SITE FREQUENCY SPECTRUM IN STATIONARY BRANCHING POPULATIONS

ROMAIN ABRAHAM, JEAN-FRANÇOIS DELMAS, AND PATRICK HOSCHEIT

ABSTRACT. This paper explores the Site Frequency Spectrum (SFS) in stationary branching populations. We derive estimates for the SFS associated with a sample from a continuous-state branching process conditioned to never go extinct, utilizing a quadratic branching mechanism. The genealogy of such processes is represented by a real tree with a semi-infinite branch, and we compute the expectation of the SFS under the infinitely-many-sites assumption as the sample size approaches infinity. Additionally, we present a continuum version of the SFS as a random point measure on the positive real line and compute the density of its expected measure explicitly. Finally, we derive estimates for the size of the clonal subpopulation carrying the same genotype as the most recent common ancestor of the whole population at a given time.

1. INTRODUCTION

The Site Frequency Spectrum (SFS) of a genetic sample is a summary statistic of the full alignment that characterizes each mutation found in the sample by the number of individuals carrying it. It has been shown to reflect many features of the past dynamics of the population from which the sample was taken, including variations in ancestral population size, selection, or the existence of population structure. In this paper, we will give estimates for the expectation of the SFS associated to a sample from a continuous-state branching process conditioned never to go extinct. Such processes, first described in [21], are representing the size of an infinite stationary population undergoing branching. In the general case, the branching mechanism is described by the Laplace exponent of a spectrally positive Lévy process. Here, we will focus on the quadratic case in which the underlying Lévy process is a Brownian motion with positive drift. The genealogy of such stationary branching processes can be represented by a metric space $(\mathcal{T}, d)$ which is a *real tree* with an infinite branch. We will use the distribution of the subtree spanned by a uniform sample of n leaves at a given time, which was given in [1], to compute the expectation of the SFS of such a sample (Theorem 4.3), under the infinitely-many-sites assumption, as n goes to ∞. We also present a continuum version of the SFS as a random point measure on $\mathbb{R}_+$, following the framework introduced in [8]. We show that the expected measure has a density with respect to Lebesgue measure, which we compute explicitly (Theorem 5.1). Finally, we study the part of the population at a given time that carries no additional mutations compared to its most recent common ancestor. We compute the expectation of the size of this subpopulation, as well as the ratio between its size and the size of the whole extant population (Theorem 5.4).

We will now review some results from the literature on the SFS. Given a rooted real tree $\mathcal{T}$ with n leaves, let $\mathcal{M}$ be an independent Poisson point process on $\mathcal{T}$ with intensity $\mu > 0$. Each atom of $\mathcal{M}$ is a mutation that is carried by the whole subpopulation descended from it. Each mutation occurs at a different locus (*i.e.* we assume the *infinitely-many-sites model*), and we assume that the ancestral allele at each locus is known. Thus we can define the SFS of a

n-sample as the vector:

$$\xi^{(n)} = (\xi_1^{(n)}, \ldots, \xi_{n-1}^{(n)}),$$

where $\xi_k^{(n)}$ is the number of mutations carried by exactly k individuals in the sample.

When $\mathcal{T}$ is a Kingman coalescent tree with n leaves, the first moment of the SFS can be explicitly computed:

$$(1) \qquad\qquad \mathbb{E}[\xi_k^{(n)}] = \frac{\theta}{k}, \quad k = 1, \ldots, n-1,$$

where θ is the population-scaled mutation rate $\theta = 4N_e\mu$. In this expression, N_e is the effective population size parameter and μ as above, the per-lineage mutation rate. See [14] for a derivation of this expression, as well as results on second moments. This result was subsequently extended to accommodate relaxations of the strict assumptions underlying the Kingman coalescent. Notably, Griffiths and Tavaré [15] established the following formula for the expectation of the SFS:

$$(2) \qquad\qquad \mathbb{E}[\xi_k^{(n)}] = \frac{\theta}{2} \sum_{i=2}^{n-k+1} i\, p^{(n)}(i,k)\, \mathbb{E}[T_i^{(n)}],$$

where $p^{(n)}(i,k)$ is the probability that at the time the coalescent has i blocks, a given one of them contains exactly k leaves, and where $T_i^{(n)}$ is the amount of time when the coalescent has exactly i blocks. This formula holds for variable population sizes, but the expectations might not be explicitly computable.

Equation (2) can be generalized to the case of Λ-coalescents [5], or even Ξ-coalescents [6]. Asymptotic results for Λ-coalescents in the case where Λ is regularly varying at 0 with index $1 < \alpha < 2$ (meaning that $\Lambda(\mathrm{d}x) = f(x)\,\mathrm{d}x$ with $f(x) \sim Ax^{1-\alpha}$ as $x \to 0$) are found in [2]:

$$\lim_{n\to\infty} n^{\alpha-2} \xi_k^{(n)} = \frac{\theta}{2} C_{A,\alpha} \frac{(2-\alpha)\Gamma(k+\alpha-2)}{k!\,\Gamma(\alpha-1)},$$

almost surely for fixed $k \geq 1$, where the constant $C_{A,\alpha}$ is explicit. Recently, Kersting et al. [16] were able to obtain a closed integral formula for the SFS in the special case of the Bolthausen-Sznitman coalescent:

$$\mathbb{E}[\xi_k^{(n)}] = \theta n \int_0^1 \frac{\Gamma(k-p)}{\Gamma(k+1)} \frac{\Gamma(n-k+p)}{\Gamma(n-k+1)} \frac{dp}{\Gamma(1-p)\Gamma(1+p)},$$

which leads to the following asymptotics for large values of n:

$$(3) \qquad\qquad \mathbb{E}[\xi_k^{(n)}] \sim \begin{cases} \frac{\theta n}{\log n} & \text{if } k = 1, \\ \frac{\theta n}{k(k-1)} \frac{1}{\log^2(n/k)} & \text{if } k \geq 2, \ k/n \to 0, \\ \frac{\theta}{n} f_1(u) & \text{if } k/n \to u \in (0,1), \\ \frac{\theta}{n-k} \frac{1}{\log(n/(n-k))} & \text{if } 1 - k/n \to 0, \end{cases}$$

where $f_1(u) = \int_0^1 u^{-1-p}(1-u)^{p-1} \sin(\pi p)/(\pi p)\,dp$ is the asymptotic profile of the SFS.

Another vein of research has focused on the use of the SFS to infer parameters of the coalescent, such as the Λ or Ξ measure for exchangeable coalescents or ancestral demographic fluctuations when effective population size is not assumed to be constant through time. Starting with [26], several negative and positive identifiability results have been proven, see [3, 17, 19, 30], that put the theory of ancestral demographic reconstruction on solid statistical footing. This has also led to numerical methods to efficiently compute the SFS under a given coalescent model

and with a given demography [29] and to use them for hypothesis testing [11, 18] or parameter inference [20, 25].

The study of site frequency spectra in branching processes, which is the subject of the present paper, is facilitated by the description of the genealogy of extant populations using *coalescent point processes* (CPP), starting with [27]. CPPs are random genealogies defined using sequences of random variables $(H_i, \ i \geq 1)$ representing the time to the most recent common ancestor (TMRCA) of consecutive individuals:

$$\text{TMRCA}(i, i+1) = H_{i+1}, \quad i \geq 1.$$

Lambert [22] proved that for an independent CPP, under mild moment conditions and assuming uniform mutations along lineages with rate θ, the following holds for fixed $k \geq 1$:

$$\lim_{n \to \infty} \frac{\xi_n^{(k)}}{n} = \theta \int_0^\infty \frac{1}{W(x)^2} \left(1 - \frac{1}{W(x)} \right)^{k-1} dx \quad \text{a.s.,}$$

where $W(x) = 1/\mathbb{P}(H > x)$ is the *scale function* of the random variables underlying the CPP. This result was later extended to more general mutation distributions in [8].

Most recently, Schweinsberg and Shuai [28] have examined site frequency spectra for critical or supercritical birth and death processes, using CPP representations of the genealogy of n sampled individuals at a given time $T_n \to \infty$, due to [23]. In the critical case, they found Kingman-like asymptotics for the total length of branches with exactly k sampled leaves in their descendance, and proved asymptotic normality.

In this work, we will study the SFS associated to a neutral, time-homogeneous mutation process at rate μ in populations modelled by a stationary continuous-state branching process $(Z_t, t \in \mathbb{R})$, for quadratic branching mechanisms given by:

$$\psi(u) = \beta u^2 + 2\beta\theta u,$$

where $\beta > 0$ is a time scaling parameter and $1/\theta > 0$ can be seen as a population size scaling parameter. In such a population, sampling n individuals at time 0, representing the present, we show in Theorem 4.3 that the asymptotic SFS $\xi_k^{(n)}$ satisfies, for $1 \leq k \leq n - 1$:

$$\frac{\beta}{\mu Z_0} \mathbb{E}[\xi_k^{(n)} | Z_0] = \frac{1}{k} + \frac{1}{k} g_1 \left(\theta Z_0, \frac{k}{n} \right) + \frac{\sqrt{k}}{n^2} g_2 \left(\theta Z_0, \frac{k}{n}, n \right),$$

where the function g_1 is explicitly given in (32) and represented in Fig. 3, and where g_2 is uniformly bounded. The function g_1, which can take positive or negative values, represents the distortion of the expected SFS with respect to the classical Kingman-coalescent case (1), for a given value of the present population size Z_0. We expect that, when averaging $\xi_k^{(n)}/Z_0$ over Z_0 (which is distributed as the sum of two independent exponentials with parameter 2θ), this contribution will vanish, leaving only the constant Kingman term $\mu/(\beta k)$. Note that, by analogy with equation (1), Theorem 4.3 gives an expression for the effective population size in a stationary continuous-state branching process:

$$N_e(Z_0) = \frac{Z_0}{4\beta},$$

which simplifies to $N_e = 1/(4\beta\theta)$ when integrating over Z_0. Higher stochasticity in the infinitesimal branching mechanism, reflected by a higher diffusion coefficient β thus leads to lower effective population size.

Our result (Theorem 4.3) relies mostly on a representation theorem of the genealogy of a sample of n individuals using a construction similar to the CPP, obtained in [1]. This construction is similar to the coalescent point processes (CPPs) extensively studied by Lambert since

their introduction in [22]. Most notably, in [8], the authors consider general CPPs associated to Poisson processes and compute asymptotic features of the SFS for these trees. The stationary setting used in the present work breaks the Poissonian structure of the coalescent point process and leads to the more involved construction of [1].

We will also present a version of the SFS defined directly on the continuum random tree representing the genealogy of the whole population, in the spirit of the construction of [8]. We compute the expected intensity of the continuum SFS and give results about the fraction of the population carrying the same alleles as its most recent common ancestor, called the *clonal subpopulation*: if Z_{cl} is the absolute size of the clonal subpopulation, and $R = Z_{cl}/Z_0$ its relative size, we compute moments of these quantities, showing in particular that R and Z_0 are negatively correlated, see Theorem 5.4.

An interesting extension of the present work would be to generalize Theorem 4.3 to branching mechanisms containing an infinite jump measure, such as stable branching mechanisms $\psi(u) = u^\alpha$, $1 < \alpha < 2$. This would require an ancestral construction similar to the CPP, but allowing for multiple branches coalescing at the same time. We expect that the SFS of such populations would not have a Kingman-like form, but possibly exhibit a U-shape, typical of genealogies described by Λ-coalescents [13].

The rest of the paper is organized as follows: in Section 2 below, we will introduce the objects and notations used in the paper, in particular the constructions of [1]. In Section 4, we will give the proof of Theorem 4.3, then, in Section 5, we will present the continuum version of the SFS and compute the density of its expected measure. Finally, in Section 5.2, we prove the results concerning the continuum SFS and the clonal subpopulation.

2. Preliminaries

2.1. **Stationary continuous branching processes.** We consider a critical quadratic branching mechanism $\psi(u) = \beta u^2$ and the associated family $(\psi_\theta, \ \theta > 0)$ of sub-critical branching mechanisms:

$$(4) \qquad\qquad \psi_\theta(u) = \psi(u + \theta) - \psi(\theta) = \beta u^2 + 2\beta\theta u.$$

Let $\theta > 0$ be fixed. We note $\mathbb{P}_x$ the distribution of a continuous-state branching (CB) process $\mathbf{Y} = (Y_t, \ t \geq 0)$ started at $x > 0$, with branching mechanism ψ_θ. The process $\mathbf{Y}$ is the solution of the following Feller diffusion equation where $(B_t, \ t \geq 0)$ is a standard Brownian lmotion:

$$\mathrm{d}Y_t = \sqrt{2\beta Y_t}\,\mathrm{d}B_t - 2\beta\theta Y_t\,\mathrm{d}t \quad \text{and} \quad Y_0 = x.$$

We also consider the associated canonical measure N, which is a σ-finite measure on the space $\mathbb{D}$ of nonnegative continuous functions f such that if $f(s) = 0$ for some $s > 0$, then $f(t) = 0$ for all $t \geq s$. The Laplace transform of the one-dimensional distributions of $\mathbf{Y}$ is given by, for $\lambda \geq 0$ and $t \geq 0$:

$$\mathbb{E}_x[\exp(-\lambda Y_t)] = \exp(-x u(t, \lambda)),$$

where

$$u(t, \lambda) = \mathrm{N}[1 - \exp(-\lambda Y_t)] = \frac{2\theta\lambda}{(2\theta + \lambda)\exp(2\beta\theta t) - \lambda}.$$

In particular, we get $\mathrm{N}[Y_t] = \exp(-2\beta\theta t)$. The tail distribution of the extinction time $\zeta = \inf\{t > 0, \ Y_t = 0\}$ under the canonical measure is given by, for $t \geq 0$:

$$(5) \qquad\qquad c(t) = \mathrm{N}[\zeta > t] = \frac{2\theta}{\exp(2\beta\theta t) - 1}.$$

It is possible to construct a stationary version of this CB process using an immigration process. Let:

$$\mathcal{N}(\mathrm{d}t, \mathrm{d}Y) = \sum_{i \in I} \delta_{(t_i, Y_i)}(\mathrm{d}t, \mathrm{d}Y)$$

be a Poisson point process on $\mathbb{R} \times \mathbb{D}$ with intensity $2\beta \, \mathrm{d}t \, \mathrm{N}[\mathrm{d}Y]$. The stationary CB process $\mathbf{Z} = (Z_t, \, t \in \mathbb{R})$ is then defined as:

$$Z_t = \sum_{t_i \leq t} Y_{t-t_i}^i.$$

This stationary CB process appears also as the limit of the CB process conditioned not to be extinct [21]. It is distributed as the stationary Feller diffusion, which is a solution of the following equation:

$$\mathrm{d}Z_t = \sqrt{2\beta Z_t} \, \mathrm{d}B_t + 2\beta(1 - \theta Z_t) \, \mathrm{d}t, \quad t \in \mathbb{R},$$

and the (stationary) one-dimensional marginal Z_t is distributed as the sum of two independent exponential random variables with parameter 2θ. In particular we have $\mathbb{E}[Z_t] = 1/\theta$.

