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 nonextinction), 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 Kurtz et al.) 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

When looking at the time evolution of a population there are two possible perspectives: either forward or backward in time. In the first case we observe the characteristics of the population at a given time t and ask 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 a review, see [8]). By way of contrast, the backward—or retrospective—view 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] and [18], and [22] for a review), the backward version of the Moran model. As was observed by Jagers [14] and Jagers and Nerman [15] for example, it is also rewarding to study the backward aspects of multitype branching processes; this point of view has turned out to be crucial in recent biological applications [12]. It is the aim of this article to pursue this last line of research further. We do so in continuous time because this gives us the opportunity

Received 7 February 2003; revision received 25 April 2003.

∗ Postal address: Institut für Mathematik, Universität München, Theresienstraße 39, D-80333 München, Germany. Email address: georgii@mathematik.uni-muenchen.de

** Postal address: Institut für Mathematik und Informatik, Universität Greifswald, Jahnstraße 15a, D-17487 Greifswald, Germany.
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 $t$ or both $t$ and $u$ tend to infinity, both the time average and the population average of ancestral types converge to a particular distribution $\alpha$ almost surely on nonextinction (Theorems 3.1 and 3.2).

This $\alpha$ will be called the *ancestral distribution of types*; its components are $\alpha_i = \pi_i h_i$, where $\pi$ 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 call the *retrospective mutation chain*, a particular continuous-time Markov chain on the type space with $\alpha$ as its invariant distribution. We will show the following:

- 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 nonextinction (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 et al. [20] and Kurtz et al. [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 [6], [7]. 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 by Athreya and Ney [2, Chapter 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 children of type $j$, and $\mathbb{Z}_+ = \{0, 1, \ldots\}$. We assume that the mean offspring matrix $M = [m_{ij}]_{i,j \in S}$ is irreducible.

According to Harris [11, Chapter VI], the associated random family tree can be constructed as follows. Let $X = \bigcup_{n \geq 0} X_n$, where $X_n$ describes the virtual $n$th generation. That is, $X_0 = S$, and $i_0 \in X_0$ specifies the type of the root, i.e. the founding ancestor. Next, $X_1 = S \times N$, and the element $x = (i_1; \ell_1) \in X_1$ is the $\ell_1$th child of type $i_1$ of the root. Finally, for $n > 1$, $X_n = S^n \times \mathbb{N}^n$, and $x = (i_1, \ldots, i_n; \ell_1, \ldots, \ell_n) \in X_n$ is the $\ell_n$th child of type $i_n$ of its parent $\tilde{x} = (i_1, \ldots, i_{n-1}; \ell_1, \ldots, \ell_{n-1})$; see Figure 1. We write $\sigma(x) = i_n$ for the type of $x \in X_n$. 
Figure 1: A realization of the branching process. Here, $S = \{a, b, c\}$, with solid lines indicating individuals of type $a$, dashed lines individuals of type $b$ and dotted lines individuals of type $c$. Siblings of the same type are counted from top to bottom. The set $X(s)$ consists of all edges that intersect the vertical line at $s$ (indicated by $\times$). The type frequencies in the population are counted by $Z(s)$.

The set $X(x, t) = \{(b, a; 1, 1), (b, b; 1, 1), (b, b; 1, 2), (c; 1)\}$ consists of all edges that emanate from $x = (b, a; 2, 1)$ and hit the vertical line at $t$ (indicated by $\circ$).

With each $x \in X$ we associate

(i) its random lifetime $\tau_x$, distributed exponentially with parameter $a_\sigma(x)$, and

(ii) its random offspring $N_x = [N_{x,j}]_{j \in S} \in \mathbb{Z}_+^S$ with distribution $p_\sigma(x)$ such that the family $\{\tau_x, N_x : x \in X\}$ is independent.

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

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

where $i_0$ is the prescribed type of the root. The random variables $\tau_x$ provide the proper time scale. Namely, for $x \in X$, let the splitting times $T_x$ be defined recursively by $T_{\tilde{x}} = T_{\tilde{x}} + \tau_x$ with $T_{i_0} := 0$. 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 $t$. The resulting tree can be visualized by identifying each $x \in X$ with an edge from $\tilde{x}$ to $x$ with length $\tau_x$ in the direction of time.

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

When $0 < s < t$ 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

$$X(x, t) = \{y \in X : xy \in X(t)\} \quad (2.1)$$
denote the set of descendants of $x$ living at time $t$; see Figure 1. In the above, the concatenation $xy$ of two strings $x, y \in X$ is defined in the obvious way, and the empty string is considered as an ancestor of type $\sigma(x)$; that is, $X(x, t) = \{\sigma(x)\}$ 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 $P^{\sigma(x)}$. We will also consider the counting measures

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

on $S$, where $\delta_i$ is the Dirac measure at $i$. The measures $Z(t)$ and $Z(x, t)$ count the type frequencies in the population $X(t)$ and the subpopulation $X(x, t)$ of descendants of $x$ respectively. 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 (see [2, p. 202, Equation (9)]) that $E^i(Z_j(t)) = (e^{tA})_{ij}$ for all $i, j \in S$, where the generator matrix $A = [aij]_{i,j \in S}$ is given by

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

By the irreducibility of $M$, $A$ is also irreducible, so that the first-moment matrix $[E^i(Z_j(t))]_{i,j \in S}$ 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 $\lambda$ (namely a real eigenvalue exceeding the real parts of all other eigenvalues) and associated positive left and right eigenvectors $\pi$ and $h$ which will be normalized so that $\langle \pi, 1 \rangle = 1 = \langle \pi, h \rangle$. Here we think of the row vector $\pi$ 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

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

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 global quantities $\lambda, \pi, h$. 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 theorem.) Consider the supercritical case $\lambda > 0$.

(a) For all $i \in S$,

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

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

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

and $P^i(W > 0) > 0$ for all $i$ if and only if

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

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

**Note:** The superscript $i$ denotes the type of the process. The notation $X(t)$ represents the state of the population at time $t$, where $X(t)$ is the set of individuals alive at time $t$. The empty set $\emptyset$ represents the absence of any individuals. The exponential distributions are used to model the time between events, such as births, deaths, or other type changes.
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 $\pi$ 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 the 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 individual of type $i$. In fact, $h_i$ is also characterized by the limiting relation

$$E^i(\lvert X(t) \rvert) e^{-\lambda t} \to h_i \quad \text{as } t \to \infty; \quad (2.5)$$

see 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 $\lambda, \pi, h$. 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, Proposition 1], and Hermisson et al. [12], this probability vector describes the distribution of ancestral types of an equilibrium population with type frequencies given by $\pi$. The vector $\alpha$ will therefore be called the \textit{ancestral distribution}. Our results below shed some additional light on the significance of $\alpha$.

