Applied Probability Trust (10 June 2003)

SUPERCRITICAL MULTITYPE BRANCHING PROCESSES: THE ANCESTRAL TYPES OF TYPICAL INDIVIDUALS

HANS-OTTO GEORGII,* *Universität München*

ELLEN BAAKE,** *Universität Greifswald*

Abstract

For supercritical multitype Markov branching processes in continuous time, we investigate the evolution of types along those lineages that survive up to some time t. We establish almost-sure convergence theorems for both time and population averages of ancestral types (conditioned on non-extinction), and identify the mutation process describing the type evolution along typical lineages. An important tool is a representation of the family tree in terms of a suitable size-biased tree with trunk. As a by-product, this representation allows a 'conceptual proof' (in the sense of [19]) of the continuous-time version of the Kesten-Stigum theorem.

Keywords: Multitype branching process; type history; ancestral distribution; size-biased tree; empirical process; large deviations; Kesten-Stigum theorem

AMS 2000 Subject Classification: Primary 60J80
Secondary 60F10

1. Introduction

Looking at the time evolution of a population one has two possible perspectives: either forward or backward in time. In the first case one observes the characteristics of the population at a given time t and asks for its behaviour as t increases to infinity. A classical model that describes the unrestricted reproduction of independent individuals is the (multitype) branching process, and a principal result in the supercritical case is the Kesten-Stigum theorem [16], which describes the population size and relative frequencies of types; see Theorem 2.1 for the precise statement. A different situation arises if the population size is kept constant; this leads to certain interacting particle systems, like the Moran model and its relatives (for review, see [7]). By way of contrast, the backwards – or retrospective – aspect of the population concerns the lineages extending back into the past from the presently living individuals and asks for the characteristics of the ancestors along such lineages. One famous example is Kingman's coalescent (see [17, 18], and [22] for a review), the backward version of the Moran model. As was observed e.g. by Jagers [14] and Jagers and Nerman [15], it is also rewarding to study the backward aspects of multitype branching processes; this point of view has turned out as crucial in recent biological applications [11]. It is the aim

* Postal address: Institut für Mathematik, Universität München, Theresienstr. 39, D-80333 München, Germany, email: georgii@mathematik.uni-muenchen.de
** Postal address: Institut für Mathematik und Informatik, Universität Greifswald, Jahnstr. 15a, D-17487 Greifswald, Germany, email: ellen.baake@uni-greifswald.de

of this article to pursue this last line of research further. We do so in continuous time because this gives us the opportunity to transfer some powerful methods recently developed for discrete time. We also concentrate on the supercritical case.

Specifically, we consider the individuals alive at some time t and investigate the types of their ancestors at an earlier time, $t-u$. We will show the following.

- When t resp. t and u tend to infinity, both time average and population average of ancestral types converge to a particular distribution α almost surely on non-extinction (Theorems 3.1 and 3.2).

This α will be called the *ancestral distribution of types*; its components are $\alpha_i = \pi_i h_i$, where π and h are the (properly normalized) left and right Perron-Frobenius eigenvectors of the generator of the first-moment matrix.

More detailed information about the evolution of types along ancestral lineages is obtained through what we would like to call the *retrospective mutation chain*, a particular continuous-time Markov chain on the type space with α as its invariant distribution. We will show:

- For all individuals alive at time t up to an asymptotically negligible fraction, the time averaged empirical type evolution process tends in distribution to the stationary retrospective mutation chain, in the limit as $t \to \infty$, almost surely on non-extinction (Theorem 3.3).

One basic ingredient of our reasoning is a law of large numbers for population averages; see Proposition 5.1. A second crucial ingredient is a representation of the family tree in terms of a size-biased tree with trunk (with the retrospective mutation chain running along the trunk); see Theorem 4.1. This representation is the continuous-time analogue of the size-biased tree representation introduced by Lyons, Pemantle and Peres [20] and Kurtz, Lyons, Pemantle and Peres [19]. In passing, it allows us to extend their conceptual proof of the Kesten-Stigum theorem to continuous time. The third ingredient is the Donsker-Varadhan large deviation principle for the retrospective mutation chain [5, 6]. This implies a large deviation principle for the typical type evolution along the surviving lineages in the tree – see Theorem 5.1.

This paper is organized as follows. In the next section we recall the construction of the family tree for multitype branching processes in continuous time. Section 3 contains the precise statement of results. Section 4 is devoted to the size-biased tree with trunk, and the proofs of the main results are collected in Section 5.

2. The branching process and basic facts

We consider a continuous-time multitype branching process as described in Athreya and Ney [2, Ch. V.7]. To fix the notation we recall the basic setting here.

Let S be a finite set of types. An individual of type $i \in S$ lives for an exponential time with parameter $a_i > 0$, and then splits into a random offspring $N_i = (N_{ij})_{j\in S}$ with distribution p_i on $\mathbb{Z}_+^S$ and finite means $m_{ij} := \mathbb{E}(N_{ij})$ for all $i, j \in S$; here, N_{ij} is the number of j-children, and $\mathbb{Z}_+ = \{0, 1, \ldots\}$. We assume that the mean offspring matrix $\mathsf{M} = (m_{ij})_{i,j\in S}$ is irreducible.

According to Harris [10, Ch. VI], the associated random family tree can be constructed as follows. Let $\mathbb{X} = \bigcup_{n\geq 0} \mathbb{X}_n$, where $\mathbb{X}_n$ describes the virtual n'th generation. That is, $\mathbb{X}_0 = S$, and $i_0 \in \mathbb{X}_0$ specifies the type of the root, i.e., the founding ancestor.

Next, $\mathbb{X}_1 = S \times \mathbb{N}$, and the element $x = (i_1, \ell_1) \in \mathbb{X}_1$ is the ℓ_1'th i_1-child of the root. Finally, for $n > 1$, $\mathbb{X}_n = S^n \times \mathbb{N}^n$, and $x = (i_1, \ldots, i_n; \ell_1, \ldots, \ell_n) \in \mathbb{X}_n$ is the ℓ_n-th i_n-child of its parent $\tilde{x} = (i_1, \ldots, i_{n-1}; \ell_1, \ldots, \ell_{n-1})$; see Fig. 1. We write $\sigma(x) = i_n$ for the type of $x \in \mathbb{X}_n$. With each $x \in \mathbb{X}$ we associate

- its random life time τ_x, distributed exponentially with parameter $a_{\sigma(x)}$, and
- its random offspring $N_x = (N_{x,j})_{j \in S} \in \mathbb{Z}_+^S$ with distribution $\mathsf{p}_{\sigma(x)}$

such that the family $\{\tau_x, N_x : x \in \mathbb{X}\}$ is independent.

The random variables N_x indicate which of the virtual individuals $x \in \mathbb{X}$ are actually realized, namely those in the random set $X = \bigcup_{n \geq 0} X_n$ defined recursively by

$$X_0 = \{i_0\}, \quad X_n = \{x = (\tilde{x}; i_n, \ell_n) \in \mathbb{X}_n : \tilde{x} \in X_{n-1}, \ell_n \leq N_{\tilde{x}, i_n}\},$$

where i_0 is the prescribed type of the root. The random variables τ_x provide the proper time scale. Namely, for $x \in X$, let the splitting times T_x be defined recursively by $T_x = T_{\tilde{x}} + \tau_x$ with $T_{\tilde{i}_0} := 0$. The lifetime interval of $x \in X$ is then $[T_{\tilde{x}}, T_x[$. Hence $X(t) = \{x \in X : T_{\tilde{x}} \leq t < T_x\}$ is the population at time t. One may visualize the resulting tree by identifying each $x \in X$ with an edge from $\tilde{x}$ to x with length τ_x in the direction of time.

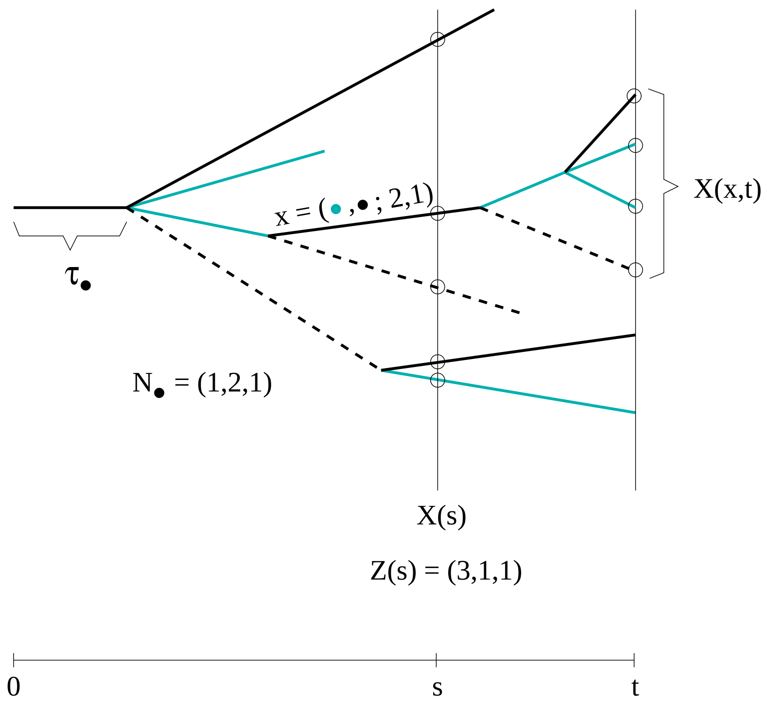

FIGURE 1: A realization of the branching process. Types are indicated by different line types, indexed in the order (black, grey, dashed), counted from top to bottom, and symbolized by filled circles. The set $X(s)$ consists of all edges that intersect the vertical line at s; the set $X(x,t)$ consists of all edges that emanate from x and hit the vertical line at t. $Z(s)$ counts the type frequencies in the population $X(s)$.

The family tree is completely determined by the process $X[0,\infty[\; := \big(X(t)\big)_{t\geq 0}$ which is a random element of $\Omega := D\big([0,\infty[, \mathfrak{P}_f(\mathbb{X})\big)$, the Skorohod space of all càdlàg functions on $[0,\infty[$ taking values in the (countable) set $\mathfrak{P}_f(\mathbb{X})$ of all finite subsets of $\mathbb{X}$. We write $\mathbb{P}^i$ for the distribution of $X[0,\infty[$ on Ω when the type of the root is $i_0 = i$, and $\mathbb{E}^i$ for the associated expectation. If i_0 is chosen randomly with distribution ν, we write $\mathbb{P}^\nu$ and $\mathbb{E}^\nu$. We will often identify $X[0,\infty[$ with the canonical process on Ω.

