Splitting trees with neutral Poissonian mutations II: Largest and oldest families.

Nicolas Champagnat[1,2], Amaury Lambert[3]

December 4, 2012

Abstract

We consider a supercritical branching population, where individuals have i.i.d. lifetime durations (which are not necessarily exponentially distributed) and give birth (singly) at constant rate. We assume that individuals independently experience neutral mutations, at constant rate θ during their lifetimes, under the infinite-alleles assumption: each mutation instantaneously confers a brand new type, called allele or haplotype, to its carrier. The type carried by a mother at the time when she gives birth is transmitted to the newborn.

We are interested in the sizes and ages at time t of the clonal families carrying the most abundant alleles or the oldest ones, as $t \to \infty$, on the survival event. Intuitively, the results must depend on how the mutation rate θ and the Malthusian parameter $\alpha > 0$ compare. Hereafter, $N \equiv N_t$ is the population size at time t, constants a, c are scaling constants, whereas k, k' are explicit positive constants which depend on the parameters of the model.

When $\alpha > \theta$, the most abundant families are also the oldest ones, they have size $cN^{1-\theta/\alpha}$ and age $t - a$.

When $\alpha < \theta$, the oldest families have age $(\alpha/\theta)t + a$ and tight sizes; the most abundant families have sizes $k \log(N) - k' \log\log(N) + c$ and all have age $(\theta - \alpha)^{-1} \log(t)$.

When $\alpha = \theta$, the oldest families have age $kt - k' \log(t) + a$ and tight sizes; the most abundant families have sizes $(k \log(N) - k' \log\log(N) + c)^2$ and all have age $t/2$.

Those informal results can be stated rigorously in expectation. Relying heavily on the theory of coalescent point processes [13, 16], we are also able, when $\alpha \leq \theta$, to show convergence in distribution of the joint, properly scaled ages and sizes of the most abundant/oldest families and to specify the limits as some explicit Cox processes.

This is in deep contrast with the largest/oldest families in the standard coalescent with Poissonian mutations, which converge to some point processes after being rescaled by N [3, 4, 5].

MSC 2000 subject classifications: Primary 60J80; secondary 92D10, 60J85, 60G70, 60G51, 60G55, 60K15.

[1]Université de Lorraine, IECN, Campus Scientifique, B.P. 70239, Vandœuvre-lès-Nancy Cedex, F-54506, France

[2]Inria, Villers-lès-Nancy, F-54600, France, E-mail: `Nicolas.Champagnat@inria.fr`

[3]Laboratoire de Probabilités et Modèles Aléatoires, UMR 7599 CNRS and UPMC Univ Paris 06, Case courrier 188, 4 Place Jussieu, F-75252 Paris Cedex 05, France, Email: `amaury.lambert@upmc.fr`

Key words and phrases: branching process – coalescent point process – splitting tree – Crump-Mode-Jagers process – linear birth-death process – allelic partition – infinite alleles model – extreme values – mixed Poisson point process – Cox process – Lévy process – scale function.

1 Introduction and motivation

We consider a general branching population, where individuals reproduce independently of each other, have i.i.d. lifetime durations with arbitrary distribution, and give birth at constant rate during their lifetime. We also assume that each birth gives rise to a single newborn. The genealogical tree associated with this construction is known as a splitting tree [6, 7, 13]. The process $(N_t; t \geq 0)$ counting the population size is a non-Markovian birth–death process belonging to the class of general branching processes, or *Crump–Mode–Jagers* (CMJ) processes. Since births arrive singly and at constant rate, these processes are sometimes called *homogeneous, binary CMJ processes*.

Also, individuals are given a type, called allele or haplotype. They inherit their type at birth from their mother, and (their germ line) can change type throughout their lifetime, at the points of independent Poisson point processes with rate θ, conditional on lifetimes (neutral mutations). The type conferred by a mutation is each time an entirely new type, an assumption known as the *infinitely-many alleles model*.

We are interested in the so-called *allelic partition* (partition into types) of the population alive at time t. In [2, 12], we obtained explicit formulae for the expected *frequency spectrum* of the allelic partition. The frequency spectrum is a convenient way of describing this partition without labelling types. It is defined as the sequence of numbers $(A_\theta(k,t), k \geq 0)$, where $A_\theta(k,t)$ is the number of types carried by k individuals at time t. For example in [2], we have derived explicit formulae for the expectation of $A_\theta(k,t)$ conditional on population size at t. From these formulae, using the theory of branching processes counted by random characteristics, we have specified the a.s. limit, as $t \to \infty$, of the fraction of types carried by a fixed number k of individuals.

If we call *clonal families*, or simply *families*, the equivalence classes associated to identity by type (i.e., the components of the allelic partition), it is usual to call *small families* the families of sizes $k = 1, 2, 3, \ldots$ Here, *large families* will refer to families with most frequent (i.e., abundant) types having alive representatives, and *old families* to families having oldest types with alive representatives, where the age of a type is the time elapsed since its original mutation. In the present work, we are interested in the asymptotic behavior, as $t \to \infty$, of the sizes and ages of large and of old families.

The most celebrated mathematical result regarding allelic partitions is Ewens' sampling formula, which provides the distribution of the frequency spectrum for the Kingman coalescent tree with neutral Poissonian mutations [5]. It has notably been shown [3, 4] that under this model, the largest (resp. oldest) families converge, after being rescaled by the population size N, to the Poisson–Dirichlet (resp. GEM) distribution. We will see here, that, for example, the largest families are never of the order of N, but depending on how the Malthusian parameter α scales with the mutation rate θ, of the order of $N^{1-\theta/\alpha}$ (case $\theta < \alpha$), order of $(\log N)^2$ (case $\theta = \alpha$), or order of $\log N$ (case $\theta > \alpha$). The first case ($\theta < \alpha$) shows more similarities with the frequency spectrum of the Beta-coalescent [1].

We refer the reader to [2] for more references on the topic of allelic partitions, especially regarding branching processes. For example, similar questions were studied for general CMJ processes, when mutations occur *at birth*, in the monography due to Z. Taïb [18]. These results rely heavily on the theory of branching processes counted by random characteristics, due to P. Jagers and O. Nerman [8, 9, 10, 15], and more specifically on *time dependent random characteristics* as developed in [10]. Z. Taïb obtains results of convergence in distribution of the (correctly rescaled) point process of ages, similar to the results we obtain in Sections 4 and 5. However, the techniques of [18] do not apply to the case where mutations occur during individuals' lifetimes, since the genealogical tree of types is not the one of a CMJ process in this case.

One of the initial motivations of [2] and of the present work was the following model inspired from the works of P.C. Sabeti and her coauthors (see e.g. [17]). Inside a large population, consider a subpopulation consisting of individuals carrying a specific selective gene called 'core haplotype' and thus experiencing demographic growth. The haplotypic structure of the subpopulation restricted to a portion of length L around the core haplotype on the chromosome carrying it, is assumed to be altered by recombination. As long as the total population is sufficiently large w.r.t. the growing subpopulation, each time a sequence belonging to (an individual in) this subpopulation recombines with another sequence, with high probability this sequence will be a new sequence belonging to the rest of the population. Therefore, the new sequence obtained after recombination can be treated as a mutant under the infinitely-many alleles model. In this setting, mutation rate is an increasing function of L. In [17], a tree representation of the allelic partition as a function of L is given for each core haplotype in a given set of genes suspected to have been selected in humans. The tree obtained this way is called "recombination tree". An interesting question is to develop statistical methods allowing to detect positive selection from the knowledge of this tree. Here, we assume that our (sub)population grows at a Malthusian rate α (supercritical CMJ process). We are able to give the asymptotic distribution of the rightmost part of the frequency spectrum for a given mutation rate (this corresponds to fixing L in the recombination setting). Since θ can be seen as a death rate when restricting the count to individuals carrying the same allele, the phase transition at $\theta = \alpha$ is intuitive. In the recombination tree, this phase transition should translate into a transition, at a certain locus length L_0, from a small number of thick branches ($L < L_0$) to a large number of thin branches ($L > L_0$). We plan to extend this study to a full description of the structure of the recombination tree.

In the next section, we define rigorously the model and recall some chosen results from [2]. Section 3 is concerned with the asymptotic behavior, as $t \to \infty$, of the expected sizes and ages of the most abundant/oldest families. Sections 4 and 5 deal with the joint convergence in distribution of these sizes and ages in the respective cases when clonal families are subcritical or critical. A final appendix is devoted to some technical lemmas for the control of moments of order 2 of largest sizes and ages.

2 Model, preliminary results and statement of the main results

2.1 Model

In this work, we consider genealogical trees satisfying the branching property and called *splitting trees* [6, 7]. Splitting trees are those random trees where individuals' lifetime durations are i.i.d. with an arbitrary distribution, but where birth events occur at Poisson times during each individual's lifetime. We call b this constant birth rate and we denote by V a r.v. distributed as the lifetime duration. Then set $\Lambda(dr) := b\mathbb{P}(V \in dr)$ a finite measure on $(0, \infty]$ with total mass b called the *lifespan measure*. We will always assume that a splitting tree is started with one unique progenitor born at time 0.

The process $(N_t; t \geq 0)$ counting the number of alive individuals at time t is a homogeneous, binary *Crump–Mode–Jagers process*, which is not Markovian unless Λ has an exponential density or is a Dirac mass at $\{+\infty\}$.

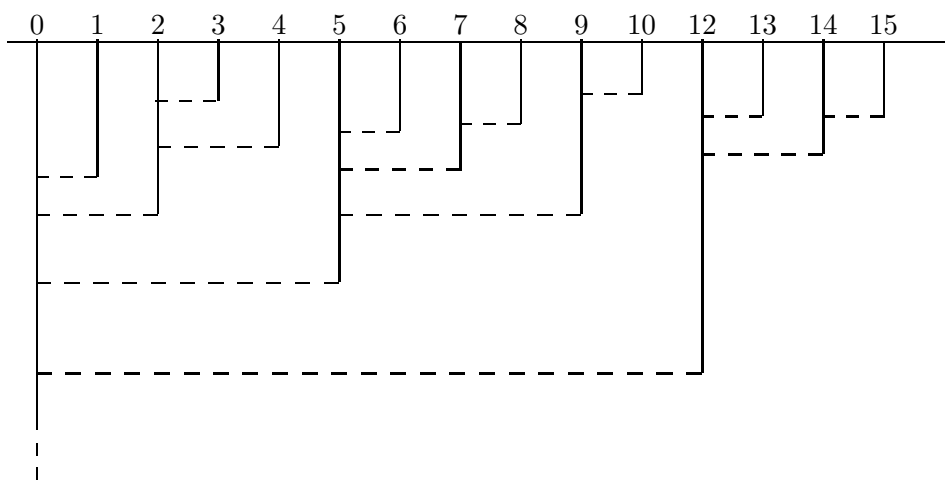

Figure 1: A coalescent point process.

In [13], it is shown that the genealogy of a splitting tree conditioned to be extant at a fixed time t is given by a *coalescent point process*, that is, a sequence of i.i.d. random variables H_i, $i \geq 1$, killed at its first value greater than t. In particular, conditional on $N_t \neq 0$, N_t follows a geometric distribution with parameter $\mathbb{P}(H \geq t)$, where H is distributed as the H_i. More specifically, for any $0 \leq i < j \leq N_t - 1$, the *coalescence time* between the i-th individual alive at time t and the j-th individual alive at time t (i.e., the time elapsed since the common lineage to both individuals split into two distinct lineages) is the maximum of $H_{i+1}, \ldots, H_j$. The graphical representation in Figure 1 is straightforward. The common law of these so-called *branch lengths* is given by

$$\mathbb{P}(H > s) = \frac{1}{W(s)}, \tag{2.1}$$

where the nondecreasing function W is such that $W(0) = 1$ and is characterized by its Laplace transform. More specifically, these branch lengths are the depths of the excursions of the jump

contour process, say $Y^{(t)}$, of the splitting tree truncated below level t. They are i.i.d. because $Y^{(t)}$ is a Markov process. Indeed, it is shown in [13] that $Y^{(t)}$ has the law of a Lévy process, say Y, with no negative jumps, reflected below t and killed upon hitting 0. The function W is called the *scale function* of Y, and is defined from the Laplace exponent ψ of Y:

$$\psi(x) = x - \int_{(0,+\infty]} \left(1 - e^{-rx}\right) \Lambda(dr) \qquad x \in \mathbb{R}_+. \tag{2.2}$$

Let α denote the largest root of ψ. In the supercritical case (i.e. $\int_{(0,\infty]} r\Lambda(dr) > 1$), and in this case only, α is positive and called the *Malthusian parameter*, because the population size grows exponentially at rate α on the survival event. Then the function W is characterized by

$$\int_0^\infty e^{-xr} W(r)\, dr = \frac{1}{\psi(x)} \qquad x > \alpha.$$

Actually, it is possible to show by path decompositions of the process Y that

$$W(x) = \exp\left(b \int_0^x dt\, \mathbb{P}(J > t)\right), \tag{2.3}$$

where J is the maximum of the path of Y killed upon hitting 0 and started from a random initial value, distributed as V. Note that since Y is also the contour process of a splitting tree, J has the law of the extinction time of the CMJ process N started from one individual.

Throughout this work, we further assume that individuals independently experience mutations at Poisson times during their lifetime, that each new mutation event confers a brand new type (called haplotype, or allele) to the individual, and that a newborn holds the same type as her mother at birth time. The mutation rate is denoted by θ. From now on, $\mathbb{P}_t$ (resp. $\mathbb{P}^\star$) will denote the conditional probability on survival up to time t (resp. on the survival event).

2.2 Expected frequency spectrum

Recall from the introduction that $A_\theta(k,t)$ is the number of types carried by k individuals at time t. Also denote by $Z_0(t)$ the number of individuals carrying the ancestral type at time t.

In [2, Cor. 4.3], we obtained the following explicit formulae for the expected frequency spectrum in the population at time t:

$$\mathbb{E}_t A_\theta(k,t) = W(t) \int_0^t e^{-\theta x} \frac{1}{W_\theta(x)^2} \left(1 - \frac{1}{W_\theta(x)}\right)^{k-1} (\theta\, dx + \delta_t(dx)), \tag{2.4}$$

and

$$\mathbb{P}_t\left(Z_0(t) = k\right) = W(t) \frac{e^{-\theta t}}{W_\theta(t)^2} \left(1 - \frac{1}{W_\theta(t)}\right)^{k-1}, \tag{2.5}$$

and

$$W_\theta(x) := e^{-\theta x} W(x) + \theta \int_0^x W(y) e^{-\theta y} dy, \tag{2.6}$$

i.e. $W'_\theta(x) = e^{-\theta x} W'(x) \geq 0$ and $W_\theta(0) = W(0) = 1$. Note that W_θ is the scale function associated to the clonal splitting tree [2, Thm. 3.1], i.e.

$$\int_0^\infty e^{-xr} W_\theta(r) dr = \frac{1}{\psi_\theta(x)}, \qquad x > \alpha - \theta,$$

where

$$\psi_\theta(x) := x - \int_{(0,+\infty]} (1 - e^{-rx}) b\mathbb{P}(V_\theta \in dr) = \frac{x\psi(x+\theta)}{x+\theta}, \tag{2.7}$$

where V_θ denotes the minimum of V and of an independent r.v. with parameter θ.

In addition, we were able to obtain the explicit expected age density of the frequency spectrum [2, Eq. (4.5)]: for all $y < t$, defining $A_\theta(k, t, y; dy)$ as the number of haplotypes of age in the interval $(y, y + dy)$ represented by exactly k alive individuals at time t, we have

$$\mathbb{E}_t A_\theta(k, t, y; dy) = \theta \, dy \, W(t) \frac{e^{-\theta y}}{W_\theta(y)^2} \left(1 - \frac{1}{W_\theta(y)}\right)^{k-1}. \tag{2.8}$$

In [2], denoting by $A_\theta(t) = \sum_{k \geq 1} A_\theta(k, t)$ the total number of haplotypes at time t, we deduced from these expressions the a.s. large time convergence of the fraction $A_\theta(k, t)/A_\theta(t)$. Recall that families with given size $k = 1, 2, 3, ...$ are referred to as *small families*. *Large families* are those who have the largest sizes and *old families* are those whose original mutation is among the oldest. We are interested in the sizes and ages of large and of old families. For example, the size of the largest family is the largest k such that $A_\theta(k, t) \geq 1$.

2.3 Statement of results

Recall that we always assume that the Malthusian parameter α is positive. The asymptotic size of the most frequent and oldest haplotypes strongly depends on the way α and the mutation rate θ compare. Since θ is an additional death rate for clonal families, the case $\alpha > \theta$ corresponds to supercritical clonal families, the case $\theta = \alpha$ corresponds to critical clonal families, and the case $\theta > \alpha$ corresponds to subcritical clonal families.

In the whole paper, we are going to use the following notation: for all $x, s > 0$, define

- $L_t(x)$ the number of haplotypes carried by at least (or exactly) x individuals alive at time t (L for *large*);
- $O_t(s)$ the number of haplotypes with alive representatives at time t older than s, i.e. whose original mutation has age greater than s (O for *old*);
- $M_t(x, s)$ the number of haplotypes carried by at least x individuals alive at time t and whose original mutation has age greater than s. By convention, we set $M_t(x, s) = M_t(x, 0) = L_t(x)$ when $s < 0$. For $0 \leq s_1 \leq s_2$, we also define

  $$M_t(x, s_1, s_2) = M_t(x, s_1) - M_t(x, s_2),$$

  the number of haplotypes carried by at least x individuals alive at time t, whose original mutation has age in $(s_1, s_2]$.

Our convergence results are of two kinds: convergence in expectation of $L_t(x_t)$ and $O_t(s_t)$ for conveniently chosen x_t and s_t, which are directly obtained from (2.4), (2.5) and (2.8) (see Section 3), and convergence in distribution of the point process of (correctly rescaled) largest families or oldest families, which require to combine the previous equations with the theory of coalescent point processes [13, 16] (see Sections 4 and 5). We obtain different results depending on whether the clonal families are supercritical, subcritical or critical.

2.3.1 Supercritical clonal families

Our only result in the supercritical case ($\alpha > \theta$) is the following (Proposition 3.3): for all $0 \leq a_0 < a_1 \leq +\infty$ and $c \geq 0$,

$$\lim_{t\to+\infty} \mathbb{E}_t M_t(ce^{(\alpha-\theta)t}, t-a_1, t-a_0) = \frac{\alpha-\theta}{\alpha} \int_{a_0}^{a_1} \exp\left(\alpha y - c\psi'_\theta(\alpha-\theta)e^{(\alpha-\theta)y}\right) (\theta dy + \delta_0(dy)), \quad (2.9)$$

where δ_0 denotes the Dirac measure at 0. Note that (2.7) yields

$$\psi'_\theta(\alpha-\theta) = \frac{(\alpha-\theta)\psi'(\alpha)}{\alpha}.$$

This result means that the largest families at time t have a size of the order of $e^{(\alpha-\theta)t}$, and that their age is of the order of t minus a constant, i.e. were born in the first moments of the population growth. In particular, the largest and oldest families are the same.

This result can be interpreted as follows: If $N_\theta(t)$ denotes the size of a clonal family started at time 0 from one individual, then conditional on its survival at time t, $N_\theta(t)\, e^{-(\alpha-\theta)t}$ converges in distribution to an exponential r.v. with parameter $\psi'_\theta(\alpha-\theta)$ [13]. If we restrict the limit in the last statement to its Dirac term, we recover the previous convergence stated for the ancestral type

$$\lim_{t\to+\infty} \mathbb{P}_t(Z_0(t) > ce^{(\alpha-\theta)t}) = \frac{\alpha-\theta}{\alpha}\, e^{-c\psi'_\theta(\alpha-\theta)}.$$

The prefactor $\frac{\alpha-\theta}{\alpha}$ is the probability of survival of the ancestral family conditional on the survival of the whole population, which is the ratio of $(\alpha-\theta)/b$ (survival probability of the ancestral family) with α/b (survival probability of the whole population).

In order to recover $\mathbb{E}_t M_t(ce^{(\alpha-\theta)t}, t-a_1, t-a_0)$, we need to integrate the mutation rate per branch $\theta\, dy$ against the expected number of individuals alive at time y *having at least one alive (not necessarily clonal) descendant at time* t, times the probability that the splitting tree spanned by one of these individuals has at least $ce^{(\alpha-\theta)t}$ *clonal* descendants, which is exactly $\mathbb{P}_{t-y}(Z_0(t-y) > ce^{(\alpha-\theta)t})$ since this splitting tree is not extinct after $t-y$ time units, which converges to

$$\frac{\alpha-\theta}{\alpha}\, e^{-c\psi'_\theta(\alpha-\theta)e^{(\alpha-\theta)y}}$$

as $t \to +\infty$. Now, the number of individuals alive at time y having descendants at time t is given by the number of branches deeper than $t-y$ in the coalescent point process, i.e. has a geometric distribution of parameter $\mathbb{P}(H > t \mid H > t-y) = W(t-y)/W(t)$. As will appear in Lemma 3.1

below, this quantity converges to $e^{-\alpha y}$ as $t \to +\infty$, which completes the interpretation of each term of (2.9).

Using the theory of time-dependent random characteristics, Z. Taïb [18] was able to obtain more precise results (but without considering ages of haplotypes) in the case of CMJ processes where mutations occur *at birth*. In this case, using the notation $\alpha-\theta$ for the (positive) Malthusian parameter clonal families (to be consistent with our notation), he obtained in Theorem (4.6) the convergence of the number of haplotypes carried by more than $ce^{(\alpha-\theta)t}$ individuals to a mixed Poisson r.v. with parameter $Cw_\infty/\alpha c^{\alpha/(\alpha-\theta)}$ where the constant C is explicit and w_∞ is the limit of $N_t e^{-\alpha t}$ when $t \to +\infty$, where N_t is the population size at time t. This result is consistent with ours, although the precise value of the constant C is not the same.

Although not stated in [18], one can easily extend this result using [14, Thm. A.1.] to obtain the convergence in distribution on the event of non-extinction of the point measure $\eta_t(\cdot)$, where $\eta_t([a_0, a_1])$ is the number of haplotypes carried at time t by a number of individuals in $[a_0 e^{(\alpha-\theta)t}, a_1 e^{(\alpha-\theta)t}]$, towards a mixed Poisson point process (also known as *Cox process*) on $\mathbb{R}_+$ with intensity measure

$$\mu(dx) := \frac{Cw_\infty}{(\alpha-\theta)x^{1+\alpha/(\alpha-\theta)}}\,dx.$$

This is actually the kind of results that we are able to prove when mutations occur during the life of individuals and when clonal families are subcritical or critical (see below). Unfortunately, the method we develop in Section 4 does not apply to the supercritical case.