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 $\alpha$. Specifically, let $0 < u < t$ and let

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

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

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

$$\lim_{u \to \infty} \lim_{t \to \infty} A_u^x(t) = \alpha \quad \mathcal{P}^i\text{-almost surely on } \Omega_{\text{surv}}. \quad (3.2)$$

The proof will be given in Section 5.2. A slightly weaker result under slightly stronger conditions (convergence in probability under the assumption (2.4)) follows immediately from Corollary 4 of [15], where very general population averages are considered.

\textbf{Remark 3.1.} The assertion (3.2) means that, for each $j \in S$, the average

$$A_u^j(t) = \frac{1}{|X(t)|} \sum_{x \in X(t)} 1_{\{\sigma(x(t-u)) = j\}}$$
Multitype branching: the ancestral types

1095

(with $1_{\cdot}$ denoting the indicator function) converges to $\alpha_j$ $P^i$-almost surely on $\Omega_{\text{surv}}$ as $t \to \infty$ and $u \to \infty$ in this order. Letting $s = t - u$, we can rewrite this average in the form

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

where $X(x, t)$ is given by (2.1). The numbers $|X(x, t)|$ with $x \in X_j(s)$ are independent and identically distributed (i.i.d.) with mean $E_j(|X(u)|)$. Assuming the validity of a law of large numbers and using Theorem 2.1(a) and (2.5), we can conclude that the above average 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 $\alpha$.

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 $\sigma(x[0, t]) = (\sigma(x(s)))_{0 \leq s \leq t}$. 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

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

**Remark 3.2.** (a) According to the portmanteau theorem [9, p. 108, Theorem 3.1], (3.3) is equivalent to the assertion that, $P^i$-almost surely on $\Omega_{\text{surv}},$

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

and it is sufficient to check this in the case when $F = \{v \in \mathcal{P}(S) : \|v - \alpha\| \geq \epsilon\}$ with arbitrary $\epsilon > 0$. 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 $\alpha$.

(b) Theorem 3.2 involves a population average of time averages. So we 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 $\Omega_{\text{surv}}$.

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 3.1. 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 \(G = [g_{ij}]_{i,j \in S}\) 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
\[
a_i \sum_{j \in S} m_{ij} h_j = \sum_{j \in S} (a_i \delta_{ij} + a_{ij}) h_j = (a_i + \lambda) h_i \quad \text{by (2.3)}.
\]
Since \(M\) is irreducible by assumption, \(G\) is irreducible as well. It is also immediate that the ancestral distribution \(\alpha\) 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 \(\pi\) and following their lines of descent backward in time. This gives the transition rates
\[
\bar{g}_{ij} = \pi_j (a_{ji} - \lambda \delta_{ij}) \pi_i^{-1} = a_j g_{ji} a_i^{-1},
\]
which correspond to the time reversal of the retrospective mutation chain.

To set the stage for Theorem 3.3, we let \(\Sigma = D(\mathbb{R}, S)\) denote the space of all doubly infinite càdlàg paths in \(S\). The space \(\Sigma\) will be equipped with the usual Skorohod topology which turns \(\Sigma\) into a Polish space; see e.g. Section 3.5 of \[9\], and in particular Theorem 5.6, for the case of the time interval \([0, \infty[\). The associated Borel \(\sigma\)-algebra coincides with the \(\sigma\)-algebra generated by the evaluation maps \(\sigma \mapsto \sigma(t)\) from \(\Sigma\) to \(\mathbb{R}\), \(t \in \mathbb{R}\) \[9, p. 127, Proposition 7.1\].

The time shift \(\vartheta_s\) on \(\Sigma\) is defined by
\[
\vartheta_s \sigma(t) = \sigma(t + s), \quad s, t \in \mathbb{R}, \sigma \in \Sigma.
\]
We write \(\mathcal{P}_\Theta(\Sigma)\) for the set of all probability measures on \(\Sigma\) 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 \[9, p. 101, Theorem 1.7\].

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
\[
\sigma(x)_{t, \text{per}}(s) = \sigma(x(s_t)), \quad s \in \mathbb{R},
\]
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
\[
R^x(t) = \frac{1}{t} \int_0^t \delta_{\vartheta_s \sigma(x)_{t, \text{per}}} \, ds \in \mathcal{P}_\Theta(\Sigma).
\]

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 the empirical distribution of the mapping \(x \mapsto R^x(t)\) from \(X(t)\) to \(\mathcal{P}_\Theta(\Sigma)\), i.e. the population average
\[
\Gamma(t) := \frac{1}{|X(t)|} \sum_{x \in X(t)} \delta_{R^x(t)}.
\]
Multitype branching: the ancestral types

(As before, this definition requires that $X(t) \neq \emptyset$.) The population average $\Gamma(t)$ is a random element of $\mathcal{P}(\mathcal{P}_0(\mathcal{S}))$, the set of all probability measures on the Polish space $\mathcal{P}_0(\mathcal{S})$, which is again equipped with the weak topology. In Section 5.3 we will prove the following result.

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

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

where $\mu \in \mathcal{P}_0(\mathcal{S})$ is the distribution of the stationary (doubly infinite) retrospective mutation chain $(\sigma(t))_{t \in \mathbb{R}}$ with generator $G$ and invariant distribution $\alpha$.

