Species abundance distributions in neutral models with immigration or mutation and general lifetimes.

By Amaury Lambert, UPMC Univ Paris 06

September 2, 2010

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
E-mail: amaury.lambert@upmc.fr
URL: http://www.proba.jussieu.fr/pageperso/amaury/index.htm

Abstract

We consider a general, neutral, dynamical model of biodiversity. Individuals have i.i.d. lifetime durations, which are not necessarily exponentially distributed, and each individual gives birth independently at constant rate λ. Thus, the population size is a *homogeneous, binary Crump–Mode–Jagers process* (which is not necessarily a Markov process). We assume that types are clonally inherited.

We consider two classes of speciation models in this setting. In the *immigration model*, new individuals of an entirely new species singly enter the population at constant rate μ (e.g., from the mainland into the island). In the *mutation model*, each individual independently experiences point mutations in its germ line, at constant rate θ.

We are interested in the *species abundance distribution*, i.e., in the numbers, denoted $I_n(k)$ in the immigration model and $A_n(k)$ in the mutation model, of species represented by k individuals, $k = 1, 2, \dots, n$, when there are n individuals in the total population.

In the immigration model, we prove that the numbers $(I_t(k); k \geq 1)$ of species represented by k individuals *at time* t, are independent Poisson variables with parameters as in Fisher's log-series. When conditioning on the total size of the population to equal n, this results in species abundance distributions given by *Ewens' sampling formula*. In particular, $I_n(k)$ converges as $n \to \infty$ to a Poisson r.v. with mean γ/k, where $\gamma := \mu/\lambda$.

In the mutation model, as $n \to \infty$, we obtain the almost sure convergence of $n^{-1}A_n(k)$ to a nonrandom explicit constant. In the case of a critical, linear birth–death process, this constant is given by Fisher's log-series, namely $n^{-1}A_n(k)$ converges to α^k/k, where $\alpha := \lambda/(\lambda + \theta)$.

In both models, the abundances of the most abundant species are briefly discussed.

Running head. Neutral models of biodiversity with general lifetimes.
Key words and phrases. Species abundance distribution – Crump–Mode–Jagers process – splitting tree – branching process – linear birth–death process – immigration – mutation – infinitely-many alleles model – Fisher logarithmic series – Ewens sampling formula – coalescent point process – scale function.

1 Introduction

Our goal is to study two models of speciation in the vein of the neutral theory of biodiversity [14], an *immigration model* and a *mutation model*, both in a same general birth/death dynamical setting. A specific feature of our results is that no assumption is made on the distribution of lifetime durations, contrasting with usual Markovian dynamics where this distribution is exponential.

We assume that particles behave independently from one another, that each particle gives birth at constant rate λ during its lifetime (interbirth durations are i.i.d. exponential random variables with parameter λ), and that lifetime durations are i.i.d.. Then the process $(N_t; t \geq 0)$ giving the number of extant individuals at time t, belongs to a wide class of branching processes called *Crump–Mode–Jagers processes*. Actually, the processes we consider are homogeneous (constant birth rate) and binary (one birth at a time) but differ in generality from classic birth–death processes in that the lifetimes durations may follow a general distribution.

Now each individual bears some type (or, equivalently, belongs to some species), and we will assume that, at each birth time t, the type of the mother at time t is passed on to their offspring without modification. However, new species can arise in this population. These new types can arise in two fashions, whence defining either speciation model.

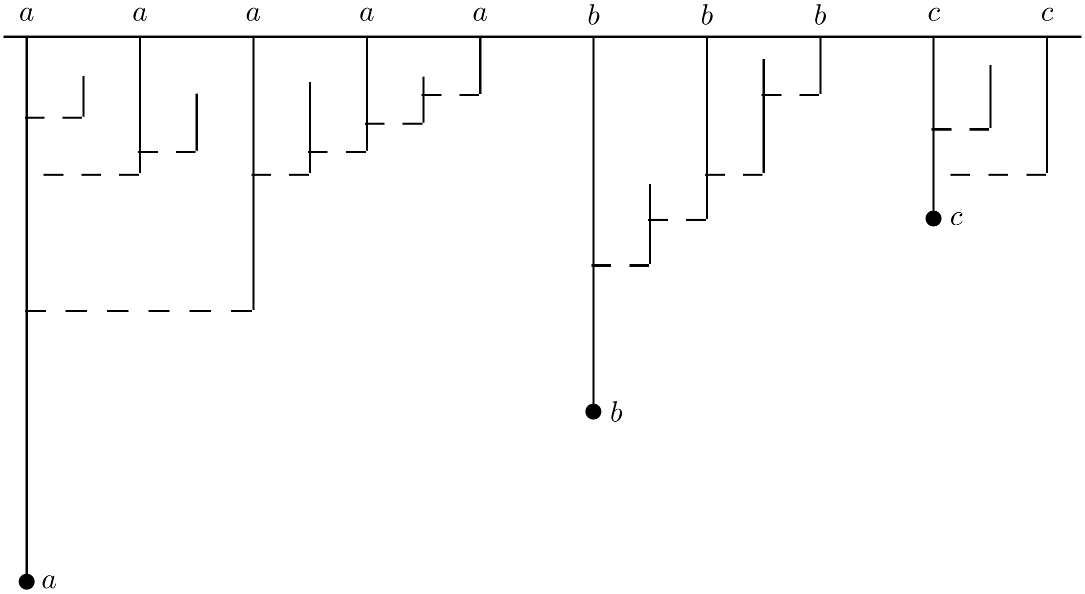

Figure 1: The immigration model. Time axis is vertical; horizontal axis shows filiation. Solid dots show the arrival times of immigrants, who all have distinct types labelled by letters a, b, c. The type of each extant individual is also shown.

