The ancestral population size conditioned on the reconstructed phylogenetic tree with occurrence data

Marc Manceau, Ankit Gupta, Timothy Vaughan, Tanja Stadler*

Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland.

Abstract

We consider a homogeneous birth-death process with three different sampling schemes. First, individuals can be sampled through time and included in a reconstructed tree. Second, they can be sampled through time and only recorded as a point ‘occurrence’ along a timeline. Third, extant individuals are sampled and included in the reconstructed tree with a fixed probability. We further consider that sampled individuals can be removed or not from the process, upon sampling, with fixed probability.

Given an outcome of the process, composed of the joint observation of a reconstructed phylogenetic tree and a record of occurrences not included in the tree, we derive the conditional probability distribution of the population size any time in the past. We additionally provide an algorithm to simulate ancestral population size trajectories given the observation of a reconstructed tree and occurrences.

This distribution can readily be used to perform inferences of the ancestral population size in the field of epidemiology and macroevolution. In epidemiology, these results will pave the way towards jointly considering data from case count studies and reconstructed transmission trees. In macroevolution, it will foster the joint examination of the fossil record and extant taxa to reconstruct past biodiversity.

Keywords: birth-death process, fossilized birth-death model, epidemiology, macroevolution, phylogenetics

1. Introduction

Since seminal papers by [Yule (1925), Kendall (1948)], and much later by [Nee et al. (1994)], birth-death models have become ubiquitous in evolutionary biology. They are used as an underlying tree prior, or as an underlying population dynamics prior in recent studies spanning fields as diverse as macroevolution, linguistics, or epidemiology (see e.g. [Heath et al. 2014], [Gray et al. 2009], [Ratmann et al. 2016]). Computing
reliable estimates of the ancestral number of species, languages or infected individuals, i.e. estimates of past population size, is key to the understanding of past processes in these fields.

In both macroevolution and epidemiology, these inferences have initially relied on the fossil record and the case counts record, modeled as a sampling of individuals from the full process through time (Foote 2000; Starrfelt and Liow 2016).

In the recent years, impressive sequencing efforts targeting present-day species and pathogens have enabled the reconstruction of phylogenies, the branch-lengths of which can be used to infer ancestral population dynamics. Two main approaches have been introduced to estimate past population sizes from reconstructed evolutionary trees.

The first one, in a maximum likelihood framework, consists of estimating the birth and death rate of the population using all branch-lengths of the tree ahead. Analytical estimates of past population size are then computed in a second step, based on simplifications of the data, such as, e.g., taking only into account the number of individuals at the beginning (Morlon et al. 2011) or more recently at the beginning and end of the process (Billaud et al. 2019).

The second one is a variation of the first method in a Bayesian setting. Birth and death rates are supposed to be piecewise constant with a given a priori distribution, and their posterior is sampled using all branch-lengths. Past population sizes are then computed by integrating the expected number of individuals over the rates posterior distribution (Stadler et al. 2013).

While this second method accounts for uncertainty in the estimated population sizes, it also relies on numerical Monte-Carlo methods which make estimates much more computationally expensive to compute. Additionally, the statistical signal in the number of lineages through time is only used to compute rate estimates, whereas it could in principle also more precisely inform the past population size distribution.

Statistical approaches stemming from the analysis of case count data or from the analysis of reconstructed evolutionary trees have been part of separate literature bodies for years, historically yielding conflict between biodiversity estimates based on the fossil record and estimates based on phylogenies of extant taxa (Quental and Marshall 2010; but see also Morlon et al. 2011). A first path towards merging these disparate data was introduced by the fossilized birth-death model of Stadler (2010), who considered a birth-death model with sampling and inclusion of individuals in the tree through time. This allowed taking into account infection trees reconstructed from pathogen sequences sampled throughout an epidemic (Stadler et al. 2011). In macroevolution, it paved the way to more precise phylogenetic dating using well-conserved fossil taxa which could be placed on a reconstructed phylogeny using morphological characters (Gavryushkina et al. 2016). Not so well-conserved fossils (i.e. occurrences) have also been used with this model, using a Markov Chain Monte Carlo (MCMC) scheme to integrate over
all possible placements along a fixed tree (Heath et al., 2014). Very recently, Vaughan et al. (2019) introduced another extension to the fossilized birth-death model, considering that individuals sampled through time could either be included or not included in the reconstructed evolutionary tree. This variation has profound consequences for data analysis, for it allows one to combine the occurrence record that cannot be placed along a tree (e.g. poorly preserved fossils, or case count epidemiological record) with samples for which molecular sequences or phenotypic measurements allow reconstructing a phylogenetic tree. They also provide a Monte-Carlo particle filtering algorithm to reconstruct past population sizes by sampling from the distribution of the number of ancestors conditioned on the joint observation of the reconstructed tree and the occurrence record. Analytical developments around this new model have further been made by Gupta et al. (2019), who derived an analytical formula for the probability density of an outcome of the process.

In this paper, we build on the analytical developments presented by Gupta et al. (2019), and aim at providing a more analytical, and faster, way to compute the same distribution as originally targeted by Vaughan et al. (2019). The efficiency of our method paves the way towards considering much bigger datasets, or towards extending the method to multi-type or density-dependent birth-death processes.

In section 2, we present the model, notation, and an overview of the strategy to express the targeted distribution. In section 3, we adapt the main results of Gupta et al. (2019) to compute the probability density of observations made after a given time, conditioned on the past population size. In section 4, we provide a way to compute the joint density of the past population size and observations made before a given time. Combining results of sections 3 and 4, we compute the distribution of past population sizes conditional on the full outcome of the process in section 5. We finally discuss applications and potential extensions of the model.

2. Model and notation

2.1. Parameters of the process

We consider a population of individuals, any of which can give birth to another individual at rate $\lambda$ or die at rate $\mu$. The process starts at time $t_{\text{of}}$ in the past with one individual, and unfolds until reaching present time 0, i.e. time is oriented from the present towards the past.

We superimpose to this background population dynamics three different sampling schemes. First, individuals can be $\psi$-sampled at rate $\psi$ throughout their lifetime. When $\psi$-sampled, the individual will be included in the reconstructed tree. Second, individuals can be $\omega$-sampled at rate $\omega$ throughout their lifetime. When $\omega$-sampled, the individual is not included in the reconstructed tree, but its sampling time is nevertheless recorded and thereafter called ‘an occurrence’. Last, the process is finished upon reaching the
present time 0, and each extant individual at that time is \( \rho \)-sampled with fixed probability \( \rho \), leading to their inclusion in the reconstructed tree. The sum of all per-capita rates will be called for short \( \gamma = \lambda + \mu + \psi + \omega \).

Following [Vaughan et al. (2019)], we also include in the model an effect of the \( \psi \)- and \( \omega \)-sampling through time on the population dynamics. We consider that, upon sampling, an individual is either removed from the process with probability \( r \in (0, 1) \), or is unaffected by the sampling with probability \( (1 - r) \). The overall number of individuals, denoted \( (I_t) \), thus follows a linear birth-death process with birth rate \( \lambda \) and death rate \( \mu + (\psi + \omega)r \). Because this work aims at fitting the model to empirical data, we only consider here a supercritical process with \( \lambda > \mu + (\psi + \omega)r \).

2.2. Introducing useful probabilities

This process has already been thoroughly studied. In the following sections, we will use two key probabilities which analytical solutions are well known. First, we will call \( u_t \) the probability that a process starting at time \( t \) with only one individual is never sampled before reaching present time 0. We recall that \( u_t \) satisfies the Master equation

\[
\begin{align*}
    u_0 &= z \\
    \dot{u}_t &= \lambda u_t^2 - \gamma u_t + \mu 
\end{align*}
\]

The solution of this for a particular initial condition \( z \) being the following

\[
\begin{align*}
    u(t, z) &= \frac{x_1(x_2 - z) - x_2(x_1 - z)e^{-\sqrt{\Delta}t}}{(x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta}t}} \\
    \text{where } \Delta &= \gamma^2 - 4\lambda \mu > (\lambda + \mu)^2 - 4\lambda \mu \geq (\lambda - \mu)^2 > 0 \text{ and } x_1, x_2 \text{ are the two roots of the polynomial } \\
    \lambda x^2 - \gamma x + \mu, \\
    x_1 &= \frac{\gamma - \sqrt{\Delta}}{2\lambda} \text{ and } x_2 = \frac{\gamma + \sqrt{\Delta}}{2\lambda} .
\end{align*}
\]

Second, we call \( p_t \) the probability that a process starting at time \( t \) with one individual leads to one sampled particle at present time 0. Writing the Master equation governing the evolution of this quantity leads to

\[
\begin{align*}
    p_0 &= 1 - z \\
    \dot{p}_t &= (2\lambda u(t, z) - \gamma)p_t 
\end{align*}
\]

The solution of this being the following

\[
p(t, z) = (1 - z)^{\frac{\Delta}{\lambda^2}} (x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta}t})^{-2} e^{-\sqrt{\Delta}t} \]

\(4\)
These formulas are well known, and correspond respectively to quantities called $p_0(t)$ and $p_1(t)$ in Stadler (2010). When $z = 1 - \rho$, we will drop the dependence on $z$ and use the shorter notation $u_t, p_t$. We recall standard ways to derive these expressions in Appendix A.

2.3. Strategy of the paper

The process with sampling leads to the observation of two distinct objects $(\mathcal{T}, \mathcal{O})$, as illustrated in Figure 1. The reconstructed tree $\mathcal{T}$, on the one hand, represents the evolutionary relationships between all $\psi$-sampled and $\rho$-sampled individuals. We further consider that $\psi$-sampled individuals are labeled either as ‘removed’ or ‘non-removed’. All $\psi$-sampled removed individuals are necessarily leaves of $\mathcal{T}$, whereas $\psi$-sampled non-removed ones can either stand as leaves (when the descent of the individual is not sampled) or as vertices along a branch (when the descent of the individual is further sampled), in which case they are referred to as sampled ancestors.

The record of occurrences $\mathcal{O}$, on the other hand, is an ordered list of all $\omega$-sampling times. We also consider that these sampling times are labeled as either ‘removed’ or ‘non-removed’.

In this paper, we are interested in computing the probability distribution of the number of individuals in the past, conditioned on the observed outcome $(\mathcal{T}, \mathcal{O})$ of the process. If $k_t$ denotes the number of sampled lineages in $\mathcal{T}$ at time $t$, we call our target distribution,

$$\forall t \geq 0, \forall i \in \mathbb{N}, \quad K^{(i)}_t := P(I_t = k_t + i \mid \mathcal{T}, \mathcal{O})$$ (2.5)

We will refer to epochs as the maximal time slices within which no sampling event in $\mathcal{O}$, nor branching event in $\mathcal{T}$, happened. These epochs are delimited by the union of sampling times in $\mathcal{O}$, branching times in the tree $\mathcal{T}$, and sampling times of leaves and sampled ancestors in $\mathcal{T}$. All pooled together, we call these ordered times $(t_h)_{h=0}^n$, starting at present time $t_0 = 0$ and ending at the root time $t_n = t_{or}$.

At any time $t \geq 0$ we also introduce:

- $\mathcal{T}_t^\uparrow$ := the tree $\mathcal{T}$ cut at time $t$
- $\mathcal{T}_t^\downarrow$ := the collection of trees (or forest) obtained by cutting $\mathcal{T}$ at time $t$, and considering all subtrees descending from cut lineages
- $\mathcal{O}_t^\uparrow$ := $\mathcal{O}_{(t, \infty)}$
- $\mathcal{O}_t^\downarrow$ := $\mathcal{O}_{(0, t)}$

The general strategy – and outline – of the paper is the following. In a breadth-first backward traversal we will compute the probability density of observations made down any time $t$, conditioned on the population
size at that time, as time increases towards the past. We call this probability density,

$$\forall i \in \mathbb{N}, \quad L_t^{(i)} := \mathbb{P}(T^i_t, O_t^i | I_t = k_t + i)$$  \hspace{1cm} (2.6)$$

In a *breadth-first forward traversal* we will then compute the joint probability density of the observations made up to any time $t$ and the population size at that time, as time decreases towards present. We call this density,

$$\forall i \in \mathbb{N}, \quad M_t^{(i)} := \mathbb{P}(T^i_t, O_t^i, I_t = k_t + i)$$  \hspace{1cm} (2.7)$$

Provided we get expressions of $(L_t)^{\text{for}}_{t=0}$ and $(M_t)^{\text{for}}_{t=0}$, our target distribution can then be expressed by combining both, noting that:

$$K_t^{(i)} := \mathbb{P}(I_t = k_t + i | T, O)$$
\[ P( I_t = k_t + i, T^\uparrow_t, O^\uparrow_t, T^\downarrow_t, O^\downarrow_t) \]
\[ \propto P(T^\downarrow_t, O^\downarrow_t \mid I_t = k_t + i, T^\uparrow_t, O^\uparrow_t) P(I_t = k_t + i, T^\uparrow_t, O^\uparrow_t) \]
\[ \propto L_t^{(i)} M_t^{(i)} \tag{2.8} \]

where the last line holds because, conditionally on \( I_t = k_t + i \), the future of the (Markov) process is independent of what happened before.

In the process of getting the likelihood of observations under the same model, [Gupta et al. (2019)] provided an analytical formula and an algorithm to compute the first ingredient \( L_t \) in the case where all individuals are removed upon sampling (i.e. \( r = 1 \)). We thus recall their main result, and adapt it to our slightly different framework, in the next section.

3. The density of observations below conditioned on past population size

We start this section by presenting the Master equations satisfied by the probability density \( L_t \). This provides us with a numerical algorithm to compute \( L_t \), which we subsequently simplify with analytical results for specific sets of parameters.