**Remark 3.3.** As in Remark 3.2(a), the portmanteau theorem implies that (3.5) is equivalent to the assertion that, $\mathbb{P}^i$-almost surely on $\Omega_{\text{surv}}$, $\Gamma(t)(F) \to 0$ for every closed $F \subset \mathcal{P}_0(\mathcal{S})$ such that $\mu / \not\in F$. Writing $d(\cdot, \cdot)$ for any metric metrizing the weak topology on $\mathcal{P}_0(\mathcal{S})$, this in turn means that, for each $\varepsilon > 0$,

$$\lim_{t \to \infty} \frac{1}{|X(t)|} \sum_{x \in X(t)} 1_{\{d(Rx(t), \mu) \geq \varepsilon\}} = 0 \text{ $\mathbb{P}^i$-almost surely on } \Omega_{\text{surv}}.$$  

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 $Rx(t)$ is close to $\mu$ in the weak topology. Theorem 3.3 also highlights the retrospective nature of our mutation chain: it describes the evolution of types along those lines of descent which survive until time $t$ (and can thus be seen when an individual alive at time $t$ 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 et al. [20] and Kurtz et al. [19]. Informally, this is a tree with a randomly selected trunk (or spine) along which time runs at a different rate and offspring are weighted according to their 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 $\lambda$ can have arbitrary sign.

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

$$\hat{p}_i(\kappa) = \frac{\langle \kappa, h \rangle p_i(\kappa)}{c_i h_i}, \quad \kappa \in \mathbb{Z}^S_+,$$

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

$$\sum_{\kappa \in \mathbb{Z}^S_+} \langle \kappa, h \rangle p_i(\kappa) = \sum_{j \in S} m_{ij} h_j = \sum_{j \in S} \left( \delta_{ij} + \frac{a_{ij}}{a_i} \right) h_j = c_i h_i$$

by (2.3); note that $c_i$ is automatically positive. Next, when an individual of type $i$ on the trunk has offspring $N_i = [N_j]_{j \in S}$ with distribution $\hat{p}_i$, 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 \( \hat{N}_{ij} h_j / \langle \hat{N}_i, h \rangle \) for a given offspring, and with probability

\[
p_{ij} = E \left( \frac{\hat{N}_{ij} h_j}{\langle \hat{N}_i, h \rangle} \right) = \frac{m_{ij} h_j}{c_{ij} h_i}
\]
on average. These are precisely the jump probabilities of the retrospective mutation chain. Finally, the lifetime of an individual of type \( i \) 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 backward; 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, let \( \{ \hat{\tau}_n, \hat{N}_n, \hat{\xi}_n : n \geq 0 \} \) be a sequence of random variables with values in \([0, \infty[ \times \mathbb{Z}_+ \times \mathbb{X}\) respectively such that, for a given type \( i_0 = i \) of the root,

(i) \( \hat{\xi}_0, \hat{N}_0 \) are independent, \( \hat{\tau}_0 \) has exponential distribution with parameter \( a_i + \lambda \), \( \hat{N}_0 \) has distribution \( \hat{p}_i \), and \( \hat{\xi}_1 \) has conditional distribution

\[
Pr(\xi_1 = (i_1; \ell_1) \mid \hat{N}_0, \hat{\tau}_0) = \frac{h_{i_1}}{\langle \hat{N}_0, h \rangle} 1_{\{\ell_1 \leq \hat{\tau}_0 \}}
\]

for all \( (i_1; \ell_1) \in \mathbb{X}_1 \);

(ii) for any \( n \geq 1 \), conditionally on \( \mathbb{F}_{n-1} = \sigma\{ \hat{\tau}_k, \hat{N}_k, \hat{\xi}_k : k < n\} \), \( \hat{\xi}_n \) and \( \hat{N}_n \) are independent, \( \hat{\tau}_n \) follows an exponential law with parameter \( a_{\sigma(\hat{\xi}_n)} + \lambda \), \( \hat{N}_n \) follows the law \( \hat{p}_{\sigma(\hat{\xi}_n)} \), and

\[
Pr(\xi_{n+1} = (\xi_n; \ell_{n+1}, \ell_{n+1}) \mid \mathbb{F}_{n-1}, \hat{\tau}_n, \hat{N}_n) = \frac{h_{\ell_{n+1}}}{\langle \hat{N}_n, h \rangle} 1_{\{\ell_{n+1} \leq \hat{\tau}_n \}}
\]

for all \( (\xi_n; \ell_{n+1}, \ell_{n+1}) \in \mathbb{F} \times \mathbb{N} \), that is, \( \xi_{n+1} \) is a child of \( \xi_n \) selected randomly with weight proportional to \( h_{\sigma(\xi_{n+1})} \).

Define \( \hat{X} = \bigcup_{n \geq 0} \hat{X}_n \subset \mathbb{X} \) recursively by \( \hat{X}_0 = \{ i \} \) and \( \hat{X}_n = \hat{X}_n^+ \cup \hat{X}_n^- \) with

\[
\hat{X}_n^+ = \{ (\xi_{n-1}; i_n, \ell_n) \in \mathbb{X} : \ell_n \leq \hat{N}_{n-1,i_n} \}
\]

the offspring of \( \xi_{n-1} \), and

\[
\hat{X}_n^- = \{ (\hat{x}; i_n, \ell_n) \in \mathbb{X} : \hat{x} \in \hat{X}_{n-1} \setminus \{ \xi_{n-1} \}, \ell_n \leq N_{\hat{x}, i_n} \}
\]

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

\[
\hat{X}(t) = \{ x \in \hat{X} : \hat{T}_{\hat{x}} \leq t < \hat{T}_{\hat{x}} \}.
\]
The selected trunk individual at time \( t \) is \( \xi(t) = \xi_n \) if \( \hat{T}_{k_{n-1}} \leq t < \hat{T}_{k_n} \), and the process \((\hat{X}(t), \xi(t))_{t \geq 0} \) in \( \Omega := \{0, \infty\} \times \Omega \times D([0, \infty), \mathbb{X}) \) describes the size-biased tree with trunk \((\xi(t))_{t \geq 0} \). As we have emphasized above, the type process along the trunk, \( \sigma(t) := \sigma(\xi(t)) \), 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 \( \hat{P}^\delta_\alpha \) for the distribution of \((\hat{X}(t), \xi(t))_{t \geq 0} \) on \( \Omega_\alpha \), and \( \hat{P}^\delta \) for its marginal, the distribution of \((\hat{X}(t))_{t \geq 0} \) on \( \Omega \). The representation theorem below establishes the relationship between \( P^\delta_\alpha \), \( \hat{P}^\alpha \), and the retrospective mutation chain. We use the shorthand \( (y(s))_{0 \leq s \leq t} \) for a path.

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

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

Recall that this theorem is valid for \( \lambda \) with arbitrary sign. The proof is postponed until Section 5.1. Here we discuss some immediate consequences and possible extensions.