The immigration model is a generalization of Karlin and McGregor's model [18] to general lifetimes. It intends to model a population on an island receiving immigrants from the mainland, as in the theory of island biogeography [23]. We assume that new propagules singly enter the island population at the instants of a Poisson process with rate μ, called the *immigration rate*, and behave from then on, as the other particles on the island. Each of these immigrating particles is of an entirely new species, but their whole descendance is entirely clonal. See Figure 1.

In the mutation model, we assume that the germ line of each particle experiences mutations during the whole lifetime of the particle. At the instants of a Poisson process with rate θ, the type of the particle changes to an entirely new type, as in the *infinitely-many alleles model* [9]. See Figure 2.

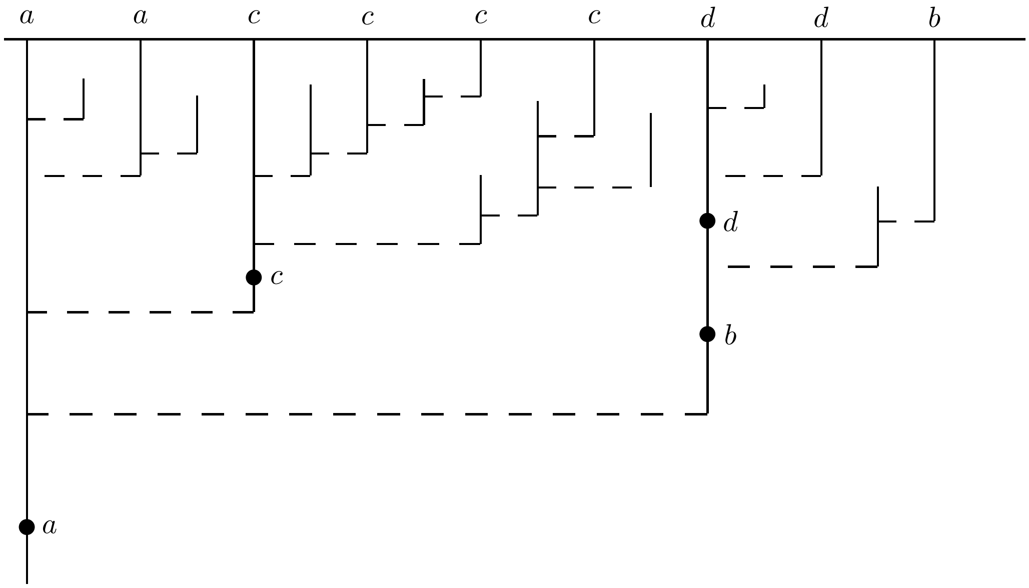

Figure 2: The mutation model. Time axis is vertical; horizontal axis shows filiation. Solid dots show the mutation events. Each mutation yields a new type, labelled by letters a, b, c, d. The type of each extant individual is also shown.

Another way of seeing the model is to replace the word particle with the word colony, and the word population with the word metapopulation. Then in our model, all individuals of a colony are of the same species, lifetimes are extinction times of colonies, and birth events correspond to propagules sent out by a colony to found a brand new colony. Immigration events correspond to propagules immigrating from the mainland and founding simultaneously a brand new colony. Mutation events correspond to mutants appearing in a colony and getting to fixation instantaneously. This way of modeling speciation is more satisfactory, but we stick to the first terminology not to obscure reading.

2 Statements of results and Fisher's logarithmic series

In [10, 11], R.A. Fisher and his coauthors suggested a simple model of species count where the probability of observing k individuals of a given species is $c\alpha^k/k$ for some constant $\alpha \in (0,1)$. Following this, a number of authors proposed dynamical models where this so-called *log-series* not only gives the distribution of the number of individuals of a single species, but also the multivariate species abundance distribution of a community, in the sense that the number of species represented by k individuals follows independently a Poisson distribution with parameter $c\alpha^k/k$. For example, Karlin and McGregor [18] studied various dynamical models of structured populations, including a critical birth–death process with immigration which is a particular case of our immigration model (i.e., where the lifespan is exponentially distributed), satisfying the previously described property. See also [19, 20], and [28] for a very nice and comprehensive account on these models and on their associated multivariate distributions.

Let us fix some time t. In the immigration model (resp. in the mutation model), we let $I_t(k)$ (resp. $A_t(k)$) denote the number of species represented by k individuals at time t. When conditioning on the total number of individuals being n at this fixed time t, we will write $I_t(k)$ instead of $I_n(k)$ and $A_t(k)$ instead of $A_n(k)$. The vectors $(I_{.}(k))_k$ and $(A_{.}(k))_k$ are called *frequency spectra*.

In the immigration model, we actually provide a rather accurate result (Theorem 4.1) on the spectrum at any time t, without conditioning on the number of individuals, stating that the random variables $(I_t(k))_k$ are independent Poisson variables with parameters as in Fisher's log-series, with a parameter α depending on time t. In Corollary 4.2, we prove that the random vector $(I_n(1), \dots, I_n(n))$ has the same law as a vector of independent Poisson variables $(Y_1, \dots, Y_n)$ conditioned on $\sum_{k=1}^n kY_k = n$, where Y_k follows the Poisson distribution with parameter γ/k, γ being defined as the immigration-to-birth rate ratio μ/λ. These two results are known in the case of a critical, linear birth–death process [18]. Notice that the conditioning in the corollary not only removes the dependence upon the origination time t, but also on the distribution of lifetime durations. This spectrum is exactly the one described by *Ewens' sampling formula* [6, 8, 9]. The asymptotic behaviour of this spectrum is well-known (see for example [5, 6]): for any fixed j,

$$\lim_{n\to\infty} (I_n(1), I_n(2), \dots, I_n(j)) \overset{\mathcal{L}}{=} (Y_1, Y_2, \dots, Y_j)$$

where the Y_k's are *independent* Poisson variables with parameter γ/k.