2.3.2 Subcritical clonal families

When $\alpha < \theta$, we prove in Proposition 3.4 that for all $a \in \mathbb{R}$

$$\lim_{t\to+\infty} \mathbb{E}_t O_t\left(\frac{\alpha t}{\theta} + a\right) = k\, e^{-\theta a}$$

for an explicit constant k, and that the maximal size of families older than $\alpha t/\theta + a$ is tight when $t \to +\infty$ for all $a \in \mathbb{R}$. We also prove in Proposition 3.5 that for all $c \in \mathbb{R}$,

$$\mathbb{E}_t L_t(x_t(c)) \sim k\,\varphi(\theta)^{c+\{-x_t(c)\}}, \tag{2.10}$$

where $\{x\}$ denotes the fractional part of the real number x, i.e. $\{x\} = x - \lfloor x \rfloor = x + \lceil -x \rceil$, where $\lfloor\cdot\rfloor$ (resp. $\lceil\cdot\rceil$) is the *integer part* (resp. *ceiling*) function, and

$$x_t(c) = k't - k''\log t + c$$

and

$$\varphi(\theta) := 1 - \frac{\psi(\theta)}{\theta}, \tag{2.11}$$

for explicit constants k, k', k''. In addition, we prove that the age of these large families is of the order of $\log t/(\theta-\alpha)$. Hence the largest and oldest families are different in the subcritical case.

Note that, both for large ages and large sizes, random fluctuations are of order 1 (the parameters a and c are not multiplied by a function of t). This explains why the right-hand side of (2.10), while remaining bounded, depends on t: on the one hand, the size of the largest families grows with time as $x_t(0)$ and has fluctuations of order 1; on the other hand the size of the largest families is an integer and hence $L_t(x_t(c))$ only depends on $\lceil x_t(c) \rceil$. Therefore, as a function of c, the right-hand side of (2.10) must only depend on $\lceil x_t(c) \rceil$, which is clear since $c + \{-x_t(c)\} = -x_t(0) + \lceil x_t(c) \rceil$.

This suggests that, given any sequence of times $(t_k)_{k\geq 0}$ such that $t_k \to +\infty$ and $\{x_{t_k}(0)\} =: v$ does not depend on k, the r.v. $L_{t_k}(x_{t_k}(c))$ should converge in distribution, or that the r.v. $X_{t_k}^{(1)} - x_{t_k}(0)$, where $X_t^{(1)}$ is the size of the largest family at time t, should converge in distribution to some r.v. with values in $\mathbb{Z} - v = \{b - v, b \in \mathbb{Z}\}$. This is what we prove in Section 4. For example, we shall state here Corollary 4.4.

Let us denote by $X_t^{(1)} \geq X_t^{(2)} \geq \dots$ the ordered sequence of family sizes in the population at time t. Let also $\mathcal{M}(\mathbb{R})$ be the set of nonnegative σ-finite measures on $\mathbb{R}$, finite on $\mathbb{R}_+$, and let us define the *semi-vague* topology on $\mathcal{M}(\mathbb{R})$ as the one induced by all maps of the form

$$\nu \in \mathcal{M}(\mathbb{R}) \mapsto \int_{\mathbb{R}} u(x)\nu(dx),$$

for all bounded continuous functions u such that, for some $x_0 \in \mathbb{R}$, $u(x) = 0$ for all $x \leq x_0$. Then, Corollary 4.4 states that the sequence of point measures $(Z_k)_{k\geq 0}$ on $\mathbb{Z} - v$ defined by

$$Z_k := \sum_{n\geq 1} \delta_{X_{t_k}^{(n)} - x_{t_k}(0)}$$

converges in $\mathbb{P}^\star$-distribution on $\mathcal{M}(\mathbb{R})$ equipped with the semi-vague topology to a mixed Poisson point measure on $\mathbb{Z} - \{x_{t_0}(0)\}$ with intensity measure

$$\mathcal{E}\, k \sum_{c\in\mathbb{Z}-\{x_{t_0}(0)\}} \varphi(\theta)^c\ \delta_c,$$

where k is an explicit constant and the mixture coefficient $\mathcal{E}$ has exponential distribution with parameter 1.

We obtain similar results for the oldest families (Theorem 4.5): if $A_t^{(1)} \geq A_t^{(2)} \geq \dots$ denotes the ordered sequence of family ages in the population at time t, the family of point measures $(Z_t, t \geq 0)$ on $\mathbb{R}$ defined by

$$Z_t := \sum_{n\geq 1} \delta_{A_t^{(n)} - \alpha t/\theta}$$

converges in $\mathbb{P}^\star$-distribution on $\mathcal{M}(\mathbb{R})$ equipped with the semi-vague topology to a mixed Poisson measure on $\mathbb{R}$ with intensity measure

$$\mathcal{E}\, k\, e^{-\theta a} da, \tag{2.12}$$

where k is an explicit constant.

In the case of CMJ processes with mutations occuring at birth, Jagers and Nerman [10, Applic. C] and Taib [18, Prop. (4.2)] obtained similar results for the extremes of the empirical age distribution.

Proposition (4.2) of [18] states the convergence of the point measure Z_t on the event of non-extinction to a mixed Poisson measure on $\mathbb{R}$ with intensity measure

$$w_\infty \, k' \, e^{-\theta a} da. \tag{2.13}$$

As already noted, in the case of splitting trees, the distribution of w_∞ conditional on survival is exponential [13], so that (2.12) has the same form as (2.13), although the constant k' is different from k.

The technique used in [18] makes use of so-called *individual* time-dependent random characteristics: for any individual i in the population, we define a nonnegative random process, called *random characteristic* $(\chi_{ti}(u), u \in \mathbb{R})$, assigning some score to the individual at age u, such that $\chi_{ti}(u) = 0$ for all $u < 0$. The random characteristic depends on an extra parameter t. We can then define the branching process counted with the time-dependent random characteristic χ_t as

$$Z_t^{\chi_t} := \sum_i \chi_{ti}(t - \sigma_i),$$

where the sum covers the set of all individuals which lived at some time in the population, and σ_i is the birth date of individual i. In such situations, the results of [10] allow one to prove the convergence in distribution of $Z_t^{\chi_t}$ as $t \to +\infty$, under a set of assumptions, among which the more stringent is that the random characteristic is *individual*, i.e. that for all $t \geq 0$, the random processes (ξ_i, χ_{ti}) for i running in the set of all individuals in the population *are i.i.d.*, where $(\xi_i(u), u \geq 0)$ is the process counting the number of children of individual i before age u.

In [18], this method is applied to the population of *haplotypes* (i.e. individuals above have to be understood as haplotypes), defining for any haplotype i (a variant of) the time-dependent random characteristic

$$\chi_{ti}(u) := \mathbb{1}_{\left\{\frac{\alpha t}{\theta} + a \leq u < \lambda_i\right\}},$$

where λ_i is the life length of haplotype i.

This method cannot be used when mutations occur during the life of individuals, since the age of this individual when the mutation occurs influences the distribution of the progeny of the new haplotype (except when lifetime durations are exponential r.v.), which contradicts the assumption that (ξ_i, χ_{ti}) are i.i.d. One may think of defining another random characteristic based on individuals rather than haplotypes, counting the number of mutations experienced by each individual, which occured more than $\alpha t/\theta + a$ time units ago and which has descendants living at present time. With this choice, the random characteristic does not depend on the age of the individual's mother. However, it depends on the whole progeny of the individual, so that (ξ_i, χ_i) and $(\xi_{i'}, \chi_{i'})$ are independent only if the individuals i does not descend from i' and conversely. Therefore, the method of [18] cannot be applied to our case.

Proposition (4.2) of [18] makes use of precise estimates on the tail distribution of the extinction time of a clonal family, which are well-known in this context. No results are given in [18] on the size of large families in the subcritical case, presumably because their method would require precise estimates on the tail distribution of the size of a clonal population at any time, which are to our knowledge not known in general for CMJ processes. In our case of splitting trees with mutations

occurring during the life of individuals, our formulae (2.4) and (2.5) for the expected frequency spectrum are *exact*. This allows us to obtain precise estimates for the tail distribution of the size of a clonal population at some time, *conditionally on the survival of the (clonal or not clonal) progeny of this population at time t*. We are then able to deduce exact asymptotics for the tail distribution of the size of the largest family using a different method than in [18] (see the proof of Theorem 4.2).

We chose here to present results on largest and oldest haplotypes, but our method easily applies to intermediate regimes. For example, one can easily adapt our calculations to prove that, for all $\gamma \in [0,1]$ and $c \in \mathbb{R}$,

$$\mathbb{E}_t M_t \left(x_t(c), \frac{\alpha\gamma}{\theta} t \right) \sim k\, \varphi(\theta)^{c+\{-x_t(c)\}},$$

where

$$x_t(c) = \frac{\alpha(1-\gamma)t}{-\log \varphi(\theta)} + c.$$

Similarly, one can prove the convergence of the point measure of sizes of families older than $(\alpha\gamma/\theta)t$ as $t \to +\infty$ and compute the limit as a mixed Poisson point measure.

2.3.3 Critical clonal families

When $\alpha = \theta$, we prove in Proposition 3.6 that, for all $a \in \mathbb{R}$,

$$\lim_{t\to+\infty} \mathbb{E}_t O_t \left(t - \frac{\log t}{\alpha} + a \right) = k e^{-\alpha a}$$

for an explicit constant k, and that the maximal size of families older than $t - \log t/\alpha + a$ is tight as $t \to +\infty$ for all $a \in \mathbb{R}$. As in the subcritical case, fluctuations are of order 1 here. We also prove in Proposition 3.7 that, for all $c \in \mathbb{R}$,

$$\lim_{t\to+\infty} \mathbb{E}_t L_t(x_t(c)) = k\, \exp\left(-\frac{2\psi'(\alpha)}{\alpha}\, c \right),$$

where

$$x_t(c) = k' t^2 + k'' t \log t + ct \tag{2.14}$$

and the constants k, k', k'' are explicit. In addition, we prove that the age of these large families is of the order of $t/2$. Here, since the fluctuations are of the order of t, the limit does not involve $\{-x_t(c)\}$ as in (2.10)

We are then able to deduce from these estimates the convergence of correctly rescaled point measures of the size of the largest and the age of the oldest families (Theorems 5.1 and 5.3): using the same notation as in Section 2.3.2, the family of point measures $(Z_t, t \geq 0)$ defined by

$$Z_t := \sum_{n\geq 1} \delta_{\frac{X_t^{(n)} - x_t(0)}{t}}$$

converges in $\mathbb{P}^\star$-distribution on $\mathcal{M}(\mathbb{R})$ equipped with the semi-vague topology to a mixed Poisson measure on $\mathbb{R}$ with intensity measure

$$\mathcal{E}\, k \, \exp\left(-\frac{2\psi'(\alpha)}{\alpha}c\right) dc,$$

where the constant k is explicit.

Note that this statement and the definition (2.14) are actually a little bit different than those of Sections 3.4 and 5.1. However, the results stated here can be proved by slightly modifying the proofs of Proposition 3.7 and Theorem 5.1. We have chosen to leave this to the interested reader.

Similarly, we obtain the convergence of the family of point measures $(Z_t, t \geq 0)$ defined by

$$Z_t := \sum_{n\geq 1} \delta_{A_t^{(n)} - t + \frac{\log t}{\alpha}}$$

to a mixed Poisson point measure on $\mathbb{R}$ with intensity measure

$$\mathcal{E}\, e^{-\alpha a}. \tag{2.15}$$

Again, in [10] and [18], a similar result is stated only for extreme ages. It takes the same form as our result with $\mathcal{E}$ replaced by w_∞ and with a different multiplicative constant in the intensity measure (2.15).

3 Large time asymptotics for the expected number of frequent or old haplotypes

Our goal here is to prove the convergence results on the expectation of $L_t(x_t)$ and $O_t(s_t)$ stated above, when clonal families are supercritical, subcritical or critical. We start with preliminary estimates on the scale functions W and W_θ.

3.1 Preliminary results on scale functions

Lemma 3.1 *The survival probability of the splitting tree is α/b, and the scale function W has the following asymptotic behavior*

$$W(t)e^{-\alpha t} = \frac{1 + O(e^{-\gamma t})}{\psi'(\alpha)}$$

as $t \to +\infty$, for some constant $\gamma > 0$.

Proof. The expression of the survival probability and the fact that $W(t) \sim e^{\alpha t}/\psi'(\alpha)$ were already proved in [13]. In order to get the higher-order term, we use the fact that

$$\mathbb{P}(J > t) = \frac{\alpha}{b} + O(e^{-\gamma t}) \tag{3.1}$$

as $t \to +\infty$, where J is the extinction time of the splitting tree started from one individual with random lifespan, distributed as V. Indeed, we know from [13] that the law $\mathbb{P}^\natural := \mathbb{P}(\cdot \mid J < \infty)$

is that of a subcritical splitting tree with lifespan measure $e^{-\alpha r}\Lambda(dr)$. In particular, under $\mathbb{P}^\natural$, the lifetime V of a single individual has exponential moments, and the first hitting time τ_0 of 0 by the contour process of the spliting tree also has exponential moments (because its Laplace exponent is the inverse of $\psi^\natural(\cdot) := \psi(\cdot + \alpha)$). Since $J \leq \tau_0$ a.s., J has exponential moments, that is, there is some $\gamma > 0$ such that $\mathbb{E}^\natural(e^{2\gamma J}) < \infty$. As a consequence, also since $\alpha/b = \mathbb{P}(J = \infty)$,

$$\mathbb{P}(J > t) - \frac{\alpha}{b} = \left(1 - \frac{\alpha}{b}\right) \mathbb{P}^\natural(J > t) = O(e^{-\gamma t}).$$

Therefore, it follows from (2.3) that

$$W(t)e^{-\alpha t} = \frac{\exp\left(-\int_t^\infty (b\mathbb{P}(J > x) - \alpha)dx\right)}{\psi'(\alpha)},$$

and the result then follows from (3.1). □

From this result and the definition (2.6) of W_θ, we can deduce the following lemma. We recall that

$$\psi_\theta(x) = \frac{x\psi(x+\theta)}{x+\theta} \qquad \text{and} \qquad \psi'_\theta(\alpha - \theta) = \frac{(\alpha-\theta)\psi'(\alpha)}{\alpha}.$$

Lemma 3.2 (i) *Assume $\alpha > \theta \geq 0$. Then*

$$W_\theta(t) \sim \frac{e^{(\alpha-\theta)t}}{\psi'_\theta(\alpha-\theta)}$$

as $t \to +\infty$.

(ii) *Assume $0 < \alpha < \theta$. Then*

$$\frac{\theta}{\psi(\theta)} - W_\theta(t) \sim \frac{e^{-(\theta-\alpha)t}}{|\psi'_\theta(\alpha-\theta)|} \tag{3.2}$$

as $t \to \infty$, and

$$1 - \frac{1}{W_\theta(t)} = \varphi(\theta)(1 - \rho(t)),$$

where $\varphi(\theta)$ is defined in (2.11) and $\rho(\cdot)$ is a non-increasing function such that

$$\rho(t) \sim \frac{\psi^2(\theta)}{\theta^2\varphi(\theta)|\psi'_\theta(\alpha-\theta)|}\, e^{-(\theta-\alpha)t} \tag{3.3}$$

as $t \to \infty$.

(iii) *Assume $\alpha = \theta > 0$. Then*

$$W_\alpha(t) = \frac{\alpha t}{\psi'(\alpha)} + \frac{1}{\psi'(\alpha)} + \alpha \int_0^{+\infty} \left(W(y)e^{-\alpha y} - \frac{1}{\psi'(\alpha)}\right) dy + o(1)$$

as $t \to +\infty$.

Proof. Points (i) and (iii) are easy consequences of Lemma 3.1 and the definition (2.6) of W_θ. Point (i) can also be seen as a trivial corollary of Lemma 3.1 since when $\alpha > \theta$, W_θ is the scale function of a supercritical splitting tree.

For Point (ii), by Tauberian theorems (see [13]), we have $W_\theta(t) \to 1/\psi'_\theta(0) = \theta/\psi(\theta)$ as $t \to \infty$. Since

$$\frac{1}{\psi(\theta)} = \int_0^\infty W(y)e^{-\theta y}dy,$$

(2.6) yields

$$W_\theta(t) - \frac{\theta}{\psi(\theta)} = e^{-\theta t}W(t) - \theta \int_t^\infty W(y)e^{-\theta y}dy.$$

Since $W(t) \sim e^{\alpha t}/\psi'(\alpha)$, we have

$$\int_t^\infty W(y)e^{-\theta y}dy \sim \frac{1}{\psi'(\alpha)} \int_t^\infty e^{-(\theta-\alpha)y}dy$$

as $t \to +\infty$. This entails (3.2). Since one has

$$\rho(t) = \frac{1}{\varphi(\theta)} \left(\frac{1}{W_\theta(t)} - \frac{\psi(\theta)}{\theta} \right)$$

and

$$\frac{1}{W_\theta(t)} - \frac{\psi(\theta)}{\theta} = \frac{\frac{\theta}{\psi(\theta)} - W_\theta(t)}{W_\theta(t)\,\frac{\theta}{\psi(\theta)}},$$

the proof of (ii) is easily completed. □

3.2 The case of supercritical families ($\alpha > \theta$)

In the case of supercritical clonal families, the asymptotic expected number of frequent haplotypes can be explicitly computed. Note that in the next statement and elsewhere in the paper, the Dirac measure at time 0 corresponds to the contribution of the family carrying the ancestral type.

Proposition 3.3 *Assume $\alpha > \theta \geq 0$ and let $0 \leq a_0 < a_1 \leq +\infty$ and $c \geq 0$. For all $t \geq 0$, let*

$$x_t(c) = c\exp((\alpha - \theta)t).$$

Then

$$\lim_{t\to+\infty} \mathbb{E}_t[M_t(x_t(c), t - a_1, t - a_0)] = \frac{\alpha - \theta}{\alpha} \int_{a_0}^{a_1} \exp\left(\alpha y - c\,\psi'_\theta(\alpha - \theta)\,e^{(\alpha-\theta)y}\right)(\theta\,dy + \delta_0(dy)).$$

Proof. Using (2.5), (2.8) and Lemma 3.1, for all $t \geq a$, we have

$$\begin{aligned}
\mathbb{E}_t[M_t(x_t(c), t-a_1, t-a_0)] &= \sum_{k=\lceil ce^{(\alpha-\theta)t}\rceil}^{+\infty} \left[W(t)\frac{e^{-\theta t}}{W_\theta(t)^2}\left(1-\frac{1}{W_\theta(t)}\right)^{k-1} \mathbb{1}_{\{a=0\}} \right. \\
&\qquad \left. + \int_{(t-a_1)\vee 0}^{t-a_0} \theta W(t)\frac{e^{-\theta x}}{W_\theta(x)^2}\left(1-\frac{1}{W_\theta(x)}\right)^{k-1} dx \right] \\
&= \mathbb{1}_{\{a_0=0\}}\frac{W(t)}{W_\theta(t)}e^{-\theta t}\left(1-\frac{1}{W_\theta(t)}\right)^{\lceil ce^{(\alpha-\theta)t}\rceil-1} + W(t)\int_{(t-a_1)\vee 0}^{t-a_0}\theta\frac{e^{-\theta x}}{W_\theta(x)}\left(1-\frac{1}{W_\theta(x)}\right)^{\lceil ce^{(\alpha-\theta)t}\rceil-1} dx \\
&\sim \frac{1}{\psi'(\alpha)}\int_{a_0}^{a_1\wedge t}\frac{e^{\theta y}}{W_\theta(t-y)e^{-(\alpha-\theta)t}}\exp\Big((\lceil ce^{(\alpha-\theta)t}\rceil-1)\log(1-1/W_\theta(t-y))\Big)(\theta dy+\delta_0(dy)),
\end{aligned} \tag{3.4}$$

where we recall that $\lceil\cdot\rceil$ is the ceiling function. Since W_θ is nondecreasing and $W_\theta(0)=1$, it follows from Lemma 3.2 (i) that there exists a constant $C>0$ such that

$$\frac{1}{C}e^{(\alpha-\theta)t} \leq W_\theta(t) \leq Ce^{(\alpha-\theta)t}, \qquad \forall t \geq 0.$$

Therefore, for all $y \geq 0$, the quantity inside the integral in the r.h.s. of (3.4) is smaller than

$$C'e^{\alpha y}\exp\left(-C''e^{(\alpha-\theta)y}\right)$$

for some constants $C', C'' > 0$. Now, using Lemma 3.2 (i) again, for all $y \geq 0$, the quantity inside the integral in the r.h.s. of (3.4) converges to

$$\frac{\psi'(\alpha)(\alpha-\theta)}{\alpha}e^{\alpha y}\exp\left(-c\psi'_\theta(\alpha-\theta)e^{(\alpha-\theta)y}\right)$$

when $t \to +\infty$. Proposition 3.3 then follows from the dominated convergence theorem. □

3.3 The case of subcritical families ($\alpha < \theta$)

Our first result deals with ages and sizes of the oldest clonal families. Note that the scaling constant a varies in $\mathbb{R}$.

Proposition 3.4 *When $\alpha < \theta$, for any $a \in \mathbb{R}$,*

$$\lim_{t\to+\infty}\mathbb{E}_t\left[O_t\left(\frac{\alpha\, t}{\theta}+a\right)\right] = \frac{\psi(\theta)}{\theta\psi'(\alpha)}e^{-\theta a}. \tag{3.5}$$

In addition, for any $x_t \to +\infty$ as $t \to +\infty$,

$$\lim_{t\to+\infty}\mathbb{E}_t\left[M_t\left(x_t, \frac{\alpha\, t}{\theta}+a\right)\right] = 0. \tag{3.6}$$

Proof. Using (2.5) and (2.8) as in the proof of Proposition 3.3, we have

$$\begin{aligned}\mathbb{E}_t[O_t(\alpha t/\theta + a)] &= W(t)\int_{\frac{\alpha t}{\theta}+a}^{t} \frac{e^{-\theta x}}{W_\theta(x)}(\theta dx + \delta_t(dx)) \\ &\sim \frac{\psi(\theta)e^{\alpha t}}{\psi'(\alpha)}\int_{\frac{\alpha t}{\theta}+a}^{t} e^{-\theta x}\Big(dx + \frac{1}{\theta}\delta_t(dx)\Big),\end{aligned}$$

where we used that $W_\theta(x) \to \theta/\psi(\theta)$ as $x \to +\infty$ (Lemma 3.2). Eq. (3.5) then easily follows.

