Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2012, Vol. 48, No. 3, 706–720
DOI: 10.1214/11-AIHP436

ANNALES
DE L'INSTITUT
HENRI
POINCARÉ
PROBABILITÉS
ET STATISTIQUES

www.imstat.org/aihp

Genealogies of regular exchangeable coalescents with applications to sampling

Vlada Limic[1]

39, rue F. Joliot-Curie, UMR 6632 CMI-LATP, CNRS and Université de Provence, Marseille, France. E-mail: vlada@cmi.univ-mrs.fr

Received 28 June 2010; revised 3 May 2011; accepted 9 May 2011

Abstract. This article considers a model of genealogy corresponding to a regular exchangeable coalescent (also known as Ξ-coalescent) started from a large finite configuration, and undergoing neutral mutations. Asymptotic expressions for the number of active lineages were obtained by the author in a previous work. Analogous results for the number of active mutation-free lineages and the combined lineage lengths are derived using the same martingale-based technique. They are given in terms of convergence in probability, while extensions to convergence in moments and convergence almost surely are discussed. The above mentioned results have direct consequences on the sampling theory in the Ξ-coalescent setting. In particular, the regular Ξ-coalescents that come down from infinity (i.e., with locally finite genealogies) have an asymptotically equal number of families under the corresponding infinite alleles and infinite sites models. In special cases, quantitative asymptotic formulae for the number of families that contain a fixed number of individuals can be given.

Résumé. Cet article considère un modèle de généalogie qui correspond à un processus de coalescence échangeable régulier (appelé aussi un Ξ-coalescent) démarré d'une configuration à taille finie et grande, et subissant des mutations neutres. Des expressions asymptotiques pour le nombre de lignées actives ont été obtenues par l'auteur dans un travail précédent. Des résultats analogues pour le nombre de lignées actives et la longueur totale des lignées sont dérivés par les mêmes techniques martingales. Ils sont donnés en terme de la convergence en probabilité, pendant que des extensions à la convergence au sens des moments et la convergence presque sûre sont examinées. Ces résultats ont des conséquences directes sur la théorie d'échantillonnage dans le cadre de Ξ-coalescence. En particulier, les Ξ-coalescents réguliers qui descendent de l'infini (c.-à-d. qui ont des généalogies localement finies) ont des nombres de familles égaux au sens asymptotique sous le modèle d'allèles infinies et le modèle de site infinis. Dans des cas particuliers, on peut ainsi dériver des formules asymptotiques quantitatives pour le nombre de familles contenant un nombre fixe d'individus.

MSC: 60J25; 60F99; 92D25

Keywords: Exchangeable coalescents; Ξ-coalescent; Λ-coalescent; regularity; sampling formula; small-time asymptotics; coming down from infinity; martingale technique; random mutation rate

1. Introduction

Kingman's coalescent [15,16] is one of the central models of mathematical population genetics. From the theoretical perspective, its importance is linked to the duality with the Fisher–Wright diffusion (and more generally with the Fleming–Viot process). Therefore the Kingman coalescent emerges in the scaling limit of genealogies of all evolutionary models that are asymptotically linked to Fisher–Wright diffusions. From the practical perspective, its elementary nature allows for exact computations and fast simulation, making it amenable to statistical analysis.

[1] Supported in part by ANR MANEGE grant.

Assume that the original sample has n individuals, labeled $\{1, 2, \ldots, n\}$. The *genealogy* is a process in which ancestral lineages coalesce in continuous time. One can identify the original sample with the trivial partition $\{\{1\}, \ldots, \{n\}\}$, and moreover, at any positive time, one can identify each of the active ancestral lineages with a unique equivalence class of $\{1, 2, \ldots, n\}$ that consists of the labels of all the individuals that descend from this lineage. In this way, each coalescent event of two (or more) ancestral lineages can be perceived as the merging of the corresponding equivalence classes. Ignoring the partition structure information, one can now view the coalescent as a *block* (rather than equivalence class) merging process.

Kingman's coalescent corresponds to the dynamics where each pair of blocks coalesces at rate 1. In particular, while there are n blocks present in the configuration, the total number of blocks decreases by 1 at rate $\binom{n}{2}$. In 1972 Ewens [12] derived a sampling formula that holds for several neutral population evolution models, and in particular for the Kingman coalescent. Its importance is well indicated by almost 900 scientific citations in less than four decades.[2] The Ewens sampling formula will be recalled in Section 3.1.

The fact that in the Kingman coalescent dynamics only pairs of blocks can merge at any given time makes it less suitable to model evolution of marine populations or viral populations under strong selection. In fact, it is believed (and argued to be observed in experiments, see e.g. [17]) that in such settings the reproduction mechanism allows for a proportion of the population to have the same parent (i.e., first generation ancestor). This translates to having multiple collisions of the ancestral lineages in the corresponding coalescent mechanism.

A family of mathematical models with the above property was independently introduced and studied by Pitman [23] and Sagitov [24] (see also the remark on page 195 of Donnelly and Kurtz [8]) under the name *Λ-coalescents* or *coalescents with multiple collisions*. Almost immediately emerged an even more general class of models, named *Ξ-coalescents* or *coalescents with simultaneous multiple collisions* or *exchangeable coalescents*. The Ξ-coalescent processes were initially studied by Möhle and Sagitov [21], and introduced by Schweinsberg [25] in their full generality. In particular, it is shown in [21] that any limit of genealogies arising from a population genetics model with (time-homogeneous) exchangeable reproduction mechanism must be a Ξ-coalescent. For additional pointers to the recent coalescent literature see [2,7].

The present article can be considered as a sequel to Berestycki et al. [3] and Limic [18]. It demonstrates once again the power of martingale techniques in the study of exchangeable coalescents. In fact, the main result is proved by a variation of the technique from [3,18]. The primary interest of the current work however is its potential for applications in studies of sampling statistics.

Unlike [3,4,18], the present work does not focus primarily on the coalescents that "come down from infinity." Indeed, the starting configuration is finite in the current context, so as long as a certain regularity condition (see [18] or (R) at the beginning of Section 3) holds, the small-time asymptotic results derived here (in Theorem 1 and Proposition 2) apply. All the Λ-coalescents (as well as certain general exchangeable coalescents of particular interest in mathematical population genetics [11,27]) are regular in this sense (see also Remark 12 in [18]). If a regular Ξ-coalescent has a locally finite genealogy (or equivalently, if its standard version comes down from infinity), its small-time asymptotics determines the asymptotic growth of the corresponding combined lineage length, and in turn, the asymptotic number of families in the infinite alleles (resp. sites) model (see Theorem 3). It is worth pointing out that Proposition 2 is the first example of a "meta-theorem" that applies in the more general context of regular exchangeable coalescents that have (potentially random but) uniformly bounded rate (per unit lineage length) of mutation, recombination and/or migration (see also Remark 6). Some consequences on the asymptotic frequency spectrum are discussed in Section 3.1.

This paper inherits the basic notation from [18]. For the benefit of the reader we recall it here. The set of real numbers is denoted by $\mathbb{R}$ and $(0, \infty)$ by $\mathbb{R}_+$. For $a, b \in \mathbb{R}$, denote by $a \wedge b$ (resp. $a \vee b$) the minimum (resp. maximum) of the two numbers. Let

$$\Delta := \left\{(x_1, x_2, \ldots,):\ x_1 \geq x_2 \geq \cdots \geq 0, \sum_i x_i \leq 1\right\}, \tag{1}$$

be the infinite unit simplex. For $\mathbf{x} = (x_1, x_2, \ldots) \in \Delta$ and $c \in \mathbb{R}$, let

$$c\mathbf{x} = (cx_1, cx_2, \ldots).$$

[2] Source: ISI Web of Knowledge http://apps.isiknowledge.com/.

The notation log is reserved for the *natural* logarithm, that is, the inverse of $\mathbb{R} \ni x \mapsto \mathrm{e}^x \in (0,\infty)$. If f is a function, defined in a left-neighborhood $(s-\varepsilon, s)$ of a point s, denote by $f(s-)$ the left limit of f at s. Given two functions $f, g : \mathbb{R}_+ \to \mathbb{R}_+$, write $f = \mathrm{O}(g)$ if $\limsup f(x)/g(x) < \infty$, $f = \mathrm{o}(g)$ if $\limsup f(x)/g(x) = 0$, and $f \sim g$ if $\lim f(x)/g(x) = 1$. Furthermore, write $f = \Theta(g)$ if both $f = \mathrm{O}(g)$ and $g = \mathrm{O}(f)$. The point at which the limits are taken is determined from the context. If $\mathcal{F} = (\mathcal{F}_t, t \geq 0)$ is a filtration, and T a stopping time relative to $\mathcal{F}$, denote by $\mathcal{F}_T$ the standard filtration generated by T, see for example [10], p. 389.

The rest of the paper is organized as follows: the model of interest is described in Section 2, in Section 3 the main results are stated, followed by a discussion of some consequences, while the arguments are postponed until Section 4.