3.1. Master equation driving \( L_t \)

One way to get the probability density \( L_t \) consists in studying its evolution through time, relying on its Master equation. First, remark that we can express \( L_0 \) at present time 0. Indeed, provided we know the exact number of individuals living at time 0, the probability to see the tips of the tree is directly driven by the \( \rho \)-sampling,
\[ L_0^{(i)} = \rho^k (1 - \rho)^i \tag{3.9} \]

We now derive the Master equation driving the evolution of \( L_t \) through time across any given epoch. We consider an infinitesimal time step \( \delta t \) and list the events which could have happened in the full process between \( t \) and \( t + \delta t \), leading to our observations. Suppose the number of observed lineages in this epoch is \( k \), and the total number of individuals alive is \( k + i \). We emphasize three cases, illustrated in Figure 2:

1. nothing happened with probability \( (1 - \gamma(k + i)\delta t) \)
2. a birth event happened
   - (a) among the \( k \) sampled lineages in \( T^\uparrow_t \), and it leads to an extinct or unsampled subtree to the left or to the right, with probability \( 2\lambda \delta t \).
   - (b) among the \( i \) other individuals, with probability \( \lambda i \delta t \).
3. a death event happened among the $i$ particles, with probability $\mu i \delta t$.

These allow us to write, $\forall i \in \mathbb{N}$,

$$L_{i+\delta t}^{(i)} = (1 - \gamma (k + i) \delta t) L_{i}^{(i)} + \lambda (2k + i) \delta t L_{i}^{(i+1)} + \mu i \delta t L_{i}^{(i-1)}$$

Note that for $i = 0$, $L_{i}^{(i-1)}$ is not defined, but the term cancels out thanks to the factor $i$.

Letting $\delta t \to 0$ and combining this with the initial condition on $L_0$, we get the following set of ordinary differential equations driving the evolution of $L_t$,

$$\forall i \in \mathbb{N}, \quad L_{i}^{(i)}|_{t=0} = \rho^{k_0} (1 - \rho)^i $$

$$\dot{L}_{i}^{(i)} = -\gamma (k + i) L_{i}^{(i)} + \lambda (2k + i) L_{i}^{(i+1)} + \mu i L_{i}^{(i-1)}$$

(3.10)

Last, we need to study how $L_t$ changes at punctual events. There are 6 types of punctual events that we can come across at time $t$ in the past, listed below and illustrated in Figure 3. We denote $L_{t+}$ the probability just before (i.e. up) the punctual event and $L_{t-}$ the probability immediately after (i.e. down). We can either find,

1. a leaf of $T^\downarrow_{t}$, labeled as removed. This is a $\psi$-sampling with removal event for which the number of unsampled lineages remains constant, and the number of sampled lineages increases by one. It thus gives,

$$L_{t+}^{(i)} = \psi r L_{t-}^{(i)}$$

(3.11)

2. a leaf of $T^\downarrow_{t}$, labeled as non-removed. This is a $\psi$-sampling without removal event for which one of the unsampled lineage becomes a sampled one. It thus gives,

$$L_{t+}^{(i)} = \psi (1 - r) L_{t-}^{(i+1)}$$

(3.12)

3. a sampled ancestor along a branch of $T^\downarrow_{t}$, necessarily labeled as non-removed. This is a $\psi$-sampling without removal event, not impacting the number of sampled or unsampled lineages. It thus gives,

$$L_{t+}^{(i)} = \psi (1 - r) L_{t-}^{(i)}$$

(3.13)
4. an occurrence in $O_t^↓$, labeled as removed. This is a $\omega$-sampling with removal event, for which the number of unsampled lineages increases by one. It thus gives,

$$L_t^{(i)} = \omega r i L_t^{(i-1)}$$  \hspace{1cm} (3.14)

Note that here also, for $i = 0$, $L_t^{(-1)}$ is not defined but the term cancels out thanks to the factor $i$.

5. an occurrence in $O_t^↓$, labeled as non-removed. This is a $\omega$-sampling without removal event, not impacting the number of sampled or unsampled lineages. It thus gives,

$$L_t^{(i)} = \omega (k + i)(1 - r) L_t^{(i-1)}$$  \hspace{1cm} (3.15)

6. a branching event between two branches of $T_t^↓$. The number of sampled lineages decreases by one. It thus gives,

$$L_t^{(i)} = \lambda L_t^{(i-1)}$$  \hspace{1cm} (3.16)

Figure 3: Six observable punctual events in the data.

This Master equation can be numerically approximated. To do so, we fix a finite upper bound $N$ on the number of hidden individuals and numerically integrate a truncated ODE system. We detail this in the following algorithm to compute an approximation of $L_t$ at any time $t$. 

9
Algorithm 1 Computes a numerical approximation of $L_t$ for a specific set of times

**Input:**
- Observed tree and occurrence data $(T, O)$,
- parameters $(t_{or}, \lambda, \mu, \psi, \omega, \rho, r)$,
- set of time points $(\tau_j)_{j=1}^S$ for which we want to compute the density,
- and the truncation $N$ setting the accuracy of the algorithm.

**Output:** A numerical approximation of $L_t$ at times $(\tau_j)_{j=1}^S$, $(\tilde{L}_{t_j}^{(i)})_{i\in\{0,1,\ldots,N\}}$.

1. Pool all $(\tau_j)$ and all branching and sampling times of $(T, O)$ in an ordered list $(t_h)_{h=1}^n$.
2. Set $j = 1$ and initialize $B$ as a $S \times (N + 1)$ empty matrix.
3. Set $\forall i \in \{0, 1, \ldots, N\}, \tilde{L}_0^{(i)} = \rho^k(1 - \rho)^i$.
4. For $h = 1, 2, \ldots, n$ do
5. Numerically solve the ODE $\dot{\tilde{L}}_t = A\tilde{L}_t$ on $(t_{h-1}, t_h)$, where matrix $A$ is a $(N + 1) \times (N + 1)$ tridiagonal matrix with entries given by,

   $\forall i \in \{0, 1, \ldots, N\} \; A^{(i,i)} = -\gamma(k + i)$

   $\forall i \in \{0, 1, \ldots, N - 1\} \; A^{(i,i+1)} = \lambda(2k + i)$

   $\forall i \in \{1, 2, \ldots, N\} \; A^{(i,i-1)} = \mu i$

6. if $t_h = \tau_j$ then
7. Record $\forall i, B^{(j,i)} = \tilde{L}_t^{(i)}$
8. Set $j = j + 1$
9. end if
10. if $t_h = t_{or}$ or $t_h = \tau_S$ then
11. return $B$
12. else if $t_h$ is a removed leaf then
13. Set $\dot{L}_{t_h}^+ = r \dot{L}_{t_h}^-$
14. else if $t_h$ is a non-removed leaf then
15. Set $\forall i < N, \tilde{L}_{t_h}^{(i)} = \psi(1 - r)\tilde{L}_{t_h}^{(i+1)}$ and $\tilde{L}_{t_h}^{(N)} = 0$
16. else if $t_h$ is a sampled ancestor then
17. Set $\dot{L}_{t_h}^+ = \psi(1 - r)\dot{L}_{t_h}^-$
18. else if $t_h$ is a removed occurrence then
19. Set $\forall i > 0, \dot{L}_{t_h}^{(i)} = \omega r \dot{L}_{t_h}^{(i-1)}$ and $\dot{L}_{t_h}^{(0)} = 0$
20. else if $t_h$ is a non-removed occurrence then
21. Set $\dot{L}_{t_h}^{(i)} = \omega (1 - r)(k + i)\dot{L}_{t_h}^{(i)}$
22. else $t_h$ is a branching event
23. Set $\dot{L}_{t_h}^+ = \lambda \dot{L}_{t_h}^-$
24. end if
25. end for

We also define a slight variation of this algorithm, that we will refer to as algorithm $[]$, where no set of time points $(\tau_j)$ is required, and the values of $\tilde{L}_t$ are not recorded through time (i.e. matrix $B$ disappears). Instead, when reaching $t_n = t_{or}$ we simply return $\tilde{L}_t^{(0)}$, which by definition is an estimate of the probability density of $(T, O)$. Note that this strategy is very close to what has been used to compute the probability density of a reconstructed tree under a logistic birth-death process (Leventhal et al. 2013).
These two algorithms will prove useful to deal with the general case. However, we can further get analytical expressions for $L_t$ when $\omega = 0$ as well as when $r = 1$ (Gupta et al., 2019). We expose these in the next two subsections.

3.2. Special case $\omega = 0$

Suppose we can express $L_i^{(t)}$ as the product $L_i^{(t)} = u_i^t W_t$ where $W_t$ is a function of time only, and $u_i^t$ is defined as in equation 2.2. We first get, from the initialization in equation (3.10), that $W_0 = \rho k_0$. Moreover, substituting $u_i^t W_t$ in the ODE leads to

$$\dot{L}_i^{(t)} = iu_i^{t-1}u_i^t W_t + u_i^t \dot{W}_t$$

$$= (\lambda iu_i^{t+1} - \gamma iu_i^t + \mu iu_i^{t-1}) W_t + u_i^t \dot{W}_t$$

Thus leading to the following ODE for $W_t$, on any epoch $(t_h, t_{h+1})$ where the number of sampled lineages remains fixed and equal to $k$,

$$u_i^t \dot{W}_t = (-\gamma (i+k) u_i^t + \lambda (2k+i) u_i^{t+1} + \mu iu_i^{t-1} - \lambda iu_i^{t+1} + \gamma iu_i^t - \mu iu_i^{t-1}) W_t$$

$$\Rightarrow \dot{W}_t = (2\lambda u_t - \gamma) k W_t$$

This is very close to the Master equation (2.3) governing the evolution of $p_t$, and it leads to (see derivation in Appendix A),

$$\forall t \in (t_h, t_{h+1}), \quad W_t = W_{t_h} \left( \frac{p_t}{p_{t_h}} \right)^k \quad (3.17)$$

Last, because $\omega = 0$, updates (3.11) to (3.16) simplify to only the following $\psi$- and $\lambda$-events,

$$\begin{align*}
\text{if } t \text{ is a removed leaf, } & \quad W_{t+} = \psi r W_{t-} \quad (3.18) \\
\text{if } t \text{ is a non-removed leaf, } & \quad W_{t+} = \psi (1-r) u_t W_{t-} \quad (3.19) \\
\text{if } t \text{ is a sampled ancestor, } & \quad W_{t+} = \psi (1-r) W_{t-} \quad (3.20) \\
\text{if } t \text{ is a branching time, } & \quad W_{t+} = \lambda W_{t-} \quad (3.21)
\end{align*}$$

Combining these updates with equation (3.17) leads to the following proposition.

**Proposition 3.1.** At any time $t$ across epoch $(t_h, t_{h+1})$, considering that we observed so far –i.e. on $(0, t_{h+1})$ – $v$ sampled ancestors, $w$ removed leaves at times $t_j \in \mathcal{W}$, $x$ branching events at times $t_j \in \mathcal{X}$, $y$ non-removed leaves at times $t_j \in \mathcal{Y}$, we get,

$$L_i^{(t)} = u_i^t W_t$$
where \( W_t = \lambda^x \psi^{v+w+y} (1-r)^{v+w+y} p_{t_j} \prod_{t_j \in \mathcal{X}} p_t \prod_{t_j \in \mathcal{Y}} u_t \prod_{t_j \in \mathcal{W}} p_t^{-1} \).

**Proof** We prove this proposition by induction across the epochs in Appendix D, using as the main arguments the equation updates (3.18) to (3.21), combined with equation (3.17). □

Note that this proposition is very similar to what is presented in section 3 by Gupta et al. (2019). We nevertheless need to highlight two differences. The first one is that we allow here for removal or not of the individual upon sampling, with a given probability \( r \), whereas Gupta et al. (2019) considered that all individuals were removed upon sampling (\( r = 1 \)), and Stadler (2010) considered that individuals were not removed upon sampling (\( r = 0 \)).

The second difference concerns the underlying framework under which we derive our results. In Gupta et al. (2019), individuals where distinguishable (say, each one is assigned a number and they can be ordered), whereas in the present paper they are not. When individuals are ordered, the probability density \( L_i(t) \) is changed by a factor \( \frac{(k+i)!}{i!} \), which is the number of ways we can arrange \( k+i \) elements in a list of size \( k \), i.e. the number of ordered configurations of hidden individuals.

Note that, when reaching the root of the tree, it reduces to a very similar formula for the probability density of \( \mathcal{T} \) because \( i = 0 \) and \( k = 1 \). We summarize this as the following corollary.

**Corollary 3.2.** When \( \omega = 0 \), the probability density of a reconstructed tree \( \mathcal{T} \) with \( v \) sampled ancestors, \( w \) removed leaves at times \( t_j \in \mathcal{W} \), \( y \) non-removed leaves at times \( t_j \in \mathcal{Y} \), and branching events at times \( t_j \in \mathcal{X} \), is

\[
\mathcal{L}(\mathcal{T}) = \lambda^w \psi^{v+w+y} (1-r)^{v+w+y} p_{t_j} \prod_{t_j \in \mathcal{X} \cup \{t_o\}} p_t \prod_{t_j \in \mathcal{Y}} u_t \prod_{t_j \in \mathcal{W}} p_t^{-1} \prod_{t_j \in \mathcal{X} \cup \{t_o\}} p_t^{-1} \quad (3.22)
\]

**Proof** It directly follows from Proposition 3.1 by noting that \( \mathcal{L}(\mathcal{T}) = \mathcal{L}_{t_o}(0) \). Note also that a rooted binary tree with \( w+y+k_0 \) leaves shows necessarily \( x = w+y+k_0 - 1 \) branching times. □

### 3.3. Special case \( r = 1 \)

When \( r = 1 \), only three kinds of punctual events, corresponding to updates (3.11), (3.14) and (3.16) need to be taken into account. Because the number of unsampled individuals \( i \) goes into formula (3.14), the simple expression \( L_i(t) = w_i W_t \) cannot be considered anymore, and one needs to find another expression.

