, \ldots)$, (22) has the form $q(1, 0, 0, \ldots) = 1$ and

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

for $n = \sum_i i a_i > 1$, with $q(\mathbf{a}) := 0$, if at least one of the entries $a_1, a_2, \ldots$ is negative.

From the sampling probabilities $p(\mathbf{n})$ or $q(\mathbf{a})$ many important characteristics of the sample can be derived. For example, the distribution of the number K_n of alleles (types) in the sample satisfies $P(K_n = k) = \sum_{\mathbf{n}} p(\mathbf{n}) = \sum_{\mathbf{a}} q(\mathbf{a})$, $k \in \{1, \ldots, n\}$, where the first sum $\sum_{\mathbf{n}}$ extends over all $\mathbf{n} = (n_1, \ldots, n_k) \in \mathbb{N}^k$ with $n_1 + \ldots + n_k = n$ and the second sum $\sum_{\mathbf{a}}$ extends over all $\mathbf{a} = (a_1, a_2, \ldots) \in \mathbb{N}_0^\infty$ with $\sum_i i a_i = n$ and $\sum_i a_i = k$. Numerical results, for example for the distribution, the mean, and the variance of K_n, can be easily compared with the values (Ewens 1972)

$$P(K_n = k) = \frac{\theta^k s(n,\, k)}{[\theta]_n}, \qquad \mathrm{E}(K_n) = \sum_{i=0}^{n-1} \frac{\theta}{\theta + i}, \qquad \mathrm{var}(K_n) = \sum_{i=1}^{n-1} \frac{\theta i}{(\theta + i)^2},$$

which hold if the underlying genealogical tree is the Kingman coalescent.

In the following we present another method to compute $\mathrm{E}(K_n)$. From the consistency condition (20) it follows by an application of Kolmogorov's extension theorem that there exists a random equivalence relation R_∞ on $\mathbb{N}$ with distribution determined via

$$P(\varrho_n R_\infty = \xi) = \frac{k! n_1! \cdots n_k!}{n!}\, p(n_1, \ldots, n_k), \qquad \xi \in \mathcal{E}_n, \tag{23}$$

where $n_1, \ldots, n_k$ are the sizes of the equivalence classes of ξ. As the right-hand side in (23) is a symmetric function of $(n_1, \ldots, n_k)$, R_∞ is exchangeable. For $r, n \in \mathbb{N}$, let $\lambda_r(n)$ denote the size of the rth largest equivalence class of $\varrho_n R_\infty$. For convenience, let $\lambda_r(n) = 0$ if $\varrho_n R_\infty$ has fewer than r classes. It is well known (Kingman 1982c, Theorem 2) that for each $r \in \mathbb{N}$ the limiting relative frequency $X_r := lim_{n\to\infty} \lambda_r(n)/n$ exist almost surely. Note that $X_1 \geqslant X_2 \geqslant \ldots$ and that $\sum_{r=1}^\infty X_r \leqslant 1$. For example, if $\Lambda = \delta_0$ is the Dirac measure concentrated at 0, then $(X_r)_{r\in\mathbb{N}}$ is the Poisson–Dirichlet process (Joyce *et al.* 2002; Kingman 1977). Moreover, if $\xi \in \mathcal{E}_n$ is such that $n_j \geqslant 2$ for all $1 \leqslant j \leqslant k$, then Kingman's paintbox representation implies that

$$P(\varrho_n R_\infty = \xi) = \mathrm{E}\left(\sum_{\substack{r_1, \ldots, r_k \in \mathbb{N} \\ \text{distinct}}} X_{r_1}^{n_1} \cdots X_{r_k}^{n_k} \right). \tag{24}$$

In particular, for $k=1$ we conclude from (23) and (24) that $p(n)=\mathrm{E}(\sum_{r=1}^{\infty}X_r^n)$, $n\geqslant 2$. Assume now that the frequencies are proper, that is, that $\sum_{r=1}^{\infty}X_r=1$ almost surely. Then, for any polynomial of the form $g(x)=\sum_{i=1}^{n}a_{ni}x^i$, we have $\mathrm{E}(\sum_{r=1}^{\infty}g(X_r))=\sum_{i=1}^{n}a_{ni}p(i)$. The special choice $g(x)=1-(1-x)^n$, that is, $a_{ni}=\binom{n}{i}(-1)^{i-1}$, yields

$$\mathrm{E}(K_n)=\mathrm{E}\left(\sum_{r=1}^{\infty}g(X_r)\right)=\sum_{i=1}^{n}\binom{n}{i}(-1)^{i-1}p(i), \tag{25}$$

that is, if the frequencies $X_1, X_2, \ldots$ are proper, we can compute the mean of K_n via (25) and (19). In the following section we analyse the situation, where the genealogical tree is star-shaped. Other examples, in particular the Bolthausen–Sznitman coalescent, are presented in Section 6.