2. Definitions and preliminaries

2.1. *The model*

For the purposes of this note, the precise definition of exchangeable coalescent processes (and their standard infinite version) is not essential. An interested reader can consult any of the references [2,7,18,25] for details of the construction, and further properties.

Instead, we next construct a related "genealogical" process which, together with a suitable enrichment, will be in the focus of the present study. Suppose that Ξ is a probability measure on Δ. Assume that we are given a Poisson Point Process on $\mathbb{R}_+ \times \Delta$

$$\pi(\cdot) = \sum_{k\in\mathbb{N}} \delta_{t_k, \mathbf{x}_k}(\cdot), \tag{2}$$

with intensity measure $\mathrm{d}t \otimes \Xi'(\mathrm{d}\mathbf{x})/\sum_{i=1}^{\infty} x_i^2$, where $\Xi(d\mathbf{x}) = (1-\Xi'(\Delta))\delta_{(0,0,\ldots)}(\mathrm{d}\mathbf{x}) + \Xi'(\mathrm{d}\mathbf{x})$ and $\Xi'((0,0,\ldots)) = 0$, $\Xi'(\Delta) \in (0,1]$ (the trivial case $\Xi'(\Delta) = 0$ corresponds to the Kingman coalescent).

Let n be a finite (typically large) integer. The *n-genealogy* (associated to Ξ) evolves as follows: at the initial time 0, there are n branches present in the system, each having trivial length 0. As time increases, the length of each branch increases at unit rate. An atom $(t, \mathbf{x})$ of π influences the evolution as follows: at time t let each branch present in the system at time $t-$ choose an i.i.d. color from $\mathbb{N} \cup (0,1)$, independently of the past evolution, according to the common law

$$P_{\mathbf{x}}\big(\{i\}\big) = x_i, \quad i \geq 1 \quad \text{and} \quad P_{\mathbf{x}}(\mathrm{d}u) = \left(1 - \sum_{i=1}^{\infty} x_i\right) du, \quad u \in (0,1).$$

For each $j \geq 1$, all the branches of color j collapse immediately (at time t) into a single branch. In addition, each pair of branches coalesces at rate $1 - \Xi'(\Delta)$ independently of the above color and collapse procedure. The readers might notice that whenever $\int_\Delta \Xi'(\mathrm{d}\mathbf{x})/\sum_{i=1}^{\infty} x_i^2 = \infty$, the time-projection of the set of atoms of π is dense on any interval of positive length. This could a priori cause difficulties in the construction, however, the above branch growing and collapsing process is always well-defined, as can be directly verified (or check [18,25] for a similar construction of exchangeable coalescents). Furthermore, it is important to note that the just described construction can be coupled over $n \in \mathbb{N}$ in such a way that for any two m, n, where $m > n$, restricting the process started from m distinct branches onto the first n branches gives precisely the process started from n branches. We call this coupling the *full genealogy* or the *full Ξ-genealogy*.

Remark 1. *Setting $\Xi(\Delta) = 1$ is convenient but arbitrary. It is easily seen that the time scale of any exchangeable coalescent can be multiplied by a constant factor, so that the law of the resulting process matches that of an exchangeable coalescent corresponding to a driving measure Ξ of total mass* 1.

One can enrich the above construction as follows: let υ be a completely independent Poisson Point Process of marks that arrive at rate γ per unit length. This means that while there are m branches present in the system, the next mark (mutation) arrives at rate $m\gamma$, and when it arrives it is placed uniformly at random (and independently from everything else) onto one of the branches. The parameter γ is usually called the *mutation rate*. In this enriched

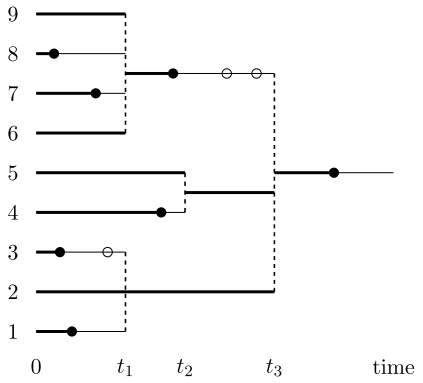

Fig. 1. A realization with $n = 9$ particles labeled by $\{1, \ldots, 9\}$, and coalescent events occurring at times t_1, t_2 and t_3. Open branches are indicated as thicker lines, and any mutation that falls onto an open (resp. closed) line is depicted as filled (resp. unfilled) circle.

construction, each branch is in one of the two states, *open* or *closed*, at any given time. Initially all the branches are open. A branch becomes closed starting from the moment a mark arrives to it, additional marks that possibly arrive afterward onto it do not change its state. Immediately after two or more branches collapse into one, the state of this newly formed branch is determined as follows: if at least one of the contributing branches is open, the new branch is open, otherwise it is closed. It is clear that one can couple this enriched genealogy again over $n \in \mathbb{N}$, and we call the resulting process the *full marked genealogy*.

Let $\mathcal{F} = (\mathcal{F}_t, t \geq 0)$ be the filtration generated by the full marked genealogy. Denote by $N \equiv N^n$ the branch counting process, and by $N^{\mathrm{o}} \equiv N^{n;\mathrm{o}}$ (resp. $N^{\mathrm{c}} \equiv N^{n;\mathrm{c}}$) the open (resp. closed) branch counting process. Note that, for each n, all the processes N^n, $N^{n;\mathrm{o}}$, $N^{n;\mathrm{c}}$ are $\mathcal{F}$-adapted. It is clear that

$$N(t) = N^{\mathrm{o}}(t) + N^{\mathrm{c}}(t), \quad t \geq 0.$$

Moreover, note that both N and N^{o}, unlike N^{c}, are monotone (decreasing) processes. In addition we have that, almost surely, for each $t \geq 0$ and $n \geq 1$,

$$N^n(t) \leq N^{n+1}(t) \quad \text{and} \quad N^{n;\mathrm{o}}(t) \leq N^{n+1;\mathrm{o}}(t),$$

but it is not necessarily true that $N^{n;\mathrm{c}}(t) \leq N^{n+1;\mathrm{c}}(t)$ (on Fig. 1, $N^{1;\mathrm{c}}(t_1) > N^{2;\mathrm{c}}(t_1)$). If for each $t > 0$ the family $(N^n(t), n \geq 1)$ is tight, or equivalently if $\lim_n N^n(t)$ exists and is a finite random variable, we will say that the full Ξ-genealogy is *locally finite*. This property is equivalent to the above mentioned coming down from infinity property for the (standard) Ξ-coalescent.

Call a mutation (or mark) that arrives onto an open (resp. closed) branch *open* (resp. *closed*). Furthermore, denote by $M \equiv M^n$ the mutation counting process, and by $M^{\mathrm{o}} \equiv M^{n;\mathrm{o}}$ (resp. $M^{\mathrm{c}} \equiv M^{n;\mathrm{c}}$) the open (resp. closed) mutation counting process.

Clearly,

$$M(t) = M^{\mathrm{o}}(t) + M^{\mathrm{c}}(t), \quad t \geq 0.$$

For the realization in Fig. 1, $N(t_1) = 4 = N^{\mathrm{o}}(t_1)$, while $M(t_1) = 5$ and $M^{\mathrm{o}}(t_1) = 4$. We also have $N^{\mathrm{c}}(t_1-) = 4$, while $N^{\mathrm{c}}(t_1) = 0$ and $N^{\mathrm{c}}(t_2-) = 2$.

Let

$$\tau^n \equiv \tau_1^n := \inf\bigl\{t \geq 0\colon\ N^n(t) = 1\bigr\}, \tag{3}$$

be the time of collapse to a single lineage, and

$$\tau_*^n := \inf\bigl\{t \geq \tau^n\colon\ \Delta\upsilon(t) > 0\bigr\},$$

be the arrival time of the first mutation to this unique lineage.

In [4] it was noted that $M(\tau^n_*) = M(\tau^n) + 1$ can be interpreted as the total number of "families" in the corresponding infinite sites model. For the realization on Fig. 1 we have $\tau^n = t_3$, and the "families" are $\{1\}, \{3\}, \{3\}^*, \{4\}, \{7\}, \{8\}, \{6, 7, 8, 9\}, \{6, 7, 8, 9\}^*, \{6, 7, 8, 9\}^{**}, \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, where the superscripts indicate the difference of the two families even though their contents, as subsets of $\mathbb{N}$, are identical.

Note in addition that $M^{\mathrm{o}}(\tau^n_*)$ can similarly be interpreted as the total number of families in the infinite alleles model. For the realization on Fig. 1, the (infinite alleles) families are $\{1\}, \{2, 5\}, \{3\}, \{4\}, \{6, 9\}, \{7\}, \{8\}$.

3. Main results

Suppose that the driving measure Ξ satisfies

$$\int_{\Delta} \frac{(\sum_{i=1}^{\infty} x_i)^2}{\sum_{i=1}^{\infty} x_i^2} \Xi(\mathrm{d}\mathbf{x}) < \infty. \tag{R}$$

This regularity condition already appeared in [18]. As observed in the introduction, all Λ-coalescents (the Kingman coalescent included) are regular.

