An exploration of genealogical coalescent models using phase-type theory

Tayla Broadbridge

Supervisor: Oscar Peralta

June 4, 2021

Thesis submitted for the degree of
Honours degree of the Bachelor of Mathematical Sciences
in
Applied Mathematics
at The University of Adelaide
Faculty of Engineering, Computer and Mathematical Sciences
School of Mathematical Sciences

THE UNIVERSITY
of ADELAIDE

Contents

Declaration

Except where stated this thesis is, to the best of my knowledge, my own work and my supervisor has approved its submission.

Signed by student: Tayla Broadbridge

Date: June 4, 2021

Signed by supervisor:

Date:

Acknowledgements

I would like to thank my supervisor, Dr. Oscar Peralta, for his constant encouragement and mentorship throughout this project. I will always be grateful for the opportunity to have worked alongside him. Thank you also to my family and friends who have supported me through the ups and downs of this year, and especially to my mum for making me a coffee every morning.

Abstract

Phase-type theory has great potential as a tool for studying the evolution of genealogies. From a coalescent theory perspective, we observe such genealogies backwards-in-time and model their ancestral process using a continuous-time Markov chain. With the application of a phase-type framework, we are able to concisely describe such coalescent models, while maintaining their underlying dynamics. The compact matrix notation of phase-type theory allows us to calculate important descriptors of genealogical quantities including tree height and total branch length with the use of simple matrix operations. We firstly demonstrate the tractability of phase-type distributions with a comprehensive self-contained explanation. This is followed by exploring the application of phase-type theory to a range of coalescent models. Phase-type representations are able to be extended upon, which we demonstrate by applying multivariate phase-type theory in a study of a two-island model. The phase-type representation for this two-island model is developed and subsequently used to compute previously mentioned genealogical quantities. In addition, we determine the correlation of the branch lengths for each island. This allows us to study the dynamics between these two island populations, which we find to be more sensitive to a change in migration rates in contrast to the coalescent rates.

Chapter 1

Introduction

The ancestry of a population can be modelled using coalescent theory, an area of population genetics which has quite recently surged to become a prominent area of research. Population genetics models were first developed throughout the 1920s by R.A. Fisher [8], J.B.S. Haldane [10] and Sewall Wright [28], and involved the study of population *evolution.* Coalescent theory was later discovered independently by researchers in the 1980s including John Kingman [15], and instead aims to describe the *ancestral* process of such populations. Observing such genealogical processes backwards-in-time provides a more computationally efficient framework for data simulation [16], and so we ask the question, how may a group of genes sampled from a population have originated from a common ancestor? The application of coalescent theory is in its early stages, but is expected to grow significantly in the near future, particularly within the area of human evolution and modelling biological phenomena [11]. Theoreticians, however, face challenges in modelling and improving algorithms that simulate such genealogical models. Theoretical advancements are needed in order to deal with the increasing interest of such problems.

To make advancements to coalescent theory, in this thesis we make use of phase-type theory which involves matrix manipulations, whereby typically complicated and tedious calculations are avoided. Phase type distributions can be traced back to researchers A.K. Erlang (1909) [7] and A. Jensen (1954) [14], and throughout the 1970s by M.F. Neuts (1975) [20]. Throughout its earlier stages, phase-type theory was used as a tool to study queuing theory, and then risk theory [2]. More recently, phase-type theory has also been applied to problems within finance [3]. Researchers including A. Hobolth, A. Siri-Jégousse and M. Bladt (2018) [13] have most recently discovered that key population genetics quantities are phase-type distributed, a finding that is extremely useful since properties of such phase-type distributions are very well understood. Such key quantities include the height and total branch length of a genealogical tree in the basic coalescent model. Tree height has the interpretation that it represents the time until a group of genes have found their most recent common ancestor. Total branch length gives us an idea of how much history a group of genes share. Study in to such quantities can be greatly useful in understanding the origin of many diseases and illnesses which are attributed to genetics [22], for example. Descriptors for these important phase-type distributed quantities are expressed in closed form, and as a result, are able to be computed almost instantaneously and without the requirement of simulations.

In this work we study simple descriptors including means, higher order moments and cross moments for the tree height and total branch length of some well known coalescent models such as the Kingman's, ψ, and β-coalescent. A correction has been made to the paper by Hobolth et. al. [13], where we recognise an error in the computation for the second moment of tree height for the ψ and β-coalescent. Together with phase-type theory and population genetics, we then study the two-island model. This model considers populations that are subdivided into two discrete finite populations, between which some migration occurs. A generalisation of this population model was developed by Sewall Wright (1943) [29], and has been studied since, by researchers including T. Nagylaki (1979) [19] and B. Rannala (1997) [23]. Most recently, the demographic history of genes has been researched using the island model by A. Arredondo (2021) [1]. To demonstrate the tractability of phase-type distributions, we look at studying the ancestry of the two-island model. Dependence between the two islands can also be investigated using multivariate phase-type theory, where we compute correlation between the total branch lengths for each island. We once again demonstrate the advantage of phase-type distributions by avoiding unnecessary simulations.

The structure of this thesis is as follows:

Chapter 2 introduces us to continuous-time Markov chains (CTMCs), which are stochastic processes whereby the holding times in each state are exponentially distributed and the process changes state according to specified probabilities. We also introduce phase-type distributed random variables, along with some examples. The structure of such random variables is explained along with definitions and derivations of important descriptors. These include the density function and Laplace transform. Finally, we discuss multivariate phase-type distributions with an emphasis on the bivariate case, which is of importance in our Chapter 5 extension.

In Chapter 3 we discuss the Wright-Fisher model and its development into the continuous-time coalescent, and investigate the dynamics of the Kingman's coalescent, which is the most straightforward continuous-time coalescent. We detail the meaning and interpretation of *tree height* and *total branch length*, which are two important quantities within population genetics. Comparisons are made between a simulated Kingman's coalescent to theoretical results derived from coalescent theory, and we also look in to further generalisations including the Λ, ψ and β coalescent. Finally, we look at the seed-bank coalescent; a more complex coalescent process whereby the underlying genealogical process can contain dormant genes.

In Chapter 4 it is demonstrated how phase-type theory can be used to compute descriptors for tree height and total branch length, as based on the work of Hobolth et. al. (2019) [13]. Results for the Kingman's coalescent and further generalisations of coalescent models seen within this paper are verified and elaborated upon using code from Sections A.2, A.3 and A.4 respectively. A correction has also been made to the paper, whereby we provide the correct computation of second order moments for the ψ and β coalescent models, which is demonstrated in Sections 4.3 and 4.4.

In Chapter 5, we introduce a new phase-type representation for a coalescent model with two-island structure, whereby migration occurs as well as coalescent events. As a result, we are able to investigate the tree heights and total branch lengths for this island case. With multivariate phase-type theory, the dependence between the branch lengths of these two islands is computed. We vary the rates of migration and coalescence to draw comparisons between genealogical structure, and determine which rates have a larger impact on such quantities. These new developments allow room for many extensions.

Within Chapter 6 we outline a three potential extensions that can follow from the phase-type representation of the two-island model, as well as summarise the results of the manuscript.

Chapter 2

Background

Chapter 2 introduces us to continuous-time Markov chains (CTMCs), which are continuous-time stochastic processes whereby the holding times in each state are exponentially distributed and the process changes state according to specified probabilities. With this underlying CTMC structure, we characterise phase-type distributions. We provide insight into the properties of phase-type distributions with examples, and derive expressions for their density functions, Laplace transforms and higher order moments. The tractability of phase-type distributions is demonstrated in order to facilitate its application in Chapters 4 and 5. Finally, we discuss multivariate phase-type distributions with an emphasis on the bivariate case, which is of importance for our Chapter 5 extension, as well as further applications of phase-type theory.

2.1 Continuous-time Markov Chains

A cádlág stochastic process, $\{X_t\}_{t\geq 0}$ with discrete state space, $\mathcal{S}$, is called a *continuous-time Markov chain* (CTMC) if, $\forall\ i, j, i_0, ..., i_n \in \mathcal{S},\ s, t \geq 0$ and
$0 \leq s_0 \leq s_1 \leq ... \leq s_n \leq s$,

$$P(X_{t+s} = j | X_s = i, X_{s_n} = i_n, ..., X_{s_1} = i, X_{s_0} = i_0) = P(X_{t+s} = j | X_s = i).$$

This means that the probability of landing in any future states only depends on the current state of the process. That is, given the present state, the rest of the past is irrelevant for predicting the future.

A CTMC is said to be *time-homogeneous* if it satisfies

$$P(X_{t+s} = j \mid X_s = i) = P(X_t = j \mid X_0 = i) \quad \forall s, t \geq 0,\ \forall i, j \in \mathcal{S}.$$

Heuristically, for a time-homogeneous CTMC, the probability of transitioning from state i at time s to state j at time $s + t$ is only dependent on the difference in times, t. We denote this probability by

$$p_{ij}^{t} = P(X_t = j \mid X_0 = i) \quad \forall t \geq 0,\ \forall i, j \in \mathcal{S}, \tag{2.1}$$

which we call the *transition probability.* Throughout this manuscript we are concerned only with the time-homogeneous case, and so in future sections we drop the term 'time-homogeneous' for brevity. In the following we study some properties of $\{p_{ij}^t\}$.

2.1.1 Transition probabilities

The transition probability $\{p_{ij}^t\}$ satisfies the *Chapman-Kolmogorov equation*

$$p_{ij}^{s+t} = \sum_{k \in \mathcal{S}} p_{ik}^{s} p_{kj}^{t} \quad \forall s, t \geq 0,\ \forall i, j \in \mathcal{S}. \tag{2.2}$$

This is true since for the chain to transition from state i to j in time $s + t$, it must be in some state k at time s. More precisely,

$$\begin{aligned} p_{ij}^{s+t} &= P(X_{s+t} = j | X_0 = i) \\ &= \sum_{k \in \mathcal{S}} P(X_{s+t} = j, X_s = k | X_0 = i) \\ &= \sum_{k \in \mathcal{S}} P(X_{s+t} = j | X_s = k, X_0 = i) P(X_s = k | X_0 = i) \\ &= \sum_{k \in \mathcal{S}} P(X_{s+t} = j | X_s = k) P(X_s = k | X_0 = i) \quad \text{by the Markov property} \\ &= \sum_{k \in \mathcal{S}} p_{kj}^{t} p_{ik}^{s}. \end{aligned}$$

As we will see later, continuous-time Markov chains are described by the *transition (jump) rate matrix*, $Q = \{q_{ij}\}_{i,j \in \mathcal{S}}$, where

$$\begin{array}{lll} q_{ij} = \lim_{h \to 0} \frac{p_{ij}^h}{h} & \text{for} & j \neq i \in \mathcal{S}, \text{ and} \\ q_{ii} = -\sum_{j \neq i} q_{ij} & \text{for} & i \in \mathcal{S}. \end{array}$$

For $i \neq j$, the quantity q_{ij} represents the transition rate from state i to state j, while $-q_{ii}$ represents the total exit rate from i. For consistency with the literature, we will from now on suppose that $\lambda_i = -q_{ii}$. We assume that $\lim_{h \to 0} p_{ij}^h / h$ exists and is finite, so that all the transition rates q_{ij}, $i \neq j$, are well-defined. Also note that the way these diagonal elements are defined ensures that the row sums of Q are equal to 0. The transition probabilities can be expressed in terms of the transition rate matrix, Q, by utilising the Kolmogorov forward and backward equations, which we will now investigate. To simplify computations, from now on we suppose that the state space $\mathcal{S}$ is finite.

Theorem 2.1.1 (Kolmogorov's differential equations). *If $\{p_{ij}\}$ represents the transition probability of the Markov process $\{X_t\}_{t\geq 0}$, then Kolmogorov's differential equations*

$$\frac{\mathrm{d}p_{ij}}{\mathrm{d}t} = \sum_{k\neq i} q_{ik}p^t_{kj} - \lambda_i p^t_{ij} \quad \forall t \geq 0,\ i,j \in \mathcal{S} \tag{2.3}$$

$$\frac{\mathrm{d}p_{ij}}{\mathrm{d}t} = \sum_{k\neq i} p^t_{ik}q_{kj} - p^t_{ij}\lambda_j \quad \forall t \geq 0,\ i,j \in \mathcal{S} \tag{2.4}$$

hold, where (2.3) is referred to as Kolmogorov's backward equation, and (2.4) is Kolmogorov's forward equation.

Proof. Fix $t \geq 0$ and $i,j \in \mathcal{S}$. By the Chapman-Kolmogorov equation,

$$\begin{aligned} p^{t+h}_{ij} &= \sum_{k\in\mathcal{S}} p^t_{kj}p^h_{ik}, \\ \implies p^{t+h}_{ij} - p^t_{ij} &= \sum_{k\in\mathcal{S}} p^t_{kj}p^h_{ik} - p^t_{ij} \\ &= \sum_{k\neq i} p^t_{kj}p^h_{ik} - (1-p^h_{ii})p^t_{ij}. \end{aligned}$$

Recall by our definition of a jump rate for $k \neq i \in \mathcal{S}$,

$$\begin{aligned} q_{ik} &= \lim_{h\to 0}\frac{p^h_{ik}}{h}, \\ \implies q_{ik}p^t_{kj} &= \lim_{h\to 0}\frac{p^h_{ik}p^t_{kj}}{h}, \\ \implies \sum_{k\neq i} q_{ik}p^t_{kj} &= \sum_{k\neq i}\lim_{h\to 0}\frac{p^h_{ik}p^t_{kj}}{h}, \\ &= \lim_{h\to 0}\sum_{k\neq i}\frac{p^h_{ik}p^t_{kj}}{h}, \end{aligned}$$

where the sum and the limit may be interchanged in this case because we are considering a finite state space. Given that

$$[1-p^h_{ii}] = \sum_{k\neq i} p^h_{ik},$$

then

$$\begin{aligned}\lim_{h\to 0}\frac{1}{h}[1-p_{ii}^h] &= \lim_{h\to 0}\frac{1}{h}\sum_{k\neq i} p_{ik}^h\\ &= \lim_{h\to 0}\frac{1}{h}\sum_{k\neq i} p_{ik}^h\\ &= \sum_{k\neq i} q_{ik}\\ &= \lambda_i,\end{aligned}$$

so that

$$\lim_{h\to 0}\frac{1-p_{ii}^h}{h}p_{ij}^t = \lambda_i p_{ij}^t.$$

Putting these results together with the definition of the derivative we have

$$\begin{aligned}\frac{\mathrm{d}p_{ij}^t}{\mathrm{d}t} &= \lim_{h\to 0}\frac{1}{h}(p_{ij}^{t+h}-p_{ij}^t)\\ &= \lim_{h\to 0}\frac{1}{h}\left(\sum_{k\neq i} p_{ik}^h p_{kj}^t - [1-p_{ii}^h]p_{ij}^t\right)\\ &= \lim_{h\to 0}\frac{1}{h}\left(\sum_{k\neq i} p_{ik}^h p_{kj}^t\right) - \lim_{h\to 0}\frac{1}{h}[1-p_{ii}^h]p_{ij}^t\\ &= \sum_{k\neq i} q_{ik}p_{kj}^t - \lambda_i p_{ij}^t,\end{aligned}$$

representing Kolmogorov's backward equation (2.3).