This has already been done in Gupta et al. (2019) and we only need to adapt here their result to our slightly different framework. For this we define the distinguishable version of the probability \( L_i(t) \) as

\[
\mathcal{L}_i(t) = \frac{(k+i)!}{i!} L_i(t). \quad (3.23)
\]
We now derive the Master equation for $T_t^{(i)}$. Multiplying both sides of (3.10) by $\frac{(k+i)!}{i!}$ we obtain

$$L_t^{(i)} = -\gamma (k+i) \frac{(k+i)!}{i!} L_t^{(i)} + \lambda (2k+i) \frac{(k+i)!}{i!} L_t^{(i+1)} + \mu (k+i) \frac{(k+i)!}{i!} L_t^{(i-1)}$$

$$= -\gamma (k+i) \frac{(k+i)!}{i!} L_t^{(i)} + \frac{(2k+i)(i+1)}{(k+i+1)(k+i)} \frac{(k+i+1)!}{(i+1)!} L_t^{(i+1)} + \frac{(k+i)(k+i-1)}{(i-1)!} L_t^{(i-1)}$$

$$= -\gamma (k+i) T_t^{(i)} + \lambda (k+i) \phi_{i,k} T_t^{(i+1)} + \mu (k+i) T_t^{(i-1)},$$

(3.24)

where

$$\phi_{i,k} = \frac{(2k+i)(i+1)}{(k+i+1)(k+i)} = 1 - \frac{k(k-1)}{(k+i+1)(k+i)}$$

is the probability that a coalescing pair of randomly chosen lineages (from $(k+i+1)$ total lineages) does not consist of two sampled lineages. This shows that $T_t^{(i)}$ satisfies the Master equation (3.24) across any epoch.

One can see that at punctual events the transition conditions (3.11) and (3.16) hold for $T_t^{(i)}$ for $\psi$-sampling and branching events respectively. Moreover at $\omega$-sampling events the transition condition (3.14) transforms to

$$T_t^{(i)} = \omega (k+i) T_{t-}^{(i-1)}.$$ 

With these transition conditions and initial condition $T_0^{(i)} = \frac{(k_0+i)!}{i!} L_0^{(i)} = \frac{(k_0+i)!}{i!} (1-\rho)^{k_0} \rho^i$, the Master equation (3.24) was solved explicitly in [Gupta et al., 2019] and the solution is of the form

$$T_t^{(i)} = \sum_{\ell=0}^{q} \frac{(k+i)!}{\ell!(i-\ell)!} W_t^{(\ell)}$$

where $q$ is the number of $\omega$-sampling events in the time-interval $[0,t]$ and the $q+1$-dimensional time-varying vector $W_t = (W_t^{(0)}, \ldots, W_t^{(q)})$ can be analytically computed following the approach in [Gupta et al., 2019].

Therefore from (3.23) we state the following proposition.

**Proposition 3.3.** At any time $t$, we can compute the $L_t^{(i)}$ values as

$$L_t^{(i)} = \sum_{\ell=0}^{q} \frac{i!}{\ell!(i-\ell)!} W_t^{(\ell)}.$$ 

where $W_t$ is a $q$ dimensional time-varying vector which can be computed following Algorithm 2 in [Gupta et al., 2019].

Note that when there is no $\omega$-sampling, then $q = 0$ for all times and $W_t^{(0)}$ is the same as $W_t$ defined in the previous section.

This ends our section on the computation of $L_t$. It thus remains to (i) present a way to compute $M_t$ and (ii) combine $L_t$ and $M_t$ to get the target distribution $K_t$ at any time $t$. We do this in turn in the next two sections.
4. The joint density of observations above and past population size

Recall that we are now interested in computing, \( \forall i \in \mathbb{N}, \ M_t^{(i)} := \mathbb{P}(T_t^\uparrow, O_t^\uparrow, I_t = k_t + i) \). We start by presenting the Master equation satisfied by \( M_t \), before turning to its resolution for specific parameter sets. The approach is very similar to the one presented in the previous section to compute \( L_t \), with the slight difference that we will need to traverse the tree forward in time instead of backward in time.

4.1. Master equation driving \( M_t \)

At the time of origin of the process \( t_{or} \), we only observe one starting lineage in \( T_{t_{or}}^\downarrow \). This provides us with the following initialization condition on \( M_t \),

\[
M_{t_{or}}^{(i)} = \mathbb{P}(N_{t_{or}} = 1 + i) = 1_{i=0}
\]

We then derive the Master equation driving the evolution of \( M_t \) across an epoch on which the number of observed lineages is fixed and equal to \( k \). Suppose we know \( M_t \), and we observe no punctual event on the infinitesimal time interval \( (t - \delta t, t) \). Unobservable events have already been illustrated in Figure 2. It allows us to get

\[
M_{t-\delta t}^{(i)} = (1 - \gamma(i + k)\delta t) M_t^{(i)} + \lambda(2k + i - 1)\delta t 1_{i>0} M_t^{(i-1)} + \mu(i + 1)\delta t M_t^{(i+1)}
\]

Letting \( \delta t \to 0 \) and combining this with the initialization, we get the following set of ODEs driving the evolution of \( M_t \),

\[
\forall i \in \mathbb{N}, \ M_{t_{or}}^{(i)} = 1_{i=0} \quad M_t^{(i)} = \gamma(i + k_t)M_t^{(i)} - \lambda(2k_t + i - 1)1_{i>0}M_t^{(i-1)} - \mu(i + 1)M_t^{(i+1)}
\]

Last, we need to take into account the evolution of \( M_t \) at punctual events. Again, there are 6 types of punctual events that we can come across at time \( t \) in the past, listed below and illustrated in Figure 2. We denote \( M_t^- \) the probability just after (i.e. below) the punctual event and \( M_t^+ \) the probability immediately before (i.e. up). Because we are here deriving \( M_t \) forward in time, one needs to carefully note differences with results derived in section 3 relating to the number of lineages before and after the event. We can indeed find the same punctual events, namely,

1. a leaf of \( T_t^\downarrow \), labeled as removed. This is a \( \psi \)-sampling with removal event for which the number of sampled lineages decreases by one and the number of unsampled lineages remains unchanged. This gives,

\[
M_{t^-}^{(i)} = \psi_r M_{t^+}^{(i)}
\]

(4.26)
2. a leaf of $T^i_t$, labeled as non-removed. This is a $\psi$-sampling without removal event for which one sampled lineages becomes unsampled. This gives,

$$M^{(i)}_t = \psi(1 - r)M^{(i)}_{t^-}$$  \hspace{1cm} (4.27)

3. a sampled ancestor along a branch of $T^i_t$, necessarily labeled as non-removed. This is a $\psi$-sampling without removal event which does not affect the number of lineages. It gives,

$$M^{(i)}_t = \psi(1 - r)M^{(i)}_{t^+}$$  \hspace{1cm} (4.28)

4. an occurrence in $O^i_t$, labeled as removed. This is a $\omega$-sampling with removal event, for which the number of unsampled lineages decreases by one. This gives,

$$M^{(i)}_t = \omega r(i + 1)M^{(i+1)}_{t^-}$$  \hspace{1cm} (4.29)

5. an occurrence in $O^i_t$, labeled as non-removed. This is a $\omega$-sampling without removal event which does not affect the number of lineages. It gives,

$$M^{(i)}_t = (k + i)\omega(1 - r)M^{(i)}_{t^+}$$  \hspace{1cm} (4.30)

6. a branching event between two branches of $T^i_t$.

This is a $\lambda$-event increasing the number of sampled lineages by one. This gives,

$$M^{(i)}_t = \lambda M^{(i)}_{t^-}$$  \hspace{1cm} (4.31)

Finally, upon reaching present time 0, one needs to take into account the $\rho$-sampling, leading to the following update,

$$M^{(i)}_0 = (1 - \rho)^i \rho^k M^{(i)}_{0^-}$$  \hspace{1cm} (4.32)

As already exposed for $L_t$, we can build a similar algorithm to compute $M_t$ in the general case, relying on a numerical ODE solver for approximating equation (4.25). As for algorithm 1, a slight variation of this algorithm would allow one to compute an estimate of the probability density of $(T, O)$. Note that this strategy is very close to what has been used to compute the probability density of a reconstructed tree under a logistic birth-death process [Etienne et al. 2012].

In the remainder of this section, we derive analytical results to avoid resorting to a numerical ODE solver.
4.2. The corresponding probability generating function

We introduce now the probability generating function corresponding to the density $M_t$, which will prove useful to get analytical results.

$$\hat{M}(t, z) := \sum_{i=0}^{\infty} z^i M_t^{(i)}$$

The initial condition on $M$ translates in $\hat{M}$ as, $\forall z, \hat{M}(t_0, z) = 1$. The ODE (4.25) furthermore translates as the following partial differential equation (PDE),

$$\partial_t \hat{M} = \sum_{i=0}^{\infty} z^i \left( \gamma (i + k) M_t^{(i)} - \lambda (2k + i - 1) \mathbb{1}_{i>0} M_t^{(i-1)} - \mu (i + 1) M_t^{(i+1)} \right)$$

$$= \gamma k \sum_{i=0}^{\infty} z^i M_t^{(i)} + \gamma \sum_{i=1}^{\infty} i z^i M_t^{(i)} - \lambda \sum_{i=0}^{\infty} z^{i+1} (2k + i) M_t^{(i)} - \mu \sum_{i=1}^{\infty} i z^{i-1} M_t^{(i)}$$

$$= -k (2\lambda z - \gamma) \hat{M} - (\lambda z^2 - \gamma z + \mu) \partial_z \hat{M}$$

Our target probability generating function $\hat{M}$ is thus the solution of the following PDE problem across a given epoch $(t_{h-1}, t_h)$, on which the number of observed lineages remains constant and equal to $k$,

$$\hat{M}(t_h, z) = F(z)$$

$$\partial_t \hat{M} + (\lambda z^2 - \gamma z + \mu) \partial_z \hat{M} + k(2\lambda z - \gamma) \hat{M} = 0 \quad (4.33)$$

Solving this PDE problem allows us to come with an analytical expression of $\hat{M}$ any time across an epoch, provided we now the expression of $\hat{M}(t_h, z)$ at the end of the epoch.

Proposition 4.1. The solution to the PDE problem (4.33) is given by

$$\hat{M}(t, z) = F(u(t_h - t, z)) R(t_h - t, z)^k$$

where we introduce $R(t, z) = p(t, z)/(1 - z)$ to ease notation.

Proof We used the method of characteristics, see derivations in Appendix B.

Between epochs, one must also update $\hat{M}$ according to punctual events taking place. Previously presented updates of $M$ (equations (4.26) to (4.31)) translate as the following for $\hat{M}$,

if $t$ is a removed leaf, $\hat{M}(t^-, z) = \sum_{i=0}^{\infty} z^i \left( \psi_r M_t^{(i)} \right) = \psi_r \hat{M}(t^+, z)$ \quad (4.34)
if \( t \) is a non-removed leaf,  
\[
\hat{M}(t^-_i, z) = \sum_{i=0}^{\infty} z^i \left( (1-r)z \hat{M}_{i+1}^{(i)} \right) = (1-r)z\hat{M}(t^+, z) \tag{4.35}
\]

if \( t \) is a sampled ancestor,  
\[
\hat{M}(t^-_i, z) = \sum_{i=0}^{\infty} z^i \left( (1-r)M_{i+1}^{(i)} \right) = (1-r)\hat{M}(t^+, z) \tag{4.36}
\]

if \( t \) is a removed occurrence,  
\[
\hat{M}(t^-_i, z) = \sum_{i=0}^{\infty} z^i \left( (\omega_{i+1})M_{i+1}^{(i+1)} \right) = \omega\hat{M}(t^+, z) \tag{4.37}
\]

if \( t \) is a non-removed occurrence,  
\[
\hat{M}(t^-_i, z) = \sum_{i=0}^{\infty} z^i \left( \omega (1-r)(k+i)M_{i+1}^{(i)} \right) = \omega(1-r) \left( k\hat{M}(t^+, z) + z\hat{M}(t^+, z) \right) \tag{4.38}
\]

if \( t \) is a branching event,  
\[
\hat{M}(t^-_i, z) = \sum_{i=0}^{\infty} z^i \left( \lambda M_{i+1}^{(i)} \right) = \lambda\hat{M}(t^+, z) \tag{4.39}
\]

If we are interested in the distribution at some point, we can thus start the formula at \( t_{sw} \) with \( F(z) = 1 \), and then iteratively alternate between the updates at punctual events and the use of Proposition 4.1 over each epoch. When reaching present time 0, the step of \( \rho \)-sampling expressed in equation (4.32) moreover translates here as,
\[
\hat{M}(0^-, z) = \sum_{i=0}^{\infty} z^i (1-\rho)^i \rho^{i+1} M_{i+1}^{(i)} = \rho^{k_0} \hat{M}(0^+, (1-\rho)z) \tag{4.40}
\]

While this procedure in theory allows us to get the analytical formula of \( \hat{M} \) at any time, updates (4.37) and (4.38) require differentiating the probability generating function, greatly complicating the expression of the function after a few occurrences. When \( \omega = 0 \), these two updates disappear and a nice recursion leads to a closed-form formula that we will detail in Proposition 4.3.