Remark 2. *In view of these observations, the ratio $(\sum_i x_i)^2 / \sum_{i=1}^{\infty} x_i^2$ should be interpreted as 0 at $(0, 0, \ldots)$, so that the regularity of any Ξ-coalescent is determined by its "non-Kingman" part $\Xi'(d\mathbf{x}) = \Xi(d\mathbf{x})\mathbf{1}_{\Delta\setminus\{(0,0,\ldots)\}}$.*

Presently, (R) seems necessary for the martingale analysis below to work, although certain modifications of the technique might lead to stronger results, see Remark 22 in [18]. We refer the reader to Section 3.2 in [18] for examples of non-regular coalescents.

In analogy to (3) define

$$\tau^n_b := \inf\bigl\{t \geq 0\colon\ N^n(t) \leq b\bigr\}, \qquad \tau^{n;\mathrm{o}}_b := \inf\bigl\{t \geq 0\colon\ N^{n;\mathrm{o}}(t) \leq b\bigr\}. \tag{4}$$

It is easy to see that (cf. [3,18]) $\tau^n_b \nearrow \tau_b$, almost surely, where

$$\mathbb{P}\bigl(\tau_b \in (0, \infty]\bigr) = 1 \tag{5}$$

(note that τ_b takes value ∞ precisely in the cases where the genealogy is not locally finite). Similarly, one can show via the the full (marked) genealogy coupling that $\tau^{n;\mathrm{o}}_b \nearrow \tau^{\mathrm{o}}_b$, where

$$\mathbb{P}\bigl(\tau^{\mathrm{o}}_b \in (0, \infty]\bigr) = 1. \tag{6}$$

The fact that $\tau^{n;\mathrm{o}}_b$ is non-decreasing in n is a direct consequence of the coupling. To see that the limit cannot take value 0, note that $\{N^{n;\mathrm{o}}(s) > b\} = \{\tau^{n;\mathrm{o}}_b > s\}$, and that for each $s \geq 0$, $b \in \mathbb{N}$,

$$\mathbb{P}\bigl(N^{n;\mathrm{o}}(s) > b\bigr) \geq \mathbb{P}\bigl(N^n(s) > b,\ \text{the first } b+1 \text{ branches stay open during } [0, s]\bigr),$$

where the right-hand side is non-decreasing in n, and where its limit $p(s, b)$ (monotone decreasing in both variables) satisfies $\lim_{s\to 0} p(s, b) = 1$, $b \geq 1$.

Let $\alpha \in (0, 1/2)$ be arbitrary but fixed. The main results of this article are given next, starting with the most general ones.

Theorem 1. *Under (R), there exists $n_0 \in \mathbb{N}$ such that for each $t > 0$ and $\beta \in (0, \alpha \wedge 1 - 2\alpha)$*

$$\sup_{n\in\mathbb{N}} E\left(\frac{M^{n;\mathrm{c}}(t \wedge \tau^{n;\mathrm{o}}_{n_0})}{\gamma \int_0^{t\wedge\tau^{n;\mathrm{o}}_{n_0}} N^n(u)\,\mathrm{d}u}\right) = \mathrm{O}\bigl(t^{\alpha} + t^{1-2\alpha}\bigr), \tag{7}$$

$$\limsup_n \mathbb{P}\left(\frac{M^{n;\mathrm{c}}(t \wedge \tau^{n;\mathrm{o}}_{n_0})}{M^n(t \wedge \tau^{n;\mathrm{o}}_{n_0})} \geq t^{\beta}\right) = \mathrm{O}\bigl(t^{\alpha-\beta} \wedge t^{1-2\alpha-\beta}\bigr). \tag{8}$$

Remark 3. *Since the estimates are non-trivial only for $t \leq 1$, the optimal choice of α is $1/3$.*

This theorem is a consequence of the next result.

Proposition 2. *Under* (R), *there exists $n_0 \in \mathbb{N}$ such that for each $s > 0$*

$$P\left(\sup_{t\in[0,s]} \left| 1 - \frac{N^{n;\mathrm{o}}(t \wedge \tau^{n;\mathrm{o}}_{n_0})}{N^n(t \wedge \tau^{n;\mathrm{o}}_{n_0})} \right| > 8s^{\alpha} \right) = \mathrm{O}(s^{1-2\alpha}) \quad \textit{uniformly over } n \in \mathbb{N}. \tag{9}$$

Remark 4. *Note that it is not true in general that $\lim_{n\to\infty} M^{n;\mathrm{o}}(t)/M^n(t) = 1$ for a fixed $t > 0$. For example, in the case where Ξ is a Dirac measure $\delta_{(x,0,\ldots)}$ for some $x \in (0,1)$, all the original n branches remain in the genealogy for an exponential (rate 1) amount of time (regardless of n), therefore $N^n(t) \sim n$ and $N^{n;\mathrm{o}}(t) \sim n\mathrm{e}^{-\gamma t}$ at $t \approx 0$ (uniformly in n), implying $M^n(t) \sim nt$ and $M^{n;\mathrm{o}}(t) \sim (1 - \mathrm{e}^{-\gamma t})/\gamma n$. In this case, $\lim_{n\to\infty} M^{n;\mathrm{o}}(t)/M^{n,t} = f(t)$ where $\lim_{t\to 0} f(t) = 1$.*

All the Ξ-coalescents without proper frequencies (cf. Möhle [20]) are regular, but they do not have locally finite genealogy. In [20] it is shown that the appropriately scaled total lineage length of coalescents without proper frequencies converges in distribution to a non-trivial random variable.

However, if the underlying full genealogy is locally finite, one has $M^{n;\mathrm{o}}(t) \sim M^n(t)$ for a fixed time t. This result is stated next in a slightly different form, that may be more interesting from the perspective of applications. Define

$$\psi(q) \equiv \psi_{\Xi}(q) := \int_{\Delta} \frac{\sum_{i=1}^{\infty}(\mathrm{e}^{-qx_i} - 1 + qx_i)}{\sum_{i=1}^{\infty} x_i^2} \Xi(\mathrm{d}\mathbf{x}). \tag{10}$$

Note that the above integral converges since $\mathrm{e}^{-z} - 1 + z \leq z^2/2$, $z \geq 0$, and, in particular,

$$\psi_{\Xi}(q) \leq q^2/2. \tag{11}$$

Define $t \mapsto v^n(t) \in \mathbb{R}_+$ by

$$\int_{v^n(t)}^{n} \frac{\mathrm{d}q}{\psi(q)} = t, \tag{12}$$

and let

$$\ell(n) := \int_1^n \frac{q}{\psi(q)}\,\mathrm{d}q.$$

Theorem 3. *If the full genealogy is locally finite and* (R) *holds, then*

$$\lim_{n\to\infty} \frac{M^n(\tau^n)}{M^{n;\mathrm{o}}(\tau^n)} = \lim_{n\to\infty} \frac{M^n(\tau^n)}{\gamma \cdot \ell(n)} = \lim_{n\to\infty} \frac{M^{n;\mathrm{o}}(\tau^n)}{\gamma \cdot \ell(n)} = 1 \quad \textit{in probability.} \tag{13}$$

Remark 5. *In the Λ-coalescent setting (where (R) automatically holds) this result coincides with the initial part of Theorem 2 in [4]. The proof of this and related results in [4] is based on studying the asymptotic behavior of the arrival time of a (uniformly chosen at) random mutation, as $n \to \infty$. Under additional assumptions on Λ, this asymptotics can be quite precisely determined, leading to the convergence almost surely results in [4] (also discussed in the paragraph containing (15) in Section 3.1). The martingale-based technique presented here is not suitable for deducing precise information about the randomly chosen mutation. However, it is more compact than the technique from [4] (that still relies on martingale estimates from [3]), provides partial information even for non-locally finite genealogies, and is better suited for finding error bounds as well as extensions (like Theorem 4 or the case of random uniformly bounded mutation rates as discussed in Remark 6).*

Define

$$\ell_t(n) := \int_0^t v^n(u)\,\mathrm{d}u = \int_{v^n(t)}^n \frac{q}{\psi(q)}\,\mathrm{d}q.$$

Similar arguments as for Theorem 3 yield the following generalization.

Theorem 4. *If the full genealogy is locally finite and* (R) *holds, then for any bounded sequence* $(t_n)_{n\geq 1}$ *of positive numbers such that* $\ell_{t_n}(n)$ *diverges as* $n \to \infty$

$$\lim_{n\to\infty} \frac{M^n(t_n)}{M^{n;\mathrm{o}}(t_n)} = \lim_{n\to\infty} \frac{M^n(t_n)}{\gamma\cdot \ell_{t_n}(n)} = \lim_{n\to\infty} \frac{M^{n;\mathrm{o}}(t_n)}{\gamma\cdot \ell_{t_n}(n)} = 1 \quad \textit{in probability.}$$

3.1. *Consequences, extensions, and further comments*

Theorem 3 can be restated as follows: for any regular Ξ-coalescent that comes down from infinity, the number of families in the infinite alleles and the infinite sites model are asymptotically equal, in probability.