4. Star-shaped tree

A tree is called star-shaped if all the n branches coalesce at the root of the tree. More formally, the star-shaped tree corresponds to the case where the measure $\Lambda=\delta_1$ is concentrated at 1. The infinitesimal rates are therefore $g_n=g_{n1}=1$ and $g_{ni}=0$ for $i\in\{2,\ldots,n-1\}$. The length ε_0 of the tree is hence exponentially distributed with parameter 1. Conditioned on the length on the tree, mutations appear independently on the n branches at the points of a Poisson process with rate r. Obviously, $K_n=min(L_n+1, n)$, where L_n denotes the number of new types or, equivalently, the number of j with $1\leqslant j\leqslant n$ and $\varepsilon_j/r<\varepsilon_0$, where the $\varepsilon_0, \varepsilon_1, \ldots$ are independent and exponentially distributed with parameter 1. Obviously $P(L_n=k\,|\,T)=B(n, 1-\mathrm{e}^{-rt}, k)$, with $B(n, p, k):=\binom{n}{k}p^k(1-p)^{n-k}$, and hence

$$\begin{aligned}P(L_n=k)&=\int_0^{\infty}B(n, 1-\mathrm{e}^{-rt}, k)\mathrm{e}^{-t}\,\mathrm{d}t=\int_0^1 B(n, p, k)\frac{(1-p)^{1/r-1}}{r}\,\mathrm{d}p\\&=\frac{\binom{-1}{k}\binom{-1/r}{n-k}}{\binom{-1-1/r}{n}}=\frac{n!\,r^k}{(n-k)!}\prod_{i=n-k}^{n}\frac{1}{1+ir}, \qquad k\in\{0,\ldots,n\}\end{aligned} \tag{26}$$

(see also Johnson *et al.* 1992, p. 242, equation (6.18)). The distribution of L_n is hence binomial with random parameter p which in turn has a beta distribution with parameters 1 and $1/r$. It is well known that $((1+L_nr)/(1+r+nr))_{n\in\mathbb{N}_0}$ is a martingale and that L_n/n (or equivalently K_n/n) converges almost surely for $n\to\infty$ to a random variable W, beta-distributed with parameters 1 and $1/r$, that is $P(W>x)=(1-x)^{1/r}$, $0<x<1$. Note that W has moments $\mathrm{E}(W^l)=\prod_{i=1}^{l}(ir/(1+ir))$, $l\in\mathbb{N}_0$. In particular, $\mathrm{var}(W)=r^2/((1+r)^2(1+2r))$.

The almost sure convergence $K_n/n\to W$ differs significantly from the well-known convergence in distribution $(K_n-\theta\log n)/\sqrt{\theta\log n}\to N(0, 1)$, which holds if the underlying genealogical tree is the Kingman coalescent.

The distribution of K_n is easily obtained from (26) via $P(K_n=k)=P(L_n=k-1)$ for

$k \in \{1, \ldots, n-1\}$ and $P(K_n = n) = P(L_n \in \{n, n-1\}) = \prod_{i=2}^{n} (ir/(1+ir))$. As the partition has at most one block of size larger than 1, we obtain

$$p(\mathbf{n}) = \begin{cases} P(K_n = n) = \displaystyle\prod_{i=2}^{n} \frac{ir}{1+ir}, & \text{if } b(\mathbf{n}) = 0, \\ \dfrac{P(K_n = k)}{k} = \dfrac{n!r^{k-1}}{k(n-k+1)!} \displaystyle\prod_{i=n-k+1}^{n} \frac{1}{1+ir}, & \text{if } b(\mathbf{n}) = 1, \\ 0, & \text{if } b(\mathbf{n}) \geqslant 2, \end{cases} \tag{27}$$

