Adv. Appl. Prob. **38,** 750–767 (2006)
Printed in Northern Ireland
© *Applied Probability Trust* 2006

ON THE NUMBER OF SEGREGATING SITES FOR POPULATIONS WITH LARGE FAMILY SIZES

M. MÖHLE,* *Eberhard Karls Universität Tübingen*

Abstract

We present recursions for the total number, S_n, of mutations in a sample of n individuals, when the underlying genealogical tree of the sample is modelled by a coalescent process with mutation rate $r > 0$. The coalescent is allowed to have simultaneous multiple collisions of ancestral lineages, which corresponds to the existence of large families in the underlying population model. For the subclass of Λ-coalescent processes allowing for multiple collisions, such that the measure $\Lambda(\mathrm{d}x)/x$ is finite, we prove that $S_n/(nr)$ converges in distribution to a limiting variable, S, characterized via an exponential integral of a certain subordinator. When the measure $\Lambda(\mathrm{d}x)/x^2$ is finite, the distribution of S coincides with the stationary distribution of an autoregressive process of order 1 and is uniquely determined via a stochastic fixed-point equation of the form $S \overset{\mathrm{D}}{=} AS + B$, with specific independent random coefficients A and B. Examples are presented in which explicit representations for (the density of) S are available. We conjecture that $S_n/\mathrm{E}(S_n) \to 1$ in probability if the measure $\Lambda(\mathrm{d}x)/x$ is infinite.

Keywords: Autoregressive process; coalescent process; infinitely-many-sites model; multiple collisions; segregating sites; stochastic difference equation; total number of mutations

2000 Mathematics Subject Classification: Primary 60J27; 92D10
Secondary 60F05; 92D15

1. Introduction and main results

Kingman's coalescent process has proved to be a powerful tool in ancestral population genetics. This process describes the ancestry of a sample of n particles, individuals, genes, or DNA sequences taken from a large population (see [12], [13], [14], and [15]). More precisely, Kingman's coalescent is a continuous-time Markov process $(R_t)_{t\geq 0}$ whose state space is the set of all equivalence relations on $\{1, \dots, n\}$ such that i and j are in the same equivalence class (block) of R_t if and only if the ith and the jth individuals in the sample have a common ancestor at time t in the past. All transitions involve exactly two classes of R_t merging together, and each such merging occurs at rate 1.

It is assumed that each individual is of a certain type. A mutation process is superimposed on the genealogical tree as follows. Mutations appear independently of the genealogical tree at the points of a 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 brand new type, never seen before.

In ancestral population genetics certain statistical functionals are of fundamental importance. For example, given a sample of n strands of DNA, let Δ_{ij} be the number of sites at which the ith

Received 23 September 2005; revision received 12 April 2006.

* Postal address: Mathematisches Institut der Universität Tübingen, Auf der Morgenstelle 10, 72076 Tübingen, Germany. Email address: martin.moehle@uni-tuebingen.de

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

and the jth segments differ. Then $\Delta_n := \binom{n}{2}^{-1} \sum_{i \neq j} \Delta_{ij}$ is the number of pairwise differences. Another important statistic is the number, S_n, of segregating sites in the sample, i.e. the number of sites at which at least one pair of segments differ. These statistics are important, for example, to test the hypothesis of neutrality [25] or for the existence of selective sweeps [8]. It is well known that, for the Kingman coalescent, S_n is asymptotically normal with mean $\mathrm{E}(S_n) \sim \theta \log n$ and variance $\mathrm{var}(S_n) \sim \theta \log n$, where $\theta := 2r$.

Within the last decade, progress has been made in describing the genealogy of populations which allow for large offspring sizes. Examples of such populations are known from the study of marine organisms [11]. In the corresponding coalescent process, many equivalence classes of R_t can merge at once into a single class. These coalescent processes with multiple collisions were introduced by Pitman [18] and Sagitov [22]. Even more generally, if the population is allowed occasionally to have many very large families, many such multiple mergers can occur simultaneously. These coalescent processes with simultaneous multiple collisions were introduced by Schweinsberg [24] and Möhle and Sagitov [16]. In this paper we study the number, S_n, of segregating sites under the assumption that the underlying genealogical tree is modelled by a coalescent process with simultaneous multiple collisions.

The paper is organized as follows. In Section 2 we present recursions for S_n, in particular for the distribution, the probability generating function, and the mean of S_n.

From Section 3 on we focus on coalescent processes with multiple collisions. The most simple way to introduce these coalescent processes is via a finite measure, Λ, on [0, 1]. Coalescent processes with multiple collisions are therefore also called Λ-coalescents. Transitions from a given equivalence relation ξ to an equivalence relation $\eta \neq \xi$ with $\xi \subset \eta$ occur (by definition) with rate

$$q_{\xi\eta} = \int_{[0,1]} x^{n-k-1}(1-x)^{k-1}\Lambda(\mathrm{d}x), \tag{1.1}$$

where $n := |\xi|$ and $k := |\eta|$ denote the number of classes (blocks) of ξ and η, respectively. We first restrict our consideration to measures Λ satisfying the conditions

$$\Lambda(\{0\}) = 0 \quad \text{and} \quad \mu_{-2} := \int_{(0,1]} x^{-2}\Lambda(\mathrm{d}x) < \infty. \tag{1.2}$$

In the spirit of Bertoin and Le Gall [3, Lemmas 3 and 4], we call measures Λ satisfying (1.2) *simple measures*, and we speak of the *simple case* whenever (1.2) is satisfied. Conditions (1.2) prevent Λ from having too much mass near 0. A typical simple measure is $\Lambda = \delta_u$, the Dirac measure at $u \in (0, 1]$.

For simple measures Λ, our main result (Theorem 3.1) states that $S_n/(nr)$ converges in distribution to a limiting random variable S which is characterized by a stochastic functional equation of the form $S \overset{\mathrm{D}}{=} AS + B$, where the coefficients A and B are certain independent random variables. Such functional equations are well known from the theory of autoregressive processes and linear recursions. For more details we refer the reader to [6] or [26]. Basic examples are presented in Section 4. In Section 5 the convergence in distribution $S_n/(nr) \to S$ is extended (see Theorem 5.1) to the more general class of measures Λ satisfying the conditions

$$\Lambda(\{0\}) = 0 \quad \text{and} \quad \mu_{-1} := \int_{(0,1]} x^{-1}\Lambda(\mathrm{d}x) < \infty. \tag{1.3}$$

In this case the characteristic equation for $S_n/(nr)$ degenerates in the limit and, hence, gives no information about the distribution of the limiting variable S. We show that the distribution of S can be characterized via an exponential integral of a subordinator with zero drift and Laplace

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

exponent $\Phi(x) = \int_{[0,1]}(1-(1-y)^x)y^{-2}\Lambda(\mathrm{d}y)$. Finally (in Section 6) we comment on the open case in which the measure $\Lambda(\mathrm{d}x)/x$ is infinite.

2. Recursions for the total number of mutations

Let Λ be a finite measure on [0, 1] and let $R = (R_t)_{t\geq 0}$ be a Λ-coalescent process, i.e. a time-continuous Markovian process with state space $\mathcal{E}$, i.e. the set of all equivalence relations on $\mathbb{N}$, and rates (1.1). Let $D_t := |R_t|$ denote the number of blocks (equivalence classes) of R_t. It is well known that the block-counting process $D = (D_t)_{t\geq 0}$ is a Markovian death process with state space $\mathbb{N}$, infinitesimal rates

$$g_{nk} := \binom{n}{k-1}\int_{[0,1]} x^{n-k-1}(1-x)^{k-1}\Lambda(\mathrm{d}x), \qquad 1 \leq k < n, \tag{2.1}$$

and total rates

$$g_n = \sum_{k=1}^{n-1} g_{nk} = \int_{[0,1]} \frac{1-(1-x)^n - nx(1-x)^{n-1}}{x^2}\Lambda(\mathrm{d}x), \qquad n \in \mathbb{N}. \tag{2.2}$$

For $n \in \mathbb{N}$, let $(\mathcal{D}_k^{(n)})_{k\in\mathbb{N}_0}$ (where $\mathbb{N}_0 = \mathbb{N}\cup\{0\}$) denote the jump chain of the process $(|\varrho_n R_t|)_{t\geq 0}$, where $\varrho_n\colon \mathcal{E} \to \mathcal{E}_n$ denotes the natural projection to the set $\mathcal{E}_n$ of all equivalence relations on $\{1,\dots,n\}$. Note that the jump chain has initial state $\mathcal{D}_0^{(n)} = n$. The first jump will be to the state k, $1 \leq k < n$, with probability

$$r_{nk} := \mathrm{P}(\mathcal{D}_1^{(n)} = k) = \frac{g_{nk}}{g_n}, \qquad n, k \in \mathbb{N},\ 1 \leq k < n.$$