If more is known about Ξ, the convergence of Theorem 3 can be extended to convergence almost surely, using arguments very similar to those of Section 4.4.1 in [3]. Furthermore, one could prove that

$$\lim_{t\to 0} \frac{N^n(t)}{v^n(t)} = 1 \quad \text{in } L^p, \forall p \geq 1.$$

Theorems 3 and 4 could then be extended to convergence in the mean (see Remark 9 at the end of the article for details).

The asymptotics (13) holds (see Drmota et al. [9] and Basdevant and Goldschmidt [1]) also for the Bolthausen–Sznitman coalescent, which does not have locally finite genealogy. The author is convinced that the arguments of Theorem 3 could be extended to cover this special setting (where $\Lambda(\mathrm{d}x) = \Xi(\mathrm{d}(x,0,0,\ldots)) = \mathrm{d}x$, $x \in [0,1]$), and provide a shorter (and more generic) proof of this result (see Remark 8).

For each fixed $r \in \mathbb{N}$, set $M_r^n(t) := \#\{\text{mutations that arrive before time } t \text{ and affect precisely } r \text{ individuals}\}$, and $M_r^{n;\mathrm{o}}(t) := \#\{\text{open mutations that arrive before time } t \text{ and affect precisely } r \text{ individuals}\}$, and

$$M_r^n := M_r^n(\tau_*^n), \qquad M_r^{n;\mathrm{o}} := M_r^{n;\mathrm{o}}(\tau_*^n).$$

In the case of Kingman's coalescent, the Ewens sampling formula [12] gives the joint law of the above family of random variables as follows: given a_i, $i = 1,\ldots,n$ such that $a_i \geq 0$ and $\sum_{i=1}^n i a_i = n$,

$$P\big(M_i^{n,\mathrm{o}} = a_i, i = 1,\ldots,n\big) = \frac{n!}{2\gamma(2\gamma+1)\cdots(2\gamma+n-1)} \prod_{i=1}^n \frac{(2\gamma)^{a_i}}{i^{a_i} a_i!}. \tag{14}$$

For a general exchangeable coalescent one can only hope for asymptotic analogues to (14). The first such approximations were given by Berestycki et al. [5,6] for all the Beta-coalescents with locally finite genealogies.

Corollary 5. *Suppose that* $\ell(n) \sim^* n^\beta$ *for some* $\beta \in (0,1)$, *where* $\sim^*$ *stands for* $\sim$ *up to a slowly varying multiple.*

$$\textit{both } M_r^n \textit{ and } M_r^{n;\mathrm{o}} \textit{ are asymptotic to } \beta\Gamma(r-\beta)\ell(n)/r!, \textit{ in probability,} \tag{15}$$

where due to assumptions $\Gamma(z) = \int_0^\infty t^{z-1}\mathrm{e}^{-t}\,\mathrm{d}t$ *is well defined at* $r-\beta$, $r \in \mathbb{N}$.

Sketch of the proof. Note that both $M_r^n - M_r^n(\tau^n) \in \{0,1\}$, $M_r^{n;\mathrm{o}} - M_r^{n;\mathrm{o}}(\tau^n) \in \{0,1\}$, where these differences are positive for at most one $r \in \{1,\ldots,n\}$. One can combine Theorem 3 with Theorem 1.2 in Schweinsberg [26] to conclude the claim (see also [4]). □

A stronger version of this result (in the sense of almost sure convergence) was obtained recently in [4], applying the main result of [13], for a class of Λ-coalescent with "β-regular variation at 0," that comprises the class of the

above mentioned Beta-coalescents. However, $\ell(n)$ might be asymptotic to n^β even without the β-regular variation condition.

More generally, if $\ell(n_k) \sim^* (n_k)^\beta$ for some $\beta \in (0,1)$ along a given subsequence $(n_k)_{k\geq 1}$, we obtain the same asymptotics for $M_r^{n_k}$ and $M_r^{n_k;\mathrm{o}}$ as in Corollary 5. In some (rather vague) sense, for each fixed r, $M_r^{n_k}$ and $M_r^{n_k;\mathrm{o}}$ should both be close to $\frac{\log \ell(n_k)}{\log n_k}\Gamma(r - \frac{\log \ell(n_k)}{\log n_k})\ell(n_k)/r!$

One can generalize (15) in a different way: given any $(t_n)_{n\geq 1}$ such that $\ell_{t_n}(n) \sim^* n^\beta$, for $\beta \in (0,1)$, one can combine Theorem 4 again with Theorem 2 in [26] to conclude that both $M_r^n(t_n)$ and $M_r^{n;\mathrm{o}}(t_n)$ are asymptotic to $\beta\Gamma(r-\beta)\ell_{t_n}(n)/r!$, in probability. Again, this convergence could hold only along subsequences.

It would seem useful for applications to not only know the first order asymptotics but also the error terms, and there seems to be no good general approach for obtaining these. However, for specific Ξ (that is ψ) one could get concrete error bounds by redoing the general calculations (in the next section) with this given ψ.

It is not difficult to see that the asymptotic results given here depend only on the behavior of measure Ξ "close to $(0,0,\ldots)$". More precisely, if Ξ_1 and Ξ_2 are two measures on Δ satisfying

$$\Xi_1(\mathrm{d}\mathbf{x})\mathbf{1}_{\{\sum_i x_i \leq \varepsilon\}} = \Xi_2(\mathrm{d}\mathbf{x})\mathbf{1}_{\{\sum_i x_i \leq \varepsilon\}} \tag{16}$$

for some $\varepsilon > 0$, then the asymptotic quantities in Theorems 1–4 corresponding to Ξ_1 and Ξ_2 will be identical. This will continue to hold even if (16) is true only in the $\sim$ sense, in fact, as long as $\psi_{\Xi_1}(q) \sim \psi_{\Xi_2}(q)$ as $q \to \infty$.

What might be surprising (even disturbing) is that for any probability measure Ξ on Δ there exists a probability measure Λ on $[0,1]$ such that

$$\psi_\Xi(q) = \psi_\Lambda(q) \quad \forall q \geq 0,$$

where $\psi_\Lambda(q)$ is defined as $\int_{[0,1]}(\mathrm{e}^{-qx} - 1 + qx)\Lambda(\mathrm{d}x)/x^2$. The question of whether the measures Ξ and Λ can be connected in terms of a stochastic coupling construction remains open. To verify the above identity, one can check that $f := \psi'_\Xi$ is a completely monotone function, meaning that $(-1)^n \frac{\mathrm{d}^n}{\mathrm{d}q^n} f(q) \geq 0$ on $(0,\infty)$. Then due to Bernstein's theorem f is a Laplace transform of a positive Borel measure on $[0,\infty)$, and it can be written in the form

$$\psi'_\Xi(q) = f(q) = a + bq + \int_0^\infty \bigl(1 - \mathrm{e}^{-qx}\bigr)\mu(\mathrm{d}x), \tag{17}$$

where μ is a measure on $[0,\infty)$ such that $\int(1 \wedge x)\mu(\mathrm{d}x) < \infty$, and $\mu(\{0\}) = 0$. Differentiating (10) twice gives $\psi'_\Xi(0) = 0$ and $\psi''_\Xi(0) = 1$, so it must be $a = 0$ and

$$1 = b + \int_0^\infty x\mu(\mathrm{d}x).$$

Moreover, $\psi_\Xi(q)/q^2 \to \Xi((0,0,\ldots))/2$, thus (17) implies that $b = \Xi((0,0,\ldots))$. Then the above identity gives $\Xi((0,0,\ldots)) + \int_0^\infty x\mu(\mathrm{d}x) = 1$, so $\Lambda(\mathrm{d}x) := x\mu(\mathrm{d}x) + \Xi((0,0,\ldots))\delta_0(\mathrm{d}x)$ is a probability measure on $[0,\infty)$. By (17) and the fundamental theorem of integral calculus one now has identity

$$\psi_\Xi(q) = b\frac{q^2}{2} + \int_0^\infty \int_0^q \bigl(1 - \mathrm{e}^{-tx}\bigr)\mu(\mathrm{d}x) = b\frac{q^2}{2} + \int_0^\infty \frac{qx + \mathrm{e}^{-qx} - 1}{x}\mu(\mathrm{d}x).$$

The last quantity is identical to ψ_Λ, provided that the support of Λ (that is, of μ) is the unit interval. This last claim can be checked by using the inverse Laplace transform formula (see, e.g., [10], Example 5.4) together with the definition of ψ_Ξ.