For $0 < s < t$ and $y \in X(t)$ we write $y(s)$ for its unique ancestor living at time s. On the other hand, for $x \in X(s)$ we let

$$X(x,t) = \big\{y \in \mathbb{X} : xy \in X(t)\big\} \tag{2.1}$$

denote the set of descendants of x living at time t; cf. Fig. 1. In the above, the concatenation xy of two strings $x, y \in \mathbb{X}$ is defined in the obvious way, and the empty string is considered as an ancestor of type $\sigma(x)$; i.e., $X(x,t) = \{\sigma(x)\}$ as long as $x \in X(t)$. By the loss-of-memory property of the exponential distributions, the descendant trees $X(x,[s,\infty[) = (X(x,t))_{t\geq s}$ with $x \in X(s)$ are conditionally independent given $X[0,s]$, with distribution $\mathbb{P}^{\sigma(x)}$. We will also consider the counting measures

$$Z(t) = \sum_{x\in X(t)} \delta_{\sigma(x)}\,, \quad Z(x,t) = \sum_{y\in X(x,t)} \delta_{\sigma(y)} \tag{2.2}$$

on S, where δ_i is the Dirac measure at i. $Z(t)$ and $Z(x,t)$ count the type frequencies in the population $X(t)$ resp. the subpopulation $X(x,t)$ of x-descendants. In particular, $Z_j(t)$ is the cardinality of $X_j(t) = \{x \in X(t) : \sigma(x) = j\}$, the subpopulation of type $j \in S$, and $\|Z(t)\| := \sum_{j\in S} Z_j(t) = |X(t)|$ is the total size of the population.

It is well-known (cf. [2], p. 202, Eq. 9) that $\mathbb{E}^i(Z_j(t)) = (e^{t\mathsf{A}})_{ij}$ for all $i,j \in S$, where the generator matrix $\mathsf{A} = (a_{ij})_{i,j\in S}$ is given by

$$a_{ij} = a_i(m_{ij} - \delta_{ij})\,. \tag{2.3}$$

By the irreducibility of M, A is also irreducible, so that the first moment matrix $(\mathbb{E}^i(Z_j(t)))_{i,j\in S}$ has positive entries for any $t > 0$. (This property is often called 'positive regularity', see [2, p. 202].) Perron-Frobenius theory then tells us that the matrix A has a principal eigenvalue λ (i.e., a real eigenvalue exceeding the real parts of all other eigenvalues), and associated positive left and right eigenvectors π and h which will be normalized s.t. $\langle \pi, 1\rangle = 1 = \langle \pi, h\rangle$. Here we think of the row vector π as a probability measure, of the column vectors h and $1 = (1,\ldots,1)^T$ as functions on S, and of the scalar product $\langle \pi, h\rangle = \sum_i \pi_i h_i$ as the associated expectation. We are mainly interested in the supercritical case $\lambda > 0$. In this case we write

$$\Omega_{\text{surv}} := \{X(t) \neq \varnothing \text{ for all } t > 0\}$$

for the event that the population survives for all times.

It is a remarkable fact that the almost-sure behaviour of the family tree is, to a large extent, already determined by the the global quantities λ, π, h. One prominent example is the following continuous-time version of the Kesten-Stigum theorem (see [16] for the discrete-time original, [1] for the continuous-time version, and [19] for the recent discrete-time conceptual proof).

Theorem 2.1. (Kesten-Stigum.) *Consider the supercritical case $\lambda > 0$.*

(a) *For all $i \in S$ we have*

$$\frac{1}{|X(t)|} \sum_{x \in X(t)} \delta_{\sigma(x)} = \frac{Z(t)}{\|Z(t)\|} \underset{t\to\infty}{\longrightarrow} \pi \quad \mathbb{P}^i\text{-almost surely on } \Omega_{\mathrm{surv}}.$$

(b) *There is a nonnegative random variable W such that*

$$\lim_{t\to\infty} Z(t)\, e^{-\lambda t} = W\,\pi \quad \mathbb{P}^i\text{-almost surely for any } i \in S\,,$$

and $\mathbb{P}^i(W > 0) > 0$ for all i if and only if

$$\mathbb{E}(N_{ij} \log N_{ij}) < \infty \quad \text{for all } i,j \in S. \tag{2.4}$$

In this case, $\{W > 0\} = \Omega_{\mathrm{surv}}$ $\mathbb{P}^i$-almost surely, and $h_i = \mathbb{E}^i(W)$.

For the sake of reference we provide here a full proof extending the conceptual discrete-time proof of [19] to our continuous-time setting. Assertion (a) reveals that the left eigenvector π holds the asymptotic proportions of the types in the population, and statement (b) implies that

$$\lambda = \lim_{t\to\infty} \frac{1}{t} \log |X(t)|$$

is the almost sure exponential growth rate of the population in the case of survival. In fact, this statement does not require condition (2.4); see the proof of Theorem 3.3 in Section 5.3. The i-th coordinate h_i of the right eigenvector h measures the long-term fertility of an i-individual. In fact, h_i is also characterized by the limiting relation

$$\mathbb{E}^i\big(|X(t)|\big)\, e^{-\lambda t} \to h_i \quad \text{as } t \to \infty; \tag{2.5}$$

cf. Remark 4.1(a) below.

3. Results

We still consider the supercritical case $\lambda > 0$. We are interested in the mutation behaviour of the population tree. More specifically, we ask for the behaviour of the sequence of types along a typical branch of this tree. It turns out that this behaviour is again completely determined by the global quantities λ, π, h. A key role is played by the probability vector $\alpha = (\alpha_i)_{i\in S}$ with components $\alpha_i = \pi_i h_i$. As observed by Jagers [14, Corollary 1], Jagers and Nerman [15, Prop. 1], and Hermisson et al. [11], this probability vector describes the distribution of ancestral types of an equilibrium population with type frequencies given by π. The vector α will therefore be called the *ancestral distribution.* Our results below shed some additional light on the significance of α.

To begin, we consider a typical individual $x \in X(t)$ alive at some large time t and ask for the type $\sigma(x(t-u))$ of its ancestor $x(t-u)$ living at some earlier time $t-u$. We find that $\sigma(x(t-u))$ is asymptotically distributed according to α. Specifically, let $0 < u < t$ and

$$A^u(t) = \frac{1}{|X(t)|} \sum_{x \in X(t)} \delta_{\sigma(x(t-u))} \tag{3.1}$$

be the empirical ancestral type distribution at time $t-u$ taken over the population $X(t)$. (Of course, this definition requires that $X(t) \neq \varnothing$.)

Theorem 3.1. (Population average of ancestral types.) *Let $\lambda > 0$ and $i \in S$. Then*

$$\lim_{u\to\infty} \lim_{t\to\infty} A^u(t) = \alpha \quad \mathbb{P}^i\text{-almost surely on } \Omega_{\mathrm{surv}}. \tag{3.2}$$

The proof will be given in Section 5.2. We would like to remark that a slightly weaker result under slightly stronger conditions (convergence in probability under assumption (2.4)) follows immediately from Corollary 4 of Jagers and Nerman [15], where very general population averages are considered.

Remark 3.1. Assertion (3.2) means that, for each $j \in S$, the average

$$A^u_j(t) = \frac{1}{|X(t)|} \sum_{x\in X(t)} I\{\sigma(x(t-u)) = j\}$$

(with $I\{.\}$ denoting the indicator function) converges to α_j $\mathbb{P}^i$-almost surely on Ω_{surv} as $t \to \infty$ and $u \to \infty$ in this order. Letting $s = t-u$, we can rewrite this average in the form

$$\sum_{x\in X_j(s)} |X(x,t)| \Big/ \sum_{x\in X(s)} |X(x,t)| \,,$$

where $X(x,t)$ is given by (2.1). The numbers $|X(x,t)|$ with $x \in X_j(s)$ are i.i.d. with mean $\mathbb{E}^j(|X(u)|)$. Assuming the validity of a law of large numbers and using Theorem 2.1(a) and Eq. (2.5), we can conclude that the average above converges to $\pi_j h_j / \langle \pi, h\rangle = \alpha_j$ as $s, u \to \infty$. This explains the particular structure of the ancestral distribution α.

In our next theorem we ask for the time average of types along the line of descent leading to a typical $x \in X(t)$. This time average is given by the empirical distribution

$$L^x(t) = \frac{1}{t} \int_0^t \delta_{\sigma(x(s))}\, ds$$

of the process $\sigma(x[0,t]) = \big(\sigma(x(s))\big)_{0\le s\le t}$. Note that $L^x(t)$ belongs to the simplex $\mathcal{P}(S)$ of all probability vectors on S; $\mathcal{P}(S)$ will be equipped with the usual total variation distance $\|\cdot\|$. To describe the behaviour of $L^x(t)$ for a typical $x \in X(t)$ we have to step one level higher and to consider the empirical distribution of $L^x(t)$ taken over the population $x \in X(t)$. This empirical distribution belongs to $\mathcal{P}(\mathcal{P}(S))$, the set of probability measures on $\mathcal{P}(S)$, which will be equipped with the weak topology.

Theorem 3.2. (Time average of ancestral types.) *Let $\lambda > 0$ and $i \in S$. Then*

$$\lim_{t\to\infty} \frac{1}{|X(t)|} \sum_{x\in X(t)} \delta_{L^x(t)} = \delta_\alpha \quad \mathbb{P}^i\text{-almost surely on } \Omega_{\mathrm{surv}}. \tag{3.3}$$

Remark 3.2. (a) According to the portmanteau theorem [8, p. 108, Th. 3.1], statement (3.3) is equivalent to the assertion that

$$\lim_{t\to\infty} \frac{1}{|X(t)|} \sum_{x\in X(t)} I\{\, L^x(t) \in F\} = 0 \quad \text{for each closed } F \subset \mathcal{P}(S) \text{ with } \alpha \notin F$$

$\mathbb{P}^i$-almost surely on Ω_{surv}, and it is sufficient to check this in the case when $F = \{\nu \in \mathcal{P}(S) : \|\nu - \alpha\| \geq \varepsilon\}$ with arbitrary $\varepsilon > 0$. The theorem therefore asserts that, for all individuals $x \in X(t)$ up to an asymptotically negligible fraction, the ancestral type average $L^x(t)$ is close to α.

(b) Theorem 3.2 involves a population average of time averages. So one may ask whether the averaging of population and time can be interchanged. It follows from Theorem 3.1 that this is indeed the case:

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \delta_{A^u(t)}\, du = \delta_\alpha$$

almost surely on Ω_{surv}.

Theorem 3.2 is in fact a corollary of our next theorem which considers the complete mutation history along a typical line of descent. To state this result we need some preparations. We introduce first the mutation process on S which will turn out to describe the time-averaged mutation behaviour along an ancestral line.