Similarly,

$$\begin{aligned}\mathbb{E}_t[M_t(x_t,\, \alpha t/\theta + a)] &= W(t)\int_{\frac{\alpha t}{\theta}+a}^{t} \frac{e^{-\theta x}}{W_\theta(x)}\left(1 - \frac{1}{W_\theta(x)}\right)^{\lceil x_t\rceil - 1}(\theta dx + \delta_t(dx)) \\ &\le \left(1 - \frac{1}{W_\theta(t)}\right)^{\lceil x_t\rceil - 1} W(t)\int_{\frac{\alpha t}{\theta}+a}^{t} \frac{e^{-\theta x}}{W_\theta(x)}(\theta dx + \delta_t(dx)),\end{aligned}$$

since W_θ is nondecreasing. Eq. (3.6) then follows from (3.5) and the fact that $(1 - 1/W_\theta(t))^{\lceil x_t\rceil} \to 0$ as $t \to +\infty$, since $W_\theta(t) \to \theta/\psi(\theta) > 1$. □

Our next result gives the asymptotics of the expected number of large families (notice again that the scaling constant c varies in $\mathbb{R}$) and states that their ages all are asymptotically equivalent to $(\theta - \alpha)^{-1}\log(t)$. We recall that $\varphi(\theta) < 1$, so that $|\log\varphi(\theta)| = -\log\varphi(\theta)$.

Proposition 3.5 *Assume $\alpha < \theta$. For all $c \in \mathbb{R}$, let*

$$x_t(c) := \frac{\alpha t - \frac{\theta}{\theta-\alpha}\log t}{|\log\varphi(\theta)|} + c.$$

Then, for all $\varepsilon > 0$,

$$\mathbb{E}_t[L_t(x_t(c))] \sim \mathbb{E}_t\left[M_t\left(x_t(c),\, \frac{1-\varepsilon}{\theta-\alpha}\log t,\, \frac{1+\varepsilon}{\theta-\alpha}\log t\right)\right] \sim A(\theta)\,\varphi(\theta)^{c-1+\{-x_t(c)\}} \tag{3.7}$$

as $t \to +\infty$, where we recall that $\{\cdot\}$ is the fractional part function and

$$A(\theta) := \frac{\Gamma\left(\frac{\theta}{\theta-\alpha}\right)\psi(\theta)}{\alpha}\,|\psi'_\theta(\alpha-\theta)|^{\frac{\alpha}{\theta-\alpha}}\left(\frac{\theta^2}{\alpha\psi(\theta)^2}\,\varphi(\theta)\;|\log\varphi(\theta)|\right)^{\frac{\theta}{\theta-\alpha}}, \tag{3.8}$$

where $\Gamma(s) = \int_0^\infty y^{s-1}e^{-y}dy$ is the Gamma function.

Proof. Proceeding similarly as in the proof of Proposition 3.3, we deduce from Lemma 3.2 (ii) that

$$\begin{aligned}\mathbb{E}_t[L_t(x_t(c))] &= W(t)\,\varphi(\theta)^{\lceil x_t(c)\rceil - 1}\int_0^t \frac{e^{-\theta x}}{W_\theta(x)}\,(1-\rho(x))^{\lceil x_t(c)\rceil - 1}\,(\theta dx + \delta_t(dx)) \\ &\sim \frac{t^{\theta/(\theta-\alpha)}\,\varphi(\theta)^{c-1+\{-x_t(c)\}}}{\psi'(\alpha)}\int_0^t \frac{e^{-\theta x + (\lceil x_t(c)\rceil - 1)\log(1-\rho(x))}}{W_\theta(x)}(\theta dx + \delta_t(dx))\end{aligned} \tag{3.9}$$

as $t \to +\infty$, where we used the relation $\lceil x_t(c) \rceil - 1 = x_t(0) + c - 1 + \{-x_t(c)\}$. We will now abridge $x_t(c)$ into x_t.

Let $\varepsilon > 0$. Let us first bound from above the previous integral restricted to the complement of $\left[\frac{1-\varepsilon}{\theta-\alpha}\log t, \frac{1+\varepsilon}{\theta-\alpha}\log t\right]$. On the one hand, using the inequality $\log(1-\rho(x)) \le 0$ and the fact that $W_\theta(x)$ converges to a positive constant when $x \to +\infty$, we have for all $\varepsilon > 0$

$$\int_{\frac{1+\varepsilon}{\theta-\alpha}\log t}^{+\infty} \frac{e^{-\theta x + (\lceil x_t \rceil - 1)\log(1-\rho(x))}}{W_\theta(x)} (\theta dx + \delta_t(dx)) \le C\left(e^{-\theta t} + e^{-\theta \frac{1+\varepsilon}{\theta-\alpha}\log t}\right) = o(t^{-\frac{\theta}{\theta-\alpha}})$$

as $t \to +\infty$.

On the other hand, since $\rho(x)$ is non-increasing and using (3.3), $\rho(x) \ge \rho(\frac{1-\varepsilon}{\theta-\alpha}\log t) \ge Ct^{-1+\varepsilon}$ for all $0 \le x \le (1-\varepsilon)/(\theta-\alpha)\ \log t$. Therefore,

$$\int_0^{\frac{1-\varepsilon}{\theta-\alpha}\log t} \frac{e^{-\theta x + (\lceil x_t \rceil - 1)\log(1-\rho(x))}}{W_\theta(x)} (\theta dx + \delta_t(dx)) \le C' \int_0^{\frac{1-\varepsilon}{\theta-\alpha}\log t} e^{-\theta x}\, e^{-C(\lceil x_t \rceil - 1)t^{-1+\varepsilon}} dx.$$

Since $(\lceil x_t \rceil - 1)t^{-1+\varepsilon} \ge C'' t^{\varepsilon}$ for t large enough, we deduce that the previous integral is also $o(t^{-\frac{\theta}{\theta-\alpha}})$. In conclusion, $\mathbb{E}_t[L_t(x_t(a))]$ and $\mathbb{E}_t\left[M_t\left(x_t(a), \frac{1-\varepsilon}{\theta-\alpha}\log t, \frac{1+\varepsilon}{\theta-\alpha}\log t\right)\right]$ are both asymptotically equivalent to $E_t(\varepsilon)$, provided that $E_t(\varepsilon)$ is uniformly bounded from below, where

$$E_t(\varepsilon) := \frac{\psi(\theta)\, t^{\theta/(\theta-\alpha)}\, \varphi(\theta)^{a-1+\{-x_t(a)\}}}{\psi'(\alpha)} \int_{\frac{1-\varepsilon}{\theta-\alpha}\log t}^{\frac{1+\varepsilon}{\theta-\alpha}\log t} e^{-\theta x + (\lceil x_t \rceil - 1)\log(1-\rho(x))} dx. \tag{3.10}$$

Note that, since $W_\theta(x) \to \psi(\theta)/\theta$ when $x \to +\infty$, the replacement of $W_\theta(x)$ by its limit in the r.h.s. of (3.9) is justified. The proof of Proposition 3.5 will hence be completed if we can prove that

$$E_t(\varepsilon) \sim A(\theta)\, \varphi(\theta)^{c-1+\{-x_t(c)\}}.$$

This is the aim of the rest of the proof. Set

$$B := \frac{\psi^2(\theta)}{\theta^2 \varphi(\theta) |\psi'_\theta(\alpha-\theta)|}, \tag{3.11}$$

and we make the change of variable $y := B(\lceil x_t \rceil - 1)e^{-(\theta-\alpha)x}$ in (3.10):

$$E_t(\varepsilon) \sim \frac{\psi(\theta)\, t^{\theta/(\theta-\alpha)}\, \varphi(\theta)^{c-1+\{-x_t\}}}{\psi'(\alpha)(\theta-\alpha)(B(\lceil x_t \rceil - 1))^{\theta/(\theta-\alpha)}} \int_{B(\lceil x_t \rceil - 1)t^{-1-\varepsilon}}^{B(\lceil x_t \rceil - 1)t^{-1+\varepsilon}} y^{\theta/(\theta-\alpha)-1}$$
$$\exp\left((\lceil x_t \rceil - 1)\log\left(1 - \rho\left(\frac{\log \frac{B(\lceil x_t \rceil - 1)}{y}}{\theta-\alpha}\right)\right)\right) dy. \tag{3.12}$$

Now, using Lemma 3.2 (ii) again, we have for all $y > 0$

$$\rho\left(\frac{\log \frac{B(\lceil x_t \rceil - 1)}{y}}{\theta-\alpha}\right) \sim B e^{\log \frac{y}{B(\lceil x_t \rceil - 1)}} = \frac{y}{\lceil x_t \rceil - 1}$$

as $t \to +\infty$. Therefore, the exponential in the r.h.s. of (3.12) converges for all $y > 0$ to e^{-y} when $t \to +\infty$. In addition, using the inequalities $\log(1-\rho(x)) \le -\rho(x)$ and $\rho(x) \ge Ce^{-(\theta-\alpha)x}$ for all x large enough, we have

$$y^{\frac{\theta}{\theta-\alpha}-1} \exp\left((\lceil x_t \rceil - 1) \log\left(1 - \rho\left(\frac{\log \frac{B(\lceil x_t \rceil - 1)}{y}}{\theta - \alpha} \right) \right) \right) \le y^{\frac{\alpha}{\theta-\alpha}} e^{-Cy}.$$

Since $x_t t^{-1-\varepsilon} \to 0$ and $x_t t^{-1+\varepsilon} \to +\infty$ when $t \to +\infty$, Lebesgue's theorem finally yields

$$E_t(\varepsilon) \sim \frac{\psi(\theta)\varphi(\theta)^{c-1+\{-x_t\}}\Gamma(\frac{\theta}{\theta-\alpha})}{\psi'(\alpha)(\theta-\alpha)B^{\theta/(\theta-\alpha)}} \left(\frac{t}{\lceil x_t \rceil - 1} \right)^{\theta/(\theta-\alpha)},$$

Remembering that $x_t \sim \alpha t / |\log \varphi(\theta)|$ as $t \to +\infty$ concludes the proof of Proposition 3.5. □

3.4 The case of critical families ($\alpha = \theta$)

The following result gives the asymptotic expected number of old families and states that their sizes are tight.

Proposition 3.6 *Assume $\alpha = \theta > 0$. For all $a \in \mathbb{R}$, we have*

$$\lim_{t \to +\infty} \mathbb{E}_t \left[O_t \left(t - \frac{\log t}{\alpha} + a \right) \right] = \frac{e^{-\alpha a}}{\alpha}.$$

In addition, for all $x_t \to +\infty$,

$$\lim_{t \to +\infty} \mathbb{E}_t \left[M_t \left(x_t,\, t - \frac{\log t}{\alpha} + a \right) \right] = 0.$$

Proof. Using Lemma 3.2 (iii), a similar computation as for Proposition 3.4 yields

$$\mathbb{E}_t[O_t(t - \log t/\alpha + a)] \sim e^{\alpha t} \int_{t - \frac{\log t}{\alpha} + a}^{t} \frac{e^{-\alpha x}}{x} (dx + \frac{1}{\alpha}\delta_t(dx)) \sim \frac{e^{\alpha t}}{t} \int_{t - \frac{\log t}{\alpha} + a}^{t} e^{-\alpha x} dx + \frac{1}{\alpha t}.$$

The first limit easily follows. The second limit is obtained exactly as in the proof of Proposition 3.4. □

The computations for the most frequent haplotypes are more involved. The following result gives the asymptotics of the expected number of large families and states that their ages are all asymptotically equivalent to $t/2$.

Proposition 3.7 *Assume $\alpha = \theta > 0$. For all $c \in \mathbb{R}$, we define*

$$x_t(c) := \frac{\alpha^2}{4\psi'(\alpha)} \left(t - \frac{\log t}{2\alpha} + c \right)^2.$$

Then, for all $\varepsilon > 0$,

$$\lim_{t \to +\infty} \mathbb{E}_t[L_t(x_t(c))] = \lim_{t \to +\infty} \mathbb{E}_t \left[M_t \left(x_t(c),\, \frac{1-\varepsilon}{2} t,\, \frac{1+\varepsilon}{2} t \right) \right] = \sqrt{\frac{2\pi}{\alpha}}\, e^{B - \frac{\psi'(\alpha)}{2}}\, e^{-\alpha c},$$

where

$$B = 1 + \alpha \int_0^{+\infty} \left(\psi'(\alpha) W(y) e^{-\alpha y} - 1 \right) dy. \tag{3.13}$$

Proof. Similarly as in the proof of Proposition 3.5, we have

$$\mathbb{E}_t[L_t(x_t(c))] \sim \frac{e^{\alpha t}}{\psi'(\alpha)} \int_0^t \frac{e^{-\alpha x}}{W_\alpha(x)} e^{(\lceil x_t(c)\rceil - 1)\log\left(1-\frac{1}{W_\alpha(x)}\right)} (\alpha dx + \delta_t(dx)). \tag{3.14}$$

Let $\varepsilon > 0$. Let us first bound from above the previous integral restricted to the complement of $\left[\frac{1-\varepsilon}{2}t, \frac{1+\varepsilon}{2}t\right]$. Fix $\eta \in (0,1)$. By Lemma 3.2 (iii), for all x large enough, $1 \le W_\alpha(x) \le \frac{\alpha t}{\psi'(\alpha)(1-\eta)}$. Hence, using the fact that $x_t(c) \sim \alpha^2 t^2/4\psi'(\alpha)$, for t sufficiently large,

$$\begin{aligned}\int_{\frac{1+\varepsilon}{2}t}^t \frac{e^{-\alpha x}}{W_\alpha(x)} e^{(\lceil x_t(c)\rceil - 1)\log\left(1-\frac{1}{W_\alpha(x)}\right)} (\alpha dx + \delta_t(dx)) &\le \int_{\frac{1+\varepsilon}{2}t}^t e^{-\alpha x - (\lceil x_t(c)\rceil - 1)\frac{(1-\eta)\psi'(\alpha)}{\alpha x}} (\alpha dx + \delta_t(dx)) \\ &\le \alpha \int_{\frac{1+\varepsilon}{2}t}^t e^{-\alpha\left(x + \frac{(1-2\eta)t^2}{4x}\right)} dx + e^{-\alpha t(1+(1-2\eta)/4)}.\end{aligned}$$

The quantity inside the integral in the integral of the r.h.s. is maximal for $x = (1-2\eta)^{1/2}t/2$, which is outside the integration domain, so that

$$\int_{\frac{1+\varepsilon}{2}t}^t e^{-\alpha\left(x + \frac{(1-2\eta)t^2}{4x}\right)} dx \le t \exp\left(-\alpha\left(\frac{t(1+\varepsilon)}{2} + \frac{t(1-2\eta)}{2(1+\varepsilon)}\right)\right).$$

This last quantity is $o(e^{-\alpha t})$ if one chooses $\eta < \varepsilon^2/2$.

Using the fact that $W_\alpha(x)$ is non-decreasing, larger than (or equal to) 1 and that $e^{-\alpha x} \le 1$, we have

$$\int_0^{\log t} \frac{e^{-\alpha x}}{W_\alpha(x)} e^{(\lceil x_t(c)\rceil - 1)\log\left(1-\frac{1}{W_\alpha(x)}\right)} \alpha dx \le \alpha e^{-\frac{\lceil x_t(c)\rceil - 1}{W_\alpha(\log t)}} \log t.$$

Since $x_t(c) \sim \alpha^2 t^2/4\psi'(\alpha)$, using Lemma 3.2 (iii) again, for t large enough, the previous integral is smaller than

$$\alpha \log(t)\, e^{-Ct^2/\log t} = o(e^{-\alpha t}),$$

for a constant $C > 0$ independent of t.

Finally, using Lemma 3.2 (iii) similarly as above, for all $\eta \in (0,1)$, for t large enough,

$$\int_{\log t}^{\frac{1-\varepsilon}{2}t} \frac{e^{-\alpha x}}{W_\alpha(x)} e^{(\lceil x_t(c)\rceil - 1)\log\left(1-\frac{1}{W_\alpha(x)}\right)} \alpha\, dx \le \alpha \int_{\log t}^{\frac{1-\varepsilon}{2}t} e^{-\alpha\left(x + \frac{(1-2\eta)t^2}{4x}\right)} dx.$$

Taking η small enough so that $(1-\eta)^{1/2}t/2 > t(1-\varepsilon)/2$, the previous quantity can be bounded from above by

$$\alpha t e^{-\frac{\alpha t}{2}\left(1-\varepsilon+\frac{1-2\eta}{1-\varepsilon}\right)} = o(e^{-\alpha t})$$

if one takes again $\eta < \varepsilon^2/2$. In conclusion, $\mathbb{E}_t[L_t(x_t(c))]$ and $\mathbb{E}_t\left[M_t\left(x_t(c), \frac{1-\varepsilon}{2}t, \frac{1+\varepsilon}{2}t\right)\right]$ are both asymptotically equivalent to $E_t(\varepsilon)$, provided that $E_t(\varepsilon)$ is uniformly bounded from below, where

$$E_t(\varepsilon) := \frac{\alpha e^{\alpha t}}{\psi'(\alpha)} \int_{\frac{1-\varepsilon}{2}t}^{\frac{1+\varepsilon}{2}t} \frac{e^{-\alpha x}}{W_\alpha(x)} e^{(\lceil x_t(c)\rceil - 1)\log\left(1-\frac{1}{W_\alpha(x)}\right)} dx.$$

Therefore, it only remains to prove that

$$\lim_{t\to+\infty} E_t(\varepsilon) = \sqrt{\frac{2\pi}{\alpha}}\, e^{B-\frac{\psi'(\alpha)}{2}}\, e^{-\alpha\, c}.$$

Using the facts that $W_\alpha(x) \sim \alpha x/\psi'(\alpha)$ and that $|\log(1-1/W_\alpha(x))| \le C/t$ for all x large enough, we have

$$E_t(\varepsilon) \sim e^{\alpha t} \int_{\frac{1-\varepsilon}{2}t}^{\frac{1+\varepsilon}{2}t} e^{-\alpha x}\, e^{x_t(c)\log\left(1-\frac{1}{W_\alpha(x)}\right)}\, \frac{dx}{x},$$

when $t \to +\infty$.

It follows from Lemma 3.2 (iii) that

$$\log\left(1-\frac{1}{W_\alpha(x)}\right) = -\frac{\psi'(\alpha)}{\alpha x} + \frac{\psi'(\alpha)(B-\psi'(\alpha)/2)}{\alpha^2 x^2} + o\left(\frac{1}{x^2}\right) \tag{3.15}$$

as $x \to +\infty$. Therefore,

$$F(t) := x_t(c) \sup_{\frac{1-\varepsilon}{2}t \le x \le \frac{1+\varepsilon}{2}t} \left[\log\left(1-\frac{1}{W_\alpha(x)}\right) + \frac{\psi'(\alpha)}{\alpha x} - \frac{\psi'(\alpha)(B-\psi'(\alpha)/2)}{\alpha^2 x^2}\right] \to 0$$

as $t \to +\infty$. Hence, using the facts that $x_t(c) \sim \frac{\alpha^2 t^2}{4\psi'(\alpha)}$ and that $x \in [\frac{1-\varepsilon}{2}t, \frac{1+\varepsilon}{2}t]$, for t large enough,

$$\begin{aligned}\frac{2e^{-C\varepsilon}}{(1+\varepsilon)t}\, e^{B-\frac{\psi'(\alpha)}{2}}\, e^{\alpha t} \int_{\frac{1-\varepsilon}{2}t}^{\frac{1+\varepsilon}{2}t} e^{-\alpha x - \frac{x_t(c)\psi'(\alpha)}{\alpha x}} dx &\le E_t(\varepsilon) \\ &\le \frac{2e^{C\varepsilon}}{(1-\varepsilon)t}\, e^{B-\frac{\psi'(\alpha)}{2}}\, e^{\alpha t} \int_{\frac{1-\varepsilon}{2}t}^{\frac{1+\varepsilon}{2}t} e^{-\alpha x - \frac{x_t(c)\psi'(\alpha)}{\alpha x}} dx, \end{aligned} \tag{3.16}$$

for a constant C independent of ε and t.