We also implemented this procedure in a programming language able to deal with symbolic calculus, but were not able to provide an implementation efficient enough to apply to standard datasets in the field. Instead, when \( \omega \neq 0 \), we suggest another strategy for computing the \( M_{i+1}^{(i)} \)'s, by approximating \( \hat{M} \) across punctual events by a polynomial of order \( N \), while still relying on Proposition 4.1 to drive the evolution of the probability generating function between events. This is a more efficient alternative to numerically solving the ODE system. We only need to derive the expression of the generating function at punctual events as given in the following proposition 4.2.

**Proposition 4.2.** The derivatives in \( z = 0 \) of a generative function which expression is
\[
\hat{M}(t_h - t, z) := R(t_h - t, z) \sum_{l=0}^{N} M_{l}^{(l)} u(t_h - t, z)
\]

is
can be numerically computed using the formula
\[
\left( \partial_z^i \tilde{M}(t_h - t, z) \right)_{z=0} = \left( \frac{\Delta}{z} e^{-\sqrt{\Delta}(t_h - t)} \right)^k \sum_{\alpha=0}^i \sum_{l=\alpha}^N \tilde{M}_l^{(i)} \left( \begin{array}{}{\alpha} \\ {l - \alpha} \end{array} \right) \frac{l! \prod_{m=0}^{l-\alpha-1} (2k + l + m)}{(l - \alpha)!} \\
(\alpha \left( x_1 x_2 \right)^{l-\alpha} \left( 1 - e^{-\sqrt{\Delta}(t_h - t)} \right)^{l+i-2\alpha} \left( x_2 - x_1 e^{-\sqrt{\Delta}(t_h - t)} \right)^{(2k+l+i-\alpha)})
\]

**Proof** The derivation is detailed in Appendix C.1.

This derivation is at the heart of algorithm 2, for it allows to drive the evolution of the \( \tilde{M}_t^{(i)} \)'s through each epoch, as well as at times when we want to record them.

We will refer to algorithm 2 as the slight variation of this algorithm aimed at computing the density of \((T, O)\). No set of time points \((\tau_j)\) is required, and the values of \(\tilde{M}_t\) are not recorded through time (i.e. matrix \(B'\) disappears). Instead, when reaching \(t_h = t_0\) we simply return \(\sum_{i=0}^N (1 - \rho)^i \tilde{M}^{(i)}\).
Algorithm 2 Computes a numerical approximation of $M_t$ for a specific set of times

**Input:**
- Observed tree and occurrence data $(\mathcal{T}, \mathcal{O})$,
- parameters $(l_{or}, \lambda, \mu, \psi, \omega, \rho)$,
- set of time points $(\tau_j)_{j=1}^S$ for which we want to compute the density,
- and the truncation $N$ setting the accuracy of the algorithm.

**Output:** A numerical approximation of $M_t$ at times $(\tau_j)_{j=1}^S$, $(\bar{M}^{(i)}_{\tau_j})_{i \in \{0,1,...,N\}, j \in \{1,2,...,S\}}$.

1. Pool all $(\tau_j)$ and all branching and sampling times of $(\mathcal{T}, \mathcal{O})$ in an ordered list $(t_h)_{h=1}^n$
2. Set $j = S$ and $B'$ as a $S \times (N+1)$ empty matrix
3. Set $\forall i \in \{0,1,\ldots,N\}$, $\bar{M}^{(i)} = I_{i=0}$
4. Set $k = 1$
5. for $h = n-1,n-2,\ldots,0$ do
6. Compute the values right before the punctual event,

$$\bar{M}^{(i)} = \left( \frac{\lambda}{\sqrt{2\pi}} \right)^{i+1} \sum_{a=0}^{\infty} \sum_{l=0}^{\infty} \bar{M}^{(l)}_{t_h} \left( \frac{l}{\alpha} \right) \frac{1}{(i - \alpha)!} \left( \prod_{m=0}^{i-\alpha} \frac{1}{2k + l + m} \right) \left( -x_1 + x_2 e^{-\sqrt{2}t_h} \right)^{\alpha} (x_1 x_2)^{1-\alpha} \left( 1 - e^{-\sqrt{2}t_h} \right)^{l+i-2\alpha} \left( x_2 - x_1 e^{-\sqrt{2}t_h} \right)^{-(2k+l+i-\alpha)}$$
7. if $t_h = \tau_j$ then
8. Record the result in $B'$: $\forall i, B^{(i,j)} = \bar{M}^{(i)}$
9. Set $j = j - 1$.
10. end if
11. if $t_h = 0$ or $t_h = \tau_S$ then
12. return $B'$
13. else if $t_h$ is a removed leaf then
14. Update $\forall i$, $\bar{M}^{(i)} = \psi_r \bar{M}$
15. Set $k = k - 1$
16. else if $t_h$ is a non-removed leaf then
17. Update $\bar{M}^{(0)} = 0$ and $\forall i > 0$, $\bar{M}^{(i)} = \psi(1-r) \bar{M}$
18. Set $k = k - 1$
19. else if $t_h$ is a sampled ancestor then
20. Update $\forall i$, $\bar{M}^{(i)} = \psi(1-r) \bar{M}$
21. else if $t_h$ is a removed occurrence then
22. Update $\forall i < N$, $\bar{M}^{(i)} = \omega(r(i+1)) \bar{M}$ and $\bar{M}^{(N)} = 0$
23. else if $t_h$ is a non-removed occurrence then
24. Update $\forall i$, $\bar{M}^{(i)} = \omega(1-r)(k+i) \bar{M}$
25. else $t_h$ is a branching event
26. Update $\forall i$, $\bar{M}^{(i)} = \lambda \bar{M}$
27. Set $k = k + 1$
28. end if
29. end for

One may wonder why we introduced a probability generating function in this section and not in the
previous one to fasten Algorithm 1. It turns out that we tried to follow the same approach to derive $L_t$, as
described in Appendix E, but it leads to another PDE problem that will require further work to be solved.

4.3. Special case $\omega = 0$

This corresponds to the special case leading to the observation of $O = \emptyset$. In that case, a nice recursion
leads to a closed-form formula for $\hat{M}$.

**Proposition 4.3.** At any time $t$, considering that we observed so far – i.e. on $(t, t_{ow})$ – $v$ sampled ancestors,
$w$ removed leaves at times $t_j \in W$, $x$ branching events at times $t_j \in X$, $y$ non-removed leaves at times $t_j \in Y$, we get,

$$
\hat{M}(t, z) = \lambda^x \psi^{w+y} r^w (1-r)^{v+y} \prod_{t_j \in X \cup \{t_{ow}\}} R(t_j - t, z) \prod_{t_j \in W} R(t_j - t, z)^{-1} \prod_{t_j \in Y} u(t_j - t, z) R(t_j - t, z)^{-1}
$$

**Proof** We prove this result by induction across the epochs of $T$ in Appendix D, using as the main arguments
the update equations (4.34), (4.35), (4.36), (4.39), combined with Proposition 4.1 driving the evolution across
an epoch.

As a simple corollary of this result, when $t_{ow} = 0$ is the present, we get back the same probability density
formula of $T$ as provided, e.g. in theorem 3.5 in Stadler (2010) (when $r = 0$), in section 3 in Gupta et al.
(2019) (when $r = 1$), or in our previous corollary 3.2.

**Proof** Proposition 4.3 offers yet another proof of corollary 3.2 by noting that

$$
L(T) = \sum_{i=0}^{\infty} M_i(t) = \hat{M}(0^+, 1) = \rho^w \hat{M}(0^+, 1 - \rho)
$$

where the last equality follows from equation (4.40) taking into account the $\rho$-sampling at present.

When $\omega = 0$, Proposition 4.3 also offers an alternative to algorithm 2 for deriving $M_t$. Indeed, resorting
to the probability generating function to get back the probability density, one can get the following corollary.

**Corollary 4.4.** At any time $t$, considering that we observed so far – i.e. on $(t, t_{ow})$ – $v$ sampled ancestors,
$w$ removed leaves at times $t_j \in W$, $x$ branching events at times $t_j \in X$, $y$ non-removed leaves at times $t_j \in Y$, we can compute $M_i^{(i)}$ using the following recursion,

$$
M_i^{(0)}(t) = \lambda^x \psi^{w+y} r^w (1-r)^{v+y} \prod_{t_j \in X \cup \{t_{ow}\}} R(t_j - t, 0) \prod_{t_j \in W} R(t_j - t, 0)^{-1} \prod_{t_j \in Y} u(t_j - t, 0) R(t_j - t, 0)^{-1}
$$

$$
M_i^{(i)} = \frac{1}{\lambda^x} \sum_{\alpha=1}^{i} M_i^{(i-\alpha)} C(\alpha)
$$

20
where we define

\[
C^{(\alpha)} = 2 \sum_{t_j \in X \cup \{t_{\infty}\}} a_{t_{j-\ell}} - 2 \sum_{t_j \in W} a_{t_{j-\ell}} - \sum_{t_j \in Y} (a_{t_{j-\ell}} + b_{t_{j-\ell}})
\]

\[
a_{t} = \left(1 - e^{-\sqrt{\Delta t}}\right) \left(x_2 - x_1 e^{-\sqrt{\Delta t}}\right)^{-1}
\]

\[
b_{t} = \left(x_1 - x_2 e^{-\sqrt{\Delta t}}\right) \left(x_1 x_2 - x_2 x_1 e^{-\sqrt{\Delta t}}\right)^{-1}
\]

**Proof** The probability density \(M_{t}^{(i)}\) can be found back by taking

\[
M_{t}^{(i)} = \frac{1}{\mathcal{A}} \left(\partial_t \tilde{M}(t, z)\right)_{z=0}
\]

The result follows from the derivation of these derivatives in Appendix C.2. \(\square\)

This special case ends the section. In the next one, we will combine results from sections 3 and 4 and use our ability to compute \(L_{t}\) and \(M_{t}\) to compute \(K_{t}\), the probability distribution of the population size given \((T, O)\).

5. The distribution of past population size conditioned on observations

5.1. The distribution at fixed times

In section 3 we exposed how to compute \(L_{t}\), the probability density of the observations below time \(t\) conditioned on the population size. This relies either on Algorithm 1 in the general case, or on the more optimized Proposition 3.1 in case \(\omega = 0\), or Proposition 3.3 in case \(r = 1\).

In section 4 we exposed how to compute \(M_{t}\), the probability density of the observations above time \(t\) and the population size. This relies either on Algorithm 2 in the general case, or on the more optimized Corollary 4.4 when \(\omega = 0\).

We now combine our ability to compute \(L_{t}\) and \(M_{t}\) to derive the probability distribution of the population size given \((T, O)\). Recall from the first section that,

\[
K_{t}^{(i)} = \mathbb{P}(I_t = k_t + i \mid T, O) \propto L_{t}^{(i)} M_{t}^{(i)}
\]

Provided we have stored numerical values \(B = \{\tilde{L}_{t_{j}}^{(i)}\}_{i \in \{0, 1, \ldots, N\}}^{j \in \{1, 2, \ldots, S\}}\) and \(B' = \{\tilde{M}_{t_{j}}^{(i)}\}_{i \in \{0, 1, \ldots, N\}}^{j \in \{1, 2, \ldots, S\}}\) for a set of time points \((t_{j})_{j=1}^{S}\), it only remains to compute

\[
\forall j \in \{1, 2, \ldots, S\} \quad D^{(j)} = \frac{B^{(j-\cdot)} \odot B'^{(j-\cdot)}}{B^{(j-\cdot)^{tr}} B'^{(j-\cdot)^{tr}}}
\]

where \(\odot\) refers to the Hadamard product, and \(^{tr}\) is the transpose of the vector.
Depending on the parameter space that one wants to consider, it thus remains to arrange pieces stemming from the previous sections. We provide a flowchart in Figure 4 to guide the reader to choose the most efficient path.

$$\Pr (\mathcal{T}, \mathcal{O})$$ Corollary 3.2 Proposition 3.3 Algorithm 2’

I want to compute

$$K_t$$ Proposition 3.1 Proposition 3.3 & Algorithm 1 & Corollary 4.4 & Algorithm 2 & Algorithm 2

Figure 4: The most efficient results depending on the parameter space considered. In red, results already described in Stadler (2010) and Gupta et al. (2019). In blue, the new contribution of this manuscript.

5.2. Generator of trajectories

The previous result gives us the distribution of the population size at any time in the past, but does not state anything about population size trajectories. We provide now an approximate way of simulating population size trajectories conditioned on $$(T, O)$$. 