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

\[
\mathbb{E}^i(Z_j(t)e^{-\lambda t} = h_i \hat{P}^\alpha_\alpha(\sigma(\xi(t)) = j) h_j^{-1} \to h_i \alpha_j h_j^{-1} = h_i \pi_j \quad \text{as } t \to \infty. \tag{4.3}
\]

In particular, (2.5) follows by summation over \( j \).

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

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

with

\[
W(t) := (Z(t), h) e^{-\lambda t}.
\]

In particular, \( h_i = \mathbb{E}^i(W(t)) \). Thus, on the \( \sigma \)-algebra \( \mathcal{F}_t \) generated by \( X[0, t], \hat{P}^\delta \) is absolutely continuous with respect to \( P^\delta \) with density \( W(t)/h_i \), and \( (W(t))_{t \geq 0} \) is a martingale with respect to \( P^\delta \). 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 let

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

the finite random subset of \( D([0, t], \mathbb{X}) \) which describes the lineages that survive until time \( t \).

Also, let \( C_j \) be the measure on \( \mathbb{P}_f(D([0, t], \mathbb{X})) \times D([0, t], \mathbb{X}) \) with Radon–Nikodym density \( e^{\lambda t} h_i h_{\sigma(\xi(t))}^{-1} \), relative to the joint distribution of \( \Phi(t) = \{ x[0, t] : x \in \hat{X}(t) \} \) and \( \xi[0, t] \) under \( \hat{P}^\delta \). Theorem 4.1 then implies that

\[
\mathbb{E}^i \left( \sum_{\psi \in \Phi(t)} F(\Phi(t), \psi) \right) = \int F(\psi, \psi) C_j(d\psi, d\psi)
\]

for any measurable \( F \geq 0 \), that is, \( C_j \) is the Campbell measure of \( \Phi(t) \) under \( P^\delta \). 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 [10, Theorem 2] in the context of critical multitype branching. The size-biased offspring distribution associated with this case is

\[
\tilde{p}_i(\kappa) = \frac{\|\kappa\|p_i(\kappa)}{m_i}, \quad \kappa \in \mathbb{Z}^S_+,
\]

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 individual of type \( i \) 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 circumflex) to characterize all quantities of the associated size-biased tree, we arrive at the following counterpart of (4.2):

\[
E_i\left( \sum_{x \in X(t)} e^{-t(L^X(t),r)} F(X[0,t],x[0,t]) \right) = \tilde{E}_i(F(\tilde{X}[0,t],\tilde{\xi}[0,t])). \tag{4.4}
\]

In the above, \( r \) is the vector with \( i \)th 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, \( \tilde{\sigma}(t) := \sigma(\tilde{\xi}(t)) \), 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 that 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 \( \zeta_i \) of split events during the sojourn in state \( i \) is then Poisson with parameter \( a_i \nu_i \), where \( \nu_i \) 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

\[
\prod_{i \in S} E(m_{\zeta_i}) = e^{t\langle \nu,r \rangle}.
\]

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 \( F(X[0,t],x[0,t]) = \exp[t(L^X(t),r)] 1_{\{\sigma(x(t))=j\}} \), (4.4) becomes

\[
E'_i(Z_j(t)) = \tilde{E}_i(e^{t(L^X(t),r)} 1_{\{\tilde{\sigma}(\tilde{\xi}(t))=j\}}), \tag{4.5}
\]

which is a version of the Feynman–Kac formula. Indeed, consider the function \( u(t,i) = E'_i(Z_j(t)) \) for fixed \( j \). Since \( u(t,i) = (e^{tA})_{ij} \), it follows that \( u(t,i) \) is the unique solution of the Cauchy problem

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

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 [4, p. 32] or [23, Theorem 2.1]) together with (2.5) we arrive at the variational principle

$$
\lambda = \lim_{t \to \infty} \frac{1}{t} \log E^t(|X(t)|) = \max_{v \in \mathcal{P}(S)} \{(v, r) - I_G(v)\},
$$

where $I_G$ is the large deviation rate function for the empirical distribution of the Markov chain with transition rates $\tilde{g}_{ij}$; see (5.10) below 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 $\alpha$. This variational principle is behind that found in [12].

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

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

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 Theorem 4.1 and Theorem 2.1(b). For the former we do not require that $\lambda$ is positive.

**Proof of Theorem 4.1.** It is sufficient to show that

$$
\hat{E}^i(F(\hat{X}[0, t], \xi[0, t]); \xi(t) = x) = e^{-\lambda t} h_{\sigma(x)} E^i(F(X[0, t], x[0, t]); x \in X(t))
$$

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

Consider the right-hand side of (5.1) and write $e^{-\lambda t} h_{\sigma(x)} = q_1 q_2 q_3$ 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 in 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 $\tau_{x_k}$ is replaced by $\hat{\tau}_k$ for $k = 0, \ldots, n$.

Indeed, the associated Radon–Nikodym density is

$$
\tilde{q}_1 = e^{-\lambda \hat{T}_x} \prod_{k=0}^{n} c_{i_k}.
$$

Conditioning $\tilde{q}_1$ on the tree $X[0, t]$ up to time $t$ and using the loss-of-memory property of the exponential law of $\tau_x$, we find that, almost surely on $\{T_{\hat{\tau}_k} \leq t\}$,

$$
E^i(\tilde{q}_1 | X[0, t]) = e^{-\lambda t} E^i(e^{-\lambda (T_{\hat{\tau}_k} - t)} | T_{\hat{\tau}_k} > t) \prod_{k=0}^{n} c_{i_k} = q_1.
$$
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 $\hat{N}_k$ for $k = 0, \ldots, n - 1$. Finally, $q_3$ is the conditional selection probability for the trunk:

$$q_3 = \tilde{P}^i(\xi_{k+1} = x_{k+1} \mid \hat{X}[0, t]).$$

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

$$\hat{E}_i^j(F(\hat{X}[0, t], x[0, t]); \hat{T}_x \leq t < \hat{T}_x, \xi_{k+1} = x_{k+1} \mid 0 \leq k < n) = \hat{E}_i^j(F(\hat{X}[0, t], \xi[0, t]); \xi(t) = x),$$

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, $P^i$-almost surely for each $i$, to a finite limiting variable $W \geq 0$. 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 the condition (2.4) holds. There are two possible routes to achieve this.