Now, let us compute the asymptotic behavior of the integral involved in these inequalities: first, the change of variable $x = \beta_t y$ with $\beta_t = \sqrt{x_t(c)\psi'(\alpha)}/\alpha$ yields

$$\int_{\frac{1-\varepsilon}{2}t}^{\frac{1+\varepsilon}{2}t} e^{-\alpha x - \frac{x_t(c)\psi'(\alpha)}{\alpha x}} dx = \beta_t \int_{\frac{1-\varepsilon}{2\beta_t}t}^{\frac{1+\varepsilon}{2\beta_t}t} e^{-\alpha\beta_t\left(y+\frac{1}{y}\right)} dy.$$

Next, we introduce the new change of variable

$$y = \frac{2+z^2+z\sqrt{z^2+4}}{2}.$$

This defines a C^1-diffeomorphism from $z \in (-\infty, +\infty)$ to $y \in (0, +\infty)$ such that

$$y + \frac{1}{y} = 2 + z^2.$$

Note that $z > 0$ if and only if $y > 1$, which means that

$$z = \operatorname{sgn}(y-1)\sqrt{y+\frac{1}{y}-2},$$

where $\mathrm{sgn}(x) = 1$ if $x \geq 0$ and -1 if $x < 0$. Since

$$\beta_t = \frac{t - \frac{\log t}{2\alpha} + c}{2} \sim \frac{t}{2}, \tag{3.17}$$

the inequality $(1-\varepsilon)t/2\beta_t < 1 < (1+\varepsilon)t/2\beta_t$ holds for t large enough, which yields

$$\int_{\frac{1-\varepsilon}{2}t}^{\frac{1+\varepsilon}{2}t} e^{-\alpha x - \frac{x_t(c)\psi'(\alpha)}{\alpha x}} dx = \beta_t \int_{-\sqrt{\frac{(1-\varepsilon)t}{2\beta_t} + \frac{2\beta_t}{(1-\varepsilon)t} - 2}}^{\sqrt{\frac{(1+\varepsilon)t}{2\beta_t} + \frac{2\beta_t}{(1+\varepsilon)t} - 2}} e^{-\alpha\beta_t(2+z^2)} \left(z + \frac{z^2+2}{\sqrt{z^2+4}} \right) dz.$$

Now,

$$\lim_{t\to+\infty} \sqrt{\frac{(1-\varepsilon)t}{2\beta_t} + \frac{2\beta_t}{(1-\varepsilon)t} - 2} = \frac{\varepsilon}{\sqrt{1-\varepsilon}} \qquad \text{and} \qquad \lim_{t\to+\infty} \sqrt{\frac{(1+\varepsilon)t}{2\beta_t} + \frac{2\beta_t}{(1+\varepsilon)t} - 2} = \frac{\varepsilon}{\sqrt{1+\varepsilon}}.$$

Since $z + \frac{z^2+2}{\sqrt{z^2+4}}$ is C^1 in the neighborhood of 0, with value 1 at $z = 0$, we obtain

$$\begin{aligned}(1 - C\varepsilon)\beta_t e^{-2\alpha\beta_t} \int_{-\sqrt{\frac{(1-\varepsilon)t}{2\beta_t} + \frac{2\beta_t}{(1-\varepsilon)t} - 2}}^{\sqrt{\frac{(1+\varepsilon)t}{2\beta_t} + \frac{2\beta_t}{(1+\varepsilon)t} - 2}} e^{-\alpha\beta_t z^2} dz &\leq \int_{\frac{1-\varepsilon}{2}t}^{\frac{1+\varepsilon}{2}t} e^{-\alpha x - \frac{x_t(c)\psi'(\alpha)}{\alpha x}} dx \\ &\leq (1 + C\varepsilon)\beta_t e^{-2\alpha\beta_t} \int_{-\sqrt{\frac{(1-\varepsilon)t}{2\beta_t} + \frac{2\beta_t}{(1-\varepsilon)t} - 2}}^{\sqrt{\frac{(1+\varepsilon)t}{2\beta_t} + \frac{2\beta_t}{(1+\varepsilon)t} - 2}} e^{-\alpha\beta_t z^2} dz\end{aligned}$$

for t large enough. Making the last change of variable $u = \sqrt{2\alpha\beta_t}\, z$ finally yields

$$(1 - C'\varepsilon)\sqrt{\frac{\pi\beta_t}{\alpha}} e^{-2\alpha\beta_t} \leq \int_{\frac{1-\varepsilon}{2}t}^{\frac{1+\varepsilon}{2}t} e^{-\alpha x - \frac{x_t(c)\psi'(\alpha)}{\alpha x}} dx \leq (1 + C'\varepsilon)\sqrt{\frac{\pi\beta_t}{\alpha}} e^{-2\alpha\beta_t}$$

for t large enough. Combining this with (3.16), we obtain that, for all $\varepsilon > 0$ small enough, there exists $t_0 > 0$ such that, for all $t > t_0$,

$$(1 - C''\varepsilon)\frac{2}{t} e^{B - \frac{\psi'(\alpha)}{2}} e^{\alpha t} \sqrt{\frac{\pi\beta_t}{\alpha}} e^{-2\alpha\beta_t} \leq E_t(\varepsilon) \leq (1 + C''\varepsilon)\frac{2}{t} e^{B - \frac{\psi'(\alpha)}{2}} e^{\alpha t} \sqrt{\frac{\pi\beta_t}{\alpha}} e^{-2\alpha\beta_t},$$

where the constant C'' is independent of ε and t. It then follows from (3.17) that

$$(1 - C''\varepsilon)\sqrt{\frac{2\pi}{\alpha}} e^{B - \frac{\psi'(\alpha)}{2}} e^{-\alpha c} \leq E_t(\varepsilon) \leq (1 + C''\varepsilon)\sqrt{\frac{2\pi}{\alpha}} e^{B - \frac{\psi'(\alpha)}{2}} e^{-\alpha c}.$$

Since we have shown in the beginning of the proof that $E_t(\varepsilon) = E_t(\varepsilon') + o(1)$ for all $\varepsilon' < \varepsilon$, the previous inequality applied to $E_t(\varepsilon')$ concludes the proof of Proposition 3.7. □

4 Large or old families: convergence in distribution for subcritical clonal families

In all this section, we assume $\alpha < \theta$. Our goal is to compute the joint limiting distribution when $t \to +\infty$ of the sizes of the largest families living at time t, and of the ages of the oldest families living at time t. Before this, we give estimates on the second factorial moment of the number of large families, used repeatedly in the sequel.

4.1 A preliminary lemma

Let us recall that, under $\mathbb{P}_t$, we can adopt the representation of the genealogy at time t by the coalescent point process $H_0, H_1, H_2, \ldots$, where $H_0 = +\infty$ and the $(H_i; i \geq 1)$ are i.i.d., killed at their first value $(= H_{N_t})$ larger than t. For all $i \geq 0$, we call *branch* i the lineage represented by H_i.

For all $t > 0$, $x \geq 1$, $0 \leq s_1 < s_2 \leq +\infty$, we define $K_t(x, s_1, s_2)$ as the number of haplotypes carried by more than x individuals alive at time t, whose original mutation occurred on the ancestral lineage (branch 0), and has age in $(s_1, s_2]$ (or, equivalently, in $(s_1, s_2 \wedge t]$).

Lemma 4.1 *For all $t > 0$, $x \geq 1$, $0 \leq s_1 < s_2 \leq t$, we have*

$$\mathbb{E}_t[M_t(x,s_1,s_2)(M_t(x,s_1,s_2)-1)] \leq 2\ \mathbb{E}_t K_t(x,s_1,s_2)\ \left[1 + \frac{\frac{1}{W(s_1)} - \frac{1}{W(t)}}{1 - \frac{1}{W(t)}}\ \mathbb{E}_t N_t\right] \times$$
$$\times \left[\frac{b(1+\theta(s_2-s_1))}{\alpha}\left(1 - \frac{1}{W_\theta(s_2)}\right)^{\lceil x \rceil - 1}\left(e^{-\theta s_1} + \int_0^{s_1} \theta e^{-\theta z}\frac{W_\theta(s_2) - W_\theta(s_1 - z)}{W_\theta(s_2)}dz\right)\right.$$
$$\left. +4\ \frac{\frac{1}{W(s_1)} - \frac{1}{W(t)}}{1 - \frac{1}{W(t)}}\ \mathbb{E}_t N_t\ \mathbb{E}_t K_t\ (\lceil x/2 \rceil, s_1, s_2)\right]. \quad (4.1)$$

and

$$\mathbb{E}_t K_t(x,s_1,s_2) \leq \frac{b}{\alpha}\int_{s_1}^{s_2}\left(1 - \frac{1}{W_\theta(y)}\right)^{\lceil x \rceil - 1}\left(e^{-\theta y} + \int_0^y \theta e^{-\theta z}\frac{W_\theta(y) - W_\theta(y-z)}{W_\theta(y)}dz\right)(\theta dy + \delta_t(dy)) \quad (4.2)$$

$$\leq \frac{b}{\alpha}\int_{s_1}^{s_2}\left(1 - \frac{1}{W_\theta(y)}\right)^{\lceil x \rceil - 1}(\theta dy + \delta_t(dy)). \quad (4.3)$$

This lemma is proved in the Appendix.

4.2 Convergence in distribution of the size of the most frequent haplotype

Let us recall the notation $X_t^{(1)} \geq X_t^{(2)} \geq \ldots \geq X_t^{(k)} \geq \ldots$ for the ordered sequence of sizes of all living families in the population at time t (with the convention that $X_t^{(k)} = 0$ when k is larger than the number of living haplotypes at time t). Our first goal is to prove the convergence in distribution of $X^{(1)}$ using only the exact formulae (2.4) and (2.5) and the *coalescent point process* construction of the genealogy of the splitting tree.

The general idea is the following one. We divide the population at time t into several sub-populations corresponding to distinct ancestors at a given time s, as shown in Fig. 2. This gives a sequence of sub-trees $(\mathcal{T}_i)_{1 \leq i \leq N_{t,s}}$, where $N_{t,s}$ is the number of individuals alive at time s having descendants at time t. These individuals are represented by crosses in Fig. 2. We choose $s = s_t$ in such a way that $N_{t,s_t} \to \infty$ and the event under consideration (here, the event that there exists a haplotype carried by more than x_t individuals at time t) has a small probability in each sub-tree and are "nearly independent" (in a sense specified in the proofs below) in distinct sub-trees. The key argument of the proof, for which Lemma 4.1 is needed, consists in checking that, in each subtree, the

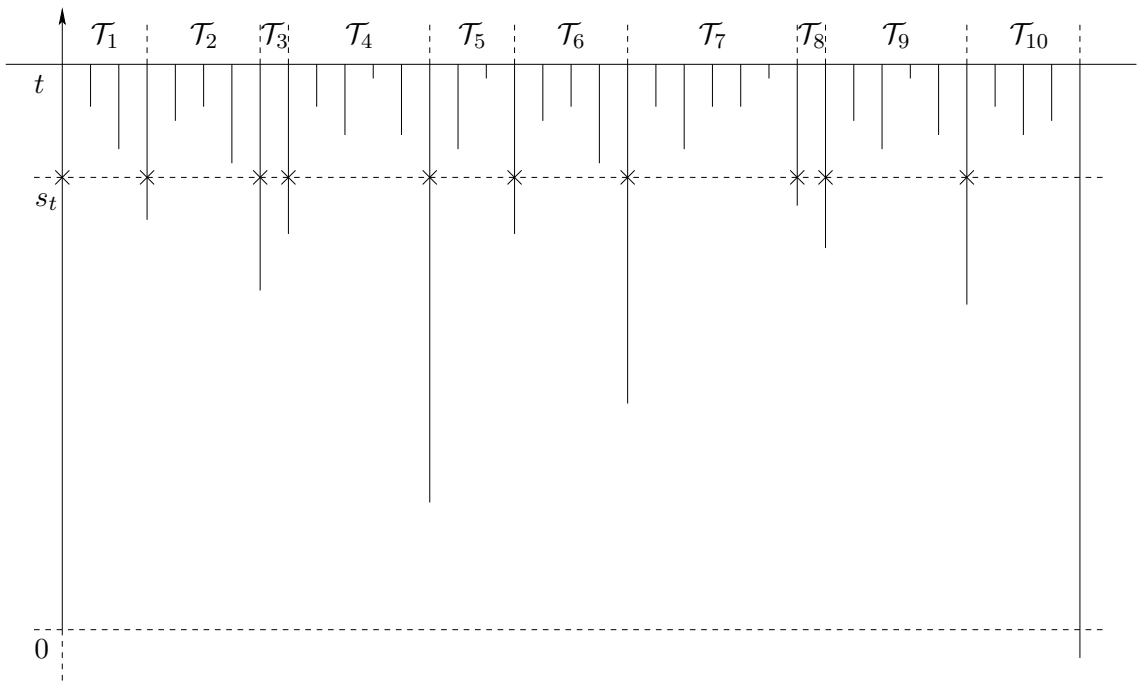

Figure 2: The definition of the sub-trees $(\mathcal{T}_i)_{1\leq i\leq N_{t,s_t}}$. The first vertical line represents the ancestral lineage (branch 0) and the other vertical lines have i.i.d. lengths $H_1, H_2, \ldots$ The rightmost vertical line is the first one higher than t, with length H_{N_t}. The crosses represent the N_{t,s_t} individuals alive at time s_t and having descendants at time t. Here, $N_{t,s_t} = 10$.

unknown probability that there exists a haplotype satisfying the property under consideration (here, carried by more than x_t individuals) is close to the expected number of such haplotypes, which is known explicitly. Here, this reads

$$\mathbb{P}_t[L_{t-s_t}(x_t) \geq 1] \sim \mathbb{E}_t L_{t-s_t}(x_t).$$

Theorem 4.2 *Assume $\alpha < \theta$. For all $c \in \mathbb{R}$, let*

$$x_t(c) := \frac{\alpha t - \frac{\theta}{\theta-\alpha}\log t}{|\log\varphi(\theta)|} + c. \tag{4.4}$$

Then

$$\mathbb{P}_t(X_t^{(1)} < x_t(c)) \sim \frac{1}{1 + A(\theta)\varphi(\theta)^{c-1+\{-x_t(c)\}}}$$

as $t \to +\infty$, where $A(\theta)$ is defined in (3.8).

Proof. Set

$$F(t,x) := \mathbb{P}_t(X_t^{(1)} \geq x) = \mathbb{P}_t[L_t(x) \geq 1], \qquad G(t,x) := \mathbb{E}_t[L_t(x)]$$

and

$$F(t,x,s) := \mathbb{P}_t[M_t(x,s) \geq 1].$$

We have for all $s < t$

$$0 \leq F(t,x) - \mathbb{P}_t\Big(\exists i \in \{1,\dots,N_{t,s}\} : \mathcal{T}_i \text{ contains a haplotype carried by more than } x \text{ individuals}\Big) \leq F(t,x,t-s).$$

which also reads

$$0 \leq F(t,x) - 1 + \mathbb{E}_t[(1-F(t-s,x))^{N_{t,s}}] = F(t,x) - \frac{1}{1+\mathbb{P}(H>t \mid H>t-s)\left(\frac{1}{F(t-s,x)}-1\right)} \leq F(t,x,t-s), \quad (4.5)$$

since $N_{t,s}$ is a geometric random variable of parameter $\mathbb{P}(H > t \mid H > t-s)$. In view of this, in order to find a non-trivial limit for $F(t,x_t)$, we need to find s_t and x_t such that $F(t,x_t,t-s_t) = o(1)$ and $F(t-s_t,x_t) = o(1)$ and is asymptotically equivalent to

$$\mathbb{P}(H > t \mid H > t-s_t) = W(t-s_t)/W(t) \sim e^{-\alpha s_t} \quad (4.6)$$

as $t \to +\infty$. In order to find an explicit asymptotic equivalent of $F(t-s_t,x_t)$, we will compare it with $G(t-s_t,x_t)$.

Let us check that the choice $x_t(c)$ in (4.4) for x_t and

$$s_t = s_t(b) = t - b\log t$$

satisfy the above properties for all $b > 1/(\theta-\alpha)$.

On the one hand, the fact that $F(t,x_t(c),t-s_t(b)) = o(1)$ is an immediate consequence of Proposition 3.5, since $b\log t > (1+\varepsilon)\log t/(\theta-\alpha)$ for some $\varepsilon > 0$. On the other hand, we can compute an asymptotic equivalent of $G(t-s_t(b),x_t(c))$ following closely the computation of the proof of Proposition 3.5. Here are the main steps of the computation: we have

$$\begin{aligned} G(t-s_t(b),x_t(c)) &= W(t-s_t(b))\left(1-\frac{\psi(\theta)}{\theta}\right)^{\lceil x_t(c)\rceil-1}\int_0^{t-s_t(b)}\frac{e^{-\theta x}}{W_\theta(x)}(1-\rho(x))^{\lceil x_t(c)\rceil-1}(\theta dx+\delta_{t-s_t(b)}(dx)) \\ &\sim \frac{e^{-\alpha s_t(b)}\,t^{\theta/(\theta-\alpha)}\,\varphi(\theta)^{c-1+\{-x_t(c)\}}}{\psi'(\alpha)}\int_0^{t-s_t(b)}\frac{e^{-\theta x}}{W_\theta(x)}(1-\rho(x))^{\lceil x_t(c)\rceil-1}(\theta dx+\delta_{t-s_t(b)}(dx)). \end{aligned}$$

It is easy to deduce from Lemma 3.2 (ii) that the contribution of the Dirac mass and of the integral on the interval $[0,(1-\varepsilon)\log t/(\theta-\alpha)]$ for any fixed $\varepsilon \in (0,1)$ are both $o(e^{-\alpha s_t(b)})$. Using the fact that $W_\theta(x) \to \theta/\psi(\theta)$ when $x \to +\infty$ and the change of variable $y = B(\lceil x_t(c)\rceil-1)e^{-(\theta-\alpha)x}$, the contribution of the integral on the interval $[(1-\varepsilon)\log t/(\theta-\alpha), b\log t]$ is asymptotically equivalent to

$$\frac{\psi(\theta)e^{-\alpha s_t(b)}\varphi(\theta)^{c-1+\{-x_t(c)\}}}{\psi'(\alpha)(\theta-\alpha)}\left(\frac{t}{B(\lceil x_t(c)\rceil-1)}\right)^{\theta/(\theta-\alpha)} \int_{B(\lceil x_t(c)\rceil-1)\ t^{-b(\theta-\alpha)}}^{B(\lceil x_t(c)\rceil-1)\ t^{-(1-\varepsilon)}} y^{\alpha/(\theta-\alpha)}\exp\left((\lceil x_t(c)\rceil-1)\log\left(1-\rho\left(\frac{\log\frac{B(\lceil x_t(c)\rceil-1)}{y}}{\theta-\alpha}\right)\right)\right)dy.$$

As in the proof of Proposition 3.5, Lebesgue's dominated convergence theorem then yields

$$G(t-s_t(b),x_t(c)) \sim A(\theta)\,\varphi(\theta)^{c-1+\{-x_t(c)\}}\mathbb{P}(H>t \mid H>t-s_t(b)) \tag{4.7}$$

as $t\to+\infty$, where we used the fact that $x_t(c)t^{-b(\theta-\alpha)}\to 0$ when $t\to+\infty$ since $b>1/(\theta-\alpha)$.

Now, it only remains to check that

$$G(t-s_t(b),x_t(c)) \sim F(t-s_t(b),x_t(c)) \tag{4.8}$$

when $t\to+\infty$. Since for all $t,x>0$

$$0\le G(t,x)-F(t,x)=\mathbb{E}_t(L_t(x))-\mathbb{P}_t(L_t(x)\ge 1)\le \mathbb{E}_t[L_t(x)(L_t(x)-1)],$$

it is sufficient to prove that

$$\mathbb{E}_t\left[L_{t-s_t(b)}(x_t(c))\left(L_{t-s_t(b)}(x_t(c))-1\right)\right]=o\left(G(t-s_t(b),x_t(c))\right)=o(e^{-\alpha s_t(b)}). \tag{4.9}$$

Taking $s_1=0$ and $s_2=t-s_t(b)$ in Lemma 4.1 (4.1) yields

$$\begin{aligned}\mathbb{E}_t\left[L_{t-s_t(b)}(x_t(c))\left(L_{t-s_t(b)}(x_t(c))-1\right)\right] &\le 2\mathbb{E}_tK_{t-s_t(b)}(x_t(c))(1+\mathbb{E}_tN_{t-s_t(b)})\times\\ &\times\left[\frac{b(1+\theta(t-s_t(b)))}{\alpha}\left(1-\frac{1}{W_\theta(t-s_t(b))}\right)^{\lceil x_t(c)\rceil-1}+4\mathbb{E}_tN_{t-s_t(b)}\mathbb{E}_tK_{t-s_t(b)}(x_t(c)/2)\right],\end{aligned}$$

where we used the notation $K_s(x):=K_s(x,0,s)$. Using the inequality $\mathbb{E}_tK_s(x)\le\mathbb{E}_tL_s(x)=G(s,x)$ and the estimates

$$\begin{aligned}\mathbb{E}_tN_{t-s_t(b)}&=\frac{1}{\mathbb{P}(H>t-s_t(b))}=W(t-s_t(b))\sim\frac{e^{\alpha(t-s_t(b))}}{\psi'(\alpha)},\\ \left(1-\frac{1}{W_\theta(t-s_t(b))}\right)^{\lceil x_t(c)\rceil-1}&\le\varphi(\theta)^{\lceil x_t(c)\rceil-1}\le Ce^{-\alpha t}t^{\theta/(\theta-\alpha)},\end{aligned} \tag{4.10}$$

and

$$G(t-s_t(b),x_t(c)/2)\sim 2^{\theta/(\theta-\alpha)}\,A(\theta)\,\varphi(\theta)^{c/2-1+\{-x_t(c)/2\}}\,e^{-\alpha(s_t(b)-t/2)}\,t^{-\theta/2(\theta-\alpha)},$$

which can be proved following the same computation as above, we finally obtain for t large enough

$$\begin{aligned}&\mathbb{E}_t\left[L_{t-s_t(b)}(x_t(c))\left(L_{t-s_t(b)}(x_t(c))-1\right)\right]\\ &\quad\le C\,G(t-s_t(b),x_t(c))\left(e^{\alpha(t-s_t(b))}\,\log t\,e^{-\alpha t}t^{\theta/(\theta-\alpha)}+e^{2\alpha(t-s_t(b))}\,e^{-\alpha(s_t(b)-t/2)}\,t^{-\theta/2(\theta-\alpha)}\right),\end{aligned}$$

which entails (4.9) and concludes the proof of Theorem 4.2. □

4.3 Joint convergence in law of the sizes of the most abundant families

The general idea of the previous argument can be summarized as follows: we construct a random number N_{t,s_t} of i.i.d. random variables $X_1, \ldots, X_{N_{t,s_t}}$—the sizes of the most frequent haplotype in each sub-tree—such that $X_t^{(1)} = \max\{X_1, \ldots, X_{N_{t,s_t}}\}$ with high probability. Our previous result then corresponds to a classical argument of extreme value theory, which is known to extend easily to compute the extremes statistics and the joint law of the largest random variables among $X_1, \ldots, X_{N_{t,s_t}}$. The object of this subsection is to prove that this argument is valid in our situation.

Theorem 4.3 *Assume $\alpha < \theta$ and recall the definition of $x_t(c)$ in (4.4). For all $n \in \mathbb{N}$, $k_1, \ldots, k_n \in \mathbb{Z}_+$ and $c_1, \ldots, c_n$ such that $c_i \geq c_{i+1} + 1$ for all $i \in \{1, \ldots, n-1\}$, we have, as $t \to +\infty$,*

$$\mathbb{P}_t\Big[L_t(x_t(c_1)) = k_1,\ L_t(x_t(c_2)) - L_t(x_t(c_1)) = k_2, \ldots,\ L_t(x_t(c_n)) - L_t(x_t(c_{n-1})) = k_n\Big] \sim \binom{k_1 + \ldots + k_n}{k_1, \ldots, k_n} \frac{\tau_t(c_1)^{k_1}(\tau_t(c_2) - \tau_t(c_1))^{k_2} \ldots (\tau_t(c_n) - \tau_t(c_{n-1}))^{k_n}}{(1 + \tau_t(c_n))^{k_1 + \ldots + k_n + 1}}, \quad (4.11)$$

with

$$\tau_t(c) := A(\theta)\varphi(\theta)^{c-1+\{-x_t(c)\}}$$

for all $c \in \mathbb{R}$, where the constant $A(\theta)$ is defined in (3.8).