2.2. Genealogical tree of the ψ_θ CB process.

The genealogy of the CB process $\mathbf{Y}$ (under the canonical measure) can be described as a random tree encoded by a Brownian excursion as follows. For a function $g \in \mathbb{D}$ define a pseudo-distance on $\mathbb{R}_+$ by, for $s, t \in \mathbb{R}_+$:

$$d_g(s, t) = g(s) + g(t) - 2m_g(s, t) \quad \text{with} \quad m_g(s, t) = \inf_{[s \wedge t, s \vee t]} g.$$

The quotient metric space $\mathcal{T}_g = \mathbb{R}_+ / \{d_g = 0\}$, with the metric d_g, is then a *real tree* [12]; the equivalence class of 0, denoted by ϱ_g, is called the root of the tree $\mathcal{T}_g$. The height of $x \in \mathcal{T}_g$ is defined as its distance to the root, that is, as $g(t)$ for any $t \in \mathbb{R}_+$ in the equivalence class x. We say that $s \in \mathbb{R}_+$ is an ancestor of $t \in \mathbb{R}_+$ if $g(s) = m_g(s, t)$; this defines a partial order on $\mathcal{T}_g$, and we write $s \preceq t$ identifying s and t with their equivalence class.

Recall $\theta > 0$ is given. Consider a Brownian motion with negative drift $B^{(\theta)} = (B_t^{(\theta)} = \sqrt{2/\beta} B'_t - 2\theta t, \, t \geq 0)$, where B' is a standard Brownian motion. Let $\mathbb{N}[d\mathcal{T}]$ denote the push-forward measure of the Itô positive excursion measure of the Brownian motion $B^{(\theta)}$ through the application $g \mapsto \mathcal{T}_g$. The σ-finite measure $\mathbb{N}[d\mathcal{T}]$ is defined on the Polish space $\mathbb{T}$ of compact rooted real trees endowed with the so-called Gromov-Hausdorff distance (where pointed compact metric spaces are identified up to an isomorphic transformation). We simply denote the root of $\mathcal{T}$ by ϱ.

We define a local time process of the tree $\mathcal{T}$ denoted by $\mathcal{Y} = (\mathcal{Y}_a, a \geq 0)$ where $\mathcal{Y}_a$ is a random measure on $\mathcal{T}$ which informally is the uniform measure on the elements of $\mathcal{T}$ at distance a from the root. More formally, let $(\ell^a(\mathrm{d}s), a \geq 0)$ be the local time process of $B^{(\theta)}$, with $\ell^a(\mathbb{R}_+)$ the total local time at level a. For any $g \in \mathbb{D}$, let Π_g be the natural projection from $\mathbb{R}_+$ on $\mathcal{T}_g$. Then, for any $a \geq 0$, denote by $\mathcal{Y}_a$ the push-forward measure of ℓ^a on $\mathcal{T}$ through the map $\Pi_{B^{(\theta)}}$. According to [9, Theorem 1.4.1], the total mass process $(\mathcal{Y}_a(1), \, a \geq 0)$ is distributed under $\mathbb{N}$ as the CB process $\mathbf{Y}$ with branching mechanism ψ_θ under the canonical measure N. For this reason, we shall identify Y_a with $\mathcal{Y}_a(1)$ for all $a \geq 0$, and thus see the tree $\mathcal{T}$ as the genealogical tree associated to the CB $\mathbf{Y}$. See also [10] for a direct construction of measures $(\mathcal{Y}_a, a \geq 0)$ from the tree $\mathcal{T}$. The maximal height (distance from the root) of a $\mathcal{T}$ is distributed as the lifetime ζ of the CB process $\mathbf{Y}$, and it will also be denoted by ζ.

We now informally describe the genealogical tree associated to the stationary CB process $\mathbf{Z}$, see [7]. Let $\sum_{i \in I} \delta_{(h_i, \mathcal{T}_i)}(\mathrm{d}h, \mathrm{d}\mathbf{t})$ be a Poisson point measure on $\mathbb{R} \times \mathbb{T}$ with intensity $2\beta \, \mathrm{d}h \mathbb{N}[\mathrm{d}\mathbf{t}]$. The tree $\mathcal{T}^{\mathrm{st}}$ is obtained by grafting the trees $\mathcal{T}_i$ at height h_i along the infinite spine $\mathbb{R}$ (and the root of $\mathcal{T}_i$ is identified with h_i on the infinite spine $\mathbb{R}$). The local time process $(\mathcal{Z}_t, t \in \mathbb{R})$

associated to $\mathcal{T}^{\text{st}}$ is the sum at each level t of the local times at level $t - h_i$ of all trees $\mathcal{T}_i$ with $h_i \leq t$: $\mathcal{Z}_t = \sum_{h_i \leq t} \mathcal{Y}^i_{t-h_i}$, where $\mathcal{Y}^i$ is the local time process of the tree $\mathcal{T}_i$. Then, the total mass process $(\mathcal{Z}_t(1), t \in \mathbb{R})$ is distributed as the stationary CB process $\mathbf{Z}$ with branching mechanism ψ_θ. As above, we shall identify $\mathcal{Z}_t(1)$ with Z_t. Notice that the measure $\mathcal{Z}_t$ puts mass only on the set of leaves of $\mathcal{T}^{\text{st}}$ at level t.

2.3. **Quantities related to the genealogical tree.** The height of $x \in \mathcal{T}^{\text{st}}$, say $H(x)$, is defined as x if x belongs to the infinite spine $\mathbb{R}$ or, if x belongs to the $\mathcal{T}_i$ grafted at height h_i, as its height in $\mathcal{T}_i$ plus h_i. We define a partial order on $\mathcal{T}^{\text{st}}$ by $x \preceq y$ for $x, y \in \mathcal{T}^{\text{st}}$ if either (i) x and y belong to the infinite spine $\mathbb{R}$ and $H(x) \leq H(y)$, or (ii) x belongs to the infinite spine $\mathbb{R}$ and y to the tree $\mathcal{T}_i$ grafted at level h_i with $H(x) \leq h_i$, or (iii) x and y belong to the same tree $\mathcal{T}_i$ and x is an ancestor of y in $\mathcal{T}_i$. For $x \preceq y$ we define $[\![x, y]\!] = \{z \in \mathcal{T}^{\text{st}} : x \preceq z \preceq y\}$ the branch from x to y. It can be isometrically identified with the segment $[H(x), H(y)]$ of $\mathbb{R}$. The length measure $\mathscr{L}(\mathrm{d}x)$ on $\mathcal{T}^{\text{st}}$ is defined through its restriction to $[\![x, y]\!]$ for all $x \preceq y$ as the image of the Lebesgue measure on $[H(x), H(y)]$.

For a set $(x_j, j \in J)$ of elements of $\mathcal{T}^{\text{st}}$, we define the set of its ancestors as $\{x \in \mathcal{T}^{\text{st}} : x \preceq x_j$ for all $j \in J\}$. If this set is not empty, then it has a maximal element (for the partial order $\preceq$) which is called the most recent common ancestor (MRCA) of $(x_j, j \in J)$ and its height is the time to the MRCA (TMRCA). We shall consider the time $-A$ of the MRCA of the extant population at time 0. Denoting by ζ_i for the maximal height of the tree $\mathcal{T}_i$, it is also defined as:
$$A = -\min\{h_i : \zeta_i + h_i \geq 0\}.$$

We also define N_t as the number of ancestors at time $-t$ of the extant population living at time 0 minus 1 (that is, we don't take into account the infinite spine):
$$N_t = \operatorname{Card} \{i \in I : , h_i < -t \quad \text{and} \quad \zeta_i + h_i \geq 0\}.$$

In particular, we have that a.s.:
$$\{A > t\} = \{N_t \geq 1\}. \tag{6}$$

According to [7], we have that N_t is, conditionally on Z_{-t}, distributed as a Poisson random variable with mean $c(t)Z_{-t}$. In particular, we have:
$$\mathbb{E}[N_t] = \frac{c(t)}{\theta}. \tag{7}$$

2.4. **The Kesten tree.** We shall also use the so-called Kesten tree $\mathcal{T}^{\text{Kesten}}$ which is obtained by grafting the trees $\mathcal{T}_i$ at height $h_i > 0$ along the semi-infinite spine $\mathbb{R}_+$ rooted at $\varrho = 0 \in \mathbb{R}_+$. The local time process $(\mathcal{Z}_t^{\text{Kesten}}, t \in \mathbb{R})$ associated to $\mathcal{T}^{\text{Kesten}}$ is then defined as: $\mathcal{Z}_t^{\text{Kesten}} = \sum_{0 < h_i \leq t} \mathcal{Y}^i_{t-h_i}$. Then, the one-dimensional marginal of the total mass process $Z_t^{\text{Kesten}} = \mathcal{Z}_t^{\text{Kesten}}(1)$ is distributed as the size biased distribution of Y_t under the excursion measure, that is, for $t > 0$ and h a measurable non-negative function:
$$\mathbb{E}[h(Z_t^{\text{Kesten}})] = \frac{\mathrm{N}[Y_t h(Y_t)]}{\mathrm{N}[Y_t]} = \mathrm{e}^{2\beta\theta t} \mathrm{N}[Y_t h(Y_t)]. \tag{8}$$

3. Coalescent Point Process of sampled stationary trees

We recall the following construction from [1]. Let $\mathcal{T}^{\text{st}}$ be the genealogical tree associated to the stationary CB process $\mathbf{Z}$ defined in the previous section. We shall consider the genealogical sub-tree $\mathcal{T}_n$ spanned by n individual uniformly chosen among the population at time 0. More precisely, let $(\mathcal{X}_k, k \in \mathbb{N}^*)$ be independent leaves of $\mathcal{T}^{\text{st}}$ at a given level, say 0 for simplicity, chosen uniformly, that is according to the probability measure $\mathcal{Z}_0/Z_0$. For $n \in \mathbb{N}^*$, let $\mathcal{T}_n$ be

the subtree spanned by the leaves $\mathcal{X}_1, \ldots, \mathcal{X}_n$ (that is, the smallest subtree of $\mathcal{T}^{\mathrm{st}}$ containing $\mathcal{X}_1, \ldots, \mathcal{X}_n$) rooted at the MRCA of $\mathcal{X}_1, \ldots, \mathcal{X}_n$. We refer to [1] for a more formal definition. We now give an elementary representation of the tree $\mathcal{T}_n$.

(i) Let $(E_{\mathrm{g}}, E_{\mathrm{d}})$ be independent exponential random variables with parameter 2θ; so that Z_0 is distributed as $E_{\mathrm{g}} + E_{\mathrm{d}}$. For simplicity, we identify Z_0 with $E_{\mathrm{g}} + E_{\mathrm{d}}$. Let also $(U_k, \ k \in \mathbb{N}^*)$ be independent random variables, uniformly distributed on $[0,1]$, independent of $E_{\mathrm{g}}, E_{\mathrm{d}}$. We define the positions $X_0 = 0$ and $X_k = Z_0 U_k - E_{\mathrm{g}}$ for $k \in \mathbb{N}^*$. The position X_0 corresponds to the individual alive at time 0 of the immortal lineage.

(ii) Let $n \in \mathbb{N}^*$ be fixed. We consider the set of "leaves" $\mathcal{L}_n = \{-E_{\mathrm{g}}, E_{\mathrm{d}}, X_0, \ldots, X_{n-1}\}$ and the corresponding order statistics $X_{(0)} = -E_{\mathrm{g}} < X_{(1)} < \ldots < X_{(n)} < X_{(n+1)} = E_{\mathrm{d}}$. For $k \in \{0, \ldots, n+1\}$ we consider the interval $[X_{(k)}, X_{(k+1)}]$ for $X_{(k)} < 0$, $[X_{(k-1)}, X_{(k)}]$ for $X_{(k)} > 0$, and the singleton $\{X_{(k)}\}$ for $X_{(k)} = 0$, and denote by I_k its length. Notice that $\sum_{k=0}^{n+1} I_k = Z_0$.

(iii) Recall the function c defined in (5). For $\delta > 0$, let $\zeta^*(\delta)$ be a random variable on $(0, \infty)$ whose distribution is given by:

$$\mathbb{P}(\zeta^*(\delta) \leq t) = \mathrm{e}^{-\delta c(t)} \quad \text{for} \quad t > 0.$$

In particular, $\zeta^*(\delta)$ is distributed as:

(9)
$$\frac{1}{2\beta\theta} \log\left(1 + \frac{2\theta\delta}{E}\right),$$

where E is an exponential random variable with mean 1. Notice that if $(\zeta_i, \ i \in I)$, with I at most countable, are independent random variables with ζ_i distributed as $\zeta^*(\delta_i)$, then $\sup_{i \in I} \zeta_i$ is distributed as $\zeta^*(\delta)$ with $\delta = \sum_{i \in I} \delta_i$.

Conditionally on $\mathcal{L}_n$, let $(\zeta_k, \ 0 \leq k \leq n+1)$ be independent random variables such that ζ_k is distributed as $\zeta^*(I_k)$, with E in (9) independent of I_k, for $0 \leq k \leq n+1$, and consider the ancestral point measure on $\mathbb{R} \times \mathbb{R}_+$ (notice the sum is from 1 to n):

(10)
$$\mathcal{A}_n = \sum_{k=1}^{n} \delta_{(X_{(k)}, \zeta_k)}.$$

Notice that $(0, 0)$ is an atom of $\mathcal{A}_n$.