4. The arguments

The proof of Proposition 2 is based on the martingale technique from [18], that originated in [3] in the Λ-coalescent setting. More precisely, Proposition 17 in [18] shows existence of $n_0 \in \mathbb{N}$ and $C \in (0,\infty)$, such that

$$E\bigl[\mathrm{d}\log\bigl(N^n(s)\bigr)|\mathcal{F}_s\bigr] = \left(-\frac{\psi(N^n(s))}{N^n(s)} + h^n(s)\right)\mathrm{d}s, \tag{18}$$

where $(h^n(s), s \geq z)$ is an $\mathcal{F}$-adapted process satisfying $\sup_{s\in[z,z\wedge\tau^n_{n_0}]} |h^n(s)| \leq C$, uniformly over n, and where

$$E\big[\big[\mathrm{d}\log\big(N^n(s)\big)\big]^2|\mathcal{F}_s\big]\mathbf{1}_{\{s\leq\tau^n_{n_0}\}} \leq C\,\mathrm{d}s \quad \text{almost surely.} \tag{19}$$

A crucial new observation is that $N^{n;\mathrm{o}}$ behaves analogously. Indeed, for $\mathbf{x} \in \Delta$, let $(X_j, j \in \mathbb{N})$ be a family of i.i.d. (generalized) random variables with law $\mathbb{P}_{\mathbf{x}}$, where

$$\mathbb{P}_{\mathbf{x}}(X_1 = i) = x_i, \quad i \in \mathbb{N}, \quad \text{and} \quad \mathbb{P}_{\mathbf{x}}(X_1 = \infty) = 1 - \sum_{i=1}^{\infty} x_i.$$

Furthermore, for each $b \in \mathbb{N}$, let the family $(Y^{(b)}_\ell, i \in \mathbb{N})$ of random variables be defined by

$$Y^{(b)}_\ell := \sum_{j=1}^{b} \mathbf{1}_{\{X_j=\ell\}}, \quad \ell \in \mathbb{N}. \tag{20}$$

Then we have, on the event $\{N^{n;\mathrm{o}}(s) = b\}$ (see the proof of [18], Proposition 17),

$$\begin{aligned} E\big(\mathrm{d}\log\big(N^{n;\mathrm{o}}(s)\big)|\mathcal{F}_s\big) = \int_\Delta E_{\mathbf{x}}\bigg[\log\frac{b - \sum_{\ell=1}^{\infty}(Y^{(b)}_\ell - \mathbf{1}_{\{Y^{(b)}_\ell>0\}})}{b}\bigg]\frac{1}{\sum_{i=1}^{\infty} x_i^2}\Xi'(\mathrm{d}\mathbf{x})\,\mathrm{d}s \\ + \big(1 - \Xi'(\Delta)\big)\binom{b}{2}\log\frac{b-1}{b} + \log\frac{b-1}{b}\cdot\gamma b. \end{aligned} \tag{21}$$

The first two terms in the drift above are identical to the only terms in the expression for the infinitesimal drift of $\log(N^n)$ (at time s on the event $\{N^n(s) = b\}$), leading to (18). The third term comes from the additional loss (that is, closure) of open branches at constant rate γ. Since $b\log((b-1)/b) = -1 + \mathrm{o}(1)$, one obtains

$$E\big[\mathrm{d}\log\big(N^{n;\mathrm{o}}(s)\big)|\mathcal{F}_s\big] = \bigg(-\frac{\psi(N^{n;\mathrm{o}}(s))}{N^{n;\mathrm{o}}(s)} + h^n_{\mathrm{o}}(s)\bigg)\mathrm{d}s, \tag{22}$$

for n_0, C from (18) and $C_{\mathrm{o}} = C + \gamma$, and where $\sup_{s\in[z,z\wedge\tau^{n;\mathrm{o}}_{n_0}]} |h^n_{\mathrm{o}}(s)| \leq C_{\mathrm{o}}$, uniformly over n. Moreover, since $b\log^2((b-1)/b) = \mathrm{o}(1)$, the bound from (19) holds with $N^{n;\mathrm{o}}$ (resp. C_{o}) in place of N^n (resp. C).

Recall the map v^n defined in (12). The argument of [18], Section 4.1, Part I already yields (cf. display (40) in [18])

$$P\bigg(\sup_{t\in[0,s]}\bigg|\log\frac{N^n(t\wedge\tau^n_{n_0})}{v^n(t\wedge\tau^n_{n_0})}\bigg| > 2s^\alpha\bigg) = \mathrm{O}\big(s^{1-2\alpha}\big), \tag{23}$$

but it clearly carries over to imply

$$P\bigg(\sup_{t\in[0,s]}\bigg|\log\frac{N^{n;\mathrm{o}}(t\wedge\tau^{n;\mathrm{o}}_{n_0})}{v^n(t\wedge\tau^{n;\mathrm{o}}_{n_0})}\bigg| > 2s^\alpha\bigg) = \mathrm{O}\big(s^{1-2\alpha}\big). \tag{24}$$

Since for $\varepsilon \in (0,1]$ we have $|\log(x)| \leq \varepsilon$ iff $1 - x \in [0, 1 - \mathrm{e}^{-\varepsilon}]$ implying $1 - x \in [0,\varepsilon]$, or $x - 1 \in [0, \mathrm{e}^\varepsilon - 1]$ implying $x - 1 \in [0, 2\varepsilon]$, one obtains (9) by combining the estimates in (23) and (24), and noting that $\tau^{n;\mathrm{o}}_{n_0} \leq \tau^n_{n_0}$, $\forall n, n_0$ in the full genealogy coupling. Alternatively, the same could be concluded by applying the argument leading to (23) and (24) in terms of the process $\log(N^n/N^{n;\mathrm{o}})$, that is, by comparing directly N^n and $N^{n;\mathrm{o}}$.

Remark 6. *It is important to note that if the mutation rate per unit length were not constant but given instead as a non-negative $\mathcal{F}$-adapted stochastic process $(\gamma_t, t \geq 0)$ such that $\mathbb{P}(\sup_{t\geq 0}\gamma_t \leq \gamma) = 1$ for some $\gamma < \infty$, then* (21)

would become

$$E\big(\mathrm{d}\log\big(N^{n;\mathrm{o}}(s)\big)|\mathcal{F}_s\big) = \int_\Delta E_{\mathbf{x}}\bigg[\log\frac{b-\sum_{\ell=1}^\infty (Y_\ell^{(b)} - \mathbf{1}_{\{Y_\ell^{(b)}>0\}})}{b}\bigg]\frac{1}{\sum_{i=1}^\infty x_i^2}\,\Xi'(\mathrm{d}\mathbf{x})\,\mathrm{d}s$$
$$+\big(1-\Xi'(\Delta)\big)\binom{b}{2}\log\frac{b-1}{b} + \log\frac{b-1}{b}\cdot\gamma_s\cdot b. \tag{25}$$

Since $\log\frac{b-1}{b}\cdot\gamma_s\cdot b\le\gamma$, *the rest of the proof of Proposition 2 would remain identical. Moreover, it is simple to see that the forthcoming arguments would easily carry over to yield appropriate analogues of Theorems 1, 3 and 4 in this general setting, where* $\gamma\int N^n(u)\,\mathrm{d}u$ *is being replaced by* $\int\gamma_u N^n(u)\,\mathrm{d}u$ *and* $\gamma\cdot\ell(n)$ *by* $\ell(n;\gamma):=\int_0^1\mathbb{E}(\gamma_u)v^n(u)\,\mathrm{d}u$, *under additional hypotheses on the process* γ *that would guarantee* $\int_0^1\gamma_u v^n(u)\,\mathrm{d}u\sim\ell(n;\gamma)$, *as well as the divergence of the sequence* $(\ell(n;\gamma))_{n\ge1}$. *For example, for estimate (26) one would use the stochastic domination with the Poisson (mean* $\gamma\int_0^t N^{n;\mathrm{c}}(u)\,\mathrm{d}u$*) law, instead of the law itself.*

As mentioned in the Introduction, Proposition 2 is just an example of a fairly general result that could be stated as follows: if the full genealogy dynamics is enriched so that in the new process the branches are called either "open" or "closed," and the open branch count process satisfies (25), then (9) holds. Two recent specific examples from this class of results are [14] (*verifying a weaker coming down from infinity criterion for a* Λ*-coalescent with migration model*) *and* [22] (*deriving the speed of coming down from infinity for the Kingman coalescent with constant recombination rate*). *The proof of this "meta-theorem" is essentially given in the previous paragraph.*

Proof of Theorem 1. Due to the construction of the full marked genealogy coupling, we know that, given the path of the processes N^n up to time t, $M^n(t)$ is a Poisson (mean $\gamma\int_0^t N^n(u)\,\mathrm{d}u$) random variable, that can be obtained as a sum of $M^{n;\mathrm{o}}(t)$ and $M^{n;\mathrm{c}}(t)$. Moreover, if $N^{n;\mathrm{o}}$ (that is $N^{n;\mathrm{c}}=N^n-N^{n;\mathrm{o}}$) is given in addition, then $M^{n;\mathrm{c}}(t)$ is a Poisson (mean $\gamma\int_0^t N^{n;\mathrm{c}}(u)\,\mathrm{d}u$) random variable, conditionally independent of $M^{n;\mathrm{o}}(t)$. The last observation is true if the fixed time t above is replaced by the random time $t\wedge\tau_{n_0}^{n;\mathrm{o}}$, measurable with respect to $\sigma\{N^n(u),N^{n,\mathrm{o}}(u),u\le t\}$. Note that $M^{n;\mathrm{o}}(t)$ cannot have (conditional) Poisson distribution, since $\mathbb{P}(M^{n;\mathrm{o}}(t)\le n)=1$. Note in addition that $M^n(t\wedge\tau_{n0}^n)$ is a Poisson (mean $\gamma\int_0^{t\wedge\tau_{n0}^n}N^n(u)\,\mathrm{d}u$) random variable, given $\mathcal{F}_{\tau_{n_0}^n}$. We cannot say the same if $\tau_{n_0}^n$ is replaced here by $\tau_{n_0}^{n;\mathrm{o}}$, since knowing $N^n(\tau_{n_0}^{n;\mathrm{o}})>n_0$ (that is, at least one branch is closed at time $\tau_{n_0}^{n;\mathrm{o}}$) excludes the event $\{M^n(\tau_{n_0}^{n;\mathrm{o}})=0\}$ that no mutation arrived prior to time $\tau_{n_0}^{n;\mathrm{o}}$. Due to this, some additional technical steps are needed below (notably, in arguing (27)–(28)).