Proof. Let us denote by $A(t; c_1, \ldots, c_n; k_1, \ldots, k_n)$ the event in the probability in the l.h.s. of (4.11). Using the notation of the proof of Theorem 4.2, for all $b > 1/(\theta - \alpha)$, we define $B(t, b; c_1, \ldots, c_n; k_1, \ldots, k_n)$ the event that among the sub-trees $\mathcal{T}_1, \ldots, \mathcal{T}_{N_{t,s_t(b)}}$, there are exactly k_i haplotypes carried by a number of living individuals at time t belonging to $[x_t(c_{i+1}), x_t(c_i))$ for all $0 \leq i \leq n-1$, with the convention $c_0 = +\infty$. Then

$$\Big|\mathbb{P}_t(A(t; c_1, \ldots, c_n; k_1, \ldots, k_n)) - \mathbb{P}_t(B(t, b; c_1, \ldots, c_n; k_1, \ldots, k_n)\Big| \leq \mathbb{P}_t[M_t(x_t(c_n), t - s_t(b)) \geq 1] = F(t, x_t(c_n), t - s_t(b)) = o(1).$$

Now, for all fixed $t \geq 0$, using the notation $X_1, \ldots, X_{N_{t,s_t}}$ introduced above, we define for all $a \in \mathbb{R}$

$$S_t(a) = \#\{i \leq N_{t,s_t} : X_i \geq a\}.$$

Defining

$$C(t, b; c_1, \ldots, c_n; k_1, \ldots, k_n) := \Big\{S_t(x_t(c_1)) = k_1,\ S_t(x_t(c_2)) - S_t(x_t(c_1)) = k_2, \ldots,\ S_t(x_t(c_n)) - S_t(x_t(c_{n-1})) = k_n\Big\},$$

we have

$$
\begin{aligned}
&\Big|\mathbb{P}_t(B(t,b;c_1,\ldots,c_n;k_1,\ldots,k_n)) - \mathbb{P}_t(C(t,b;c_1,\ldots,c_n;k_1,\ldots,k_n))\Big| \\
&\le \mathbb{P}_t\Big[\exists i \le N_{t,s_t(b)} : \mathcal{T}_i \text{ contains at least 2 haplotypes carried by more than } x_t(c_n) \text{ individuals}\Big] \\
&\le \sum_{k\ge 1} \mathbb{P}(H>t \mid H>t-s_t(b))\,\mathbb{P}(H\le t \mid H>t-s_t(b))^{k-1}\,k\,\mathbb{E}_t\Big[L_{t-s_t(b)}(x_t(c_n))(L_{t-s_t(b)}(x_t(c_n))-1)\Big] \\
&= \frac{\mathbb{E}_t\left[L_{t-s_t(b)}(x_t(c_n))(L_{t-s_t(b)}(x_t(c_n))-1)\right]}{\mathbb{P}(H>t \mid H>t-s_t(b))},
\end{aligned}
$$

where we used the fact that $N_{t,s_t(b)}$ has geometric distribution with parameter $\mathbb{P}(H > t \mid H > t - s_t(b))$. Equations (4.6) and (4.9) then yield

$$
\Big|\mathbb{P}_t(A(t;c_1,\ldots,c_n;k_1,\ldots,k_n)) - \mathbb{P}_t(C(t,b;c_1,\ldots,c_n;k_1,\ldots,k_n)\Big| = o(1).
$$

Next, $\mathbb{P}_t(C(t,b;c_1,\ldots,c_n;k_1,\ldots,k_n))$ can be computed using standard extreme value techniques: conditioning on $N_{t,s_t(b)}$ and considering all the possible ways to realize this event, we have

$$
\begin{aligned}
&\mathbb{P}_t(C(t,b;c_1,\ldots,c_n;k_1,\ldots,k_n)) \\
&= \sum_{k\ge k_1+\ldots+k_n} \mathbb{P}(H>t \mid H>t-s_t(b))\,\mathbb{P}(H\le t \mid H>t-s_t(b))^{k-1}\binom{k}{k_1,\ldots,k_n}\times \\
&\qquad \mathbb{P}_t\Big[X_1,\ldots,X_{k_1}\ge x_t(c_1) > X_{k_1+1},\ldots,X_{k_1+k_2}\ge x_t(c_2) > \ldots \ge x_t(c_n) > X_{k_1+\ldots+k_n+1},\ldots,X_k\Big] \\
&= \sum_{k\ge k_1+\ldots+k_n} \mathbb{P}(H>t \mid H>t-s_t(b))\,\mathbb{P}(H\le t \mid H>t-s_t(b))^{k-1}\binom{k_1+\ldots+k_n}{k_1,\ldots,k_n}\binom{k}{k_1+\ldots+k_n}\times \\
&\qquad F(t-s_t(b),x_t(c_1))^{k_1}\left[F(t-s_t(b),x_t(c_2))-F(t-s_t(b),x_t(c_1))\right]^{k_2}\times\ldots \\
&\qquad\qquad \ldots\times\left[F(t-s_t(b),x_t(c_n))-F(t-s_t(b),x_t(c_{n-1}))\right]^{k_n}\left[1-F(t-s_t(b),x_t(c_n))\right]^{k-k_1-\ldots-k_n}.
\end{aligned}
$$

The equation

$$
\sum_{k\ge m}\binom{k}{m}x^{k-m} = \frac{1}{(1-x)^{m+1}}, \qquad \forall x\in(-1,1),\ m\in\mathbb{N}
$$

yields

$$
\begin{aligned}
&\mathbb{P}_t(C(t,b;c_1,\ldots,c_n;k_1,\ldots,k_n)) \\
&\quad = \frac{\mathbb{P}(H>t \mid H>t-s_t(b))\mathbb{P}(H\le t \mid H>t-s_t(b))^{k_1+\ldots+k_n-1}}{\left[\mathbb{P}(H>t \mid H>t-s_t(b))+F(t-s_t(b),x_t(c_n))\mathbb{P}(H\le t \mid H>t-s_t(b))\right]^{k_1+\ldots+k_n+1}}\times \\
&\quad \binom{k_1+\ldots+k_n}{k_1,\ldots,k_n}F(t-s_t(b),x_t(c_1))^{k_1}\left[F(t-s_t(b),x_t(c_2))-F(t-s_t(b),x_t(c_1))\right]^{k_2}\times\ldots \\
&\qquad\qquad \ldots\times\left[F(t-s_t(b),x_t(c_n))-F(t-s_t(b),x_t(c_{n-1}))\right]^{k_n},
\end{aligned}
$$

and Theorem 4.3 then follows from (4.6), (4.7) and (4.8). □

In order to state our next result, we define the number t_n by the equation $x_{t_n}(0) = n$. In view of (4.4), this equation has a unique solution t_n if n is large enough, say larger than n_0. In addition,

$$t_n \sim \frac{|\log \varphi(\theta)|\,(\theta - \alpha)}{\theta}\, n,$$

and hence $t_n \to +\infty$ as $n \to +\infty$.

We also recall the notation $\mathcal{M}(\mathbb{R})$ for the set of nonnegative σ-finite measures on $\mathbb{R}$, finite on $\mathbb{R}_+$, and the definition of the *semi-vague* topology as the one induced by all maps of the form

$$\nu \in \mathcal{M}(\mathbb{R}) \mapsto \int_{\mathbb{R}} u(x)\nu(dx),$$

for all continuous bounded function u on $\mathbb{R}$ such that there exists $x_0 \in \mathbb{R}$ such that $u(x) = 0$ for all $x \leq x_0$. Note that this topology is stronger than the usual vague topology, but weaker than the usual weak topology.

Corollary 4.4 *Assume $\alpha < \theta$. Then, the sequence of point measures $(Z_n)_{n \geq n_0}$ on $\mathbb{Z}$, defined by*

$$Z_n := \sum_{k \geq 1} \delta_{X_{t_n}^{(k)} - n},$$

converges as $n \to +\infty$ in $\mathbb{P}^\star$-distribution on the set $\mathcal{M}(\mathbb{R})$ equipped with the semi-vague topology to a mixed Poisson point measure on $\mathbb{Z}$ with intensity measure

$$\mathcal{E} A(\theta) \frac{\psi(\theta)}{\theta} \sum_{c \in \mathbb{Z}} \varphi(\theta)^{c-1}\, \delta_c,$$

where the mixture coefficient $\mathcal{E}$ has exponential distribution with parameter 1.

The proof of such results is quite standard (cf. [11] in the general context of point measures and [14] more specifically on extreme values). However, we shall give a proof for sake of completeness and because of the specificity of the semi-vague topology.

Note that one can also easily obtain the convergence, in the sense of finite-dimensional distributions, of any finite sequence of translated extreme family sizes towards the corresponding sequence of extreme points of the limit point measure in the previous result. We shall not prove this, but instead we refer to [14] for the proof of similar standard results.

Proof. Let us first prove the convergence in distribution when $\mathcal{M}(\mathbb{R})$ is equipped with the vague topology. This amounts to prove the joint convergence in distribution of the random variables $L_{t_n}(n+i) - L_{t_n}(n+i+1)) = L_{t_n}(x_{t_n}(i)) - L_{t_n}(x_{t_n}(i+1))$, giving the number of haplotypes represented by exactly $n+i$ individuals at time t_n, for $b \leq i \leq a$ for all $a < b$ in $\mathbb{Z}$. Fix $b < a$ in $\mathbb{Z}$ and $k_b, \ldots, k_a$

in $\mathbb{Z}_+$. On the one hand, we claim that

$$\begin{aligned}
&\lim_{n\to+\infty} \mathbb{P}_{t_n}(L_{t_n}(n+i) - L_{t_n}(n+i+1) = k_i,\ \forall b \le i \le a) \\
&\quad = \lim_{n\to+\infty} \sum_{k\ge 0} \mathbb{P}_{t_n}(L_{t_n}(x_{t_n}(i)) - L_{t_n}(x_{t_n}(i+1)) = k_i,\ \forall b \le i \le a \text{ and } L_{t_n}(x_{t_n}(a+1)) = k) \\
&\quad = \sum_{k\ge 0} \binom{k_a + \ldots + k_b + k}{k_a, \ldots, k_b, k} \frac{\tau_{t_n}(a+1)^k (\tau_{t_n}(a) - \tau_{t_n}(a+1))^{k_a} \ldots (\tau_{t_n}(b) - \tau_{t_n}(b+1))^{k_b}}{(1+\tau_t(b))^{k+k_b+\ldots+k_a+1}} \\
&\quad = \binom{k_a + \ldots + k_b}{k_a, \ldots, k_b} \frac{[A(1-\varphi)]^{k_a+\ldots+k_b}\, \varphi^{(a-1)k_a+\ldots+(b-1)k_b}}{[1 + A(\varphi^{b-1} - \varphi^a)]^{k_a+\ldots+k_b+1}},
\end{aligned}$$

where we used the fact that, for all $x \in \mathbb{Z}$ and $n \ge n_0$,

$$\tau_{t_n}(x) = A\varphi^{x-1} \qquad \text{with} \qquad A = A(\theta) \quad \text{and} \quad \varphi = \varphi(\theta) := 1 - \frac{\psi(\theta)}{\theta}.$$

This is an immediate consequence of Theorem 4.3, provided we can justify the exchange of the sum over k and the limit $n \to +\infty$, i.e. that we can control the remainder of the series uniformly over $n \ge n_0$. The following inequality, making use of Proposition 3.5, solves this question: for all $N \in \mathbb{N}$,

$$\begin{aligned}
&\mathbb{P}_{t_n}(L_{t_n}(x_{t_n}(i)) - L_{t_n}(x_{t_n}(i+1)) = k_i,\ \forall b \le i \le a \text{ and } L_{t_n}(x_{t_n}(a+1)) \ge N) \\
&\qquad \le \mathbb{P}_{t_n}(L_{t_n}(x_{t_n}(a+1)) \ge N) \le \frac{\mathbb{E}_{t_n} L_{t_n}(x_{t_n}(a+1))}{N} \le \frac{1}{N} \sup_{t\ge 0} \mathbb{E}_t L_t(x_t(a+1)) \le \frac{C}{N}
\end{aligned}$$

for some constant $C > 0$.

On the other hand, assume that $\mathcal{E}$ is an exponential random variable with parameter 1, and $(P_x)_{x\in\mathbb{Z}}$ is a sequence of r.v. with P_x distributed as a mixed Poisson with parameter $\mathcal{E}A(1-\varphi)\varphi^{x-1}$ and such that the r.v. $(P_x)_{x\in\mathbb{Z}}$ are independent conditionally on $\mathcal{E}$. Then,

$$\begin{aligned}
\mathbb{P}(P_i = k_i,\ \forall b \le i \le a) &= \int_0^\infty dx\, e^{-x}\, e^{-A(1-\varphi)x \sum_{m=b}^a \varphi^{m-1}} \frac{(xA(1-\varphi))^{k_a+\ldots+k_b}\, \varphi^{(a-1)k_a+\ldots+(b-1)k_b}}{k_a! \ldots k_b!} \\
&= \frac{[A(1-\varphi)]^{k_a+\ldots+k_b}\, \varphi^{(a-1)k_a+\ldots+(b-1)k_b}}{k_a! \ldots k_b!\, [1 + A(1-\varphi)\sum_{m=b}^a \varphi^{m-1}]^{k_a+\ldots+k_b+1}} \int_0^\infty y^{k_a+\ldots+k_b} e^{-y} dy \\
&= \binom{k_a + \ldots + k_b}{k_a, \ldots, k_b} \frac{[A(1-\varphi)]^{k_a+\ldots+k_b}\, \varphi^{(a-1)k_a+\ldots+(b-1)k_b}}{[1 + A(\varphi^{b-1} - \varphi^a)]^{k_a+\ldots+k_b+1}},
\end{aligned}$$

where we used the change of variable $y = x(1 + A(1-\varphi)\sum_{m=b}^a \varphi^{m-1})$. Observing that $\mathbb{P}_{t_n}$ converges to $\mathbb{P}^\star$ for the total variation norm, this ends the proof of Corollary 4.4 when $\mathcal{M}(\mathbb{R})$ is equipped with the vague topology.

To complete the proof of Corollary 4.4, since all the point measures Z_n have support in $\mathbb{Z}$, we need to check that, for any continuous bounded function f on $\mathbb{R}$ and any sequence $(u(k))_{k\in\mathbb{Z}}$ such that $u(k) = 0$ for all $k \le k_0$ for some $k_0 \in \mathbb{Z}$,

$$\lim_{n\to+\infty} \mathbb{E}_{t_n} f\left(\int_{\mathbb{Z}} u(x) Z_n(dx)\right) = \mathbb{E}\, f\left(\sum_{k>k_0} u(k) P_k\right).$$

Note first that the sum in the r.h.s. is almost surely finite since, conditional on $\mathcal{E}$, this is a sum of independent r.v. with only finitely many of them being non-zero by Borel-Cantelli's lemma.

Next, fix $\varepsilon > 0$ and let a and T be large enough so that

$$\frac{\tau_{t_n}(a)}{1+\tau_{t_n}(a)} = \frac{A\varphi^{a-1}}{1+A\varphi^{a-1}} \leq A\varphi^{a-1} \leq \varepsilon,$$

and, by Theorem 4.2,

$$\sup_{t\geq T} \mathbb{P}_t\left(X_t^{(1)} \geq x_t(a)\right) \leq 2\varepsilon.$$

Then, for all n such that $t_n \geq T$,

$$\left|\mathbb{E}_{t_n} f\left(\int_{\mathbb{Z}} u(x)Z_n(dx)\right) - \mathbb{E}_{t_n} f\left(\sum_{k=k_0+1}^{a} u(k)\big(L_{t_n}(x_{t_n}(k)) - L_{t_n}(x_{t_n}(k+1))\big)\right)\right| \leq 4\varepsilon\|f\|_\infty.$$

The first step of the proof then yields

$$\left|\mathbb{E}_{t_n} f\left(\int_{\mathbb{Z}} u(x)Z_n(dx)\right) - \mathbb{E}\, f\left(\sum_{k=k_0+1}^{a} u(k)P_k\right)\right| \leq (4\|f\|_\infty + 1)\varepsilon$$

for all n large enough. Since

$$\mathbb{P}(\exists k \geq a : P_k \geq 1) \leq \sum_{k\geq a} \mathbb{P}(P_k \geq 1) = \sum_{k\geq a}(1 - e^{-A\varphi^{k-1}}) \leq A\sum_{k\geq a}\varphi^{k-1} \leq \frac{\varepsilon}{1-\varphi},$$

we finally obtain

$$\left|\mathbb{E}_{t_n} f\left(\int_{\mathbb{Z}} u(x)Z_n(dx)\right) - \mathbb{E}\, f\left(\sum_{k>k_0} u(k)P_k\right)\right| \leq (4\|f\|_\infty + 1 + (1-\varphi)^{-1})\varepsilon,$$

which ends the proof of Corollary 4.4. □

4.4 Convergence in distribution of the ages of oldest families

The previous method can easily be extended to prove the convergence of the ages of the oldest families. Let us recall the notation $A_t^{(1)} \geq A_t^{(2)} \geq \ldots \geq A_t^{(k)} \geq \ldots$ for the ordered sequence of ages of all alive families at time t (with the convention that $A_t^{(k)} = 0$ when k is larger than the number of alive families at time t).

Theorem 4.5 *Assume $\alpha < \theta$ and define for all $a \in \mathbb{R}$*

$$x_t(a) = \frac{\alpha t}{\theta} + a.$$

For all $n \in \mathbb{N}$, $k_1, \ldots, k_n \in \mathbb{Z}_+$ and $a_1 > a_2 > \ldots > a_n$, we have

$$\lim_{t\to+\infty} \mathbb{P}_t\Big[O_t(x_t(a_1)) = k_1,\ O_t(x_t(a_2)) - O_t(x_t(a_1)) = k_2, \ldots,\ O_t(x_t(a_n)) - O_t(x_t(a_{n-1})) = k_n\Big]$$
$$= \frac{\theta\psi'(\alpha)}{\psi(\theta)}\binom{k_1+\ldots+k_n}{k_1,\ldots,k_n}\frac{e^{-\theta k_1 a_1}(e^{-\theta a_2} - e^{-\theta a_1})^{k_2}\ldots(e^{-\theta a_n} - e^{-\theta a_{n-1}})^{k_n}}{\left(\frac{\theta\psi'(\alpha)}{\psi(\theta)} + e^{-\theta a_n}\right)^{k_1+\ldots+k_n+1}}. \quad (4.12)$$

In addition, the family of $\mathcal{M}(\mathbb{R})$-valued random variables $(Z_t, t \geq 0)$, defined for all $t \geq 0$ by

$$Z_t := \sum_{k \geq 1} \delta_{A_t^{(k)} - \frac{\alpha t}{\theta}},$$

converges as $t \to +\infty$ in $\mathbb{P}^\star$-distribution in $\mathcal{M}(\mathbb{R})$ equipped with the semi-vague topology to a mixed Poisson point measure on $\mathbb{R}$ with intensity measure

$$\mathcal{E} \frac{\psi(\theta)}{\psi'(\alpha)} e^{-\theta a} \, da, \tag{4.13}$$

where the mixture coefficient $\mathcal{E}$ has exponential distribution with parameter 1.

The proof follows closely the lines of those of Theorems 4.2 and 4.3 and Corollary 4.4. As a first step, we prove the following lemma.

Lemma 4.6 *With the same notation as in Theorem 4.5, we have for all $a \in \mathbb{R}$*

$$\lim_{t \to +\infty} \mathbb{P}_t \left(A_t^{(1)} \leq \frac{\alpha t}{\theta} + a \right) = \frac{1}{1 + \frac{\psi(\theta)}{\theta \psi'(\alpha)} e^{-\theta a}}.$$

Proof. The proof of this lemma is similar to the one of Theorem 4.2. Defining for all $t, x \geq 0$

$$F(t,x) = \mathbb{P}_t(A_t^{(1)} \geq x) = \mathbb{P}_t[O_t(x) \geq 1] \qquad \text{and} \qquad G(t,x) = \mathbb{E}_t[O_t(x)],$$

we have for all $x < s < t$

$$\begin{aligned} 0 &\leq F(t,x) - \mathbb{P}_t\Big(\exists i \in \{1, \ldots, N_{t,s}\} : \mathcal{T}_i \text{ contains at time } t \text{ a haplotype older than } x\Big) \\ &= F(t,x) - \frac{1}{1 + \mathbb{P}(H > t \mid H > t-s) \left(\frac{1}{F(t-s,x)} - 1 \right)} \leq F(t, t-s) \leq G(t, t-s). \end{aligned} \tag{4.14}$$

Defining

$$s_t(b) = bt \qquad \text{with} \qquad b \in (0, 1 - \alpha/\theta),$$

following the proof of Proposition 3.4, one easily checks that $G(t, t - s_t(b)) = o(1)$ and

$$G(t - s_t(b), x_t(a)) \sim \frac{\psi(\theta)}{\theta \psi'(\alpha)} e^{-\theta a - \alpha s_t(b)}, \tag{4.15}$$

where we stick to the notation $x_t(a) = \frac{\alpha t}{\theta} + a$. Then Lemma 4.6 follows from (4.6) and the fact that

$$G(t - s_t(b), x_t(a)) \sim F(t - s_t(b), x_t(a))$$

as $t \to \infty$. To prove this last equation, it is sufficient to prove that

$$\mathbb{E}_t[O_{t - s_t(b)}(x_t(a))(O_{t - s_t(b)}(x_t(a)) - 1)] = o(e^{-\alpha s_t(b)}), \tag{4.16}$$

as in (4.9).