Finally, let $\mathfrak{T}_n$ be the ancestral tree associated defined as following: attach the semi-infinite branch $(-\infty, 0]$ at the position $X_0 = 0$ on the segment $[-E_{\mathrm{g}}, E_{\mathrm{d}}]$, and for all $1 \leq k \leq n$, such that $X_{(k)} \neq 0$, attach a branch with length ζ_k at the position $X_{(k)}$ on the segment $[-E_{\mathrm{g}}, E_{\mathrm{d}}]$. Then, identify the bottom of each branch such that $X_{(k)} < 0$ (resp. $X_{(k)} > 0$) with the point with depth ζ_k on the first branch with longer length on the right (resp. on the left). Eventually cut the semi-infinite branch at its last (going downwards) branching point, say ϱ_n, which is at length $\max_{1 \leq k \leq n} \zeta_k$. Then, consider ϱ_n as the root of $\mathfrak{T}_n$. An instance of the ancestral tree is represented in Fig 1.

The next result is a consequence of [1, Lemma 4.1]; notice however that in [1] the ancestral lineage (that is the position of X_0) is given, and that X_0 is not seen as a leaf of the sampled tree. In other words, the approach developed in [1] does not involve the immortal lineage and thus sees the stationary CB process $\mathbf{Z}$ as a CB process with immigration, whereas our approach here takes into account the immortal lineage as X_0 is a leaf of $\mathfrak{T}_n$.

Lemma 3.1 (Representation of the genealogical tree of n individuals). *For $n \in \mathbb{N}^*$, the rooted tree $\mathcal{T}_n$ spanned by $\mathcal{X}_1, \ldots, \mathcal{X}_n$ is distributed as the rooted tree $\mathfrak{T}_n$.*
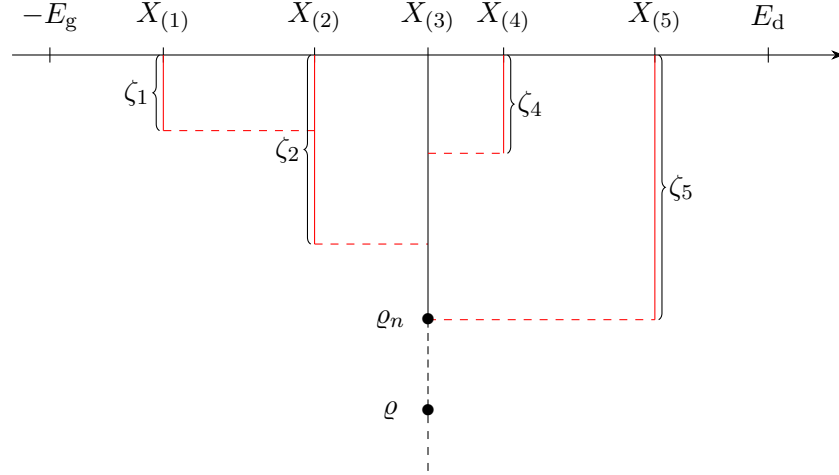
FIGURE 1. An instance for $n = 5$ of the ancestral tree $\mathfrak{T}_n$ with its root ϱ_n, which appears in Lemma 3.1. In this instance, the semi-infinite branch is attached to $X_{(3)} = X_0 = 0$ and cut at the MRCA ϱ_n of the uniformly sampled individuals $\{X_1, \ldots, X_4\}$ in the whole population $[-E_g, E_d]$ and X_0. The branch attached to $X_{(k)}$ has length ζ_k, with $\zeta_3 = 0$ by convention as $X_{(3)} = 0$. The tree $\mathfrak{T}'_n$ which appears in Lemma 5.2 is similar but for the semi-infinite branch which is now cut at the MRCA ϱ of the whole population $[-E_g, E_d]$. (Of course ϱ is an ancestor of ϱ_n and can be equal to ϱ_n.)

According to [7, Proposition 7.3], conditionally on Z_0, the time A of the grand MRCA of the entire population at time 0 is distributed as $\zeta^*(Z_0)$, and thus also distributed as $\max_{0 \leq k \leq n+1} \zeta_k$ conditionally on $\mathcal{L}_n$, that is:

$$(11) \qquad \mathbb{P}(A \leq t \mid Z_0 = z) = \mathbb{P}\left(\max_{0 \leq k \leq n+1} \zeta_k \mid (U_k, \ k \in \mathbb{N}^*), \ E_g + E_d = z\right) = \exp(-c(t)z).$$

(This formula can also be deduced from (6) and the distribution of N_t.)

We end this section with a technical lemma which will be used later on. Let $n \in \mathbb{N}^*$ be fixed. Using the ancestral process $\mathcal{A}_n$ from (10), we define for $j \leq \ell \in [\![1, n]\!]$:

$$(12) \qquad \zeta^\star_{j:\ell} = \zeta_j \vee \cdots \vee \zeta_\ell = \sup_{j \leq i \leq \ell} \zeta_i.$$

We also define $\zeta^{\mathrm{MRCA}}_{j:\ell}$ for the time to the MRCA of $X_{(j)}, \ldots, X_{(\ell)}$. By construction, we have $\zeta^{\mathrm{MRCA}}_{j:\ell} \leq \zeta^\star_{j:\ell}$, see Fig. 2 for various instances (and Fig. 2(D) for an instance of strict inequality). Notice that $\zeta^{\mathrm{MRCA}}_{j:j} = 0$ by construction and recall that $\zeta_j = 0$ if $X_{(j)} = 0$. We have the following precise result.

Lemma 3.2 (Time to the MRCA of consecutive individuals). *Let $n \in \mathbb{N}^*$ be given. Let $1 \leq j < \ell \leq n$. We have:*

$$\zeta^{\mathrm{MRCA}}_{j:\ell} = \begin{cases} \zeta^\star_{j+1:\ell} & \text{if } X_{(j)} \geq 0, \\ \zeta^\star_{j:\ell-1} & \text{if } X_{(\ell)} \leq 0, \\ \zeta^\star_{j:\ell} & \text{if } X_{(j)} X_{(\ell)} \leq 0. \end{cases}$$

Proof. In the first case (see Fig. 2 on the top left for an illustration), we consider that $X_{(j)} = 0$, and thus $\zeta_j = 0$. Then, the branch with length $\zeta^\star_{j:\ell}$ necessarily branches on the ancestral
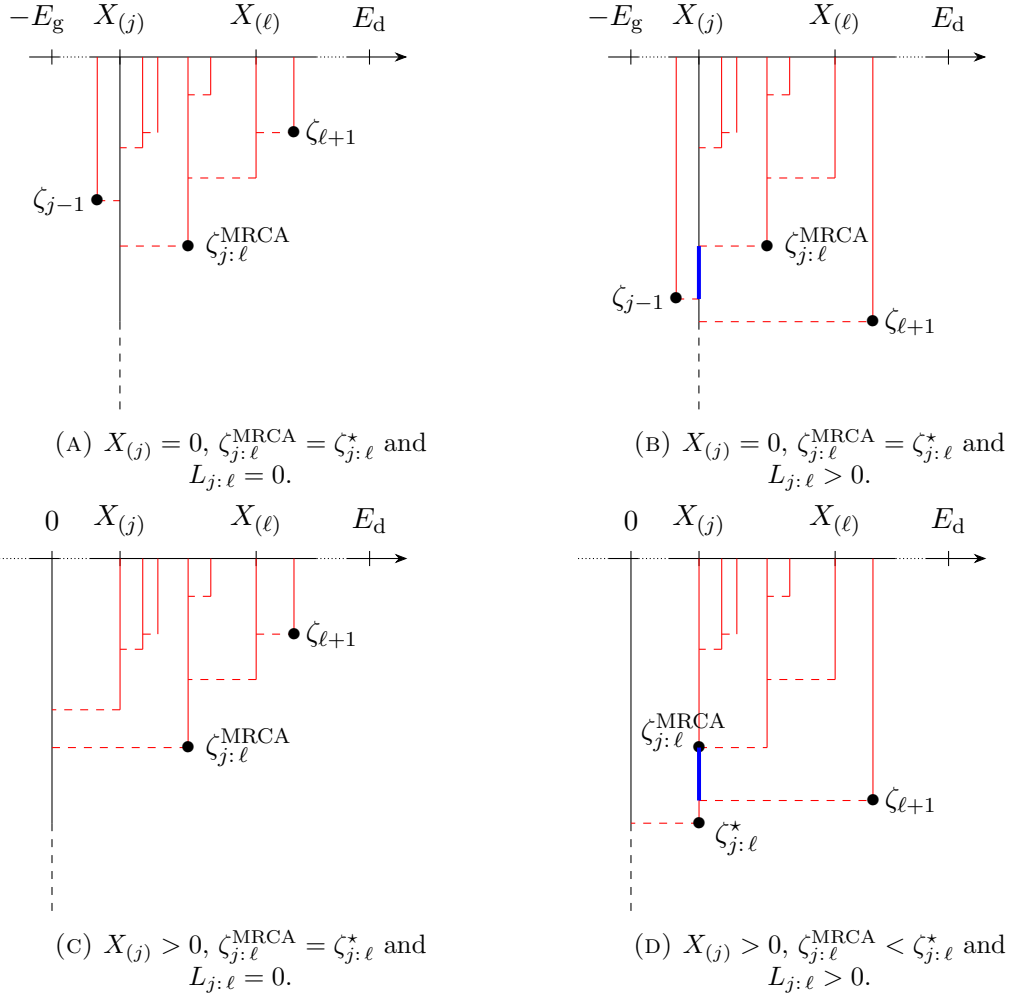
(A) $X_{(j)} = 0$, $\zeta_{j:\ell}^{\mathrm{MRCA}} = \zeta_{j:\ell}^{\star}$ and $L_{j:\ell} = 0$.

(B) $X_{(j)} = 0$, $\zeta_{j:\ell}^{\mathrm{MRCA}} = \zeta_{j:\ell}^{\star}$ and $L_{j:\ell} > 0$.

(C) $X_{(j)} > 0$, $\zeta_{j:\ell}^{\mathrm{MRCA}} = \zeta_{j:\ell}^{\star}$ and $L_{j:\ell} = 0$.

(D) $X_{(j)} > 0$, $\zeta_{j:\ell}^{\mathrm{MRCA}} < \zeta_{j:\ell}^{\star}$ and $L_{j:\ell} > 0$.

FIGURE 2. Four possible configurations of $X_{(j)}, \ldots, X_{(\ell)}$ with their TMRCA $\zeta_{j:\ell}^{\mathrm{MRCA}}$, along with locations (in blue and of length $L_{j:\ell}$) for k-admissible mutations (with $k = \ell - j + 1$) carried only by this set of leaves, whenever these exist. Notice that $\zeta_{j:\ell}^{\mathrm{MRCA}}$ is strictly less than $\zeta_{j:\ell}^{\star}$ only in the bottom left figure.

branch of $X_{(j)}$ (that is, the branch attached to $X_{(j)}$), and the branching point is the MRCA of $X_{(j)}, \ldots, X_{(\ell)}$. Thus the time to the MRCA is $\zeta_{j:\ell}^{\mathrm{MRCA}} = \zeta_{j+1:\ell}^{\star} = \zeta_{(\ell)}^{\star}$, where we used that $\zeta_j = 0$ for the last equality.

In the second case, we consider that $X_{(j)} > 0$ and $\zeta_j = \zeta_{j:\ell}^{\star}$, see an instance in Fig. 2 on the top right. Then, the branch with length $\zeta_{j+1:\ell}^{\star}$ necessarily branches on the ancestral branch of $X_{(j)}$, and the branching point is the MRCA of $X_{(j)}, \ldots, X_{(\ell)}$. This also gives $\zeta_{j:\ell}^{\mathrm{MRCA}} = \zeta_{j+1:\ell}^{\star}$.

In the third case, we consider that $X_{(j)} > 0$ and there exists $i \in [\![2, \ell]\!]$ such that $\zeta_i = \zeta_{j:\ell}^{\star}$, and thus $\zeta_i = \zeta_{j+1:\ell}^{\star}$ (see Fig. 2 bottom left for an illustration of this configuration). Let $i_g = \inf\{i' \in [\![1, j-1]\!] : \zeta_{i'} > \zeta_i \text{ or } X_{(i')} = 0\}$. By definition, the ancestral branch of the leaf $X_{(i)}$ branches onto the ancestral branch of $X_{(i_g)}$ if $X_{(i_g)} > 0$ or onto the spine if $X_{(i_g)} = 0$. In both cases, the branching point is the MRCA of $X_{(j)}, \ldots, X_{(\ell)}$. This also gives $\zeta_{j:\ell}^{\mathrm{MRCA}} = \zeta_i = \zeta_{j+1:\ell}^{\star}$.

Those three cases give a complete picture when $X_{(j)} \geq 0$. The case $X_{(\ell)} \leq 0$ is similar. So we are left with the case $X_{(j)}X_{(\ell)} < 0$, where one of the leaves $X_{(j)}, \ldots, X_{(\ell)}$ belongs to the infinite spine (see Fig. 2 bottom right). In this case, the MRCA of $X_{(j)}, \ldots, X_{(\ell)}$ is on the spine at height $\zeta_{j:\ell}^{\mathrm{MRCA}} = \zeta_{j:\ell}^\star$, $\square$

4. Discrete Frequency Spectrum

The neutral mutations on the stationary population are given by the atoms of a point measure on $\mathcal{T}^{\mathrm{st}}$ with intensity a mutation rate, say $\mu > 0$, times the length measure $\mathscr{L}(\mathrm{d}x)$ on $\mathcal{T}^{\mathrm{st}}$. We sample $n \in \mathbb{N}^*$ individuals from the extant population in a stationary branching process at a given time, say 0 for simplicity. In this section, we will first give some general results for the site frequency spectra of the ancestral tree $\mathcal{T}_n$, defined in Section 3, with n fixed. Thanks to Lemma 3.1, we can recast the problem using a point measure $\mathcal{M} = \sum_{i \in I} \delta_{m_i}$ on the random tree $\mathfrak{T}_n$ (associated with the ancestral point measure $\mathcal{A}_n$) with intensity μ times the length measure on its branches. The associated site frequency spectrum $(\xi_k^{(n)}, \ 1 \leq k \leq n-1)$ is then defined by:

$$\xi_k^{(n)} = \sum_{i \in I} \mathbf{1}_{\{c_n(m_i) = k\}}, \tag{13}$$

where for $x \in \mathfrak{T}_n$, $c_n(x) \in [\![1, n]\!]$ is the number of leaves $X_{(j)}$ among the n sampled leaves such that $x \preceq X_{(j)}$. Note that the only vertex of $\mathfrak{T}_n$ such that $c_n(x) = n$ is the root ϱ, which justifies that we are only considering the SFS up to index $k = n-1$.