Definition: The *retrospective mutation chain* is the Markov chain $(\sigma(t))_{t\geq 0}$ on S which stays in a state $i \in S$ for an exponential holding time with parameter $a_i + \lambda$ and then jumps to $j \in S$ with probability

$$p_{ij} = \frac{m_{ij}\, h_j}{(1+\lambda/a_i)h_i}\,.$$

That is, the generator $\mathsf{G} = (g_{ij})_{i,j\in S}$ of $(\sigma(t))_{t\geq 0}$ is given by

$$g_{ij} = (a_i+\lambda)(p_{ij} - \delta_{ij}) = h_i^{-1}(a_{ij} - \lambda\delta_{ij})h_j.$$

We note that G is indeed a generator because $a_i \sum_{j\in S} m_{ij}h_j = \sum_{j\in S}(a_i\delta_{ij} + a_{ij})h_j = (a_i+\lambda)h_i$ by (2.3). Since M is irreducible by assumption, G is irreducible as well. It is also immediate that the ancestral distribution α is the (unique) stationary distribution of G. The retrospective mutation chain was identified by Jagers [13, p. 195] and may be interpreted as the forward version of the backward Markov chain [15, Proposition 1] that results from picking individuals randomly from the stationary type distribution π and following their lines of descent backward in time. This gives the transition rates

$$\bar{g}_{ij} = \pi_j(a_{ji} - \lambda\delta_{ij})\pi_i^{-1} = \alpha_j g_{ji} \alpha_i^{-1}\,, \tag{3.4}$$

which corresponds to the time reversal of the retrospective mutation chain.

To set up the stage for Theorem 3.3 we let $\Sigma = D(\mathbb{R}, S)$ denote the space of all doubly infinite càdlàg paths in S. Σ will be equipped with the usual Skorohod topology which turns Σ into a Polish space; see e.g. [8], Section 3.5 and in particular Th. 5.6, for the case of the time interval $[0,\infty[$. The associated Borel σ-algebra coincides with the σ-algebra generated by the evaluation maps $\Sigma \ni \sigma \to \sigma(t)$, $t \in \mathbb{R}$ [8, p. 127, Prop. 7.1]. The time shift ϑ_s on Σ is defined by

$$\vartheta_s\sigma(t) = \sigma(t+s)\,, \qquad s,t \in \mathbb{R},\ \sigma \in \Sigma.$$

We write $\mathcal{P}_\Theta(\Sigma)$ for the set of all probability measures on Σ which are invariant under the shift group $\Theta = (\vartheta_s)_{s\in\mathbb{R}}$. Endowed with the weak topology, $\mathcal{P}_\Theta(\Sigma)$ is a Polish space [8, p. 101, Th. 1.7].

Next we introduce the time-averaged type evolution process of an individual in the population tree. For $t > 0$ and $x \in X(t)$ we let $\sigma(x)_{t,\text{per}} \in \Sigma$ be defined by

$$\sigma(x)_{t,\text{per}}(s) = \sigma(x(s_t)), \quad s \in \mathbb{R}, \tag{3.5}$$

where s_t is the unique number in $[0,t[$ with $s \equiv s_t \mod t$. That is, $\sigma(x)_{t,\text{per}} \in \Sigma$ is the periodically continued type history of x up to time t. The time-averaged type evolution of x is then described by the *empirical type evolution process*

$$R^x(t) = \frac{1}{t}\int_0^t \delta_{\vartheta_s\sigma(x)_{t,\text{per}}}\, ds \;\in \mathcal{P}_\Theta(\Sigma). \tag{3.6}$$

We are interested in the typical behaviour of $R^x(t)$ when x is picked at random from $X(t)$, the population at time t. This is captured in their empirical distribution, i.e., the population average

$$\Gamma(t) := \frac{1}{|X(t)|}\sum_{x\in X(t)} \delta_{R^x(t)}\,. \tag{3.7}$$

(As before, this definition requires that $X(t) \neq \varnothing$.) $\Gamma(t)$ is a random element of $\mathcal{P}(\mathcal{P}_\Theta(\Sigma))$, the set of all probability measures on the Polish space $\mathcal{P}_\Theta(\Sigma)$, which is again equipped with the weak topology. In Section 5.3 we will prove:

Theorem 3.3. (Typical ancestral type evolution.) *Let $\lambda > 0$ and $i \in S$. Then*

$$\lim_{t\to\infty}\Gamma(t) = \delta_{\boldsymbol{\mu}} \quad \mathbb{P}^i\textit{-almost surely on } \Omega_{\text{surv}}, \tag{3.8}$$

where $\boldsymbol{\mu} \in \mathcal{P}_\Theta(\Sigma)$ is the distribution of the stationary (doubly infinite) retrospective mutation chain $(\sigma(t))_{t\in\mathbb{R}}$ with generator G and invariant distribution α.

Remark 3.3. As in Remark 3.2(a), the portmanteau theorem implies that (3.8) is equivalent to the assertion that, $\mathbb{P}^i$-almost surely on Ω_{surv}, $\Gamma(t)(F) \to 0$ for every closed $F \subset \mathcal{P}_\Theta(\Sigma)$ such that $\boldsymbol{\mu} \notin F$. Writing $d(\cdot,\cdot)$ for any metric metrizing the weak topology on $\mathcal{P}_\Theta(\Sigma)$ this in turn means that, for each $\varepsilon > 0$,

$$\lim_{t\to\infty}\frac{1}{|X(t)|}\sum_{x\in X(t)} I\big\{d(R^x(t),\boldsymbol{\mu}) \geq \varepsilon\big\} = 0$$

$\mathbb{P}^i$-almost surely on Ω_{surv}. The theorem therefore states that, for all individuals $x \in X(t)$ up to an asymptotically negligible fraction, the time-averaged ancestral type evolution process $R^x(t)$ is close to $\boldsymbol{\mu}$ in the weak topology. Theorem 3.3 also highlights the restrospective nature of our mutation chain: it describes the evolution of types along those lines of descent which survive until time t (and thus can be seen when a time-t individual looks back into the past).

4. Size-biasing of the family tree

In this section we construct a continuous-time version of the size-biased multitype Galton-Watson tree as introduced by Lyons, Pemantle, Peres, and Kurtz [20, 19]. Informally, this is a tree with a randomly selected trunk (or spine) along which time runs at a different rate and offspring is weighted according to its size; in particular, there is always at least one offspring along the trunk so that the trunk survives forever. The children off the trunk get ordinary (unbiased) descendant trees (the bushes). It will turn out that the trunk of the size-biased tree describes the evolution along a typical ancestral line that survives up to some fixed time. The construction is not confined to the supercritical case; that is, in this section λ can have arbitrary sign.

First of all, for each type $i \in S$ we introduce the size-biased offspring distribution

$$\widehat{\mathsf{p}}_i(\kappa) = \frac{\langle \kappa, h\rangle \, \mathsf{p}_i(\kappa)}{c_i \, h_i}\,, \qquad \kappa \in \mathbb{Z}_+^S, \tag{4.1}$$

where $\langle \kappa, h\rangle = \sum_j \kappa_j h_j$ and $c_i = 1 + \lambda/a_i$ is a normalizing constant. $\widehat{\mathsf{p}}_i$ will serve as the offspring distribution of an i-individual on the trunk; it is indeed a probability distribution since

$$\sum_{\kappa\in\mathbb{Z}_+^S} \langle \kappa, h\rangle \, \mathsf{p}_i(\kappa) = \sum_{j\in S} m_{ij} h_j = \sum_{j\in S} (\delta_{ij} + a_{ij}/a_i) h_j = c_i \, h_i$$

by (2.3); note that c_i is automatically positive. Next, when an i-individual on the trunk has offspring $\widehat{N}_i = (\widehat{N}_{ij})_{j\in S}$ with distribution $\widehat{\mathsf{p}}_i$, one of these offspring is chosen as the successor on the trunk, where children are picked with probability proportional to h_j when their type is j. That is, the successor is of type j with probability $\widehat{N}_{ij}\; h_j/\langle \widehat{N}_i, h\rangle$ for a given offspring, and with probability

$$p_{ij} = \mathbb{E}\Big(\frac{\widehat{N}_{ij}\; h_j}{\langle \widehat{N}_i, h\rangle}\Big) = \frac{m_{ij}\, h_j}{c_i \, h_i}$$

on average. *These are precisely the jump probabilities of the retrospective mutation chain.* Finally, the lifetime of an i-individual on the trunk will be exponential with parameter $a_i{+}\lambda$, which coincides again with the holding time parameter of the retrospective mutation chain. A corresponding embedded chain combined with size-biased waiting times also occurs when more general non-Markovian populations (i.e., with waiting times deviating from the exponential distribution) are traced backwards, see [15, Proposition 1].

We now construct the size-biased tree in detail. Let $\{\tau_x, N_x : x \in \mathbb{X}\}$ be as in Section 2 and, independently of this, a sequence $\{\widehat{\tau}_n, \widehat{N}_n, \xi_n : n \geq 0\}$ of random variables with values in $]0,\infty[\,, \mathbb{Z}_+^S, \mathbb{X}$ respectively such that, for a given type $i_0 = i$ of the root, $\xi_0 = i$ and

- $\widehat{\tau}_0, \widehat{N}_0$ are independent, $\widehat{\tau}_0$ has exponential distribution with parameter $a_i{+}\lambda$, $\widehat{N}_0$ has distribution $\widehat{\mathsf{p}}_i$, and ξ_1 has conditional distribution

$$P\big(\xi_1 = (i_1, \ell_1) | \widehat{N}_0, \widehat{\tau}_0\big) = \frac{h_{i_1}}{\langle \widehat{N}_0, h\rangle}\, I\{\ell_1 \leq \widehat{N}_{0,i_1}\}$$

  for all $(i_1, \ell_1) \in \mathbb{X}_1$.

- For any $n \geq 1$, conditionally on $\mathbb{F}_{n-1} = \sigma\{\widehat{\tau}_k, \widehat{N}_k, \xi_{k+1} : k < n\}$, $\widehat{\tau}_n, \widehat{N}_n$ are independent and follow an exponential law with parameter $a_{\sigma(\xi_n)}+\lambda$ resp. the law $\widehat{\mathsf{p}}_{\sigma(\xi_n)}$, and

$$P\big(\xi_{n+1} = (\xi_n; i_{n+1}, \ell_{n+1})|\mathbb{F}_{n-1}, \widehat{\tau}_n, \widehat{N}_n\big) = \frac{h_{i_{n+1}}}{\langle \widehat{N}_n, h\rangle} I\{\ell_{n+1} \leq \widehat{N}_{n,i_{n+1}}\}$$

for all $(i_{n+1}, \ell_{n+1}) \in S \times \mathbb{N}$, i.e., ξ_{n+1} is a child of ξ_n selected randomly with weight proportional to $h_{\sigma(\xi_{n+1})}$.