for $\mathbf{n} = (n_1, \ldots, n_k)$ with $n := n_1 + \ldots + n_k \geqslant 2$, where $b(\mathbf{n}) := \#\{j \mid n_j > 1\}$ denotes the number of indices $j \in \{1, \ldots, k\}$ with $n_j > 1$. Note that K_n is a sufficient statistic for r, as $p(\mathbf{n})/P(K_n = k)$ does not depend on r. The frequency process $(X_r)_{r \in \mathbb{N}}$ introduced at the end of Section 3 satisfies $X_1 \overset{d}{=} 1 - W$ and $X_r = 0$ almost surely for $r \geqslant 2$. The structural distribution is that of a random variable with $(n-1)$th moment $p(n) = P(K_n = 1) = 1/(1+nr)$, that is, of a random variable of the form $(1-W)1_{\{U>W\}}$, where U is uniformly distributed on $(0, 1)$ and independent of W.

From (26) and $K_n = min(L_n + 1, n)$ it follows that the distribution of K_n satisfies the recursion $P(K_n = 1) = 1/(1+nr)$ and $P(K_n = k) = nr/(1+nr)P(K_{n-1} = k-1)$ for $n \geqslant 2$ and $k \in \{2, \ldots, n\}$. Thus, the probability generating function $f_n(s) := \sum_{k=1}^{n} P(K_n = k)s^k$ of K_n satisfies the recursion $f_1(s) = s$ and

$$f_n(s) = \frac{s}{1+nr}(1 + nrf_{n-1}(s)), \qquad n \geqslant 2,\ s \in [0, 1]. \tag{28}$$

Taking the derivative with respect to s and using $\mathrm{E}(K_n) = f_n'(1)$, it follows that the expected number of types satisfies the recursion $\mathrm{E}(K_1) = 1$ and $\mathrm{E}(K_n) = 1 + nr/(1+nr)\mathrm{E}(K_{n-1})$ for $n \geqslant 2$. Induction on n yields the explicit solution

$$\mathrm{E}(K_n) = \frac{nr}{1+r} + 1 - \prod_{i=1}^{n} \frac{ir}{1+ir} = n\mathrm{E}(W) + 1 - \mathrm{E}(W^n), \qquad n \in \mathbb{N}, \tag{29}$$

and the bounds $(1+nr)/(1+r) \leqslant \mathrm{E}(K_n) \leqslant 1 + nr/(1+r)$, $n \in \mathbb{N}$. Expressions for the higher moments of K_n can be derived similarly. For example, taking the second derivative with respect to s on both sides of (28), it follows that the second factorial moment of K_n satisfies the recursion $\mathrm{E}((K_1)_2) = 0$ and $\mathrm{E}((K_n)_2) = nr/(1+nr)\,(\mathrm{E}((K_{n-1})_2) + 2\,\mathrm{E}(K_{n-1}))$, $n \geqslant 2$, with solution

$$\mathrm{E}((K_n)_2) = n^2\mathrm{E}(W^2) + n(2\mathrm{E}(W) - \mathrm{E}(W^2) - 2\mathrm{E}(W^n)), \qquad n \in \mathbb{N}. \tag{30}$$

5. Extensions to simultaneous multiple collisions

If $\Lambda_j \neq 0$ for some $j \geqslant 2$, the recursion for the sampling probabilities $p(\mathbf{n})$ becomes more complicated, as multiple collisions of ancestral lineages are allowed to occur simultaneously with positive probability. In order to present this recursion, it is helpful to introduce the following notation. For a permutation π of the set $\{1, \ldots, a\}$ ($a \in \mathbb{N}$) and for k ($\in \mathbb{N}$) positive integers $t_1, \ldots, t_k \in \mathbb{N}$ with $t_1 + \ldots + t_k = a$, define $T_1 := \{\pi i \,|\, 1 \leqslant i \leqslant t_1\}$ and

$$T_j := \{\pi i \,|\, t_1 + \ldots + t_{j-1} + 1 \leqslant i \leqslant t_1 + \ldots + t_j\} \qquad \forall j \in \{2, \ldots, k\}.$$

Obviously, the sets $T_1, \ldots, T_k$ form a partition of $\{1, \ldots, a\}$ with $|T_j| = t_j$ for all $j \in \{1, \ldots, k\}$. We call $T_1, \ldots, T_k$ the partition of $\{1, \ldots, a\}$ induced from π and $t_1, \ldots, t_k$. The following theorem presents a recursion for the sampling probabilities $p(\mathbf{n})$.