We stress that if a mutation is present in exactly $k \in [\![1, n-1]\!]$ leaves of $\mathfrak{T}_n$, then those leaves necessarily have consecutive positions, in the sense that the leaves carrying that mutation are exactly $X_{(j)}, \ldots, X_{(j+k-1)}$ for some $j \in [\![1, n-k+1]\!]$. In order to be carried by exactly k consecutive leaves, a mutation has to be ancestral to their MRCA, but no ancestral to any other leaf. We will call such mutations k-admissible.

Lemma 4.1 (k-admissible mutations). *Let $n \in \mathbb{N}^*$ and $k \in [\![1, n-1]\!]$ be given. Conditionally on the ancestral point measure $\mathcal{A}_n$, the number of k-admissible mutations carried by the k-tuple $X_{(j)}, \ldots, X_{(\ell)}$, for $j \in [\![1, n-k+1]\!]$ and $\ell = j + k - 1$, is Poisson distributed with mean $\mu L_{j:\ell}$, where:*

$$L_{j:\ell} = \begin{cases} [\zeta_j \wedge \zeta_{\ell+1} - \zeta_{j:\ell}^{\mathrm{MRCA}}]_+ & \text{if} \quad X_{(j)} > 0, \\ [\zeta_{j-1} \wedge \zeta_\ell - \zeta_{j:\ell}^{\mathrm{MRCA}}]_+ & \text{if} \quad X_{(\ell)} < 0, \\ [\zeta_{j-1} \wedge \zeta_{\ell+1} - \zeta_{j:\ell}^{\mathrm{MRCA}}]_+ & \text{if} \quad X_{(j)}X_{(\ell)} \leq 0, \end{cases} \tag{14}$$

where in (14) we set $\zeta_0 = \zeta_{n+1} = +\infty$ by convention and $\zeta_{j:\ell}^{\mathrm{MRCA}} = 0$ if $k = 1$ by construction.

Intuitively, the first two cases in equation (14) represent the two symmetric situations in which all of the leaves $X_{(j)}, \ldots, X_{(j+k-1)}$ are on one side of the infinite spine. In those cases, k-admissible mutations are possible only on ancestral branches of the $X_{(j)}, \ldots, X_{(j+k-1)}$, see Fig. 2(D). The third case represents the contribution of the spine, which is nonzero if and only if both ζ_{j-1} and ζ_{j+k} are greater than the longest ancestral branch among the $\zeta_1, \ldots, \zeta_k$ and if $X_{(j-1)}$ and $X_{(j+k)}$ lie on opposite sides of the spine, see Fig. 2(B).

Proof. We first assume that $X_{(j)} > 0$, meaning that all the k consecutive leaves $X_{(j)}, \ldots, X_{(\ell)}$ are on the right side of the spine. The mutations carried only by the k leaves need to lie on the stem of the genealogical tree, say $\mathfrak{T}_{j:\ell}$, of $X_{(j)}, \ldots, X_{(\ell)}$, which is of length $\zeta_{j:\ell}^\star - \zeta_{j:\ell}^{\mathrm{MRCA}}$. By Lemma 3.2, this length is also equal to $[\zeta_j - \zeta_{j+1:\ell}^\star]_+$. We shall now assume it is positive, that is, $\zeta_{j:\ell}^{\mathrm{MRCA}} < \zeta_j$, see Fig. 2(D) for an instance.

If $\zeta_{\ell+1} \leq \zeta_{j:\ell}^{\mathrm{MRCA}}$, all mutations on the stem will also be carried by $X_{(\ell+1)}$ since the ancestral branch of $X_{(\ell+1)}$ will be grafted on $\mathfrak{T}_{j:\ell}$, providing no k-admissible mutations. If $\zeta_{j:\ell}^{\mathrm{MRCA}} < \zeta_{\ell+1} \leq \zeta_j$, then the ancestral branch of $X_{(\ell+1)}$ will be grafted on the stem of $\mathfrak{T}_{j:\ell}$, and only mutations between the root of this sub-tree and that branching point will be k-admissible. If $\zeta_{\ell+1} > \zeta_j$, then the ancestral branch of $X_{(\ell+1)}$ will be grafted below the stem, and all mutations on the stem are then k-admissible.

In conclusion the part of branch carrying the k-admissible mutations is of length $[\zeta_{\ell+1} \wedge \zeta_j - \zeta_{j:\ell}^{\mathrm{MRCA}}]_+$.

The case $X_{(\ell)} < 0$ is similar. So, we now consider the case $X_{(j)}X_{(\ell)} \leq 0$, see Fig. 2(B) for an instance of $L_{j:\ell} > 0$. In particular, there exists $i \in [\![j,\ell]\!]$ (random) such that $X_{(i)} = 0$, and the MRCA of $X_{(j)}, \ldots, X_{(\ell)}$ belongs to the ancestral lineage of $X_{(i)}$, that is the spine. The k-admissible mutations then need to be on the spine below the MRCA but above the MRCA of $X_{(j-1)}, \ldots, X_{(\ell)}$ and the MRCA of $X_{(j)}, \ldots, X_{(\ell+1)}$. Using the convention $\zeta_0 = \zeta_{n+1} = +\infty$, we deduce the part of the branch carrying the k-admissible mutations is of length $[\zeta_{j-1} \wedge \zeta_{\ell+1} - \zeta_{j:\ell}^{\mathrm{MRCA}}]_+$. $\qquad\square$

The number of k-admissible mutations carried by $\mathfrak{T}_n$ is Poisson distributed with mean μL_k with:

$$(15) \qquad L_k = \sum_{j=1}^{n-k+1} L_{j:k}.$$

We have a simple closed formula for the expectation of L_k. Recall (9). Let $U_{(1)} < \cdots < U_{(n)}$ be the order statistics of n independent uniform random variable on $[0,1]$ which are also independent of Z_0 and of an independent exponential random variable E with mean 1. Set $S_0 = 0$ and for $\ell \in [\![1,n]\!]$:

$$S_\ell = \mathbb{E}\left[\frac{1}{2\beta\theta} \log\left(1 + \frac{2\theta Z_0 U_{(\ell)}}{E}\right) \Big| Z_0\right]$$

Lemma 4.2 (Mean of L_k). *Let $n \in \mathbb{N}^*$ and $k \in [\![1, n-1]\!]$ be given. We have:*

$$(16) \qquad \mathbb{E}[L_k \,|\, Z_0] = (n-k)(2S_k - S_{k-1} - S_{k+1}) + S_{k+1} - S_{k-1}.$$

Proof. Since $x \wedge y = x + y - x \vee y$ and $[x - z]_+ = x \vee z - z$, we get that:

$$[x \wedge y - z]_+ = x \vee z + y \vee z - z - x \vee y \vee z \quad \text{for all} \quad x, y, z \in \mathbb{R}.$$

Set $J \in [\![1, n]\!]$ such that $X_{(J)} = 0$. Using Lemmas 3.2 and 4.1, we obtain for $k > 1$ and that:

$\mathbb{E}[L_k \,|\, Z_0]$

$$= \mathbb{E}\Big[\sum_{J < j \leq n-k+1} \big(\zeta_{j:j+k-1}^\star + \zeta_{j+1:j+k}^\star \mathbf{1}_{\{j+k \leq n\}} - \zeta_{j+1:j+k-1}^\star - \zeta_{j:j+k}^\star \mathbf{1}_{\{j+k \leq n\}}\big) \,\big|\, Z_0\Big]$$

$$+ \mathbb{E}\Big[\sum_{1 \leq j < J-k+1} \big(\zeta_{j-1:j+k-2}^\star \mathbf{1}_{\{j \geq 2\}} + \zeta_{j:j+k-1}^\star - \zeta_{j:j+k-2}^\star - \zeta_{j-1:j+k-1}^\star \mathbf{1}_{\{j \geq 2\}}\big) \,\big|\, Z_0\Big]$$

$$+ \mathbb{E}\Big[\sum_{1 \vee (J-k+1) \leq j \leq J \wedge (n-k+1)} \big(\zeta_{j-1:j+k-1}^\star \mathbf{1}_{\{j \geq 2\}} + \zeta_{j:j+k}^\star \mathbf{1}_{\{j+k \leq n\}} - \zeta_{j:j+k-1}^\star - \zeta_{j-1:j+k}^\star \mathbf{1}_{\{j \geq 2, j+k \leq n\}}\big) \,\big|\, Z_0\Big].$$

Let $U_{(1)} < \cdots < U_{(n)}$ be the order statistics of n independent uniform random variable on $[0,1]$ which are also independent of Z_0. We simply denote by W_ℓ the random variable given by (9) with δ replaced by $Z_0 U_{(\ell)}$ and E independent of $Z_0, U_1, \ldots, U_n$. In particular, conditionally on

J and Z_0, we have that $\zeta_{j:\ell}^\star$ is distributed as $W_{\ell-j+1}$ if $J < j \leq \ell \leq n$ or $1 \leq j \leq \ell < J$ but simply as $W_{\ell-j}$ if $1 \leq j \leq J \leq \ell \leq n$ and $j < \ell$ as $\zeta_J = 0$. We thus deduce that:

$$\mathbb{E}[L_k \,|\, Z_0] = \mathbb{E}\Big[\sum_{J<j\leq n-k+1} \big(W_k + W_k \mathbf{1}_{\{j+k\leq n\}} - W_{k-1} - W_{k+1}\mathbf{1}_{\{j+k\leq n\}}\big) \,|\, Z_0 \Big]$$

$$+ \mathbb{E}\Big[\sum_{1\leq j<J-k+1} \big(W_k + W_k \mathbf{1}_{\{j\geq 2\}} - W_{k-1} - W_{k+1}\mathbf{1}_{\{j\geq 2\}}\big) \,|\, Z_0 \Big]$$

$$+ \mathbb{E}\Big[\sum_{1\vee(J-k+1)\leq j\leq J\wedge(n-k+1)} \big(W_k \mathbf{1}_{\{j\geq 2\}} + W_k \mathbf{1}_{\{j+k\leq n\}} - W_{k-1} - W_{k+1}\mathbf{1}_{\{j\geq 2, \, j+k\leq n\}}\big) \,|\, Z_0 \Big].$$

By definition, we have $S_\ell = \mathbb{E}[W_\ell \,|\, Z_0]$ for $\ell \in [\![1,n]\!]$. We get:

$$\mathbb{E}[L_k \,|\, Z_0] = 2(n-k)S_k - (n-k+1)S_{k-1} - (n-k-1)S_{k+1}$$

$$= (n-k)(2S_k - S_{k-1} - S_{k+1}) + S_{k+1} - S_{k-1}.$$

It is easy to check that this formula also holds for $k = 1$ as $S_0 = 0$. $\qquad\square$

We now compute the SFS of the ancestral tree $\mathcal{T}_n$ of $n \in \mathbb{N}^*$ individuals sampled from the extant population in a stationary branching process at a given time, say 0 for simplicity.

Theorem 4.3 (Site frequency spectra of n individuals at a given generation). *The expected number of mutations carried by exactly $k \in [\![1, n-1]\!]$ individuals among $n \geq 2$ individuals sampled uniformly in the population at a fixed time for a stationary subcritical branching process satisfies:*

$$\frac{\beta}{\mu Z_0} \,\mathbb{E}[\xi_k^{(n)} \,|\, Z_0] = \frac{1}{k} + \frac{1}{k}\, g_1\left(\theta Z_0, \frac{k}{n}\right) + \frac{\sqrt{k}}{n^2}\, g_2\left(\theta Z_0, \frac{k}{n}, n\right),$$

where the function g_1 is continuous on $\mathbb{R}_+^* \times [0,1]$ with $g_1(z,0) = 0$ and for all z, there exists a constant C such that $g_1(z,u) \leq Cu(|\log(u)| + 1)$ and $g_2(z,u,n) \leq C$ for all $u \in [0,1]$ and $n \geq 2$. In particular, if $(k_n, n \in \mathbb{N}^*)$ is a sequence such that $\lim_{n\to\infty} k_n/n = u \in [0,1]$ and $k_n \in [\![1, n-1]\!]$, then we have:

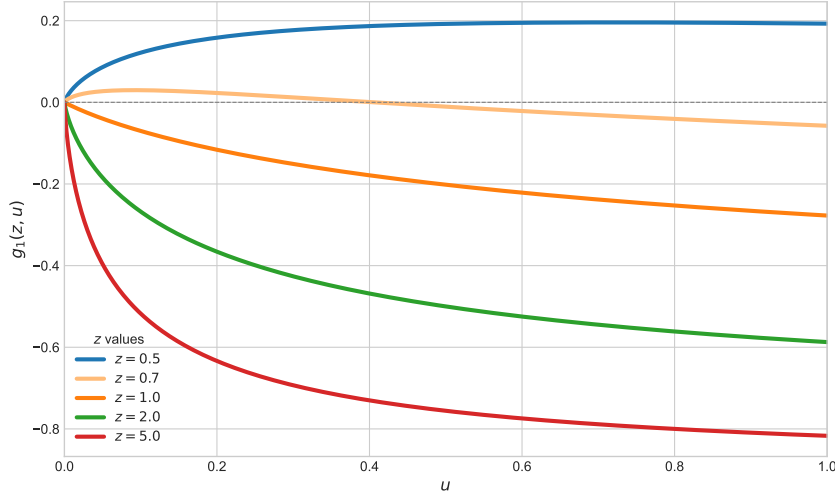$$(17) \qquad\qquad \lim_{n\to\infty} k_n\, \mathbb{E}[\xi_{k_n}^{(n)} \,|\, Z_0] = \frac{\mu Z_0}{\beta}\left(1 + g_1\left(\theta Z_0, u\right)\right).$$

The function g_1 is explicitly given in (32) and drawn in Fig. 3 for various values of z. Note that $g_1(z,u) = 2(z-1)u\log(u) + O(u)$ for small u, hence g_1 is not differentiable at $u = 0$ except for the singular value $z = 1$, which corresponds to the case in which Z_0 is equal to its mean $\mathbb{E}[Z_0] = 1/\theta$.