Now, we observe that

$$\mathbb{P}(H > s \mid H \le t) = \frac{\frac{1}{W(s)} - \frac{1}{W(t)}}{1 - \frac{1}{W(t)}} \sim \frac{1}{W(s)} \sim \psi'(\alpha)e^{-\alpha s} \quad \text{when } s, t \to +\infty \text{ with } t - s \to +\infty \tag{4.17}$$

and, by Lemma 3.2 (ii), for a constant C that may change from line to line,

$$\begin{aligned}
e^{-\theta x_t(a)} + \int_0^{x_t(a)} \theta e^{-\theta z} \frac{W_\theta(t - s_t(b)) - W_\theta(x_t(a) - z)}{W_\theta(t - s_t(b))} dz \\
&= e^{-\theta x_t(a)} + \int_0^{x_t(a)} \theta e^{-\theta z} W_\theta(x_t(a) - z)\varphi(\theta)[\rho(x_t(a) - z) - \rho(t - s_t(b))]dz \\
&\le e^{-\theta x_t(a)} + C \int_0^{x_t(a)} e^{-\theta z}\, e^{-(\theta-\alpha)(x_t(a)-z)} dz \\
&\le C e^{-(\theta-\alpha)x_t(a)}.
\end{aligned}$$

Combining these two facts with (4.10) and Lemma 4.1 (4.1) in which we take $x = 1$, $s_1 = x_t(a)$ and $s_2 = t - s_t(b)$, we have

$$\begin{aligned}
\mathbb{E}_t \left[O_{t-s_t(b)}(x_t(a)) \left(O_{t-s_t(b)}(x_t(a)) - 1 \right) \right] \le C\mathbb{E}_t K_{t-s_t(b)}(1, x_t(a), t - s_t(b)) \left(1 + e^{-\alpha x_t(a)}\, e^{\alpha(t-s_t(b))} \right) \times \\
\times \left[(1+t)e^{-(\theta-\alpha)x_t(a)} + e^{-\alpha x_t(a)}\, e^{\alpha(t-s_t(b))}\, \mathbb{E}_t K_{t-s_t(b)}(1, x_t(a), t - s_t(b)) \right].
\end{aligned} \tag{4.18}$$

Here, in contrast with the proof of Theorem 4.2, the bound $K_s(1, x, s) \le O_s(x)$ is not sufficient to obtain the desired result. Instead, we use Lemma 4.1 (4.2):

$$\begin{aligned}
&\mathbb{E}_t K_{t-s_t(b)}(1, x_t(a), t - s_t(b)) \\
&\quad \le \frac{b}{\alpha} \int_{x_t(a)}^{t-s_t(b)} \left(e^{-\theta y} + \int_0^y \theta e^{-\theta z} W_\theta(y - z)\varphi(\theta)[\rho(y - z) - \rho(y)]dz \right) (\theta dy + \delta_{t-s_t(b)}(dy)).
\end{aligned}$$

Hence, by Lemma 3.2 (ii) again,

$$\begin{aligned}
\mathbb{E}_t K_{t-s_t(b)}(1, x_t(a), t - s_t(b)) &\le C \int_{x_t(a)}^{t-s_t(b)} \left(e^{-\theta y} + \int_0^y e^{-(\theta-\alpha)(y-z)} e^{-\theta z} dz \right) (\theta dy + \delta_{t-s_t(b)}(dy)) \\
&\le C \int_{x_t(a)}^{t-s_t(b)} e^{-(\theta-\alpha)y} (\theta dy + \delta_{t-s_t(b)}(dy)) \\
&\le C e^{-(\theta-\alpha)x_t(a)}.
\end{aligned}$$

Together with (4.18), this yields

$$\mathbb{E}_t \left[O_{t-s_t(b)}(x_t(a)) \left(O_{t-s_t(b)}(x_t(a)) - 1 \right) \right] \le C(a) e^{-\alpha s_t(b)} \left(t e^{-(1-\alpha/\theta)\alpha t} + e^{-\alpha b t} \right) = o(e^{-\alpha s_t(b)}),$$

where the constant $C(a)$ depends on a but not on t. This concludes the proof of Lemma 4.6. □

Proof of Theorem 4.5 Equation (4.12) can be deduced from Lemma 4.6 exactly as Theorem 4.3 was deduced from Theorem 4.2. We leave the details to the reader.

In view of (4.12), the computation in the proof of Corollary 4.4 immediately proves (replacing $\tau_{t_n}(a)$ with $\frac{\psi(\theta)}{\theta\psi'(\alpha)}\,e^{-\alpha a}$) that, for all $a_1 > a_2 > \ldots > a_n$, the random vector

$$\Big(O_t(x_t(a_2)) - O_t(x_t(a_1)), \ldots, O_t(x_t(a_n)) - O_t(x_t(a_{n-1}))\Big)$$

converges in distribution as $t \to +\infty$ to a vector whose coordinates are independent conditionally on $\mathcal{E}$ and have mixed Poisson distributions with mixture coefficient $\mathcal{E}$ and parameters

$$\mathcal{E}\frac{\psi(\theta)}{\theta\psi'(\alpha)}(e^{-\theta a_2} - e^{-\theta a_1}), \ldots, \mathcal{E}\frac{\psi(\theta)}{\theta\psi'(\alpha)}(e^{-\theta a_n} - e^{-\theta a_{n-1}}).$$

It is then standard to deduce the convergence in distribution of Z_t to P on $\mathcal{M}(\mathbb{R})$ equipped with the vague topology (cf. e.g. [11, Thm. 4.7]). The semi-vague topology can then be handled similarly as in the proof of Corollary 4.4. Again, we leave the details to the reader. □

5 Large or old families: convergence in distribution for critical clonal families

The method that we used in the previous section can also be applied to the case where $\alpha = \theta$. All the proofs are similar, and we will only give details at places where the proofs differ. We use the same notation as in the previous section.

5.1 Frequent haplotypes

Theorem 5.1 *Assume $\alpha = \theta$. For all $c \in \mathbb{R}$, let*

$$x_t(c) = \frac{\alpha^2}{4\psi'(\alpha)}\left(t - \frac{\log t}{2\alpha} + c\right)^2.$$

For all $n \in \mathbb{N}$, $k_1, \ldots, k_n \in \mathbb{Z}_+$ and $c_1 > c_2 > \ldots > c_n$, we have

$$\begin{aligned}\lim_{t\to+\infty} \mathbb{P}_t\Big[L_t(x_t(c_1)) = k_1,\ L_t(x_t(c_2)) - L_t(x_t(c_1)) = k_2, \ldots,\ L_t(x_t(c_n)) - L_t(x_t(c_{n-1})) = k_n\Big]\\ = \sqrt{\frac{\alpha}{2\pi}}\,e^{-B+\frac{\psi'(\alpha)}{2}}\binom{k_1+\ldots+k_n}{k_1,\ldots,k_n}\frac{e^{-\alpha k_1 c_1}(e^{-\alpha c_2} - e^{-\alpha c_1})^{k_2}\ldots(e^{-\alpha c_n} - e^{-\alpha c_{n-1}})^{k_n}}{\left(\sqrt{\frac{\alpha}{2\pi}}\,e^{-B+\frac{\psi'(\alpha)}{2}} + e^{-\alpha c_n}\right)^{k_1+\ldots+k_n+1}},\end{aligned}$$

where the constant B is defined in (3.13). In addition, the family of $\mathcal{M}(\mathbb{R})$-valued random variables $(Z_t, t \geq 0)$, defined for all $t \geq 0$ by

$$Z_t := \sum_{k\geq 1} \delta_{\sqrt{X_t^{(k)}} - \frac{\alpha}{2\sqrt{\psi'(\alpha)}}\left(t - \frac{\log t}{2\alpha}\right)},$$

converges as $t \to +\infty$ in $\mathbb{P}^\star$-distribution in $\mathcal{M}(\mathbb{R})$ equipped with the semi-vague topology to a mixed Poisson point measure on $\mathbb{R}$ with intensity measure

$$\mathcal{E}\sqrt{2\pi\alpha}\, e^{B-\frac{\psi'(\alpha)}{2}}\, e^{-\alpha c}\, dc, \tag{5.1}$$

where the mixture coefficient $\mathcal{E}$ has exponential distribution with parameter 1.

The proof of this result is exactly the same as for Theorems 4.3 and 4.5, provided we can prove the following lemma.

Lemma 5.2 *With the same notation as in Theorem 5.1, for all $c \in \mathbb{R}$,*

$$\lim_{t\to+\infty} \mathbb{P}_t(X_t^{(1)} < x_t(c)) = \frac{1}{1+\sqrt{\frac{2\pi}{\alpha}}\, e^{B-\frac{\psi'(\alpha)}{2}}\, e^{-\alpha c}}.$$

Proof. The proof of this result is similar to the one of Theorem 4.2. Fix $\varepsilon > 0$. We first observe that Proposition 3.7 implies that

$$\mathbb{P}_t\left(M_t\Big(x_t(c), 0, \frac{1-\varepsilon}{2}t\Big) \geq 1\right) \leq \mathbb{E}_t M_t\Big(x_t(c), 0, \frac{1-\varepsilon}{2}t\Big) = o(1),$$

and thus

$$\mathbb{P}_t(X_t^{(1)} \geq x_t(c)) = \mathbb{P}_t\left(M_t\Big(x_t(c), \frac{1-\varepsilon}{2}t\Big) \geq 1\right) + o(1),$$

so that it is enough for us to study $\mathbb{P}_t(M_t(x_t(c), s_t^{(1)}) \geq 1)$, where we put

$$s_t^{(1)} := \frac{1-\varepsilon}{2}t.$$

Defining

$$F(t,x) = \mathbb{P}_t[M_t(x, s_t^{(1)}) \geq 1], \qquad G(t,x) = \mathbb{E}_t[M_t(x, s_t^{(1)})]$$

and

$$F(t,x,s) = \mathbb{P}_t[M_t(x,s) \geq 1],$$

we can make the same computation as in the proof of Theorem 4.2 to show that (4.5) holds true if $t - s > s_t^{(1)}$. So let us define

$$s_t(b) = bt, \qquad \text{where} \qquad b \in (0, 1/2).$$

By Proposition 3.7, we immediately have $F(t, x_t(c), t - s_t(b)) = o(1)$. We observe that

$$G(t - s_t(b), x_t(c)) = W(t - s_t(b)) \int_{s_t^{(1)}}^{t-s_t(b)} \frac{e^{-\alpha x}}{W_\alpha(x)} e^{(\lceil x_t(c)\rceil - 1)\log(1 - 1/W_\alpha(x))}(\alpha dx + \delta_{t-s_t(b)}(dx)).$$

Using the inequality $\log(1-x) \leq -x$, Lemma 3.2 (iii) entails that the contribution of the Dirac mass is

$$O\left(\frac{1}{t}\, e^{-x_t(c)/W_\alpha(t-s_t(b))}\right).$$

Fix $\eta \in (0,1)$. Using the expression of $x_t(c)$ and the fact that $1/W_\alpha(t) \geq (1-\eta)\psi'(\alpha)/(\alpha t)$ for t large enough, this last quantity is

$$O\left(e^{-\frac{\alpha(1-\eta)}{4(1-b)}t}\, t^{-\left(1-\frac{1-\eta}{4(1-b)}\right)}\right) = o\left(e^{-\alpha s_t(b)}\right),$$

where the last equality is valid if one chooses $\eta < (1-2b)^2$.

Hence

$$G(t-s_t(b), x_t(c)) = W(t-s_t(b)) \int_{s_t^{(1)}}^{t-s_t(b)} \frac{e^{-\alpha x}}{W_\alpha(x)} e^{(\lceil x_t(c)\rceil - 1)\log(1-1/W_\alpha(x))} \alpha dx + o\left(e^{-\alpha s_t(b)}\right).$$

Now, the integral in the r.h.s. is exactly the same as in (3.14), except for the interval of integration. We actually proved in the proof of Proposition 3.7 that, since $(1-b) > 1/2$, this integral is equivalent to

$$\int_{\frac{1-\varepsilon}{2}t}^{\frac{1+\varepsilon}{2}t} \frac{e^{-\alpha x}}{W_\alpha(x)} e^{(\lceil x_t(c)\rceil - 1)\log(1-1/W_\alpha(x))} \alpha dx,$$

which is itself equivalent to

$$\frac{1}{W(t)} \sqrt{\frac{2\pi}{\alpha}}\, e^{B - \frac{\psi'(\alpha)}{2}}\, e^{-\alpha c}.$$

Therefore,

$$G(t-s_t(b), x_t(c)) \sim e^{-\alpha s_t(b)} \sqrt{\frac{2\pi}{\alpha}}\, e^{B-\frac{\psi'(\alpha)}{2}}\, e^{-\alpha c}, \tag{5.2}$$

and, recalling that (4.5) holds (with our current notation), the proof of Lemma 5.2 will be completed if we can prove that

$$G(t-s_t(b), x_t(c)) \sim F(t-s_t(b), x_t(c))$$

as $t \to +\infty$. Again, this is implied by the estimate

$$\mathbb{E}_t[M_{t-s_t(b)}(x_t(c), s_t^{(1)})(M_{t-s_t(b)}(x_t(c), s_t^{(1)}) - 1)] = o(e^{-\alpha s_t(b)}), \tag{5.3}$$

which we now prove.

Applying Lemma 4.1 (4.1) with $x = x_t(c)$, $s_1 = s_t^{(1)}$ and $s_2 = t - s_t(b)$, and combining the result with (4.17) and the fact that, for all $s \leq t$,

$$e^{-\alpha s} + \int_0^s \alpha e^{-\alpha z} \frac{W_\alpha(t) - W_\alpha(s-z)}{W_\alpha(t)} dz \leq e^{-\alpha s} + \int_0^s \alpha e^{-\alpha z} = 1,$$

we obtain

$$\begin{aligned}
&\mathbb{E}_t[M_{t-s_t(b)}(x_t(c), s_t^{(1)})(M_{t-s_t(b)}(x_t(c), s_t^{(1)}) - 1)] \\
&\qquad \leq C\, \mathbb{E}_t K_{t-s_t(b)}(x_t(c), s_t^{(1)}, t-s_t(b)) \left(1 + e^{-\alpha s_t^{(1)}}\, e^{\alpha(t-s_t(b))}\right) \times \\
&\left[(1+t)\left(1 - \frac{1}{W_\alpha(t-s_t(b))}\right)^{\lceil x_t(c)\rceil - 1} + e^{-\alpha s_t^{(1)}}\, e^{\alpha(t-s_t(b))}\, \mathbb{E}_t K_{t-s_t(b)}\left(\frac{x_t(c)}{2}, s_t^{(1)}, t-s_t(b)\right)\right].
\end{aligned} \tag{5.4}$$

Fix again $\eta \in (0,1)$. Using the inequality $\log(1-x) \le -x$ and Lemma 3.2 (iii), we have for t large enough,

$$\begin{aligned}\left(1-\frac{1}{W_\alpha(t-s_t(b))}\right)^{\lceil x_t(c)\rceil-1} &\le C\exp\left(-\frac{\alpha^2}{4\psi'(\alpha)}\left(t^2-\frac{t\log t}{\alpha}+2ct\right)\frac{(1-\eta)\psi'(\alpha)}{\alpha(t-s_t(b))}\right)\\ &\le C\exp\left(-\frac{\alpha(1-\eta)}{4(1-b)}\,t\right)t^{1/2},\end{aligned}$$

where we used the inequality $1/(1-b)<2$ to upper bound the exponent of t in the last inequality. Using Lemma 4.1 (4.3), this last inequality yields

$$\begin{aligned}\mathbb{E}_t K_{t-s_t(b)}(x_t(c),s_t^{(1)},t-s_t(b)) &\le \frac{b}{\alpha}\int_{s_t^{(1)}}^{t-s_t(b)}\left(1-\frac{1}{W_\alpha(y)}\right)^{\lceil x_t(c)\rceil-1}(\alpha dy+\delta_0(dy))\\ &\le \frac{b}{\alpha}\left(1-\frac{1}{W_\alpha(t-s_t(b))}\right)^{\lceil x_t(c)\rceil-1}\int_{s_t^{(1)}}^{t-s_t(b)}(\alpha dy+\delta_0(dy))\\ &\le C\exp\left(-\frac{\alpha(1-\eta)}{4(1-b)}\,t\right)t^{3/2}.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}_t K_{t-s_t(b)}(x_t(c)/2,s_t^{(1)},t-s_t(b)) &\le \frac{b}{\alpha}\left(1-\frac{1}{W_\alpha(t-s_t(b))}\right)^{\lceil x_t(c)/2\rceil-1}\int_{s_t^{(1)}}^{t-s_t(b)}(\alpha dy+\delta_0(dy))\\ &\le C\exp\left(-\frac{\alpha(1-\eta)}{8(1-b)}\,t\right)t^{5/4}.\end{aligned}$$

Combining the previous inequalities with (5.4), we finally obtain

$$\begin{aligned}&\mathbb{E}_t[M_{t-s_t(b)}(x_t(c),s_t^{(1)})(M_{t-s_t(b)}(x_t(c),s_t^{(1)})-1)]\\ &\qquad\le C\,t^3\,e^{-\alpha s_t(b)}\exp\left(-\alpha t\left(\frac{1-\varepsilon}{2}+\frac{1-\eta}{4(1-b)}-1\right)\right)\times\\ &\qquad\qquad\left[\exp\left(-\alpha t\,\frac{1-\eta}{4(1-b)}\right)+\exp\left(-\alpha t\left(b+\frac{1-\varepsilon}{2}+\frac{1-\eta}{8(1-b)}-1\right)\right)\right].\end{aligned}$$

Remember now that ε and η are free parameters in $(0,1)$. We may assume that they are linked to b by the equation

$$\frac{1-\eta}{4(1-b)}=\frac{1-\varepsilon}{2},\qquad\text{or}\qquad b=\frac{1}{2}-\frac{\varepsilon-\eta}{2(1-\varepsilon)}$$

which is always possible since $b<1/2$. This yields

$$\begin{aligned}&\mathbb{E}_t[M_{t-s_t(b)}(x_t(c),s_t^{(1)})(M_{t-s_t(b)}(x_t(c),s_t^{(1)})-1)]\\ &\qquad\le C\,t^3\,e^{-\alpha s_t(b)}\left[\exp\left(-\alpha t\,\frac{1-3\varepsilon}{2}\right)+\exp\left(-\alpha t\left(\frac{1}{4}-\frac{\varepsilon-\eta}{2(1-\varepsilon)}-\frac{7\varepsilon}{4}\right)\right)\right].\end{aligned}$$

Taking b close enough to $1/2$ allows to take both ε and η as close to 0 as desired. Therefore, (5.3) is proved and the proof of Lemma 5.2 is completed. □

5.2 Old haplotypes

Theorem 5.3 *Assume $\alpha = \theta$ and define for all $a \in \mathbb{R}$*

$$x_t(a) = t - \frac{\log t}{\alpha} + a.$$

For all $n \in \mathbb{N}$, $k_1, \ldots, k_n \in \mathbb{Z}_+$ and $a_1 > a_2 > \ldots > a_n$, we have

$$\lim_{t\to+\infty} \mathbb{P}_t\Big[O_t(x_t(a_1)) = k_1,\ O_t(x_t(a_2)) - O_t(x_t(a_1)) = k_2, \ldots,\ O_t(x_t(a_n)) - O_t(x_t(a_{n-1})) = k_n\Big]$$
$$= \alpha \binom{k_1 + \ldots + k_n}{k_1, \ldots, k_n} \frac{e^{-\alpha k_1 a_1}(e^{-\alpha a_2} - e^{-\alpha a_1})^{k_2} \ldots (e^{-\alpha a_n} - e^{-\alpha a_{n-1}})^{k_n}}{(\alpha + e^{-\alpha a_n})^{k_1+\ldots+k_n+1}}. \quad (5.5)$$

In addition, the family of $\mathcal{M}(\mathbb{R})$-valued random variables $(Z_t, t \geq 0)$, defined for all $t \geq 0$ by

$$Z_t := \sum_{k\geq 1} \delta_{A_t^{(k)} - t + \frac{\log t}{\alpha}},$$

converges as $t \to +\infty$ in $\mathbb{P}^\star$-distribution in $\mathcal{M}(\mathbb{R})$ equipped with the semi-vague topology to a mixed Poisson point measure on $\mathbb{R}$ with intensity measure

$$\mathcal{E}\, e^{-\theta a}\, da, \quad (5.6)$$

where the mixture coefficient $\mathcal{E}$ has exponential distribution with parameter 1.

Again, this result follows from the next lemma exactly as Theorem 4.5 followed from Lemma 4.6.

Lemma 5.4 *For all $a \in \mathbb{R}$*

$$\lim_{t\to+\infty} \mathbb{P}_t\left(A_t^{(1)} \leq t - \frac{\log t}{\alpha} + a\right) = \frac{1}{1 + \frac{1}{\alpha} e^{-\alpha a}}.$$

Proof. We define $F(t,x)$ and $G(t,x)$ exactly as in the proof of Lemma 4.6 and we put

$$s_t(b) = \frac{b}{\alpha} \log t \qquad \text{with} \qquad b \in (0,1).$$

With this new notation, (4.14) holds true and Proposition 3.6 implies that $G(t, t - s_t(b)) = o(1)$. In addition, one checks exactly as in the proof of Proposition 3.6 that

$$G(t - s_t(b), x_t(a)) \sim \frac{e^{-\alpha a}}{\alpha}\, e^{-\alpha s_t(b)}$$

when $t \to +\infty$. The proof will then be completed if we can prove (4.16). We first observe that $W'_\alpha(x) = e^{-\alpha x} W'(x)$ is bounded thanks to (2.3). Therefore, by Lemma 3.2 (iii),

$$e^{-\alpha x_t(a)} + \int_0^{x_t(a)} \alpha e^{-\alpha z} \frac{W_\alpha(t - s_t(b)) - W_\alpha(x_t(a) - z)}{W_\alpha(t - s_t(b))} dz$$
$$\leq C\left(te^{-\alpha t} + \int_0^{x_t(a)} e^{-\alpha z} \frac{t - s_t(b) - x_t(a) + z}{t - s_t(b)} dz\right)$$
$$\leq C \frac{\log t}{t}.$$

Combining this inequality with (4.17) and Lemma 4.1 (4.1) in which we take $x = 1$, $s_1 = x_t(a)$ and $s_2 = t - s_t(b)$ yields

$$\mathbb{E}_t\left[O_{t-s_t(b)}(x_t(a))\left(O_{t-s_t(b)}(x_t(a))-1\right)\right] \leq C\mathbb{E}_t K_{t-s_t(b)}(1, x_t(a), t-s_t(b))\left(1+e^{\alpha(t-s_t(b)-x_t(a))}\right)\times \\ \left[(1+\log t)\frac{\log t}{t} + e^{\alpha(t-s_t(b)-x_t(a))}\mathbb{E}_t K_{t-s_t(b)}(1, x_t(a), t-s_t(b))\right]. \quad (5.7)$$

By Lemma 4.1 (4.2), we have

$$\mathbb{E}_t K_{t-s_t(b)}(1, x_t(a), t-s(b)) \leq \frac{b}{\alpha}\int_{x_t(a)}^{t-s_t(b)}(\alpha dy + \delta_{t-s_t(b)}(dy))\left(e^{-\alpha y} + \int_0^y \alpha e^{-\alpha z}\frac{W_\alpha(y)-W_\alpha(y-z)}{W_\alpha(y)}dz\right).$$

Using again the fact that $W'_\alpha(x)$ is bounded and that $1/W_\alpha(y) \leq C/y$ for all y large enough, we deduce that

$$\begin{aligned}\mathbb{E}_t K_{t-s_t(b)}(1, x_t(a), t-s_t(b)) &\leq \int_{x_t(a)}^{t-s_t(b)}(\alpha dy + \delta_{t-s_t(b)}(dy))\left(e^{-\alpha y} + \frac{C}{y}\int_0^y ze^{-\alpha z}dz\right) \\ &\leq C\int_{x_t(a)}^{t-s_t(b)}\frac{1}{y}(\alpha dy + \delta_{t-s_t(b)}(dy)) \\ &\leq C\frac{\log t}{t}.\end{aligned}$$

Therefore, it follows from (5.7) that

$$\mathbb{E}_t\left[O_{t-s_t(b)}(x_t(a))\left(O_{t-s_t(b)}(x_t(a))-1\right)\right] \leq C\, e^{-\alpha s_t(b)}\,\log t\left(\frac{(\log t)^2}{t} + \frac{\log t}{t^b}\right) = o(e^{-\alpha s_t(b)}).$$

This completes the proof of Lemma 5.4. □

A Proof of Lemma 4.1

Recall the notation introduced in Section 4.1.