Indeed, recall we have,

$$K_t^{(i)} := \Pr (N_t = k_t + i \mid T, O) \propto L_t^{(i)} M_t^{(i)}$$

$$\dot{L}_t^{(i)} = -\beta (k_t + i) L_t^{(i)} + \lambda (2k_t + i) L_t^{(i+1)} + \mu i L_t^{(i-1)}$$

$$\dot{M}_t^{(i)} = \beta (k_t + i) M_t^{(i)} - \mu (i+1) M_t^{(i+1)} - \lambda (2k_t + i - 1) \mathbb{1}_{i>0} M_t^{(i-1)}$$

We thus get,

$$K_t^{(i)} \propto \dot{L}_t^{(i)} M_t^{(i)} + L_t^{(i)} \dot{M}_t^{(i)}$$

$$\propto -\beta (k_t + i) L_t^{(i)} M_t^{(i)} + \lambda (2k_t + i) L_t^{(i+1)} M_t^{(i)} + \mu i L_t^{(i-1)} M_t^{(i)}$$

$$+ \beta (k_t + i) M_t^{(i)} L_t^{(i)} - \mu (i+1) M_t^{(i+1)} L_t^{(i)} - \lambda (2k_t + i - 1) \mathbb{1}_{i>0} M_t^{(i-1)} L_t^{(i)}$$

$$\propto \lambda (2k_t + i) \frac{L_t^{(i+1)}}{L_t^{(i)}} K_t^{(i)} + \mu \frac{L_t^{(i-1)}}{L_t^{(i)}} K_t^{(i)} - \lambda (2k_t + i - 1) \mathbb{1}_{i>0} K_t^{(i-1)} \frac{L_t^{(i)}}{L_t^{(i-1)}} - \mu (i+1) K_t^{(i+1)} \frac{L_t^{(i)}}{L_t^{(i+1)}}$$

$$\propto Q_t^{(i, i)} K_t^{(i)} + Q_t^{(i+1, i)} K_t^{(i+1)} + Q_t^{(i-1, i)} K_t^{(i-1)}$$

(5.41)
Where we introduced in the last line the following notation,
\[ Q^{(i+1,i)}_t = -\mu (i+1) \frac{L^{(i)}_t}{L^{(i+1)}_t}, \]
\[ Q^{(i-1,i)}_t = -\lambda (2k_t + i - 1) \mathbb{1}_{i>0} \frac{L^{(i)}_t}{L^{(i-1)}_t}, \]
\[ Q^{(i,i)}_t = \lambda (2k_t + i) \frac{L^{(i+1)}_t}{L^{(i)}_t} + \mu i \frac{L^{(i-1)}_t}{L^{(i)}_t}. \]

Using these, we see that \( Q^{(i,i)}_t = -\left( Q^{(i,i+1)}_t + Q^{(i,i-1)}_t \right) \). This allows us to draw trajectories of the number of ancestors in the past as a continuous in time Markov process with the (inhomogeneous) rates \( Q_t \) written above.

Remark that we could equally write these ODE coefficients using the \( M_t^{(i)} \)'s, depending on which traversal we prefer to perform prior to the simulation. This gives,
\[
\dot{K}^{(i)}_t \propto \lambda (2k_t + i) \frac{M^{(i+1)}_t}{M^{(i)}_t} K^{(i+1)}_t + \mu i \frac{M^{(i-1)}_t}{M^{(i)}_t} K^{(i)}_t - \lambda (2k_t + i - 1) \mathbb{1}_{i>0} \frac{M^{(i-1)}_t}{M^{(i)}_t} K^{(i)}_t
\]
\[ \propto R^{(i+1,i)}_t K^{(i)}_t + R^{(i,i+1)}_t K^{(i+1)}_t + R^{(i-1,i)}_t K^{(i)}_t \tag{5.42} \]

where we introduced in the last line the following notation,
\[ R^{(i+1,i)}_t = \lambda (2k_t + i) \frac{M^{(i)}_t}{M^{(i+1)}_t}, \]
\[ R^{(i-1,i)}_t = \mu i \frac{M^{(i)}_t}{M^{(i-1)}_t}, \]
\[ R^{(i,i)}_t = -\lambda (2k_t + i - 1) \mathbb{1}_{i>0} \frac{M^{(i-1)}_t}{M^{(i)}_t} - \mu (i+1) \frac{M^{(i+1)}_t}{M^{(i)}_t}. \]

This is a standard result for Markov chains that are conditioned on a final state, and the shape of the newly derived transition kernel is called a Doob’s transform. Note that these transitions simplify for special cases when we have an analytical expression of either \( L^{(i)}_t \) or \( M^{(i)}_t \).

5.3. Numerical implementation

Results of this paper have been implemented numerically and the code is freely available on GitLab: https://gitlab.com/MMarc/popsize-distribution/.

We used the numerical implementation to verify the correctness of the results in several ways:

1. We verified that the values of the probability density of \((T,O)\) computed using \( L_t \) and \( M_t \) (algorithms 1 and 2) were equivalent to values computed using already known formulas when \( \omega = 0 \) (Stadler, 2010) or when \( r = 1 \) (Gupta et al., 2019). See result in Figure 5AB.
2. We verified that the values of the probability density of \((T, O)\) computed using \(L_t\) or \(M_t\) (algorithms 1 and 2) were identical on examples for which no previous formula was known. See result in Figure S1.

3. We assessed the distribution of the population size against the only numerical method performing the same goal, the particle filtering developed in Vaughan et al. (2019). We compared values of a few quantiles computed using the two methods, see result in Figure 5DEF).

![Graphs](image-url)

Figure 5: Assessment of the accuracy of the methods presented in this paper, on toy datasets. First row, probability density of data, A) against known analytical formula when \(\omega = 0\) and \((\mu, \rho, \psi, r) = (1, 0.5, 0.3, 0.6); B) against known analytical formula when \(r = 1\) and \((\mu, \rho, \psi, \omega) = (1, 0.5, 0.3, 0.6); C) obtained using algorithms 1’ or 2’ otherwise, with \((\mu, \rho, \psi, r, \omega) = (1, 0.5, 0.3, 0.2, 0.6). Second row, quantiles of the population size distribution, against the particle filter in Vaughan et al. (2019), with parameters \((\mu, \rho, \psi, r, \omega) = (1, 0.1, 0.001, 0.5, 0.001). D) quantile of level 0.2; E) median; F) quantile of level 0.8.

We also illustrate in Figure 6 our target distribution \(K_t\) of the past population size conditioned on \((T, O)\), on a few simulated examples.

24
Figure 6: Hints of the true population size (in green) under three different processes. A) A homogeneous birth-death with \( \rho \)-sampling at present; B) A homogeneous birth-death with \( \rho \)-sampling at present and \( \psi \)-sampling through time; C) A homogeneous birth-death process with \( \rho \), \( \psi \)– and \( \omega \)-sampling.
6. Discussion

The results we have derived in this paper fit into two main categories. The first category concerns results allowing one to compute the probability density \( L \) of \((T, O)\). While results allowing to compute \( L \) when \( \omega = 0 \) or \( r = 1 \) were already known, our two algorithms 1' and 2' have the potential to improve the computation time of \( L \) when \( \omega \neq 0 \) and \( r \neq 1 \).

The second category concerns our main results, allowing one to compute the probability distribution of the population size in the past and to generate population size trajectories conditioned on \((T, O)\) (section 5). These constitute a dramatically more efficient alternative to the Monte-Carlo particle filtering algorithm developed by [Vaughan et al. (2019)] to perform the same task.

We discuss first how these results can be used in the context of current phylodynamics, before turning to future extensions.

6.1. Use for inference purposes

When applying phylodynamics methods to empirical data, it is common practice in the field to condition the model generating the phylogeny on either (i) the observation of at least one individual at present, descending from time \( t_{\text{ar}} \), an event that we call \( S_{t_{\text{ar}}}^1 \); or (ii) the observation of at least two individuals at present, descending from time \( t_{\text{mrca}} \), an event that we call \( S_{t_{\text{mrca}}}^2 \), where \( \text{mrca} \) stands for most recent common ancestor. Indeed, we can study a dataset precisely because it survived to the present.

The probability density of \((T, O)\) needs be conditioned on \( S_{t_{\text{ar}}}^1 \) or \( S_{t_{\text{mrca}}}^2 \) when used in empirical applications, which can be done by dividing \( L \) by, respectively, \((1 - u_{t_{\text{ar}}})\) and \((1 - u_{t_{\text{mrca}}})^2\).

However, since \( S_{t_{\text{ar}}}^1 \), \( S_{t_{\text{mrca}}}^2 \) are always satisfied on empirical data, and we already condition on the observation of empirical data, nothing more needs to be done to use the probability distribution \( K_2 \) to retrieve past population sizes.

If the tree is directly fixed and observed, results from the first category can be used to perform parameter inferences in a maximum likelihood framework. Taken as a function of the parameters, the probability density of \((T, O)\) is called the likelihood, and can be numerically optimized to look for parameters maximizing its value (i.e. maximum likelihood estimators).

If one also knows the true parameter values or has inferred estimates of those, results from the second category can also be used to compute the distribution of the population size in the past. Any quantile of interest can then be numerically computed from this distribution. As an example, we chose to represent on Figure 6 the 95% envelope of the distribution, comprised between the level 2.5% and the level 97.5% quantiles. Alternatively, if one needs a point estimate of the population size, the median of the distribution is a good candidate.
If the tree is not directly observed, but sequencing data or any other set of phenotypic measurements, called \( A \), is available for the set of \( \psi \)-sampled individuals, then one might consider analyzing those in a Bayesian framework. For instance, consider a model describing the evolution of molecular sequences or phenotypic measurements along a fixed tree is chosen, relying on parameters that we call \( \theta \). We are then interested in quantifying the posterior distribution

\[
f(T, \theta, \lambda, \mu, \rho, \psi, \omega, t_{or} | O, A \propto f(A | T, \theta) f(\mathcal{O}, T | \lambda, \mu, \rho, \psi, \omega, t_{or}) f(\lambda, \mu, \rho, \psi, \omega, \theta, t_{or}),
\]

with \( f(\lambda, \mu, \rho, \psi, \omega, \theta, t_{or}) \) being the prior distribution on the model parameters. Standard MCMC algorithms can be used to sample from this posterior distribution. Results from the second category can then be used and applied to a sample from the posterior distribution to integrate the distribution of the population size over the posterior distribution of trees and parameters values.

In addition, the distribution \( K_t \) of population sizes (or a trajectory sampled from this distribution using the results of section 5.2) can be computed directly for each sampled combination of tree and parameters, and thus used to compute the full joint posterior of both population size trajectories, trees and model parameters. While this is similar to the capability of the method of Vaughan et al. (2019), we expect the method presented here to be far more computationally efficient as it cuts out the necessity of simulating particle trajectories within the MCMC step. (The trajectory specific calculations or simulations in our method need only be done after the MCMC algorithm is complete, and only for a subsample of the full MCMC chain states.)

### 6.2. Future extensions

This process lends itself well for various biologically realistic extensions to allow closer fit to empirical data in a variety of situations.

The first extension that we envision is to relax the assumption of rate homogeneity and instead work with time-varying rates. This has already been considered in different studies relying on birth-death processes, either with exponentially varying functions (Morlon et al., 2011) or with piecewise constant rates (a model dubbed as skyline birth-death process, see Stadler et al., 2013; Gavryushkina et al., 2016). Because most of what we wrote can be straightforwardly adapted to such a framework, this would not require much theoretical work. However, the challenge would be to do so without overfitting the data.

Another popular extension that has been described in the literature on birth-death processes for phylo-dynamics is to consider multi-type birth-death processes (Maddison et al., 2007). Each individual is assigned a type, which impacts its propensity to give birth to other types. All sampling-related parameters can also be considered type-dependent. The challenge here boils down to dealing with an increase of dimensionality,
where we would be interested in the joint distribution of subpopulation sizes. Recent work by Freyman and Höhna (2018) could first allow us to find the distribution of the types along sampled lineages in the tree. If we have \( j \) types, we would then be interested in
\[
K_t^{(i_1), (i_2), \ldots, (i_j)} = P \left( I_t^{(1)} = k_t^{(1)} + i^{(1)}, I_t^{(2)} = k_t^{(2)} + i^{(2)}, \ldots, I_t^{(j)} = k_t^{(j)} + i^{(j)} \mid T_c, O \right)
\]
where all quantities are type-dependent analogs of quantities defined in the current paper, and \( T_c \) would be a type-coloured tree, i.e. a tree where each sampled lineage is assigned a type. This extension is particularly interesting for epidemiological applications, when different populations of infected individuals, clustered according to some characteristic (e.g. patient behaviour or geography) might have very different dynamics (Stadler and Bonhoeffer, 2013).

Finally, we are very hopeful that this piece of work could be applied as well to density-dependent birth-death processes, also known as logistic birth-death models. Indeed, very similar ideas to the breadth-first forward and backward traversals as applied in algorithms 1' and 2' appear in the context of logistic birth-death models (Etienne et al., 2012; Leventhal et al., 2013). Preliminary results obtained by adapting our numerical algorithms to this framework are very encouraging, and we are currently in the process of deriving as much analytical results as we can to speed up the method. We are hoping to present this in a subsequent paper.

6.3. Conclusion

This manuscript presents a way to efficiently compute the distribution of the number of ancestors in a linear birth-death process, conditioned on the observation of a reconstructed tree and a record of occurrences through time. The method paves the way to the consideration of more complex demographic scenarios, with either time-dependent or density-dependent parameters. We hope that it will foster important research advances for unravelling demographic histories in epidemiology, macroevolution, and any other fields where birth-death processes form a relevant model framework.

Acknowledgements

The authors are very grateful to Rachel Warnock for helpful comments on potential applications of the model.

References

Billaud, O., Moen, D.S., Parsons, T.L., Morion, H., 2019. Estimating diversity through time using molecular phylogenies: Old and species-poor frog families are the remnants of a diverse past. bioRxiv doi:10.1101/698620
Etienne, R.S., Haegeman, B., Stadler, T., Aze, T., Pearson, P.N., Purvis, A., Phillimore, A.B., 2012. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. Proceedings of the Royal Society B: Biological Sciences 279, 1300–1309.

Foote, M., 2000. Origination and extinction components of taxonomic diversity: general problems. Paleobiology 26, 74–102.

Freyman, W.A., Höhna, S., 2018. Stochastic character mapping of state-dependent diversification reveals the tempo of evolutionary decline in self-compatible onagraceae lineages doi: 10.1093/sysbio/syy078

Gavryushkina, A., Heath, T.A., Ksepka, D.T., Stadler, T., Welch, D., Drummond, A.J., 2016. Bayesian total-evidence dating reveals the recent crown radiation of penguins. Systematic Biology 66, 57–73. doi: 10.1093/sysbio/syu060

Gray, R.D., Drummond, A.J., Greenhill, S.J., 2009. Language phylogenies reveal expansion pulses and pauses in pacific settlement. Science 323, 479–483.