Due to Proposition 2, on the event $A_t=\{N^{n;\mathrm{c}}(u)\le 8t^\alpha N^n(u),\forall u\in[0,t\wedge\tau_{n_0}^{n;\mathrm{o}}]\}=\{N^{n;\mathrm{o}}(u)\ge(1-8t^\alpha)N^n(u),\ \forall u\in[0,t\wedge\tau_{n_0}^{n;\mathrm{o}}]\}$ of probability greater than $1-\mathrm{O}(t^{1-2\alpha})$ we have

$$\int_0^{t\wedge\tau_{n_0}^{n;\mathrm{o}}}N^{n;\mathrm{c}}(u)\,\mathrm{d}u\le 8t^\alpha\cdot\int_0^{t\wedge\tau_{n_0}^{n;\mathrm{o}}}N^n(u)\,\mathrm{d}u.$$

On the complement of A_t we have $\int_0^{t\wedge\tau_{n_0}^{n;\mathrm{o}}}N^{n;\mathrm{c}}(u)\,\mathrm{d}u\le\int_0^{t\wedge\tau_{n_0}^{n;\mathrm{o}}}N^n(u)\,\mathrm{d}u$, since $N^{n;\mathrm{c}}\le N^n$ as a consequence of the full marked genealogy coupling. Hence we compute

$$E\bigg[\frac{M^{n;\mathrm{c}}(t\wedge\tau_{n_0}^{n;\mathrm{o}})}{\int_0^{t\wedge\tau_{n_0}^{n;\mathrm{o}}}N^n(u)\,\mathrm{d}u}\bigg]=E\bigg[E\bigg(\frac{M^{n;\mathrm{c}}(t\wedge\tau_{n_0}^{n;\mathrm{o}})}{\int_0^{t\wedge\tau_{n_0}^{n;\mathrm{o}}}N^n(u)\,\mathrm{d}u}\bigg|\big((N^n(u),N^{n;\mathrm{o}}(u)\big),u\in\big[0,t\wedge\tau_{n_0}^{n;\mathrm{o}}\big]\big)\bigg)\bigg]$$
$$=E\bigg[\frac{\gamma\int_0^{t\wedge\tau_{n_0}^{n;\mathrm{o}}}N^{n;\mathrm{c}}(u)\,\mathrm{d}u}{\int_0^{t\wedge\tau_{n_0}^{n;\mathrm{o}}}N^n(u)\,\mathrm{d}u}\bigg]$$
$$\le\gamma\big(8t^\alpha\mathbb{P}(A_t)+\mathbb{P}\big(A_t^c\big)\big), \tag{26}$$

and the estimate (7) readily follows.

In order to obtain (8), we first note that

$$\left\{M^n\big(t\wedge\tau^{n;\mathrm{o}}_{n_0}\big)<\frac{\gamma}{4}\int_0^{t\wedge\tau^{n;\mathrm{o}}_{n_0}}N^n(u)\,\mathrm{d}u\right\}\subset\left\{M^n\big(t\wedge\tau^{n}_{n_0}\big)<\frac{\gamma}{2}\int_0^{t\wedge\tau^{n}_{n_0}}N^n(u)\,\mathrm{d}u\right\}$$
$$\cup\left\{M^n\big(t\wedge\tau^{n}_{n_0}\big)-M^n\big(t\wedge\tau^{n;\mathrm{o}}_{n_0}\big)>\frac{\gamma}{4}\int_0^{t\wedge\tau^{n;\mathrm{o}}_{n_0}}N^n(u)\,\mathrm{d}u\right\}.$$

The estimate (8) will now simply follow from (7), the Markov inequality, and

$$\limsup_n\mathbb{P}\left(M^n\big(t\wedge\tau^{n}_{n_0}\big)<\frac{\gamma}{2}\int_0^{t\wedge\tau^{n}_{n_0}}N^n(u)\,\mathrm{d}u\right)=0,\tag{27}$$

$$\limsup_n\mathbb{P}\left(M^n\big(t\wedge\tau^{n}_{n_0}\big)-M^n\big(t\wedge\tau^{n;\mathrm{o}}_{n_0}\big)>\frac{\gamma}{4}\int_0^{t\wedge\tau^{n;\mathrm{o}}_{n_0}}N^n(u)\,\mathrm{d}u\right)=0.\tag{28}$$

To obtain (27), we first note that

$$\mathbb{P}\left(M^n\big(t\wedge\tau^{n}_{n_0}\big)<\frac{\gamma}{2}\int_0^{t\wedge\tau^{n}_{n_0}}N^n(u)\,\mathrm{d}u\,\Big|\,\big(N^n(u),u\le t\wedge\tau^{n}_{n_0}\big)\right)\le E\big(\mathrm{e}^{-c\int_0^{t\wedge\tau^{n}_{n_0}}N^n(u)\,\mathrm{d}u}\big),$$

for some $c>0$, uniformly over n (and γ), and also that $(\int_0^{t\wedge\tau^{n}_{n_0}}N^n(u)\,\mathrm{d}u)_{n\ge1}$ diverges, almost surely, as $n\to\infty$. Indeed, this sequence of random variables is monotone increasing in n, so it must be converging, almost surely, to a possibly generalized (i.e., taking value ∞) random variable. Moreover, the divergence with probability 1 is clear if the full genealogy is not locally finite. Otherwise, note that $(\int_0^t N^n(u)\,\mathrm{d}u)_{n\ge1}$ diverges while $(\int_{t\wedge\tau^{n}_{n_0}}^t N^n(u)\,\mathrm{d}u)_{n\ge1}$ converges, almost surely, as $n\to\infty$. The latter statement is clear due to (5). The former follows from the fact that $\lim_n v^n(t)=v(t)<\infty$, the inequality (11), the asymptotics (23), and

$$\int_0^t v^n(u)\,\mathrm{d}u=\int_{v^n(t)}^n\frac{q}{\psi(q)}\,\mathrm{d}q\ge\int_{v^n(t)}^n\frac{2}{q}\,\mathrm{d}q\approx\int_{v(t)}^\infty\frac{2}{q}\,\mathrm{d}q.\tag{29}$$

The first identity above, as noted in [4], is due to the change of variables $q=v^n(u)$, $\mathrm{d}q=-\psi(v^n(u))\,\mathrm{d}u$.

To obtain (28), note that $M^n(t\wedge\tau^{n}_{n_0})-M^n(t\wedge\tau^{n;\mathrm{o}}_{n_0})$ is a Poisson (mean $\gamma\int_{t\wedge\tau^{n;\mathrm{o}}_{n_0}}^{t\wedge\tau^{n}_{n_0}}N^n(u)\,\mathrm{d}u$) random variable given $\mathcal{F}_{\tau^{n}_{n_0}}$. Moreover, in the setting of locally finite full genealogy, due to (6), the sequence $(\int_{t\wedge\tau^{n;\mathrm{o}}_{n_0}}^{t\wedge\tau^{n}_{n_0}}N^n(u)\,\mathrm{d}u)_{n\ge1}$ is convergent (tight), while $(\int_0^{t\wedge\tau^{n;\mathrm{o}}_{n_0}}N^n(u)\,\mathrm{d}u)_{n\ge1}$ is always divergent, implying (28). Otherwise, both $\tau^{n;\mathrm{o}}_{n_0}$ and $\tau^{n}_{n_0}$ diverge to ∞ as $n\to\infty$, almost surely, so $M^n(t\wedge\tau^{n}_{n_0})-M^n(t\wedge\tau^{n;\mathrm{o}}_{n_0})=0$, except on an event of negligible probability, again implying (28). □

Proof of Theorem 3. The same change of variables as in (29) gives

$$\int_0^{(v^n)^{-1}(1)}v^n(u)\,\mathrm{d}u=\int_1^n\frac{q}{\psi(q)}\,\mathrm{d}q=\ell(n),$$

where

$$(v^n)^{-1}(1)=\int_1^n\frac{\mathrm{d}q}{\psi(q)}.$$

The hypotheses of locally finite full genealogy and regularity imply (see [25], Proposition 33 or [18], Theorem 1) that $\int_a^\infty\mathrm{d}q/\psi(q)<\infty$ (for any $a>0$), and therefore that $(v^n)^{-1}(1)$ increases towards a finite limit 1^*. As a consequence, $\ell(n)$ diverges as $n\to\infty$ (see (29)).