Define $\widehat{X} = \bigcup_{n\geq 0} \widehat{X}_n \subset \mathbb{X}$ recursively by $\widehat{X}_0 = \{i\}$ and $\widehat{X}_n = \widehat{X}_n^\sharp \cup \widehat{X}_n^\flat$ with

$$\widehat{X}_n^\sharp = \{(\xi_{n-1}; i_n, \ell_n) \in \mathbb{X}_n : \ell_n \leq \widehat{N}_{n-1,i_n}\},$$

the offspring of ξ_{n-1}, and

$$\widehat{X}_n^\flat = \{(\tilde{x}; i_n, \ell_n) \in \mathbb{X}_n : \tilde{x} \in \widehat{X}_{n-1} \setminus \{\xi_{n-1}\}, \ell_n \leq N_{\tilde{x},i_n}\}$$

the offspring of all other individuals in $\widehat{X}_{n-1}$. (Note that in the last display there is no hat on N; that is, the bushes have unbiased offspring.) The split times $\widehat{T}_x$ are given by $\widehat{T}_{\xi_0} = \widehat{\tau}_0$, $\widehat{T}_{\xi_n} = \widehat{T}_{\xi_{n-1}} + \widehat{\tau}_n$ for $n \geq 1$, and $\widehat{T}_x = \widehat{T}_{\tilde{x}} + \tau_x$ if $x \in \widehat{X} \setminus \{\xi_n : n \geq 0\}$. (Again, in the latter case there is no hat on τ, meaning that the individuals off the trunk have unbiased life times.) The total population at time t is then given by

$$\widehat{X}(t) = \{x \in \widehat{X} : \widehat{T}_{\tilde{x}} \leq t < \widehat{T}_x\}.$$

The selected trunk individual at time t is $\xi(t) = \xi_n$ if $\widehat{T}_{\xi_{n-1}} \leq t < \widehat{T}_{\xi_n}$, and the process $\big(\widehat{X}(t), \xi(t)\big)_{t\geq 0}$ in $\Omega_* := D\big([0,\infty[, \mathfrak{P}_f(\mathbb{X}) \times \mathbb{X}\big) = \Omega \times D\big([0,\infty[, \mathbb{X}\big)$ describes the size-biased tree with trunk $\big(\xi(t)\big)_{t\geq 0}$. As we have emphasized above, the type process along the trunk, $\sigma(t) := \sigma\big(\xi(t)\big)$, is a copy of the retrospective mutation chain as defined in Section 3. In contrast, the individuals off the trunk may be understood as a branching process with immigration.

We write $\widehat{\mathbb{P}}_*^i$ for the distribution of $\big(\widehat{X}(t), \xi(t)\big)_{t\geq 0}$ on Ω_*, and $\widehat{\mathbb{P}}^i$ for its marginal, the distribution of $\big(\widehat{X}(t)\big)_{t\geq 0}$ on Ω. The representation theorem below establishes the relationship between $\mathbb{P}^i$, $\widehat{\mathbb{P}}_*^i$ and the retrospective mutation chain. We use the shorthand $y[0,t]$ for a path $\big(y(s)\big)_{0\leq s\leq t}$.

Theorem 4.1. *Let $t > 0$, $i \in S$, and $F : D\big([0,t], \mathfrak{P}_f(\mathbb{X}) \times \mathbb{X}\big) \to [0,\infty[$ be any measurable function. Then one has*

$$h_i^{-1}\, \mathbb{E}^i \left(e^{-\lambda t} \sum_{x\in X(t)} F\big(X[0,t], x[0,t]\big)\, h_{\sigma(x)}\right) = \widehat{\mathbb{E}}_*^i \left(F\big(\widehat{X}[0,t], \xi[0,t]\big)\right). \tag{4.2}$$

Recall that this theorem is valid for arbitrary sign of λ. The proof is postponed until Section 5.1. Here we discuss some immediate consequences and possible extensions.

Remark 4.1. (a) Setting $F(X[0,t],x[0,t]) = I\{\sigma(x(t)) = j\}\, h_j^{-1}$ in (4.2) and using the ergodic theorem for the retrospective mutation chain $\sigma\big(\xi(t)\big)$ we obtain the Perron-Frobenius result

$$\mathbb{E}^i\left(Z_j(t)\right) e^{-\lambda t} = h_i\, \widehat{\mathbb{P}}_*^i\big(\sigma(\xi(t)) = j\big)\, h_j^{-1} \underset{t\to\infty}{\longrightarrow} h_i\, \alpha_j\, h_j^{-1} = h_i\, \pi_j\,. \tag{4.3}$$

In particular, Eq. (2.5) follows by summing over j.

(b) Taking any F of the form $F(X[0,t],x[0,t]) = g(X[0,t])$ we conclude that

$$h_i^{-1}\, \mathbb{E}^i\Big(W(t)\, g(X[0,t])\Big) = \widehat{\mathbb{E}}^i\Big(g(\widehat{X}[0,t])\Big)$$

with

$$W(t) := \langle Z(t), h\rangle\, e^{-\lambda t}\,.$$

In particular, $h_i = \mathbb{E}^i\big(W(t)\big)$. Thus, on the σ-algebra $\mathcal{F}_t$ generated by $X[0,t]$, $\widehat{\mathbb{P}}^i$ is absolutely continuous with respect to $\mathbb{P}^i$ with density $W(t)/h_i$, and $(W(t))_{t\geq 0}$ is a martingale with respect to $\mathbb{P}^i$. The latter statement is one of the standard facts of branching process theory; see e.g. [2], p. 209, Theorem 1.

(c) Theorem (4.1) has the appearance of the Campbell theorem of point process theory; see, e.g., [21], pp. 14 & 228. To clarify the relation let $t > 0$ be fixed and

$$\Phi(t) = \big\{x[0,t] : x \in X(t)\big\}$$

the finite random subset of $D\big([0,t],\mathbb{X}\big)$ which describes the lineages that survive until time t. Also, let C_t^i be the measure on $\mathfrak{P}_f\big(D\big([0,t],\mathbb{X}\big)\big) \times D\big([0,t],\mathbb{X}\big)$ with Radon-Nikodym density $e^{\lambda t}\, h_i\, h_{\sigma(\xi(t))}^{-1}$ relative to the joint distribution of $\widehat{\Phi}(t) = \big\{x[0,t] : x \in \widehat{X}(t)\big\}$ and $\xi[0,t]$ under $\widehat{\mathbb{P}}_*^i$. Theorem (4.1) then implies that

$$\mathbb{E}^i\Big(\sum_{\psi\in\Phi(t)} F\big(\Phi(t),\psi\big)\Big) = \int F(\Psi,\psi)\, C_t^i(d\Psi, d\psi)$$

for any measurable $F \geq 0$, i.e., C_t^i is the Campbell measure of $\Phi(t)$ under $\mathbb{P}^i$. This assertion, however, is slightly weaker than Theorem (4.1) because $X[0,t]$ also includes the lineages that die out before time t.

Remark 4.2. In the above, the size-biased tree was constructed using the right eigenvector h as a weight on the types. As a matter of fact, the same construction can be carried out when h is replaced by an arbitrary weight vector $\gamma \in\,]0,\infty[^S$, and a representation theorem analogous to Theorem (4.1) can be obtained. We discuss here only the special case $\gamma \equiv 1$ which is of particular interest, and already appears in [9, Theorem 2] in the context of critical multitype branching. The size-biased offspring distribution associated with this case is

$$\widetilde{\mathsf{p}}_i(\kappa) = \|\kappa\|\, \mathsf{p}_i(\kappa)/m_i\,, \qquad \kappa \in \mathbb{Z}_+^S,$$

where $\|\kappa\| = \sum_j \kappa_j$ is the total offspring and $m_i = \sum_j m_{ij}$ its expectation under p_i. The lifetime of an i-individual on the trunk is exponential with parameter $a_i m_i$, and the successor on the trunk is chosen among the children with *equal* probability. Writing a tilde (instead of a hat) to characterize all quantities of the associated size-biased tree, one arrives at the following counterpart of (4.2):

$$\mathbb{E}^i\Big(\sum_{x\in X(t)} e^{-t\langle L^x(t),r\rangle}\, F\big(X[0,t],x[0,t]\big)\Big) = \widetilde{\mathbb{E}}^i_*\Big(F\big(\widetilde{X}[0,t],\tilde{\xi}[0,t]\big)\Big)\,. \tag{4.4}$$

In the above, r is the vector with i-coordinate $r_i = a_i(m_i-1) = \sum_j a_{ij}$, the mean reproduction rate of type i. Accordingly, the expectation $\langle L^x(t),r\rangle$ is the *mean reproduction rate along the lineage leading to x at time t*. The type process along the trunk, $\tilde{\sigma}(t) := \sigma\big(\tilde{\xi}(t)\big)$, is the Markov chain with transition rates $\tilde{g}_{ij} = a_i\, m_{ij} - m_i\,\delta_{ij}$. In view of the decomposition $a_{ij} = \tilde{g}_{ij} + r_i\delta_{ij}$, this Markov chain describes the pure mutation part of the type evolution.

On the left-hand side of (4.4), each individual is weighted according to the mean fertility of its lineage. Indeed, suppose we are given a lineage up to time t of which we know only the intervals of time spent in each state $i \in S$, and imagine that random split events and independent random offspring sizes are distributed over $[0,t]$ with the appropriate rates and distributions. The number ζ_i of split events during the sojourn in state i is then Poisson with parameter $a_i t\nu_i$, where ν_i is the fraction of time spent in state i; and the expected total offspring at each of these events is m_i. Since offspring sizes are independent, the expected product of offspring sizes along the lineage then amounts to $\prod_{i\in S}\mathbb{E}\big(m_i^{\zeta_i}\big) = e^{t\langle\nu,r\rangle}$. A result similar to (4.4), with an analogous interpretation of the exponential factor, already appears in [3, p. 127] in the context of Palm trees for spatially inhomogeneous branching.

Here are some consequences of (4.4):

(a) For $F(X[0,t],x[0,t]) = \exp\big[\,t\,\langle L^x(t),r\rangle\big]\, I\{\sigma(x(t))=j\}$, Eq. (4.4) becomes

$$\mathbb{E}^i\Big(Z_j(t)\Big) = \widetilde{\mathbb{E}}^i_*\Big(e^{t\,\langle L^{\tilde{\xi}}(t),r\rangle}\, I\{\sigma(\tilde{\xi}(t))=j\}\Big)\,, \tag{4.5}$$

which is a version of the *Feynman-Kac formula.* Indeed, consider the function $u(t,i) = \mathbb{E}^i(Z_j(t))$ for fixed j. Since $u(t,i) = (e^{t\mathsf{A}})_{ij}$, it follows that $u(t,i)$ is the unique solution of the Cauchy problem

$$\frac{d}{dt}\,u(t,i) = \sum_{k\in S}\tilde{g}_{ik}\,u(t,k) + r_i\,u(t,i), \quad u(0,i) = \delta_{ij}\,,$$

which is given by the Feynman-Kac formula.