Gupta, A., Manceau, M., Vaughan, T., Khammash, M., Stadler, T., 2019. The probability distribution of the reconstructed phylogenetic tree with occurrence data. bioRxiv , 679365.

Heath, T.A., Huelsembeck, J.P., Stadler, T., 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences 111, -2957.

Kendall, D.G., 1948. On the generalized ‘birth-and-death’ process. Ann. Math. Stat. 19, 1–15.

Leventhal, G.E., Günthard, H.F., Bonhoeffer, S., Stadler, T., 2013. Using an epidemiological model for phylogenetic inference reveals density dependence in hiv transmission. Molecular biology and evolution 31, 6–17.

Maddison, W.P., Midford, P.E., Otto, S.P., 2007. Estimating a binary character’s effect on speciation and extinction. Systematic Biology 56, 701–710. doi: 10.1080/10635150701607033

Morlon, H., Parsons, T.L., Plotkin, J.B., 2011. Reconciling molecular phylogenies with the fossil record. P. Natl. Acad. Sci. USA 108, 16327–16332.

Nee, S., May, R.M., Harvey, P.H., 1994. The reconstructed evolutionary process. Philosophical Transactions of the Royal Society of London B: Biological Sciences 344, 305–311.

Quental, T.B., Marshall, C.R., 2010. Diversity dynamics: molecular phylogenies need the fossil record. Trends in Ecology & Evolution 25, 434–441.

Ratmann, O., Hodcroft, E.B., Pickles, M., Cori, A., Hall, M., Lycett, S., Didelot, X., 2016. Phylogenetic tools for generalized hiv-1 epidemics: findings from the pangea-hiv methods comparison. Molecular biology and evolution 34, 185–203.

Stadler, T., 2010. Sampling-through-time in birth–death trees. Journal of theoretical biology 267, 396–404.

Stadler, T., Bonhoeffer, S., 2013. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. Philosophical Transactions of the Royal Society of London B: Biological Sciences 368, 20120198.

Stadler, T., Kühnert, D., Bonhoeffer, S., Drummond, A.J., 2013. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proceedings of the National Academy of Sciences 110, 228–233.

Starrfelt, J., Liow, L.H., 2016. How many dinosaur species were there? fossil bias and true richness estimated using a poisson sampling model. Philosophical Transactions Of The Royal Society B-Biological Sciences 371, 20150219. doi: 10.1098/rstb.2015.0219.

Vaughan, T.G., Leventhal, G.E., Rasmussen, D.A., Drummond, A.J., Welch, D., Stadler, T., 2019. Estimating epidemic incidence and prevalence from genomic data. Molecular Biology and Evolution doi: 10.1093/molbev/msz106

Yule, G.U., 1925. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Phil. Trans. R. Soc. Lond. B .

### Appendix A. Solving well-known ODEs

#### Appendix A.1. The extinction probability

We first deal with equation (2.1) governing $u_t$, and start by studying the polynomial $\lambda x^2 - \gamma x + \mu$. This polynomial has discriminant $\Delta = \gamma^2 - 4\mu > (\lambda + \mu)^2 - 4\lambda \mu \geq (\lambda - \mu)^2 > 0$. Roots are

$$ x_1 = \frac{\gamma - \sqrt{\Delta}}{2\lambda} \quad \text{and} \quad x_2 = \frac{\gamma + \sqrt{\Delta}}{2\lambda} $$

Moreover, we know that both roots are positive because $\Delta < \gamma^2 \Rightarrow \sqrt{\Delta} < \gamma \Rightarrow x_1 > 0$.

On an interval including zero and where the polynomial remains positive (as $(-\infty, x_1)$ for example), we can write,

$$ \frac{du}{\lambda u^2 - \gamma u + \mu} = dt $$
as not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

\[ \frac{du}{(x_1 - u)(x_2 - u)} = \lambda dt \]

\[ \frac{1}{x_2 - x_1} \left( \frac{1}{x_1 - u} - \frac{1}{x_2 - u} \right) du = \lambda dt \]

\[ \left( \frac{1}{x_1 - u} - \frac{1}{x_2 - u} \right) du = \sqrt{\Delta} dt \]

Integrating both sides between time 0 and \( t \), we get

\[ \frac{x_2 - u_t}{x_1 - u_t} = \frac{x_2 - z}{x_1 - z} e^{\sqrt{\Delta} t} \]

\[ \iff x_2(x_1 - z)e^{-\sqrt{\Delta} t} - u_t(x_1 - z)e^{-\sqrt{\Delta} t} = x_2(x_1 - z) - u_t(x_2 - z) \]

\[ \iff u_t \left( (x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta} t} \right) = x_2(x_1 - z) - x_2(x_1 - z)e^{-\sqrt{\Delta} t} \]

\[ \iff u_t = \frac{x_1(x_2 - z) - x_2(x_1 - z)e^{-\sqrt{\Delta} t}}{(x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta} t}} \]

This is the result stated in equation \( (2.2) \). Note that this quantity is called \( p_0(t) \) in Stadler (2010), or \( E(t) \) in Maddison et al. (2007).

\textbf{Appendix A.2. Probability to leave only one sampled descendant}

We aim here to integrate a slight variation of equation \( (2.3) \) governing \( p_t \) when \( k = 1 \). The equation we are interested in is,

\[ \frac{dW_s}{ds} = (2\lambda u(s, z) - \gamma)kW_s \quad (A.1) \]

\[ \frac{dW_s}{W_s} = \left( \frac{2\lambda k x_1(x_2 - z) - x_2(x_1 - z)e^{-\sqrt{\Delta} s}}{(x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta} s}} - \gamma k \right) ds \]

\[ = \left( \frac{2\lambda k x_1}{\sqrt{\Delta}} \frac{\sqrt{\Delta}(x_2 - z)e^{\Delta t} - (x_1 - z)}{(x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta} t}} - \frac{2\lambda k x_2}{\sqrt{\Delta}} \frac{(x_1 - z)\sqrt{\Delta}e^{-\sqrt{\Delta} t} - (x_1 - z)}{(x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta} t}} - \gamma k \right) ds \]

All these three terms can be integrated visually between some time \( t_h \) and \( t \), leading to,

\[ \ln \frac{W_t}{W_{t_h}} = \frac{2\lambda k x_1}{\sqrt{\Delta}} \ln \left( \frac{(x_2 - z)e^{\Delta t} - (x_1 - z)}{(x_2 - z) - (x_1 - z)e^{-\Delta t}} \right)_{t_h}^{t} - \frac{2\lambda k x_2}{\sqrt{\Delta}} \ln \left( \frac{(x_2 - z) - (x_1 - z)e^{-\Delta t}}{(x_2 - z) - (x_1 - z)e^{-\Delta t}} \right)_{t_h}^{t} - \gamma k(t - t_h) \]

\[ = \frac{2\lambda k x_1}{\sqrt{\Delta}} \ln \left( \frac{(x_2 - z)e^{\Delta t} - (x_1 - z)}{(x_2 - z) - (x_1 - z)e^{-\Delta t}} \right)_{t_h}^{t} + \frac{2\lambda k x_2}{\sqrt{\Delta}} \ln \left( \frac{(x_2 - z) - (x_1 - z)e^{-\Delta t}}{(x_2 - z) - (x_1 - z)e^{-\Delta t}} \right)_{t_h}^{t} + \gamma k(t - t_h) \]

\[ = -\frac{2\lambda k(x_2 - x_1)}{\sqrt{\Delta}} \ln \left( \frac{(x_2 - z) - (x_1 - z)e^{-\Delta t}}{(x_2 - z) - (x_1 - z)e^{-\Delta t}} \right)_{t_h}^{t} + 2\lambda k x_1(t - t_h) \]

\[ = -2k \ln \left( \frac{(x_2 - z) - (x_1 - z)e^{-\Delta t}}{(x_2 - z) - (x_1 - z)e^{-\Delta t}} \right)_{t_h}^{t} + k\sqrt{\Delta}(t - t_h) \]

Leading to the final expression below

\[ W_t = W_{t_h} \left( \frac{(x_2 - z) - (x_1 - z)e^{-\Delta t}}{(x_2 - z) - (x_1 - z)e^{-\Delta t}} \right)^{-2k} e^{-k\sqrt{\Delta}(t - t_h)} \quad (A.2) \]
Note that the case \( k = 1, \ t_h = 0 \) and \( W_0 = 1 - z \) corresponds to the probability \( p_t \) given as equation (2.4),

\[
p(t, z) = (1 - z) \frac{\Delta}{\lambda^2} \left( (x_2 - z) - (x_1 - z)e^{-\sqrt{\lambda}t} \right)^{-2} e^{-\sqrt{\lambda}t}
\]

while the general case can be expressed using function \( p \) as

\[
W_t = W_{th} \left( \frac{p(t, z)}{p(t_h, z)} \right)^k
\]

**Appendix A.3. A few useful properties**

Solutions \( u(t, z) \) and \( p(t, z) \) to ODEs (2.1) and (2.3) satisfy two properties relying on the semi-group property of solutions of ODEs, namely,

\[
\begin{align*}
  u(t_2 - t_1, u(t_1, z)) &= u(t_2, z) \\
p(t_2 - t_1, u(t_1, z)) &= \frac{p(t_2, z) p(0, u(t_1, z))}{p(t_1, z)} = \frac{p(t_2, z)(1 - u(t_1, z))}{p(t_1, z)}
\end{align*}
\]

These two properties are useful in many calculations throughout this document, e.g.

- Solving the main PDE in [Appendix B](#) requires inverting \( u \), using the first property with,
  \[
  z = u(t - t_h, z_0)
  \]
  \[
  \implies u(t - t_h, z) = u(t - t_h, u(t - t_h, z_0))
  \]
  \[
  \implies u(t - t_h, z) = u(0, z_0)
  \]
  \[
  \implies z_0 = u(t - t_h, z)
  \]

- The same Appendix section requires also composing function \( p \) and \( u \), using
  \[
  \frac{p(t - t_h, u(t_h - t, z))}{p(0, u(t_h - t, z))} = \frac{p(0, z)}{p(t_h - t, z)}
  \]

- In the proof of Proposition 4.3, we switch to the notation \( R(t, z) = p(t, z)/(1 - z) \) and again compose \( R \) and \( u \) in the same way,
  \[
  \begin{align*}
  \frac{p(t_j - t, u(t_h - t, z))}{p(0, u(t_h - t, z))} &= \frac{p(t_j - t, z)}{p(t_h - t, z)} \\
  \frac{p(t_j - t, u(t_h - t, z))}{1 - u(t_h - t, z)} &= \frac{p(t_j - t, z)}{p(t_h - t, z)} \\
  \implies R(t_j - t_h, u(t_h - t, z)) &= \frac{R(t_j - t, z)}{R(t_h - t, z)}
  \end{align*}
  \]
Appendix B. Solving the main PDE

We aim now at finding an analytical solution for the following PDE, driving the evolution of $\hat{M}$ across a given epoch $(t_{h-1}, t_{h})$, on which the number of observed lineages remains constant and equal to $k$.

\[
\hat{M}(t_{h}, z) = F(z)
\]
\[
\partial_t \hat{M} + (\lambda z^2 - \gamma z + \mu) \partial_z \hat{M} + k(2\lambda z - \gamma) \hat{M} = 0
\]

Appendix B.1. Principle of the method of characteristics

This problem can be solved by the method of characteristics. We suppose that we can write $\hat{M}(t, z) = \hat{M}(t(s), z(s))$ where functions $t$ and $z$ satisfy the ODEs,

\[
\frac{dz}{ds} = \lambda z^2 - \gamma z + \mu
\]
\[
\frac{dt}{ds} = 1
\]

This way, the function $g(s) = \hat{M}(t(s), z(s))$ satisfies another ODE, that we will have to solve,

\[
\frac{dg}{ds} = \frac{dz}{ds} \partial_z \hat{M} + \frac{dt}{ds} \partial_t \hat{M} = (\lambda z^2 - \gamma z + \mu) \partial_z \hat{M} + \partial_t \hat{M} = -k(2\lambda z - \gamma) \hat{M}
\]
\[
\Leftrightarrow \frac{dg}{ds} + k(2\lambda z - \gamma) g = 0
\]

Appendix B.2. Step 1, solve for $t(s), z(s)$ and $g(s)$

We start by integrating $t(s)$. We moreover fix that $t(0) = t_{h}$, thus leading to $t(s) = t_{h} + s$.

We now turn to $z$, and remark that it satisfies previously studied ODE \([2.1]\). Integrating between 0 and $s$ leads to,

\[
z(s) = u(s, z_0) = \frac{x_1(x_2 - z_0) - x_2(x_1 - z_0)e^{-\sqrt{\Delta t}}}{(x_2 - z_0) - (x_1 - z_0)e^{-\sqrt{\Delta t}}}
\]

Last, $g$ satisfies an ODE very similar to \([A.1]\). Taking care of the minus sign, it leads to the following result,

\[
g_s = g_0 \left( \frac{1}{R(s, z_0)} \right)^k \tag{B.1}
\]

Appendix B.3. Step 2, express $\hat{M}$ back as a function of $t, z$

We want to express our two unknown quantities $s$ and $z_0$ as functions of $t$ and $z$. 