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

Or, more naturally, we 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], we observe first that $W$ is nontrivial if and only if $\hat{P}^i$ is absolutely continuous with respect to $P^i$ (with Radon–Nikodym density $W/h_i$), which is the case if and only if

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

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

To check that (5.2) is equivalent to (2.4), notice first that (2.4) is equivalent to $E(\log(\hat{N}_i, h)) < \infty$ for all $i \in S$, (5.3)

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

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

$$\limsup_{t \to \infty} \hat{W}(\hat{T}_n) \geq \limsup_{l \to \infty} (\hat{N}_n, h)e^{-l\hat{T}_n} = \infty \quad \text{almost surely,}$$

so that (5.2) fails.

Conversely, suppose that (5.3) holds. As in Section 4, we consider the offspring $\hat{X}_{n+1}$ of the trunk created at time $\hat{T}_n$ having type counting measure $\hat{N}_n$. We also introduce the
\(\sigma\)-algebra \(\mathcal{T}\) generated by the trunk variables \(\{\tilde{T}_{(n)}, \tilde{N}_{(n)} : n \geq 0\}\) 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),

\[
\tilde{Z}_{t}^\delta(x) = \sum_{n: \tilde{T}_{(n)} \leq t} e^{-\lambda \tilde{T}_{(n)}} \mathbb{E}_{x}^{\tilde{T}_{(n)}} \left( \sum_{x \in X_{n+1}^{\delta}} \langle Z(x, t), h \rangle e^{-\lambda t - \tilde{T}_{(n)}} \right) \mathcal{T}
\]

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 \(n^{-1} \log(\tilde{N}_{(n)}, h) \to 0\) almost surely. On the other hand, \(\lim \inf_{n \to \infty} \tilde{T}_{(n)} / n \to 0\) by the law of large numbers, whence

\[
\lim \inf_{n \to \infty} \sum_{n \geq 0} e^{-\lambda \tilde{T}_{(n)}} < \infty \quad \text{almost surely.}
\]

This means that, conditionally on \(\mathcal{T}\), \(\tilde{W}(t)\) 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, \(\mathbb{E}[W > 0] = \Omega_{\text{surr}}\), almost surely, follows from the trivial inclusion \(\mathbb{E}[W > 0] \subseteq \Omega_{\text{surr}}\) and the well-known fact that \(q_i = \mathbb{P}(W = 0)\) solves the equation \(q_i = \mathbb{E}(\prod_{j \in S} q_{N_j}^{\delta})\) which has the extinction probabilities as unique nontrivial solution [2, p. 205, Equation (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 \(X(x, [t, t+u]) = (X(x, t+s))_{0 \leq s \leq u}\) describes the subtree of descendants of \(x\) during the time interval \([t, t+u]\).

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

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

**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 \(\delta\) is so large that \(u < \delta\) and \(\rho := \mathbb{E}^j(Z_j(\delta)) > 1\). Such a \(\delta\) exists because \(\lambda > 0\) and \(A\) is irreducible. Let \(\mathcal{F}_{n\delta}\) denote the \(\sigma\)-algebra generated by \(X[0, n\delta]\). Since \(u < \delta\), for each \(n \geq 1\) 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 \((\varphi_l)_{l \geq 1}\) on \(\Omega_{\text{surr}}\) 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 \(\mathbb{P}^i\)-almost surely on \(\Omega_{\text{surr}}\), \(\lim_{k \to \infty}(1/k) \sum_{l=1}^{k} \varphi_l = c_j\), 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 \(\Omega_{\text{surr}}\) as \(n \to \infty\); here \(\Psi_n = \sum_{l=1}^{n} \varphi_l\) with \(\varphi_l = Z_j(l\delta)\).
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 \(\Omega_{\text{surv}}\). By Lemma 4 of [19], it follows that \(\lim \inf_{l \to \infty} \frac{\psi_{l+1}}{\psi_l} \geq \rho\) almost surely on \(\Omega_{\text{surv}}\). This implies that

\[
\lim \sup_{n \to \infty} \frac{\psi_{n-1}}{\psi_n} = \lim \sup_{n \to \infty} \sum_{l=1}^{n-1} \frac{\psi_l}{\psi_n} < \infty
\]

almost surely on \(\Omega_{\text{surv}}\). As

\[
\frac{1}{\psi_n} \sum_{x \in X_j(n\delta)} \varphi_{n,x} = A_n + (A_n - A_{n-1}) \frac{\psi_{n-1}}{\psi_n},
\]

the proposition follows in the case of large \(\delta\).

If \(\delta > 0\) is arbitrary, we choose some \(k \in \mathbb{N}\) such that \(\delta' := k\delta\) is as 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

\[
\lim_{n \to \infty} \frac{1}{\psi_{nk+l}} \sum_{x \in X_j((nk+l)\delta)} \varphi_{nk+l,x} = c_j P^i\text{-almost surely on } \Omega_{\text{surv}},
\]

and the proof is complete.

A typical application of Proposition 5.1 is the following corollary. Consider the \(X_j(s)\)-averaged type counting measure

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

(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 \(i, j \in S\), as \(n \to \infty\),

\[
C_{j,u}(n\delta) \to E^i(Z(u)) \quad P^i\text{-almost surely on } \Omega_{\text{surv}}.
\]

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

**Lemma 5.1.** Given \(\epsilon > 0\), there exists some \(\delta > 0\) such that, for all \(i, j \in S\) and \(k \in \mathbb{N}\), we have, \(P^i\text{-almost surely on } \Omega_{\text{surv}},

\[
\lim \sup_{n \to \infty} \sup_{n \delta \leq s \leq (n+1)\delta} \frac{\|Z(s)\|}{\|Z(n\delta)\|} < 1 + \epsilon,
\]

(5.5)

and

\[
\lim \inf_{n \to \infty} \inf_{n \delta \leq s \leq (n+1)\delta} \frac{Z_j(s)}{Z_j(n\delta)} > 1 - \epsilon,
\]

(5.6)

and

\[
\lim \inf_{n \to \infty} \inf_{n \delta \leq s \leq (n+1)\delta} \inf_{k \delta \leq u \leq (k+1)\delta} \frac{\sum_{y \in X_j(s)} Z(y, s + u)}{\sum_{y \in X_j(n\delta)} Z(y, n\delta + k\delta)} > 1 - \epsilon.
\]