Proof. Thanks to Lemma 3.1, we recast the problem using a point measure $\mathcal{M}$ on the random tree $\mathfrak{T}_n$ (associated with the ancestral point measure $\mathcal{A}_n$) with intensity μ times the length measure on its branches. Let $J \in [\![1,n]\!]$ such that $X_{(J)} = 0$. We also recall the definition (12) of $\zeta_{j:\ell}^\star$ and that $\zeta_J = 0$.

Let $k \in [\![1, n-1]\!]$. The number of k-admissible mutations carried by $\mathfrak{T}_n$ is Poisson distributed with mean μL_k given in (15). Recall that S_k is distributed as $\mathbb{E}[\zeta^*(Z_0 U_{(k)})]$ with $U_{(k)}$ the k-th order statistics of n independent random variables $U_1, \dots, U_n$ uniformly distributed over $[0,1]$ also independent of Z_0, and $(Z_0, U_{(k)})$ is independent of E in (9). We also recall formula (5.15) from [1]:

$$\mathbb{E}\left[\zeta^*(\delta)\right] = \delta\, H(2\theta\delta) \quad \text{with} \quad H(x) = \int_0^\infty \frac{1 - e^{-u}}{u}\, \frac{du}{u+x}.$$

FIGURE 3. Plot of $g_1(z, u)$ for various values of z.

Let γ be the Euler constant. Using that:

$$1 - \gamma = \int_0^1 \left(1 - \mathrm{e}^{-u} - u\right) \frac{\mathrm{d}u}{u^2} + \int_1^\infty \left(1 - \mathrm{e}^{-u}\right) \frac{\mathrm{d}u}{u^2},$$

and elementary computations, we get that:

$$
\begin{aligned}
H(x) &= \int_0^1 \left(1 - \mathrm{e}^{-u} - u + \frac{u^2}{2} - \frac{u^3}{6}\right)\left(\frac{1}{u+x} - \frac{1}{u}\right)\frac{\mathrm{d}u}{u} + \int_1^\infty (1 - \mathrm{e}^{-u})\left(\frac{1}{u+x} - \frac{1}{u}\right)\frac{\mathrm{d}u}{u} \\
&\quad + \int_0^1 \left(1 - \frac{u}{2} + \frac{u^2}{6}\right)\frac{\mathrm{d}u}{u+x} + \int_0^1 \left(\frac{1}{2} - \frac{u}{6}\right)\mathrm{d}u + 1 - \gamma \\
&= h_0(x) - \left(1 + \frac{x}{2} + \frac{x^2}{6}\right)\log(x) - \frac{x}{6} + 1 - \gamma,
\end{aligned}
$$

where:

$$(18) \qquad h_0(x) = \left(1 + \frac{x}{2} + \frac{x^2}{6}\right)\log(1 + x) - \int_{(0,\infty)} f(u)\,\frac{x\,\mathrm{d}u}{u+x},$$

and:

$$(19) \qquad f(u) = \frac{1}{u^2}\left(1 - \mathrm{e}^{-u} - u + \frac{u^2}{2} - \frac{u^3}{6}\right)\mathbf{1}_{\{u \le 1\}} + \frac{1}{u^2}\left(1 - \mathrm{e}^{-u}\right)\mathbf{1}_{\{u > 1\}}.$$

This decomposition is motivated by the fact that $f(u)/u^2$ is integrable. So we get:

$$\beta \mathbb{E}\left[\zeta^*(\delta)\right] = -\delta\left(1 + \theta\delta + \frac{2}{3}\theta^2\delta^2\right)\log(2\theta\delta) + (1 - \gamma)\delta - \frac{\theta}{3}\delta^2 + \delta h_0(2\theta\delta).$$

Let $\log_+(x) = \max(0, \log(x))$. For simplicity, we set:

$$(20) \qquad\qquad\qquad\qquad\qquad h_1(x) = x h_0(x),$$

and get the following bounds on the derivatives of h_1: there exists a finite constant C such that for $x \ge 0$:

$$(21) \qquad\qquad |h_1^{(i)}(x)| \le C(1 + x^{3-i}\log_+(x)) \quad \text{and} \quad i \in \{0, \ldots, 3\}.$$

Now, in the computation of S_k, the random variable $U_{(k)}$, which is independent of Z_0, has a Beta distribution with parameter $(k, n-k+1)$. We recall that if V has a Beta distribution with parameter (a, b) and $i \in \mathbb{N}$:

$$\mathbb{E}[V] = \frac{a}{a+b},$$

$$\mathbb{E}[V^2] = \frac{a(a+1)}{(a+b)(a+b+1)},$$

$$\mathbb{E}[V^{i+1} \log(V)] = \frac{a \cdots (a+i)}{(a+b) \cdots (a+b+i)} \big(\Psi(a+i+1) - \Psi(a+b+i+1)\big),$$

with $\Psi(x) = \Gamma'(x)/\Gamma(x)$ the digamma function. We shall use that for $x > 0$:

$$(22) \qquad \log(x) - \frac{1}{x} \leq \Psi(x) \leq \log(x) - \frac{1}{2x} \quad \text{and} \quad \Psi(x+1) = \Psi(x) + \frac{1}{x}.$$

Let us mention that $\Psi(\ell+1) = H_\ell - \gamma$, with $H_0 = 0$ and $H_\ell = \sum_{i=1}^{\ell} i^{-1}$ the harmonic sum for $\ell \in \mathbb{N}^*$ and γ the Euler constant.

We set:

$$B_{n,k} = \Psi(n+2) - \Psi(k+1) + 1 - \gamma - \log(2\theta Z_0),$$

$$C_{n,k} = \Psi(n+3) - \Psi(k+2) - \frac{1}{3} - \log(2\theta Z_0),$$

$$D_{n,k} = \Psi(n+4) - \Psi(k+3) - \log(2\theta Z_0)$$

and

$$A_{n,k} = \frac{k}{(n+1)} B_{n,k} + \theta Z_0 \frac{k(k+1)}{(n+1)(n+2)} C_{n,k} + \frac{2\theta^2 Z_0^2}{3} \frac{k(k+1)(k+2)}{(n+1)(n+2)(n+3)} D_{n,k},$$

as well as:

$$F_{n,k} = \mathbb{E}\left[U_{(k)} h_0(2\theta Z_0 U_{(k)}) \,|\, Z_0\right] = \frac{1}{2\theta Z_0} \mathbb{E}\left[h_1(2\theta Z_0 U_{(k)}) \,|\, Z_0\right].$$

In particular, we have:

$$(23) \qquad \frac{\beta}{Z_0} S_k = A_{n,k} + F_{n,k}.$$

By convention we set $U_{(0)} = 0$ so that the formula (23) also holds for $k = 0$ as by convention $S_0 = 0$. Recall $k \in [\![1, n-1]\!]$. Using (22), we also get:

$$\frac{\beta}{Z_0} S_{k-1} = A_{n,k} + F_{n,k-1} - \frac{1}{(n+1)}\left(B_{n,k} - 1 + \frac{1}{k}\right)$$
$$- \theta Z_0 \frac{k}{(n+1)(n+2)}\left(2C_{n,k} - \frac{k-1}{k+1}\right)$$
$$- \frac{2\theta^2 Z_0^2}{3} \frac{k(k+1)}{(n+1)(n+2)(n+3)}\left(3D_{n,k} - \frac{k-1}{k+2}\right),$$

$$\frac{\beta}{Z_0} S_{k+1} = A_{n,k} + F_{n,k+1} + \frac{1}{(n+1)}(B_{n,k} - 1)$$
$$+ \theta Z_0 \frac{k+1}{(n+1)(n+2)}(2C_{n,k} - 1)$$
$$+ \frac{2\theta^2 Z_0^2}{3} \frac{(k+1)(k+2)}{(n+1)(n+2)(n+3)}(3D_{n,k} - 1).$$

So for $k \in [\![1, n-1]\!]$, we have:

$$(24) \quad \frac{\beta}{Z_0} \mathbb{E}[L_k \mid Z_0] = \frac{1}{k} + \frac{B_{n,k}^{(1)}}{(n+1)} + \theta Z_0 \frac{C_{n,k}^{(1)}}{(n+1)(n+2)}$$
$$+ \frac{2\theta^2 Z_0^2}{3} \frac{(k+1)D_{n,k}^{(1)}}{(n+1)(n+2)(n+3)} + \left(R_{n,k}^{(1)} + (n-k)R_{n,k}^{(2)} \right).$$

with:

$$B_{n,k}^{(1)} = 2B_{n,k} - 3,$$
$$C_{n,k}^{(1)} = 2(-n + 3k + 1)C_{n,k} + \frac{n(3k+1) - (5k^2 + 2k + 1)}{k+1},$$
$$D_{n,k}^{(1)} = 6(-n + 2k + 1)D_{n,k} + \frac{1}{k+2}((n-k)(5k+4) - (2k^2 + 3k + 4)),$$
$$R_{n,k}^{(1)} = F_{n,k+1} - F_{n,k-1},$$
$$R_{n,k}^{(2)} = \left(2F_{n,k} - F_{n,k-1} - F_{n,k+1} \right).$$

So we get with $u = k/n \in (0, 1)$:

$$\frac{B_{n,k}^{(1)}}{(n+1)} = \frac{u}{k} \left(-2\log(u) - 1 - 2\gamma - 2\log(2\theta Z_0) \right) + O\left(\frac{u}{k^2} \right),$$
$$\frac{C_{n,k}^{(1)}}{(n+1)(n+2)} = \frac{u}{k} \left(2(1 - 3u)\left(\log(u) + \log(2\theta Z_0) \right) + \frac{11}{3} - 7u \right) + O\left(\frac{u}{k^2} \right),$$
$$\frac{(k+1)D_{n,k}^{(1)}}{(n+1)(n+2)(n+3)} = \frac{u^2}{k} \left(6(1 - 2u)\left(\log(u) + \log(2\theta Z_0) \right) + 5 - 7u \right) + O\left(\frac{u}{k^2} \right),$$

where $O\left(u/k^2 \right)$ has to be understood as a function of θ, Z_0, n, and k which is bounded by C/nk, with C a constant depending only on θ and Z_0. We first consider the term $R_{n,k}^{(1)}$:

$$2\theta Z_0 R_{n,k}^{(1)} = \mathbb{E}\left[h_1(2\theta Z_0 (\Delta + U_{(k-1)})) - h_1(2\theta Z_0 (U_{(k-1)})) \mid Z_0 \right],$$

where $\Delta = U_{(k+1)} - U_{(k-1)}$ is distributed as $U_{(2)}$. Recall $u = k/n$, and notice that:

$$(25) \qquad \mathbb{E}\left[(U_{(k-1)} - u)^2 \right] = \frac{u(1-u)}{n} + O(n^{-2}).$$

Notice that (25) holds indeed for $k = 1$ as by convention $U_{(0)} = 0$ and the left hand-side of (25) is equal to $O(1/n^2)$. Since:

$$h_1(\delta + x) - h_1(x) = \delta h_1'(x) + \int_0^\delta (\delta - t)h_1''(t + x)\, dt,$$

and, thanks to (21) for the control of the second derivative of h_1:

$$|h_1'(2\theta Z_0 U_{(k-1)}) - h_1'(2\theta Z_0 u)| \le C(1 + \theta Z_0)^3 |U_{(k-1)} - u|$$

we deduce, using Cauchy-Schwartz inequality and (25), that:

$$(26) \qquad R_{n,k}^{(1)} = \mathbb{E}[\Delta]h_1'(2\theta Z_0\, u) + \theta Z_0\, \mathbb{E}[\Delta^2]^{1/2}\, \mathbb{E}\left[(U_{(k-1)} - u)^2 \right]^{1/2} O(1) + \theta Z_0\, O\left(\mathbb{E}[\Delta^2] \right)$$
$$= \frac{2}{k}\, u h_1'(2\theta Z_0\, u) + O\left(\frac{u^2}{k^{3/2}} \right).$$

We now control the term $R_{n,k}^{(2)}$. We have:

$$2h_1(x + \delta) - h_1(x) - h_1(x + \delta + \delta') = (\delta - \delta')h_1'(x) + H(x, \delta, \delta'), \tag{27}$$

with:

$$(28) \quad H(x, \delta, \delta')$$

$$= 2 \int_0^\delta (\delta - t) h_1''(x + t) \, dt - \int_0^{\delta + \delta'} (\delta + \delta' - t) h_1''(t) \, dt$$

$$= \left(\delta^2 - \frac{(\delta + \delta')^2}{2} \right) h_1''(x) + \int_0^\delta (\delta - t)^2 h_1'''(t + x) \, dt - \frac{1}{2} \int_0^{\delta + \delta'} (\delta - t)^2 h_1'''(t + x) \, dt.$$

Take $X = 2\theta Z_0 U_{(k-1)}$, $\delta = 2\theta Z_0 \Delta$ and $\delta' = 2\theta Z_0 \Delta'$ with $\Delta = U_{(k)} - U_{(k-1)}$ and $\Delta' = U_{(k+1)} - U_{(k)}$. Notice that Δ and Δ' are distributed as $U_{(1)}$ and that $(\Delta, U_{(k-1)})$ and $(\Delta', U_{(k-1)})$ have the same distribution. This implies that:

$$\mathbb{E}[(\Delta - \Delta')h_1'(X) \mid Z_0] = 0,$$

and thus:

$$2\theta Z_0 \, R_{n,k}^{(2)} = \mathbb{E}[2h_1(X + \delta) - h_1(X) - h_1(X + \delta + \delta')] = \mathbb{E}[H(X, \delta, \delta') \mid Z_0]. \tag{29}$$

We also have, thanks to (21):

$$|h_1''(2\theta Z_0 \, U_{(k-1)}) - h_1''(2\theta Z_0 u)| \leq C(1 + \theta Z_0)^2 |U_{(k-1)} - u|.$$

Using (29), (25) and that $\Delta + \Delta'$ is distributed as $U_{(2)}$, we obtain similarly that:

$$R_{n,k}^{(2)} = 2\theta Z_0 \left(\mathbb{E}[\Delta^2] - \frac{\mathbb{E}[(\Delta + \Delta')^2]}{2} \right) h_1''(2\theta Z_0 \, u) + \frac{1}{n} O\left(\frac{u^2}{k^{3/2}} \right)$$

$$= -2\theta Z_0 \frac{u}{nk} h_1''(2\theta Z_0 \, u) + \frac{1}{n} O\left(\frac{u^2}{k^{3/2}} \right).$$

We thus obtain that:

$$R_{n,k}^{(1)} + (n - k)R_{n,k}^{(2)} = \frac{2}{k} uh_1'(2\theta Z_0 \, u) - \frac{2\theta Z_0}{k} u(1 - u) h_1''(2\theta Z_0 \, u) + O\left(\frac{u^2}{k^{3/2}} \right). \tag{30}$$

In conclusion, we get that for $k \in [\![1, n-1]\!]$:

$$\frac{\beta k}{Z_0} \mathbb{E}[L_k \mid Z_0] = 1 + g_1(\theta Z_0, u) + O\left(\frac{u^2}{k^{1/2}} \right) = 1 + g_1(\theta Z_0, u) + \frac{k^{3/2}}{n^2} O(1), \tag{31}$$

with g_1 given for $u \in [0, 1]$ and $z > 0$ by:

$$g_1(z, u) = u(-2\log(u) - 1 - 2\gamma - 2\log(2z)) \tag{32}$$

$$+ zu \left(2(1 - 3u)(\log(u) + \log(2z)) + \frac{11}{3} - 7u \right)$$

$$+ \frac{2}{3} z^2 u^2 (6(1 - 2u)(\log(u) + \log(2z)) + 5 - 7u)$$

$$+ 2u h_1'(2zu) - 2zu(1 - u) h_1''(2zu),$$

where h_1 is defined in (20) through h_0 from (18) and f from (19). Thanks to (21), we get that g is continuous on $\mathbb{R}_+^* \times [0, 1]$, that $g_1(z, 0) = 0$ and that for all $z > 0$, there exists a constant C such that $g_1(z, u) \leq Cu(|\log(u)| + 1)$. Set $g_2(z, u, n)$ as $n^2/\sqrt{k}$ the very last right hand side term of (31) so that $g_2(z, u, n) = O(1)$. Then, use Lemma 4.1, to get $\mathbb{E}[\xi_k^{(n)} \mid Z_0] = \mu \mathbb{E}[L_k \mid Z_0]$.