(b) Summing over j in (4.5) and using Varadhan's lemma of large deviation theory (see [12, p. 32] or [23, Theorem 2.1]) together with (2.5) we arrive at the *variational principle*

$$\lambda = \lim_{t\to\infty}\frac{1}{t}\log\mathbb{E}^i\big(|X(t)|\big) = \max_{\nu\in\mathcal{P}(S)}\big[\langle\nu,r\rangle - I_{\widetilde{\mathsf{G}}}(\nu)\big]\,,$$

where $I_{\widetilde{\mathsf{G}}}$ is the large deviation rate function for the empirical distribution of the Markov chain with transition rates $\tilde{g}_{ij}$; cf. (5.10) for its definition in the case of the transition rates g_{ij}. In fact, it is not difficult to see that the maximum is attained at (and only at) the ancestral distribution α. This variational principle is behind the one found in [11].

(c) Just as in Remark 4.1(b) we find that

$$\widetilde{W}(t) := \sum_{x \in X(t)} e^{-t\langle L^x(t), r\rangle}$$

is a martingale. In this martingale (which does not seem to have been considered so far), each individual at time t is weighted according to the mean fertility of its lineage.

5. Proofs

5.1. Transforming the tree

Here we prove Theorems 4.1 and 2.1(b). For the former we do not need that λ is positive.

Proof of Theorem 4.1. It is sufficient to show that

$$\widehat{\mathbb{E}}_*^i \Big(F(\widehat{X}[0,t], \xi[0,t])\,;\, \xi(t) = x \Big) = e^{-\lambda t}\, h_i^{-1}\, h_{\sigma(x)}\, \mathbb{E}^i \Big(F(X[0,t], x[0,t])\,;\, x \in X(t) \Big) \tag{5.1}$$

for all $x \in \mathbb{X}$; the theorem then follows by summation over all $x \in \mathbb{X}$. Suppose that $x = (i_1, \dots, i_n; \ell_1, \dots, \ell_n) \in \mathbb{X}_n$, and let $x_k = (i_1, \dots, i_k; \ell_1, \dots, \ell_k)$ be its ancestor in generation k, $0 \le k \le n$.

Consider the right-hand side of (5.1) and write $e^{-\lambda t}\, h_i^{-1}\, h_{\sigma(x)} = q_1\, q_2\, q_3$ with

$$q_1 = e^{-\lambda t} \prod_{k=0}^{n-1} c_{i_k}\,, \quad q_2 = \prod_{k=0}^{n-1} \frac{\langle N_{x_k}, h\rangle}{c_{i_k}\, h_{i_k}}\,, \quad q_3 = \prod_{k=0}^{n-1} \frac{h_{i_{k+1}}}{\langle N_{x_k}, h\rangle}\,;$$

of course, the random quantities q_2 and q_3 must then be included into the expectation. The factor q_1 corresponds to the time change obtained when the exponential parameter a_{i_k} is replaced by $a_{i_k} + \lambda = a_{i_k} c_{i_k}$ along the ancestral line of x, i.e., when τ_{x_k} is replaced by $\widehat{\tau}_k$ for $k = 0, \dots, n$. Indeed, the associated Radon-Nikodym density is

$$\bar{q}_1 = e^{-\lambda T_x} \prod_{k=0}^{n} c_{i_k}\,.$$

Conditioning $\bar{q}_1$ on the tree $X[0,t]$ up to time t and using the loss of memory property of the exponential law of τ_x we find that, almost surely on $\{T_{\tilde{x}} \le t\}$,

$$\mathbb{E}^i \big(\bar{q}_1 \,\big|\, X[0,t] \big) = e^{-\lambda t}\, \mathbb{E}^i \big(e^{-\lambda(T_x - t)} \,\big|\, T_x > t \big) \prod_{k=0}^{n} c_{i_k} = q_1\,.$$

Next, it is immediate from (4.1) that the factor q_2 is precisely the Radon-Nikodym density corresponding to a change from N_{x_k} to the size-biased offspring $\widehat{N}_k$ for $k = 0, \dots, n-1$. Finally, q_3 is the conditional selection probability for the trunk:

$$q_3 = \widehat{\mathbb{P}}^i \big(\xi_{k+1} = x_{k+1} \text{ for } 0 \le k < n \,\big|\, \widehat{X}[0,t] \Big)\,.$$

The right-hand side of (5.1) is therefore equal to

$$\begin{aligned} &\widehat{\mathbb{E}}_*^i \Big(F(\widehat{X}[0,t], x[0,t])\,;\, \widehat{T}_{\tilde{x}} \le t < \widehat{T}_x,\, \xi_{k+1} = x_{k+1} \text{ for } 0 \le k < n \Big) \\ &\qquad = \widehat{\mathbb{E}}_*^i \Big(F(\widehat{X}[0,t], \xi[0,t])\,;\, \xi(t) = x \Big)\,, \end{aligned}$$

as was to be shown.

In the rest of this paper we assume that $\lambda > 0$.

Proof of Theorem 2.1(b). The basic observation is that the martingale $W(t) = \langle Z(t), h\rangle\, e^{-\lambda t}$ considered in Remark 4.1(b) converges to a finite limiting variable $W \geq 0$ $\mathbb{P}^i$-almost surely for each i. When combined with Theorem 2.1(a) to be proved below, this implies the asserted convergence result. The essential part of the proof consists in showing that W is nontrivial if and only if condition (2.4) holds. There are two possible routes to achieve this.

Either one can consider a discrete time skeleton $\delta\mathbb{N}$ and simply apply the discrete-time version of the Kesten-Stigum theorem. For this one has to check that condition (2.4) holds if and only if $\mathbb{E}^i(Z_j(\delta)\log Z_j(\delta)) < \infty$ for all $i, j \in S$, which can be done.

Or, more naturally, one can use Theorem 4.1 to extend the conceptual proof of Lyons et al. [20] and Kurtz et al. [19] directly to continuous time. We spell out some details for the convenience of the reader. As in [20], one observes first that W is nontrivial if and only if $\widehat{\mathbb{P}}^i$ is absolutely continuous with respect to $\mathbb{P}^i$ (with Radon-Nikodym density W/h_i), which is the case if ond only if

$$\limsup_{t\to\infty} \widehat{W}(t) < \infty \quad \widehat{\mathbb{P}}^i\text{-almost surely;} \tag{5.2}$$

here we have put a hat on W to stress the change of the underlying measure.

To check that (5.2) is equivalent to (2.4) one notices first that (2.4) is equivalent to

$$\mathbb{E}\big(\log\langle \widehat{N}_i, h\rangle\big) < \infty \quad \text{for all } i \in S, \tag{5.3}$$

by the properties of log and Eq. (4.1). Next one observes that $\widehat{X}(t) \setminus \{\xi(t)\}$ is a branching process with immigration at the split times of the trunk $\xi(t)$. Specifically, let $\widehat{T}_{(n)} := \widehat{T}_{\xi_n}$ be the n-th split time and $\widehat{N}_{(n)} = \widehat{N}_{\xi_n}$ the n-th offspring of the trunk. The $\widehat{N}_{(n)}$ are independent (conditionally on the trunk), with distribution $\widehat{\mathsf{p}}_{\sigma(\xi_n)}$.

Suppose first (5.3) fails, and pick any $j \in S$ with $\mathbb{E}\big(\log\langle \widehat{N}_j, h\rangle\big) = \infty$. Consider the subsequence $(\widehat{T}_{(n_l)})_{l\geq 1}$ of split times of the trunk for which $\sigma(\xi_{n_l}) = j$. Since the random variables $\log\langle \widehat{N}_{(n_l)}, h\rangle$ are i.i.d. with infinite mean, a standard Borel-Cantelli argument shows that $\limsup_{l\to\infty} l^{-1}\log\langle \widehat{N}_{(n_l)}, h\rangle = \infty$ almost surely. On the other hand, $\limsup_{l\to\infty} \widehat{T}_{(n_l)}/l < \infty$ a.s. because the differences $\widehat{T}_{(n_{l+1})} - \widehat{T}_{(n_l)}$ are i.i.d. with finite mean. This gives

$$\limsup_{l\to\infty} \widehat{W}(\widehat{T}_{(n_l)}) \geq \limsup_{l\to\infty} \langle \widehat{N}_{(n_l)}, h\rangle\, e^{-\lambda \widehat{T}_{(n_l)}} = \infty \quad \text{a.s.,}$$

so that (5.2) fails.

Conversely, suppose (5.3) holds. As in Section 4, we consider the offspring $\widehat{X}^\sharp_{n+1}$ of the trunk created at time $\widehat{T}_{(n)}$ having type counting measure $\widehat{N}_{(n)}$. We also introduce the σ-algebra $\mathcal{T}$ generated by the trunk variables $\{\widehat{T}_{(n)}, \widehat{N}_{(n)} : n \geq 0\}$ and use a tilde to characterize the trunk-reduced quantities obtained by removing the trunk individuals from the population. Then for each $t > 0$ we obtain, with the notation (2.2),

$$\begin{aligned}\widehat{\mathbb{E}}^i_*\big(\widetilde{W}(t)\big|\mathcal{T}\big) &= \sum_{n:\,\widehat{T}_{(n)}\le t} e^{-\lambda \widehat{T}_{(n)}}\, \widehat{\mathbb{E}}^i_*\Big(\sum_{x\in\widetilde{X}^\sharp_{n+1}} \langle Z(x,t),h\rangle\, e^{-\lambda(t-\widehat{T}_{(n)})}\Big|\,\mathcal{T}\Big)\\ &= \sum_{n:\,\widehat{T}_{(n)}\le t} e^{-\lambda \widehat{T}_{(n)}}\langle \widetilde{N}_{(n)},h\rangle\end{aligned}$$

by the martingale property of $W(t)$ applied to the descendant trees $X(x,\cdot)$. Now, (5.3) and a Borel-Cantelli argument imply that $n^{-1}\log\langle \widehat{N}_{(n)},h\rangle \to 0$ almost surely. On the other hand, $\liminf_{n\to\infty}\widehat{T}_{(n)}/n > 0$ by the law of large numbers, whence

$$\sum_{n\ge 0} e^{-\lambda \widehat{T}_{(n)}}\langle \widetilde{N}_{(n)},h\rangle < \infty \quad \text{a.s.}$$

This means that, conditionally on $\mathcal{T}$, $\widetilde{W}(t)$ is a submartingale with bounded expectation, which gives (5.2) by the submartingale convergence theorem and finishes the proof of Theorem 2.1(b). The final identity $\{W > 0\} = \Omega_{\rm surv}$ a.s. follows from the trivial inclusion $\{W > 0\} \subset \Omega_{\rm surv}$ and the well-known fact that $q_i = \mathbb{P}^i(W = 0)$ solves the equation $q_i = \mathbb{E}\big(\prod_{j\in S} q_j^{N_{ij}}\big)$ which has the extinction probabilities as unique non-trivial solution [2, p. 205, Eq. (25)].