Similarly, to derive Kolmogorov's forward equation (2.4),

$$\begin{aligned}p_{ij}^{t+h}-p_{ij}^t &= \sum_k p_{ik}^t p_{kj}^h - p_{ij}^t\\ &= \sum_{k\neq j} p_{ik}^t p_{kj}^h - [p_{jj}^h - 1]p_{ij}^t,\end{aligned}$$

and with analogous steps as before, we get

$$\frac{\mathrm{d}p_{ij}^t}{\mathrm{d}t} = \sum_{k\neq j} p_{ik}^t q_{kj} - p_{ij}^t\lambda_j.$$

□

For $t \geq 0$, define $P_t = \{p_{ij}^t\}_{i,j \in \mathcal{S}}$, the matrix representation of our transition probability $\{p_{ij}^t\}$. We can therefore express Kolmogorov's forward equation (2.4) as

$$\frac{\mathrm{d}P_t}{\mathrm{d}t} = QP_t, \quad t \geq 0, \tag{2.5}$$

and Kolmogorov's backward equation (2.3) as

$$\frac{\mathrm{d}P_t}{\mathrm{d}t} = P_t Q, \quad t \geq 0. \tag{2.6}$$

The forward and backward equations with initial condition $P_0 = I$ have a unique solution of the form $P_t = e^{Qt} = \sum_{n=0}^{\infty} (Qt)^n/n!$ (see [Corollary 1.3.11, [4]]).

2.1.2 Path behaviour of CTMCs

Throughout the manuscript, S_n will denote the time of the n^{th} transition in the CTMC $\{X_t\}_{t\geq 0}$ where $S_0 = 0$, and

$$S_n = \inf\{t > S_{n-1} : X_t \neq X_{S_{n-1}}\} \quad \forall n \geq 1. \tag{2.7}$$

Also suppose that T_n denotes the time between the $(n-1)^{th}$ and the n^{th} jump in the CTMC $\{X_t\}_{t\geq 0}$, that is,

$$T_n = S_n - S_{n-1} \quad \forall n \geq 1.$$

For $i \in \mathcal{S}$, recall that $\lambda_i = -q_{ii}$ the exit rate from state i. It can be shown that given $\{X_{S_n}\}_{n\geq 0}$, the holding times $\{T_n\}_{n\geq 1}$ are conditionally independent. Furthermore, if $X_{S_n} = i$, then T_n is exponentially distributed with rate λ_i, and $\{X_{S_{n+1}} = j\}$ with probability $m_{ij} := q_{ij}/\lambda_i$. The proof of this follows from the Markov property (see [Corollary 1.3.6, [4]]).

Given a transition rate matrix Q and initial state $j_0 \in \mathcal{S}$, a CTMC can be constructed by following the algorithm:

Algorithm 1 Constructing the CTMC

```
0: Set $X_0 = j_0$

1:    a: The time until the first jump (time spent in state $X_0$) is $T_1 \sim \text{Exp}(\lambda_{j_0})$
      b: The next state $X_1 = j_1$ is chosen with probability $m_{j_0,j_1}$

2:    a: The time between the first and second jump, $S_2 - S_1$ is $T_2 \sim \text{Exp}(\lambda_{j_1})$
      b: The next state $X_2 = j_2$ is chosen with probability $m_{j_1,j_2}$
   ⋮

n:    a: The time between the $n^{th}$ and the $(n+1)^{th}$ jump, $S_{n+1} - S_n$ in $T_{n+1} \sim \text{Exp}(\lambda_{j_n})$
      b: The next state $X_{n+1} = j_{n+1}$ is chosen with probability $m_{j_n,j_{n+1}}$.
```

An example simulation of Algorithm 2 is shown in Figure 2.1.

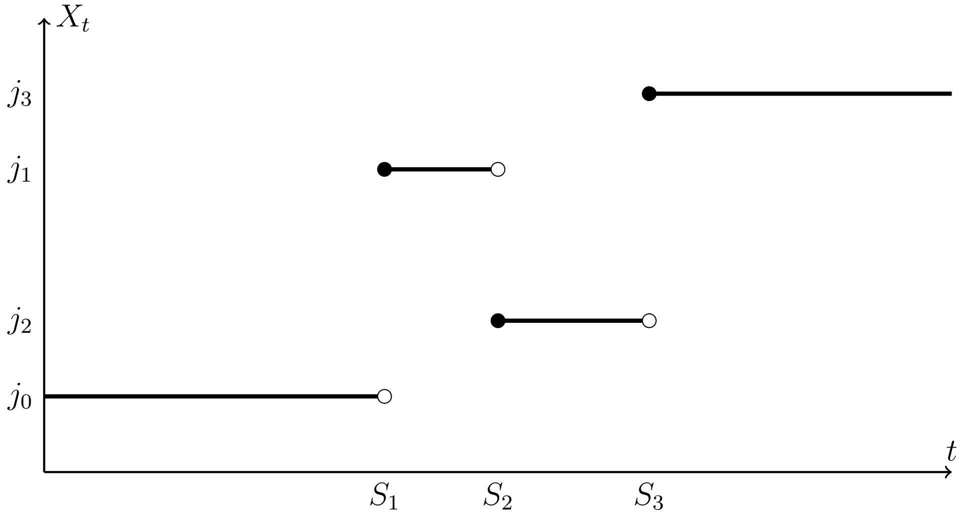

Figure 2.1: A sample construction of the CTMC $\{X_t\}_{t\geq 0}$ by Algorithm 2. The Markov chain is initialised in state j_0, so that $X_0 = j_0$. It then transitions to state j_1, where it remains for time $S_2 - S_1$. It subsequently transitions to states j_2 and finally j_3.

Since the holding time between the $(n-1)^{th}$ and n^{th} jump, $S_n - S_{n-1}$, is exponentially distributed with parameter λ_i given that $X_{S_{n-1}} = i$, the conditional probability that there will be a jump in the process $\{X_t\}_{t\geq 0}$ during the interval $[t, t+h]$ is approximately $\lambda_i h$.

To see this, recall the definition of the off-diagonal elements of matrix Q,

$$q_{ij} = \lim_{h \to 0} \frac{p_{ij}^h}{h},$$

from which we derive

$$\begin{aligned}\lambda_i = -q_{ii} &= \sum_{j \neq i} \lim_{h \to 0} \frac{P(X_h = j \mid X_0 = i)}{h} \\ &= \lim_{h \to 0} \sum_{j \neq i} \frac{P(X_h = j \mid X_0 = i)}{h} \\ &= \lim_{h \to 0} \frac{P(\text{There is a jump in the interval } [0,h] \mid X_0 = i)}{h}.\end{aligned}$$

Thus

$$P(\text{There is a transition in } [0,h] \mid X_0 = i) = \lambda_i h + o(h),$$

where $o(h)$ denotes any function $f(h)$ such that $f(h)/h \to 0 \quad \text{as} \quad h \to 0$. Using the standard infinitesimal convention with $o(\mathrm{d}t) = 0$, the above can therefore be expressed as

$$P(\text{There is a transition in } [0,\mathrm{d}t] \mid X_0 = i) = \lambda_i \mathrm{d}t.$$

Similarly, we can argue that $P(X_{t+\mathrm{d}t} = j \mid X_t = i) = q_{ij}\mathrm{d}t$ for $i \neq j$, which is a common interpretation of the jump intensities $\{q_{ij}\}$.

2.1.3 Classification of states

In the following we present a classification of states for the CTMC $\{X_t\}_{t \geq 0}$ which is of importance in following sections.

Let $N_i = \sum_{j=1}^{\infty} \mathbb{1}\{X_j = i\}$ be the total number of visits to state i. The state i is said to be *transient* if $N_i < \infty$ almost surely, and *recurrent* otherwise. An important subclass of recurrent states is that of *absorbing* states, which are states of the CTMC $\{X_t\}_{t \geq 0}$ that once entered, cannot be left. This absorption property translates to a 0 intensity jump rate. Indeed, suppose $X(S_n) = i$ and $\lambda_i = 0$. That is, the CTMC $\{X_t\}_{t \geq 0}$ is in state i at the time of the n^{th} transition where $\lambda_i = 0$ implies that the time between the n^{th} and $(n+1)^{th}$ transition is $T_{n+1} \sim \text{Exp}(\lambda_i) = \text{Exp}(0)$, or in other words, $T_{n+1} = +\infty$ and no further jumps occur. Note that when the diagonals of the matrix Q have entry 0, the entire row must be populated by zeros, since $q_{ii} = -\sum_{j \neq i} q_{ij}$ for $i \in \mathcal{S}$. In the forthcoming section we will be particularly interested in CTMCs that consist of transient states and a single absorbing state.

2.2 Phase-type distributions

Phase-type distributions are in essence, a matrix-based generalisation of the exponential distribution. The distribution can be represented by a random variable describing the time until absorption of a Markov process with one absorbing state, where every other state represents one transient phase. Phase-type distributions offer a particularly useful property in that important descriptors can be explicitly expressed in terms of matrix notation. Furthermore, phase-type distributions are dense in the class of distributions on $\mathbb{R}_+$, meaning every distribution with support in $[0, \infty)$ may be approximated arbitrarily closely by a phase-type distribution (see Section 3.2.1, [4]).

Definition 2.2.1. *Let $\{X_t\}_{t\geq 0}$ be a continuous time Markov chain with finite state space $\mathcal{S} = \{1, 2, ..., p, p{+}1\}$ where states $1, 2, ..., p$ are transient and state $p{+}1$ is absorbing. In block-matrix notation, the intensity matrix has the form*

$$Q = \begin{pmatrix} A & \mathbf{a} \\ \mathbf{0} & 0 \end{pmatrix}, \tag{2.8}$$

where the sub-intensity matrix A represents transitions between transient states $\{1, 2, ..., p\}$ and the column vector $\mathbf{a}$ represents the transitions from $\{1, 2, \ldots, p\}$ into the absorbing state $p + 1$. The time until absorption

$$\tau = \{\inf t \geq 0 : X_t = p + 1\}$$

is said to have a phase-type distribution. *This is denoted $\tau \sim PH(\boldsymbol{\pi}, A)$ where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$ is the initial distribution of $\{X_t\}_{t\geq 0}$ over states $\{1, 2, \ldots, p\}$ with $\sum_i^p \pi_i = 1$.*

Below we include some common examples of phase-type distributions.

Exponential distribution

The exponential distribution with parameter λ is the simplest example of a phase-type distribution. For such a case we let $A = -\lambda$ and $\pi = 1$. Then the jump time from 1 to 2 is $\text{Exp}(\lambda)$ distributed.

Erlang distribution

Let $X_1, X_2, ..., X_k$ be independent, identically exponentially distributed random variables, $X_i \sim \text{Exp}(\lambda_i)$ for $i = 1, 2, ..., k$. Then $Z = \sum_{i=1}^k X_i$ has an Erlang distribution characterised by the density

$$f(z) = \begin{cases} \frac{\lambda^k z^{k-1} e^{-\lambda z}}{(k-1)!} & k = 1, 2, ..., \ z \geq 0 \\ 0 & z < 0 \end{cases}$$

where $k > 0$ is a shape parameter and $\lambda > 0$ is the rate parameter.

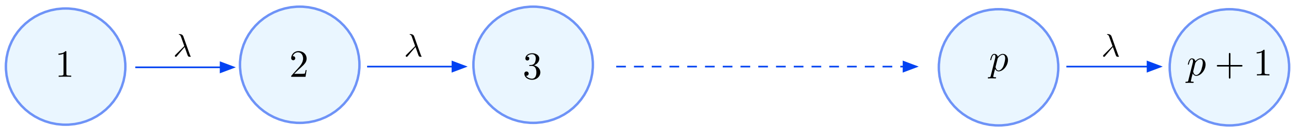

Figure 2.2: A state space representation for the Erlang distribution. In this case, the process only transitions forward until reaching absorbing state $p+1$.

The phase-type representation of the Erlang distribution is $\boldsymbol{\pi} = (1, 0, ..., 0)$ and

$$A = \begin{pmatrix} -\lambda & \lambda & & & \\ & -\lambda & \lambda & & \\ & & \ddots & \ddots & \\ & & & & \lambda \\ & & & & -\lambda \end{pmatrix},$$

with $\boldsymbol{a} = (0, 0, \dots, \lambda)^\intercal$ which contains the transition into the absorbing state $p+1$. See Figure 2.2 for a representation.

Hyperexponential distribution

Let $X_1, X_2, ..., X_p$ be independent, $X_i \sim \text{Exp}(\lambda_i)$ and f_i denote the exponential density with parameter λ_i. Let $f = \sum_{i=1}^{n} \pi_i f_i$ where $\pi_i > 0$ and $\sum_{i=1}^{n} \pi_i = 1$. The state space for the Hyperexponential distribution is represented in Figure 2.3 below.

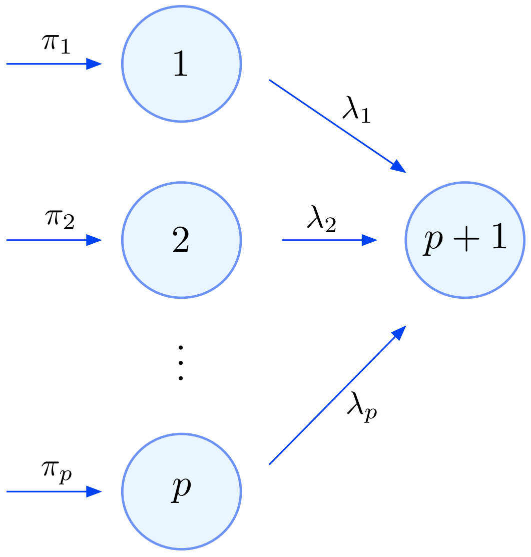

Figure 2.3: A state space representation of the Hyperexponential distribution. In this case, all states communicate and the process can transition into the absorbing state from any state.

Then f corresponds to the density of $\mathrm{PH}(\boldsymbol{\pi}, A)$, with $\pi = (\pi_1, \pi_2, \ldots, \pi_p)$,

$$A = \begin{pmatrix} -\lambda_1 & & & \\ & -\lambda_2 & & \\ & & \ddots & \\ & & & -\lambda_p \end{pmatrix}$$

and $\boldsymbol{a} = (\lambda_1, \lambda_2, \ldots, \lambda_p)^\intercal$.

Coxian distribution

Coxian distributions arise from the convolution of exponential distributions and are a generalisation of the Erlang distribution. The difference to Erlang is that in the case of the Coxian distribution, the chain can reach the absorbing state through any of the $1, 2, ..., p-1$ states, not just state p. The state space diagram is shown in Figure 2.4 below.

