# Notes

David Helekal

June 17, 2020

# 1 Simulation

## 1.1 Coalescent Preliminaries

The coalescent is a CTMC defined on the set $\{1...n\}$, parametrised via the coalescent rate, in our case $1/Ne(t)g$, where $g$ is a scale parameter and $Ne(t)$ the population size at time $t$. The transition rates of the coalescent process are given by

$$\rho(j, j-1) = \binom{j}{2} \cdot \frac{1}{Ne(t)g}$$

The waiting times in the homogenous case are exponentially distributed

$$P[W_j \leq s] = 1 - \exp\left(-s\frac{\binom{j}{2}}{Ne(t)g}\right)$$

Furthermore, the waiting times for individual coalescent events, conditioned on being less than the time between two consecutive sampling events $\Delta t$ are distributed as follows

$$P[W_j \leq s \mid W_j \leq \Delta t] = \frac{P[W_j \leq s]}{P[W_j \leq \Delta t]} \quad \forall s \leq \Delta t \tag{1}$$

In the inhomogenous case, the waiting times can be derived as follows: For an inhomogenous CTMC, let $E_j(t)$ be the total exit rate from state $j$ at time $t$. By the markov property individual exit events from a given state only depend on the state and given time, i.e. they form a time-inhomogenous poisson process. As such the probability of no events in an interval $[t, t+s] \quad s \in \mathbb{R}^+$ is

$$\exp\left(-\int_t^{t+s} E_j(\tau)d\tau\right) = \exp\left(-\int_0^s E_j(t+\tau)d\tau\right) \tag{2}$$

The waiting times are defined as

$$W_j(t) = \inf\{s : X(t+s) \neq j \mid X(t) = j\} \tag{3}$$

As such

$$W_j(t) > s \Rightarrow \forall \tau \in [t, t+s] X(\tau) = j \tag{4}$$

Furthermore the above relation doesn't hold iff an exit event has occured in the time interval $[t, t+s]$. As such:

$$P[W_j(t) > s] = P[\text{no exit events in } [t, t+s]] = \exp\left(-\int_0^s E_j(t+\tau)d\tau\right)$$

$$P[W_j(t) < s] = 1 - \exp\left(-\int_0^s E_j(t+\tau)d\tau\right)$$

In the case of phylodynamic coalescent this becomes

$$P[W_j(t) \leq s] = 1 - \exp\left(-\int_0^s \frac{\binom{j}{2}}{Ne(t+\tau)g}d\tau\right) \tag{5}$$

Note, the waiting times a re still memoryless:

$$P\left[W_j(t) > s + u \mid W_j(t) > s\right] = P\left[W_j(t) > s + u \mid X(s) = j\right] \tag{6}$$

By markov property

$$P\left[W_j(t) > s + u \mid X(s) = j\right] = P\left[W_j(t + s) > u\right] \tag{7}$$

## 1.2 Homogenous case

The sampling process conditioned on sampling times follows a modified gillespie scheme. In order to facilitate the computation of the likelihoods of the individual simulated trees, it is preferred to avoid rejection sampling. As such we require sampling the conditional likelihood 1. This is achieved by inverse transform sampling. Let:

$$\begin{aligned} u \sim & U([0,1]) \\ T(u) : P[T(u) \leq s] = & \frac{P[T(u) \leq s]}{P[T(u) \leq \Delta t]} \quad \forall s \leq \Delta t \end{aligned}$$

Where $T(u)$ is assumed to be monotone increasing and invertible.

$$\begin{aligned} & P[T(u) \leq s] = P[u \leq T^{-}1(s)] \\ \Rightarrow \quad & P[u \leq T^{-1}(s)] = \frac{\int_0^s \lambda \exp(-\lambda t) dt}{\int_0^{\Delta t} \lambda \exp(-\lambda t) dt} \\ \Rightarrow \quad & T^{-1}(s) = \frac{1 - \exp(-\lambda s)}{1 - \exp(-\lambda \Delta t)} \end{aligned}$$

Defining $y \triangleq T^{-1}(s)$, we obtain the transform:

$$T(y) = \frac{-1}{\lambda} \log[1 - y(1 - \exp(-\lambda \Delta t))] \tag{8}$$

The corresponding pdf evaluated at $u$ is

$$f_{\mathbf{T}(u)}(T(u)) = \lambda \left( \frac{1}{1 - \exp(\lambda \Delta t)} - u \right) \tag{9}$$

```
f <- (sampling_times, Ne): //Sampling times in descending order
  extant_lineages <- 1
  future_lineages <- length(sampling_times)-1
  t <- sampling_times[1]
  idx <- 1

  while extant_lineages > 1 or future_lineages > 0:
    if extant_lineages < 2:
```

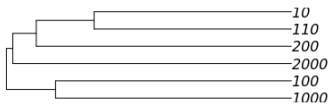


Figure 1: An example simulated coalescent tree

```
    idx++
    t <- sampling_times[idx]
    extant_lineages++
    future_lineages--
  else:
    delta_t <- t-sampling_times[idx+1]
    rate <- binom(extant_lineages,2)/Ne
    w_t ~ exp(rate)

    if w_t < delta_t:
      coalesce_lineages
      extant_lineages--
      t <- t+w_t
    else:
      idx++
      t <- sampling_times[idx]
      extant_lineages++
      future_lineages--
```

### 1.3 Inhomogenous Case

In the inhomogenous case, the scheme is similar, with the key difference that the sampling times now follow a much more complex distribution. As such a sampling scheme such as rejection sampling will be required (?)

### 1.4 Multistrain+Inhomogenous Case

In this case, coalescent nodes have an added colour property, and each colour coalesces according to a colour specific, time dependent case. Nodes of non-identical colour can coalesce iff at least one of them is the last remaining node of a given colour.
Given $M$ colours, $M$ population size functions $\{Ne_j(t)\}_{1\leq j\leq M}$, and initial population size $N$, Let $Y(t)$ be a CTMC with the state space:

$$S = \left\{\mathbf{s} \in \mathbb{Z}_+^N : |\mathbf{s}| \leq N, |\mathbf{s}| \geq 1\right\} \tag{10}$$

and the transition rates

$$\mathbf{s} \to \mathbf{s} - \mathbf{e_j} \quad \binom{s_j}{2} N e_j(t) + \delta_{s_j,1} N e_j(t) \sum_{i \neq j} s_i \tag{11}$$