Note that for $k = n-1$, formula (14) reduces to $\mathbb{E}[L_{n-1}|Z_0] = 2(S_{n-1} - S_{n-2})$. The derivations above are still valid in that case, except for $R^{(1)}_{n,n-1} = 2(F_{n,n-1} - F_{n,n-2})$ and $R^{(2)}_{n,n-1} = 0$. The same computations (with $\Delta = U_{(n-1)} - U_{(n-2)}$ distributed as $U_{(1)}$) give the same asymptotic (26) for $R^{(1)}_{n,n-1}$. Since for $u = 1$, the second derivative term $u(1-u)h''_1(2\theta Z_0 u)$ vanishes in (30), this enables us to recover the asymptotic (17). $\qquad\square$

5. Continuous Frequency Spectrum

In this section, we will consider the continuous frequency spectrum of the genealogical tree $\mathcal{T}^{\mathrm{st}}$ associated to the stationary CB process $\mathbf{Z}$. We consider a (neutral) mutation process given by a Poisson point process on $\mathcal{M} = \sum_{i \in I} \delta_{m_i}$ on $\mathcal{T}^{\mathrm{st}}$ with intensity $\mu\mathscr{L}(\mathrm{d}x)$, where $\mu > 0$ is the individual mutation rate and $\mathscr{L}(\mathrm{d}x)$ the length measure on $\mathcal{T}^{\mathrm{st}}$.

The total offspring subtree of $x \in \mathcal{T}^{\mathrm{st}}$ is defined by $\mathcal{T}^{\mathrm{st}}(x) = \{y \in \mathcal{T}^{\mathrm{st}} : x \preceq y\}$ and the corresponding clonal sub-tree is defined by:

$$(33) \qquad \mathcal{T}^{\mathrm{st}}_{\mathrm{clonal}}(x) = \{y \in \mathcal{T}^{\mathrm{st}}(x) : \; \mathcal{M}(\llbracket x, y \rrbracket) = 0\}.$$

In the following sections we shall study the mean measure of the size of the population at time 0 carrying a mutation and the size of the clonal population at time 0 of the MRCA of the extant population at time 0.

5.1. **The mean site frequency measure.** Following [8], we consider the site frequency point measures on $(0, +\infty)$ of the extant population at time 0:

$$(34) \qquad \Phi = \sum_{i \in I} \delta_{\mathcal{Z}_0(\mathcal{T}^{\mathrm{st}}(m_i))}.$$

In other words, we associate to each mutation on the tree the size of the population at time 0 carrying it. The main result of this section describes the mean measures Λ of this point measure:

$$(35) \qquad \Lambda(\mathrm{d}r) = \mathbb{E}[\Phi(\mathrm{d}r)].$$

Let $\Gamma(0, r) = \int_r^\infty v^{-1}\, \mathrm{e}^{-v}\, \mathrm{d}v$ denote the incomplete Gamma function.

Theorem 5.1 (The mean SFS measure). *The mean site frequency measure Λ of the genealogical tree $\mathcal{T}^{\mathrm{st}}$ (associated to the stationary CB process $\mathbf{Z}$) is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}_+$, with density given by:*

$$(36) \qquad f(r) = \frac{\mu}{\beta}\left(\frac{\mathrm{e}^{-2\theta r}}{\theta r} + \mathrm{e}^{-2\theta r} + 2\theta r\, \Gamma(0, 2\theta r)\right).$$

It is worth noticing that

$$f(r) \sim_{r \to 0+} \frac{\mu}{\beta\theta r} \quad \text{and} \quad f(r) \sim_{r \to +\infty} \frac{2\mu}{\beta}\, \mathrm{e}^{-2\theta r}.$$

In other terms, for small r, that is, for mutations shared by a small fraction of the extant population at time 0, the only significant contribution comes from the mutations not located on the spine. By contrast, for large r, corresponding to mutations shared by a large number of the extant population, only spine mutations are significant.

The discrete equivalent of the site frequency point measure is the following measure, defined on $(0, 1)$, with $\xi^{(n)}_k$ as in (13):

$$\Phi^{(n)}_{\mathrm{d}} = \sum_{1 \le k \le n-1} \xi^{(n)}_k \delta_{\frac{k}{n}}.$$

As n goes to ∞, conditionally on $\mathcal{T}^{\mathrm{st}}$, this measure converges a.s. to the normalized site frequency point measure:

$$\Phi_{\mathrm{d}}^{(\infty)} = \sum_{i \in I} \delta_{\mathcal{Z}_0(\mathcal{T}^{\mathrm{st}}(m_i))/\mathcal{Z}_0}.$$

Unfortunately, due to the lack of a branching structure for the normalized process $\mathcal{Z}_t/\mathcal{Z}_0$, it is not straightforward to obtain an expression for the mean measure of $\Phi_{\mathrm{d}}^{(\infty)}$ as in Theorem 5.1.

Proof. Recall $-A$ denotes the TMRCA and N_t the number of the at time $-t$ of the extant population living at time 0. Recall the construction of the genealogical tree $\mathcal{T}^{\mathrm{st}}$ from Section 2.2. We shall identify $s \in \mathbb{R}$ with the element on the infinite spine $\mathbb{R}$ of $\mathcal{T}^{\mathrm{st}}$ at height s.

Using the branching property, we get for h a non-negative measurable function defined on $[0, +\infty)$ with $h(0) = 0$:

$$\begin{aligned}
\Lambda(h) &= \mathbb{E}\left[\sum_{i \in I} h(\mathcal{Z}_0(\mathcal{T}^{\mathrm{st}}(m_i)))\right] \\
&= \mathbb{E}\left[\int_{\mathcal{T}^{\mathrm{st}}} \mathcal{M}(\mathrm{d}x)\, h(\mathcal{Z}_0(\mathcal{T}^{\mathrm{st}}(x)))\right] \\
&= \mu\mathbb{E}\left[\int_0^\infty \mathrm{d}t\, N_t\, \mathrm{N}[h(Y_t) \,|\, \zeta > t]\right] + \mu\mathbb{E}\left[\int_0^A \mathrm{d}t\, \mathbb{E}[h(\mathcal{Z}_0(\mathcal{T}^{\mathrm{st}}(-t))) \,|\, A > t]\right].
\end{aligned}$$

In this formula, the first term represents the contributions at time 0 of the N_t individuals at time t before the present that are ancestral to the population at time 0, whereas the second term is the contribution of the infinite spine, that is, the descendants at time 0 of populations immigrating between time $-t$ and 0.

The density distribution q_t of Y_t under the excursion measure N, see [24, p. 63], is given by:

$$\mathrm{N}[\mathrm{d}Y_t = r] = q_t(r)\,\mathrm{d}r = \frac{4\theta^2\, \mathrm{e}^{-2\beta\theta t}}{(1 - \mathrm{e}^{-2\beta\theta t})^2} \exp\left(-\frac{2\theta r}{1 - \mathrm{e}^{-2\beta\theta t}}\right)\,\mathrm{d}r.$$

Using also the expectation of N_t in (7) and $\mathrm{N}[\zeta > t] = c(t)$, we obtain for the first term that:

$$\begin{aligned}
\mathbb{E}\left[\int_0^\infty \mathrm{d}t\, N_t\, \mathrm{N}[h(Y_t) \,|\, \zeta > t]\right] &= \frac{1}{\theta} \int_{(0,\infty)} h(r)\,\mathrm{d}r \int_0^\infty \mathrm{d}t\, q_t(r) \\
&= \int_{(0,\infty)} h(r)\, \frac{\mathrm{e}^{-2\theta r}}{\beta\theta r}\,\mathrm{d}r.
\end{aligned}$$

For the second term, we notice that $\mathcal{T}^{\mathrm{st}}(-t)$ is distributed as the Kesten tree (rooted at $-t$) defined in Section 2.2. Using (8), we get that:

$$\mathbb{E}[h(\mathcal{Z}_0(\mathcal{T}^{\mathrm{st}}(-t)))] = \mathbb{E}[h(Z_t^{\mathrm{Kesten}})] = \mathrm{e}^{2\beta\theta t}\, \mathrm{N}[Y_t\, h(Y_t)].$$

Using also (11) and that Z_0 is distributed as the sum of two independent exponential random variable with parameter 2θ, we get:

$$
\begin{aligned}
\mathbb{E}\left[\int_0^A \mathrm{d}t\, \mathbb{E}[h(\mathcal{Z}_0(\mathcal{T}^{\mathrm{st}}(-t)))\,|\,A>t]\right] &= \int_0^\infty \mathrm{d}t\, \mathbb{P}(A>t)\, \mathrm{e}^{2\beta\theta t}\, \mathrm{N}[Y_t\, h(Y_t)] \\
&= \int_0^\infty \mathrm{d}t \int_{(0,\infty)} \mathrm{d}r\,(1-(1-\mathrm{e}^{-2\beta\theta t})^2)\,\mathrm{e}^{2\beta\theta t}\, r q_t(r) h(r) \\
&= \frac{2\theta}{\beta}\int_{(0,\infty)} h(r)\mathrm{d}r\, r \int_0^1 \frac{1+u}{u^2}\,\mathrm{e}^{-2\theta r/u}\,\mathrm{d}u \\
&= \frac{1}{\beta}\left(\mathrm{e}^{-2\theta r}+2\theta r \int_0^1 \mathrm{e}^{-2\theta r/u}\,\frac{du}{u}\right) \\
&= \frac{1}{\beta}\left(\mathrm{e}^{-2\theta r}+2\theta r\, \Gamma(0,2\theta r)\right).
\end{aligned}
$$

$\square$

5.2. **The clonal subpopulation size.** In this section, we consider the size Z_{cl} of the clonal population at time 0, meaning the individuals sharing the same type as the MRCA ϱ of the extant population at time 0:

$$Z_{\mathrm{cl}} = \mathcal{Z}_0(\mathcal{T}^{\mathrm{st}}_{\mathrm{clonal}}(\varrho)),$$

with the clonal sub-tree defined by (33). Of course, we have $Z_{\mathrm{cl}} \leq Z_0$ a.s.. By definition of the mutation point measure $\mathcal{M}$, we get that for all $n \geq 1$:

$$\mathbb{E}[Z_{\mathrm{cl}}^n\,|\,Z_0] = \mathbb{E}\left[\int_{(\mathcal{T}^{\mathrm{st}})^n} \mathrm{e}^{-\mu L(\mathcal{X}_1,\ldots,\mathcal{X}_n)}\prod_{i=1}^n \mathcal{Z}_0(\mathrm{d}\mathcal{X}_i)\,\Bigg|\,Z_0\right],$$

where $L(\mathcal{X}_1,\ldots,\mathcal{X}_n)$ is the length of the tree $\mathcal{T}'_n$ spanned by the n leaves $\mathcal{X}_1,\ldots,\mathcal{X}_n$ uniformly sampled in the extant population at time 0 and the MRCA, say ϱ', of the extant population. We recall the tree $\mathcal{T}_n$ spanned by the n leaves $\mathcal{X}_1,\ldots,\mathcal{X}_n$ is rooted at the MRCA of $\mathcal{X}_1,\ldots,\mathcal{X}_n$, and is thus a sub-tree of $\mathcal{T}'_n$ obtained by removing the (possibly empty) branch from ϱ' to just before the MRCA of $\mathcal{X}_1,\ldots,\mathcal{X}_n$.

Following Section 3, we consider the tree $\mathfrak{T}'_n$ defined as $\mathfrak{T}_n$ but for the last step where we cut the semi-infinite branch not at its last (going downwards) branching point ϱ_n, which is at length $\max_{1\leq k\leq n}\zeta_k$, but at ϱ which is at length $\max_{0\leq k\leq n+1}\zeta_k$. Notice that the distribution of $\max_{0\leq k\leq n+1}\zeta_k$ does not depend on n, see (11), which explain why we do not stress the dependence of ϱ in n. See Fig. 1 for an instance of $\mathfrak{T}'_n$. Similarly to Lemma 3.1, using [1, Lemma 4.1], we get the following result.