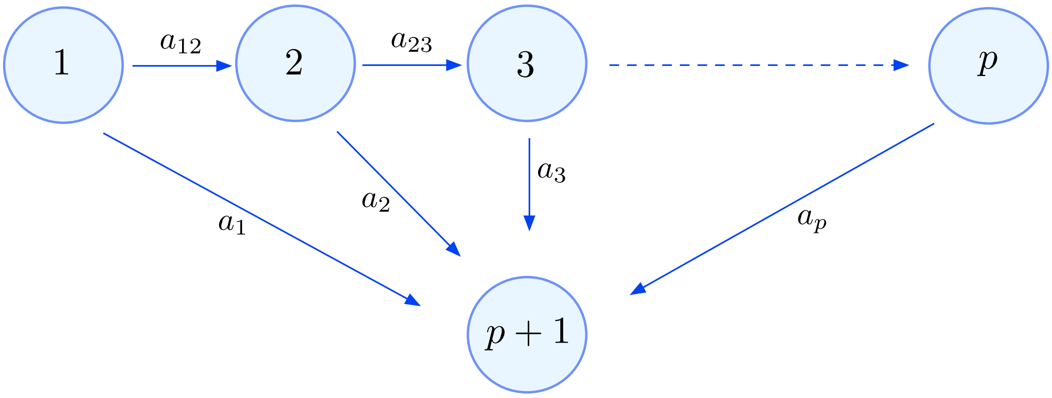

Figure 2.4: State transition diagram for a Coxian distribution. Denote $\lambda_i = a_{i,i+1} + a_i$ for $i = 1, 2, \ldots, p-1$ and $\lambda_p = a_p$. Here, $a_{i,i+1}$ is the transition rate from state i to state $i+1$, and a_i is the transition rate from state i to absorbing state $p+1$.

Thus, a Coxian distribution has phase-type representation with $\pi = (1, 0, \ldots, 0)$, and

$$A = \begin{pmatrix} -\lambda_1 & a_{12} & & & \\ & -\lambda_2 & a_{23} & & \\ & & \ddots & \ddots & \\ & & & & a_{p-1,p} \\ & & & & -\lambda_p \end{pmatrix}.$$

Here, we have that $\boldsymbol{a} = (\lambda_1 - a_{12}, \lambda_2 - a_{23}, \ldots, \lambda_p)^\intercal$.

2.2.1 Properties of phase-type distributions

From the definition of the matrix-exponential and the fact that $\boldsymbol{a} = -A\mathbf{1}$ where $\mathbf{1} = (1, 1, \dots, 1)^\intercal$, the exponential of the matrix (2.8) is given by

$$e^{Qs} = \begin{pmatrix} e^{As} & \mathbf{1} - e^{As}\mathbf{1} \\ \mathbf{0} & 1 \end{pmatrix}. \tag{2.9}$$

Indeed, it is straightforward to verify that

$$Q^n = \begin{pmatrix} A^n & A^{n-1}\boldsymbol{a} \\ \mathbf{0} & 0 \end{pmatrix}.$$

so by definition of the matrix exponential,

$$\begin{aligned} e^{Qs} &= \sum_{n=0}^{\infty} \frac{(Qs)^n}{n!} \\ &= \begin{pmatrix} \sum_{n=0}^{\infty} \frac{(As)^n}{n!} & \sum_{n=0}^{\infty} \frac{(As)^{n-1}}{n!}\boldsymbol{a} \\ \mathbf{0} & 1 \end{pmatrix} \\ &= \begin{pmatrix} e^{As} & \mathbf{1} - e^{As}\mathbf{1} \\ \mathbf{0} & 1 \end{pmatrix}, \end{aligned}$$

Equation (2.9) allows us to compute some descriptors of PH$(\boldsymbol{\pi}, A)$ via probabilistic arguments as follows.

Theorem 2.2.2. *If $\tau \sim$ PH$(\boldsymbol{\pi}, A)$, the density f of τ is given by $f(s) = \boldsymbol{\pi} e^{As}\boldsymbol{a}$.*

Proof. Recall that e^{Qt} is the transition matrix P^t of the Markov process $\{X_t\}_{t\geq 0}$. The restriction of P^s to the transient states $\{1, 2, ..., p\}$ is given by e^{As}, as seen in Equation 2.9. Therefore,

$$\begin{aligned} p_{ij}^s &= P(X_s = j \mid X_0 = i) \\ &= (e^{As})_{ij} \quad \text{for} \quad i, j = 1, 2, \dots, p. \end{aligned}$$

Since $\tau \sim$ PH$(\boldsymbol{\pi}, A)$, we can interpret $f(s)\mathrm{d}s$ as $P(\tau \in [s, s + \mathrm{d}s))$. Then $f(s)\mathrm{d}s$ represents the probability that the absorption time, τ, is within the infinitesimal time interval, $[s, s + \mathrm{d}s)$. If $\tau \in [s, s + \mathrm{d}s)$, the CTMC $\{X_t\}_{t\geq 0}$ is in some transient state at time s. If $X_s = j$, the probability of a jump to the absorbing state $p + 1$ in the time interval $[s, s + \mathrm{d}s)$ is $a_j\mathrm{d}s$ where $\boldsymbol{a} = (a_1, a_2, \dots, a_p)^\intercal$. Furthermore, by definition, the probability of $\{X_t\}_{t\geq 0}$ starting in state

$i \in \{1, 2, \ldots, p\}$ is π_i. Collecting these facts, we get that

$$\begin{aligned} f(s)\mathrm{d}s &= P(\tau \in [s, s + \mathrm{d}s)) \\ &= \sum_{i=1}^{p} \pi_i \sum_{j=1}^{p} p_{ij}^{s} a_j \mathrm{d}s \\ &= \sum_{i=1}^{p} \sum_{j=1}^{p} \pi_i p_{ij}^{s} a_j \mathrm{d}s \\ &= \sum_{i=1}^{p} \sum_{j=1}^{p} \pi_i \left(e^{As}\right)_{ij} a_j \mathrm{d}s \\ &= \boldsymbol{\pi} e^{As} \boldsymbol{a} \mathrm{d}s, \end{aligned}$$

and the proof is finished. □

Theorem 2.2.3. *If $\tau \sim PH(\boldsymbol{\pi}, A)$, the distribution function F of τ is given $F(s) = 1 - \boldsymbol{\pi} e^{As\mathbf{1}}$.*

Proof. As $F(s)$ represents the probability that $\{X_t\}_{t\geq 0}$ is absorbed by time $s \geq 0$, that is $F(s) = \mathbb{P}(\tau \leq s)$, then $1 - F(s)$ represents the probability that $\{X_t\}_{t\geq 0}$ has not been absorbed by time s. This quantity can therefore be represented as

$$\begin{aligned} 1 - F(s) &= P(X_s \in \{1, 2, ..., p\}) \\ &= \sum_{i=1}^{p} \pi_i \sum_{j=1}^{p} p_{ij}^{s} \\ &= \sum_{i=1}^{p} \sum_{j=1}^{p} \pi_i p_{ij}^{s} \\ &= \sum_{i=1}^{p} \sum_{j=1}^{p} \pi_i \left(e^{As}\right)_{ij} \\ &= \boldsymbol{\pi} e^{As} \mathbf{1}, \end{aligned}$$

$$\implies F(s) = 1 - \boldsymbol{\pi} e^{As} \mathbf{1}.$$

□

The following technical result is of importance for forthcoming results: For any real (square) matrix B whose eigenvalues have strictly negative real part,

(P1) $e^{Bx} \to 0$ at an exponential rate as $x \to \infty$, (P2) B is invertible, and

(P3)

$$\int_0^\infty e^{Bs}\mathrm{d}s = (-B)^{-1}.$$

In particular, since any sub-intensity matrix A has eigenvalues with strictly negative real part (See Corollary 3.1.12, [4]) then P1, P2 and P3 hold with $B = A$. With this in mind, we are able to compute the Laplace transform of a phase-type distribution as follows.

Theorem 2.2.4. *1. The Laplace transform $L_\tau(s)$ of τ is given by*

$$L_\tau(s) := \mathbb{E}[e^{-s\tau}] = \boldsymbol{\pi}(sI - A)^{-1}\boldsymbol{a}, \quad \forall s \geq 0,$$

where I denotes the $p \times p$ identity matrix.

2. For $n \geq 1$, the n^{th} moment of τ is given by

$$\mathbb{E}[\tau^n] = (-1)^n n! \boldsymbol{\pi} A^{-n}\mathbf{1}.$$

Proof. Since A has eigenvalues with strictly negative real part, then so does $A - sI$ for any $s \geq 0$. Even more, $e^{(A-sI)x}$ decays at an exponential rate as $x \to 0$. Therefore, verifying the Laplace transform for τ gives

$$\begin{aligned}
L_\tau(s) &= E[e^{-s\tau}] \\
&= \int_0^\infty e^{-sx} f(x)\,\mathrm{d}x \\
&= \int_0^\infty e^{-sx}\boldsymbol{\pi} e^{Ax}\boldsymbol{a}\,\mathrm{d}x \\
&= \boldsymbol{\pi}\Big(\int_0^\infty e^{-sx}e^{Ax}\,\mathrm{d}x\Big)\boldsymbol{a} \\
&= \boldsymbol{\pi}\Big(\int_0^\infty (e^{(A-sI)x})\mathrm{d}x\Big)\,\boldsymbol{a} \\
&= \boldsymbol{\pi}(sI - A)^{-1}\boldsymbol{a}.
\end{aligned}$$

Regarding the n^{th} moment of τ,

$$\begin{aligned}\mathbb{E}[\tau^n] &= (-1)^n \frac{d^n}{\mathrm{d}s^n} L_\tau(s) \mid_{s=0} \\ &= (-1)^n(-1)^n \boldsymbol{\pi} n!(sI - A)^{-(n+1)} \mid_{s=0} \boldsymbol{a} \\ &= (-1)^{2n} \boldsymbol{\pi} n!(-A)^{-(n+1)} \boldsymbol{a} \\ &= \boldsymbol{\pi} n!(-A)^{-n}(-A)^{-1} \boldsymbol{a} \\ &= \boldsymbol{\pi} n!(-A)^{-n}(-A)^{-1}(-A)\mathbf{1} \\ &= \boldsymbol{\pi} n!(-A)^{-n}\mathbf{1} \\ &= (-1)^n n! \boldsymbol{\pi} A^{-n}\mathbf{1}.\end{aligned}$$

□

2.2.2 Phase-type renewal theory

Renewal processes are the prime example of how phase-type distributions solve complex problems easily. We highlight that, in terms of phase-type theory, the concept of concatenating random variables (which is central to this subsection and further chapters) arises in a natural way. Furthermore, it allows for simple expressions for complex problems, as we will see next.

Definition 2.2.5 (Renewal Process). *Let $W_1, W_2, \ldots$ be a sequence of i.i.d. positive random variables with $F(z) := F(W_i \leq w)$. Let $Y_n = W_1 + W_2 + \ldots + W_n, \quad n \geq 1$. Define the counting process $\{N(t)\}_{t\geq 0}$ by $N(t) = \sup\{n : Y_n \leq t\}$. We say that $\{N(t)\}_{t\geq 0}$ is a renewal process with arrival epochs Y_i and inter-arrival times W_i for $i = 1, 2, \ldots$.*

The reason these processes are called renewal processes is because probabilistically, since W_i are i.i.d., the process 'starts over' after each arrival epoch, Y_i. We will now consider renewal processes within the phase-type framework.

Definition 2.2.6 (Phase-type renewal process). *A phase-type renewal process $\{N(t)\}_{t\geq 0}$ is a renewal process such that the inter-arrival times have a phase-type distribution, $PH(\boldsymbol{\pi}, A)$.*

Of particular interest is the renewal density, $u : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ of $\{N(s)\}_{s\geq 0}$, defined by

$$u(s) = \frac{d}{\mathrm{d}s}\mathbb{E}[N(s)], \quad s \geq 0.$$

This quantity has the alternative interpretation that $u(s)\mathrm{d}s$ is the probability of a renewal (or arrival) during the infinitesimal time interval $[s, s + \mathrm{d}s)$. We will now discuss the concept of *concatenation*, and then demonstrate how it can be used to compute the renewal density.

Suppose that each $W_m \sim \mathrm{PH}(\boldsymbol{\pi}, A)$ with underlying CTMC $\{\hat{X}^m(s)\}_{s\geq 0}$. Concatenation of absorbing CTMCs means to join the chains end-by-end. By concatenating the underlying

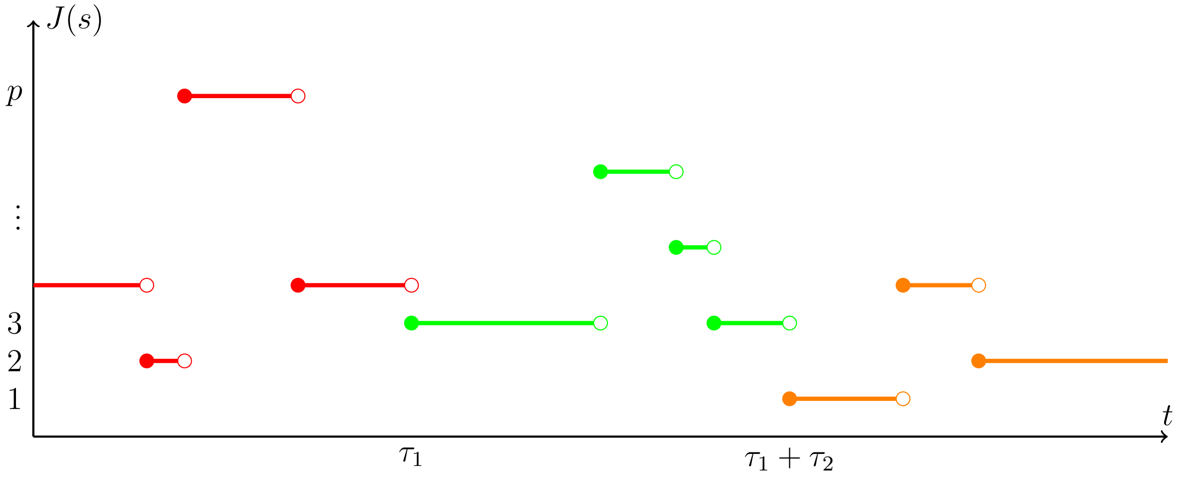

Figure 2.5: An example representation of a concatenated Markov process.

CTMCs $\{\hat{X}^m(s)\}_{s\geq 0}$ of the W_m (times between arrivals) up to the time of absorption, we obtain a new CTMC $\{J(s)\}_{s\geq 0}$ on the state space $\{1, 2, \ldots, p\}$. To visualise the concatenated process, see Figure 2.5. Note that each colour in the diagram denotes one of the underlying Markov jump process. The time of absorption of the first two Markov chains are τ_1 and τ_2 respectively.

For the concatenated process $\{J(s)\}_{s\geq 0}$, a jump from state i to state j in a time interval of length $\mathrm{d}t$ can occur in two possible ways:

1. Through the underlying phase-type distributed process, for which a transition will occur with probability $a_{ij}\mathrm{d}t$, where $A = \{a_{ij}\}$ or
2. By a process $\{\hat{X}^m(s)\}_{s\geq 0}$ exiting from state i to an absorbing state for some $m \geq 1$, and the next process $\{\hat{X}^{m+1}(s)\}_{s\geq 0}$ initiating in state j. This has probability $(a_i\mathrm{d}t)\pi_j$ where $\boldsymbol{a} = (a_1, a_2, \ldots, a_p)^\intercal$ and $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_p)$.