This result contrasts with the mutation model, where species with abundance k are shown to accumulate linearly with population size, instead of stabilizing as previously. First, Theorem 5.1 gives the expected number of species with a fixed age and with abundance k. Then Theorem 5.3 gives exact formulae for the almost-sure asymptotic accumulation of species with given abundances. In the case of a critical birth–death process with (birth/death rate λ and) mutation rate θ, we get

$$\lim_{n\to\infty} n^{-1} A_n(k) = c\frac{\alpha^k}{k} \qquad \text{a.s.},$$

where $\alpha := \lambda/(\lambda+\theta)$, and $c = (1-\alpha)/\alpha$. We also have the a.s. convergence of the total number of species A_n divided by n to $-c\ln(1-\alpha)$.

Thus, species with k individuals tend to accumulate linearly with sample size in the mutation model, while their cardinality converges to a finite random variable in the immigration model. This has an important consequence for the species with a large number of individuals. In the immigration model, it can be shown that the oldest j species on the island have a number of individuals of the order of n, as n grows [26]. In the mutation model, in contrast, the proportion $B_n(k)$ of individuals belonging to species with more than k individuals is

$$B_n(k) = 1 - n^{-1}\sum_{j=1}^{k-1} jA_n(j) \longrightarrow 1 - \sum_{j=1}^{k-1} c\alpha^j = 1-(1-\alpha)\sum_{j=1}^{k-1}\alpha^{j-1} = \alpha^{k-1}.$$

As a consequence, for any $\varepsilon > 0$, there is an integer k such that $\limsup_n B_n(k) \le \varepsilon$. Actually, independent calculations [4] show that the most abundant species have abundances of the order of n^β, with $\beta = 1-\theta/\eta$, where η is the exponential growth rate of the total population, in the case when the mutation rate θ is smaller than η. In the case when $\theta > \eta$, these abundances are of the order of $\log(n)$.

3 Splitting trees and coalescent point processes

The genealogical trees that we consider here are usually called splitting trees [12]. Splitting trees are those random trees where individuals give birth at constant rate λ during a lifetime with general distribution $\pi(\cdot)/\lambda$, to i.i.d. copies of themselves, where π is a positive measure on $(0,\infty]$ with total mass λ called the *lifespan measure*. We assume that they are started with one unique progenitor born at time 0. We denote by $\mathbb{P}$ their law, and the subscript s in $\mathbb{P}_s$ means conditioning on the lifetime of the progenitor being s. Of course if $\mathbb{P}$ bears no subscript, this means that the lifetime of the progenitor follows the usual distribution $\pi(\cdot)/\lambda$.

In [22], we have considered the so-called jumping chronological contour process (JCCP) of the splitting tree truncated up to height (time) t, which starts at $\min(s,t)$, where s is the death time of the progenitor, visits all existence times (smaller than t) of all individuals exactly once and terminates at 0. We have shown [22, Theorem 4.3] that the JCCP is a Markov process, more specifically, it is a compound Poisson process X with jump measure π, compensated at rate -1, reflected below t, and killed upon hitting 0. We denote the law of X by P, to make the difference with the law $\mathbb{P}$ of the CMJ process. As seen previously, we record the lifetime duration, say s, of the progenitor, by writing P_s for its conditional law on $X_0 = s$.

Let us be a little more specific about the JCCP. Recall that this process visits all existence times of all individuals of the truncated tree. For any individual of the tree, we denote by α its birth time and by ω its death time. When the visit of an individual v with lifespan $(\alpha(v), \omega(v)]$ begins, the value of the JCCP is $\omega(v)$. The JCCP then visits all the existence times of v's lifespan at constant speed -1. If v has no child, then this visit lasts exactly the lifespan of v; if v has at least one child, then the visit is interrupted each time a birth time of one of v's daughters, say w, is encountered (youngest child first since the visit started at the death level). At this point, the JCCP jumps from $\alpha(w)$ to $\omega(w)\wedge t$ and starts the visit of the existence times of w. Since the tree has finite length, the visit of v has to terminate: it does so at the chronological level $\alpha(v)$ and continues the exploration of the existence times of v's mother, at the height (time) where it had been interrupted. This procedure then goes on recursively as soon as 0 is encountered (birth time of the progenitor). See Figure 3 for an example.