The infinitely-many-alleles model is assumed, i.e. mutations appear on each branch of the tree at the points of a Poisson process with rate $r > 0$ and each mutation leads to a brand new type. Let S_n denote the total number of mutations along the genealogical tree back to the most recent common ancestor of a sample of size n. Note that, in the infinitely-many-sites model, S_n is equal to the number of segregating sites. Obviously, $S_1 = 0$ and

$$S_n = Y_n + S_{\mathcal{D}_1^{(n)}} = Y_n + \sum_{k=1}^{n-1} \mathbf{1}_{\{\mathcal{D}_1^{(n)}=k\}} S_k, \qquad n \geq 2, \tag{2.3}$$

where Y_n is the number of mutations that arise during the time, $\tau_n \overset{\mathrm{D}}{=} \mathrm{Exp}(g_n)$, that the sample has n ancestors. (By '$\overset{\mathrm{D}}{=}$' we denote equality in distribution.) Conditional on τ_n, Y_n is Poisson-distributed with parameter $nr\tau_n$. Thus, Y_n has probability generating function

$$\mathrm{E}(s^{Y_n}) = \mathrm{E}(\mathrm{E}(s^{Y_n} \mid \tau_n)) = \mathrm{E}(\mathrm{e}^{-nr\tau_n(1-s)}) = \frac{g_n}{g_n + nr(1-s)}, \qquad s \in [0,1],$$

i.e. Y_n is geometrically distributed with parameter $g_n/(g_n + nr)$. The recursion (2.3) for S_n is useful to derive recursions for functionals of S_n. For example, the probability generating function of S_n satisfies the recursion

$$\begin{aligned}\mathrm{E}(s^{S_n}) &= \mathrm{E}(s^{Y_n})\sum_{k=1}^{n-1} r_{nk}\,\mathrm{E}(s^{S_k}) \\ &= \frac{1}{g_n + nr(1-s)}\sum_{k=1}^{n-1} g_{nk}\,\mathrm{E}(s^{S_k}), \qquad s \in [0,1],\ n \geq 2,\end{aligned} \tag{2.4}$$

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

with initial condition $\mathrm{E}(s^{S_1}) = 1$. Substituting the argument s of the probability generating function by $\mathrm{e}^{-\lambda}$, $\lambda \geq 0$, turns (2.4) into a recursion for the Laplace transform of S_n, which we will use later in this article. The mean and the second mean of S_n respectively satisfy the recursions

$$\mathrm{E}(S_n) = \mathrm{E}(Y_n) + \sum_{k=1}^{n-1} r_{nk}\,\mathrm{E}(S_k) = \frac{nr}{g_n} + \frac{1}{g_n}\sum_{k=2}^{n-1} g_{nk}\,\mathrm{E}(S_k)$$

and

$$\begin{aligned}\mathrm{E}(S_n^2) &= \mathrm{E}(Y_n^2) + 2\,\mathrm{E}(Y_n)\,\mathrm{E}(S_{\mathcal{D}_1^{(n)}}) + \mathrm{E}(S^2_{\mathcal{D}_1^{(n)}})\\ &= \mathrm{E}(Y_n^2) - 2\,\mathrm{E}(Y_n)^2 + 2\,\mathrm{E}(Y_n)\,\mathrm{E}(S_n) + \mathrm{E}(S^2_{\mathcal{D}_1^{(n)}})\\ &= \frac{nr}{g_n} + 2\frac{nr}{g_n}\,\mathrm{E}(S_n) + \frac{1}{g_n}\sum_{k=2}^{n-1} g_{nk}\,\mathrm{E}(S_k^2),\end{aligned}$$

from which a recursion for the variance of S_n may be derived. Slightly more complicated is the recursion for the distribution of S_n. We have $\mathrm{P}(S_1 = j) = \delta_{j0}$ (the Kronecker symbol) and, for $n \geq 2$ and $j \in \mathbb{N}_0$,

$$\begin{aligned}\mathrm{P}(S_n = j) &= \sum_{k=1}^{n-1} r_{nk}\,\mathrm{P}(Y_n + S_k = j)\\ &= \sum_{k=1}^{n-1} r_{nk} \sum_{i=0}^{j} \mathrm{P}(Y_n = i)\,\mathrm{P}(S_k = j-i)\\ &= \sum_{k=1}^{n-1} \frac{g_{nk}}{g_n} \sum_{i=0}^{j} \frac{g_n}{g_n + nr}\left(\frac{nr}{g_n + nr}\right)^i \mathrm{P}(S_k = j-i)\\ &= \sum_{k=1}^{n-1} \frac{g_{nk}}{g_n + nr} \sum_{i=0}^{j}\left(\frac{nr}{g_n + nr}\right)^i \mathrm{P}(S_k = j-i).\end{aligned}$$

We will not use the recursion for the distribution of S_n in our further considerations, but we do have to study the mean and the variance of S_n in more detail. For this purpose it is helpful to consider the total length, L_n, of the tree $(\varrho_n R_t)_{t\geq 0}$, which satisfies a recursion similar to (2.3), namely $L_1 = 0$ and

$$L_n = n\tau_n + L_{\mathcal{D}_1^{(n)}} = n\tau_n + \sum_{k=1}^{n-1} \mathbf{1}_{\{\mathcal{D}_1^{(n)}=k\}} L_k, \qquad n \geq 2, \tag{2.5}$$

where $\tau_n \overset{\mathrm{D}}{=} \mathrm{Exp}(g_n)$ is the time during which the sample has n ancestors. Let $M = (M(t))_{t\geq 0}$ denote the mutation Poisson process with parameter $r > 0$. Then we have the distributional relation

$$S_n \overset{\mathrm{D}}{=} M(L_n) \tag{2.6}$$

between S_n and L_n. In our model (genetic neutrality), the mutation process M is independent of the underlying genealogical tree and, hence, in particular, independent of the tree length L_n. As a consequence of (2.6), the mathematical analysis of S_n is essentially equivalent to

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

that of L_n. Many functionals of S_n can be expressed in terms of L_n. For example, for $k \in \mathbb{N}$, $\mathrm{P}(S_n < k) = \mathrm{P}(L_n < T_k)$, where T_k, the time of the kth jump of the Poisson process M, is independent of L_n. Moreover, $\mathrm{E}(S_n) = r\,\mathrm{E}(L_n)$,

$$\begin{aligned}\mathrm{var}(S_n) &= \mathrm{E}(\mathrm{var}(S_n \mid L_n)) + \mathrm{var}(\mathrm{E}(S_n \mid L_n))\\ &= \mathrm{E}(rL_n) + \mathrm{var}(rL_n)\\ &= r\,\mathrm{E}(L_n) + r^2\,\mathrm{var}(L_n),\end{aligned}$$

and, hence, $\mathrm{E}(S_n^2) = r\,\mathrm{E}(L_n) + r^2\,\mathrm{E}(L_n^2)$. From (2.5) it follows that the Laplace transform of L_n satisfies the recursion

$$\mathrm{E}(\mathrm{e}^{-\lambda L_n}) = \mathrm{E}(\mathrm{e}^{-\lambda n\tau_n}) \sum_{k=1}^{n-1} r_{nk}\,\mathrm{E}(\mathrm{e}^{-\lambda L_k}) = \frac{1}{g_n + n\lambda} \sum_{k=1}^{n-1} g_{nk}\,\mathrm{E}(\mathrm{e}^{-\lambda L_k}), \qquad \lambda \geq 0,\, n \geq 2,$$

with initial condition $\mathrm{E}(\mathrm{e}^{-\lambda L_1}) = 1$. The first and second raw moments, $a_n := \mathrm{E}(L_n)$ and $b_n := \mathrm{E}(L_n^2)$, satisfy the recursions $a_1 := b_1 := 0$ and

$$a_n := \frac{n}{g_n} + \frac{1}{g_n} \sum_{k=2}^{n-1} g_{nk} a_k \quad \text{and} \quad b_n := \frac{2n}{g_n} a_n + \frac{1}{g_n} \sum_{k=2}^{n-1} g_{nk} b_k, \qquad n \geq 2. \tag{2.7}$$

Note that neither sequence $(a_n)_{n\in\mathbb{N}}$ nor sequence $(b_n)_{n\in\mathbb{N}}$ depends on the mutation rate r. Let

$$r_{nk}^* := \sum_{l\in\mathbb{N}_0} r_{nk}^{(l)} = \sum_{l\in\mathbb{N}_0} \mathrm{P}(\mathcal{D}_l^{(n)} = k) = \sum_{l\in\mathbb{N}_0} \mathrm{E}(\mathbf{1}_{\{\mathcal{D}_l^{(n)} = k\}}) = \mathrm{E}\bigg(\sum_{l\in\mathbb{N}_0} \mathbf{1}_{\{\mathcal{D}_l^{(n)} = k\}}\bigg)$$

denote the expected number of visits of the jump chain $(\mathcal{D}_l^{(n)})_{l\in\mathbb{N}_0}$ to the state k. Straightforward induction on n shows that a_n is given by $a_n = \sum_{k=2}^n (k/g_k) r_{nk}^*$. This formula can be found in [8, Equation (4.2)]. Unfortunately, closed expressions for r_{nk}^* are rarely available. However, under additional assumptions, upper bounds for r_{nk}^* are available and lead to asymptotic results for S_n for large n. Explicit solutions for the distribution of S_n are only known for special cases; in particular for the Kingman coalescent, for which the measure $\Lambda = \delta_0$ is the Dirac measure at 0, and for the star-shaped coalescent with $\Lambda = \delta_1$, the point measure at 1.