Note that these two types of transition for the concatenated process are mutually exclusive, and so the rate of a transition in $[t, t + \mathrm{d}t)$ is

$$\gamma_{ij}\mathrm{d}t := a_{ij}\mathrm{d}t + a_i\pi_i\mathrm{d}t. \tag{2.10}$$

Now, summing over all possible states in the Markov chain,

$$\begin{aligned} \Gamma\mathrm{d}t &= \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}\mathrm{d}t + \sum_{i=1}^{p}\sum_{j=1}^{p} a_i\pi_i\mathrm{d}t \\ &= A\mathrm{d}t + \boldsymbol{a\pi}\mathrm{d}t \\ \implies \Gamma &= A + \boldsymbol{a\pi}. \end{aligned}$$

That is, the concatenated process $\{J(s)\}_{s\geq 0}$ has the intensity matrix

$$\Gamma = A + \boldsymbol{a}\boldsymbol{\pi}. \tag{2.11}$$

The transition matrix of $\{J(s)\}_{s\geq 0}$ at time s is hence given by $e^{(A+\boldsymbol{a}\boldsymbol{\pi})s}$, which is key to finding an expression for the renewal density, u.

Theorem 2.2.7 (Renewal density)**.** *The renewal density, $u : \mathbb{R}_+ \to \mathbb{R}_+$, of a renewal process with inter arrival times which are PH($\boldsymbol{\pi}, A$) is given by $u(x) = \boldsymbol{\pi}e^{(A+\boldsymbol{a}\boldsymbol{\pi})x}$, $x \geq 0$.*

Proof. Conditioning on the initial state of $\{\hat{X}^1(s)\}_{s\geq 0}$ (the first process) and the state of the process $\{J(s)\}_{s\geq 0}$ at time x, we have that

$$\begin{aligned}
u(x)\mathrm{d}x &= P(\text{Renewal in } [x, x+\mathrm{d}x)) \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} P(\text{Renewal in } [x, x+\mathrm{d}x] \mid \hat{X}^1(0) = i, J(x) = j)P(J(x) = j \mid \hat{X}^1(0) = i)P(\hat{X}^1(0) = i) \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} P(\text{Absorption in } [x, x+\mathrm{d}x] \mid J(x) = j)P(J(x) = j \mid J^1(0) = i)P(J^1(0) = i) \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}(\text{Absorption in } [x, x+\mathrm{d}x] \mid J(x) = j)(e^{A+\boldsymbol{a}\boldsymbol{\pi}x})_{ij}\pi_{ij} \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}(a_j\mathrm{d}x)(e^{(A+\boldsymbol{a}\boldsymbol{\pi}x)})_{ij}\pi_i \\
&= \sum_{j=1}^{p} a_j(\boldsymbol{\pi}e^{(A+\boldsymbol{a}\boldsymbol{\pi})x})_j\mathrm{d}x \\
&= \boldsymbol{\pi}e^{(A+\boldsymbol{a}\boldsymbol{\pi})x}\boldsymbol{a}\mathrm{d}x,
\end{aligned}$$

where in the third equality we use that $\hat{X}^1(0) = J(0)$ by construction. □

Although the models we are interested in in forthcoming chapters are not renewal processes, similar concatenation arguments will still apply, resulting in compact formulae similar to that in Theorem 2.2.7

2.2.3 Multivariate phase-type distributions

We now extend the definition of phase-type distributions to the multivariate case. Multivariate phase-type theory will be useful in Chapter 5 where we draw comparisons between two dependent random variables.

Let $\tau \sim PH(\boldsymbol{\pi}, A)$ and let $\{X_t\}_{t\geq 0}$ be the underlying Markov jump process which generates τ. We let $Z_{i\ell}$ denote the duration of the ℓ^{th} visit of $\{X_t\}_{t\geq 0}$ to state i. We suppose $N_i \geq 0$ represents the total number of visits to state i. Then, each of the $Z_{i\ell}$ for $i = 1, 2, \ldots, p$ and $\ell = 1, 2, \ldots, N_i$ is exponentially distributed with rate $-a_{ii}$.

The random variables $Z_i := \sum_{\ell=1}^{N_i} Z_{i\ell} = \int_0^\tau 1\{X_u = i\}\mathrm{d}u$, for $i = 1, 2, \ldots, p$, are the total times the process $\{X_t\}_{t\geq 0}$ spends in the different states prior to absorption. Notice that $\tau = \sum_{i=1}^{p} 1 \cdot Z_i$, the time until absorption. By choosing constants different from 1, this idea can be generalised to reward each occupation time Z_i as follows.

We look at the construction of the corresponding Markov jump process where the rewards earned in each state correspond to the holding times in each state. Define the reward $\boldsymbol{r} = (r_1, r_2, \ldots, r_p)$, a vector of positive reward rates (constants). Also define $Y = \int_0^\tau r_{X_t}\mathrm{d}t = \sum_{i=1}^{p} r_i Z_i$, the total reward earned before absorption time τ. When $X_{S_n} = i$, the holding time of $\{X_t\}_{t\geq 0}$ in state i is exponentially distributed with rate λ_i. That is, $T_n \sim \text{Exp}(\lambda_i)$, as shown in Figure 2.6. The corresponding reward earned in state i, denoted r_i, is proportional to the time spent there. As a result, the total reward earned during this holding time for state i is exponentially distributed with rate λ_i/r_i. This can be visualised in Figure 2.6.

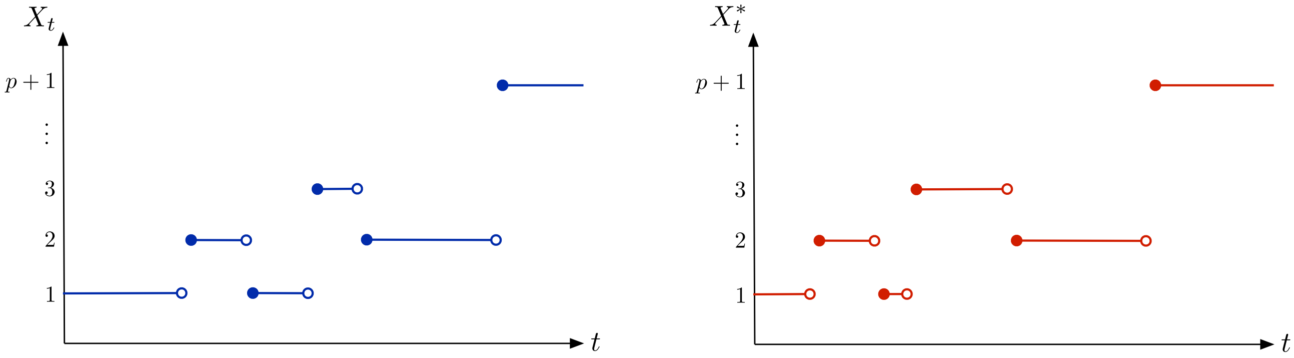

Figure 2.6: Left: Holding times of the CTMC $\{X_t\}_{t\geq 0}$ such that $T_i \sim \text{Exp}(\lambda_i)$. Right: The corresponding total rewards earned during the holding times, $r_i T_i \sim \text{Exp}(\lambda_i/r_i)$.

This transforms the original phase-type distribution in to another, which is important for further construction of multivariate phase-type distributions. Suppose that $\Delta(\boldsymbol{r})$ represents the matrix such that $\boldsymbol{r}$ lies on the diagonal. We therefore have that $Y \sim PH(\boldsymbol{\pi}, \Delta^{-1}(\boldsymbol{r})A)$. This concept can be generalised to a multivariate setting and for $r_i \geq 0$ as follows.

Definition 2.2.8 (Multivariate phase-type distribution (see Definition 8.1.1, [4])). *Let $n \geq 1$ be a positive integer and let $R = \{R_{ij}\}$ be a $p \times n$ matrix of non-negative rewards. Each column*

j of R may be considered to be a function $r_j : \{1, 2, \dots, p\} \longrightarrow \mathbb{R}_+$ defined by $r_j(i) = R_{ij}$. Let

$$Y_j = \int_0^\tau r_j(X_t)\mathrm{d}t = \sum_{i=1}^{p} R_{ij} Z_{i.} \quad j = 1, 2, \dots, n. \tag{2.12}$$

Then the random vector $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)$ is said to have a multivariate phase-type distribution parametrised by $\boldsymbol{\pi}$, S, and R, denoted $\boldsymbol{Y} \sim MPH_p^*(\boldsymbol{\pi}, S, R)$, where p represents the number of transient states in the underlying CTMC.

The joint distribution of $\boldsymbol{Y}$ can be expressed in terms of the joint moment generating function, which we will see below.

Theorem 2.2.9 (Moment Generating function). *Let $\boldsymbol{Y} \sim MPH^*(\boldsymbol{\pi}, A, R)$. Then there exists $\theta_0 > 0$ such that the multivariate moment generating function for $\boldsymbol{Y}$ exists and is given by*

$$H(\theta) = \mathbb{E}[e^{\langle Y, \theta \rangle}] = \boldsymbol{\pi}(-\Delta(R\theta) - A)^{-1}\boldsymbol{a} = \boldsymbol{\pi}(I - U\Delta(R\theta))^{-1}\mathbf{1}, \tag{2.13}$$

where $U = (-A)^{-1}$, $\boldsymbol{a} = -A\mathbf{1}$, and for any $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)$ where $\theta_i < \theta_0$. Recall also that $\Delta(\boldsymbol{a})$ represents a matrix where the vector $\boldsymbol{a}$ lies on the diagonal.

Proof. (See Theorem 8.1.2,[4]). □

This moment generating function allows us to compute cross-moments of bivariate multivariate phase-type distributions, which are defined as follows.

Theorem 2.2.10 (Cross moments for multivariate phase-type distribution). *For $Y \sim MPH_p^*(\boldsymbol{\pi}, A, R)$ and $h_j \in \mathbb{N}$, the cross moments are given by*

$$\mathbb{E}\left(\prod_{j=1}^{p} Y_j^{h_j}\right) = \boldsymbol{\pi} \sum_{\ell=1}^{h!} \left(\prod_{i=1}^{h} (-A^{-1}\Delta(R_{\cdot,\sigma_l(i)}) \right) \mathbf{1} \tag{2.14}$$

where $h = \sum_{j=1}^{n} h_j$, $R_{\cdot,j}$ is the j^{th} column of our reward matrix R, and $\sigma_1, \sigma_2, \dots, \sigma_{h!}$ are the ordered permutations of h-tuples of derivatives, and $\sigma_l(i)$ is the value among $1, 2, \dots, n$ at the i^{th} position of the permutation σ_l.

Proof. (See Theorem 8.1.5 [4]) □

Theorem 2.2.11 (Moments for bivariate phase-type distribution). *For $Y \sim MPH_p(\boldsymbol{\pi}, A, R)$ and $i, j \in \{1, 2\}$, the first, second and cross moments are given by*

$$\mathbb{E}[Y_i] = \boldsymbol{\pi} U \Delta(R_{.i})\mathbf{1} = \boldsymbol{\pi} U R_{.i}, \tag{2.15}$$

$$\mathbb{E}[Y_i^2] = 2\boldsymbol{\pi}(U\Delta(R_{.i}))^2\mathbf{1} = 2\boldsymbol{\pi}U\Delta(R_{.i})UR_{.i}, \tag{2.16}$$

and

$$\mathbb{E}[Y_iY_j] = \boldsymbol{\pi}U\Delta(R_{.i})UR_{.j} + \boldsymbol{\pi}U\Delta(R_{.j})UR_{.i}, \tag{2.17}$$

where $R_{.i}$ *denotes the* i^{th} *column of our matrix of rewards,* $R = \{R_{ij}\}$. *For this bivariate case,* $i = 1, 2$ *and* $j = 1, 2, \ldots, n$.

These explicit formulae for second and cross moments will be utilised in computations for covariance and correlation coefficient, which in turn allows us to study the dependence between phase-type distributed random variables. The use of these results is of particular importance and is demonstrated in Chapter 6. To prove these results we use the definition of the moment generating function for Y, as defined in Theorem 2.2.9.

Proof. 1. Firstly, using Equation 2.13,

$$\begin{aligned}
\mathbb{E}[Y_i] &= \frac{d}{d\theta_i}H(\theta) \mid_{\theta_i=0} \\
&= \frac{d}{d\theta_i}\boldsymbol{\pi}(I-(-A^{-1})\Delta(R\theta))^{-1}\mathbf{1} \\
&= \boldsymbol{\pi}\frac{d}{d\theta_i}((I-(-A^{-1})\Delta(R\theta))^{-1}) \mid_{\theta_i=0} \mathbf{1} \\
&= \boldsymbol{\pi}(-(I-(-A^{-1})\Delta(R\theta))^{-2}A^{-1}\Delta(R_{.i}))\mathbf{1} \mid_{\theta_i=0} \\
&= \boldsymbol{\pi}(-A^{-1}\Delta(R_{.i}))\mathbf{1} \\
&= \boldsymbol{\pi}(-A^{-1})R_{.i}.
\end{aligned}$$

2. Again, using equation 2.13

$$\begin{aligned}
\mathbb{E}[Y_i^2] &= \frac{d^2}{d\theta_i^2}H(\theta) \mid_{\theta_i=0} \\
&= \frac{d^2}{d\theta_i^2}\boldsymbol{\pi}(I-(-A^{-1})\Delta(R\theta))^{-1}\mathbf{1} \\
&= \boldsymbol{\pi}\frac{d^2}{d\theta_i^2}((I-(-A^{-1})\Delta(R\theta))^{-1}) \mid_{\theta_i=0} \mathbf{1} \\
&= \boldsymbol{\pi}\frac{d}{d\theta_i}(-(I-(-A^{-1})\Delta(R\theta))^{-2}A^{-1}\Delta(R_{.i})) \mid_{\theta_i=0} \mathbf{1} \\
&= \boldsymbol{\pi}(2(I-(-A^{-1})\Delta(R\theta))^{-3}A^{-1}\Delta(R_{.i}))(-A^{-1}\Delta(R_{.i}) \mid_{\theta_i=0} \mathbf{1} \\
&= 2\boldsymbol{\pi}A^{-1}\Delta(R_{.i})(-A^{-1})\Delta(R_{.i})\mathbf{1} \\
&= 2\boldsymbol{\pi}(-A^{-1}\Delta(R_{.i}))^2\mathbf{1}.
\end{aligned}$$

3. We refer to Theorem 2.2.10 to prove this result. Firstly we note that our ordered permutations are simply $\sigma_1 = (i, j)$ and $\sigma_2 = (j, i)$. Using Equation 2.14 from Theorem 2.2.10, we have that

$$\begin{aligned}
\mathbb{E}[Y_i Y_j] &= \boldsymbol{\pi} \sum_{\ell=1}^{2!} \left(\prod_{i=1}^{2} (-A^{-1}) \Delta(R_{.\sigma_\ell(i)}) \right) \mathbf{1} \\
&= \boldsymbol{\pi} \sum_{\ell=1}^{2} \big((-A^{-1}) \Delta(R_{.\sigma_\ell(1)}) (-A^{-1}) \Delta(R_{.\sigma_\ell(2)}) \big) \mathbf{1} \\
&= \boldsymbol{\pi} \big((-A^{-1}) \Delta(R_{.\sigma_1(1)}) (-A^{-1}) \Delta(R_{.\sigma_1(2)}) + (-A^{-1}) (R_{.\sigma_2(1)}) (-A^{-1}) \Delta(R_{.\sigma_2(2)}) \big) \mathbf{1} \\
&= \boldsymbol{\pi} \big((-A^{-1}) \Delta(R_{.i}) (-A^{-1}) \Delta(R_{.j}) + (-A^{-1}) \Delta(R_{.i}) (-A^{-1}) \Delta(R_{.j}) \big) \mathbf{1} \\
&= \boldsymbol{\pi} (-A^{-1}) \Delta(R_{.i}) (-A^{-1}) \Delta(R_{.j}) \mathbf{1} + \boldsymbol{\pi} (-A^{-1}) \Delta(R_{.i}) (-A^{-1}) \Delta(R_{.j}) \mathbf{1} \\
&= \boldsymbol{\pi} (-A^{-1}) \Delta(R_{.i}) (-A^{-1}) R_{.j} + \boldsymbol{\pi} (-A^{-1}) \Delta(R_{.i}) (-A^{-1}) R_{.j}
\end{aligned}$$

and the proof is complete. □

Chapter 3

Genealogical models

We begin this chapter by firstly motivating the need for a mathematical model that can describe the process which generates genetic data. We seek to understand the ancestral genealogy of a population by starting with a current day sample and looking backwards in time. Coalescent theory aims to answer questions about this ancestry and the behaviour of populations, which we outline, through development of the discrete and continuous-time coalescent models. Genealogical quantities of interest are introduced, including tree height and total branch length, and their biological interpretation is discussed.