Remark 7. *Note that* $\ell(n) \sim \int_0^a v^n(u)\,\mathrm{d}u$ *for any fixed* $a > 0$.

It suffices to prove the convergence in probability for $M^n(\tau^n)/(\gamma\ell(n))$ and $M^{n;\mathrm{o}}(\tau^n)/(\gamma\ell(n))$. We discuss the convergence of the former sequence in some detail below, and give at present the reasoning for the latter convergence assuming the former: since $M^n(\tau^n) = M^{n;\mathrm{o}}(\tau^n) + M^{n;\mathrm{c}}(\tau^n)$, it suffices to show that $M^{n;\mathrm{c}}(\tau^n) = \mathrm{o}(\ell(n))$, in probability. This will follow by (8), provided $M^{n;\mathrm{c}}(\tau^n) - M^{n;\mathrm{c}}(\tau_{n_0}^{n;\mathrm{o}} \wedge t) = \mathrm{o}(\ell(n))$ for some fixed t. However, $M^{n;\mathrm{c}}(\tau^n) - M^{n;\mathrm{c}}(\tau_{n_0}^{n;\mathrm{o}} \wedge t) \le M^n(\tau^n) - M^n(\tau_{n_0}^{n;\mathrm{o}} \wedge t)$ in the full (marked) genealogy coupling, and $M^n(\tau^n) - M^n(\tau_{n_0}^{n;\mathrm{o}} \wedge t)$ is (as in the proof of Proposition 2) a conditional Poisson random variable with mean $\gamma \cdot \int_{\tau_{n_0}^{n;\mathrm{o}} \wedge t}^{\tau^n} N^n(u)\,\mathrm{d}u$. So the above claim follows due to convergence (tightness) of $\int_{\tau_{n_0}^{n;\mathrm{o}} \wedge t}^{\tau^n} N^n(u)\,\mathrm{d}u$ and divergence of $\ell(n)$.

The rest of the argument is analogous to that for Theorem 5 in [3]. The bulk of it is showing that $\int_0^{\tau^n} N^n(u)\,\mathrm{d}u \sim \ell(n)$, in probability, as $n \to \infty$. Since, given N^n, $M^n(\tau^n)$ is a Poisson random variable with mean $\int_0^{\tau^n} N^n(u)\,\mathrm{d}u$, and since $\ell(n)$ diverges with n, one concludes (as a special case of LLN) that $M^n(\tau^n) \sim \gamma\ell(n)$, in probability. It therefore suffices to show that for any subsequence n_k there exists a further subsequence n_{k_j} such that $\int_0^{\tau^{n_{k_j}}} N^{n_{k_j}}(u)\,\mathrm{d}u/\ell(n_{k_j}) \to 1$, almost surely.

To simplify the notation, we rename the subsequence $(n_k)_{k\ge1}$ as $(k)_{k\ge1}$. Recall (23), and for a fixed $\alpha \in (0, 1/2)$ choose a decreasing sequence $(s_k)_{k\ge1}$ of positive numbers, such that

$$\sum_k s_k^{1-2\alpha} < \infty.$$

Then we can conclude from (5), (23) and the Borel–Cantelli lemma that

$$\lim_{k\to\infty} \sup_{t\in[0,s_k]} \left| \frac{N^k(t)}{v^k(t)} - 1 \right| = 0 \quad \text{almost surely.} \tag{30}$$

Due to the assumption of locally finite full genealogy, we can now choose a subsequence k_j such that $\int_0^{s_j} v^{k_j}(u)\,\mathrm{d}u$ diverges as $j \to \infty$, and that also both

$$\int_{s_j}^{1^*} v^{k_j}(u)\,\mathrm{d}u \Big/ \int_0^{s_j} v^{k_j}(u)\,\mathrm{d}u \quad \text{and} \quad \int_{s_j}^{\tau^{k_j}} N^{k_j}(u)\,\mathrm{d}u \Big/ \int_0^{s_j} N^{k_j}(u)\,\mathrm{d}u \tag{31}$$

tend to 0 as $j \to \infty$ (for the second sequence, the limit is taken almost surely). Joint with (30), this ensures that

$$\lim_{j\to\infty} \frac{\int_0^{\tau^{k_j}} N^{k_j}(u)\,\mathrm{d}u}{\int_0^{1^*} v^{k_j}(u)\,\mathrm{d}u} = \lim_{j\to\infty} \frac{\int_0^{s_j} N^{k_j}(u)\,\mathrm{d}u}{\int_0^{s_j} v^{k_j}(u)\,\mathrm{d}u} = 1 \quad \text{almost surely.}$$

□

Remark 8. *If* $\Xi(\mathrm{d}\mathbf{x}) = \mathbf{1}_{\{\mathbf{x}=(x,0,0,\ldots)\}}$ *for* $x \in (0, 1)$, *or equivalently, in case of the Bolthausen–Sznitman coalescent we have*

$$\psi(q) = q\log q + \mathrm{O}(q) \quad \text{as } q \to \infty,$$

with $\mathrm{O}(q) \ge 0$, *for all* $q > 0$. *Therefore*

$$v^n(t) \in \left[n^{\mathrm{e}^{-t(1+o(1))}}, n^{\mathrm{e}^{-t}}\right] \quad \text{as well as} \quad \ell(n) \sim \frac{n}{\log n} \quad \text{as } n \to \infty.$$

We conclude that $v^n(t)$ *is of order* 1 *at times of order* $\log\log(n)$ *and in turn that*

$$\int_1^{\log\log n} v^n(t)\,\mathrm{d}t = \mathrm{o}\big(\ell(n)\big) \quad \text{as } n \to \infty. \tag{32}$$

In order to show that $\int_0^{\tau^n} N^n(u)\,\mathrm{d}u \sim \ell(n)$, *it therefore suffices to start as in the paragraph which comprises* (30) *and* (31), *ensuring the following analogue of* (31)

$$\int_{s_j}^1 v^{k_j}(u)\,\mathrm{d}u \Big/ \int_0^{s_j} v^{k_j}(u)\,\mathrm{d}u \to 0 \quad \textit{and} \quad \int_{s_j}^1 N^{k_j}(u)\,\mathrm{d}u \Big/ \int_0^{s_j} N^{k_j}(u)\,\mathrm{d}u \to 0,$$

and to verify in addition that (*possibly along a further subsequence*)

$$\int_{1\wedge\tau^{k_j}}^{\tau^{k_j}} N^{k_j}(t)\,\mathrm{d}t = \mathrm{o}\big(\ell(k_j)\big) \quad \textit{as } j\to\infty. \tag{33}$$

This can all be done due to (23), *the fact that* $v^n(t)$ *is bounded by a power of* n *smaller than* 1 *for each fixed* t, *asymptotically in* n, *and finally the estimate*

$$E\tau^n = \mathrm{O}(\log\log n),$$

which can be obtained via optional stopping of the martingale $\bar{M}_t := \int_{N^n(t)}^n \frac{\mathrm{d}q}{\bar{\psi}(q)} - t$, $t\ge 0$, *where* $\bar{\psi}(q) := \int_{[0,1]}((1-x)^q - 1 + qx)/x^2 = \psi(q) + \mathrm{O}(1)$ (*see* [19] *for further use of* $\bar{M}$ *and its generalizations*).