Example 2.1. (*Kingman coalescent.*) For the Kingman coalescent ($\Lambda = \delta_0$, i.e. $g_n = g_{n,n-1} = n(n-1)/2$) the recursion (2.5) reduces to $L_n = n\tau_n + L_{n-1}$, $n \geq 2$. Thus, $L_n = \sum_{i=2}^n i\tau_i$, where the τ_i, $i \geq 2$, are independent, exponentially distributed random variables with parameter $g_i = i(i-1)/2$, and, hence,

$$\mathrm{E}(L_n) = \sum_{i=2}^n i\,\mathrm{E}(\tau_i) = \sum_{i=2}^n \frac{2}{i-1} = 2\sum_{k=1}^{n-1} \frac{1}{k} \sim 2\log n$$

and

$$\mathrm{var}(L_n) = \sum_{i=2}^n i^2\,\mathrm{var}(\tau_i) = \sum_{i=2}^n \frac{4}{(i-1)^2} = 4\sum_{k=1}^{n-1} \frac{1}{k^2} \sim \frac{2\pi^2}{3}.$$

In particular, $\mathrm{E}(S_n) = r\,\mathrm{E}(L_n) \sim \theta \log n$ and $\mathrm{var}(S_n) = r\,\mathrm{E}(L_n) + r^2\,\mathrm{var}(L_n) \sim \theta \log n$, with $\theta := 2r$. The probability generating function of S_n is given by

$$\mathrm{E}(s^{S_n}) = \prod_{k=2}^n \mathrm{E}(s^{Y_k}) = \prod_{k=2}^n \frac{k-1}{k-1+\theta(1-s)}. \tag{2.8}$$

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

Further properties of S_n, in particular its asymptotic normality, follow easily from (2.8). The exact formula

$$\mathrm{P}(S_n = k) = \frac{n-1}{\theta} \sum_{j=1}^{n-1} (-1)^{j-1} \binom{n-2}{j-1} \left(\frac{\theta}{j+\theta}\right)^{k+1}, \qquad n \geq 2,\ k \geq 0,$$

for the distribution of S_n goes back to Watterson [27].

Example 2.2. (*Star-shaped coalescent.*) For the star-shaped coalescent ($\Lambda = \delta_1$, i.e. $g_n = g_{n1} = 1$ for $n \geq 2$), we have $L_n = n\tau_n$, $\mathrm{E}(L_n) = n$, and $\mathrm{var}(L_n) = n^2$, for $n \geq 2$. In particular, $\mathrm{E}(S_n) = nr$ and $\mathrm{var}(S_n) = nr(1 + nr)$ for $n \geq 2$, which follows also from the fact that, for $n \geq 2$, $S_n = Y_n$ is geometrically distributed with parameter $1/(1 + nr)$.

We mention that all the recursions derived in this section are also valid for the general class of exchangeable coalescent processes with simultaneous multiple collisions studied by Schweinsberg [24] and Möhle and Sagitov [16]. We simply need to replace the rates g_{nk} and the total rates g_n by those of the more general block-counting process, $D = (D_t)_{t \geq 0}$, of the coalescent with simultaneous multiple collisions.

3. Asymptotics for the number of mutations

Our aim is to study the asymptotic behaviour of S_n for large n in the situation in which the underlying genealogical tree is given by a Λ-coalescent, i.e. a coalescent with multiple collisions. In this section we restrict our considerations to so-called simple measures Λ, i.e. to measures Λ satisfying conditions (1.2). From (2.2) it follows that (1.2) is equivalent to $\lim_{n\to\infty} g_n < \infty$. As already mentioned in the introduction, conditions (1.2) prevent Λ from having too much mass near 0. The Λ-coalescent is a Markov process of jump-hold type with bounded transition rates and step function paths if and only if (1.2) is satisfied (see [18, p. 1874]). Typical examples satisfying conditions (1.2) are $\Lambda = \delta_u$, the Dirac measure at $u \in (0, 1]$, or $\Lambda = \beta(p, q)$, the beta distribution with parameters $p > 2$ and $q > 0$. Typical measures which do not satisfy conditions (1.2) are beta distributions $\beta(p, q)$ with $0 < p \leq 2$. A treatment of more general measures requires extended, or other, methods. We discuss these extensions in more detail in Sections 5 and 6. Conditions (1.2) imply that

$$\nu(\mathrm{d}x) := x^{-2}\Lambda(\mathrm{d}x) \tag{3.1}$$

is a finite measure on $[0, 1]$ with $\nu(\{0\}) = 0$. For $k \in \mathbb{N}_0$, let $m_k := \int x^k \nu(\mathrm{d}x)$ denote the kth moment of ν. We furthermore exclude the trivial case in which $\Lambda \equiv 0$, i.e. we assume that $\nu((0, 1]) > 0$ and, hence, that $m_k > 0$ for all $k \in \mathbb{N}_0$ and $m_0 > m_1 > m_2 > \cdots$. For the proof of the convergence result (Theorem 3.1) we need the following two lemmas.

Lemma 3.1. *If the measure $\Lambda \neq 0$ satisfies conditions (1.2), then the following assertions hold.*

(i) $1 - \mathcal{D}_1^{(n)}/n$ *converges in distribution to the probability measure* $\nu_0 := \nu/m_0$, *where* $m_0 := \nu([0, 1])$ *is the total mass of the measure,* ν, *defined in (3.1).*

(ii) *There exists a constant,* $c > 0$ *(depending on* Λ *but not on* n *and* r*), such that* $\mathrm{E}(S_n) \leq cnr$ *for all* $n \in \mathbb{N}$ *and all* $r \geq 0$.

(iii) *The Laplace transform,* ψ_n, *of* $S_n/(nr)$ *satisfies* $-c \leq \psi_n'(\lambda) \leq 0$ *for all* $n \in \mathbb{N}$ *and all* $\lambda \geq 0$, *where* c *is the constant in (ii).*

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

Proof. (i) By assumption, ν is a finite measure with $\nu(\{0\}) = 0$. Thus, $\int (1-x)^n \nu(\mathrm{d}x) \to 0$ by dominated convergence. Therefore,

$$\begin{aligned} g_n &= \int \frac{1-(1-x)^n - nx(1-x)^{n-1}}{x^2}\Lambda(\mathrm{d}x) \\ &= \int (1-(1-x)^n - nx(1-x)^{n-1})\nu(\mathrm{d}x) \\ &= m_0 - \int (1-x)^n \nu(\mathrm{d}x) - \int nx(1-x)^{n-1}\nu(\mathrm{d}x) \\ &\to m_0. \end{aligned}$$

For $n \in \mathbb{N}$ and $0 \le x \le 1$, let $Z_n \equiv Z_n(x)$ be a random variable binomially distributed with parameters n and x. For $l \in \mathbb{N}$, we have

$$\begin{aligned} g_n \,\mathrm{E}\left(\left(1 - \frac{\mathcal{D}_1^{(n)} - 1}{n}\right)^l\right) &= g_n \sum_{k=1}^{n-1}\left(1 - \frac{k-1}{n}\right)^l r_{nk} \\ &= \sum_{k=1}^{n-1}\left(1 - \frac{k-1}{n}\right)^l g_{nk} \\ &= \int \sum_{k=1}^{n-1}\left(\frac{n-k+1}{n}\right)^l \mathrm{P}(Z_n(x) = n-k+1)\nu(\mathrm{d}x) \\ &= \int \sum_{i=2}^{n}\left(\frac{i}{n}\right)^l \mathrm{P}(Z_n(x) = i)\nu(\mathrm{d}x) \quad (\text{where } i = n-k+1) \\ &= \int [\mathrm{E}((Z_n(x)/n)^l) - (1/n)^l\,\mathrm{P}(Z_n(x) = 1)]\nu(\mathrm{d}x) \\ &\to \int x^l \nu(\mathrm{d}x) \\ &= m_l \end{aligned}$$

by dominated convergence. Thus, the moments of $1 - (\mathcal{D}_1^{(n)} - 1)/n$ converge to those of $\nu_0 := \nu/m_0$. As $1 - (\mathcal{D}_1^{(n)} - 1)/n$ (which satisfies $0 \le 1 - (\mathcal{D}_1^{(n)} - 1)/n \le 1$) is uniformly bounded, this convergence of moments implies the convergence in distribution.