3.1 The Wright-Fisher model

One of the simplest models to describe the genealogical relationship among genes over a series of generations was introduced by Wright (1931) and Fisher (1930) [11]. This reproductive model describes the evolution of a population and hence the transmission of genes through subsequent generations (in discrete time). We suppose the Wright-Fisher model follows a *haploid* reproductive model. The idea of a haploid model is that offspring are produced from a single parent. An alternative is to follow a *diploid* reproductive model whereby the population is split into N female and N male genes. In this case offspring are produced by considering one female gene and one male gene. Throughout this manuscript, we choose to assume a haploid population of $2N$ genes for simplicity, and because it has little practical consequence to do so [11]. The Wright-Fisher model follows some additional important simplifying assumptions.

1. *Constant population size.* The population size within each generation does not increase or decrease.

2. *Discrete and non-overlapping generations.* All individuals, in each generation, have the same life expectancy from conception to reproduction. Reproduction and death considered to be simultaneous events amongst all individuals.

3. *All individuals are equally fit.* We do not consider the possibility that any genes are more likely to reproduce. All genes are probabilistically equivalent.

4. *No geographical or social structure.* The Wright-Fisher model assumes that parent genes are sampled uniformly at random.

Such a genealogy is produced following Algorithm 2, and is demonstrated in Figure (3.1).

Algorithm 2 Constructing a population with the Wright-Fisher model

```
Start with $2N$ haploid genes in generation $\ell$
Sample uniformly, with replacement, one parent gene from generation $\ell$
Assign a replica copy of this selected gene to a spot in generation $\ell + 1$
Repeat steps 2-3 for subsequent generations
```

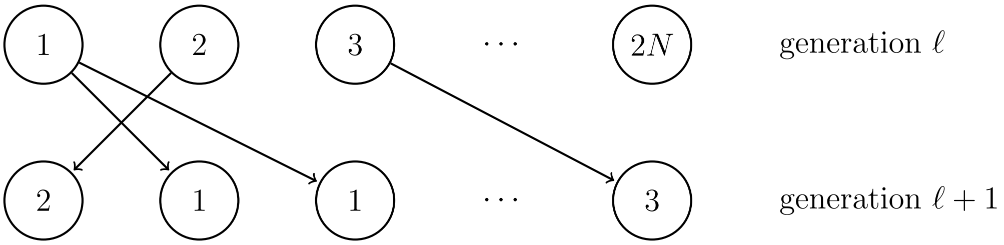

Figure 3.1: An example of one iteration of the haploid Wright-Fisher reproductive model which assumes a constant population of size $2N$.

We are able to determine the number of descendants of a particular individual in a future generation. Every gene in generation ℓ has equal probability $1/2N$ of being the parent gene to the offspring in generation $\ell + 1$. The number of descendants of an individual can therefore be represented by a binomial distribution with parameters $n = 2N$ and $p = 1/2N$. Define W_k to be the number of descendants of individual k in generation ℓ, where $k = 1, 2, \ldots, 2N$. Then,

$$P(W_k = x) = \binom{2N}{x} \frac{1}{2N}^{x} \left(1 - \frac{1}{2N}\right)^{2N-x}, \quad x = 1, 2, \ldots, 2N. \tag{3.1}$$

Since $W_k \sim \text{Binomial}\left(2N, \frac{1}{2N}\right)$,

$$\mathbb{E}[W_k] = np = 2N\frac{1}{2N} = 1, \tag{3.2}$$

and

$$\text{var}(W_k) = np(1-p) = 2N\frac{1}{2N}\left(1 - \frac{1}{2N}\right) = 1 - \frac{1}{2N}. \tag{3.3}$$

In Figure 3.1 we demonstrate a formulation of the Wright-Fisher genealogy with 8 generations, and 5 genes per generation.

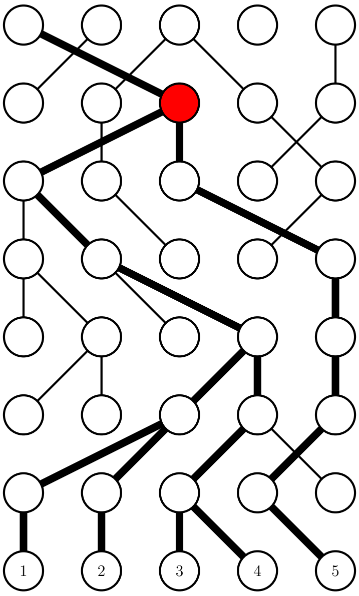

Figure 3.2: A sample genealogy of a haploid Wright-Fisher model. Each of the 8 rows represents a generation of size 5.

The genealogical relationship of the genes can be followed by tracing the lineages through subsequent generations. We will now shift our perspective. Rather than observing the evolution of this population, let us look at its ancestry. Generation 8 in Figure 3.1, which we will refer to as the current population, comprises of genes labelled $1, 2, 3, 4$ and 5. By following the bold lineages backwards in time through previous generations, we observe the ancestry of these genes. In this particular example, we notice that all 5 genes originated from an ancestor in generation 1. More importantly, however, we notice that genes $1, 2, 3, 4$ and 5 find their *most recent common ancestor (MRCA)* in generation 2 (6 generations back in time).

With this motivation, together with properties of the binomial and geometric distributions, we will derive the discrete-time (basic) coalescent process in Section 4.3.

3.2 The discrete-time coalescent

To properly understand such genealogies 'backwards in time', we need to introduce the notion of a coalescent event, which we do by the following example in Figure 3.1. We firstly extract the genealogy of genes $1, 2, 3, 4$ and 5 from Figure 3.1, and reconstruct it as outlined in Figure (3.2). To 'coalesce' means to 'merge'. This is the term we use to describe the event of two (or sometimes more) genes finding a parent gene at some point backwards in time. Here we see that genes 3 and 4 are the first to find a common parent, and their respective lineages merge in to one. In more technical terms, the first event of coalescence occurs between genes 3 and

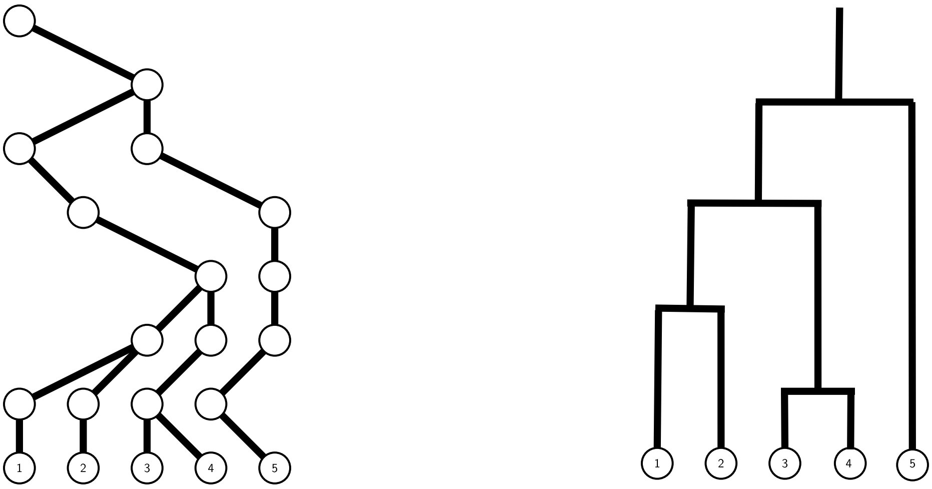

Figure 3.3: The Wright-Fisher genealogy (left) is re-structured to represent a coalescent tree diagram (right).

4 and as a result, creates the gene $\{3,4\}$. Then, genes $\{1\}$ and $\{2\}$ find a parent gene and become $\{1,2\}$. The third event of coalescence occurs when genes $\{1,2\}$ and $\{3,4\}$ merge, and become $\{1,2,3,4\}$. The fourth and final event of coalescence occurs when genes $\{1,2,3,4\}$ and $\{5\}$ merge, and their *MRCA* has been found.

3.2.1 The coalescence of two genes

To better familiarise ourselves with the concept of a coalescent, consider a present population of $2N$ genes, and suppose we are interested in the probability that two genes came from the same parent in the previous generation (one generation back in time). In this case, the first gene has $2N$ different ways of 'choosing' a parent. The probability that the second gene chooses this same parent is $1/(2N)$. The probability that any two genes find a common ancestor in one generation is therefore $1/(2N)$. Moreover, the probability that these two genes have different parents is its complement, $1 - 1/(2N)$.

We let V_k represent the waiting time (number of generations) for k genes to have $k-1$ ancestors. That is, number of generations back in time for a single pair of genes to merge. What about the probability that a single pair of genes find a common ancestor ℓ generations back in time? For two genes to have the same ancestor ℓ generations back in time, the genes must have *different* parent genes for the previous $\ell - 1$ generations, *and*, have the *same* parent in the ℓ^{th} generation. Assuming independence amongst generations, this has probability

$$P(V_2 = \ell) = \left(1 - \frac{1}{2N}\right)^{\ell-1} \frac{1}{2N}.$$

Thus, the time it takes (in generations) for two genes to find a common ancestor is a geometric random variable with parameter $1/(2N)$.

3.2.2 The coalescence of k genes

We are also interested in the probability that k individuals have the same parent in the previous generation. To derive this probability, we firstly consider the probability that k individuals have different parents. Note that for this to be the case, the first gene chooses their parent uniformly at random. The second gene then must choose a parent from the remaining $2N-1$. The third gene must then choose a parent from the remaining $2N-2$, and so on. So, the probability that k genes have different parents is given by

$$\begin{aligned}
&\frac{2N}{2N}\frac{(2N-1)}{2N}\frac{(2N-2)}{2N}\cdots\frac{(2N-(k-1))}{2N}\\
&=\prod_{i=1}^{k-1}\Big(1-\frac{i}{2N}\Big)\\
&=1-\sum_{i=1}^{k-1}\frac{i}{2N}+O\Big(\frac{1}{N^2}\Big)\\
&=1-\frac{k(k-1)}{4N}+O\Big(\frac{1}{N^2}\Big)\\
&=1-\binom{k}{2}\frac{1}{2N}+O\Big(\frac{1}{N^2}\Big).
\end{aligned}$$

Thus, for large values of N, the probability that k genes have the same parent in the previous generation is approximately

$$\binom{k}{2}\frac{1}{2N}$$

Now, what about the probability that k genes find a common ancestor ℓ generations back in time? We will again denote V_k as the number of generations back in time until the k genes find a common ancestor. Since the quantity of interest is geometrically distributed, an approximation of this probability for large values of N is

$$P(V_k=l)\approx\Big\{1-\binom{k}{2}\frac{1}{2N}\Big\}^{l-1}\binom{k}{2}\frac{1}{2N}. \tag{3.4}$$

With Equation (3.4) in mind, we are now able to derive the continuous-time coalescent which arises as the limit of the discrete-time coalescent in the following section.

3.3 The continuous time coalescent

We will now study the continuous approximation to the previously discussed discrete-time coalescent model, and detail specific coalescent models from the literature.

3.3.1 Further approximations of coalescence

In Section 4.3 we discussed the discrete time coalescent, where time is measured in generations. Throughout this section, let ℓ represent the number of generations back in time, and recall the haploid population is of size $2N$. Then, $t = \ell/2N$ represents a fraction of time which will be temporarily used as our new time scale. We now denote the waiting time for k genes to find a common ancestor as T_k.

From Equation (3.4), we have the probability that k genes find a common ancestor ℓ generations back in time. From the same equation, since V_k is geometrically distributed, the probability that k genes have *different* parents $\ell - 1$ generations back in time is approximately

$$\left\{1 - \binom{k}{2}\frac{1}{2N}\right\}^{\ell-1}.$$

This is equivalent to the probability that it takes more than $\ell - 1$ generations back in time to reach the MRCA. That is,

$$P(T_k > \ell - 1) = \left\{1 - \binom{k}{2}\frac{1}{2N}\right\}^{\ell-1} \tag{3.5}$$

Next, we will derive a further approximation by using an exponential approximation to the geometric distribution. From Equation (3.5) we see that

$$\begin{aligned} P(V_k > \ell) &= P(V_k > t2N) \\ &\approx \left\{1 - \binom{k}{2}\frac{1}{2N}\right\}^{t2N} \quad \text{from (3.5)} \\ &= \left(\left\{1 - \binom{k}{2}\frac{1}{2N}\right\}^{2N}\right)^{t} \\ &\to e^{-\binom{k}{2}t}, \quad as \quad N \to \infty \end{aligned}$$

Since $P(T_k > t2N) = P(T_k/2N > t)$, we have that

$$P\left(\frac{T_k}{2N} \leq t\right) \approx 1 - e^{-\binom{k}{2}t}. \tag{3.6}$$

Recall that we have implemented a change in how time is measured in such a way that one unit of time, t, corresponds to approximately $2N$ generations. Then, we see from the above that the waiting time for any 2 out of k genes to coalesce is exponentially distributed with rate $\binom{k}{2}$. An alternative explanation is that a *given* pair of genes merge at rate 1; since there are $\binom{k}{2}$ possible pairs, then the rate at which the first event of coalescence occurs is simply the sum of all possibilities, weighted by its corresponding merging rates, resulting in $\binom{k}{2} \cdot 1 = \binom{k}{2}$. Furthermore, Equation 3.6 motivates the need of a coalescent which evolves in continuous time, as opposed to the discrete time version studied in Section 4.3. In the rest of this chapter we study some common examples of continuous-time coalescents.