5.2. Laws of large numbers for population averages

In this section we are concerned with laws of large numbers for population averages. We state a general such law for discrete time skeletons and then use it to prove Theorems 2.1(a) and 3.1. Recall from (2.1) that, for $t, u > 0$ and $x \in X(t)$, the path $X(x,[t,t+u]) = \big(X(x,t+s)\big)_{0\le s\le u}$ describes the subtree of x-descendants during the time interval $[t,t+u]$.

Proposition 5.1. *Let $\delta, u > 0$, $i, j \in S$, and $f : D\big([0,u],\mathfrak{P}(\mathbb{X})\big) \to \mathbb{R}$ be a measurable function with existing mean $c_j = \mathbb{E}^j\big(f\circ X[0,u]\big)$. Then*

$$\lim_{n\to\infty}\frac{1}{Z_j(n\delta)}\sum_{x\in X_j(n\delta)} f\circ X(x,[n\delta,n\delta+u]) = c_j \quad \mathbb{P}^i\text{-}almost\ surely\ on\ \Omega_{\rm surv}.$$

Proof. This result follows essentially from Lemmas 3 and 4 in [19]. Since this reference contains no proof of the former, we provide a proof here for the sake of completeness.

We assume first that δ is so large that $u < \delta$ and $\rho := \mathbb{E}^j(Z_j(\delta)) > 1$. Such a δ exists because $\lambda > 0$ and A is irreducible. Let $\mathcal{F}_{n\delta}$ denote the σ-algebra generated by $X[0,n\delta]$. Since $u < \delta$, for each $n \ge 1$ the random variables $\varphi_{n,x} := f\circ X(x,[n\delta,n\delta+u])$ with $x \in X_j(n\delta)$ are $\mathcal{F}_{(n+1)\delta}$-measurable and, conditionally on $\mathcal{F}_{n\delta}$, i.i.d. with mean c_j. This implies that the sequence $(\varphi_l)_{l\ge1}$ on $\Omega_{\rm surv}$ obtained by enumerating first $\{\varphi_{1,x} : x \in X_j(\delta)\}$ in some order, then $\{\varphi_{2,x} : x \in X_j(2\delta)\}$ and so on, is still i.i.d. with mean c_j. The strong law of large numbers therefore implies that $\lim_{k\to\infty}(1/k)\sum_{l=1}^k \varphi_l = c_j$ $\mathbb{P}^i$-almost surely on $\Omega_{\rm surv}$, and thus in particular that the subsequence

$$A_n := \frac{1}{\Psi_n}\sum_{l=1}^n \sum_{x\in X_j(l\delta)} \varphi_{l,x}$$

converges to c_j $\mathbb{P}^i$-almost surely on $\Omega_{\rm surv}$ as $n \to \infty$; here $\Psi_n = \sum_{l=1}^n \psi_l$ with $\psi_l = Z_j(l\delta)$.

Next, the sequence $(\psi_l)_{l\geq 1}$ dominates a single-type discrete-time Galton-Watson process with mean $\rho > 1$, and the latter survives precisely on $\Omega_{\rm surv}$. By Lemma 4 of [19], it follows that $\liminf_{l\to\infty} \psi_{l+1}/\psi_l \geq \rho$ almost surely on $\Omega_{\rm surv}$. This implies that

$$\limsup_{n\to\infty} \Psi_{n-1}/\psi_n = \limsup_{n\to\infty} \sum_{l=1}^{n-1} \psi_l/\psi_n < \infty$$

almost surely on $\Omega_{\rm surv}$. As

$$\frac{1}{\psi_n} \sum_{x\in X_j(n\delta)} \varphi_{n,x} = A_n + (A_n - A_{n-1})\, \Psi_{n-1}/\psi_n \,,$$

the proposition follows in the case of large δ.

If $\delta > 0$ is arbitrary, we choose some $k \in \mathbb{N}$ such that $\delta' := k\delta$ is so large as required above. Let $0 \leq l < k$. Applying the preceding result to each of the subtrees $X(x, [l\delta, \infty[)$ with $x \in X(l\delta)$ and averaging, we then find that

$$\lim_{n\to\infty} \frac{1}{\psi_{nk+l}} \sum_{x\in X_j((nk+l)\delta)} \varphi_{nk+l,x} = c_j$$

$\mathbb{P}^i$-almost surely on $\Omega_{\rm surv}$, and the proof is complete.

A typical application of the preceding proposition is the following corollary. Consider the $X_j(s)$-averaged type counting measure

$$C_{j,u}(s) = \frac{1}{Z_j(s)} \sum_{x\in X_j(s)} Z(x, s{+}u) \tag{5.4}$$

at time $s{+}u$, where $Z(x, s{+}u)$ is defined by (2.2). Proposition 5.1 then immediately implies the following corollary.

Corollary 5.1. *For any $\delta, u > 0$ and $i, j \in S$,*

$$C_{j,u}(n\delta) \underset{n\to\infty}{\longrightarrow} \mathbb{E}^j(Z(u)) \quad \mathbb{P}^i\text{-almost surely on } \Omega_{\rm surv}.$$

To pass from a discrete time skeleton to continuous time we will use the following continuity lemma which follows also from Proposition 5.1.

Lemma 5.1. *Given $\varepsilon > 0$, there exists some $\delta > 0$ such that for all $i, j \in S$ and $k \in \mathbb{N}$ one has*

$$\limsup_{n\to\infty} \sup_{n\delta\leq s\leq(n+1)\delta} \frac{\|Z(s)\|}{\|Z(n\delta)\|} < 1 + \varepsilon \,, \tag{5.5}$$

$$\liminf_{n\to\infty} \inf_{n\delta\leq s\leq(n+1)\delta} \frac{Z_j(s)}{Z_j(n\delta)} > 1 - \varepsilon \,, \tag{5.6}$$

and

$$\liminf_{n\to\infty} \inf_{n\delta\le s\le(n+1)\delta} \inf_{k\delta\le u\le(k+1)\delta} \frac{\sum_{y\in X_j(s)} \|Z(y,s{+}u)\|}{\sum_{y\in X_j(n\delta)} \|Z(y,n\delta{+}k\delta)\|} > 1-\varepsilon \tag{5.7}$$

$\mathbb{P}^i$*-almost surely on* Ω_{surv}.

Proof. We begin by proving the upper bound (5.5). For $n\delta \le s \le (n{+}1)\delta$ we can write

$$\|Z(s)\| = \sum_{x\in X(n\delta)} |X(x,s)| \le \sum_{x\in X(n\delta)} M(x,[n\delta,(n{+}1)\delta])\ ,$$

where $M(x,[n\delta,(n{+}1)\delta]) = \max_{n\delta\le s\le(n+1)\delta} |X(x,s)|$. Hence

$$\sup_{n\delta\le s\le(n+1)\delta} \frac{\|Z(s)\|}{\|Z(n\delta)\|} \le \max_{j\in S} \frac{1}{Z_j(n\delta)} \sum_{x\in X_j(n\delta)} M(x,[n\delta,(n{+}1)\delta])\ .$$

By Proposition 5.1, the last expression converges to $m(\delta) := \max_{j\in S} \mathbb{E}^j\big(M(0,[0,\delta])\big)$ almost surely on Ω_{surv}. Now, $M(0,[0,\delta])$ is dominated by the total size at time δ of the modified branching process for which the random variables $N_{x,\sigma(x)}$ in Section 2 are replaced by $N_{x,\sigma(x)} \vee 1$, so that each individual has at least one offspring of its own type. The latter process has a finite generator matrix, say A^+. Hence $m(\delta) \le \max_j(e^{\delta\mathsf{A}^+}1)_j \to 1$ as $\delta\to 0$. This completes the proof of (5.5).

Next we note that (5.6) follows from (5.7) by setting $u=k=0$. So it only remains to prove (5.7). Let $n\delta \le s \le (n{+}1)\delta$ and $k\delta \le u \le (k{+}1)\delta$. Considering only those individuals $y\in X(s)$ already alive at time $n\delta$ and still alive at time $(n{+}1)\delta$, and only those descendants $z \in X(y,s{+}u)$ living during the whole period $[(n{+}k)\delta,(n{+}k{+}2)\delta]$, we obtain the estimate

$$\sum_{y\in X_j(s)} \|Z(y,s{+}u)\| \ge \sum_{x\in X_j(n\delta)} I\{\tau_{x,n\delta} > \delta\} \sum_{z\in X(x,(n+k)\delta)} I\{\tau_{z,(n+k)\delta} > 2\delta\}\ .$$

Here we write $\tau_{x,t} = \inf\{u>0 : \sigma(x) \notin X(t{+}u)\} = T_x - t$ for the remaining life time of $x\in X(t)$ after time t. Proposition 5.1 therefore implies that the left-hand side of (5.7) is at least

$$\mathbb{E}^j\Big(I\{\tau_{j,0}>\delta\} \sum_{z\in X(k\delta)} I\{\tau_{z,k\delta}>2\delta\}\Big)\Big/\mathbb{E}^j\big(|X(k\delta)|\big) \tag{5.8}$$

$\mathbb{P}^i$-almost surely on Ω_{surv}. By the Markov property, the numerator is equal to

$$\mathbb{E}^j\Big(I\{\tau_{j,0}>\delta\} \sum_{z\in X(k\delta)} \exp[-2\delta a_{\sigma(z)}]\Big) \ge e^{-2\delta a}\, \mathbb{E}^j\big(I\{\tau_{j,0}>\delta\}\,|X(k\delta)|\big)$$

with $a=\max_i a_i$. The ratio in (5.8) is therefore not smaller than $e^{-2\delta a}\,(1-\varepsilon_k)$, where

$$\varepsilon_k = \mathbb{E}^j\Big(I\{\tau_{j,0}\le\delta\}\,|X(k\delta)|\Big)\Big/\mathbb{E}^j\big(|X(k\delta)|\big)\ .$$

For $k=0$ we have $\varepsilon_0 = 1-e^{-\delta a_j}$. For $k \geq 1$ we can use Theorem 4.1 to obtain

$$\varepsilon_k = \widehat{\mathbb{E}}^j_* \left(I\{\tau_{j,0} \leq \delta\}\, h^{-1}_{\sigma(\xi(k\delta))}\right) \Big/ \widehat{\mathbb{E}}^j_* \left(h^{-1}_{\sigma(\xi(k\delta))}\right) \leq \frac{\max_i h_i}{\min_i h_i} \left(1-e^{-\delta(a+\lambda)}\right).$$

Hence, if δ is sufficiently small then the ratio in (5.8) is larger than $1-\varepsilon$.