(ii) Obviously, $\sum_{k=2}^{n-1} kr_{nk} \le \sum_{k=2}^{n-1}(n-1)r_{nk} \le n-1$ for all $n \in \mathbb{N}$. Furthermore,

$$\sum_{k=2}^{n-1} kr_{nk} \le \sum_{k=1}^{n-1} kr_{nk} = \mathrm{E}(\mathcal{D}_1^{(n)}) \sim n(1 - m_1/m_0),$$

where $1 - m_1/m_0 < 1$. Thus, there exists a constant p, $0 < p < 1$ (which might depend on Λ but not on n), such that

$$\sum_{k=2}^{n-1} kr_{nk} \le np \quad \text{for all } n \in \mathbb{N}.$$

By induction on $l \in \mathbb{N}$, it follows that

$$\sum_{k=2}^{n-1} kr_{nk}^{(l)} \le np^l \quad \text{for all } n, l \in \mathbb{N}.$$

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

The sequence, $(g_n)_{n\in\mathbb{N}}$, of total rates is monotone increasing. In particular, $1/g_k \le 1/g_2 = 1$ for $k \ge 2$. Thus,

$$\begin{aligned}\frac{\mathrm{E}(S_n)}{nr} &= \frac{1}{n}\sum_{k=2}^{n}\frac{k}{g_k}r^*_{nk} \le \frac{1}{n}\sum_{k=2}^{n}kr^*_{nk} = \frac{1}{n}\sum_{k=2}^{n}k\sum_{l\in\mathbb{N}_0}r^{(l)}_{nk}\\ &= \frac{1}{n}\sum_{l\in\mathbb{N}_0}\sum_{k=2}^{n}kr^{(l)}_{nk} = \frac{1}{n}\bigg(n + \sum_{l\in\mathbb{N}}\sum_{k=2}^{n-1}kr^{(l)}_{nk}\bigg)\\ &\le \frac{1}{n}\sum_{l\in\mathbb{N}_0}np^l = \frac{1}{1-p}\\ &=: c.\end{aligned}$$

(iii) Obviously, $\psi_n(\lambda) = f_n(s(\lambda))$, where f_n is the probability generating function of S_n and $s(\lambda) := \exp(-\lambda/(nr))$, $\lambda \ge 0$. Therefore, $\psi_n'(\lambda) = s'(\lambda)f_n'(s(\lambda)) \le 0$ and

$$\psi_n''(\lambda) = s''(\lambda)f_n'(\lambda) + (s'(\lambda))^2 f_n''(s(\lambda)) \ge 0,$$

i.e. ψ_n' is monotone increasing on $[0,\infty)$. It is hence sufficient to verify that $\psi_n'(0) \ge -c$ for all $n \in \mathbb{N}$ or, equivalently, that $\mathrm{E}(S_n) \le cnr$ for all $n \in \mathbb{N}$, which is true by (ii).

Let A and B be independent random variables. The following lemma goes back to Vervaat [26].

Lemma 3.2. *Assume that* $-\infty \le \mu := \mathrm{E}(\log|A|) < 0$. *If the stochastic functional equation*

$$X \overset{\mathrm{D}}{=} AX + B \tag{3.2}$$

has a solution, then the solution is unique in distribution.

Proof. Let $((A_n, B_n))_{n\in\mathbb{N}_0}$ be a sequence of independent and identically distributed random variables with $(A_n, B_n) \overset{\mathrm{D}}{=} (A, B)$. By iterating (3.2), we obtain

$$X \overset{\mathrm{D}}{=} A_0\cdots A_n X + \sum_{i=0}^{n-1} B_i \prod_{j=0}^{i-1} A_j. \tag{3.3}$$

By the strong law of large numbers,

$$\frac{1}{n}\sum_{k=0}^{n}\log|A_k| \xrightarrow{\text{a.s.}} \mathrm{E}(\log|A|) = \mu \in [-\infty, 0),$$

where '$\xrightarrow{\text{a.s.}}$' denotes almost-sure convergence. Therefore, $\log|A_0\cdots A_n| \xrightarrow{\text{a.s.}} -\infty$ and, hence, $A_0\cdots A_n \xrightarrow{\text{a.s.}} 0$. From (3.3) it follows that X is the weak limit of $\sum_{i=0}^{n-1} B_i \prod_{j=0}^{i-1} A_j$ as $n \to \infty$. In particular, the distribution of X is uniquely determined by A and B.

We are now able to present the convergence theorem, which clarifies the asymptotic behaviour of the number, S_n, of segregating sites.

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

Theorem 3.1. *If the measure $\Lambda \neq 0$ satisfies conditions (1.2), then $S_n/(nr)$ converges in distribution to a nonnegative limiting random variable S. The distribution of S is uniquely determined by the stochastic functional equation*

$$S \overset{\mathrm{D}}{=} AS + B, \tag{3.4}$$

where A and B are independent random variables (independent of S) distributed as follows: $1 - A \overset{\mathrm{D}}{=} \nu_0 := \nu/m_0$ and $B \overset{\mathrm{D}}{=} \mathrm{Exp}(m_0)$. The Laplace transform, ψ, of S is uniquely determined by the functional equation

$$\psi(\lambda) = \frac{m_0}{m_0 + \lambda}\mathrm{E}(\psi(A\lambda)) = \frac{1}{m_0 + \lambda}\int_{[0,1]} \psi((1-x)\lambda)\nu(\mathrm{d}x), \qquad \lambda \geq 0. \tag{3.5}$$

Before the proof of Theorem 3.1 is presented, let us make the following remark.

Remark 3.1. Stochastic functional equations of the form (3.4) or, equivalently, (3.5) are well known from the theory of autoregressive processes and stochastic affine recursions. The distribution of S (see, for example, [26] or [6]) coincides with the stationary distribution of a generalized autoregressive process $(X_n)_{n\in\mathbb{N}_0}$ defined by $X_0 := 0$ and $X_{n+1} := A_nX_n + B_n$, where $(A_n, B_n)_{n\in\mathbb{N}_0}$ is a sequence of independent, identically distributed random variables with $(A_n, B_n) \overset{\mathrm{D}}{=} (A, B)$. Note that

$$X_n = \sum_{i=0}^{n-1} B_{n-i-1} \prod_{j=n-i}^{n-1} A_j \overset{\mathrm{D}}{=} \sum_{i=0}^{n-1} B_i \prod_{j=0}^{i-1} A_j, \qquad n \in \mathbb{N}_0,$$

and, hence, that $S \overset{\mathrm{D}}{=} \sum_{i=0}^{\infty} B_i \prod_{j=0}^{i-1} A_j$.

Proof of Theorem 3.1. Fix the mutation rate $r > 0$ and let ψ_n denote the Laplace transform of $S_n/(nr)$.

Step 1. We verify the existence of a subsequence $(n_l)_{l\in\mathbb{N}}$ of integers such that the limits $\lim_{l\to\infty} \psi_{n_l}(\lambda)$, $\lambda \geq 0$, all exist.

Let $\mathbb{Q}_+ = \{q_1, q_2, \dots\}$ be a countable representation of the set of all nonnegative rational numbers. As the sequence $(\psi_n(q_1))_{n\in\mathbb{N}}$ is bounded, there exists a subsequence, $(n_{1l})_{l\in\mathbb{N}}$, such that the limit $\psi(q_1) := \lim_{l\to\infty} \psi_{n_{1l}}(q_1)$ exists. As the sequence $(\psi_n(q_2))_{n\in\mathbb{N}}$ is bounded, there exists a subsequence, $(n_{2l})_{l\in\mathbb{N}}$, of $(n_{1l})_{l\in\mathbb{N}}$ such that the limit $\psi(q_2) := \lim_{l\to\infty} \psi_{n_{2l}}(q_2)$ exists. Iteratively we find, for each k, a subsequence $(n_{kl})_{l\in\mathbb{N}}$ of $(n_{k-1,l})_{l\in\mathbb{N}}$ such that the limit $\psi(q_k) := \lim_{l\to\infty} \psi_{n_{kl}}(q_k)$ exists. Thus, for the diagonal sequence $n_l := n_{ll}$ we have $\psi(q_k) = \lim_{l\to\infty} \psi_{n_l}(q_k)$ for all k. In other words, for the subsequence $(n_l)_{l\in\mathbb{N}}$, the limits $\psi(q) = \lim_{l\to\infty} \psi_{n_l}(q)$, $q \in \mathbb{Q}_+$, all exist.