3.3.2 Kingman's coalescent

Kingman's n-coalescent is the most straightforward continuous-time coalescent in the sense that we can easily visualise and simulate this process, and it will therefore be discussed first. The coalescent structure itself follows the same genealogy that we discussed in the discrete-time coalescent section, meaning that only two genes (and no more) can merge at a given time. In a forward genealogical sense, this means that a parent gene has no more than two offspring at a time. Unlike our discrete-time coalescent, however, the time between subsequent coalescent events is no longer generational. These time increments are now randomly chosen such that they follow an exponential distribution. Algorithm 3 below outlines the simulation of a Kingman's coalescent.

Algorithm 3 Simulating Kingman's coalescent

```
Start with $n$ genes in the current population
Simulate the waiting time $T_n$ until an event of coalescence, $T_n \sim \exp(\binom{n}{2})$.
Choose randomly, two genes $a$ and $b$, uniformly among the $\binom{n}{2}$ possible genes.
Merge genes $a$ and $b$ in to one, and decrease the sample size $n \to n-1$.
If $n \geq 2$, go to step 2, otherwise stop.
```

We construct a sample simulation of Kingman's coalescent with an initial population of $n = 5$ genes, which is demonstrated in Figure 3.4. In this simulated example, we denote genes in the current generation as $\{1\}, \{2\}, \{3\}, \{4\}$ and $\{5\}$. The first event of coalescence occurs between genes $\{2\}$ and $\{3\}$ after time T_5. The second event of coalescence occurs between genes $\{1\}$ and $\{2, 3\}$ after time $T_5 + T_4$. As the algorithm states, this process continues backwards in time, until all 5 genes have found their MRCA.

Two important genealogical quantities in coalescent theory are *tree height* and the *total branch length*. If we consider the diagram as in Figure 3.4, the tree height represents the sum of the time epochs, T_k for $k = 2, 3, 4, 5$, in between the coalescent events. On the other hand, the total branch length represents the sum of these epochs, each multiplied by their respective number of lineages. These quantities can be computed in a straightforward way by the fact that the time until the most recent common ancestor (MRCA) is exponentially distributed.

Definition 3.3.1 (Tree height)**.** *The* tree height*, τ_n, of a population with initial size n is equal to the sum of the branches,*

$$\tau_n = T_2 + T_3 + ... + T_n, \tag{3.7}$$

and represents the time until the most recent common ancestor. For the case of Kingman's coalescent, $T_k \sim Exp(\lambda_k)$ where $\lambda_k = \binom{k}{2}$ for $k = 2, 3, \ldots, n$.

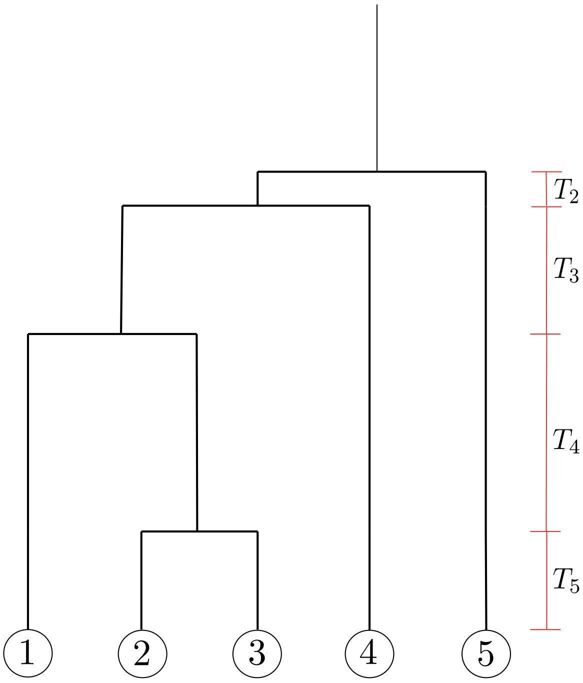

Figure 3.4: An sample simulation of the Kingman's-coalescent, starting with $n = 5$ genes in the present day population.

This quantity can be visualised with Figure (3.5).

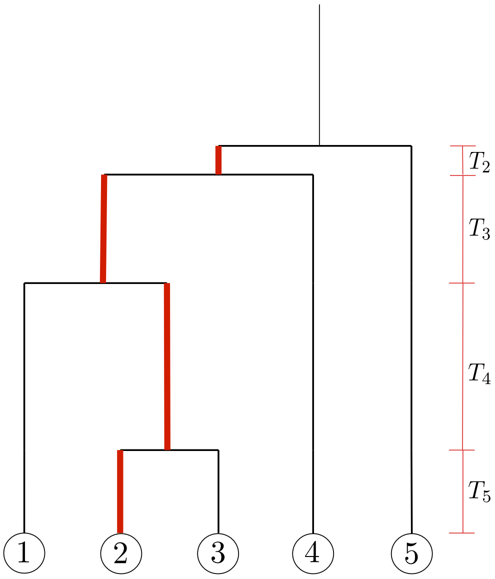

Figure 3.5: The sum of the red time epochs represent the tree height of our Kingman's coalescent model, for the case where our population is of size $n = 5$.

Since T_k is exponentially distributed for $k \in \{2, 3, \dots, n\}$, we can derive an expression for the expected tree height, $\mathbb{E}[\tau_n]$ as follows:

$$\begin{aligned}\mathbb{E}[\tau_n] &= \mathbb{E}\Big[\sum_{k=2}^{n} T_k\Big] \\ &= \sum_{k=2}^{n} T_k \mathbb{E}[T_k] \\ &= \sum_{k=2}^{n} \frac{1}{\binom{k}{2}} \\ &= \sum_{k=2}^{n} \frac{2}{k(k-1)} \\ &= 2\Big(1 - \frac{1}{n}\Big).\end{aligned}$$

Since the individual time epochs T_k for all $k \in \{2, 3, \dots, n\}$ are also independent, the variance can be expressed as follows.

$$\begin{aligned}\operatorname{var}(\tau_n) &= \operatorname{var}\Big(\sum_{k=2}^{n} T_k\Big) \\ &= \sum_{k=2}^{n} \operatorname{var}(T_k) \\ &= 4\sum_{k=2}^{n} \frac{1}{k^2(k-1)^2}.\end{aligned}$$

Definition 3.3.2 (Total branch length)**.** *The* total branch length*, L_n, of a population with initial size n is equal to*

$$L_n = nT_n + (n-1)T_{n-1} + ... + 2T_2. \tag{3.8}$$

We visualise the total branch length with Figure 3.6.

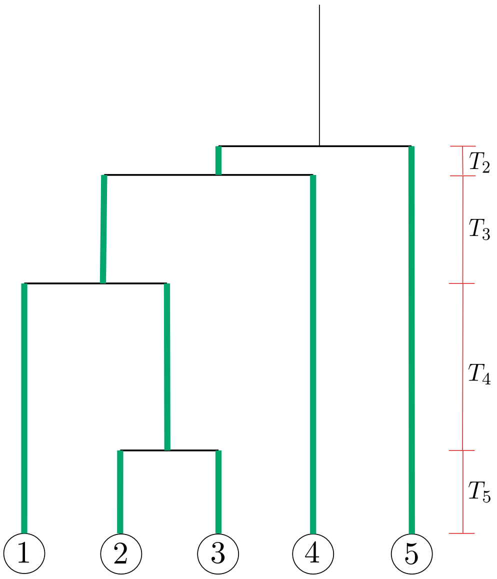

Figure 3.6: The sum of all the individual green branches represent the total branch length of our Kingman's coalescent model, for the case where our population is of size $n = 5$.

Notice that the total branch length is simply a weighting of the tree height. As a result,

$$\mathbb{E}[L_n] = \sum_{k=2}^{n} k\mathbb{E}[\tau_k] = 2\sum_{k=1}^{n-1} \frac{1}{k}$$

These approximations for $\mathbb{E}[\tau_n]$ and $\mathbb{E}[L_n]$ can be compared against the tree height and total branch length from a simulation of Kingman's coalescent, as shown in Figure 3.7.

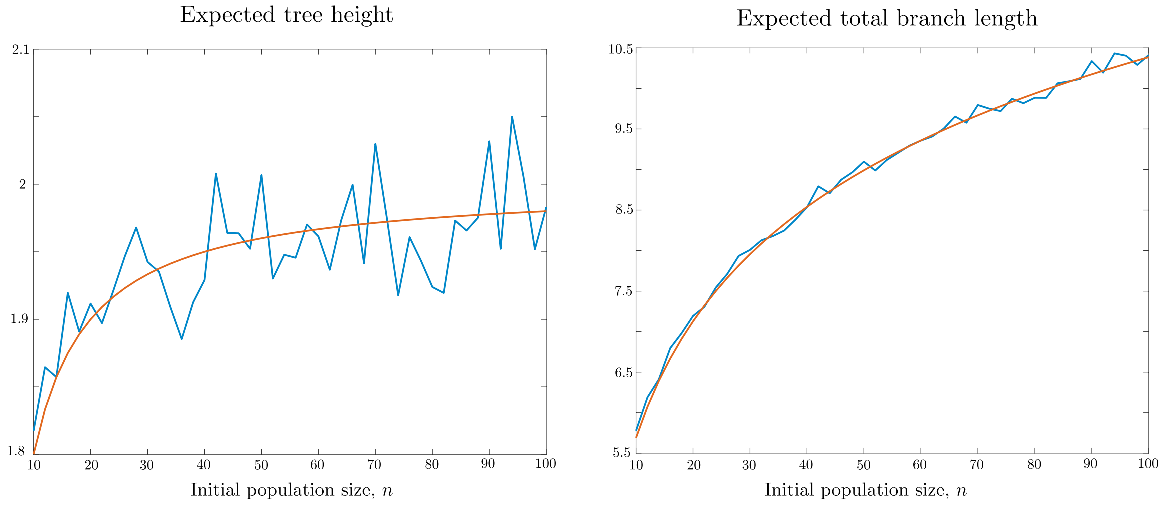

Figure 3.7: Left: Expected tree height simulation (blue) and discrete approximation $\mathbb{E}[\tau_n] = 2(1 - \frac{1}{n})$ (orange). Right: Expected total branch length simulation (blue) and discrete approximation $\mathbb{E}[L_n] = 2\sum_{k=1}^{n-1} \frac{1}{k}$. The simulated expectations were averaged over 1000 samples. Theoretical results were instantaneous and also present a much smoother curve.

The ability to approximate these quantities in such an efficient way (theoretically) is very advantageous in comparison to producing samples. This idea will be discussed further in the next chapter. Recall that tree height has the genealogical interpretation of being the time until the *MRCA* is found. Total branch length, on the other hand, gives us a sense of how much history genes within a sample share. A group of genes would share the least history if they all originated from a common ancestor some time ago and then evolved along distinct lineages.

3.3.3 The Λ-coalescent

The Λ-coalescent was introduced independently by Pitman (1999) [21] and Sagitov (1999) [24]. From a purely mathematical perspective, the Λ-coalescent characterises the class of exchangeable coalescent processes for which more than two genes can coalesce at once. It can therefore referred to as a *multiple merger coalescent*. Exchangability is an important feature of this process which will be explained next.

An exchangeable genetic process denotes a genealogy whereby all genes in a particular generation are equally likely to coalesce. There are no genes in any of the n generations that are more likely to be involved in an event of coalescence. Some of the previously discussed assumptions of the Kingman's coalescent which arises from the Wright-Fisher model are violated in many animal, plant and fungal species. Multiple-merger coalescents, such as the Λ-coalescent, are introduced to allow for large offspring numbers and have been proven to be a more accurate model for investigating certain genealogies. In particular, researchers including Menardo, Gagneuz and Freund (2018) [17] found that such a model was more accurate in investigating Mycobacterium Tuberculosis, a species of pathogenic bacteria.

The dynamics of the Λ-coalescent are characterised by a finite measure Λ on $[0, 1]$. When the process has b lineages, each subset of k lineages merge at a rate

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\Lambda(\mathrm{d}x), \quad k = 2, 3, ..., b. \tag{3.9}$$

where this integral can be interpreted as the expectation

$$\lambda_{b,k} = \lambda\mathbb{E}[X^{k-2}(1-X)^{b-k}] \quad \text{where} \quad X \sim \frac{\Lambda(\mathrm{d}x)}{\lambda} \quad \text{with the total mass} \quad \lambda = \int_0^1 \Lambda(\mathrm{d}x).$$

The tree diagram in Figure (3.8) demonstrates a sample simulation of how the Λ-coalescent evolves, assuming an initial population of $n = 8$ genes.

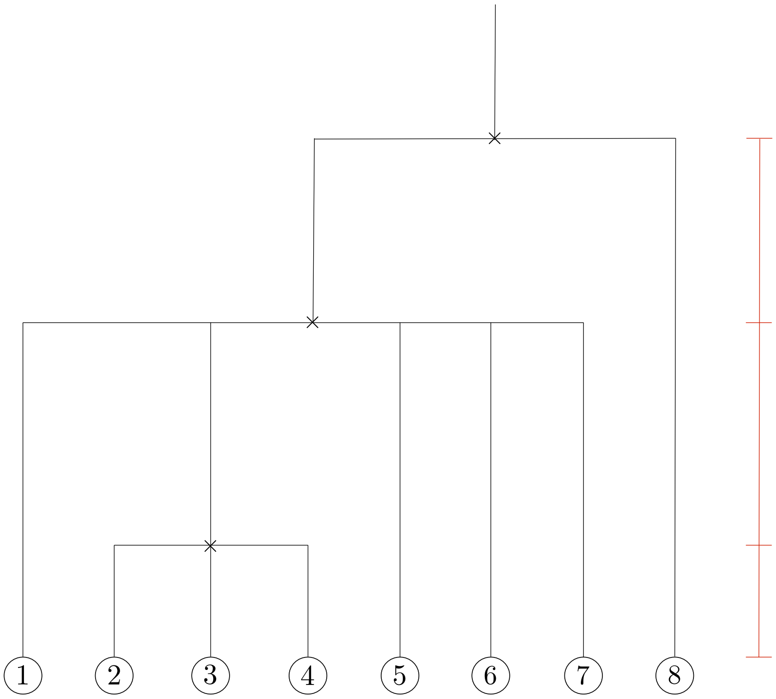

Figure 3.8: A sample simulation for the Λ-coalescent when the first generation has $n = 8$ genes. The first event of coalescence occurs amongst genes $\{2\}$, $\{3\}$ and $\{4\}$. The second event of coalescence occurs amongst five lineages, and the final event of coalescence occurs amongst two lineages.