(5.7)
Proof. We begin by proving the upper bound (5.5). When \( n \delta \leq s \leq (n + 1) \delta \), we can write

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

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

\[
\sup_{n \delta \leq s \leq (n + 1) \delta} \frac{\| Z(s) \|}{\| Z(n \delta) \|} \leq \max_{j \in S} \frac{1}{Z_J(n \delta)} \sum_{x \in X(n \delta)} M(x, [n \delta, (n + 1) \delta]).
\]

By Proposition 5.1, the last expression converges to \( m(\delta) := \max_{j \in S} E^j(M(0, [0, \delta])) \) almost surely on \( \Omega_{\text{surv}} \). Now, \( M(0, [0, \delta]) \) is dominated by the total size at time \( \delta \) of the modified branching process for which the random variables \( N_x,\sigma(x) \) in Section 2 are replaced by \( N_x,\sigma(x) \lor 1 \), so that each individual has at least one offspring of its own type. The latter process has a finite generator matrix, say \( A^+ \). Hence, \( m(\delta) \leq \max_j (e^{\delta A^+} 1)_j \rightarrow 1 \) as \( \delta \rightarrow 0 \). 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 \( n \delta \leq s \leq (n + 1) \delta \) and \( k \delta \leq u \leq (k + 1) \delta \). 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

\[
\sum_{y \in X(s)} \| Z(y, s + u) \| \geq \sum_{x \in X(n \delta)} 1_{\tau_x, n \delta > \delta} \sum_{z \in X(x, (n + k) \delta)} 1_{\tau_z, (n + k + 2) \delta > 2 \delta}.
\]

Here we write \( \tau_{x,i} = \inf\{ u > 0 : \sigma(x) \notin X(t + u) \} = T_x - t \) for the remaining lifetime of \( x \in X(t) \) after time \( t \). Proposition 5.1 therefore implies that, \( P^\delta \)-almost surely on \( \Omega_{\text{surv}} \), the left-hand side of (5.7) is at least

\[
\frac{E^\delta(1_{\tau_{x,i} > \delta} \sum_{z \in \mathcal{X}(k \delta)} 1_{\tau_{z,i} > 2 \delta})}{E^\delta(|X(k \delta)|)}.
\]

(5.8)

By the Markov property, the numerator is equal to

\[
E^\delta \left( 1_{\tau_{x,i} > \delta} \sum_{z \in \mathcal{X}(k \delta)} \exp[-2 \delta a_{\sigma(z)}] \right),
\]

which is no less than

\[
e^{-2 \delta a_{\sigma}} E^\delta(1_{\tau_{x,i} > \delta} |X(k \delta)|),
\]

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

\[
\varepsilon_k = \frac{E^\delta(1_{\tau_{x,i} > \delta} |X(k \delta)|)}{E^\delta(|X(k \delta)|)}.
\]

For \( k = 0 \) we have \( \varepsilon_0 = 1 - e^{-\delta a_{\sigma}} \). For \( k \geq 1 \) we can use Theorem 4.1 to obtain

\[
\varepsilon_k = \frac{E^\delta(1_{\tau_{x,i} > \delta} h_{\sigma(\xi(k \delta))}^{-1})}{E^\delta(h_{\sigma(\xi(k \delta))})} \leq \max_i h_i(1 - e^{-\delta a_{\sigma} + \lambda_\delta}).
\]

Hence, if \( \delta \) 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 Theorem 3.1.

**Proof of Theorem 2.1(a).** Essentially we reproduce here the argument of [19]. Let $\varepsilon > 0$ be given and let $\varepsilon' > 0$ be such that, for every $v \in \mathcal{P}(S)$, $\|v - \pi\| < \varepsilon$ whenever $\|av - \pi\| < \varepsilon'$ for some $a > 0$. Let $\delta > 0$ be as small as required in Lemma 5.1. According to (4.3), we can choose some $u > 0$ so large that
$$
\|E_j(Z(u)e^{-\lambda u}) - h_{j,\pi}\| < \varepsilon' \min_{i \in S} h_i
$$
for all $j \in S$. Corollary 5.1 then implies that, $P_i$-almost surely on $\Omega_{\text{surv}}$,
$$
\|C_{j,u}(s)e^{-\lambda u} - h_{j,\pi}\| < \varepsilon' \min_{i \in S} h_i
$$
for all sufficiently large $s \in \delta \mathbb{N}$. Writing
$$
\Pi(t) = \frac{Z(t)}{\|Z(t)\|} \quad \text{and} \quad a(t) = \frac{\|Z(t)\|e^{-\lambda u}}{(Z(t-u), h)} \quad \text{for } t > u,
$$
we conclude that
$$
\|a(t)\Pi(t) - \pi\| \leq \frac{1}{(Z(t-u), h)} \sum_{j \in S} Z_j(t-u)\|C_{j,u}(t-u)e^{-\lambda u} - h_{j,\pi}\| < \varepsilon',
$$
and therefore $\|\Pi(t) - \pi\| < \varepsilon$ for all sufficiently large $t \in \delta \mathbb{N}$ almost surely on $\Omega_{\text{surv}}$. 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 almost surely on $\Omega_{\text{surv}}$. Since $\varepsilon$ 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 $\alpha^u_j = \pi_j E_j(\|Z(u)\|)e^{-\lambda u}$. Since $\alpha^u \to \alpha$ as $u \to \infty$ by (2.5), it is sufficient to show that
$$
P_i(\text{for all } u > 0, A^u(t) \to \alpha^u \text{ as } t \to \infty \mid \Omega_{\text{surv}}) = 1. \quad (5.9)
$$
Fix any $j \in S$, $u > 0$, and $\delta > 0$. By Corollary 5.1, $P^i$-almost surely on $\Omega_{\text{surv}}$,
$$
\|C_{j,u}(s)\| \to E_j(\|Z(u)\|) \quad \text{as } s \to \infty \text{ through } \delta \mathbb{N}.
$$
Combining this with Remark 3.1 and Theorem 2.1(a), and again writing
$$
\Pi_j(s) := \frac{Z_j(s)}{\|Z(s)\|},
$$
we obtain that
$$
A^u_j(s + u) = \frac{Z_j(s)\|C_{j,u}(s)\|}{|X(s + u)|} = \frac{\Pi_j(s)\|C_{j,u}(s)\|}{\sum_{k \in S} \Pi_k(s)\|C_{k,u}(s)\|} \to \frac{\pi_j E_j(\|Z(u)\|)}{E^u(\|Z(u)\|)} = \alpha^u_j \quad \text{as } s \to \infty, s \in \delta \mathbb{N}, P^i\text{-almost surely on } \Omega_{\text{surv}}.
$$
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
$$
P^i(\text{for all } u > 0 \text{ and all } j \in S, \liminf_{t \to \infty} A^u_j(t) > (1 - 2\varepsilon)\alpha^u_j \bigg| \Omega_{\text{surv}} \bigg) = 1,
$$
where the $u$-uniformity in (5.7) allows us to bring the $u$-quantifier inside the probability operator. This gives (5.9) because $\varepsilon$ is arbitrary and $A^u(t)$ and $\alpha^u$ 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 Theorem 4.1 (the representation theorem) 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 $\nu \in \mathcal{P}_0(\Sigma)$, let