Lemma 3.1(iii) ensures that $|\psi_n'(\lambda)| \leq c$ for all $\lambda \geq 0$ and $n \in \mathbb{N}$. Thus (by the mean value theorem), for $p, q \in \mathbb{Q}_+$ there exists a $\xi \equiv \xi(n_l, p, q)$ between p and q such that

$$|\psi(p) - \psi(q)| = \lim_{l\to\infty} |\psi_{n_l}(p) - \psi_{n_l}(q)| = \lim_{l\to\infty} |\psi_{n_l}'(\xi)(p - q)| \leq c|p - q|.$$

Thus, ψ is Lipschitz continuous on $\mathbb{Q}_+$. For an arbitrary $\lambda \geq 0$, take two sequences, $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$, of rational numbers such that $a_k \nearrow \lambda$ and $b_k \searrow \lambda$. As Laplace transforms are monotone decreasing on $\mathbb{R}_+$, we have

$$\psi_{n_l}(a_k) \geq \psi_{n_l}(\lambda) \geq \psi_{n_l}(b_k).$$

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

Taking the limit as $l \to \infty$ yields

$$\psi(a_k) \geq \limsup_{l\to\infty} \psi_{n_l}(\lambda) \geq \liminf_{l\to\infty} \psi_{n_l}(\lambda) \geq \psi(b_k).$$

As ψ is continuous on $\mathbb{Q}_+$, the difference $\psi(a_k)-\psi(b_k)$ converges to 0 as $k \to \infty$. Thus, the limit $\psi(\lambda) := \lim_{l\to\infty} \psi_{n_l}(\lambda)$ exists, and ψ is continuous on $\mathbb{R}_+$.

Step 2. Assume that $(n_l)_{l\in\mathbb{N}}$ is a subsequence such that the limits $\psi(\lambda) := \lim_{l\to\infty} \psi_{n_l}(\lambda)$, $\lambda \geq 0$, all exist. Again from Lemma 3.1(iii), it follows that ψ is Lipschitz continuous. The continuity theorem for Laplace transforms ensures that ψ is the Laplace transform of some random variable S and that $(S_{n_l})_{l\in\mathbb{N}}$ converges in distribution to S. It remains to verify that the distribution of S does not depend on the subsequence $(n_l)_{l\in\mathbb{N}}$. In order to see this we proceed as follows. From (2.4) we conclude that ψ_n satisfies the recursion

$$\begin{aligned}\psi_n(\lambda) &= \mathrm{E}(\mathrm{e}^{-\lambda S_n/(nr)})\\ &= \frac{g_n}{g_n + nr(1-\mathrm{e}^{-\lambda/(nr)})}\sum_{k=1}^{n-1} r_{nk}\,\mathrm{E}(\mathrm{e}^{-\lambda S_k/(nr)})\\ &= \frac{g_n}{g_n + nr(1-\mathrm{e}^{-\lambda/(nr)})}\sum_{k=1}^{n-1} r_{nk}\psi_k\biggl(\frac{k}{n}\lambda\biggr).\end{aligned} \tag{3.6}$$

Since $\lim_{n\to\infty} g_n = m_0$, it follows that

$$\lim_{n\to\infty} \frac{g_n}{g_n + nr(1-\mathrm{e}^{-\lambda/(nr)})} = \frac{m_0}{m_0+\lambda},$$

which is the Laplace transform of an exponentially distributed random variable B, say, with parameter m_0. Now,

$$\begin{aligned}\sum_{k=1}^{n-1} \psi_k\biggl(\frac{k}{n}\lambda\biggr) r_{nk} &= \sum_{x\in\{1/n,\dots,(n-1)/n\}} \psi_{n(1-x)}((1-x)\lambda)\,\mathrm{P}(1-\mathcal{D}_1^{(n)}/n = x)\\ &= \int_{[0,1]} \psi_{[n(1-x)]}((1-x)\lambda) P_{1-\mathcal{D}_1^{(n)}/n}(\mathrm{d}x),\end{aligned}$$

where $P_{1-\mathcal{D}_1^{(n)}/n}$ denotes the distribution of $1-\mathcal{D}_1^{(n)}/n$, $[x] := \sup\{z\in\mathbb{Z} \colon z \leq x\}$ for $x\in\mathbb{R}$, and we adopt the convention that $\psi_0(\lambda) := 1$. As ψ is continuous, monotone, and bounded, the convergence

$$\lim_{l\to\infty} \psi_{[n_l(1-x)]}((1-x)\lambda) = \psi((1-x)\lambda)$$

holds uniformly in $x \in [0,1]$. By setting $n := n_l$ in (3.6) and letting $l \to \infty$, we obtain

$$\psi(\lambda) = \frac{m_0}{m_0+\lambda}\int_{[0,1]} \psi((1-x)\lambda)\frac{\nu(\mathrm{d}x)}{m_0},$$

which is the functional equation (3.5). In the language of random variables this functional equation is equivalent to $S \overset{\mathrm{D}}{=} AS + B$, where A and B are mutually independent and independent of S with $1-A \overset{\mathrm{D}}{=} \nu_0$ and $B \overset{\mathrm{D}}{=} \mathrm{Exp}(m_0)$. The random variable A takes values in $[0,1]$ almost surely. By assumption, ν_0 is not the zero measure, i.e. $\mathrm{P}(A=1) < 1$. Thus, $-\infty \leq \mathrm{E}(\log A) < 0$. Lemma 3.2 ensures the uniqueness of S in distribution. In particular, the distribution of S does not depend on the subsequence $(n_l)_{l\in\mathbb{N}}$.

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

Corollary 3.1. *Assume that the measure $\Lambda \neq 0$ satisfies conditions (1.2). Then, at least for $0 \le \lambda < m_0$, the Laplace transform ψ of S has the Taylor expansion $\psi(\lambda) = \sum_{k=0}^{\infty} c_k \lambda^k$, with $c_0 := 1$ and*

$$c_k := \prod_{i=1}^{k} \frac{1}{\int((1-x)^i - 1)\nu(\mathrm{d}x)}, \qquad k \in \mathbb{N}. \tag{3.7}$$

In particular, S has moments

$$\mathrm{E}(S^k) = \prod_{i=1}^{k} \frac{i}{\Phi(i)} = \frac{k!}{\Phi(1)\cdots\Phi(k)}, \qquad k \in \mathbb{N}, \tag{3.8}$$

where $\Phi(i) := \int_{[0,1]}(1-(1-x)^i)\nu(\mathrm{d}x)$, $i \in \mathbb{N}$.

Proof. Define c_k as in (3.7). It is straightforward to verify that, for $0 \le \lambda < m_0 := \nu([0,1])$, $\psi(\lambda) := \sum_{k=0}^{\infty} c_k \lambda^k$ is a convergent series which solves the functional equation (3.5) on $[0, m_0)$. The formula for the moments follows from $\mathrm{E}(S^k) = (-1)^k \psi^{(k)}(0+) = (-1)^k k!\, c_k$.

4. Examples

We present two basic but important examples. In both examples, explicit expressions for the density of the limiting random variable S are derived.

Example 4.1. Fix a $\beta > 0$. Suppose that Λ has density $x \mapsto \beta x^2(1-x)^{\beta-1}$ with respect to the Lebesgue measure on $[0,1]$. Obviously, conditions (1.2) are satisfied and, hence, Theorem 3.1 is applicable. The measure ν has density $x \mapsto \beta(1-x)^{\beta-1}$ with respect to the Lebesgue measure, i.e. ν is the beta distribution with parameters 1 and β. Therefore, the Laplace transform, ψ, of the limiting random variable S satisfies the functional equation

$$(1+\lambda)\psi(\lambda) = \int_{[0,1]} \psi((1-x)\lambda)\nu(\mathrm{d}x) = \beta \int_0^1 \psi((1-x)\lambda)(1-x)^{\beta-1}\mathrm{d}x$$

with unique solution $\psi(\lambda) = 1/(1+\lambda)^{\beta+1}$, $\lambda \ge 0$. Thus, S is gamma distributed with parameters $\beta + 1$ and 1, i.e. S has density $t \mapsto t^{\beta}\mathrm{e}^{-t}/\Gamma(\beta+1)$, $t \ge 0$.