Theorem 5.1. *If the underlying genealogical tree is given by a coalescent with mutation rate $\theta > 0$ and with rate functions $\phi_j : \mathbb{N}^j \to \mathbb{R}$, $j \in \mathbb{N}$, then the sampling probabilities $p(\mathbf{n})$ of non-ordered allele configurations $\mathbf{n} = (n_1, \ldots, n_k)$ satisfy the recursion $p(1) = 1$ and*

$$p(\mathbf{n}) = \frac{nr}{g_n + nr} \sum_{\substack{j=1 \\ n_j=1}}^{k} \frac{1}{k} p(\tilde{\mathbf{n}}_j) + \frac{n!}{g_n + nr} \sum_{a=1}^{n-1} \frac{1}{a!} \sum_{\substack{b_1, \ldots, b_a \in \mathbb{N} \\ b_1 + \ldots + b_a = n}} \frac{\phi_a(b_1, \ldots, b_a)}{b_1! \cdots b_a!}$$

$$\cdot \sum_{\substack{t_1, \ldots, t_k \in \mathbb{N} \\ t_1 + \ldots + t_k = a}} \sum_{(b_{\pi 1}, \ldots, b_{\pi a})} \frac{\displaystyle\prod_{j=1}^{k} \frac{t_j!}{t_{j1}! \cdots t_{jn}!}}{\displaystyle\frac{a!}{a_1! \cdots a_n!}} p(\mathbf{n} - \sum_{l=1}^{k} \sum_{i \in T_l} (b_i - 1)\mathbf{e}_l) \tag{31}$$

for $\mathbf{n} = (n_1, \ldots, n_k)$ with $n := n_1 + \ldots + n_k \geqslant 2$, where the total rates g_n are given by (12)*, $r := \theta/2$, $\tilde{\mathbf{n}}_j := (n_1, \ldots, n_{j-1}, n_{j+1}, \ldots, n_k)$ and $\mathbf{e}_j$ denotes the* j*th unit vector in $\mathbb{R}^k$. The sum $\sum_{(b_{\pi 1}, \ldots, b_{\pi a})}$ extends over all vectors $(b_{\pi 1}, \ldots, b_{\pi a})$, where π is a permutation of $\{1, \ldots, a\}$ such* that *$b_{\pi 1} \geqslant \ldots \geqslant b_{\pi t_1}$, $b_{\pi(t_1+1)} \geqslant \ldots \geqslant b_{\pi(t_1+t_2)}$ and so on, and $\sum_{i \in T_j} b_i = n_j$ for all $j \in \{1, \ldots, k\}$. $T_1, \ldots, T_k$ is the partition induced from π and $t_1, \ldots, t_k$. Furthermore, $t_{jl} := \#\{i \in T_j \,|\, b_i = l\}$ denotes the number of* $\mathrm{i} \in \mathbf{T}_j$ *with $b_i = l$ and $a_l := t_{1l} + \ldots + t_{kl} = \#\{i \,|\, b_i = l\}$ is the number of $i \in \{1, \ldots, a\}$ with $b_i = l$.*

Proof. In principle, the combinatorial arguments are the same as in the proof of (18). If the first event backwards in time is a mutation, the situation is totally identical to that in (18). Therefore, the first sum on the right-hand side in (31) has to be identical to that in (18). If the first event backwards in time is a coalescence, then exactly $a \in \{1, \ldots, n-1\}$ merging groups of sizes $b_1, \ldots, b_a \in \mathbb{N}$, $b_1 + \ldots + b_a = n$, occur with probability

$$\frac{n!}{a!\, b_1! \cdots b_a!} \frac{\phi_a(b_1, \ldots, b_a)}{g_n}.$$

Note that there exist exactly $n!/(a! b_1! \cdots b_a!)$ equivalence relations $\eta \in \mathcal{E}_n$ which correspond to a $(b_1, \ldots, b_a)$ merger. We now have to sum over all rearrangements $(b_{\pi 1}, \ldots, b_{\pi a})$ which are consistent with the type sizes $n_1, \ldots, n_k$, that is, which satisfy $b_{\pi 1} + \ldots + b_{\pi t_1} = n_1$, $b_{\pi(t_1+1)} + \ldots + b_{\pi(t_1+t_2)} = n_2$ and so on. There exist exactly