We are now ready for the proofs of Theorem 2.1(a) and 3.1.

Proof of Theorem 2.1(a). Essentially we reproduce here the argument of [19]. Let $\varepsilon > 0$ be given and $\varepsilon' > 0$ be such that, for every $\nu \in \mathcal{P}(S)$, $\|\nu - \pi\| < \varepsilon$ whenever $\|a\nu - \pi\| < \varepsilon'$ for some $a > 0$. Let $\delta > 0$ be so small as required in Lemma 5.1. According to (4.3), we can choose some $u \in \delta\mathbb{N}$ so large that

$$\left\|\mathbb{E}^j\left(Z(u)\, e^{-\lambda u}\right) - h_j\, \pi\right\| < \varepsilon' \min_{i\in S} h_i$$

for all $j \in S$. Corollary 5.1 then implies that, $\mathbb{P}^i$-almost surely on Ω_{surv},

$$\left\|C_{j,u}(s)\, e^{-\lambda u} - h_j\, \pi\right\| < \varepsilon' \min_{i\in S} h_i$$

for all sufficiently large $s \in \delta\mathbb{N}$. Writing $\Pi(t) = Z(t)/\|Z(t)\|$ and $a(t) = \frac{\|Z(t)\|\, e^{-\lambda u}}{\langle Z(t-u), h\rangle}$ for $t > u$, we conclude that

$$\left\|a(t)\, \Pi(t) - \pi\right\| \leq \frac{1}{\langle Z(t-u), h\rangle} \sum_{j\in S} Z_j(t-u)\, \left\|C_{j,u}(t-u)\, e^{-\lambda u} - h_j\, \pi\right\| < \varepsilon'$$

and therefore $\left\|\Pi(t) - \pi\right\| < \varepsilon$ for all sufficiently large $t \in \delta\mathbb{N}$ a.s. on Ω_{surv}. Finally, using (5.5) and (5.6) we find that $\Pi_j(t) > (1-2\varepsilon)\pi_j - \varepsilon$ for all $j \in S$ and all sufficiently large *real* t, again a.s. on Ω_{surv}. Since ε was arbitrary and $\Pi(t), \pi \in \mathcal{P}(S)$, this gives the desired convergence result.

Proof of Theorem 3.1. Recall the definition (3.1) of $A^u(t) \in \mathcal{P}(S)$, the $X(t)$-average of the ancestral type distribution at time $t-u$, and let $\alpha^u \in \mathcal{P}(S)$ be given by its coordinates $\alpha^u_j = \pi_j\, \mathbb{E}^j(\|Z(u)\|)\, e^{-\lambda u}$. Since $\alpha^u \to \alpha$ as $u \to \infty$ by (2.5), it is sufficient to show that

$$\mathbb{P}^i\left(\forall\, u > 0 : A^u(t) \underset{t\to\infty}{\longrightarrow} \alpha^u \,\middle|\, \Omega_{\text{surv}}\right) = 1. \tag{5.9}$$

Fix any $j \in S$, $u > 0$ and $\delta > 0$. By Corollary 5.1,

$$\left\|C_{j,u}(s)\right\| \to \mathbb{E}^j\left(\|Z(u)\|\right) \quad \text{as } s \to \infty \text{ through } \delta\mathbb{N}$$

$\mathbb{P}^i$-almost surely on Ω_{surv}. Combining this with Remark 3.1 and Theorem 2.1(a) we obtain, writing again $\Pi_j(s) := Z_j(s)/\|Z(s)\|$,

$$\begin{aligned} A^u_j(s+u) &= \frac{Z_j(s)\, \|C_{j,u}(s)\|}{|X(s+u)|} = \frac{\Pi_j(s)\, \|C_{j,u}(s)\|}{\sum_{k\in S} \Pi_k(s)\, \|C_{k,u}(s)\|} \\ &\underset{\delta\mathbb{N}\ni s\to\infty}{\longrightarrow} \pi_j\, \mathbb{E}^j\left(\|Z(u)\|\right) \Big/ \mathbb{E}^\pi\left(\|Z(u)\|\right) = \alpha^u_j \end{aligned}$$

$\mathbb{P}^i$-almost surely on Ω_{surv}.

Next let $\varepsilon > 0$ be given and $\delta > 0$ be chosen according to Lemma 5.1. Applying the above to $u = k\delta$ with arbitrary $k \in \mathbb{N}$ and using (5.5) and (5.7) we find that

$$\mathbb{P}^i\Big(\forall\, u > 0\ \forall\, j \in S : \liminf_{t\to\infty} A^u_j(t) > (1-2\varepsilon)\alpha^u_j \,\Big|\, \Omega_{\rm surv}\Big) = 1\,,$$

where the u-uniformity in (5.7) allows us to bring the u-quantifier inside of the probability. This gives (5.9) because ε is arbitrary and $A^u(t)$ and α^u are probability measures on S.

5.3. Application of large deviation theory

In this section we prove Theorems 3.2 and 3.3. The main tools are the representation theorem 4.1 and the Donsker-Varadhan large deviation principle for the empirical process of the retrospective mutation chain. In fact, these two ingredients together imply a large deviation principle for the type histories as follows. For every $\boldsymbol{\nu} \in \mathcal{P}_\Theta(\Sigma)$ let

$$H_{\mathsf{G}}(\boldsymbol{\nu}) = \sup_{t>0} H\big(\boldsymbol{\nu}_{[0,t]}; \boldsymbol{\mu}_{[0,t]}\big)/t$$

be the process-level large deviation rate function for the retrospective mutation chain. In the above, $\boldsymbol{\nu}_{[0,t]}$ and $\boldsymbol{\mu}_{[0,t]}$ are the restrictions of $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ to the time interval $[0,t]$, and $H\big(\boldsymbol{\nu}_{[0,t]}; \boldsymbol{\mu}_{[0,t]}\big)$ is their relative entropy. See [4, Eq. (4.4.28)]; alternative expressions can be found in [4, Theorem 4.4.38] and [23, Theorems 7.3 and 7.4].

Theorem 5.1. *For the empirical type evolution process $R^x(t)$ as in* (3.6) *we have, for $i \in S$ and closed $F \subset \mathcal{P}_\Theta(\Sigma)$*

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathbb{E}^i\Big(\sum_{x\in X(t)} I\{R^x(t) \in F\}\Big) \le \lambda - \inf_{\boldsymbol{\nu}\in F} H_{\mathsf{G}}(\boldsymbol{\nu})\,,$$

while for open $G \subset \mathcal{P}_\Theta(\Sigma)$

$$\liminf_{t\to\infty} \frac{1}{t} \log \mathbb{E}^i\Big(\sum_{x\in X(t)} I\{R^x(t) \in G\}\Big) \ge \lambda - \inf_{\boldsymbol{\nu}\in G} H_{\mathsf{G}}(\boldsymbol{\nu})\,.$$

Moreover, the function H_{G} is lower semicontinuous with compact level sets and attains its minimum 0 *precisely at $\boldsymbol{\mu}$.*

Proof. In view of Theorem 4.1, for every measurable $C \subset \mathcal{P}_\Theta(\Sigma)$ we have

$$\mathbb{E}^i\Big(\sum_{x\in X(t)} I\{R^x(t) \in C\}\Big) = h_i\, e^{\lambda t}\, \widehat{\mathbb{E}}^i_*\Big(I\{R^\xi(t) \in C\}\, h^{-1}_{\sigma(\xi(t))}\Big)\,.$$

Since $\max_i |\log h_i| < \infty$, the h's can be ignored on the exponential scale. The theorem thus follows from the Donsker-Varadhan large deviation principle; see [23, p.37, Theorem 7.8] or [4, Theorem 4.4.27], for example.

There is a similar large deviation principle on the level of empirical distributions. For $\nu \in \mathcal{P}(S)$ let

$$I_{\mathsf{G}}(\nu) = \sup_{v\in]0,\infty[^S} \Big[-\sum_{i\in S} \nu_i(\mathsf{G}v)_i/v_i\Big] = \inf_{\boldsymbol{\nu}\in\mathcal{P}_\Theta(\Sigma):\,\boldsymbol{\nu}_0=\nu} H_{\mathsf{G}}(\boldsymbol{\nu}) \qquad (5.10)$$

be the level-two rate function of the retrospective mutation chain; here we write $\boldsymbol{\nu}_0$ for the time-zero marginal distribution of $\boldsymbol{\nu}$. (For the second identity see [23, p.37, Theorem 7.9].) Then the following statement holds.

Corollary 5.2. *For any $i \in S$ and closed $F \subset \mathcal{P}(S)$,*

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathbb{E}^i \Big(\sum_{x\in X(t)} I\{L^x(t) \in F\}\Big) \leq \lambda - \inf_{\nu\in F} I_{\mathsf{G}}(\nu) \ ,$$

while for open $G \subset \mathcal{P}(S)$

$$\liminf_{t\to\infty} \frac{1}{t} \log \mathbb{E}^i \Big(\sum_{x\in X(t)} I\{L^x(t) \in G\}\Big) \geq \lambda - \inf_{\nu\in G} I_{\mathsf{G}}(\nu) \ .$$

Moreover, the function I_{G} is continuous and strictly convex and attains its minimum 0 precisely at α.

Proof. Simply replace the process-level large deviation principle for the retrospective mutation chain by the one for its empirical distributions. The latter can either be deduced from the former by the contraction principle, see [23, Theorems 2.3 & 7.9], or be proved directly as in [12, Section IV.4].

We are now ready for the proofs of Theorems 3.2 and 3.3.

Proof of Theorem 3.3. Let d be a metric for the weak topology on $\mathcal{P}_\Theta(\Sigma)$. To be specific, we let d_Σ denote the Skorohod metric on Σ (defined in analogy to the one-sided case considered in [8, p. 117, Eq. (5.2)]), and d be the associated Prohorov metric on $\mathcal{P}_\Theta(\Sigma)$; see [8, p. 96, Eq. (1.1)]. For any fixed $\varepsilon > 0$ we consider the set $C = \{\boldsymbol{\nu} \in \mathcal{P}_\Theta(\Sigma) : d(\boldsymbol{\nu}, \boldsymbol{\mu}) \geq \varepsilon\}$, the complement of the open ε-neigborhood of $\boldsymbol{\mu}$. In view of Remark 3.3 we need to show that

$$\Gamma(t, C) := \frac{1}{|X(t)|} \sum_{x\in X(t)} I\{R^x(t) \in C\} \underset{t\to\infty}{\longrightarrow} 0$$

$\mathbb{P}^i$-almost surely on Ω_{surv}. In the first part of the proof we will establish this convergence along a discrete time skeleton $\delta\mathbb{N}$, where $\delta > 0$ is arbitrary.