Note that for the special case $\beta = 1$, ν is the uniform distribution on $[0,1]$. The block-counting process $D = (D_t)_{t \ge 0}$ has rates $g_{nk} = 1/(n+1)$, $1 \le k < n$, and total rates $g_n = (n-1)/(n+1)$, $n \in \mathbb{N}$. Hence, the corresponding jump chain moves from n to any state k, $1 \le k < n$, with equal probability $r_{nk} = g_{nk}/g_n = 1/(n-1)$.

Example 4.2. Let $\Lambda = \delta_u$ be the Dirac measure at $u \in (0,1]$. The measure ν is then concentrated at u with total mass $m_0 = \nu([0,1]) = u^{-2}$. Thus, by Theorem 3.1, $S_n/(nr)$ converges in distribution to a nonnegative limiting random variable S. The distribution of S is uniquely determined by the stochastic functional equation $S \overset{\mathrm{D}}{=} (1-u)S + B$, where $B \overset{\mathrm{D}}{=} \mathrm{Exp}(u^{-2})$ is independent of S. The functional equation, (3.5), for the Laplace transform, ψ, of S reduces to

$$(1+u^2\lambda)\psi(\lambda) = \psi((1-u)\lambda), \qquad \lambda \ge 0,$$

and S has moments

$$\mathrm{E}(S^k) = \prod_{i=1}^{k} \frac{iu^2}{1-(1-u)^i} = k!\,u^{2k} \prod_{i=1}^{k} \frac{1}{1-(1-u)^i}, \qquad k \in \mathbb{N}_0.$$

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

In particular, $\mathrm{E}(S) = u$ and $\mathrm{var}(S) = u^3/(2-u)$. The distribution of S coincides with the stationary distribution of an autoregressive process $(X_n)_{n\in\mathbb{N}_0}$ of order 1 defined by $X_0 := 0$ and $X_{n+1} := (1-u)X_n + B_n$, $n \in \mathbb{N}_0$, where $(B_n)_{n\in\mathbb{N}_0}$ is a sequence of independent, identically distributed random variables, each exponentially distributed with parameter u^{-2}. From Remark 3.1 it follows that $S \overset{\mathrm{D}}{=} \sum_{i=0}^{\infty}(1-u)^i B_i$. For the star-shaped coalescent ($u = 1$) it follows immediately that $S \overset{\mathrm{D}}{=} B_0$ is exponentially distributed with parameter 1. Assume now that $0 < u < 1$. As $(1-u)^i B_i \overset{\mathrm{D}}{=} \mathrm{Exp}(\lambda_i)$ with $\lambda_i := 1/(u^2(1-u)^i)$, it follows that S has Laplace transform $\psi(\lambda) = \prod_{i=0}^{\infty} \lambda_i/(\lambda_i + \lambda)$, density

$$t \mapsto \sum_{i=0}^{\infty} \lambda_i \mathrm{e}^{-\lambda_i t} \prod_{\substack{j=0\\ j\neq i}}^{\infty} \frac{\lambda_j}{\lambda_j - \lambda_i}, \qquad t \geq 0,$$

and distribution function

$$t \mapsto 1 - \sum_{i=0}^{\infty} \mathrm{e}^{-\lambda_i t} \prod_{\substack{j=0\\ j\neq i}}^{\infty} \frac{\lambda_j}{\lambda_j - \lambda_i}, \qquad t \geq 0.$$

5. Extensions and further examples

In this section it is assumed that the measure Λ satisfies (1.3), which are obviously weaker conditions than (1.2). Note that in this case the measure ν defined via (3.1) is no longer necessarily finite. From (2.1) it follows that (1.3) is equivalent to $\lim_{n\to\infty} \gamma_n/n < \infty$, where (see [23, Lemma 3])

$$\gamma_n := \sum_{k=1}^{n-1} (n-k) g_{nk} = \int_{[0,1]} (nx - 1 + (1-x)^n)\nu(\mathrm{d}x).$$

Note that it follows from Lemma 25 of [18] that the Λ-coalescent remains infinite if (1.3) holds. Obviously, the right-hand sides of (3.7) and (3.8) are still defined for the wider class of measures satisfying (1.3). Thus, it is tempting to generalize the convergence result (Theorem 3.1) to measures satisfying (1.3), as follows.

Theorem 5.1. *If the measure* $\Lambda \neq 0$ *satisfies conditions (1.3), then* $S_n/(nr)$ *converges in distribution to a nonnegative limiting random variable* S *uniquely determined by its moments (see (3.8)). The Laplace transform,* ψ*, of* S *solves the integral equation*

$$\lambda\psi(\lambda) = \int_{[0,1]} [\psi((1-x)\lambda) - \psi(\lambda)]\nu(\mathrm{d}x), \tag{5.1}$$

where ν *is defined via (3.1).*

Remark 5.1. If the measure ν is finite, then (5.1) and (3.5) are equivalent, and Theorem 5.1 essentially coincides with Theorem 3.1.

Proof of Theorem 5.1. Without loss of generality (see Theorem 3.1), assume that $m_0 := \nu([0,1]) = \infty$. Define $\nu_m(\mathrm{d}x) := \mathbf{1}_{\{x>1/m\}}\,\nu(\mathrm{d}x)$ and $\Lambda_m(\mathrm{d}x) := x^2\nu_m(\mathrm{d}x)$. Let $S_n(m)$ denote the number of segregating sites in a sample of size n for the situation in which the underlying genealogical tree is modelled by a Λ_m-coalescent. Applying Theorem 3.1 to the

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

measure Λ_m yields the weak convergence of $S_n(m)/(nr)$ to a limiting random variable $S(m)$ whose distribution is uniquely determined by its moments,

$$\begin{aligned} \mathrm{E}(S(m)^k) &= \prod_{i=1}^{k} \frac{i}{\int (1-(1-x)^i)\nu_m(\mathrm{d}x)} \\ &= \prod_{i=1}^{k} \frac{i}{\int_{(1/m,1]} (1-(1-x)^i)\nu(\mathrm{d}x)}, \qquad k \in \mathbb{N}_0. \end{aligned} \tag{5.2}$$

The right-hand side of (5.2) converges to (3.8) as m tends to infinity. This convergence of moments implies the convergence of $S(m)$ in distribution to some limiting random variable S with moments (3.8). Note that, on the one hand, from $\lim_{k\to\infty} \Phi(k) = \nu([0,1]) = \infty$ it follows that $\sum_{k=1}^{\infty} t^k/(\Phi(1)\cdots\Phi(k))$ is a convergent series for all t. Thus, the moment generating function $t \mapsto \mathrm{E}(\mathrm{e}^{tS}) = 1 + \sum_{k=1}^{\infty} t^k/(\Phi(1)\cdots\Phi(k))$ exists and, hence, the distribution of S is uniquely determined by the sequence of moments (3.8). For more details on such moment problems we refer the reader to [10]. On the other hand, Λ_m converges weakly to Λ as m tends to infinity and, hence, S is the weak limit of the sequence $(S_n/(nr))_{n\in\mathbb{N}}$.

It remains to verify (5.1). Define c_k as in (3.7). Using the Taylor expansion $\psi(\lambda) = \sum_{k=0}^{\infty} c_k \lambda^k$, $0 \le \lambda < m_0 = \infty$, for the Laplace transform of S, it is straightforward to verify that ψ solves the integral equation (5.1).

In the following it is helpful to introduce the transformation $T\colon [0,1] \to [0,\infty]$ via

$$T(y) := -\log(1-y), \tag{5.3}$$

with the convention that $T(1) := \infty$. Note that $T^{-1}(y) = 1 - \mathrm{e}^{-y}$. The author would like to express his thanks to Aleksander Iksanov for pointing out the following characterization of the distribution of S in terms of an exponential integral of a subordinator.

Proposition 5.1. *Assume that the measure $\Lambda \neq 0$ satisfies conditions (1.3). Then the distribution of the limiting variable S has the representation*

$$S \overset{\mathrm{D}}{=} \int_0^\infty \mathrm{e}^{-X_t}\,\mathrm{d}t,$$

where $(X_t)_{t\ge0}$ is a subordinator (i.e. a Lévy process with nondecreasing paths) with zero drift and Lévy measure $\varrho := \nu_T$, i.e. $\varrho(A) = \nu(T^{-1}(A))$ for all Borel sets $A \subseteq [0,\infty]$, where ν is the measure defined in (3.1) and T is the transformation defined in (5.3).