$$H_G(\nu) = \sup_{t > 0} \frac{H(\nu_{[0,t]} \bowtie \mu_{[0,t]})}{t}$$

be the process-level large deviation rate function for the retrospective mutation chain. In the above, $\nu_{[0,t]}$ and $\mu_{[0,t]}$ are the restrictions of $\nu$ and $\mu$ to the time interval $[0,t]$, and $H(\nu_{[0,t]} \bowtie \mu_{[0,t]})$ is their relative entropy. See [5, Equation (4.4.28)]; alternative expressions can be found in [5, Theorem 4.4.38] and [23, Theorems 7.3 and 7.4].

**Theorem 5.1.** For the empirical type evolution process $R^i(t)$ as in (3.4) we have, for $i \in S$ and closed $F \subset \mathcal{P}_0(\Sigma)$,

$$\limsup_{t \to \infty} \frac{1}{t} \log E_i \left( \sum_{x \in X(t)} 1_{\{R^i(t) \in F\}} \right) \leq \lambda - \inf_{\nu \in F} H_G(\nu),$$

while, for open $G \subset \mathcal{P}_0(\Sigma)$,

$$\liminf_{t \to \infty} \frac{1}{t} \log E_i \left( \sum_{x \in X(t)} 1_{\{R^i(t) \in G\}} \right) \geq \lambda - \inf_{\nu \in G} H_G(\nu).$$

Moreover, the function $H_G$ is lower semicontinuous with compact level sets and attains its minimum 0 precisely at $\mu$.

**Proof.** In view of Theorem 4.1, for every measurable $C \subset \mathcal{P}_0(\Sigma)$ we have

$$E_i \left( \sum_{x \in X(t)} 1_{\{R^i(t) \in C\}} \right) = h_i e^{\lambda t} \hat{E}_\lambda (1_{\{R^i(t) \in C\}} h_{\sigma(\xi(t))}^{-1}).$$

Since $\max_i |\log h_i| < \infty$, the $h_i$ 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 [5, 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

$$I_G(\nu) = \sup_{v \in [0,\infty]^S} \left[ -\sum_{i \in S} \frac{v_i(G\nu)}{v_i} \right] = \inf_{\nu \in \mathcal{P}_0(\Sigma); \nu_0 = \nu} H_G(\nu)$$

be the level-two rate function of the retrospective mutation chain; here we write $\nu_0$ for the time-zero marginal distribution of $\nu$. (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 \( F \subset P(S) \),
\[
\limsup_{t \to \infty} \frac{1}{t} \log E\left( \sum_{x \in X(t)} 1_{\{L^i(t) \in F\}} \right) \leq \lambda - \inf_{v \in F} I_G(v),
\]
while, for open \( G \subset P(S) \),
\[
\liminf_{t \to \infty} \frac{1}{t} \log E\left( \sum_{x \in X(t)} 1_{\{L^i(t) \in G\}} \right) \geq \lambda - \inf_{v \in G} I_G(v).
\]
Moreover, the function \( I_G \) is continuous and strictly convex and attains its minimum 0 precisely at \( \alpha \).

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 and 7.9], or proved directly as in [4, 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 \( P_B(\Sigma) \). To be specific, we let \( d_{RS} \) denote the Skorohod metric on \( RS \) (defined in analogy to the one-sided case considered in [9, p. 117, Equation (5.2)]), and \( d \) be the associated Prohorov metric on \( P(\Sigma) \); see [9, p. 96, Equation (1.1)]. For any fixed \( \varepsilon > 0 \), we consider the set \( C = \{ v \in P_B(\Sigma) : d(v, \mu) \geq \varepsilon \} \), the complement of the open \( \varepsilon \)-neighborhood of \( \mu \). In view of Remark 3.3, we need to show that, \( P^\varepsilon \)-almost surely on \( \Omega_{surv} \),
\[
\Gamma(t, C) := \frac{1}{|X(t)|} \sum_{x \in X(t)} 1_{\{R^i(t) \in C\}} \to 0 \quad \text{as} \quad t \to \infty.
\]
In the first part of the proof we will establish this convergence along a discrete-time skeleton \( \delta N \), where \( \delta > 0 \) is arbitrary.

Since \( C \) is closed and \( H_G \) has compact level sets and attains its minimum 0 at \( \mu \) only, the infimum \( c := \inf_{v \in C} H_G(v) \) is strictly positive. We can therefore choose a constant \( \gamma \) such that \( \lambda > \gamma > \lambda - c \). We write
\[
\Gamma(t, C) = \left( \frac{e^{\gamma t}}{|X(t)|} \right) \left( e^{-\gamma t} \sum_{x \in X(t)} 1_{\{R^i(t) \in C\}} \right)
\]
and show that each factor tends to 0 along \( \delta N \) almost surely on \( \Omega_{surv} \). In view of Corollary 5.1 and Theorem 2.1(a),
\[
\frac{|X((n+1)\delta)|}{|X(n\delta)|} = \sum_{j \in S} \frac{Z_j(n\delta)}{Z(n\delta)} \frac{1}{Z_j(n\delta)} \sum_{x \in X_j(n\delta)} |X(x, (n+1)\delta)| \to \sum_{j \in S} \pi_j E^j(|X(\delta)|) = e^{\lambda \delta} \quad \text{as} \quad n \to \infty \quad \text{almost surely on} \quad \Omega_{surv}.
\]
Hence \( n^{-1} \log |X(n\delta)| \to \lambda \delta \) and therefore \( e^{\gamma n\delta}/|X(n\delta)| \to 0 \) almost surely on \( \Omega_{surv} \). On the other hand, using Markov’s inequality and Theorem 5.1 we obtain for any \( a > 0 \) that
\[
\limsup_{n \to \infty} \frac{1}{n \delta} \log P^\varepsilon \left( e^{-\gamma n\delta} \sum_{x \in X(n\delta)} 1_{\{R^i(n\delta) \in C\}} \geq a \right) \leq \lambda - c - \gamma < 0.
\]
The Borel–Cantelli lemma thus shows that the second factor of $\Gamma(t, C)$ also tends to 0 almost surely as $t \to \infty$ through $\delta \mathbb{N}$. We therefore conclude that $\lim_{n \to \infty} \Gamma(n\delta, C) = 0$ almost surely on $\Omega_{\text{surv}}$.