Proof of Theorem 4. Due to the assumption that $\ell_{t_n}(n)$ diverges, one can argue as in the proof of Theorem 3 that it suffices to show that

$$\frac{M^n(t_n)}{\gamma\cdot\ell_{t_n}(n)} \to 1 \quad \text{in probability.} \tag{34}$$

The argument for (34) is analogous to the last one. In fact, with the same choice of the sequence $(s_k)_{k\ge1}$ as above, one now chooses a subsequence k_j so that $\int_0^{s_j\wedge t_j} v^{k_j}(u)\,\mathrm{d}u$ diverges as $j\to\infty$, and in addition both

$$\int_{s_j}^{t_j} v^{k_j}(u)\,\mathrm{d}u\mathbf{1}_{\{t_j>s_j\}} \Big/ \int_0^{s_j} v^{k_j}(u)\,\mathrm{d}u \quad \text{and} \quad \int_{sj}^{t_j} N^{k_j}(u)\,\mathrm{d}u\mathbf{1}_{\{t_j>s_j\}} \Big/ \int_0^{s_j} N^{k_j}(u)\,\mathrm{d}u$$

tend to 0 as $j\to\infty$. Joint with (30) this ensures that

$$\lim_{j\to\infty} \frac{\int_0^{t_j} N^{k_j}(u)\,\mathrm{d}u}{\int_0^{t_j} v^{k_j}(u)\,\mathrm{d}u} = \lim_{j\to\infty} \frac{\int_0^{t_j\wedge s_j} N^{k_j}(u)\,\mathrm{d}u}{\int_0^{t_j\wedge s_j} v^{k_j}(u)\,\mathrm{d}u} = 1,$$

almost surely. Finally, (34) follows by another application of LLN. □

Remark 9. *The arguments of* [3], *Section* 4.3 *apply in the regular setting. Indeed, instead of* [3], *Lemma* 20 *one now has the following statement*: *There exists* $n_0\in\mathbb{N}$ *and* $K_0<\infty$ *such that for all* $b\ge n_0$, $\mathbf{x}\in\Delta\cap\{\mathbf{x}\colon \sum_{i=1}^\infty x_i\le 1/4\}$, $c>0$, *and* $Y_\ell^{(b)}$ *given by* (20) *we have*

$$E_{\mathbf{x}}\left[\exp\left\{c\left[\log\left(b-\sum_{\ell=1}^\infty\big(Y_\ell^{(b)}-\mathbf{1}_{\{Y_\ell^{(b)}>0\}}\big)\right)-\log b\right]^2\right\}-1\right] \le \mathrm{e}^{9c/4}K_0\left[\left(\sum_i x_i^2\right)+\left(\sum_i x_i\right)^2\right]. \tag{35}$$

Since $\sum_i x_i^2 \le (\sum_i x_i)^2$, *the RHS above could be simply bounded by* $2K_0(\sum_i x_i)^2$, *however, in doing so one might lose some information of the impact of regularity* (*or irregularity*). *Due to* (35), *in the definition of the process* $E^{(c)}$ *and the related calculations leading to* (33) *and* (34) *in* [3], *the constant* K_0 *should be replaced by*

$$\bar{K}_0 = K_0\left(1+\int\left(\sum_{i=1}^\infty x_i\right)^2\frac{\Xi(\mathrm{d}\mathbf{x})}{\sum_{i=1}^\infty x_i^2}\right).$$

In particular, one can conclude that for any $s > 0$ and $d \geq 1$

$$\sup_{n\geq 1} \mathbb{E}\left(\sup_{t\in[0,s]} \left| \frac{N^n(t)}{v^n(t)} \right|^d \right) = D(d,s) < \infty, \tag{36}$$

and moreover that $\lim_{s\to 0} D(d,s) = 0$. As indicated earlier, we then have $N^n(t)/v^n(t) \to 1$, as $t \to 0$ in L^d, for each $d \geq 1$.

As a consequence of the previous observations, and arguments very similar to those for Theorem 1, it is not difficult to check that, for each fixed $t > 0$,

$$\frac{E(M^n(t))}{\gamma \cdot \ell_t(n)} = \frac{E(\int_0^t N^n(u)\,\mathrm{d}u)}{\int_0^t v^n(u)\,\mathrm{d}u} \to 1, \tag{37}$$

as well as

$$\frac{E(M^{n;\mathrm{c}}(t))}{\gamma \cdot \ell_t(n)} = \frac{E(\int_0^t N^{n;\mathrm{c}}(u)\,\mathrm{d}u)}{\int_0^t v^n(u)\,\mathrm{d}u} \to 0,$$

implying (37) with $M^{n;\mathrm{o}}$ in place of M^n. Under the assumption of locally finite genealogies, $\ell_1(n) \sim \ell(n)$ and

$$E\big[M^n\big(\tau_1^n\big) - M^n(1)\big] = E\left[\gamma \int_1^{\tau_1^n} N^n(u)\,\mathrm{d}u\right] \leq \gamma\, E\big[N^n(1)\cdot\big(\tau_1^n - 1\big)\big] \leq C \quad \forall n \geq 1,$$

where the final uniform estimate is due to $\sup_n E(N^n(1)) < \infty$, and $\sup_n E(\tau_1^n - 1 | N^n(1)) \leq E(\sup_n \tau_1^n) < \infty$. Hence, for a locally finite Ξ-genealogy, both $M^n(\tau_1^n))/(\gamma\ell(n))$ and $M^{n;\mathrm{o}}(\tau_1^n)/(\gamma\ell(n))$ converge to 1 in the mean.

Acknowledgments

The author wishes to thank Matthias Birkner for a pointer to Bernstein's theorem, and to Julien and Nathanaël Berestycki for rewarding discussions. She is also grateful to the anonymous referee for several useful comments and pointers to the literature.

References

[1] A.-L. Basdevant and C. Goldschmidt. Asymptotics of the allele frequency spectrum associated with the Bolthausen–Sznitman coalescent. *Electron. J. Probab.* **13** (2008) 486–512. MR2386740

[2] N. Berestycki. *Recent Progress in Coalescent Theory. Ensaios matematicos [Mathematical Surveys]* **16**. Sociedade Brasileira de Matemática, Rio de Janeiro, 2009. MR2574323

[3] J. Berestycki, N. Berestycki and V. Limic. The Λ-coalescent speed of coming down from infinity. *Ann. Probab.* **38** (2010) 207–233. MR2599198

[4] J. Berestycki, N. Berestycki and V. Limic. Asymptotic sampling formulae and particle system representations for Λ-coalescents. Preprint. Available at http://www.cmi.univ-mrs.fr/~vlada/research.html, 2011.

[5] J. Berestycki, N. Berestycki and J. Schweinsberg. Beta-coalescents and continuous stable random trees. *Ann. Probab.* **35** (2007) 1835–1887. MR2349577

[6] J. Berestycki, N. Berestycki and J. Schweinsberg. Small-time behavior of beta-coalescents. *Ann. Inst. H. Poincaré Probab. Statist.* **44** (2008) 214–238. MR2446321

[7] J. Bertoin. *Random Fragmentation and Coagulation Processes*. Cambridge Univ. Press, Cambridge, 2006. MR2253162

[8] P. Donnelly and T. Kurtz. Particle representations for measure-valued population models. *Ann. Probab.* **27** (1999) 166–205. MR1681126

[9] M. Drmota, A. Iksanov, M. Möhle and U. Rösler. Asymptotic results concerning the total branch length of the Bolthausen–Sznitman coalescent. *Stochastic Process. Appl.* **117** (2007) 1404–1421. MR2353033

[10] R. Durrett. *Probability: Theory and Examples*, 3rd edition. *Duxbury Advanced Series*. Duxbury Press, Belmont, CA, 2004. MR1609153

[11] R. Durrett and J. Schweinsberg. A coalescent model for the effect of advantageous mutations on the genealogy of a population. Random partitions approximating the coalescence of lineages during a selective sweep. *Stochastic Process. Appl.* **115** (2005) 1628–1657. MR2165337

[12] W. J. Ewens. The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* **3** (1972) 87–112. MR0325177

[13] A. Gnedin, B. Hansen and J. Pitman. Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws. *Probab. Surv.* **4** (2007) 146–171. MR2318403
[14] C. Foucart. Distinguished exchangeable coalescents and generalized Fleming–Viot processes with immigration. Preprint. Available at http://arxiv.org/abs/1006.0581, 2011.
[15] J. F. C. Kingman. The coalescent. *Stochastic. Process. Appl.* **13** (1982) 235–248. MR0671034
[16] J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Probab.* **19** (1982) 27–43. MR0633178
[17] G. Li and D. Hedgecock. Genetic heterogeneity, detected by PCR SSCP, among samples of larval Pacific oysters (Crassostrea gigas) supports the hypothesis of large variance in reproductive success. *Can. J. Fish. Aquat. Sci.* **55** (1998) 1025–1033.
[18] V. Limic. On the speed of coming down from infinity for Ξ-coalescent processes. *Electron. J. Probab.* **15** (2010) 217–240. MR2594877
[19] V. Limic. Coalescent processes and reinforced random walks: A guide through martingales and coupling. Habilitation thesis. Available at http://www.cmi.univ-mrs.fr/~vlada/research.html, 2011.
[20] M. Möhle. Coalescent processes without proper frequencies and applications to the two-parameter Poisson–Dirichlet coalescent. Preprint, 2009.
[21] M. Möhle and S. Sagitov. A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.* **29** (2001) 1547–1562. MR1880231
[22] E. Pardoux and M. Salamat. On the height and length of the Ancestral Recombination Graph. *J. Appl. Probab.* **46** (2009) 669–689. MR2560895
[23] J. Pitman. Coalescents with multiple collisions. *Ann. Probab.* **27** (1999) 1870–1902. MR1742892
[24] S. Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.* **36** (1999) 1116–1125. MR1742154
[25] J. Schweinsberg. Coalescents with simultaneous multiple collisions. *Electron. J. Probab.* **5** (2000) 1–50. MR1781024
[26] J. Schweinsberg. The number of small blocks in exchangeable random partitions. *ALEA* **7** (2010) 217–242. MR2672786
[27] J. Schweinsberg and R. Durrett. Random partitions approximating the coalescence of lineages during a selective sweep. *Ann. Appl. Probab.* **15** (2005) 1591–1651. MR2152239