Proof. The function

$$\Phi(x) := \int_{[0,\infty]} (1-\mathrm{e}^{-xy})\varrho(\mathrm{d}y) = \int_{[0,1]} (1-(1-y)^x)\nu(\mathrm{d}y)$$

is a Laplace exponent (Bernstein function) of a subordinator $(X_t)_{t\ge0}$ with Lévy measure ϱ and zero drift. Setting $Z := \int_0^\infty \mathrm{e}^{-X_t}\,\mathrm{d}t$, by Proposition 3.3 of [7] (specialized to subordinators) we have

$$\mathrm{E}(Z^k) = \frac{k!}{\Phi(1)\cdots\Phi(k)}, \qquad k \in \mathbb{N},$$

the right-hand side of which is a moment sequence that uniquely determines the distribution. By comparing the latter equality with (3.8) we deduce that $\mathrm{E}(Z^k) = \mathrm{E}(S^k)$ for all k and, hence, that $Z \overset{\mathrm{D}}{=} S$.

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

Remark 5.2. When $0 < m_0 := \nu([0,1]) < \infty$, the Lévy measure ϱ is finite ($\varrho([0,\infty]) = m_0 < \infty$), which means that $(X_t)_{t\geq 0}$ is a compound Poisson process $X_t = \sum_{i=1}^{N(t)} \eta_i$, where $N := (N(t))_{t\geq 0}$ is a homogeneous Poisson process with parameter m_0 and η_i, $i \in \mathbb{N}$, are random variables, independent of each other and of N, with common distribution function $y \mapsto \mathrm{P}(\eta_i \leq y) := m_0^{-1}\varrho([0,y])$. Let $T_1, T_2, T_3, \ldots,$ $T_1 < T_2 < T_3 < \cdots$, denote the jump times of the Poisson process N. Then

$$\begin{aligned} Z = \int_0^\infty \mathrm{e}^{-X_t}\,\mathrm{d}t &= \sum_{i=0}^\infty \int_{T_i}^{T_{i+1}} \mathrm{e}^{-X_t}\,\mathrm{d}t \\ &= T_1 + (T_2 - T_1)\mathrm{e}^{-\eta_1} + (T_3 - T_2)\mathrm{e}^{-\eta_1-\eta_2} + \cdots \\ &= T_1 + \mathrm{e}^{-\eta_1} Z_1, \end{aligned}$$

where $Z_1 \overset{\mathrm{D}}{=} Z$. Therefore, we have $S \overset{\mathrm{D}}{=} \mathrm{e}^{-\eta_1} S + T_1$, which is equivalent to (3.4).

Example 5.1. Assume that Λ has density $x \mapsto x$ (i.e. the identity mapping) with respect to the Lebesgue measure on $[0,1]$. In this case the block-counting process D has rates $g_{nk} = 1/(n-k+1)$, $1 \leq k < n$, and total rates $g_n = \sum_{i=2}^n 1/i$, $n \geq 2$. Note that $g_n \sim \log n$ as $n \to \infty$. By Theorem 5.1, $S_n/(nr)$ converges in distribution to a nonnegative random variable S uniquely determined by its moments, $\mathrm{E}(S^k) = k!/(h_1 \cdots h_k)$, $k \in \mathbb{N}_0$, where

$$\begin{aligned} h_i = \int_{[0,1]} (1-(1-x)^i)\nu(\mathrm{d}x) &= \int_0^1 \frac{1-(1-x)^i}{x}\,\mathrm{d}x \\ &= \int_0^1 \sum_{j=1}^i \binom{i}{j}(-x)^{j-1}\,\mathrm{d}x = \sum_{j=1}^i \binom{i}{j}\frac{(-1)^{j-1}}{j} \\ &= \sum_{j=1}^i \frac{1}{j} \end{aligned}$$

is the ith harmonic number, $i \in \mathbb{N}$. In particular, $\mathrm{E}(S) = 1$, $\mathrm{var}(S) = \frac{1}{3}$, and S has Laplace transform

$$\psi(\lambda) := \mathrm{E}(\mathrm{e}^{-\lambda S}) = \sum_{k=0}^\infty \frac{(-\lambda)^k}{k!}\mathrm{E}(S^k) = \sum_{k=0}^\infty \frac{(-\lambda)^k}{h_1\cdots h_k}, \qquad \lambda \geq 0.$$

The Lévy measure, ϱ, of the corresponding subordinator satisfies

$$\varrho([a,b]) = \nu([1-\mathrm{e}^{-a}, 1-\mathrm{e}^{-b}]) = \int_{1-\mathrm{e}^{-a}}^{1-\mathrm{e}^{-b}} x^{-1}\,\mathrm{d}x = \log\left(\frac{1-\mathrm{e}^{-b}}{1-\mathrm{e}^{-a}}\right)$$

for $0 < a < b < \infty$. Note that ϱ has density $x \mapsto \mathrm{e}^{-x}/(1-\mathrm{e}^{-x})$ with respect to the Lebesgue measure on $(0,\infty)$.

The following example is taken from [4, p. 102].

Example 5.2. Fix an α, $0 < \alpha < 1$, and assume that Λ has density

$$x \mapsto \frac{(1-\alpha)^2}{\alpha\Gamma(\alpha+1)} \frac{x^2(1-x)^{1/\alpha-2}}{(1-(1-x)^{1/\alpha})^{2-\alpha}}$$

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

with respect to the Lebesgue measure on (0, 1). It is straightforward to check that

$$\int_{[0,1]} x^{-1}\Lambda(\mathrm{d}x) = \frac{1-\alpha}{\Gamma(\alpha+1)} < \infty.$$

Note that the Lévy measure ϱ has density

$$x \mapsto \frac{(1-\alpha)^2}{\alpha\Gamma(\alpha+1)}\frac{\mathrm{e}^{x/\alpha}}{(\mathrm{e}^{x/\alpha}-1)^{2-\alpha}}, \qquad 0 < x < \infty,$$

with respect to the Lebesgue measure on $(0, \infty)$. From the results of [4, p. 102] we conclude that $S \overset{\mathrm{D}}{=} X^\alpha$, where X is an exponential variable with unit mean.

Remark 5.3. Clearly, when S satisfies (3.4), the corresponding distribution is absolutely continuous. More generally, from Proposition 2.1 of [7] it follows that S always admits a density, f, which is infinitely differentiable on $(0, \infty)$ and solves the integral equation

$$f(x) = \int_x^\infty \varrho((\log(u/x), \infty])f(u)\,\mathrm{d}u, \qquad 0 < x < \infty.$$

6. Final remarks and open problems

Assume that the measure Λ does not satisfy (1.3) or, equivalently, that the measure $\Lambda(\mathrm{d}x)/x$ is infinite. In this case the asymptotic behaviour of S_n for large n seems to be of a different nature. Fix an $\varepsilon > 0$. Chebyshev's inequality shows that

$$\begin{aligned} \mathrm{P}\biggl(\biggl|\frac{S_n}{\mathrm{E}(S_n)} - 1\biggr| \geq \varepsilon\biggr) &= \mathrm{P}(|S_n - \mathrm{E}(S_n)| \geq \varepsilon\,\mathrm{E}(S_n)) \\ &\leq \frac{\mathrm{var}(S_n)}{\varepsilon^2\,\mathrm{E}(S_n)^2} = \frac{r\,\mathrm{E}(L_n) + r^2\,\mathrm{var}(L_n)}{\varepsilon^2 r^2\,\mathrm{E}(L_n)^2} \\ &= \frac{ra_n + r^2(b_n - a_n^2)}{\varepsilon^2 r^2 a_n^2} = \frac{1}{\varepsilon^2}\biggl(\frac{1}{ra_n} + \frac{b_n}{a_n^2} - 1\biggr), \end{aligned}$$

where $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ are the sequences defined in (2.7). Therefore, $S_n/\mathrm{E}(S_n) \to 1$ in probability if

$$a_n \to \infty \quad \text{and} \quad b_n \sim a_n^2. \tag{6.1}$$

The first example in the literature of a Λ-coalescent with multiple collisions was probably the so-called Bolthausen–Sznitman coalescent [5], which is (by definition) the Λ-coalescent with Λ the uniform distribution on [0, 1]. We refer the reader to [9] for some more recent results on this particular Λ-coalescent. Obviously, (1.3) does not hold for the Bolthausen–Sznitman coalescent. It is shown in Appendix A that $a_n \geq n/\log n$ for all $n \geq 3$. In particular, $a_n \to \infty$ for the Bolthausen–Sznitman coalescent. We conjecture that $a_n \sim n/\log n$ and that $b_n \sim a_n^2$, i.e. that (6.1) holds and, therefore, that $S_n/\mathrm{E}(S_n) \to 1$ in probability for the Bolthausen–Sznitman coalescent. For the Λ-coalescent with beta distribution $\Lambda = \beta(2-\alpha, \alpha)$, $1 < \alpha < 2$, Berestycki *et al.* [1, Theorem 1.9] verified that $S_n/n^{2-\alpha} \to c$ in probability. They also presented an expression for the limiting constant, c, in terms of gamma functions. In this context we also refer the reader to [2], where closely related exact asymptotic results for the site frequency spectrum were derived for such beta coalescents. We conjecture that (6.1) and, hence, the convergence in probability $S_n/\mathrm{E}(S_n) \to 1$ hold for all Λ-coalescent processes which do not satisfy (1.3).