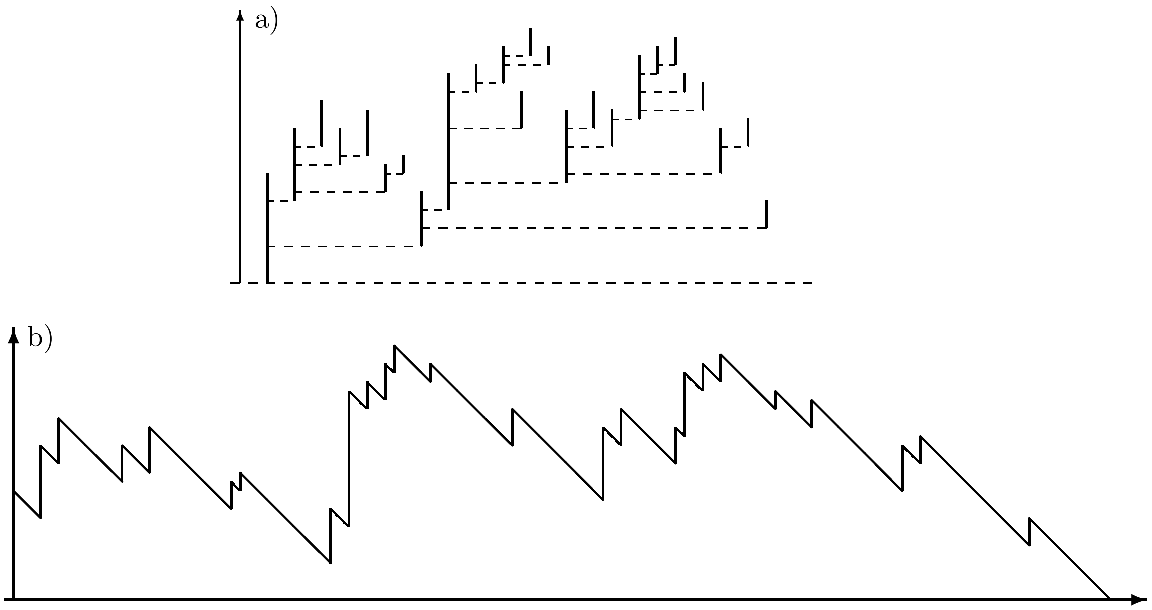

Figure 3: a) A realization of a *splitting tree* with finite extinction time. Horizontal axis has no interpretation, but horizontal arrows indicate filiation; vertical axis indicates real time; b) The associated jumping chronological *contour process* with jumps in solid line.

Since the JCCP is Markovian (as seen earlier, it is a reflected, killed Lévy process), its excursions between consecutive visits of points at height t are i.i.d. excursions of X. Observe in particular that the number of visits of t by X is exactly the number N_t of individuals alive at time t, where N is the aforementioned homogeneous, binary Crump–Mode–Jagers process. See Figure 4.

This property has two consequences, the first of which will be exploited in the immigration model, and the second one in the mutation model.

The first consequence is the computation of the one-dimensional marginals of N. Let T_A denote the first hitting time of the set A by X. Conditional on the initial progenitor to have lived s units of time, we have

$$\mathbb{P}_s(N_t = 0) = P_s(T_0 < T_{(t,+\infty)}), \tag{1}$$

and, applying recursively the strong Markov property,

$$\mathbb{P}_s(N_t = k \mid N_t \neq 0) = P_t(T_{(t,+\infty)} < T_0)^{k-1} P_t(T_0 < T_{(t,+\infty)}). \tag{2}$$

Note that the subscript s in the last display is useless.

The second consequence is that because X is (strongly) Markovian, the depths of the excursions of X away from t are i.i.d., distributed as some random variable $H := t - \inf_{0\le s\le T} X_s$, where X is started at t and T denotes the first hitting time $T_0 \wedge T_{(t,+\infty)}$ of $\{0\} \cup (t, +\infty)$ by X. We record this by letting H_i denote the depth of the excursion between the i-th visit of t and its $(i+1)$-th visit, and stating that the variables $H_1, H_2, \dots$ form a sequence of i.i.d. random variables distributed as H and killed at its first value greater than t.

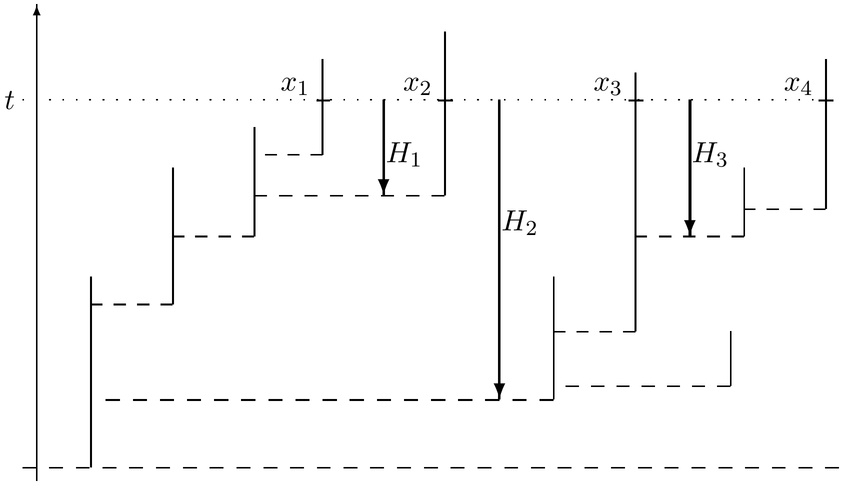

Figure 4: Illustration of a splitting tree showing the durations H_1, H_2, H_3 elapsed since *coalescence* for each of the three consecutive pairs $(x_1, x_2), (x_2, x_3)$ and (x_3, x_4) of the $N_t = 4$ individuals alive at time t.

But in the splitting tree, H_i is also the *coalescence time* (or *divergence time*) between individual i and individual $i+1$, that is, the time elapsed since the lineages of individual i and $i+1$ have diverged. Further, it can actually be shown [22] that the coalescence time $C_{i,i+k}$ between individual i and individual $i+k$ is given by

$$C_{i,i+k} = \max\{H_{i+1}, \dots, H_{i+k}\}, \tag{3}$$

so that the genealogical structure of the alive population of a splitting tree is entirely given by the knowledge of a sequence of independent random variables $H_1, H_2, \dots$ that we will call *branch lengths*, all distributed as H. We call the whole sequence the *coalescent point process*.