On a first hand, we get easily \( s = t - t_h \). We moreover can solve for \( z_0 \) in the following equation, remembering the semi-group property of \( u \),

\[
z = u(t - t_h, z_0) \iff z_0 = u(t_h - t, z)
\]

Substituting these into the previous expression \([B.1]\) of \( g_s \) then leads to,

\[
\tilde{M}(t, z) = F(u(t_h - t, z)) R(t - t_h, u(t_h - t, z))^{-k} \\
= F(u(t_h - t, z)) R(t_h - t, z)^k
\]

where the first to second equality relies on a property exposed in Appendix A.3. This gives us the final formula which is stated in Proposition 4.1.

**Appendix C. Some useful algebra**

This section of the Appendix pools together all bits of algebra that are not really digestible, but are used in the main text.

**Appendix C.1. Derivative of \( \tilde{M} \)**

We first modify a bit the expression of the generating function,

\[
\tilde{M}(t_h - t, z) = R(t_h - t, z)^k \sum_{i=0}^{N} \tilde{M}_i(t_h) u(t_h - t, z)^i
\]

\[
= \left( \frac{\Delta}{\lambda^2} e^{-\sqrt{\lambda t}} \right)^k \sum_{i=0}^{N} \tilde{M}_i(t_h) \left( x_1(x_2 - z) - x_2(x_1 - z) e^{-\sqrt{\lambda t}} \right)^i \left( (x_2 - z) - (x_1 - z) e^{-\sqrt{\lambda t}} \right)^{(2k+i)}
\]

Applying Leibniz’s derivation rule to the product, we get,

\[
\left( \partial_z^i \tilde{M}(t_h - t, z) \right)_{z=0} = \left( \frac{\Delta}{\lambda^2} e^{-\sqrt{\lambda t}} \right)^k \sum_{i=0}^{N} \tilde{M}_i(t_h) \sum_{\alpha=0}^{i} \binom{i}{\alpha} \left( \partial_z^\alpha \left( x_1(x_2 - z) - x_2(x_1 - z) e^{-\sqrt{\lambda t}} \right) \right)^i |_{z=0} \left( \partial_z^{i-\alpha} \left( (x_2 - z) - (x_1 - z) e^{-\sqrt{\lambda t}} \right)^{(2k+i)} \right)_{z=0}
\]

(C.1)

The first of the two derivatives in the sum can be computed as,

\[
\partial_z \left( x_1(x_2 - z) - x_2(x_1 - z) e^{-\sqrt{\lambda t}} \right)^l = l \left( -x_1 + x_2 e^{-\sqrt{\lambda t}} \right) \left( x_1(x_2 - z) - x_2(x_1 - z) e^{-\sqrt{\lambda t}} \right)^{l-1} 1_{l \geq 1}
\]

\[
\partial_z^2 \left( x_1(x_2 - z) - x_2(x_1 - z) e^{-\sqrt{\lambda t}} \right)^l = l(l-1) \left( -x_1 + x_2 e^{-\sqrt{\lambda t}} \right)^2 \left( x_1(x_2 - z) - x_2(x_1 - z) e^{-\sqrt{\lambda t}} \right)^{l-2} 1_{l \geq 2}
\]

\vdots

33
\[ \partial_z^\alpha \left( x_1(x_2 - z) - x_2(x_1 - z)e^{-\sqrt{z\lambda}} \right)^t = \frac{t!}{(t - \alpha)!} \left( -x_1 + x_2 e^{-\sqrt{z\lambda}} \right)^\alpha \left( x_1(x_2 - z) - x_2(x_1 - z)e^{-\sqrt{z\lambda}} \right)^{t-\alpha} \mathbb{I}_{t \geq \alpha} \]

While the second gives us,
\[ \partial_z \left( \left( x_2 - z \right) - \left( x_1 - z \right)e^{-\sqrt{z\lambda}} \right)^{(2k+l)} = (2k+l) \left( 1 - e^{-\sqrt{z\lambda}} \right) \left( x_2 - z \right) - (x_1 - z)e^{-\sqrt{z\lambda}} \]
\[ \partial_z^2 \left( \left( x_2 - z \right) - \left( x_1 - z \right)e^{-\sqrt{z\lambda}} \right)^{(2k+l)} = (2k+l)(2k+l+1) \left( 1 - e^{-\sqrt{z\lambda}} \right)^2 \left( x_2 - z \right) - (x_1 - z)e^{-\sqrt{z\lambda}} \]
\[ \vdots \]
\[ \partial_z^{i-\alpha} \left( \left( x_2 - z \right) - \left( x_1 - z \right)e^{-\sqrt{z\lambda}} \right)^{(2k+l)} = \left( \prod_{m=0}^{i-\alpha-1} (2k+l+m) \right) \left( 1 - e^{-\sqrt{z\lambda}} \right)^{i-\alpha} \left( x_2 - z \right) - (x_1 - z)e^{-\sqrt{z\lambda}} \]

Applying these derivatives in \( z = 0 \) in equation (C.1) yields,
\[ \left( \partial_z^\alpha \tilde{M}(t_h - t, z) \right)_{z=0} = \left( \frac{\lambda^\alpha}{\lambda^2} e^{-\sqrt{\lambda t}} \right)^k \sum_{i=0}^N \sum_{\alpha=0}^i \tilde{M}_{t_h}^{(i)} \left( \frac{t!}{(t - \alpha)!} \prod_{m=0}^{i-\alpha-1} (2k+l+m) \right) \left( -x_1 + x_2 e^{-\sqrt{z\lambda}} \right)^\alpha \left( x_1 x_2 \right)^{t-\alpha} \left( 1 - e^{-\sqrt{z\lambda}} \right)^{i+\alpha-2\alpha} \left( x_2 - x_1 e^{-\sqrt{z\lambda}} \right)^{(2k+l+i+\alpha)} \]

Which is the expression provided in Proposition (4.2).

Appendix C.2. Derivatives of \( \tilde{M} \) when \( \omega = 0 \)

We wish here to derive the \( \partial_z^i \tilde{M}(t, z) \) where function \( \tilde{M} \) is as given in Proposition (4.3), i.e.
\[ \tilde{M}(t, z) = \lambda^\alpha e^{i\psi + \psi'(1 - r)^+ e^\nu R(t_0, t - z) \prod_{t_j \in X} R(t_j - t, z) \prod_{t_j \in W} R(t_j - t, z)^{-1} \prod_{t_j \in Y} u(t_j - t, z) R(t_j - t, z)^{-1} \]

We take for simplicity the derivative of the logarithm of \( \tilde{M} \) and express the derivatives of \( \tilde{M} \) using these and Leibniz’s formula,
\[ \partial_z \tilde{M} = \tilde{M} \partial_z (\ln \tilde{M}) \]
\[ \partial_z^2 \tilde{M} = \partial_z \tilde{M} \partial_z (\ln \tilde{M}) + \tilde{M} \partial_z^2 (\ln \tilde{M}) \]
\[ \partial_z^3 \tilde{M} = \partial_z^2 \tilde{M} \partial_z (\ln \tilde{M}) + 2 \partial_z \tilde{M} \partial_z^2 (\ln \tilde{M}) + \tilde{M} \partial_z^3 (\ln \tilde{M}) \]
\[ \vdots \]
\[ \partial_z^i \tilde{M} = \sum_{\alpha=0}^i \left( \frac{i-\alpha}{\alpha-1} \right) \left( \partial_z^{i-\alpha} \tilde{M} \right) \left( \partial_z^\alpha (\ln \tilde{M}) \right) \]

(C.2)
In order to compute the derivatives of \( \ln \hat{M} \), one needs to get the derivatives of \( \ln R(t, z) \) and \( \ln u(t, z) \).

We have

\[
\ln R(t, z) = -2 \ln \left( (x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta t}} \right) + \ln \frac{\Lambda}{x^2} - \sqrt{\Delta t}
\]

\[
\partial_z \ln R(t, z) = 2 \left( 1 - e^{-\sqrt{\Delta t}} \right) \left( (x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta t}} \right)^{-1}
\]

\[
\partial_z^2 \ln R(t, z) = 2 \left( 1 - e^{-\sqrt{\Delta t}} \right)^2 \left( (x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta t}} \right)^{-2}
\]

\[
\partial_z^3 \ln R(t, z) = 4 \left( 1 - e^{-\sqrt{\Delta t}} \right)^3 \left( (x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta t}} \right)^{-3}
\]

\[
\vdots
\]

\[
\partial_z^a \ln R(t, z) = 2(\alpha - 1)! \left( 1 - e^{-\sqrt{\Delta t}} \right)^\alpha \left( (x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta t}} \right)^{-\alpha}
\]

Finally taking the function in \( z = 0 \) leads to

\[
\partial_z^a \ln R(t, 0) = 2(\alpha - 1)!a_t^a \tag{C.3}
\]

where we defined \( a_t := \left( 1 - e^{-\sqrt{\Delta t}} \right) (x_2 - x_1 e^{-\sqrt{\Delta t}})^{-1} \tag{C.4} \)

In the same way we get,

\[
\ln u(t, z) = \ln \left( (x_1 (x_2 - z) - x_2 (x_1 - z)e^{-\sqrt{\Delta t}} \right) - \ln \left( (x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta t}} \right)
\]

\[
\partial_z \ln u(t, z) = - \left( x_1 - x_2 e^{-\sqrt{\Delta t}} \right) \left( x_1 (x_2 - z) - x_2 (x_1 - z)e^{-\sqrt{\Delta t}} \right)^{-1}
\]

\[
+ \left( 1 - e^{-\sqrt{\Delta t}} \right) \left( (x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta t}} \right)^{-1}
\]

\[
\partial_z^2 \ln u(t, z) = - \left( x_1 - x_2 e^{-\sqrt{\Delta t}} \right)^2 \left( x_1 (x_2 - z) - x_2 (x_1 - z)e^{-\sqrt{\Delta t}} \right)^{-2}
\]

\[
+ \left( 1 - e^{-\sqrt{\Delta t}} \right)^2 \left( (x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta t}} \right)^{-2}
\]

\[
\partial_z^3 \ln u(t, z) = -2 \left( x_1 - x_2 e^{-\sqrt{\Delta t}} \right)^3 \left( x_1 (x_2 - z) - x_2 (x_1 - z)e^{-\sqrt{\Delta t}} \right)^{-3}
\]

\[
+ 2 \left( 1 - e^{-\sqrt{\Delta t}} \right)^3 \left( (x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta t}} \right)^{-3}
\]

\[
\vdots
\]

\[
\partial_z^a \ln u(t, z) = (\alpha - 1)! \left[ - \left( x_1 - x_2 e^{-\sqrt{\Delta t}} \right)^\alpha \left( x_1 (x_2 - z) - x_2 (x_1 - z)e^{-\sqrt{\Delta t}} \right)^{-\alpha}
\]

\[
+ \left( 1 - e^{-\sqrt{\Delta t}} \right)^\alpha \left( (x_2 - z) - (x_1 - z)e^{-\sqrt{\Delta t}} \right)^{-\alpha} \right]
\]

Here also, we are interested in the function in \( z = 0 \),

\[
\partial_z^a \ln u(t, 0) = (\alpha - 1)! (a_t^a - b_t^a) \tag{C.5}
\]

where we defined \( b_t := \left( x_1 - x_2 e^{-\sqrt{\Delta t}} \right) (x_1 x_2 - x_2 x_1 e^{-\sqrt{\Delta t}})^{-1} \tag{C.6} \)
Last ingredient needed to write the derivative of $\ln \hat{M}$, we get,

\[
(\frac{\partial^2}{\partial z^2} \ln (u(t, z) R(t, z)^{-1}))_{z=0} = \left( \frac{\partial^2}{\partial z^2} \ln u(t, z) \right)_{z=0} - \left( \frac{\partial^2}{\partial z^2} \ln R(t, z) \right)_{z=0} = - (\alpha - 1)! (a^\alpha_0 + b^\alpha_0)
\]

where we defined $C^{(\alpha)} := 2a^\alpha_{t, w-t} + 2 \sum_{t_j \in X} a^\alpha_{t_j - t} - 2 \sum_{t_j \in Y} a^\alpha_{t_j - t} - \sum_{t_j \in W} (a^\alpha_{t_j - t} + b^\alpha_{t_j - t})$

Finally, using equations (C.4) and (C.7) one can compute

\[
\left( \frac{\partial^2}{\partial z^2} \ln \hat{M}(t, z) \right)_{z=0} = (\alpha - 1)! C^{(\alpha)}
\]

Plugging this into equation (C.2) and noting that $\left( \frac{\partial^2}{\partial z^2} \ln \hat{M}(t, z) \right)_{z=0} = i! M^{(i)}_t$, we get

\[
M^{(i)}_t = \sum_{\alpha=1}^{i} \left( i - 1 \right) \frac{(i - \alpha)! (\alpha - 1)!}{\alpha!} M^{(i-\alpha)}_t C^{(\alpha)} = \frac{1}{i} \sum_{\alpha=1}^{i} M^{(i-\alpha)}_t C^{(\alpha)}
\]

which is the result stated in corollary 4.4.

Appendix D. Inductions across the epochs

Appendix D.1. Proof of Proposition 3.1

We prove the proposition by induction across the epochs.

If we observe only the first epoch and the $k_0$ leaves at present, then we get at any time $t$ across the first epoch $(0, t_1)$, $L^{(i)}_t = (1 - \rho)^{k_0} p_0^i u^i_0 t^{k_0}$, which satisfies Proposition 3.1

Suppose we observed so far – i.e. on $(0, t_{h+1}) - v$ sampled ancestors, $w$ removed leaves at times $t_j \in W$, $x$ branching events at times $t_j \in X$, $y$ non-removed leaves at times $t_j \in Y$. And suppose that Proposition 3.1 is verified across epoch $(t_h, t_{h+1})$. Let us have a look at what happen across epoch $(t_{h+1}, t_{h+2})$.