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

The next step would be to analyse the limiting behaviour of the standardized variable

$$S_n^* := \frac{S_n - \mathrm{E}(S_n)}{\sqrt{\mathrm{var}(S_n)}}.$$

The convergence of S_n^* in distribution to some limiting variable S is well known in the Kingman case ($\Lambda = \delta_0$), where S is standard normal distributed. It does not seem straightforward to modify the proofs presented in this article so as to hold for S_n^*. Therefore, methods other than those presented here seem to be necessary to analyse the asymptotic behaviour of S_n^* for measures Λ which do not satisfy (1.3). Contraction methods (see [19], [20], and [21]) are often helpful in the asymptotic analysis of random recursive sequences. Neininger and Rüschendorf [17] presented asymptotic results for a class of such sequences in which the characteristic equation for the scaled sequence degenerates in the limit to a trivial equation and, thus, gives no information about the limiting distribution. They explained how the normal distribution arises although the degenerate limit equation does not give any indication of asymptotic normality. When the measure $\Lambda(\mathrm{d}x)/x$ is infinite, the asymptotic behaviour of the recursion (2.3) might be tractable with similar methods.

Appendix A.

We show that, for the Bolthausen–Sznitman coalescent, the total tree length, L_n, of a sample of size $n \geq 3$ is on average no smaller than $n/\log n$.

Lemma A.1. *For the Bolthausen–Sznitman coalescent,* $a_n := \mathrm{E}(L_n) \geq n/\log n$ *for all* $n \geq 3$. *In particular,* $\lim_{n\to\infty} a_n = \infty$.

Proof. We use induction on n. For the Bolthausen–Sznitman coalescent, the block-counting process has rates $g_{nk} = n/((n-k)(n-k+1))$, $1 \leq k < n$, and total rates $g_n = n-1$, $n \in \mathbb{N}$. The recursion for the sequence $(a_n)_{n\in\mathbb{N}}$ yields $a_1 = 0$, $a_2 = 2$, and $a_3 = 3$. In particular, $a_3 \geq 3/\log(3) \approx 2.73$. Assume now that $a_k \geq k/\log k$ holds for all k, $3 \leq k < n$, for some fixed $n \geq 4$. Then $a_k \geq k/\log n$ for all k, $2 \leq k < n$, and, hence,

$$a_n = \frac{n}{g_n} + \sum_{k=2}^{n-1} r_{nk} a_k \geq \frac{n}{g_n} + \frac{1}{\log n} \sum_{k=2}^{n-1} k r_{nk}.$$

The substitution $l = n - k$ yields

$$\begin{aligned}
\sum_{k=2}^{n-1} k r_{nk} &= \sum_{l=1}^{n-2} (n-l) r_{n,n-l} = \frac{n}{n-1} \sum_{l=1}^{n-2} \frac{n-l}{l(l+1)} \\
&= \frac{n}{n-1}\left(n \sum_{l=1}^{n-2} \frac{1}{l(l+1)} - \sum_{l=1}^{n-2} \frac{1}{l+1}\right) \\
&= \frac{n}{n-1}\left(n\left(1 - \frac{1}{n-1}\right) - h_{n-1} + 1\right) \\
&= \frac{n}{n-1}\left(n - 1 - \frac{1}{n-1} - h_{n-1} + 1\right) \\
&= n + \frac{n}{n-1}\left(-\frac{1}{n-1} - h_{n-1} + 1\right),
\end{aligned}$$

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

where $h_n := \sum_{i=1}^{n} 1/i$ denotes the nth harmonic number. Thus, we have

$$\begin{aligned} a_n &\geq \frac{n}{g_n} + \frac{n}{\log n} + \frac{n}{(n-1)\log n}\biggl(-\frac{1}{n-1} - h_{n-1} + 1\biggr) \\ &= \frac{n}{\log n} + \frac{n(\log n - 1/(n-1) - h_{n-1} + 1)}{(n-1)\log n} \\ &\geq \frac{n}{\log n}, \end{aligned}$$

since

$$\log n - \frac{1}{n-1} - h_{n-1} + 1 \geq -\gamma - \frac{1}{n-1} + 1 \geq 0$$

for $n \geq 4$, where $\gamma \approx 0.577\,216$ denotes the Euler constant.

Acknowledgements

The author would like to thank Jason Schweinsberg for a number of helpful suggestions and Aleksander Iksanov for several remarks, in particular for pointing out the characterization of the limiting variable S in terms of an exponential integral of a subordinator.

References

[1] Berestycki, J., Berestycki, N. and Schweinsberg, J. (2005). Small-time behavior of beta-coalescents. Preprint.

[2] Berestycki, J., Berestycki, N. and Schweinsberg, J. (2006). Beta-coalescents and continuous stable random trees. Preprint.

[3] Bertoin, J. and Le Gall, J.-F. (2003). Stochastic flows associated to coalescent processes. *Prob. Theory Relat. Fields* **126,** 261–288.

[4] Bertoin, J. and Yor, M. (2001). On subordinators, self-similar Markov processes and some factorizations of the exponential variable. *Electron. Commun. Prob.* **6,** 95–106.

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

[6] Brandt, A. (1986). The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Adv. Appl. Prob.* **18,** 211–220.

[7] Carmona, P., Petit, F. and Yor, M. (1997). On the distribution and asymptotic results for exponential functionals of Lévy processes. In *Exponential Functionals and Principal Values Related to Brownian Motion*, ed. M. Yor, Biblioteca de la Revista Matematica Iberoamericana, Madrid, pp. 73–121.

[8] Durrett, R. and Schweinsberg, J. (2005). A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stoch. Process. Appl.* **115,** 1628–1657.

[9] Goldschmidt, C. and Martin, J. B. (2005). Random recursive trees and the Bolthausen–Sznitman coalescent. *Electron. J. Prob.* **10,** 718–745.

[10] Gut, A. (2003). On the moment problem for random sums. *J. Appl. Prob.* **40,** 797–802.

[11] Hedgecock, D. (1994). Does variance in reproductive success limit effective population sizes of marine organisms? In *Genetics and Evolution of Aquatic Organisms*, ed. A. Beaumont, Chapman and Hall, London, pp. 122–134.

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

[13] Kingman, J. F. C. (1982). On the genealogy of large populations. In *Essays in Statistical Science* (J. Appl. Prob. Spec. Vol. **19A**), eds J. Gani and E. J. Hannan, Applied Probability Trust, Sheffield, pp. 27–43.

[14] Kingman, J. F. C. (1982). The coalescent. *Stoch. Process. Appl.* **13,** 235–248.

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

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

[17] Neininger, R. and Rüschendorf, L. (2004). On the contraction method with degenerate limit equation. *Ann. Prob.* **32,** 2838–2856.

[18] Pitman, J. (1999). Coalescents with multiple collisions. *Ann. Prob.* **27,** 1870–1902.

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press

[19] Rösler, U. (1991). A limit theorem for 'Quicksort'. *RAIRO Inf. Théoret. Appl.* **25,** 85–100.

[20] Rösler, U. (1992). A fixed point theorem for distributions. *Stoch. Process. Appl.* **42,** 195–214.

[21] Rösler, U. and Rüschendorf, L. (2001). The contraction method for recursive algorithms. *Algorithmica* **29,** 3–33.

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

[23] Schweinsberg, J. (2000). A necessary and sufficient condition for the Λ-coalescent to come down from infinity. *Electron. Commun. Prob.* **5,** 1–11.

[24] Schweinsberg, J. (2000). Coalescents with simultaneous multiple collisions. *Electron. J. Prob.* **5,** 1–50.

[25] Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595.

[26] Vervaat, W. (1979). On a stochastic difference equation and a representation of non-negative infinitely divisible random variables. *Adv. Appl. Prob.* **11,** 750–783.

[27] Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoret. Pop. Biol.* **7,** 256–276.

https://doi.org/10.1239/aap/1158685000 Published online by Cambridge University Press