Here, exact formulae can be deduced for (1) and (2) from the fact that the JCCP is a Lévy process with no negative jumps. In particular, it can be convenient to handle its Laplace exponent ψ instead of its jump measure π, that is,

$$\psi(a) := a - \int_0^\infty \pi(dx)(1 - e^{-ax}) \qquad a \geq 0. \tag{4}$$

We know [22] that the process is subcritical, critical or supercritical, according to whether $m := \int_{(0,\infty]} r\pi(dr) < 1$, $= 1$ or > 1. In the latter case, the rate η at which $(N_t; t \geq 0)$ grows exponentially on the event of non-extinction, called the *Malthusian parameter*, is the only nonzero root of the convex function ψ. Furthermore, the probability of exit of an interval (from the bottom or from the top) by X has a simple expression (see e.g. [2]), in the form

$$P_s(T_0 < T_{(t,+\infty)}) = \frac{W(t-s)}{W(t)}, \tag{5}$$

where the so-called *scale function* W is the nonnegative, nondecreasing, differentiable function such that $W(0) = 1$, characterized by its Laplace transform

$$\int_0^\infty dx\, e^{-ax}\, W(x) = \frac{1}{\psi(a)} \qquad a > \eta. \tag{6}$$

As a consequence, the typical branch length H between two consecutive individuals alive at time t has the following distribution (conditional on there being at least two extant individuals at time t)

$$\mathbb{P}(H < s) = P_t(T_{(t,+\infty)} < T_s \mid T_{(t,+\infty)} < T_0) = \frac{1 - \frac{1}{W(s)}}{1 - \frac{1}{W(t)}} \qquad 0 \leq s \leq t. \tag{7}$$

Let us stress that in some examples, (6) can be inverted. When π has an exponential density, $(N_t; t \geq 0)$ is a linear birth–death process with (birth rate λ and) death rate, say ρ. If $\lambda \neq \rho$, then (see [22] for example)

$$W(x) = \frac{\rho - \lambda e^{(\lambda-\rho)x}}{\rho - \lambda} \qquad x \geq 0,$$

whereas if $\lambda = \rho$,

$$W(x) = 1 + \lambda x \qquad x \geq 0.$$

When π is a point mass at ∞, $(N_t; t \geq 0)$ is a pure-birth process, called Yule process, with birth rate λ. Then (let $\rho \to 0$)

$$W(x) = e^{\lambda x} \qquad x \geq 0.$$

In the case when $\lambda \neq \rho \neq 0$, it had already been noticed by B. Rannala [25] that the coalescence times of a population whose genealogy is given by a (linear) birth–death process started (singly) t units of time ago and whose size is conditioned to be n, are identical to those of the order statistics of n i.i.d. random variables with density

$$f(s) = \frac{(1 - p_0(s))(\rho - \lambda p_0(s))}{p_0(t)} \qquad 0 < s < t,$$

where ρ is the death rate and

$$p_0(t) := \frac{\rho\left(e^{rt} - 1\right)}{\lambda e^{rt} - \rho},$$

where $r := \lambda - \rho$. Now (7) applied to the expression of the scale function given previously for the birth–death case ($\lambda \neq \rho$) agrees with the findings of B. Rannala under the form

$$f(s)\; ds = \mathbb{P}(H \in ds) = \frac{r^2\; e^{rs}}{(\lambda e^{rs} - \rho)^2} \cdot \frac{\lambda e^{rt} - \rho}{e^{rt} - 1}\; ds \qquad 0 < s < t.$$

It is remarkable that in this case, exchanging λ and ρ leaves the distribution of H unchanged. No extension of this fact is known in the general case.

We end this section by the following lemma.

Lemma 3.1 *The one-dimensional marginal of N_t when the lifespan of the progenitor is random with law $\pi(\cdot)/\lambda$, is given by*

$$\mathbb{P}(N_t \neq 0) = \frac{W'(t)}{\lambda W(t)} \qquad t \geq 0$$

and

$$\mathbb{P}(N_t = k) = \left(1 - \frac{1}{W(t)}\right)^{k-1} \frac{W'(t)}{\lambda W(t)^2} \qquad t \geq 0.$$

Proof. From (1) and (5), we get

$$\mathbb{P}_s(N_t = 0) = \frac{W(t-s)}{W(t)}$$

and from (2) and (5), we get

$$\mathbb{P}_s(N_t = k \mid N_t \neq 0) = \left(1 - \frac{1}{W(t)}\right)^{k-1} \frac{1}{W(t)}.$$

Let us compute the unconditional law of N_t by integrating over s. First,

$$\mathbb{P}(N_t = 0) = \int_0^t \lambda^{-1}\pi(ds)\frac{W(t-s)}{W(t)} = \frac{F(t)}{\lambda W(t)},$$

where

$$F(t) := \int_0^t \pi(ds)W(t-s) \qquad t \geq 0.$$

Now by Fubini–Tonelli,

$$\int_0^\infty dt\, F(t)e^{-at} = \int_0^\infty \pi(ds) \int_s^\infty dte^{-at}W(t-s) = \frac{1}{\psi(a)}\int_0^\infty \pi(ds)e^{-as},$$

referring to (6), where we recall from (4) that

$$\psi(a) = a - \int_0^\infty \pi(dx)(1 - e^{-ax}) = a - \lambda + \int_0^\infty \pi(dx)e^{-ax} \qquad a \geq 0.$$

This yields

$$\int_0^\infty dt\, F(t)e^{-at} = 1 + \frac{\lambda - a}{\psi(a)}.$$

This Laplace transform can be inverted as follows

$$F(t) = \lambda W(t) - W'(t) \qquad t \geq 0.$$

Thus, we get the announced expression for N_t. □

4 The immigration model

Assume that we start at time 0 on the island with no individual at all. Let I_t denote the total number of extant individuals at time t. Let $I_t(k)$ denote the number of species (each corresponding to a single progenitor immigrant) with k representative individuals at time t. In particular,

$$I_t = \sum_{k \geq 1} kI_t(k).$$

We allow k to equal 0, $I_t(0)$ corresponding to the number of effective immigrants having 0 descendance at time t. Recall from the Preliminaries the scale function W.