The observed punctual event $t_{h+1}$ can either be,

1. a removed ancestral leaf. Update (3.18) then applies. Subsequently, the number of sampled lineages increases by one and formula (3.17) applies on the next epoch, leading to

\[
L^{(i)}_t = u^i_0 W_t
\]

where

\[
W_t = \sum_{t_j \in X} t^{k_0} \prod_{t_j \in Y} u^{k_0}_t \prod_{t_j \in W} p_t^{-1} \prod_{t_j \in W \cup \{t_{h+1}\}} p_t^{-1}
\]

\[
= \sum_{t_j \in X} t^{k_0} \prod_{t_j \in Y} u^{k_0}_t \prod_{t_j \in W} p_t^{-1} \prod_{t_j \in W \cup \{t_{h+1}\}} p_t^{-1}
\]

\[36\]
2. a non-removed ancestral leaf. Update (3.19) then applies. Subsequently, the number of sampled lineages increases by one and formula (3.17) applies on the next epoch, leading to

\[ L_t^{(i)} = u_i^t W_t \]

where \( W_t = \lambda^x \psi^{v+w+(y+1)}(1-r)^{(v+1)+y} \frac{p_t}{p_{h+1}} p_t^{k+1} \prod_{t_j \in X} p_{t_j} \prod_{t_j \in Y} u_{t_j} p_{t_j}^{-1} \prod_{t_j \in W} p_{t_j}^{-1} \]

3. a non-removed sampled ancestor along a branch. Update (3.20) then applies. The number of sampled lineages increases by one and formula (3.17) applies on the next epoch, leading to

\[ L_t^{(i)} = u_i^t W_t \]

where \( W_t = \lambda^x \psi^{v+w+(y+1)}(1-r)^{(v+1)+y} \frac{p_t}{p_{h+1}} p_t^{k+1} \prod_{t_j \in X} p_{t_j} \prod_{t_j \in Y} u_{t_j} p_{t_j}^{-1} \prod_{t_j \in W} p_{t_j}^{-1} \]

4. a branching event between two sampled lineages. Update (3.21) then applies. The number of sampled lineages decreases by one, and formula (3.17) applies on the next epoch, leading to

\[ L_t^{(i)} = u_i^t W_t \]

where \( W_t = \lambda^{x+1} \psi^{v+w+y} (1-r)^{(v+1)+y} \frac{p_t}{p_{h+1}} \prod_{t_j \in X} p_{t_j} \prod_{t_j \in Y} u_{t_j} p_{t_j}^{-1} \prod_{t_j \in W} p_{t_j}^{-1} \]

In all four cases, Proposition 3.1 is satisfied across epoch \((t_{h+1}, t_{h+2})\).

Appendix D.2. Proof of Proposition 4.3

This Proposition is also proven by induction across the epochs.

We start at \( t_{or} = t_n \) with \( k = 1 \) lineage. Across epoch \((t_{n-1}, t_n)\), applying Proposition 4.1 with \( F(z) = 1 \) and \( k = 1 \), we get \( \tilde{M}(t, z) = R(t_{or} - t, z) \), which verifies Proposition 4.3.

Suppose now that Proposition 4.3 is verified across epoch \((t_h, t_{h+1})\) and that we observed, on \((t_h, t_{or})\), \( v \) sampled ancestors, \( w \) removed leaves at times \( t_j \in W \), \( x \) branching events at times \( t_j \in X \), \( y \) non-removed leaves at times \( t_j \in Y \). Let us have a look at what happens on \((t_{h-1}, t_h)\).

Punctual event \( t_h \) can either be,
1. a removed leaf. The number of sampled lineages then goes from \(1 + x - y - w\) to \(x - y - w\), and applying update (4.34) followed by Proposition 4.1 leads to

\[
\tilde{M}(t, z) = \lambda^x \psi^{v+(w+1)+y} \mu^{w+1} (1 - r)^{v+y} R(t_{or} - t_h, u(t_h - t, z)) R(t_h - t, z)^{x-y-w} \prod_{t_j \in \mathcal{X}} R(t_j - t_h, u(t_h - t, z))
\]

\[
\prod_{t_j \in \mathcal{W}} R(t_j - t_h, u(t_h - t, z))^{-1} \prod_{t_j \in \mathcal{Y}} u(t_j - t_h, u(t_h - t, z)) R(t_j - t_h, u(t_h - t, z))^{-1}
\]

\[
= \lambda^x \psi^{v+(w+1)+y} \mu^{w+1} (1 - r)^{v+y} R(t_{or} - t_h, z) R(t_h - t, z)^{x-y-w} \prod_{t_j \in \mathcal{X}} R(t_j - t_h, z) \prod_{t_j \in \mathcal{W}} R(t_j - t_h, z) \prod_{t_j \in \mathcal{Y}} u(t_j - t_h, z) R(t_j - t_h, z)
\]

\[
= \lambda^x \psi^{v+(w+1)+y} \mu^{w+1} (1 - r)^{v+y} R(t_{or} - t_h, z)
\]

\[
\prod_{t_j \in \mathcal{X}} R(t_j - t_h, z) \prod_{t_j \in \mathcal{W}} R(t_j - t_h, z)^{-1} \prod_{t_j \in \mathcal{Y}} u(t_j - t_h, z) R(t_j - t_h, z)^{-1}
\]

where the first to second equality is detailed in Appendix A and the second to third comes after canceling out the \(R(t_h - t, z)\).

2. a non-removed leaf. The number of sampled lineages then goes from \(1 + x - y - w\) to \(x - y - w\), and applying update (4.35) followed by Proposition 4.1 leads to

\[
\tilde{M}(t, z) = \lambda^x \psi^{v+w+(y+1)+w} (1 - r)^{v+(y+1)} R(t_{or} - t_h, z) R(t_h - t, z)^{x-y-w} u(t_h - t, z)
\]

\[
\prod_{t_j \in \mathcal{X}} R(t_j - t_h, z) \prod_{t_j \in \mathcal{W}} R(t_h - t, z) \prod_{t_j \in \mathcal{Y}} u(t_j - t_h, z) R(t_j - t_h, z)
\]

\[
= \lambda^x \psi^{v+w+(y+1)+w} (1 - r)^{v+(y+1)} R(t_{or} - t_h, z)
\]

\[
\prod_{t_j \in \mathcal{X}} R(t_j - t_h, z) \prod_{t_j \in \mathcal{W}} R(t_j - t_h, z)^{-1} \prod_{t_j \in \mathcal{Y}} u(t_j - t_h, z) R(t_j - t_h, z)^{-1}
\]

3. a sampled ancestor. The number of sampled lineages then remains unchanged and equal to \(1 + x - y - w\).

Applying update (4.36) followed by Proposition 4.1 leads to

\[
\tilde{M}(t, z) = \lambda^x \psi^{(v+1)+v+w+y+1+y}(1 - r)^{(v+1)+y} R(t_{or} - t_h, z) R(t_h - t, z)^{x-y-w}
\]

\[
\prod_{t_j \in \mathcal{X}} R(t_j - t_h, z) \prod_{t_j \in \mathcal{W}} R(t_h - t, z) \prod_{t_j \in \mathcal{Y}} u(t_j - t_h, z) R(t_j - t_h, z)
\]

\[
= \lambda^x \psi^{(v+1)+v+w+y+1+y} R(t_{or} - t_h, z)
\]

\[
\prod_{t_j \in \mathcal{X}} R(t_j - t_h, z) \prod_{t_j \in \mathcal{W}} R(t_j - t_h, z)^{-1} \prod_{t_j \in \mathcal{Y}} u(t_j - t_h, z) R(t_j - t_h, z)^{-1}
\]

4. a branching time. The number of sampled lineages then goes from \(1 + x - y - w\) to \(2 + x - y - w\), and applying update (4.39) followed by Proposition 4.1 leads to

\[
\tilde{M}(t, z) = \lambda^{x+1} \psi^{v+w+y+1+y} (1 - r)^{v+y} R(t_{or} - t_h, z) R(t_h - t, z)^{2+x-y-w}
\]
\[
\prod_{j \in X} R(t_j - t, z) \prod_{j \in W} R(t_h - t, z) \prod_{j \in Y} u(t_j - t, z) R(t_h - t, z) = \lambda^{z+1} \frac{(1-r)^{w+y} R(t_\text{or} - t, z)}{1-z} \prod_{j \in X \cup \{t_h\}} R(t_j - t, z) \prod_{j \in Y} u(t_j - t, z) R(t_j - t, z)^{-1}
\]

In all these cases, Proposition 4.3 is verified across epoch \((t_{h-1}, t_h)\), which ends the proof.

**Appendix E. Using a probability generating function to solve for \(L_t\) as well**

**Appendix E.1. A slightly different strategy**

Recall that \(L_t\) verifies the following Master equation,

\[
\dot{L}_t^{(i)} = -\gamma(i + k)L_t^{(i)} + \lambda(2k + i)L_t^{(i+1)} + \mu i L_t^{(i-1)}
\]

\[
L_0^{(i)} = \rho_0^k (1 - \rho_0)^i
\]

If we introduce the corresponding probability generating function,

\[
\hat{L}(t, z) = \sum_{i=0}^{\infty} z^i L_t^{(i)}
\]

Then the initial condition on \(L\) translates in \(\hat{L}\) as,

\[
\hat{L}(0, z) = \sum_{i=0}^{\infty} (z(1 - \rho))^i \rho_0^k = \frac{\rho_0^k}{1 - z(1 - \rho)} , \forall z \in \left(\frac{\pm 1}{1 - \rho}\right)
\]

The ODE translates into a PDE, but not as nicely as for \(M_t\), see below,

\[
\partial_t \hat{L} = \sum_{i=0}^{\infty} z^i \left( -\gamma(i + k)L_t^{(i)} + \lambda(2k + i)L_t^{(i+1)} + \mu i L_t^{(i-1)} \right)
\]

\[
= -\gamma k \sum_{i=0}^{\infty} z^i L_t^{(i)} - \gamma \sum_{i=1}^{\infty} iz^i L_t^{(i)} + \lambda \sum_{i=1}^{\infty} z^{i-1} (2k + i - 1) L_t^{(i)} + \mu \sum_{i=0}^{\infty} (i + 1) z^{i+1} L_t^{(i)}
\]

\[
= -\gamma k \hat{L} - \gamma z \partial_z \hat{L} + (2k - 1) \lambda \frac{1}{z} (\hat{L} - L_t^{(0)}) + \lambda \partial_z \hat{L} + \mu z^2 \partial_z \hat{L} + \mu z \hat{L}
\]

\[
= \left( -\gamma k + (2k - 1) \lambda \frac{1}{z} + \mu z \right) \hat{L} + (\mu z^2 - \gamma z + \lambda) \partial_z \hat{L} - (2k - 1) \lambda \frac{1}{z} \hat{L}(t, 0)
\]

We are thus left with the following PDE problem,

\[
\hat{L}(0, z) = \frac{\rho_0^k}{1 - z(1 - \rho)}
\]

\[
- z \partial_t \hat{L} + (\mu z^3 - \gamma z^2 + \lambda z) \partial_z \hat{L} + (\mu z^2 - \gamma k z + (2k - 1) \lambda) \hat{L} - (2k - 1) \lambda \frac{1}{z} \hat{L}(t, 0) = 0 \quad (E.1)
\]

This remaining term with \(\hat{L}(t, 0)\) complicates things a little bit. However, the initial condition on \(\hat{L}\) provides us with a first candidate function to satisfy this PDE.
Appendix E.2. Solution

We introduce below function $f$, and show that it satisfies the PDE problem (E.1).

$$f(t,z) := \frac{p^k_t}{1 - zu_t}$$

First, we observe that it satisfies the initial condition. We then need to check that it satisfies the PDE, and to do so we expand each of the four components of equation (E.1).

The first one gives us,

$$-z\partial_t f = -z k p_t p_t^{-1} (1 - zu_t) + z u_t p_t^k$$

$$= -z k (2\lambda u_t - \gamma)(1 - zu_t) + (\lambda u_t^2 - \gamma u_t + \mu) z p_t^k$$

$$= \lambda (-2k zu_t - (2k - 1)zu_t^2) + \gamma (kz - (k - 1)zu_t) - \mu z^2 p_t^k$$

we then turn to the second component,

$$(\mu z^3 - \gamma z^2 + \lambda z)\partial_z f = \frac{\lambda z u_t - \gamma z^2 u_t + \mu z^3 u_t}{(1 - zu_t)^2} p_t^k$$

the third one,

$$(\mu z^2 - \gamma k z + (2k - 1)\lambda)f = \frac{(1 - zu_t)(\mu z^2 - \gamma k z + (2k - 1)\lambda)}{(1 - zu_t)^2} p_t^k$$

$$= \lambda ((2k - 1) - (2k - 1)zu_t) + \gamma (-kz + kz^2 u_t) + \mu (z^2 - z^3 u_t) p_t^k$$

and the fourth and final one,

$$-(2k - 1)\lambda f(t,0) = -\lambda (2k - 1)(1 - zu_t)^2 p_t^k$$

$$= \frac{\lambda (-2k - 1)z^2 u_t^2 + 2(2k - 1)zu_t - (2k - 1))}{(1 - zu_t)^2} p_t^k$$

Putting everything together, we can now check that indeed,

$$-z\partial_t f + (\mu z^3 - \gamma z^2 + \lambda z)\partial_z f + (\mu z^2 - \gamma k z + (2k - 1)\lambda)f - (2k - 1)f(t,0) = 0$$

While the branching and $\psi$-sampling with removal updates do not change anything to this solution, all the others do. Further work is thus needed to look for other solutions to this same PDE with different initial conditions.