Lemma 5.2 (Representation of the genealogical tree of n individuals and the MRCA of the extant population). *For $n \in \mathbb{N}^*$, the rooted tree $\mathcal{T}'_n$ spanned by $\varrho'_n, \mathcal{X}_1,\ldots,\mathcal{X}_n$ is distributed as the rooted tree $\mathfrak{T}'_n$.*

We thus deduce that:

(37) $$\mathbb{E}[Z_{\mathrm{cl}}^n\,|\,Z_0] = \mathbb{E}\left[Z_0^n\,\mathrm{e}^{-\mu L_n}\,\big|\,Z_0\right]$$

with L_n the total length of the tree $\mathfrak{T}'_n$. By construction the total length of $\mathfrak{T}'_n$ is given by the length of the segments attached to the random points $X_1,\ldots,X_{n-1}$ and the the semi-infinite spine cut at $\max_{0\leq k\leq n+1}\zeta_k$ which is attached to $X_0 = 0$, that is:

$$L_n = \max_{0\leq k\leq n+1}\zeta_k + \Lambda_{n-1} \quad \text{and} \quad \Lambda_{n-1} = \sum_{k=1}^n \zeta_k.$$

(Notice that in the above formula $\zeta_\ell = 0$ for the index $\ell \in [\![1, n]\!]$ such that $X_{(\ell)} = X_0$.)

Remark 5.3 (On the asymptotic of L_n). Let us mention that the asymptotics of Λ_{n-1} has been computed in [1, Section 5], and we have the following convergence in distribution:

$$\Lambda_{n-1} - \frac{Z_0}{\beta} \log\left(\frac{n}{2\theta Z_0}\right) \xrightarrow[n\to\infty]{(d)} \mathcal{L},$$

where the distribution of $\mathcal{L}$ is given in [4, Lemma 5.4] (with $\mathcal{L}$ denoted by W_0 therein). In fact the construction of the ζ_k's can be done in such way that this convergence is a.s., see [1, Theorem 5.1]. This provides the a.s. convergence of $L_n - Z_0 \log(n)/\beta$ in the setting of [1]. However, we did not investigate the joint law of $\mathcal{L}$ and the TMRCA of the whole population at time 0 given by $\max_{0 \le k \le n+1} \zeta_k$ (which we recall does not depend on n).

Recall that $\beta(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ for $a, b \in \mathbb{R}_+^*$. We set:

$$R = \frac{Z_{\mathrm{cl}}}{Z_0} \quad \text{and} \quad \alpha = \frac{\mu}{2\beta\theta}.$$

Theorem 5.4. *For $n \in \mathbb{N}^*$, we have:*

$$(38) \qquad \mathbb{E}[Z_{\mathrm{cl}}^{n-1} R] = \frac{\alpha}{(1 + \alpha)^n} \left[\beta\left(n, \frac{2 + \alpha}{1 + \alpha}\right) + \beta\left(n, \frac{\alpha}{1 + \alpha}\right) - \frac{2}{n}\right] \mathbb{E}\left[Z_0^{n-1}\right].$$

The formula for $\mathbb{E}[Z_{\mathrm{cl}}^n]$ is explicit and given by (44) below. In particular, we have:

$$(39) \qquad \mathbb{E}[Z_{\mathrm{cl}}] = \mathbb{E}[RZ_0] = \frac{6}{(\alpha + 1)(\alpha + 2)(\alpha + 3)} \mathbb{E}[Z_0] \quad \text{and} \quad \mathbb{E}[R] = \frac{2}{(\alpha + 1)(\alpha + 2)}$$

and thus:

$$\mathrm{Cov}(R, Z_0) = -\frac{2\alpha}{(\alpha + 1)(\alpha + 2)(\alpha + 3)} \mathbb{E}[Z_0].$$

We also have:

$$\frac{\mathbb{E}[Z_{\mathrm{cl}}^n]}{\mathbb{E}[Z_0^n]} \sim_{n\to\infty} \frac{2\alpha}{2 + \alpha} \Gamma\left(\frac{\alpha}{1 + \alpha}\right) \frac{1}{(1 + \alpha)^n n^{\alpha/(1+\alpha)}}.$$

Interestingly, Theorem 5.4 shows that R and Z_0 are negatively correlated as $\mathrm{Corr}(R, Z_0) = -1 + 3/(\alpha + 3)$: larger populations tend to have smaller clonal subpopulations, and this effect becomes stronger as the mutation rate increases.

5.3. **Proofs of Theorem 5.4.** We shall use many times the following formula for $b > 0$:

$$\beta(1, b) = \frac{1}{b} \quad \text{and} \quad \beta(a, b) \sim_{a\to\infty} \Gamma(b) a^{-b},$$

and, as $\Gamma(b + 1) = b\Gamma(b)$, for $a > 1$:

$$\beta(a - 1, b + 1) = \frac{b}{a - 1} \beta(a, b).$$

We shall also use that for U uniform on $[0, 1]$, $a \ge 0$, $k > 0$ and $b = a/(1 + \alpha)$:

$$(40) \qquad \mathcal{U}(k, a) := \mathbb{E}\left[U^{\alpha + a}\left(1 - U^{1+\alpha}\right)^{k-1}\right] = \frac{1}{1 + \alpha} \beta(k, 1 + b),$$

that for $a > 0$ and $k > 1$:

$$(41) \qquad \mathcal{U}(k - 1, a) = \frac{a}{k - 1} \frac{1}{(1 + \alpha)^2} \beta(k, b),$$

and that for $a > 1 + \alpha$ and $k > 2$:

$$(42) \qquad \mathcal{U}(k-2, a) = \frac{a(a-1-\alpha)}{(k-1)(k-2)} \frac{1}{(1+\alpha)^3} \beta(k, b-1).$$

Let $n \in \mathbb{N}^*$. As Z_0 is the sum of two independent exponential random variables with mean $1/2\theta$, we get:

$$\mathbb{E}[Z_0^{n-1}] = \frac{n!}{(2\theta)^{n-1}}.$$

Using (37), we first consider the quantity:

$$\mathbb{E}[Z_{\mathrm{cl}}^{n-1} R] = \mathbb{E}\left[Z_0^{n-1} \, \mathrm{e}^{-\mu L_n}\right].$$

We shall now go back to the definition of the random variables $(\zeta_k, 0 \leq k \leq n+1)$ from Item (ii) of Section 3 to give a nice representation of the distribution of $(\max_{0 \leq k \leq n+1} \zeta_k, \Lambda_{n-1})$ under the probability measure $\mathrm{d}\mathbb{Q}_n = Z_0^{n-1} \mathrm{d}\mathbb{P} / \mathbb{E}[Z_0^{n-1}]$. Thus, since $2\theta Z_0$ has the $\Gamma(2, 1)$ distribution, we obtain that under $\mathbb{Q}_n$ it has the $\Gamma(n+1, 1)$ distribution.

Recall the random variables $X_{(0)} = -E_{\mathrm{g}} < X_{(1)} < \ldots < X_{(n)} < X_{(n+1)} = E_{\mathrm{d}}$. For $k \in \{0, \ldots, n+1\}$ are the order statistics of $\{-E_{\mathrm{g}}, E_{\mathrm{d}}, X_0, \ldots, X_{n-1}\}$ with $X_0 = 0$ and $X_k = Z_0 U_k - E_{\mathrm{g}}$ for $k \in \mathbb{N}^*$ and $(U_k, \ k \in \mathbb{N}^*)$ be independent random variables, uniformly distributed on $[0, 1]$, independent of $E_{\mathrm{g}}, E_{\mathrm{d}}$.

In particular the random variables $(\Delta_k = 2\theta(X_{(k)} - X_{(k-1)}), 1 \leq k \leq n+1)$ are distributed as $(2\theta Z_0(U'_{(k)} - U'_{(k-1)}), 1 \leq k \leq n+1)$, where $U'_{(0)} = 0 < U'_{(1)} < \ldots < U'_{(n)} < U'_{(n+1)} = 1$ is the order statistics of $\{0, 1, U'_1, \ldots, U'_n\}$, where the random variables $(U'_k, \ k \in \mathbb{N}^*)$ are uniformly distributed on $[0, 1]$, independent and independent of Z_0. Using properties of the Poisson process, we deduce that under $\mathbb{Q}_n$, the random variables $(\Delta_k, 1 \leq k \leq n+1)$ are distributed as $(E_k, 1 \leq k \leq n+1)$, where $\mathbf{E} = (E_k, k \in \mathbb{N}^*)$ are independent exponential random variables with mean 1.

Set $(\zeta'_k, 1 \leq k \leq n+1)$ with:

$$\zeta'_k = \frac{1}{2\beta\theta} \log\left(\frac{E'_k + E_k}{E'_k}\right),$$

where the random variables $(E'_k, k \in \mathbb{N}^*)$ are distributed as $\mathbf{E}$ and independent of $\mathbf{E}$. Now recall there exists a (random) index $i \in [\![1, n]\!]$ such that $\zeta_i = 0$, so intuitively among the $n+2$ random variable $\zeta_0, \ldots, \zeta_{n+1}$, there are only $n+1$ non trivial ones. More precisely, we get that $(\max_{0 \leq k \leq n+1} \zeta_k, \Lambda_{n-1})$ is under $\mathbb{Q}_n$ distributed as:

$$\left(\max_{1 \leq k \leq n+1} \zeta'_k, \ \sum_{k=2}^{n} \zeta'_k\right).$$

The random variables $(V_k, k \in \mathbb{N})$, with:

$$V_k = \frac{E'_k}{E'_k + E_k},$$

are independent and uniformly distributed on $[0, 1]$. We deduce that:

$$(43) \qquad \mathbb{E}[Z_{\mathrm{cl}}^{n-1} R] = \mathbb{E}\left[Z_0^{n-1}\right] \mathbb{E}\left[\left(\min_{1 \leq k \leq n+1} V_k^\alpha\right) \prod_{j=2}^{n} V_j^\alpha\right].$$

Elementary computations give that:

$$\mathbb{E}\left[\left(\min_{1\le k\le n+1} V_k^\alpha\right)\prod_{j=2}^n V_j^\alpha\right] = 2A_n + (n-1)B_n,$$

with, thanks to (40):

$$A_n = \mathbb{E}\left[V_1^\alpha \prod_{k=2}^{n+1}\mathbf{1}_{\{V_1<V_k\}}\prod_{j=2}^n V_j^\alpha\right] = \frac{1}{(1+\alpha)^{n-1}}\,\mathbb{E}\left[U^\alpha(1-U)\left(1-U^{1+\alpha}\right)^{n-1}\right]$$

$$= \frac{1}{(1+\alpha)^{n-1}}\left(\mathcal{U}(n,0)-\mathcal{U}(n,1)\right)$$

$$= \frac{1}{(1+\alpha)^n}\left[\frac{1}{n}-\beta\left(n,\frac{2+\alpha}{1+\alpha}\right)\right],$$

and for $n\ge 2$, thanks to (41):

$$B_n = \mathbb{E}\left[V_2^{2\alpha}\prod_{k=1,3,\dots,n+1}\mathbf{1}_{\{V_2<V_k\}}\prod_{j=3}^{n-1}V_j^\alpha\right]$$

$$= \frac{1}{(1+\alpha)^{n-2}}\,\mathbb{E}\left[U^{2\alpha}(1-U)^2\left(1-U^{1+\alpha}\right)^{n-2}\right]$$

$$= \frac{1}{(1+\alpha)^{n-2}}\left(\mathcal{U}(n-1,\alpha)-2\mathcal{U}(n-1,1+\alpha)+\mathcal{U}(n-1,2+\alpha)\right)$$

$$= \frac{1}{n-1}\,\frac{1}{(1+\alpha)^n}\left[\alpha\beta\left(n,\frac{\alpha}{1+\alpha}\right)-\frac{2(1+\alpha)}{n}+(2+\alpha)\beta\left(n,\frac{2+\alpha}{1+\alpha}\right)\right].$$

We deduce that:

$$2A_n + (n-1)B_n = \frac{\alpha}{(1+\alpha)^n}\left[\beta\left(n,\frac{2+\alpha}{1+\alpha}\right)+\beta\left(n,\frac{\alpha}{1+\alpha}\right)-\frac{2}{n}\right].$$

We thus deduce (38) from (43). Taking $n=1$, gives the value of $\mathbb{E}[R]$ in (39).

We now compute $\mathbb{E}[Z_{\mathrm{cl}}^n]$. We have:

$$\mathbb{E}[Z_{\mathrm{cl}}^n] = \frac{1}{2\theta}\,\mathbb{E}\left[Z_0^{n-1}(2\theta Z_0)\,\mathrm{e}^{-\mu L_n}\right]$$

$$= \frac{1}{2\theta}\,\mathbb{E}[Z_0^{n-1}]\,\mathbb{E}\left[(E_1+\dots E_{n+1})\min_{1\le k\le n+1}\left(\frac{E_k'}{E_k+E_k'}\right)^\alpha\prod_{j=2}^n\left(\frac{E_j'}{E_j+E_j'}\right)^\alpha\right]$$

$$= \frac{1}{n+1}\,\mathbb{E}[Z_0^n]\left(2C_n+(n-1)D_n\right),$$

with:

$$C_n = \mathbb{E}\left[E_1\min_{1\le k\le n+1}\left(\frac{E_k'}{E_k+E_k'}\right)^\alpha\prod_{j=2}^n\left(\frac{E_j'}{E_j+E_j'}\right)^\alpha\right]$$

and for $n\ge 2$:

$$D_n = \mathbb{E}\left[E_2\min_{1\le k\le n+1}\left(\frac{E_k'}{E_k+E_k'}\right)^\alpha\prod_{j=2}^n\left(\frac{E_j'}{E_j+E_j'}\right)^\alpha\right].$$