Since C is closed and H_{G} has compact level sets and attains its minimum 0 at $\boldsymbol{\mu}$ only, the infimum $c := \inf_{\boldsymbol{\nu}\in C} H_{\mathsf{G}}(\boldsymbol{\nu})$ is strictly positive. We can therefore choose a constant $\lambda > \gamma > \lambda - c$. We write

$$\Gamma(t, C) = \Big(\frac{e^{\gamma t}}{|X(t)|}\Big) \quad \Big(e^{-\gamma t} \sum_{x\in X(t)} I\{R^x(t) \in C\}\Big)$$

and show that each factor tends to 0 along $\delta\mathbb{N}$ a.s. on Ω_{surv}. In view of Corollary 5.1 and Theorem 2.1(a),

$$\begin{aligned} \frac{|X((n{+}1)\delta)|}{|X(n\delta)|} &= \sum_{j\in S} \frac{Z_j(n\delta)}{\|Z(n\delta)\|} \quad \frac{1}{Z_j(n\delta)} \sum_{x\in X_j(n\delta)} |X(x, (n{+}1)\delta)| \\ &\underset{n\to\infty}{\longrightarrow} \sum_{j\in S} \pi_j \, \mathbb{E}^j(|X(\delta)|) = e^{\lambda\delta} \quad \text{a.s. on } \Omega_{\text{surv}}. \end{aligned}$$

Hence $n^{-1}\log|X(n\delta)| \to \lambda\delta$ and therefore $e^{\gamma n\delta}/|X(n\delta)| \to 0$ a.s. on $\Omega_{\rm surv}$. On the other hand, using Markov's inequality and Theorem 5.1 we obtain for any $a > 0$

$$\limsup_{n\to\infty} \frac{1}{n\delta} \log \mathbb{P}^i \Big(e^{-\gamma n\delta} \sum_{x\in X(n\delta)} I\{R^x(n\delta) \in C\} \geq a \Big) \leq \lambda - c - \gamma < 0 \,.$$

The Borel-Cantelli lemma thus shows that also the second factor of $\Gamma(t,C)$ tends to 0 a.s. as $t \to \infty$ through $\delta\mathbb{N}$. We therefore conclude that $\lim_{n\to\infty}\Gamma(n\delta, C) = 0$ a.s. on $\Omega_{\rm surv}$.

To extend this result to the full convergence $t \to \infty$ along all reals we pick some $0 < \varepsilon' < \varepsilon$ and let C' be defined in terms of ε' instead of ε. Also, let A be an arbitrary closed set in Σ, $\varepsilon^* = \varepsilon - \varepsilon'$, and $A^* = \big\{\sigma \in \Sigma : d_\Sigma(\sigma, A) < \varepsilon^*\big\}$ the ε^*-augmentation of A. Then for any two time instants s, t with $s \leq t \leq s+\delta$ and every $y \in X(t)$ we can write

$$\begin{aligned} R^y(t)(A) \leq\ & \frac{1}{t}\int_0^s I_A(\vartheta_u\sigma(y)_{t,\rm per})\, du + \frac{\delta}{t} \\ \leq\ & R^{y(s)}(s)(A^*) + \frac{1}{s}\int_0^s I\big\{u : d_\Sigma(\vartheta_u\sigma(y(s))_{s,\rm per}, \vartheta_u\sigma(y)_{t,\rm per}) \geq \varepsilon^*\big\}\, du + \frac{\delta}{t}\,. \end{aligned}$$

By the locality of the Skorohod metric d_Σ, there exists a constant $c = c(\varepsilon^*)$ such that $d_\Sigma(\vartheta_u\sigma(y(s))_{s,\rm per}, \vartheta_u\sigma(y)_{t,\rm per}) < \varepsilon^*$ whenever the interval $[-u, s-u]$ on which these functions agree contains $[-c, c]$. The second term in the last sum is therefore at most $2c/s$, whence

$$R^y(t)(A) \leq R^{y(s)}(s)(A^*) + \varepsilon^*$$

for sufficiently large s. This means that $d\big(R^y(t), R^{y(s)}(s)\big) < \varepsilon^*$ and therefore

$$\{R^y(t) \in C\} \subset \{R^{y(s)}(s) \in C'\}$$

when s is large enough. For such s we obtain

$$\begin{aligned} \Gamma(t,C) - \Gamma(s,C') &\leq \Big(\frac{1}{|X(t)|} - \frac{1}{|X(s)|}\Big)|X(t)| \\ &\quad + \frac{1}{|X(s)|}\sum_{x\in X(s)} I\{R^x(s)\in C'\}\big(|X(x,t)| - 1\big) \\ &\leq 1 - \inf_{s\leq t\leq s+\delta} |X(t)|/|X(s)| \\ &\quad + \frac{1}{|X(s)|}\sum_{x\in X(s)} \big(M(x,[s,s+\delta]) - 1\big)\ , \end{aligned}$$

where $M(x,[s,s+\delta]) = \max_{s\leq t\leq s+\delta}|X(x,t)|$ as in the proof of Lemma 5.1. Setting $s = n\delta$, letting $n \to \infty$ and using Theorem 2.1(a) and Proposition 5.1 we see that the last term converges to $\mathbb{E}^\pi\big(M(0,[0,\delta]) - 1\big)$ a.s. on $\Omega_{\rm surv}$. According to the proof of (5.5), this limit can be made arbitrarily small if δ is chosen small enough. In combination with (5.6) and the first part of this proof, this shows that $\limsup_{t\to\infty}\Gamma(t,C) \leq a$ for every $a > 0$ almost surely on $\Omega_{\rm surv}$. The proof is thus complete.

Proof of Theorem 3.2. There are two possible routes for the proof. One can either repeat the argument above by simply replacing Theorem 5.1 by Corollary 5.2. Or one notices that $L^x(t)$ is the time-zero marginal of $R^x(t)$ and that the marginal mapping $\boldsymbol{\nu} \to \boldsymbol{\nu}_0$ is continuous in the topologies chosen. The latter fact is used for the derivation of the level-two large deviation principle from that on the process level by means of the contraction principle; see [23, p. 34].

Acknowledgement. It is our pleasure to thank Nina Gantert, Peter Jagers, Götz Kersting, and Anton Wakolbinger for helpful discussions, and invaluable references to the branching literature. Financial support from the German Research Council (DFG) and the Erwin Schrödinger International Institute for Mathematical Physics in Vienna is gratefully acknowledged.

References

[1] Athreya, K. B. (1968). Some results on multitype continuous time Markov branching processes. *Ann. Math. Stat.* **39,** 347–357.

[2] Athreya, K. B. and Ney, P. E. (1972). *Branching Processes.* Springer, New York.

[3] Chauvin, B., Rouault, A. and Wakolbinger, A. (1991). Growing conditioned trees. *Stoch. Proc. Appl.* **39,** 117–130.

[4] Deuschel, J. D. and Stroock, D. W. (1989). *Large Deviations.* Academic Press, Boston. Reprint, AMS-Chelsea, Providence, RI, 2000.

[5] Donsker, M. D. and Varadhan, S. R. S. (1975). Asymptotic evaluation of certain Markov process expectations for large time, I. *Comm. Pure Appl. Math.* **28,** 1–47.

[6] Donsker, M. D. and Varadhan, S. R. S. (1983). Asymptotic evaluation of certain Markov process expectations for large time. IV. *Comm. Pure Appl. Math.* **36,** 183–212.

[7] Durrett, R. (2002). *Probability models for DNA sequence evolution.* Springer, New York.

[8] Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes – Characterization and Convergence.* Wiley, New York.

[9] Gorostiza, L. G., Roelly, S. and Wakolbinger (1992). *Persistence of critical multitype particle and measure branching processes. Probab. Theor. Relat. Fields* **92**, 313-335.

[10] Harris, T. E. (1963). *The Theory of Branching Processes.* Springer, Berlin. corrected reprint, Dover, New York, 2002.

[11] Hermisson, J., Redner, O., Wagner, H. and Baake, E. (2002). Mutation-selection balance: Ancestry, load, and maximum principle. *Theor. Pop. Biol.* **62,** 9–46. cond-mat/0202432.

[12] DEN HOLLANDER, F. (2000). *Large Deviations. Fields Institute Monographs* vol. 14. AMS, Providence, RI.

[13] JAGERS, P. (1989). General branching processes as Markov fields. *Stoch. Proc. Appl.* **32,** 183–242.

[14] JAGERS, P. (1992). Stabilities and instabilities in population dynamics. *J. Appl. Prob.* **29,** 770–780.

[15] JAGERS, P. AND NERMAN, O. (1996). The asymptotic composition of supercritical muti-type branching populations. In *Séminaire de Probabilités XXX*, ed. J. Azéma, M. Emery, and M. Yor. *Lecture Notes in Mathematics*, vol. 1626. Springer, Berlin pp. 40–54.

[16] KESTEN, H. AND STIGUM, B. P. (1966). A limit theorem for multidimensional Galton-Watson processes. *Ann. Math. Statist.* **37,** 1211–1233.

[17] KINGMAN, J. F. C. (1982). The coalescent. *Stoch. Proc. Appl.* **13,** 235–248.

[18] KINGMAN, J. F. C. (1982). On the genealogy of large populations. *J. Appl. Prob.* **19A,** 27–43.

[19] KURTZ, T., LYONS, R., PEMANTLE, R. AND PERES, Y. (1997). A conceptual proof of the Kesten-Stigum theorem for multi-type branching processes. In *Classical and Modern Branching Processes*, ed. K. B. Athreya and P. Jagers. Springer, New York pp. 181–185.

[20] LYONS, R., PEMANTLE, R. AND PERES, Y. (1995). Conceptual proofs of LlogL criteria for mean behaviour of branching processes. *Ann. Prob.* **23,** 1125–1138.

[21] MATTHES, K., KERSTAN, J. AND MECKE, J. (1978). *Infinitely Divisible Point Processes.* Wiley, Chichester.

[22] MOEHLE, M. (2000). Ancestral processes in population genetics - the coalescent. *J. Theor. Biol.* **204,** 629 – 638.

[23] VARADHAN, S. R. S. (1988). Large deviations. In *École d'Été de Probabilités de Saint-Flour XV-XVII, 1985*, ed. P. L. Hennequin. *Lecture Notes in Mathematics*, vol. 1362. Springer, Berlin pp. 1–49.