Theorem 4.1 *The random variables $(I_t(0), I_t(1), \ldots)$ are independent Poisson random variables. For any $k \neq 0$, the r.v. $I_t(k)$ is a Poisson r.v. with parameter*

$$\frac{\gamma}{k}\left(1 - \frac{1}{W(t)}\right)^k,$$

where $\gamma := \mu/\lambda$ is the immigration-to-birth ratio. The Poisson r.v. $I_t(0)$ has parameter

$$\mu t - \gamma \ln W(t).$$

Thanks to a standard result on independent Poisson random variables X_k with respective means $c\alpha^k/k$, conditioning on $\sum kX_k$ removes the dependence in α (see e.g. [28, p.220]). It is then remarkable that conditioning the frequency spectrum on the total number of individuals *removes the dependence in t*. In the case of exponential lifetimes, this property has been rediscovered various times, see for example [24]. Here, the conditioning does *not only* remove the dependence in t, but also in λ, W or π, that is, *in the whole dynamical scheme distribution.*

Corollary 4.2 *Let $Y_1, Y_2, \ldots$ be independent random variables, where Y_k follows the Poisson distribution with parameter γ/k. Conditional on the total number I_t of species at time t equalling n, the random vector $(I_t(1), \ldots, I_t(n))$, then also denoted $(I_n(1), \ldots, I_n(n))$, has the same law as $(Y_1, \ldots, Y_n)$ conditioned by $\sum_{k=1}^n kY_k = n$.*

Remark 1 *This conditional spectrum is exactly the same one as that obtained in the Kingman coalescent with mutations at rate γ in the infinite-alleles model (i.e., the spectrum given by Ewens' sampling formula). In the case of exponential lifetimes, this coincidence between the binary branching process with immigration and the Moran process with mutations can be explained thanks to Hoppe's urn model (see [6]). This observation has been recast in the neutral theory of biodiversity literature as a possible relaxation of the 'zero-sum assumption' [7, 13].*

Remark 2 *Theorem 4.1 is concerned with species with fixed abundances $k = 1, 2, \ldots$, i.e., the 'small' families. It is also possible to get results for the abundances $P_1, P_2, \ldots$ of the immmigrant surviving families ranked by decreasing order of ages, i.e., the 'large' families, either as the population size $n \to \infty$ or as time $t \to \infty$ in the supercritical case (mean number of offspring $m > 1$). M. Richard [26] obtains that the vector $(P_1, P_2, \cdots)$ rescaled by population size converges a.s. to the GEM distribution with parameter γ.*

Let us now prove the theorem. Let M_t be the number of immigrants having reached the island up until time t, and $T_1 < \cdots < T_{M_t} < t$ the times of arrival of these immigrants. For any integer n, let σ_n denote an independent, random (uniform) permutation on $\{1, \ldots, n\}$. Then M_t is a Poisson r.v. with parameter μt, and conditional on $M_t = n$, the random variables $(T_{\sigma_n(1)}, \ldots, T_{\sigma_n(n)})$ are i.i.d., uniformly distributed on $[0, t]$. Then we call $Z_t^{(i)}$ the number of descendants at time t of the particle having immigrated at time $T_{\sigma_n(i)}$. The random variables $(Z_t^{(i)}, i = 1, \ldots, n)$ are i.i.d. distributed as some r.v. Z_t which is the value of the Crump–Mode–Jagers process N_t started at a uniform time on $[0, t]$

$$\mathbb{P}(Z_t^{(i)} = k) = \frac{1}{t}\int_0^t du \mathbb{P}(N_u = k),$$

where it will always be understood that $N_0 = 1$. The following statement is the key result to the theorem.

Proposition 4.3 *The law of Z_t is given by the following two equations.*

$$\mathbb{P}(Z_t = k) = \frac{1}{\lambda k t}\left(1 - \frac{1}{W(t)}\right)^k$$

for $k \neq 0$, whereas

$$\mathbb{P}(Z_t = 0) = 1 - \frac{1}{\lambda t}\ln W(t).$$

Before proving the proposition, we remind the reader of an elementary lemma on multinomial distributions with Poisson randomizing parameter. The theorem follows from this lemma and the proposition.

Lemma 4.4 *Let $p := (p_0, p_1, \dots)$ be some probability distribution on the integers, let $X_1, X_2, \dots$ be i.i.d. r.v. with law p and let B be an independent Poisson r.v. with parameter β. Finally, set*

$$B_k := \#\{i = 1, \dots, B : X_i = k\} \qquad k \geq 0.$$

Then the random variables $B_0, B_1, \dots$ are independent Poisson r.v., and B_k has parameter βp_k.