$$\prod_{j=1}^{k} \frac{t_j!}{t_{j1}! \cdots t_{jn}!}$$

such arrangements which correspond to an 'ordered' arrangement, where 'ordered' means that $b_{\pi 1} \geqslant \ldots \geqslant b_{\pi t_1}$, $b_{\pi(t_1+1)} \geqslant \ldots \geqslant b_{\pi(t_1+t_2)}$ and so on. The denominator $a!/(a_1! \cdots a_n!)$ takes into account that some of the merger numbers $b_1, \ldots, b_a$ are equal. □

Remark. In order to make the recursion (31) more transparent, it is explained in this remark why this recursion reduces to the simpler recursion (18) for the case where the coalescent process allows only for multiple collisions.

Only mergers $b_1, \ldots, b_a$, $a \in \{1, \ldots, n-1\}$ which satisfy $\phi_a(b_1, \ldots, b_a) > 0$ contribute to the right-hand side in (31). As simultaneous multiple collisions do not appear, $\phi_a(b_1, \ldots, b_a) > 0$ if and only if there exists some $m \in \{1, \ldots, a\}$ with $b_m = n-(a-1)$ and $b_i = 1$ for $i \in \{1, \ldots, a\} \backslash \{m\}$. In particular,

$$\frac{a!}{a_1! \cdots a_n!} = \frac{a!}{(a-1)!} = a.$$

Now consider the partition $T_1, \ldots, T_k$ induced from some integers $t_1, \ldots, t_k \in \mathbb{N}$ and some permutation π of the set $\{1, \ldots, a\}$. Let $j \in \{1, \ldots, k\}$ be the index such that $m \in T_j$. The constraint $n_j = \sum_{i \in T_j} b_i = b_m + |T_j| - 1 = n - a + t_j$, that is, $t_j = n_j - (n-a)$, can be only satisfied if $n_j > n-a$. For all other indices $l \in \{1, \ldots, k\} \backslash \{j\}$ we have $n_l = \sum_{i \in T_l} b_i = |T_l| = t_l$. Thus

$$\sum_{\substack{t_1, \ldots, t_k \in \mathbb{N} \\ t_1 + \ldots + t_k = a}} \sum_{(b_{\pi 1}, \ldots, b_{\pi a})} \frac{\prod_{j=1}^k t_j!/(t_{j1}! \cdots t_{jn}!)}{a!/(a_1! \cdots a_n!)} p\left(\mathbf{n} - \sum_{l=1}^k \sum_{i \in T_l} (b_i - 1)\mathbf{e}_l\right)$$

$$= \sum_{\substack{j=1 \\ n_j > n-a}}^k \frac{t_j}{a} p(\mathbf{n} - (b_m - 1)\mathbf{e}_j) = \sum_{\substack{j=1 \\ n_j > n-a}}^k \frac{n_j - (n-a)}{a} p(\mathbf{n} - (n-a)\mathbf{e}_j).$$

The substitution $i = n - a$ shows that (31) reduces to (18).

Examples. For $n \leqslant 3$ it is easily seen that the formulae for $p(\mathbf{n})$ based on the recursion (31) are identical to those presented in Section 3 based on the recursion (18) induced from a coalescent with only multiple collisions. The complexity of the formula for $p(\mathbf{n})$ increases rapidly with n. For example, for $n = 4$, $k = 2$ and $n_1 = n_2 = 2$, the recursion (31),

$$p(2,2) = \frac{3\phi_2(2,2)}{g_4 + 4r} p(1,1) + \frac{\frac{2}{3} g_{43}}{g_4 + 4r} p(2,1),$$

already involves the rate $\phi_2(2,2)$ of a double binary collision.