To outline the general structure of the Λ coalescent, we consider the following particular formulation in which the first generation contains $b = 3$ genes:

1. The first event of coalescence occurs such that either two of the three genes merge, *or*, all three genes merge at once. There are $\binom{3}{2} = 3$ possible ways that $k = 2$ of the $b = 3$ total genes can merge. The rate at which the first event of coalescence occurs is therefore $3\lambda_{3,2} + \lambda_{3,3}$. The probability that two genes merge is
$$\frac{3\lambda_{3,2}}{3\lambda_{3,2} + \lambda_{3,3}},$$
whereas the probability that all three genes merge at once has probability
$$\frac{\lambda_{3,3}}{3\lambda_{3,2} + \lambda_{3,3}}.$$
2. If all three genes coalesce, then the MRCA has been found immediately. If two genes coalesce, it is equally probable that it is one of the three pairs of genes $\{1, 2\}$, $\{2, 3\}$ or $\{1, 3\}$. Each pair is equally probable of being the first to coalesce.

3. If the first event of coalescence occurred by two genes merging, there are now two lineages in the next generation. The second and final event of coalescence then occurs at a rate $\lambda_{2,2}$.

Now looking at the structure of the Λ-coalescent in a general sense, the ancestry of a population of size b behaves in the following way.

1. The first event of coalescence amongst the b total genes occurs such that any $k = 2, 3, \ldots, b$ genes could potentially merge. The rate of a coalescent event, assuming that our initial population size is $b = n$, is

$$g_n = \binom{n}{n}\lambda_{n,n} + \binom{n}{n-1}\lambda_{n,n-1} + \ldots + \binom{n}{2}\lambda_{n,2}. \tag{3.10}$$

   Any k genes, $2 \leq k < n$, merge with probability

$$\frac{\binom{n}{k}}{\lambda_{n,k}g_n},$$

   and all n genes coalesce at once with probability

$$\frac{\binom{n}{n}}{\lambda_{n,n}g_n}.$$

2. Due to exchangeability, if k_0 genes, $2 \leq k_0 < n$, coalesce in the first event, then any combination of k_0 genes merge with probability

$$\frac{1}{\binom{n}{k_0}}.$$

   In this case, the second event of coalescence occurs such that there are now $b = n - k_0 + 1$ total genes remaining in the second generation.

3. After the first event of coalescence and the population size has decreased, steps 1 and 2 are repeated until $b = 1$. That is, until the most recent common ancestor is found.

3.3.4 The ψ-coalescent

The ψ-coalescent is a particular example of the Λ-coalescent which appears as the limit of the genealogical process of *Moran models with a highly skewed offspring distribution* [6]. It is described as follows.

The ψ-coalescent differentiates two possible reproductive events in the underlying forward-in-time genealogical process. We consider the two different reproduction events to be a Moran model reproduction event (with probability $1 - \epsilon$), whereby a single individual is randomly chosen to reproduce and the single offspring replaces one randomly chosen (non-parental) individual, or a 'sweepstake' reproductive event, where a single parent replaces a proportion ψ of the $2N$ genes in the subsequent generation.
The Moran model is an alternative genealogical formulation to the Wright-Fisher model, and in contrast, contains overlapping generations. A new generation is formed from the previous generation by sampling uniformly at random one gene to give birth to a new gene, and one gene to die (see Section 1.1, [11]). The remaining genes persist as they are, until the next time step when they may be chosen to reproduce or die. Within the usual Moran model, introduced by Patrick Moran in 1958 [18], coalescence always occurs between two genes. In the general case, the number of genes involved in the coalescent event is instead a random variable that can range from 2 to $2N$, where recall that $2N$ is our population size. That is, $2, 3, ..., 2N$ genes can coalesce at any time step [6].

As mentioned in Section 4.4.3, researchers have been recently studying Mycobacterium Tuberculosis using multiple-merger coalescents. The ψ-coalescent, otherwise known as the Dirac-coalescent, was used for such an experiment and was found to be the best fitting coalescent for one of their particular datasets [17].

For the continuous-time analog, the ψ-coalescent, we consider the transition rates

$$\lambda_{b,k} = \psi^{k-2}(1-\psi)^{b-k}, \tag{3.11}$$

where again, b represents the number of total genes in the current population, k represents the number of genes to merge and $\psi \in (0, 1)$.

Similar to the Λ-coalescent case, this transition rate can also be expressed as

$$\lambda_{b,k} = \lambda \mathbb{E}[X^{k-2}(1-X)^{b-k}],$$

with $X \sim \delta_\psi$ and $\lambda = 1$, where δ_ψ denotes the point mass at ψ. Taking the limit of the ψ-coalescent as $\psi \to 0$,

$$\lim_{\psi \to 0} \lambda_{b,k} = \lim_{\psi \to 0} \psi^{k-2}(1-\psi)^{b-k},$$

we obtain the Kingman's coalescent.

3.3.5 The β-coalescent

The β-coalescent is yet another sub-class of the Λ-coalescent. In this case, it arises as the limit genealogical process of stable Galton-Watson, which we discuss next.

A Galton-Watson process (by Francis Galton and Henry William Watson [25]) is a branching process which behaves as follows. The underlying genealogy evolves by firstly considering an individual gene in generation 1. After one unit of time, this gene produces a randomly distributed number of offspring according to an assigned probability distribution. Each of the offspring in generation 2 live for one unit of time, and then also produce their own offspring. All subsequent generations behave in the same manner and the process continues in this way. By a proper time scaling, similar to that in Subsection 3.3.1, we obtain the β-coalescent, which are known to describe the genealogies of large populations where a single individual can produce a large number of offspring. Like the ψ-coalecent model, this particular class of the Λ-coalescent has also been applied to the Mycobacterium Tuberculosis study, where it was found to be the best-fitting model for nine of their data sets [17].

Now, recall that for the Λ-coalescent, when the process has b lineages, each subset of k lineages merge at the rate

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\Lambda(\mathrm{d}x), \quad k \in \{2, 3, ..., b\}.$$

In the case of the β-coalescent, the probability measure Λ is a beta$(2-\alpha, \alpha)$ distribution with $1 \leq \alpha < 2$. That is,

$$\Lambda(\mathrm{d}x) = \frac{1}{B(2-\alpha, \alpha)} x^{(2-\alpha)-1}(1-x)^{\alpha-1}\mathrm{d}x$$

where B is the beta function,

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Thus,

$$\begin{aligned}
\lambda_{b,k} &= \int_0^1 x^{k-2}(1-x)^{b-k}\Lambda(\mathrm{d}x), \quad k = 2, 3, ..., b \\
&= \frac{1}{\Gamma(2-\alpha)\Gamma(\alpha)} \int_0^1 x^{k-\alpha-1}(1-x)^{b-k+\alpha-1}(\mathrm{d}x) \\
&= \frac{1}{\Gamma(2-\alpha)\Gamma(\alpha)} \frac{\Gamma(k-\alpha)\Gamma(b-k+\alpha)}{\Gamma(b)} \\
&= \frac{B(k-\alpha, b-k+\alpha)}{B(\alpha, 2-\alpha)}.
\end{aligned}$$

Alternatively, the transition rate can also be expressed as

$$\lambda_{b,k} = \lambda\mathbb{E}[X^{k-2}(1-X)^{b-k}], \tag{3.12}$$

where $X \sim$ beta$(2-\alpha, \alpha)$ and $\lambda = 1$. Taking the limit of the β-coalescent as $\alpha \to 2$,

$$\lim_{\alpha\to 2} \lambda_{b,k} = \lim_{\alpha\to 2} \frac{B(k-\alpha, b-k+\alpha)}{B(\alpha, 2-\alpha)},$$

we arrive at Kingman's coalescent.

3.3.6 The seed-bank coalescent

The seed-bank coalescent considers a genealogy whereby genes can enter an inactive (dormant) state which arises as the in-ability for genes to reproduce in the underlying evolution process. Genes can remain in this state for arbitrarily many generations. The genes switch from active to inactive, and inactive back to active at fixed rates. Whilst active, genes coalesce according to the dynamics of Kingman's coalescent.

Seed-bank structures play an important role in the evolution of certain species. Genetic types can disappear from an active population and return later, due to the germination of seeds or activation of dormant forms. Dormancy is an important feature in the evolution of microorganisms, where some certain bacterial structures such as endospores, are non-reproductive. The addition of seed-bank structure is considered to be the addition of an evolutionary force to the previously discussed Kingman's coalescent. It is of interest how its incorporation effects population and thus, its ancestral structure. Researchers including Blath, Gonzalez, Casanova, Spanó (2018) [5] show that strong seed-bank effect can lead to genealogical behaviour that is vastly different to the long term dynamics of the Kingman's coalescent model. More specifically, they show that there are some scenarios by which the time until the MRCA is infinite, and that some genes may not even be able to find a common ancestor at all. We will now explain the dynamics of this process in terms of coalescent theory.

Let $c \geq 0$ represent the rate at which a lineage becomes inactive, and $K \geq 0$ be the rate at which an in-active lineage reactivates. Figure 3.3.6 shows a state transition diagram for the active and in-active lineages.

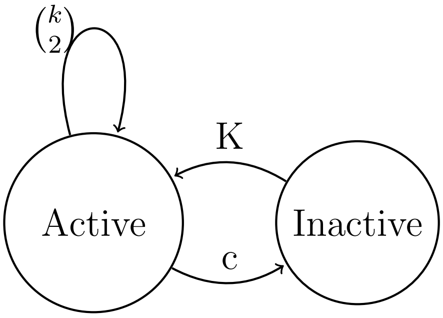

Figure 3.9: Structure of the seed-bank coalescent. An inactive gene reactivates at rate K, and an active gene becomes inactive at rate c. The active lineages can coalesce according to the Kingman's coalescent.

Denote i as the total number of lineages in the current generation, and j as the number of those which are inactive. The number of active lineages is thus $i - j$. Thus, the transition rate for any event

$$\lambda_{i,j} = \binom{i-j}{2} + (i-j)c + jK \tag{3.13}$$

where $\binom{i-j}{2}$ denotes the rate at which any two active lineages coalesce. In Figure 3.10 we show a representation of a sample genealogy for a population of size $i = 5$, where initially, $j = 2$ lineages are inactive. The dotted lines denote lineages that are presently inactive, and the × events denote the times at which the lineages switch from active to inactive, and vice versa.

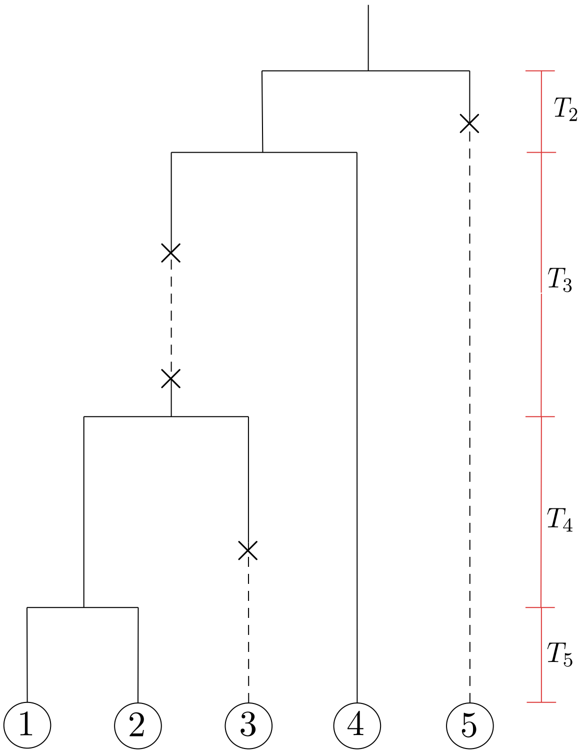

Figure 3.10: An example tree diagram for the seed bank coalescent with population size $i = 5$ where $j = 2$ genes are initially inactive.

The steps to constructing a seed-bank coalescent process are explained below, with reference to our particular example in Figure 3.10.

1. Initially we have total population size $i = 5$ and $j = 2$ inactive genes.

2. A random pair of the total 3 active genes coalesce with intensity $\binom{3}{2}$. In our example, genes $\{1\}$ and $\{2\}$ have coalesced with intensity 1 after time T_5. Now, our population size decreases and we have $i = 4$ and $j = 2$.

3. A random active gene of the total 2 in-active genes reactivates at rate $2K$. Here, gene $\{3\}$ reactivates at the rate c. Now our number of inactive genes has decreased by one, so $i = 4$ and $j = 1$.

4. Any arbitrary pair of the total 3 active genes coalesce with intensity $\binom{3}{2}$. Here, genes $\{1, 2\}$ and $\{3\}$ coalesce at time $T_5 + T_4$ with intensity 1. Our population size has decreased and now, $i = 3$ and $j = 1$.

5. Any of the remaining 2 active lineages becomes in-active at a rate $2c$. Here, gene $\{1, 2, 3\}$ becomes inactive at rate c. The number of inactive genes has increased by one, so $i = 3$ and $j = 2$.

6. Any of the 2 inactive lineages can reactivate at rate $2K$. In our example, gene $\{1, 2, 3\}$ reactivates with intensity K. The number of inactive genes has again decreased by one, so $i = 3$ and $j = 1$.

7. The 2 active lineages coalesce at time $T_5 + T_4 + T_3$ with intensity 1. Now, $i = 2$ and $j = 1$. There are 2 genes remaining in the population, one of which is inactive.

8. The remaining 1 inactive lineage, gene $\{5\}$, reactivates with rate K. Now $i = 2$ and $j = 0$.

9. The final event of coalescence occurs at time $\tau_n = T_5 + T_4 + T_3 + T_2$ with intensity 1, the MRCA has been reached.

Now, looking at the structure of the seed-bank coalescent in a general sense, the ancestry of a population of size i behaves in the following way.

1. The first event to occur, either an event of coalescence or the event that one of our genes becomes inactive is
$$\lambda_{i,j} = \binom{i-j}{2} + (i-j)c + jK.$$
   - Any two of the $i - j$ active genes will merge with probability
   $$\frac{\binom{i-j}{2}}{\binom{i-j}{2} + (i-j)c + jK}$$
   - One of the active genes will become in-active (dormant) with probability
   $$\frac{(i-j)c}{\binom{i-j}{2} + (i-j)c + jK}$$
   such that the numer of in-active genes, $i - j \geq 1$.

- One of the in-active genes will re-activate with probability

$$\frac{(jK}{\binom{i-j}{2}+(i-j)c+jK}$$

such that the number of in-active genes, $j \geq 1$.

Chapter 4

Phase-type theory in genetics

Throughout this chapter we will provide a demonstration of the previously discussed phase-type theory by applying it to a series of examples from coalescent theory, including Kingman's coalescent, the Λ-coalescent, and the seed-bank coalescent, as based on the work of Hobolth et. al. [13]. We are able to demonstrate that these coalescent models have phase-type distributed characteristics, by identifying that the time until a MRCA is found is the equivalent to reaching an absorbing state of a Markov chain. We will outline coalescent models within a phase-type framework, which in turn, allows us to derive straightforward, compact formulae for descriptors of tree height and total branch length, including expected values and densities.

4.1 Kingman's n-Coalescent

We will now discuss the phase-type representation for the Kingman's coalescent, as initially discussed in Chapter 3.

4.1.1 Tree height

Recall that the time epochs in between coalescent events are exponentially distributed units of time, and the tree height, τ_n, is equal to the sum of these epochs. That is,

$$\tau_n = T_n + T_{n-1} + ... + T_2, \tag{4.1}$$

which can be visualised for the case of $n = 5$ initial genes in Figure 3.5. This quantity has a phase-type distribution because the single absorbing state represents the generation at which the most recent common ancestor is found. That is, the transition from state $n - 1$ to state n is the last jump in the CTMC before it is absorbed.