To extend this result to the full convergence $t \to \infty$ along all reals, we pick some $\varepsilon'$ with $0 < \varepsilon' < \varepsilon$ and let $C'$ be defined in terms of $\varepsilon'$ instead of $\varepsilon$. Also, let $A$ be an arbitrary closed set in $\Sigma$, $\varepsilon^* = \varepsilon - \varepsilon'$, and $A^* = \{ \sigma \in \Sigma : d_\Sigma(\sigma, A) < \varepsilon^* \}$, the $\varepsilon^*$-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

$$R^s(t)(A) \leq \frac{1}{t} \int_0^s 1_{A(\theta_\sigma \sigma(y)(\tau), \text{per})} \, du + \frac{\delta}{t}.$$

By the locality of the Skorohod metric $d_\Sigma$, there exists a constant $c = c(\varepsilon^*)$ such that $d_\Sigma(\theta_\sigma \sigma(y)(s), \text{per}, \theta_\sigma \sigma(y)(t), \text{per}) < \varepsilon^*$ 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

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

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

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

when $s$ is large enough. For such an $s$, we obtain

$$\Gamma(t, C) - \Gamma(s, C') \leq \frac{1}{|X(t)|} \left| X(t) \right| + \frac{1}{|X(s)|} \sum_{x \in X(s)} 1_{\{R^s(t)(x) \in C'\}} (|X(x, t)| - 1)$$

$$\leq 1 - \inf_{s \leq \tau \leq s + \delta} \frac{|X(t)|}{|X(s)|} + \frac{1}{|X(s)|} \sum_{x \in X(s)} (M(x, [s, s + \delta]) - 1),$$

where $M(x, [s, s + \delta]) = \max_{s \leq \tau \leq s + \delta} |X(x, t)|$ 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 $E^\pi(M(0, [0, \delta])) - 1)$ almost surely on $\Omega_{\text{surv}}$. According to the proof of (5.5), this limit can be made arbitrarily small if $\delta$ is chosen small enough. In combination with (5.6) and the first part of this proof, this shows that $\lim \sup_{t \to \infty} \Gamma(t, C) \leq a$ for every $a > 0$ almost surely on $\Omega_{\text{surv}}$. The proof is thus complete.

Proof of Theorem 3.2. There are two possible routes for the proof. We can either repeat the argument above by simply replacing Theorem 5.1 by Corollary 5.2. Or we can notice that $L^s(t)$ is the time-zero marginal of $R^s(t)$ and that the marginal mapping $\nu \mapsto \nu_0$ 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].

Acknowledgements

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. Statist. 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. Process. Appl. 39, 117–130.
[4] D'Oliveria Cardoso, F. (2000). Large Deviations (Fields Inst. Monogr. 14). American Mathematical Society, Providence, RI.
[5] Deuschel, J. D. and Stroock, D. W. (1989). Large Deviations. Academic Press, Boston, MA. Reprinted: American Mathematical Society, Providence, RI, 2001.
[6] Donsker, M. D. and Varadhan, S. R. S. (1975). Asymptotic evaluation of certain Markov process expectations for large time. I. Commun. Pure Appl. Math. 28, 1–47.
[7] Donsker, M. D. and Varadhan, S. R. S. (1983). Asymptotic evaluation of certain Markov process expectations for large time. IV. Commun. Pure Appl. Math. 36, 183–212.
[8] Durrett, R. (2002). Probability Models for DNA Sequence Evolution. Springer, New York.
[9] Ethier, S. N. and Kurtz, T. G. (1986). Markov Processes. Characterization and Convergence. John Wiley, New York.
[10] Gorostiza, L. G., Roelly, S. and Wakolbinger, A. (1992). Persistence of critical multitype particle and measure branching processes. Prob. Theory Relat. Fields 92, 313–335.
[11] Harris, T. E. (1963). The Theory of Branching Processes. Springer, Berlin. Reprinted: Dover, New York, 2002.
[12] Hermisson, J., Redner, O., Wagner, H. and Baake, E. (2002). Mutation–selection balance: ancestry, load, and maximum principle. Theor. Pop. Biol. 62, 9–46.
[13] Jagers, P. (1989). General branching processes as Markov fields. Stoch. Process. 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 multi-type branching populations. In Séminaire de Probabilités XXX (Lecture Notes Math. 1626), eds J. Azéma, M. Emery, and M. Yor, 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. Process. Appl. 13, 235–248.
[18] Kingman, J. F. C. (1982). On the genealogy of large populations. In Essays in Statistical Science (J. Appl. Prob. Spec. Vol. 19A), eds J. Gani and E. J. Hannan, Applied Probability Trust, Sheffield, pp. 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, eds K. B. Athreya and P. Jagers, Springer, New York, pp. 181–185.
[20] Lyons, R., Pemantle, R. and Peres, Y. (1995). Conceptual proofs of L log L criteria for mean behaviour of branching processes. Ann. Prob. 23, 1125–1138.
[21] Matthies, K., Kerstan, J. and Mecke, K. (1978). Infinitely Divisible Point Processes. John Wiley, Chichester.
[22] Möhle, M. (2000). Ancestral processes in population genetics—the coalescent. J. Theoret. Biol. 204, 629–638.
[23] Varadhan, S. R. S. (1988). Large deviations. In École d'ÉtÉ de ProbabilitÉs de Saint-Flour XV–XVII, 1985 (Lecture Notes Math. 1362), ed. P. L. Hennequin, Springer, Berlin, pp. 1–49.