6. Examples and numerical studies

Assume that $\Lambda = U$ is uniformly distributed on [0, 1]. As already mentioned in Section 3, the corresponding Λ-coalescent $R = (R_t)_{t \geqslant 0}$ is the Bolthausen–Sznitman coalescent

(Bertoin and Le Gall 2000; Bolthausen and Sznitman 1998). The death process $D := (|R_t|)_{t \geqslant 0}$ has infinitesimal rates

$$g_{nk} = \binom{n}{k-1} \int_0^1 x^{n-k-1}(1-x)^{k-1}\,\mathrm{d}x = \binom{n}{k-1} B(k, n-k) = \frac{n}{(n-k)(n-k+1)},$$

$k \in \{1, \ldots, n-1\}$, where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ denotes the beta function. The total rates are

$$g_n = \sum_{k=1}^{n-1} g_{nk} = \sum_{k=1}^{n-1} \left(\frac{n}{n-k} - \frac{n}{n-k+1} \right) = n - 1.$$

In Table 1 the recursion (22) has been used to compute the sampling probabilities $p(\mathbf{n})$ for a sample of size $n = 5$ and mutation rate $\theta = 1$. The probabilities are compared with the values known for the Kingman coalescent ($\Lambda = \delta_0$) and for the star-shaped coalescent ($\Lambda = \delta_1$).

Numerical studies for several sample sizes and several values of the mutation rate θ indicate that, in comparison to the Ewens sampling formula based on the Kingman coalescent, under the Bolthausen–Sznitman coalescent configurations with low or high number of types (k close to 1 or n) appear with higher probability. Figures 2 and 3 show

Table 1. Sampling probabilities $p(\mathbf{n})$ for a sample of size $n = 5$ and mutation rate $\theta = 1$

Configuration **n**	Kingman coalescent ($\Lambda = \delta_0$)	Bolthausen–Sznitman coalescent ($\Lambda = U$)	Star-shaped coalescent ($\Lambda = \delta_1$)
(5)	0.200 00	0.222 53	0.285 71
(4,1)	0.125 00	0.134 16	0.119 05
(3,2)	0.083 33	0.035 71	0.000 00
(3,1,1)	0.055 56	0.067 16	0.063 49
(2,2,1)	0.041 67	0.023 81	0.000 00
(2,1,1,1)	0.020 83	0.032 97	0.035 71
(1,1,1,1,1)	0.008 33	0.032 97	0.142 86

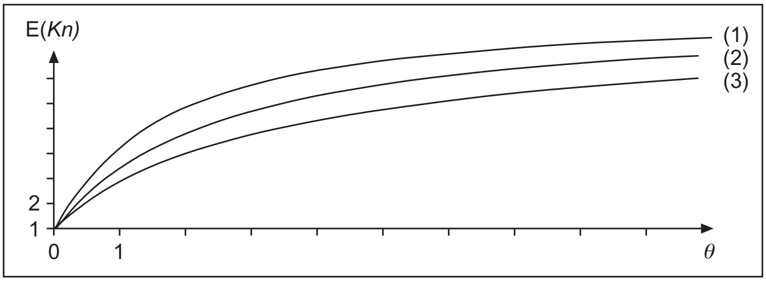

Figure 2. Mean of the number of types for a sample of size $n = 10$ as a function of θ ((1) = Star-shaped coalescent, (2) = Bolthausen–Sznitman coalescent, (3) = Kingman coalescent).

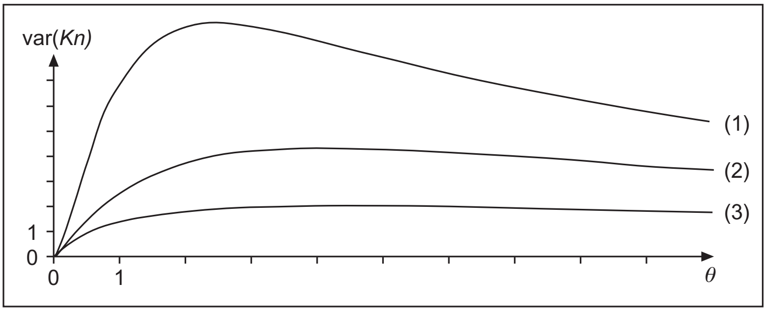

Figure 3. Variance of the number of types for a sample of size $n = 10$ as a function of θ ((1) = Star-shaped coalescent, (2) = Bolthausen–Sznitman coalescent, (3) = Kingman coalescent).

the mean and the variance of K_n, the number of types, for a sample of size $n = 10$ as a function of θ.