We have:

$$C_n = \mathbb{E}\left[(E_1 + E_1')\, (1 - V_1) \left(\min_{1 \le k \le n+1} V_k^\alpha \right) \prod_{j=2}^n V_k^\alpha \right]$$

$$= 2\mathbb{E}\left[(1 - V_1) \left(\min_{1 \le k \le n+1} V_k^\alpha \right) \prod_{j=2}^n V_k^\alpha \right]$$

$$= 2A_n^{(0)} + A_n^{(01)} + (n-1)B_n^{(0)},$$

where we used that $E_1 + E_1'$ is independent of V_1 for the third equality, and with:

$$A_n^{(0)} = \mathbb{E}\left[(1 - V_1) V_1^\alpha \prod_{k=2}^{n+1} \mathbf{1}_{\{V_1 < V_k\}} \prod_{j=2}^n V_j^\alpha \right]$$

$$= \frac{1}{(1+\alpha)^{n-1}}\, \mathbb{E}\left[U^\alpha (1 - U)^2 (1 - U^{\alpha+1})^{n-1} \right]$$

$$= \frac{1}{(1+\alpha)^{n-1}} \left(\mathcal{U}(n,0) - 2\mathcal{U}(n,1) + \mathcal{U}(n,2) \right)$$

$$= \frac{1}{(1+\alpha)^n} \left[\frac{1}{n} - 2\beta\left(n, \frac{2+\alpha}{1+\alpha} \right) + \beta\left(n, \frac{3+\alpha}{1+\alpha} \right) \right],$$

and (using elementary computations for the last equality):

$$A_n^{(01)} = 2\mathbb{E}\left[(1 - V_1)\, V_{n+1}^\alpha \prod_{k=1}^n \mathbf{1}_{\{V_{n+1} < V_k\}} \prod_{j=2}^n V_j^\alpha \right] = A_n^{(0)},$$

and for $n \ge 2$:

$$B_n^{(0)} = 2\mathbb{E}\left[(1 - V_1)\, V_2^{2\alpha} \prod_{k=1,3,\ldots,n+1} \mathbf{1}_{\{V_1 < V_k\}} \prod_{j=3}^n V_j^\alpha \right]$$

$$= \frac{1}{(1+\alpha)^{n-2}}\, \mathbb{E}\left[U^{2\alpha} (1 - U)^3 (1 - U^{1+\alpha})^{n-2} \right]$$

$$= \frac{1}{(1+\alpha)^{n-2}} \left(\mathcal{U}(n-1,\alpha) - 3\mathcal{U}(n-1,1+\alpha) + 3\mathcal{U}(n-1,2+\alpha) - \mathcal{U}(n-1,3+\alpha) \right)$$

$$= \frac{1}{n-1}\, \frac{1}{(1+\alpha)^n} \left[\alpha\beta\left(n, \frac{\alpha}{1+\alpha} \right) - 3\frac{(1+\alpha)}{n} + 3(2+\alpha)\beta\left(n, \frac{2+\alpha}{1+\alpha} \right) - (3+\alpha)\beta\left(n, \frac{3+\alpha}{1+\alpha} \right) \right].$$

Similarly, we also have for $n \ge 2$:

$$D_n = \mathbb{E}\left[(E_2 + E_2')\, (1 - V_2) \left(\min_{1 \le k \le n+1} V_k^\alpha \right) \prod_{j=2}^n V_j^\alpha \right]$$

$$= 2\mathbb{E}\left[(1 - V_2) \left(\min_{1 \le k \le n+1} V_k \right)^\alpha \prod_{j=2}^n V_j^\alpha \right]$$

$$= 2\left(2A_n^{(1)} + B_n^{(11)} + (n-2)B_n^{(1)} \right),$$

with:

$$A_n^{(1)} = \mathbb{E}\left[V_1^\alpha(1-V_2)\prod_{k=2}^{n+1}\mathbf{1}_{\{V_1<V_k\}}\prod_{j=2}^n V_j^\alpha\right] = A_n - A_n^{(2)},$$

and:

$$A_n^{(2)} = \mathbb{E}\left[V_1^\alpha V_2\prod_{k=2}^{n+1}\mathbf{1}_{\{V_1<V_k\}}\prod_{j=2}^n V_j^\alpha\right]$$

$$= \frac{1}{(2+\alpha)(1+\alpha)^{n-2}}\,\mathbb{E}\left[U^\alpha(1-U)(1-U^{2+\alpha})(1-U^{1+\alpha})^{n-2}\right]$$

$$= \frac{1}{(2+\alpha)(1+\alpha)^{n-2}}\Big(\mathcal{U}(n-1,0) - \mathcal{U}(n-1,2+\alpha) - \mathcal{U}(n-1,1) + \mathcal{U}(n-1,3+\alpha)\Big)$$

$$= \frac{1}{n-1}\,\frac{1}{(2+\alpha)(1+\alpha)^n}\left[(1+\alpha) - (2+\alpha)\beta\left(n,\frac{2+\alpha}{1+\alpha}\right) - \beta\left(n,\frac{1}{1+\alpha}\right) + (3+\alpha)\beta\left(n,\frac{3+\alpha}{1+\alpha}\right)\right],$$

and with (using elementary computations for the last equality):

$$B_n^{(11)} = \mathbb{E}\left[(1-V_2)\,V_2^{2\alpha}\prod_{k=1,3,\ldots,n}\mathbf{1}_{\{V_2<V_k\}}\prod_{j=3}^n V_j^\alpha\right] = B_n^{(0)},$$

and lastly with, for $n \geq 3$:

$$B_n^{(1)} = \mathbb{E}\left[(1-V_2)\,V_2^\alpha V_3^{2\alpha}\prod_{k=1,2,4,\ldots,n+1\}}\mathbf{1}_{\{V_3<V_k\}}\prod_{j=4}^n V_j^\alpha\right]$$

$$= B_n - B_n^{(2)},$$

and, using (42):

$$B_n^{(2)} = \frac{1}{(2+\alpha)(1+\alpha)^{n-3}}\,\mathbb{E}\left[U^{2\alpha}(1-U)^2(1-U^{2+\alpha})(1-U^{1+\alpha})^{n-3}\right]$$

$$= \frac{1}{(2+\alpha)(1+\alpha)^{n-3}}\Big(\mathcal{U}(n-2,\alpha) - \mathcal{U}(n-2,2+2\alpha)$$

$$- 2\mathcal{U}(n-2,1+\alpha) + 2\mathcal{U}(n-2,3+2\alpha) + \mathcal{U}(n-2,2+\alpha) - \mathcal{U}(n-2,4+2\alpha)\Big)$$

$$= \frac{1}{(n-1)(n-2)}\,\frac{1}{(2+\alpha)(1+\alpha)^n}\left[(n-1)\alpha(1+\alpha)\beta\left(n-1,\frac{\alpha}{1+\alpha}\right) - \frac{(2+2\alpha)(1+\alpha)}{n}\right.$$

$$- 2(1+\alpha)^2 + 2(3+2\alpha)(2+\alpha)\beta\left(n,\frac{2+\alpha}{1+\alpha}\right)$$

$$\left.+ (2+\alpha)\beta\left(n,\frac{1}{1+\alpha}\right) - (4+2\alpha)(3+\alpha)\beta\left(n,\frac{3+\alpha}{1+\alpha}\right)\right].$$

In conclusion, we obtain that:

$$(44)\quad \mathbb{E}[Z_{\text{cl}}^n]$$

$$= \frac{1}{n+1}\,\mathbb{E}[Z_0^n]\Big(2C_n + (n-1)D_n\Big)$$

$$= \frac{2}{n+1}\,\mathbb{E}[Z_0^n]\Big(3A_n^{(0)} + 2(n-1)(A_n - A_n^{(2)} + B_n^{(0)}) + (n-1)(n-2)(B_n - B_n^{(2)})\Big).$$

Taking $n = 1$ in the above formula, we get:

$$\mathbb{E}[Z_{\mathrm{cl}}] = 3A_1^{(0)}\,\mathbb{E}[Z_0] = \frac{3}{(1+\alpha)}\left[1 - 2\frac{1+\alpha}{2+\alpha} + \frac{1+\alpha}{3+\alpha}\right]\mathbb{E}[Z_0],$$

which gives the first part of (39). We now give the leading term in (44). We have:

$$(1+\alpha)^n\,A_n^{(0)} = O(n^{-1}),$$
$$(1+\alpha)^n\,A_n = O(n^{-1}),$$
$$(1+\alpha)^n\,A_n^{(2)} = O(n^{-1}),$$
$$(1+\alpha)^n\,B_n^{(0)} = o(n^{-1}),$$
$$(1+\alpha)^n\,B_n = n^{-1-\alpha/(1+\alpha)}\alpha\Gamma\left(\frac{\alpha}{1+\alpha}\right) + O(n^{-2}),$$
$$(1+\alpha)^n\,B_n^{(2)} = n^{-1-\alpha/(1+\alpha)}\frac{1+\alpha}{2+\alpha}\alpha\Gamma\left(\frac{\alpha}{1+\alpha}\right) + O(n^{-2}).$$

We deduce that:

$$\mathbb{E}[Z_{\mathrm{cl}}^n] = \frac{2\alpha}{2+\alpha}\,\Gamma\left(\frac{\alpha}{1+\alpha}\right)\mathbb{E}[Z_0^n]\,\frac{1}{(1+\alpha)^n}\left(\frac{1}{n^{\alpha/(1+\alpha)}} + O(n^{-1})\right).$$

This ends the proof of Theorem 5.4.

References

[1] R. Abraham and J.-F. Delmas. Exact simulation of the genealogical tree for a stationary branching population and application to the asymptotics of its total length. *Advances in Applied Probability*, 53(2):537–574, 2021.

[2] J. Berestycki, N. Berestycki, and V. Limic. Asymptotic sampling formulae for Λ-coalescents. *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, 50(3):715–731, 2014.

[3] A. Bhaskar and Y. S. Song. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Annals of Statistics*, 42(6):2469–2493, 2014.

[4] H. Bi and J.-F. Delmas. Total length of the genealogical tree for quadratic stationary continuous-state branching processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 52(3), 2016.

[5] M. Birkner, J. Blath, and B. Eldon. Statistical properties of the site-frequency spectrum associated with lambda-coalescents. *Genetics*, 195(3):1037–53, 2013.

[6] J. Blath, M. C. Cronjäger, B. Eldon, and M. Hammer. The site-frequency spectrum associated with Ξ-coalescents. *Theoretical Population Biology*, 110:36–50, 2016.

[7] Y.-T. Chen and J.-F. Delmas. Smaller population size at the MRCA time for stationary branching processes. *The Annals of Probability*, 40(5), 2012.

[8] J. J. Duchamps and A. Lambert. Mutations on a random binary tree with measured boundary. *Annals of Applied Probability*, 28(4):2141–2187, 2018.

[9] T. Duquesne and J.-F. Le Gall. *Random Trees, Lévy Processes and Spatial Branching Processes*, volume 281. SMF, 2002.

[10] T. Duquesne and J.-F. Le Gall. Probabilistic and fractal aspects of Lévy trees. *Probability Theory and Related Fields*, 131(4):553–603, 2005.

[11] B. Eldon, M. Birkner, J. Blath, and F. Freund. Can the Site-Frequency Spectrum Distinguish Exponential Population Growth from Multiple-Merger Coalescents? *Genetics*, 2015.

[12] S. N. Evans, J. Pitman, and A. Winter. Rayleigh processes, real trees, and root growth with re-grafting. *Probability Theory and Related Fields*, 134(1):81–126, 2005.

[13] F. Freund, E. Kerdoncuff, S. Matuszewski, M. Lapierre, M. Hildebrandt, J. D. Jensen, L. Ferretti, A. Lambert, T. B. Sackton, and G. Achaz. Interpreting the pervasive observation of U-shaped Site Frequency Spectra. *PLOS Genetics*, 19(3):e1010677, 2023.

[14] Y. X. Fu. Statistical Properties of Segregating Sites. *Theoretical Population Biology*, 48(2):172–197, 1995.

[15] R. C. Griffiths and S. Tavaré. The age of a mutation in a general coalescent tree. *Communications in Statistics. Stochastic Models*, 14(1-2):273–295, 1998.

[16] G. Kersting, A. Siri-Jégousse, and A. H. Wences. Site Frequency Spectrum of the Bolthausen-Sznitman Coalescent. *Latin American Journal of Probability and Mathematical Statistics*, 18(1):1483, 2021.

[17] J. Kim, E. Mossel, M. Z. Rácz, and N. Ross. Can one hear the shape of a population history? *Theoretical Population Biology*, 100:26–38, 2015.

[18] J. Koskela. Multi-locus data distinguishes between population growth and multiple merger coalescents. *Statistical Applications in Genetics and Molecular Biology*, 17(3), 2018.

[19] J. Koskela, P. A. Jenkins, and D. Spanò. Computational inference beyond Kingman's coalescent. *Journal of Applied Probability*, 52(2):519–537, 2015.

[20] J. Koskela, P. A. Jenkins, and D. Spanò. Bayesian non-parametric inference for Lambda-coalescents: Posterior consistency and a parametric method. *Bernoulli*, 24(3):2122–2153, 2018.

[21] A. Lambert. Quasi-Stationary Distributions and the Continuous-State Branching Process Conditioned to Be Never Extinct. *Electronic Journal of Probability*, 12, 2007.

[22] A. Lambert. The Allelic Partition for Coalescent Point Processes. *Markov Processes and Related Fields*, 15:359–386, 2009.

[23] A. Lambert. The coalescent of a sample from a binary branching process. *Theoretical Population Biology*, 122:30–35, 2018.

[24] Z. Li. *Measure-Valued Branching Markov Processes*. Springer, 2011.

[25] S. Matuszewski, M. E. Hildebrandt, G. Achaz, and J. D. Jensen. Coalescent Processes with Skewed Offspring Distributions and Nonequilibrium Demography. *Genetics*, 208(1):323–338, 2018.

[26] S. Myers, C. Fefferman, and N. Patterson. Can one learn history from the allelic spectrum? *Theoretical Population Biology*, 73(3):342–348, 2008.

[27] L. Popovic. Asymptotic genealogy of a critical branching process. *The Annals of Applied Probability*, 14(4):2120–2148, 2004.

[28] J. Schweinsberg and Y. Shuai. Asymptotics for the site frequency spectrum associated with the genealogy of a birth and death process, 2023.

[29] J. P. Spence, J. A. Kamm, and Y. S. Song. The Site Frequency Spectrum for General Coalescents. *Genetics*, 202(4):1549–1561, 2016.

[30] J. Terhorst and Y. S. Song. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proceedings of the National Academy of Sciences*, 112(25):7677–7682, 2015.

ROMAIN ABRAHAM, INSTITUT DENIS POISSON, UNIVERSITÉ D'ORLÉANS, UNIVERSITÉ DE TOURS, CNRS, FRANCE
Email address: `romain.abraham@univ-orleans.fr`

JEAN-FRANÇOIS DELMAS, CERMICS, ÉCOLE DES PONTS, FRANCE
Email address: `jean-francois.delmas@enpc.fr`

PATRICK HOCHEIT, INRAE, MAIAGE, UNIVERSITÉ PARIS-SACLAY, 78350 JOUY-EN-JOSAS, FRANCE
Email address: `patrick.hoscheit@inrae.fr`