As seen in the proof of Theorem 4.2, this result (and actually all the results of Sections 4 and 5) are consequences of estimates of the form

$$\mathbb{P}_t(M_t(x_t, s_t) \geq 1) \sim \mathbb{E}_t M_t(x_t, s_t)$$

as $t \to +\infty$, for convenient choices of x_t and s_t. We chose to prove this result using the inequality

$$0 \leq \mathbb{E}_t M_t(x_t, s_t) - \mathbb{P}_t(M_t(x_t, s_t) \geq 1) \leq \mathbb{E}_t[M_t(x_t, s_t)(M_t(x_t, s_t)-1)],$$

i.e. proving that

$$\mathbb{E}_t[M_t(x_t, s_t)(M_t(x_t, s_t)-1)] = o\left(\mathbb{E}_t M_t(x_t, s_t)\right).$$

Such results are obtained using Lemma 4.1, which is an immediate consequence of the following two lemmas.

We need to define the random variable $K'_t(x, s_1, s_2)$ by slightly modifying the definition of $K_t(x, s_1, s_2)$: introducing an independent random variable H' distributed as H_i conditional on $\{H_i < t\}$, $K'_t(x, s_1, s_2)$ is the number of haplotypes carried by more than x individuals alive at time t, whose last mutation occurred on branch 0, and is older than s_1 and younger than $s_2 \wedge H'$.

As a first step, we compute an upper bound of $\mathbb{E}_t[M_t(x, s_1, s_2)(M_t(x, s_1, s_2) - 1)]$ expressed in terms of K_t and K'_t.

Lemma A.1 *For all $t > 0$, $x \geq 1$, $0 \leq s_1 \leq s_2 \leq +\infty$, we have*

$$\begin{aligned}\mathbb{E}_t[M_t(x, s_1, s_2)(M_t(x, s_1, s_2) - 1)] \leq\ & \mathbb{E}_t[K_t(x, s_1, s_2)(K_t(x, s_1, s_2) - 1)] \\ & + (\mathbb{E}_t N_t)\mathbb{E}_t[K'_t(x, s_1, s_2)(K'_t(x, s_1, s_2) - 1)] \\ & + 8(\mathbb{E}_t N_t)(\mathbb{E}_t K_t(\lceil x/2 \rceil, s_1, s_2))(\mathbb{E}_t K'_t(x, s_1, s_2)) \\ & + 8(\mathbb{E}_t N_t)^2(\mathbb{E}_t K'_t(\lceil x/2 \rceil, s_1, s_2))(\mathbb{E}_t K'_t(x, s_1, s_2)).\end{aligned}$$

Proof. We let M_i be the number of mutations on branch i (this branch has length H_i), considering only the mutations younger than t when $i = 0$. For all $j \leq M_i$, we define ℓ_{ij} the duration elapsed *since* the j-th oldest mutation on branch i, with $\ell_{i(M_i+1)} = 0$, $\ell_{i0} = H_i$ and $\ell_{00} = t$. We also define M'_0 as the smallest $k \geq 1$ such that $\ell_{0k} \leq H'$ (and $M'_0 = 0$ there is no such $k \geq 1$).

For $0 \leq j \leq M_i$, denote by R_t^{ij} the number of individuals alive at time t descending clonally from the time interval $I_{ij} := (t - \ell_{ij}, t - \ell_{i(j+1)})$ on branch i. More specifically, for a progenitor individual alive on the time interval (a, b) and experiencing no mutation between times a and b, we refer to 'clonal descendants from the time interval (a, b)' as those individuals alive at t (including possibly the progenitor) descending from those daughters of the progenitor who were born during the time interval (a, b), and that still carry the same type the progenitor carried at time a. Using the notation $A_{ij} = A_{ij}(t, x, s_1, s_2) := \{R_t^{ij} \geq x,\ \ell_{ij} \in [s_1, s_2)\}$, we have

$$M_t(x, s_1, s_2) = \sum_{0 \leq j \leq M_0} \mathbb{1}_{A_{0j}} + \sum_{1 \leq i < N_t} \sum_{1 \leq j \leq M_i} \mathbb{1}_{A_{ij}},$$

$$K_t(x, s_1, s_2) = \sum_{0 \leq j \leq M_0} \mathbb{1}_{A_{0j}}$$

and

$$K'_t(x, s_1, s_2) = \sum_{0 \leq j \leq M'_0} \mathbb{1}_{A_{0j}}.$$

Therefore, by construction of the coalescent point process,

$$\begin{aligned}\mathbb{E}_t M_t(x, s_1, s_2) &= \mathbb{E}_t K_t(x, s_1, s_2) + \sum_{i \geq 1} \sum_{j \geq 1} \mathbb{P}_t(A_{ij}, i < N_t, j \leq M_i) \\ &= \mathbb{E}_t K_t(x, s_1, s_2) + \sum_{i \geq 1} \sum_{j \geq 1} \mathbb{P}_t(A_{ij}, j \leq M_i \mid i < N_t)\mathbb{P}_t(i < N_t) \\ &= \mathbb{E}_t K_t(x, s_1, s_2) + \sum_{i \geq 1} \sum_{j \geq 1} \mathbb{P}_t(A_{0j}, j \leq M'_0)\mathbb{P}_t(i < N_t) \\ &= \mathbb{E}_t K_t(x, s_1, s_2) + (\mathbb{E}_t N_t - 1)\mathbb{E}_t K'_t(x, s_1, s_2).\end{aligned}$$

Now,

$$M_t(x,s_1,s_2)(M_t(x,s_1,s_2)-1) = 2\sum_{0\le k<j\le M_0} \mathbb{1}_{A_{0k}}\mathbb{1}_{A_{0j}} + 2\sum_{i=1}^{N_t-1}\sum_{1\le k<j\le M_i} \mathbb{1}_{A_{ik}}\mathbb{1}_{A_{ij}}$$
$$+ 2\sum_{k=0}^{M_0}\sum_{i=1}^{N_t-1}\sum_{j=1}^{M_i} \mathbb{1}_{A_{0k}}\mathbb{1}_{A_{ij}} + 2\sum_{1\le l<i<N_t}\sum_{k=1}^{M_l}\sum_{j=1}^{M_i} \mathbb{1}_{A_{lk}}\mathbb{1}_{A_{ij}}.$$

Hence, using a similar computation as above,

$$\begin{aligned}
&\mathbb{E}_t M_t(x,s_1,s_2)(M_t(x,s_1,s_2)-1)\\
&\quad= \mathbb{E}_t K_t(x,s_1,s_2)(K_t(x,s_1,s_2)-1) + 2\sum_{i\ge1}\sum_{k\ge1}\sum_{j>k}\mathbb{P}_t(A_{0k}\cap A_{0j}, j\le M_0')\mathbb{P}_t(i<N_t)\\
&\quad+ 2\sum_{k\ge0}\sum_{i\ge1}\sum_{j\ge1}\mathbb{P}_t(A_{0k}\cap A_{ij}, k\le M_0, i<N_t, j\le M_i)\\
&\quad+ 2\sum_{l\ge1}\sum_{k\ge1}\sum_{i>l}\sum_{j\ge1}\mathbb{P}_t(A_{0k}\cap A_{(i-l)j}, k\le M_0', i-l<N_t, j\le M_{i-l})\mathbb{P}_t(l<N_t)\\
&\quad= \mathbb{E}_t K_t(x,s_1,s_2)(K_t(x,s_1,s_2)-1) + (\mathbb{E}_t N_t-1)\mathbb{E}_t K_t'(x,s_1,s_2)(K_t'(x,s_1,s_2)-1)\\
&\quad+ 2\sum_{k\ge0}\sum_{i\ge1}\sum_{j\ge1}\mathbb{P}_t(A_{0k}\cap A_{ij}, k\le M_0, i<N_t, j\le M_i)\\
&\quad+ 2(\mathbb{E}_t N_t-1)\sum_{k\ge1}\sum_{i\ge1}\sum_{j\ge1}\mathbb{P}_t(A_{0k}\cap A_{ij}, k\le M_0', i<N_t, j\le M_i).
\end{aligned}$$

For short, we write

$$\begin{aligned}
&\mathbb{E}_t M_t(x,s_1,s_2)(M_t(x,s_1,s_2)-1)\\
&\quad\le \mathbb{E}_t K_t(x,s_1,s_2)(K_t(x,s_1,s_2)-1) + (\mathbb{E}_t N_t)\mathbb{E}_t K_t'(x,s_1,s_2)(K_t'(x,s_1,s_2)-1)\\
&\quad+ 2\sum_{k\ge0}\sum_{i\ge1}\sum_{j\ge1}\mathbb{P}_t(A_{0k}\cap A_{ij}\cap B_{ijk}) + 2(\mathbb{E}_t N_t)\sum_{k\ge1}\sum_{i\ge1}\sum_{j\ge1}\mathbb{P}_t(A_{0k}\cap A_{ij}\cap B_{ijk}') \qquad \text{(A.1)}
\end{aligned}$$

where $B_{ijk} := \{k\le M_0, i<N_t, j\le M_i\}$ and $B_{ijk}' := \{k\le M_0', i<N_t, j\le M_i\}$.

Now for any positive integers i,k, define the three following events

$$\alpha_{ik} := \{\max_{1\le j\le i} H_j > \ell_{0k}\}, \qquad \beta_{ik} := \{\ell_{0(k+1)} < \max_{1\le j\le i} H_j \le \ell_{0k}\}, \qquad \gamma_{ik} := \{\max_{1\le j\le i} H_j \le \ell_{0(k+1)}\}.$$

We are going to state and prove six inequalities, where the left-hand side is obtained by intersecting each event $A_{0k}\cap A_{ij}\cap B_{ijk}$ or $A_{0k}\cap A_{ij}\cap B_{ijk}'$ with each of the preceding ones α_{ik}, β_{ik}, γ_{ik}, and summing over $i,j\ge1$ and $k\ge0$ for the events involving B_{ijk}, and $k\ge1$ for the events involving

B'_{ijk}.

$$\sum_{i,j,k} \mathbb{P}_t(\alpha_{ik} \cap A_{0k} \cap A_{ij} \cap B_{ijk}) \leq (\mathbb{E}_t N_t)(\mathbb{E}_t K_t(x, s_1, s_2))(\mathbb{E}_t K'_t(x, s_1, s_2)), \tag{A.2}$$

$$\sum_{i,j,k} \mathbb{P}_t(\alpha_{ik} \cap A_{0k} \cap A_{ij} \cap B'_{ijk}) \leq (\mathbb{E}_t N_t)(\mathbb{E}_t K'_t(x, s_1, s_2))^2, \tag{A.3}$$

$$\sum_{i,j,k} \mathbb{P}_t(\gamma_{ik} \cap A_{0k} \cap A_{ij} \cap B_{ijk}) \leq (\mathbb{E}_t N_t)(\mathbb{E}_t K_t(x, s_1, s_2))(\mathbb{E}_t K'_t(x, s_1, s_2)), \tag{A.4}$$

$$\sum_{i,j,k} \mathbb{P}_t(\gamma_{ik} \cap A_{0k} \cap A_{ij} \cap B'_{ijk}) \leq (\mathbb{E}_t N_t)(\mathbb{E}_t K'_t(x, s_1, s_2))^2, \tag{A.5}$$

$$\sum_{i,j,k} \mathbb{P}_t(\beta_{ik} \cap A_{0k} \cap A_{ij} \cap B_{ijk}) \leq 2(\mathbb{E}_t N_t)(\mathbb{E}_t K_t(\lceil x/2 \rceil, s_1, s_2))(\mathbb{E}_t K'_t(x, s_1, s_2)), \tag{A.6}$$

$$\sum_{i,j,k} \mathbb{P}_t(\beta_{ik} \cap A_{0k} \cap A_{ij} \cap B'_{ijk}) \leq 2(\mathbb{E}_t N_t)(\mathbb{E}_t K'_t(\lceil x/2 \rceil, s_1, s_2))(\mathbb{E}_t K'_t(x, s_1, s_2)). \tag{A.7}$$

Combining these six equations with (A.1) and with the inequalities $K_t(x, s_1, s_2) \leq K_t(\lceil x/2 \rceil, s_1, s_2)$ and $K'_t(x, s_1, s_2) \leq K'_t(\lceil x/2 \rceil, s_1, s_2)$ yields the inequality given in the lemma.

We are going to detail the proof of the inequalities (A.2), (A.4) and (A.6) (in this order). The other inequalities can be proved using the same computations. Let us start with (A.2). Hereafter we denote by $A_{0k}^{(i)}$ the event $\{I_{0k}$ has more than x clonal descendants within $\{0, \ldots, i-1\}$ and $\ell_{0k} \in [s_1, s_2)\}$.

$$\begin{aligned}
\mathbb{P}_t(\alpha_{ik} \cap A_{0k} \cap A_{ij} \cap B_{ijk}) &= \mathbb{P}_t(\alpha_{ik} \cap A_{0k}^{(i)} \cap A_{ij} \cap B_{ijk}) \\
&\leq \mathbb{P}_t(k \leq M_0, A_{0k}^{(i)}, i < N_t, j \leq M_i, A_{ij}) \\
&= \mathbb{P}_t(k \leq M_0, A_{0k}^{(i)}, i < N_t, j \leq M_i, A_{ij} \mid i < N_t)\mathbb{P}_t(i < N_t) \\
&= \mathbb{P}_t(k \leq M_0, A_{0k}^{(i)} \mid i < N_t)\mathbb{P}_t(A_{ij}, j \leq M_i \mid i < N_t)\mathbb{P}_t(i < N_t) \\
&= \mathbb{P}_t(k \leq M_0, A_{0k}^{(i)}, i < N_t)\mathbb{P}_t(A_{0j}, j \leq M'_0).
\end{aligned} \tag{A.8}$$

Then, denoting ρ_k the index of the $\lceil x \rceil$-th individual carrying the type of interval I_{0k} ($:= +\infty$ if there is no such individual),

$$\begin{aligned}
\sum_{i,j,k} \mathbb{P}_t(\alpha_{ik} \cap A_{0k} \cap A_{ij} \cap B_{ijk}) &\leq \mathbb{E}_t \left(\sum_{0 \leq k \leq M_0} \sum_{i \geq 1} \mathbb{1}_{\ell_{0k} \in [s_1, s_2)} \mathbb{1}_{\rho_k \leq i < N_t} \right) \mathbb{E}_t \sum_{j \leq M'_0} \mathbb{1}_{A_{0j}} \\
&= \mathbb{E}_t \left(\sum_{0 \leq k \leq M_0} \mathbb{1}_{\ell_{0k} \in [s_1, s_2)} (N_t - \rho_k)^+ \right) \mathbb{E}_t K'_t(x, s_1, s_2).
\end{aligned}$$

Now, the lack of memory property of geometric distributions yields

$$\begin{aligned}
&\mathbb{E}_t\left(\sum_{0\le k\le M_0} \mathbb{1}_{\ell_{0k}\in[s_1,s_2)}(N_t-\rho_k)^+\right)\\
&\qquad= \sum_{k\ge 0}\mathbb{E}_t\left(N_t-\rho_k \mid k\le M_0, \ell_{0k}\in[s_1,s_2), \rho_k<\infty\right)\mathbb{P}_t\left(k\le M_0, \ell_{0k}\in[s_1,s_2), \rho_k<\infty\right)\\
&\qquad= (\mathbb{E}_t N_t)\sum_{k\ge 0}\mathbb{P}_t\left(k\le M_0, A_{0k}\right)\\
&\qquad= (\mathbb{E}_t N_t)(\mathbb{E}_t K_t(x,s_1,s_2)),
\end{aligned}$$

which entails (A.2).

Next, let us proceed with (A.4) and let σ_k denote the label of the first branch with length greater than $\ell_{0(k+1)}$. Observe that conditionally on $\ell_{0(k+1)}$, A_{0k} is independent of the branch lengths occurring before σ_k, and further, the events $\{i\le\sigma_k\}=\{\max_{j\le i-1}H_j\le\ell_{0(k+1)}\}$, $\{H_i\le\ell_{0(k+1)}, j\le M_i, A_{ij}\}$ and $\{k\le M_0, A_{0k}\}$ are independent. As a consequence,

$$\begin{aligned}
\mathbb{P}_t(\gamma_{ik}\cap A_{0k}\cap A_{ij}\cap B_{ijk}) &= \mathbb{P}_t(i<\sigma_k, k\le M_0, A_{0k}, j\le M_i, A_{ij})\\
&= \mathbb{P}_t(i\le\sigma_k, k\le M_0, A_{0k}, H_i\le\ell_{0(k+1)}, j\le M_i, A_{ij})\\
&= \mathbb{E}_t(\mathbb{P}_t(i\le\sigma_k \mid \ell_{0(k+1)})\mathbb{P}_t(A_{0k}, k\le M_0 \mid \ell_{0(k+1)})\mathbb{P}_t(H_i\le\ell_{0(k+1)}, j\le M_i, A_{ij} \mid \ell_{0(k+1)}))\\
&\le \mathbb{E}_t(\mathbb{P}_t(i\le\sigma_k, A_{0k}, k\le M_0 \mid \ell_{0(k+1)})\mathbb{P}_t(j\le M_i, A_{ij} \mid \ell_{0(k+1)}))\\
&= \mathbb{E}_t(\mathbb{P}_t(i\le\sigma_k, A_{0k}, k\le M_0 \mid \ell_{0(k+1)})\mathbb{P}_t(j\le M_i, A_{ij}))\\
&= \mathbb{P}_t(j\le M_0', A_{0j})\mathbb{P}_t(i\le\sigma_k, A_{0k}, k\le M_0).
\end{aligned}$$

As a consequence, since σ_k and $\{k\le M_0,\ A_{0k}\}$ are independent conditionally on $\ell_{0(k+1)}$,

$$\begin{aligned}
\sum_{i,j,k}\mathbb{P}_t(\gamma_{ik}\cap A_{0k}\cap A_{ij}\cap B_{ijk}) &\le (\mathbb{E}_t K_t'(x,s_1,s_2))\sum_{k\ge 0}\mathbb{E}_t\left(\sum_{i\le\sigma_k}\mathbb{1}_{k\le M_0, A_{0k}}\right)\\
&= (\mathbb{E}_t K_t'(x,s_1,s_2))\sum_{k\ge 0}\mathbb{E}_t\left(\mathbb{P}_t(k\le M_0, A_{0k} \mid \ell_{0(k+1)})\mathbb{E}_t(\sigma_k \mid \ell_{0(k+1)})\right)\\
&\le (\mathbb{E}_t K_t'(x,s_1,s_2))\sum_{k\ge 0}\mathbb{E}_t\left(\mathbb{P}_t(k\le M_0, A_{0k} \mid \ell_{0(k+1)})\mathbb{E}_t(N_t)\right)\\
&= (\mathbb{E}_t K_t'(x,s_1,s_2))(\mathbb{E}_t K_t(x,s_1,s_2))(\mathbb{E}_t N_t),
\end{aligned}$$

which is (A.4).