For the Bolthausen–Sznitman coalescent, numerical studies support the conjecture that K_n is asymptotically normal and that $\mathrm{E}(K_n)$ and $\mathrm{Var}(K_n)$ are both of order n. The author has not been able to verify this conjecture rigorously.

Numerical studies for other choices of the measures $\Lambda_1, \Lambda_2, \ldots$ indicate that $\mathrm{E}(K_n)$ is never smaller than the mean of K_n under the genealogy of the Kingman coalescent, and never larger than the mean of K_n under the genealogy of the star-shaped coalescent. The variance of K_n shows the same behaviour compared to the corresponding variances under the Kingman coalescent and the star-shaped tree. In this sense, the Kingman coalescent and the star-shaped coalescent can be viewed as the two extreme genealogies, which bound all other cases.

Acknowledgement

The author would like to thank Bob Griffiths for fruitful comments on combinatorial aspects of the Ewens sampling formula during a workshop at the Banff International Research Station (BIRS) in Canada in August 2004. Furthermore, the author thanks anonymous referees for helpful comments on the first version of the manuscript.

References

Bertoin, J. and Le Gall, J.F. (2000) The Bolthausen–Sznitman coalescent and the genealogy of continuous-state branching processes. *Probab. Theory Related Fields*, **117**, 249–266.

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

Cannings C. (1974) The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. *Adv. Appl. Probab.*, **6**, 260–290.

Cannings, C. (1975) The latent roots of certain Markov chains arising in genetics: a new approach, II. Further haploid models. *Adv. Appl. Probab.*, **7**, 264–282.

De Iorio, M. and Griffiths, R.C. (2004) Importance sampling on coalescent histories. *Adv. Appl. Probab.*, **36**, 417–433.

Ewens, W.J. (1972) The sampling theory of selectively neutral alleles. *Theoret. Popul. Biol.*, **3**, 87–112.

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

Griffiths, R.C. and Lessard, S. (2005) Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles. *Theoret. Popul. Biol.* To appear.

Griffiths, R.C. and Tavaré, S. (1996) Monte Carlo inference methods in population genetics. *Math. Comput. Modelling*, **23**, 141–158.

Johnson, N.L., Kotz, S. and Kemp A.W. (1992) *Univariate Discrete Distributions*, 2nd edn. New York: Wiley.

Joyce, P., Krone, S.M. and Kurtz, T.G. (2002) Gaussian limits associated with the Poisson–Dirichlet distribution and the Ewens sampling formula. *Ann. Appl. Probab.*, **12**, 101–124.

Kingman, J.F.C. (1977) The population structure associated with the Ewens sampling formula. *Theoret. Popul. Biol.*, **11**, 274–283.

Kingman, J.F.C. (1982a) On the genealogy of large populations. *J. Appl. Probab.*, **19A**, 27–43.

Kingman, J.F.C. (1982b) Exchangeability and the evolution of large populations. In G. Koch and F. Spizzichino (eds), *Exchangeability in Probability and Statistics*, pp. 97–112. Amsterdam: North-Holland.

Kingman, J.F.C. (1982c) The coalescent. *Stochastic Process. Appl.*, **13**, 235–248.

Kingman, J.F.C. (2000) Origins of the coalescent: 1974–1982. *Genetics*, **156**, 1461–1463.

Möhle, M. and Sagitov, S. (2001) A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.*, **29**, 1547–1562.

Pitman, J. (1995) Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields*, **102**, 145–158.

Pitman, J. (1999) Coalescents with multiple collisions. *Ann. Probab.*, **27**, 1870–1902.

Pitman, J. (2002) Combinatorial stochastic processes. Technical Report 621, Department of Statistics, University of California, Berkeley.

Rosenblatt, M. (1959) Functions of a Markov process that are Markovian. *J. Math. Mech.*, **8**, 585–596.

Sagitov, S. (1999) The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.*, **36**, 1116–1125.

Schweinsberg, J. (2000a) A necessary and sufficient condition for the Λ-coalescent to come down from infinity. *Electron. Comm. Probab.*, **5**, 1–11.

Schweinsberg, J. (2000b) Coalescents with simultaneous multiple collisions. *Electron. J. Probab.*, **5**, 1–50.

Received September 2004 and revised April 2005