Proof of the proposition. Thanks to Lemma 3.1, we have

$$\mathbb{P}(N_t \neq 0) = \frac{W'(t)}{\lambda W(t)} \qquad t \geq 0,$$

and

$$\mathbb{P}(N_t = k) = \left(1 - \frac{1}{W(t)}\right)^{k-1} \frac{W'(t)}{\lambda W(t)^2} \qquad t \geq 0.$$

Let us now turn to Z_t, which has the law of N_t with origination time uniform on $[0,t]$. First,

$$\mathbb{P}(Z_t \neq 0) = \frac{1}{t}\int_0^t du \mathbb{P}(N_u \neq 0) = \frac{1}{t}\int_0^t du \frac{W'(u)}{\lambda W(u)} = \frac{1}{\lambda t}\ln W(t).$$

Second,

$$\mathbb{P}(Z_t = k) = \frac{1}{t}\int_0^t du \mathbb{P}(N_u = k) = \frac{1}{t}\int_0^t du \left(1 - \frac{1}{W(u)}\right)^{k-1} \frac{W'(u)}{\lambda W(u)^2} = \frac{1}{\lambda k t}\left(1 - \frac{1}{W(t)}\right)^k,$$

which ends the proof of the proposition. □

5 The mutation model

Recall from the section on splitting trees and coalescent point processes that the genealogy at a fixed time t of the N_t extant individuals of the splitting tree, originating from a single progenitor individual born at time 0, is characterized by the branch lengths H_i, $i = 1, \dots N_t - 1$, where H_i is the divergence time between individual i and individual $i+1$. In addition, these r.v. are i.i.d. with common distribution

$$\mathbb{P}(H < s) = \frac{1 - \frac{1}{W(s)}}{1 - \frac{1}{W(t)}} \qquad 0 \leq s \leq t,$$

where the so-called scale function W depends on the birth rate λ and on the lifespan measure π, and is characterized by its Laplace transform.

In the critical or supercritical cases, where W is unbounded, we can define the long-lived tree asymptotics, by letting $t \to \infty$. This leads to

$$\mathbb{P}(H < s) = 1 - \frac{1}{W(s)} \qquad s \geq 0,$$

and the *stationary* genealogy is then given by an infinite sequence of branches with i.i.d. lengths, with tail as in the last display. In the subcritical case, W has a finite limit equal to $1/(1-m)$ (see [22]). Then conditioning on the population being still extant at time t and letting $t \to \infty$, the *quasi-stationary* genealogy is given by a parameter m geometric number of branches with i.i.d. lengths distributed as follows

$$\mathbb{P}^{\star}(H < s) = m^{-1}\left(1 - \frac{1}{W(s)}\right) \qquad s \geq 0,$$

where the star superscript serves to remind the conditioning.

In this section, individuals experience mutations at rate θ during their lifetime, and each mutation yields a brand new type. This assumption corresponds to what is usually called the *infinitely-many alleles model*. We now introduce the function W_θ, which is the scale function associated to the so-called *clonal process*. More specifically, if one restricts the tree to points bearing the same type (e.g., the same type as the progenitor's type), then one retrieves a new splitting tree, whose birth rate remains equal to λ and whose lifetime durations are distributed as a r.v. V^θ defined as the minimum of V and of an independent exponential variable with parameter θ (i.e., the first mutation event). As in [21], we can then define H^θ as the divergence time between consecutive individuals in the clonal splitting tree. In the (more general) coalescent point process, H^θ is defined as the divergence time between individual 0 and the first individual whose type satisfies the following property: it is one of the successive types that appeared across time in the history of the lineage of individual 0. We have proved [21] that the function W_θ (either defined as the scale function of the clonal splitting tree or equivalently, in the coalescent point process, as the inverse of the tail of H^θ) satisfies

$$W_\theta(x) = 1 + \int_0^x W'(s)e^{-\theta s}\, ds \qquad x \geq 0. \tag{8}$$

Now consider the standing population at time t conditioned on being nonempty, whose probability law we denote by $\mathbb{P}^\star$. For any real number $y \in (0,t)$, define $A_t(k; dy)$ as the number of species originating in a point mutation having occurred during the time interval $(y, y+dy)$ and represented by exactly k alive individuals at time t. The following proposition gives the expectation under $\mathbb{P}^\star$ of $A_t(k; dy)$ and is extracted from [3].

Theorem 5.1 *For any $k \geq 1$, the expected number of species of age in dy and abundance k is*

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

In [3], we provide arguments giving an intuition of this result. To be more specific, the last expression can be seen as the product of the three following terms :

$$\theta\, dy\, \frac{W(t)}{W(y)}$$

which is the sum over $i = 1, 2 \dots$ of the probabilities that the i-th branch length has size $H_i \geq y$ and (is the one that) carries a mutation with age in $(y, y + dy)$, multiplied by

$$\frac{W(y)\, e^{-\theta y}}{W_\theta(y)}$$

which is the probability that the type carried by the lineage of the i-th individual at time $t - y$ has at least one alive representative, finally multiplied by

$$\frac{1}{W_\theta(y)} \left(1 - \frac{1}{W_\theta(y)}\right)^{k-1}$$

which is the probability that the type carried by the lineage of the i-th individual at time $t - y$ has exactly k alive representatives, conditional on having at least 1.