Finally, let us turn to (A.6). Denote by $A_{0k}^{\prime(i)}$ (resp. $A_{0k}^{\prime\prime(i)}$) the event that there exists at least $\lceil x/2\rceil$ individual with label *smaller* (resp. *larger*) than i descending clonally from the time interval I_{0k} and that $\ell_{0k}\in[s_1,s_2)$. Then

$$\mathbb{P}_t(\beta_{ik}\cap A_{0k}\cap A_{ij}\cap B_{ijk}) \le \mathbb{P}_t(\beta_{ik}\cap A_{0k}^{\prime(i)}\cap A_{ij}\cap B_{ijk}) + \mathbb{P}_t(\beta_{ik}\cap A_{0k}^{\prime\prime(i)}\cap A_{ij}\cap B_{ijk}).$$

Let us deal with the first term of the right-hand side of this last inequality. Exactly as in the proof of (A.2),

$$\begin{aligned}\mathbb{P}_t(\beta_{ik}, k \le M_0, A'^{(i)}_{0k}, j \le M_i, A_{ij}, i < N_t) &\le \mathbb{P}_t(k \le M_0, A'^{(i)}_{0k}, j \le M_i, A_{ij}, i < N_t)\\ &= \mathbb{P}_t(k \le M_0, A'^{(i)}_{0k}, i < N_t)\mathbb{P}_t(A_{ij}, j \le M_i \mid i < N_t)\\ &= \mathbb{P}_t(k \le M_0, A'^{(i)}_{0k}, i < N_t)\mathbb{P}_t(A_{0j}, j \le M'_0),\end{aligned}$$

and we finally get

$$\sum_{i,j,k} \mathbb{P}_t(\beta_{ik} \cap A'^{(i)}_{0k} \cap A_{ij} \cap B_{ijk}) \le (\mathbb{E}_t N_t)(\mathbb{E}_t K_t(\lceil x/2 \rceil, s_1, s_2))(\mathbb{E}_t K'_t(x, s_1, s_2)). \tag{A.9}$$

As for the second term, we need to define J_i the unique integer satisfying $\ell_{0(J_i+1)} < \max_{1 \le j \le i} H_j \le \ell_{0J_i}$ ($J_i := +\infty$ on $\{i \ge N_t\}$ and $J_i = k$ on β_{ik}). Then

$$\sum_{i,j,k} \mathbb{P}_t(\beta_{ik} \cap A''^{(i)}_{0k} \cap A_{ij} \cap B_{ijk}) = \sum_{i,j} \mathbb{P}_t(A''^{(i)}_{0J_i}, A_{ij}, j \le M_i, i < N_t). \tag{A.10}$$

Set also $\ell^*_i := \ell_{iM_i}$ the age of the oldest mutation on branch H_i ($\ell^*_i = 0$ if $M_i = 0$). Then conditional on $\{i < N_t\}$ and on the value of ℓ^*_i, the numbers of clonal descendants R^{ij}_t of the interval I_{ij} and the number, say $K^{(i)}$, of haplotypes whose last mutation is older than ℓ^*_i and s_1, younger than s_2, and occurred on lineage 0, and with more than $\lceil x/2 \rceil$ alive clonal descendants with labels larger than i, are independent, so that

$$\begin{aligned}\mathbb{P}_t(A''^{(i)}_{0J_i}, A_{ij}, j \le M_i, i < N_t) &\le \mathbb{P}_t(K^{(i)} \ge 1, A_{ij}, j \le M_i, i < N_t)\\ &= \mathbb{E}_t(\mathbb{1}_{i<N_t}\mathbb{P}_t(K^{(i)} \ge 1 \mid i < N_t, \ell^*_i)\mathbb{P}_t(A_{ij}, j \le M_i \mid i < N_t, \ell^*_i))\\ &\le \mathbb{E}_t(\mathbb{1}_{i<N_t}\mathbb{P}_t(K_t(\lceil x/2 \rceil, s_1, s_2) \ge 1)\mathbb{P}_t(A_{ij}, j \le M_i \mid i < N_t, \ell^*_i))\\ &= \mathbb{P}_t(K_t(\lceil x/2 \rceil, s_1, s_2) \ge 1)\mathbb{P}_t(A_{ij}, j \le M_i \mid i < N_t)\mathbb{P}_t(i < N_t) \quad \text{(A.11)}\\ &\le (\mathbb{E}_t K_t(\lceil x/2 \rceil, s_1, s_2))\mathbb{P}_t(A_{0j}, j \le M'_0)\mathbb{P}_t(i < N_t). \quad \text{(A.12)}\end{aligned}$$

We finally obtain

$$\sum_{i,j,k} \mathbb{P}_t(\beta_{ik} \cap A''^{(i)}_{0k} \cap A_{ij} \cap B_{ijk}) \le (\mathbb{E}_t N_t)(\mathbb{E}_t K_t(\lceil x/2 \rceil, s_1, s_2))(\mathbb{E}_t K'_t(x, s_1, s_2)),$$

which completes the proof of (A.6) by summing the last inequality and inequality (A.9).

The proof of (A.7) is very similar, but needs further explanation. Let us define the events $A'^{(i)}_{0k}$ and $A''^{(i)}_{0k}$ similarly as above, with the additional condition that $\ell_{0k} \le H'$. Then, we first prove that

$$\sum_{i,j,k} \mathbb{P}_t(\beta_{ik} \cap A'^{(i)}_{0k} \cap A_{ij} \cap B'_{ijk}) \le (\mathbb{E}_t N_t)(\mathbb{E}_t K'_t(\lceil x/2 \rceil, s_1, s_2))(\mathbb{E}_t K'_t(x, s_1, s_2))$$

following the very same computation as for (A.9). Next, we observe that $\mathbb{P}_t(A''^{(i)}_{00}) = 0$ since $H' < t$ a.s. and $\ell_{00} = t$. Therefore, (A.10) also holds true with our new definition of $A''^{(i)}_{0k}$. Thus, defining $K^{(i)}$ as the number of haplotypes whose last mutation is older than ℓ^*_i and s_1, younger than s_2 **and** H',

and occurred on lineage 0, and with more than $\lceil x/2 \rceil$ alive clonal descendants with labels larger than i, the computation of (A.12) is true, provided that $K_t(\lceil x/2 \rceil, s_1, s_2)$ is replaced by $K'_t(\lceil x/2 \rceil, s_1, s_2)$. We then obtain

$$\sum_{i,j,k} \mathbb{P}_t(\beta_{ik} \cap A''^{(i)}_{0k} \cap A_{ij} \cap B'_{ijk}) \leq (\mathbb{E}_t N_t)(\mathbb{E}_t K'_t(\lceil x/2 \rceil, s_1, s_2))(\mathbb{E}_t K'_t(x, s_1, s_2)),$$

and the proof of (A.7) is completed. □

Lemma 4.1 follows from the combination of the previous lemma with the following estimates on $K'_t(x, s_1, s_2)$ and $K_t(x, s_1, s_2)$.

Lemma A.2 *For all $t > 0$, $x \geq 1$, $0 \leq s_1 < s_2 \leq t$, we have*

$$\mathbb{E}_t K'_t(x, s_1, s_2) \leq \frac{\frac{1}{W(s_1)} - \frac{1}{W(t)}}{1 - \frac{1}{W(t)}} \, \mathbb{E}_t K_t(x, s_1, s_2), \tag{A.13}$$

$$\mathbb{E}_t K_t(x, s_1, s_2) \leq \frac{b}{\alpha} \int_{s_1}^{s_2} \left(1 - \frac{1}{W_\theta(y)}\right)^{\lceil x \rceil - 1} \left(e^{-\theta y} + \int_0^y \theta e^{-\theta z} \frac{W_\theta(y) - W_\theta(y-z)}{W_\theta(y)} dz\right) (\theta dy + \delta_t(dy)), \tag{A.14}$$

$$\mathbb{E}_t K'_t(x, s_1, s_2)(K'_t(x, s_1, s_2) - 1) \leq \frac{\frac{1}{W(s_1)} - \frac{1}{W(t)}}{1 - \frac{1}{W(t)}} \, \mathbb{E}_t K_t(x, s_1, s_2)(K_t(x, s_1, s_2) - 1), \tag{A.15}$$

$$\begin{aligned} \mathbb{E}_t K_t(x, s_1, s_2)(K_t(x, s_1, s_2) - 1) \leq \frac{2b}{\alpha} \left(\mathbb{E}_t K_t(x, s_1, s_2)\right) (1 + \theta(s_2 - s_1)) \times \\ \times \left(1 - \frac{1}{W_\theta(s_2)}\right)^{\lceil x \rceil - 1} \left(e^{-\theta s_1} + \int_0^{s_1} \theta e^{-\theta z} \frac{W_\theta(s_2) - W_\theta(s_1 - z)}{W_\theta(s_2)} dz\right). \end{aligned} \tag{A.16}$$

Proof. With the notation of the proof of Lemma A.1, we have

$$\begin{aligned} \mathbb{E}_t K'_t(x, s_1, s_2) &= \sum_{k \geq 1} \mathbb{P}_t(A_{0k}, k \leq M'_0) \\ &= \sum_{k \geq 1} \mathbb{P}_t(A_{0k}, k \leq M_0, H' \geq \ell_{0k}) \\ &\leq \sum_{k \geq 1} \mathbb{P}_t(A_{0k}, k \leq M_0) \mathbb{P}(H' \geq s_1) \\ &\leq \mathbb{E}_t K_t(x, s_1, s_2) \mathbb{P}(H \geq s_1 \mid H < t), \end{aligned}$$

which is inequality (A.13). Similarly,

$$\begin{aligned} \mathbb{E}_t K'_t(x, s_1, s_2)(K'_t(x, s_1, s_2) - 1) &= 2 \sum_{k \geq 1} \sum_{j > k} \mathbb{P}_t(A_{0k}, A_{0j}, j \leq M'_0) \\ &\leq \sum_{k \geq 1} \sum_{j > k} \mathbb{P}_t(A_{0k}, A_{0j}, j \leq M_0) \mathbb{P}(H' \geq s_1) \\ &\leq \mathbb{E}_t[K_t(x, s_1, s_2)(K_t(x, s_1, s_2) - 1] \mathbb{P}(H \geq s_1 \mid H < t), \end{aligned}$$

which is inequality (A.15).

For the two other inequalities, let us define $R_t^{(a,b)}$ the number of individuals alive at time t descending clonally from the time interval (a,b). More specifically, given a progenitor individual alive on the time interval (a,b) and experiencing no mutation between times a and b, $R_t^{(a,b)}$ is the number of individuals alive at time t (including this progenitor if $b \geq t$) descending from those daughters (including the daughters themselves) of the progenitor who were born during the time interval (a,b), and that still carry the same type that the progenitor carried at time a. Since W_θ is the scale function associated with the clonal reproduction process, for all $k \geq 0$,

$$\begin{aligned}\mathbb{P}(R_t^{(a,b)} = k) &= \mathbb{P}(N_{t-a}^\theta = k \mid \zeta = b-a) \\ &= \mathbb{P}(N_{t-a}^\theta \neq 0 \mid \zeta = b-a)\mathbb{P}(N_{t-a}^\theta = k \mid N_{t-a}^\theta \neq 0) \\ &= \left(1 - \mathbb{1}_{t>b}\frac{W_\theta(t-b)}{W_\theta(t-a)}\right)\left(1 - \frac{1}{W_\theta(t-a)}\right)^{k-1}\frac{1}{W_\theta(t-a)}, \end{aligned} \tag{A.17}$$

where N^θ is the population size process of a clonal splitting tree and ζ is the lifetime of the progenitor. (This result is actually Eq. (5.3) of [2].)

Note that, by construction of the splitting tree, replacing in the definition of $R_t^{(a,b)}$ the progenitor individual alive on the time interval (a,b) by a *clonal lineage* alive on the time interval (a,b), does not change anything to the distribution of $R_t^{(a,b)}$. By *lineage* alive on the time interval (a,b), we mean here a finite sequence of individuals $(i_k)_{1\leq k\leq K}$ such that individual i_1 was alive at time a, individual i_K was alive at time b, and for all $1 \leq k \leq K-1$, individual i_{k+1} was born from individual i_k at some time a_k such that $a_1 > a$ and $a_{K-1} < b$. By *clonal* lineage alive on the time interval (a,b) we mean in addition that for all $1 \leq k \leq K$, individual i_k experienced no mutation during the time interval (a_{k-1}, a_k), where $a_0 = a$ and $a_K = b$.

Now, by definition of $K_t(x, s_1, s_2)$, we have

$$\mathbb{E}_t K_t(x, s_1, s_2) = \sum_{k\geq 0} \mathbb{E}_t\left[\mathbb{1}_{\ell_{0k}\in[s_1,s_2]}\,\mathbb{P}_t\left(R_t^{0k} \geq x \mid \ell_{0j}, j \geq 0\right)\right],$$

where

$$\begin{aligned}\mathbb{P}_t\left(R_t^{0k} \geq x \mid \ell_{0j},\, j \geq 0\right) &= \frac{\mathbb{P}\left(R_t^{0k} \geq x,\ N_t \geq 1 \mid \ell_{0k},\, \ell_{0(k+1)}\right)}{\mathbb{P}(N_t \geq 1)} \\ &\leq \frac{b}{\alpha}\,\mathbb{P}\left(R_t^{0k} \geq x,\ N_t \geq 1 \mid \ell_{0k},\, \ell_{0(k+1)}\right),\end{aligned}$$

since $\mathbb{P}(N_t \geq 1) \geq \mathbb{P}(N_s \geq 1,\ \forall s \geq 0)$ and the survival probability of the splitting tree is α/b. Now, the event $\{N_t \geq 1,\ R_t^{0k} \geq x\}$ is the event where $N_t \geq 1$ and the clonal lineage on branch 0 on the time interval $(t-\ell_{0k}, t-\ell_{0(k+1)})$ has more than x clonal descendants alive at time t. Therefore,

$$\mathbb{P}_t\left(R_t^{0k} \geq x \mid \ell_{0j},\, j \geq 0\right) \leq \frac{b}{\alpha}\,\mathbb{P}\left(R_t^{(t-\ell_{0k}, t-\ell_{0(k+1)})} \geq x \mid \ell_{0k},\, \ell_{0(k+1)}\right). \tag{A.18}$$

Now, for all $k \geq 0$, $t - \ell_{0k}$ is distributed as the minimum of t and a sum of k i.i.d. exponential random variables of parameter θ, and $t - \ell_{0(k+1)}$ as the minimum of t and the sum of $t - \ell_{0k}$ and an

exponential random variable of parameter θ, independent of $t-\ell_{0k}$. Therefore, it follows from (A.17) that

$$\begin{aligned}
\frac{\alpha}{b}\,\mathbb{E}_t K_t(x,s_1,s_2) &\le \int_0^\infty dz\,\theta e^{-\theta z}\,\mathbb{1}_{s_2=t}\,\mathbb{P}(R_t^{(0,z)}\ge x)\\
&\qquad + \sum_{k\ge 1}\int_0^\infty dz\,\theta e^{-\theta z}\int_0^z dy\,\frac{\theta^k y^{k-1}}{(k-1)!}\,\mathbb{1}_{y\in[t-s_2,t-s_1]}\,\mathbb{P}(R_t^{(y,z)}\ge x)\\
&= \mathbb{1}_{s_2=t}\int_0^\infty dz\,\theta e^{-\theta z}\left(1-\mathbb{1}_{z<t}\frac{W_\theta(t-z)}{W_\theta(t)}\right)\left(1-\frac{1}{W_\theta(t)}\right)^{\lceil x\rceil-1}\\
&\qquad + \int_{t-s_2}^{t-s_1} dy\,\theta e^{\theta y}\int_y^\infty dz\,\theta e^{-\theta z}\left(1-\mathbb{1}_{z<t}\frac{W_\theta(t-z)}{W_\theta(t-y)}\right)\left(1-\frac{1}{W_\theta(t-y)}\right)^{\lceil x\rceil-1}\\
&= \int_{t-s_2}^{t-s_1}(\theta dy+\delta_0(dy))\,\left(1-\frac{1}{W_\theta(t-y)}\right)^{\lceil x\rceil-1}\left(1-\int_y^t dz\,\theta e^{-\theta(z-y)}\frac{W_\theta(t-z)}{W_\theta(t-y)}\right).
\end{aligned}$$

Equation (A.14) then follows from the changes of variables $z'=z-y$ and $y'=t-y$, and the identity $1=e^{-\theta y}+\int_0^y \theta e^{-\theta z}dz$.

Finally, let us turn to (A.16): first,

$$\mathbb{E}_t K_t(x,s_1,s_2)(K_t(x,s_1,s_2)-1) = 2\sum_{0\le j<k}\mathbb{P}_t(A_{0j},A_{0k},k\le M_0). \tag{A.19}$$

Now, fix $k>l\ge 0$. Since $\ell_{0j}>\ell_{0(j+1)}\ge\ell_{0k}>\ell_{0(k+1)}$, on the event $\{A_{0j},A_{0k},k\le M_0\}$, we have $\ell_{0(j+1)}-\ell_{0k}\le s_2-s_1$, $\ell_{0k}\le s_2$ and $\ell_{0(k+1)}\ge s_1-(\ell_{0k}-\ell_{0(k+1)})$. Therefore, using (A.18) as before,

$$\begin{aligned}
&\mathbb{P}_t(A_{0j},A_{0k},k\le M_0) = \mathbb{E}_t\left[\mathbb{1}_{\ell_{0j}\in[s_1,s_2]}\mathbb{P}_t\left(R_t^{0j}\ge x\mid \ell_{0j},\,\ell_{0(j+1)}\right)\mathbb{1}_{\ell_{0k}\in[s_1,s_2]}\mathbb{P}_t\left(R_t^{0k}\ge x\mid \ell_{0k},\,\ell_{0(k+1)}\right)\right]\\
&\le \mathbb{E}_t\Bigg[\mathbb{1}_{\ell_{0j}\in[s_1,s_2]}\,\mathbb{P}_t\left(R_t^{0j}\ge x\mid \ell_{0j},\,\ell_{0(j+1)}\right)\mathbb{1}_{\ell_{0(j+1)}-\ell_{0k}\le s_2-s_1}\times\\
&\qquad\times\,\mathbb{1}_{\ell_{0k}\ge s_1}\,\frac{b}{\alpha}\left(1-\mathbb{1}_{\ell_{0k}-\ell_{0(k+1)}<s_1}\frac{W_\theta(s_1-(\ell_{0k}-\ell_{0(k+1)}))}{W_\theta(s_2)}\right)\left(1-\frac{1}{W_\theta(s_2)}\right)^{\lceil x\rceil-1}\mathbb{1}_{\ell_{0k}>0}\Bigg],
\end{aligned}$$

where the last indicator comes from the fact that $1-1/W_\theta(\ell_{0k})=0$ when $\ell_{0k}=0$. Now, on the event $\{\ell_{0k}>0\}$, one has

$$\ell_{0n}=t-E_1-\ldots-E_n,\qquad \forall 0\le n\le k$$

and

$$\ell_{0(k+1)}=0\vee(t-E_1-\ldots-E_{k+1}),$$

where $(E_n)_{n\ge1}$ is a sequence of i.i.d. exponential r.v. of parameter θ. In addition, on the event $\{\ell_{0k}\ge s_1,\ \ell_{0k}-\ell_{0(k+1)}<s_1\}$, one has $\ell_{0k}-\ell_{0(k+1)}=E_{k+1}$. Hence,

$$\begin{aligned}
\mathbb{P}_t(A_{0j},A_{0k},k\le M_0) &\le \frac{b}{\alpha}\,\mathbb{E}_t\Bigg[\mathbb{1}_{L_{0j}\in[s_1,s_2]}\,\mathbb{P}_t\left(R_t^{0j}\ge x\mid \ell_{0j}=L_{0j},\ell_{0(j+1)}=L_{0(j+1)}\right)\times\\
&\qquad\times\,\mathbb{1}_{E_{j+2}+\ldots+E_k\le s_2-s_1}\left(1-\mathbb{1}_{E_{k+1}<s_1}\frac{W_\theta(s_1-E_{k+1})}{W_\theta(s_2)}\right)\left(1-\frac{1}{W_\theta(s_2)}\right)^{\lceil x\rceil-1}\Bigg],
\end{aligned}$$

where

$$L_{0j} = 0 \vee (t - E_1 - \ldots - E_j) \qquad \text{and} \qquad L_{0(j+1)} = 0 \vee (t - E_1 - \ldots - E_{j+1}).$$

Since $(L_{0j}, L_{0(j+1)})$, $(E_{j+2}, \ldots, E_k)$ and E_{k+1} are independent, we finally obtain

$$\begin{aligned}
\sum_{0 \le j < k} \mathbb{P}_t(A_{0j}, A_{0k}, k \le M_0) &\le \frac{b}{\alpha} \sum_{0 \le j < k} \mathbb{P}_t(A_{0j}, j \le M_0) \mathbb{P}(E_{j+2} + \ldots + E_k \le s_2 - s_1) \times \\
&\quad \times \left(1 - \mathbb{E}\left(\mathbb{1}_{E_{k+1} < s_1} \frac{W_\theta(s_1 - E_{k+1})}{W_\theta(s_2)}\right)\right) \left(1 - \frac{1}{W_\theta(s_2)}\right)^{\lceil x \rceil - 1} \\
&= \frac{b}{\alpha} \left(\mathbb{E}_t K_t(x, s_1, s_2)\right) \sum_{i \ge 0} \mathbb{P}(E_1 + \ldots + E_i \le s_2 - s_1) \times \\
&\quad \times \left(1 - \int_0^{s_1} \theta e^{-\theta z} \frac{W_\theta(s_1 - z)}{W_\theta(s_2)} dz\right) \left(1 - \frac{1}{W_\theta(s_2)}\right)^{\lceil x \rceil - 1}. \qquad \text{(A.20)}
\end{aligned}$$

Now, we have

$$\sum_{i \ge 0} \mathbb{P}(E_1 + \ldots + E_i \le s_2 - s_1) = 1 + \mathbb{E}(P) = 1 + \theta(s_2 - s_1),$$

where P is a Poisson r.v. of parameter $\theta(s_2 - s_1)$. Combining this equation with (A.19) and (A.20) ends the proof of (A.16). □

Acknowledgments. This work was funded by project MANEGE 'Modèles Aléatoires en Écologie, Génétique et Évolution' 09-BLAN-0215 of ANR (French national research agency).

References

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

[2] Champagnat, N., Lambert, A. (2012)
Splitting trees with neutral Poissonian mutations I: Small families. *Stoch. Proc. Appl.* **122(3)**, 1003–1033.

[3] Donnelly, P., Tavaré, S. (1986)
The ages of alleles and a coalescent. *Adv. in Appl. Prob.* **18** 1–19.

[4] Durrett, R. (2008)
Probability Models for DNA Sequence Evolution. Springer–Verlag, Berlin. 2nd revised ed.

[5] Ewens, W.J. (2005)
Mathematical Population Genetics. 2nd edition, Springer–Verlag, Berlin.

[6] Geiger, J. (1996)
Size-biased and conditioned random splitting trees. *Stoch. Proc. Appl.* **65** 187–207.

[7] Geiger, J., Kersting, G. (1997)
Depth-first search of random trees, and Poisson point processes, in *Classical and modern branching processes* (Minneapolis, 1994) IMA Math. Appl. Vol. 84. Springer-Verlag, New York.

[8] Jagers, P. (1974)
Convergence of general branching processes and functionals thereof. *J. Appl. Prob.* **11** 471–478.

[9] Jagers, P., Nerman, O. (1984)
The growth and composition of branching populations. *Adv. Appl. Prob.* **16** 221–259.

[10] Jagers, P., Nerman, O. (1984)
Limit theorems for sums determined by branching processes and other exponentially growing processes. *Stoch. Proc. Appl.* **17** 47–71.

[11] Kallenberg, O. (1986)
Random Measures. Academic Press, New York, London; Akademie-Verlag, Berlin, 4th ed.

[12] Lambert, A. (2009)
The allelic partition for coalescent point processes. *Markov Proc. Relat. Fields* **15** 359–386.

[13] Lambert, A. (2010)
The contour of splitting trees is a Lévy process. *Ann. Probab.* **38** 348—395.

[14] Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983)
Extremes and related properties of random sequences and processes. Springer-Verlag, New York.

[15] Nerman, O. (1981)
On the convergence of supercritical general (CMJ) branching processes. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **57** 365—395.

[16] Popovic, L. (2004)
Asymptotic genealogy of a critical branching process. *Ann. Appl. Prob.* **14** 2120–2148.

[17] Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., Ackerman, H.C., Campbell, S.J., Altshuler, D., Cooperk, R., Kwiatkowski, D., Ward, R., Lander, E.S. (2002)
Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419** 832–837.

[18] Taïb, Z. (1992)
Branching processes and neutral evolution. Lecture Notes in Biomathematics Vol. 93. Springer-Verlag, Berlin.