Recall that A_t denotes the number of species in the population at time t and that $A_t(k)$ denotes the number of species represented by exactly k extant individuals. We can record the last theorem under its integral representation :

Proposition 5.2 *For any $k \geq 1$,*

$$\mathbb{E}^\star A_t(k) = W(t) \int_0^t dy\, \theta\, e^{-\theta y} \frac{1}{W_\theta(y)^2} \left(1 - \frac{1}{W_\theta(y)}\right)^{k-1}$$

and

$$\mathbb{E}^\star A_t = W(t) \int_0^t dy\, \theta\, e^{-\theta y} \frac{1}{W_\theta(y)}.$$

Furthermore, we got the following asymptotic result, extracted from [3] and [21]. Here, $A_n(k)$ denotes the number of species with k individuals in the coalescent point process with population size n. Recall that coalescent point processes with different population sizes can be constructed on the same space by merely adding new independent branches. This allows us to state pathwise convergences for A_n as $n \to \infty$.

Theorem 5.3 *For all $k \geq 1$, the following convergence holds a.s., as $n \to \infty$ for the coalescent point process, and as $t \to \infty$ for the splitting tree in the supercritical case and on the event of non-extinction :*

$$\lim_{n\to\infty} n^{-1} A_n(k) = \lim_{t\to\infty} N_t^{-1} A_t(k) = \int_0^\infty dy\, \theta\, e^{-\theta y} \frac{1}{W_\theta(y)^2} \left(1 - \frac{1}{W_\theta(y)}\right)^{k-1}$$

and

$$\lim_{n\to\infty} n^{-1} A_n = \lim_{n\to\infty} N_t^{-1} A_t(k) = \int_0^\infty dy\, \theta\, e^{-\theta y} \frac{1}{W_\theta(y)}.$$

Remark 3 *The a.s. result for coalescent point processes relies on laws of large numbers (see [21]). The a.s. result for splitting trees relies on the theory of random characteristics (see [3]) introduced in the seminal paper [15] and further developed in [16, 17] and especially in [27].*

Remark 4 *As in the last section, one could ask about the behaviour of large families, as the number n of individuals grows. In contrast to the immigration case, here there are no families with abundances $O(n)$. Preliminary calculations [4] show that there are two possible regimes, depending on the respective positions of the mutation rate θ and of the Malthusian parameter η (see section on splitting trees). In the case when $\theta < \eta$ the abundance of the largest family is of order $O(n^\beta)$, where $\beta = 1 - \theta/\eta$, otherwise it is of order $O(\log(n))$.*

As in the previous section, we have displayed results holding for a general lifespan measure π. On the other hand, here the quantities displayed in the theorem can only be computed in the case of critical birth–death processes, that is, when the death rate of individuals is constant, equal to their birth rate λ, so that $W(x) = 1 + \lambda x$. In that case, $W'_\theta(x) = \lambda e^{-\theta x}$, and we can integrate the quantities in the theorem.

Corollary 5.4 *In the case of a critical birth–death process with birth and death rate λ,*

$$\lim_{n\to\infty} n^{-1} A_n(k) = (\alpha^{-1} - 1)\frac{\alpha^k}{k} \qquad a.s.,$$

where

$$\alpha := \frac{\lambda}{\lambda + \theta}.$$

In addition,

$$\lim_{n\to\infty} n^{-1} A_n = -(\alpha^{-1} - 1)\ln(1 - \alpha) \qquad a.s.$$

References

[1] Athreya, K.B., Ney, P.E. (1972)
Branching processes. Springer-Verlag, New York.

[2] Bertoin, J. (1996)
Lévy processes. Cambridge University Press, Cambridge.

[3] Champagnat, N., Lambert, A. (2010)
Splitting trees with neutral Poissonian mutations I: Small families. Submitted.

[4] Champagnat, N., Lambert, A. (2010)
Splitting trees with neutral Poissonian mutations II: Large families. In preparation.

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

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

[7] Etienne, R.S., Alonso, D., McKane, A.J. (2007)
The zero-sum assumption in neutral biodiversity theory. *J. Theoret. Biol.* **248** 522–536.

[8] Ewens, W.J. (1972)
The sampling theory of selectively neutral alleles. *Theoret. Popul. Biol.* **3** 87–112, and erratum, p.376.

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

[10] Fisher, R.A. (1943)
A theoretical distribution for the apparent abundance of different species. *J. Anim. Ecol.* **12** 54–58.

[11] Fisher, R.A., Corbet, S.A., Williams, C.B. (1943)
The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12** 42–58.

[12] 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.

[13] Haegeman, B., Etienne, R.S. (2008)
Relaxing the zero-sum assumption in neutral biodiversity theory. *J. Theoret. Biol.* **252** 288–294.

[14] Hubbell, S.P. (2001)
The Unified Neutral Theory of Biodiversity and Biogeography. Princeton U. Press, NJ.

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

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

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

[18] Karlin, S., McGregor (1967)
The number of mutant forms maintained in a population. *Proc. 5th Berkeley Symposium Math. Statist. Prob.* **IV** 415–438.

[19] Kendall, D.G. (1948)
On some modes of population growth leading to R.A. Fisher's logarithmic series distribution. *Biometrika* **35** 6–15.

[20] Kimura, M., Crow, J.F. (1964)
The number of alleles that can be maintained in a finite population. *Genetics* **49** 725–738.

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

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

[23] MacArthur, R.H., Wilson, E.O. (1967)
The Theory of Island Biogeography. Princeton U. Press, NJ.

[24] Rannala, B. (1996)
The sampling theory of neutral alleles in an island population of fluctuating size. *Theoret. Popul. Biol.* **50** 91–104.

[25] Rannala, B. (1997)
Gene genealogy in a population of variable size. *Heredity* **78** 417—423.

[26] Richard, M. (2010)
Limit theorems for splitting trees with structured immigration and applications to biogeography. Submitted.

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

[28] Watterson, G.A. (1974)
Models for the logarithmic species abundance distributions. *Theoret. Popul. Biol.* **6** 217–250.