More specifically, we claim that the tree height is phase-type distributed, $\tau_n \sim \mathrm{PH}(\boldsymbol{\pi}, A)$, where $\boldsymbol{\pi} = (1, 0, ..., 0)$ and

$$A = \begin{pmatrix} -\binom{n}{2} & \binom{n}{2} & & & & \\ & -\binom{n-1}{2} & \binom{n-1}{2} & & & \\ & & -\binom{n-2}{2} & \binom{n-2}{2} & & \\ & & & \ddots & \ddots & \\ & & & & & \\ & & & & & -\binom{2}{2} \end{pmatrix}. \tag{4.2}$$

We will now discuss more thoroughly the intensities in (4.2). Recall that the time between coalescent events are exponentially distributed with a rate which depends on the number of genes in the current population. For the case of the Kingman's coalescent, we have that $T_k \sim \mathrm{Exp}\left(\binom{k}{2}\right)$. That is, any 2 of the k genes in the current population can coalesce. Thus, we have

1. The first exit intensity is
$$\binom{n}{2} = \frac{n(n-1)}{2},$$
representing the rate at which the first event of coalescence occurs whereby the population decreases from n to $n-1$.

2. Similarly, the second exit intensity is
$$\binom{n-1}{2} = \frac{(n-1)(n-2)}{2},$$
representing the rate at which the second event of coalescence occurs, whereby the population size decreases from $n-1$ to $n-2$.

$\vdots$

k. In general, the k^{th} event of coalescence occurs at rate
$$\binom{n-k+1}{2} = \frac{(n-k+1)(n-k)}{2},$$
which represents the rate at which the population decreases from size $n-k+1$ to $n-k$.

The underlying CTMC can only transition forward from state k to state $(k+1)$, for all $k = 1, 2, ..., n-1$, since we are counting the tree height one generation at a time. This is why the time until absorption is the sum of the individual occupation times, T_k, for $k = n, n-1, ..., 2$.

4.1.2 Total branch length

Recall that the total branch length is defined as

$$L_n = nT_n + (n-1)T_{n-1} + ... + 2T_2 \tag{4.3}$$

where $T_k \sim \text{Exp}(\lambda_k)$ for $k \in \{n, n-1, \ldots, 2\}$. Since $T_k \sim \text{Exp}(\lambda_k)$ (the time it takes for one pair of the total k genes to coalesce), where $\lambda_k = \binom{k}{2} = k(k-1)/2$, then $kT_k \sim \text{Exp}(\lambda_k/k)$ by the scaling property of exponential distributions.

The quantity L_n therefore also has a phase-type distribution, $\text{PH}(\boldsymbol{\pi}, S)$, where $\boldsymbol{\pi} = (1, 0, ..., 0)$ and

$$S = \begin{pmatrix} -\frac{(n-1)}{2} & \frac{(n-1)}{2} & & & & \\ & -\frac{(n-2)}{2} & \frac{(n-2)}{2} & & & \\ & & -\frac{(n-3)}{2} & \frac{(n-3)}{2} & & \\ & & & \ddots & \ddots & \\ & & & & & \\ & & & & & -\frac{2}{2} \end{pmatrix} \tag{4.4}$$

The total branch length has phase-type distribution because like tree height, the time at which we reach the MRCA represents the absorbing state. The sub-intensity matrix, S, is described as follows.

1. The first exit intensity is $\frac{(n-1)}{2}$ since the expected branch length before the first event of coalescence is
$$\mathbb{E}[nT_n] = \frac{n}{\lambda_n},$$
where $\lambda_n = \binom{n}{2}$. That is,
$$\mathbb{E}[nT_n] = \frac{n}{\binom{n}{2}} = \frac{2}{(n-1)}.$$

2. The second exit intensity is $\frac{(n-2)}{2}$, again since the expected branch length before the second event of coalescence is
$$\mathbb{E}[(n-1)T_{n-1}] = \frac{(n-1)}{\lambda_{n-1}},$$
where $\lambda_{n-1} = \binom{n-1}{2}$. That is,
$$\mathbb{E}[(n-1)T_{n-1}] = \frac{(n-1)}{\binom{n-1}{2}} = \frac{2}{n-2}$$

3. In general, the k^{th} exit intensity is $\frac{(n-k+1)}{2}$ because the expected branch length before the k^{th} coalescence is
$$\mathbb{E}[(n-k+1)T_{n-k+1}] = \frac{(n-k+1)}{\lambda_{n-k+1}},$$
where $\lambda_{n-k+1} = \binom{n-k+1}{2}$. That is,
$$\mathbb{E}[(n-k+1)T_{n-k+1}] = \frac{(n-k+1)}{\binom{n-k+1}{2}}.$$

As previously mentioned, the underlying CTMC can only transition forward, from state k to state $(k+1)$, for all $k \in \{1, 2, \dots, n-1\}$. Since we are also counting the total branch length one generation at a time, the time until absorption is the sum of the individual occupation times, scaled by their respective lengths, kT_k, for $k \in \{n, n-1, \dots, 2\}$.

4.1.3 Theoretical results

Recall that in Chapter 3 we derived explicit formulae for descriptors of phase-type distributed random variables, including expectation and density. We will focus primarily on the expectation within the following sections, for the sake of its interpretability. With our phase-type results we derive theoretical values for the expected tree height and expected total branch length for the Kingman's coalescent, such that there are n genes in our initial population. From Theorem 2.2.4, we have that
$$\mathbb{E}[\tau_n] = \boldsymbol{\pi}(-A^{-1})\mathbf{1}$$
and
$$\mathbb{E}[L_n] = \boldsymbol{\pi}(-S^{-1})\mathbf{1}.$$
Recall that we also have the following equivalent results from Chapter 4,
$$\mathbb{E}[\tau_n] = 2\left(1 - \frac{1}{n}\right) \quad \text{and} \quad \mathbb{E}[L_n] = 2\sum_{k=1}^{n-1} \frac{1}{k},$$
and so instead of plotting our theoretical phase-type results, we refer back to Figure 3.7. We notice that the expected tree height, $\mathbb{E}[\tau_n]$ converges to 2 as $n \to \infty$.

It is important to specify here that since total branch length is simply a transformation of the tree height, its descriptors can also be computed using *rewards* which were discussed in Chapter 3. For upcoming examples, we will switch to using rewards rather than working with the scaled matrix S. We use the result from Section 3.2.3, where $Y \sim PH(\boldsymbol{\pi}, \Delta^{-1}(\boldsymbol{r})A)$ such that $\Delta(\boldsymbol{r})$ represents the matrix such that $\boldsymbol{r}$ lies on the diagonal, and A is our sub-intensity matrix for tree height. In the case of the Kingman's coalescent, as well as the ψ and β-coalescent models, we have $\boldsymbol{r} = (n, n-1, \dots, 2)^{\intercal}$. This concept will become more clear as we explore some more coalescent models.

4.2 The Λ-coalescent

Within this section we will focus on tree height for the two special cases of the Λ-coalescent already introduced in Chapter 4: the ψ-coalescent and the β-coalescent. Firstly we will define our phase-type parameters in terms of the general Λ-coalescent.

4.2.1 Tree height

Tree height for the case of the Λ-coalescent does not have a simple closed form expression as in the Kingman's case, since we need to consider the possibility that more than two genes will coalesce at once. This means that the total number of coalescent events is now random. As before, the tree height is still defined as the sum of the time epochs between coalescent events, however, since the number of coalescent events are random, we refrain from providing a formula analogous to that of the Kingman's coalescent. We simply define the tree height to be the time until the MRCA is reached. Despite this apparent difficulty, we are still able to provide simple expressions for its descriptors as follows.

Recall that in the case of the Λ-coalescent, when the process has b lineages, k lineages coalesce at rate

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\Lambda(\mathrm{d}x).$$

The height of the tree of a sample of n genes following dynamics of the Λ-coalescent is phase-type distributed, $\tau_n \sim \mathrm{PH}(\boldsymbol{\pi}, A)$ with $\boldsymbol{\pi} = (1, 0, \dots, 0)$,

$$A = \begin{pmatrix} -g_n & g_{n,2} & g_{n,3} & \cdots & g_{n,n-1} \\ & -g_{n-1} & g_{n-1,2} & \cdots & g_{n-1,n-2} \\ & & -g_{n-2} & \cdots & g_{n-2.n-3} \\ & & & \ddots & \\ & & & & \\ & & & & -g_2 \end{pmatrix}, \text{ and } \quad \boldsymbol{a} = \begin{pmatrix} g_{n,n} \\ g_{n-1,n-1} \\ g_{n-2,n-2} \\ \vdots \\ \\ g_2 \end{pmatrix} \tag{4.5}$$

where $g_{k,i} = \binom{k}{i}\lambda_{k,i}$ for $k \in \{2, 3, ..., n\}$ and $i \in \{2, 3, ..., k\}$. This quantity represents the rate at which any i of the total k remaining genes merge, where

$$\begin{aligned} g_k &= \sum_{i=2}^{k} g_{k,i} \\ &= g_{k,2} + g_{k,3} + \dots + g_{k,k-1} + g_{k,k}, \end{aligned}$$

representing the exit intensity from the state at which there are k genes remaining in the population. In the following we will discuss the first row of the above matrix, A, to give a more clear understanding of these intensities.

The first row of matrix A contains the rates at which the first coalescent event occurs amongst the initial n genes in the population, with the exception of the absorption event in which all n genes coalesce at once. The time until the first event of coalescence is exponentially distributed,

$$\begin{aligned}\mathrm{Exp}(g_n) &= \mathrm{Exp}\Big(\sum_{i=2}^{n} g_{n,i}\Big)\\ &= \mathrm{Exp}\big(g_{n,2}+g_{n,3}+\ldots+g_{n,n-1}+g_{n,n}\big).\end{aligned}$$

Elaborating on this exit intensity $g_n = g_{n,2} + g_{n,3} + \ldots + g_{n,n-1} + g_{n,n}$,

- $g_{n,2} = \binom{n}{2}\lambda_{n,2}$ represents the rate at which any 2 of the total n genes coalesce. In other words, $g_{n,2}$ is the rate of the first coalescent event such that only two genes merge.
- $g_{n,3} = \binom{n}{3}\lambda_{n,3}$ is the rate at which any 3 of the total n genes coalesce.
- $g_{n,n-1} = \binom{n}{n-1}\lambda_{n,n-1}$ is the rate at which $n-1$ of the total n genes merge. That is, all but one of the genes coalesce.
- $g_{n,n} = \binom{n}{n}\lambda_{n,n} = \lambda_{n,n}$is the rate at which all the genes coalesce at once. That is, the rate at which we find the most recent common ancestor after the very first coalescent event and hence the process is absorbed.

It is important to emphasise here that, unlike in the case of the Kingman's coalescent, the second event of coalescence does not necessarily correspond to the 'second coalescence of two genes'. We are referring to an event of coalescence as the coalescence of i genes, where $i \in \{2, 3, \ldots, k)$. In other words, the next state of our process is dependent on the type of coalescence that occurred previously. Suppose that $k < n$ genes merged in the first generation. That is, the population size drops from size n to size $n-k+1$ after the first coalescent event. In this case, the time between the first and second events of coalescence is

$$\mathrm{Exp}(g_{n-k+1}) = \mathrm{Exp}\left(\sum_{i=2}^{n-k+1} g_{n-k+1,i}\right)$$

where $g_{n-k+1} = g_{n-k+1,2} + g_{n-k+1,3} + \ldots + g_{n-k+1,n-k} + g_{n-k+1,n-k+1}$.

It is clear that the underlying Markov chain can only transition forward. That is, the process only jumps from state k to state $(k+j)$ for all $k \in \{2, 3, \ldots, n-1\}$ and $j \in \{1, 2, \ldots, n\}$

such that $(k+j) \leq n$. The absorbing state of this process again corresponds to the time at which the most recent common ancestor is found, so the tree height for the Λ-coalescent is phase-type distributed.

Now that we have defined the general model, we will focus on our two specific cases: the ψ-coalescent and $\beta-$coalescent.

4.3 The ψ-coalescent

As mentioned, the ψ-coalescent is a special case of the Λ-coalescent. In (4.5) we defined the sub-intensity matrix for the tree height of the Λ-coalescent. In the case of a ψ-coalescent, the intensities take the form

$$g_{b,k} = \binom{b}{k} \lambda_{b,k}$$

with

$$\lambda_{b,k} = \psi^{k-2}(1-\psi)^{b-k} \quad \text{for} \quad \psi \in (0,1).$$

4.3.1 Tree height and total branch length

With the aid of (4.5), we can again explicitly evaluate the expected tree height $\mathbb{E}[\tau_n] = \boldsymbol{\pi}(-A^{-1})\mathbf{1}$ in a variety of situations. Results are shown below in Figure 4.1.

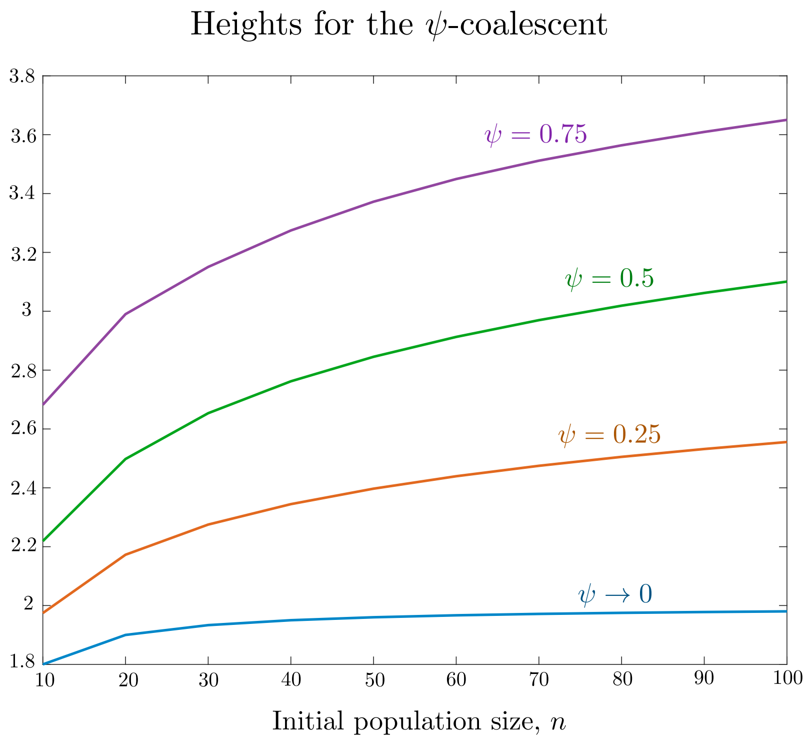

Figure 4.1: Expected tree heights, $\mathbb{E}[\tau_n]$, for the ψ-coalescent with varying values of ψ. We notice that as $\psi \to 1$, the expected tree height increases. The expected tree height also increases as population size increases.

The expected tree heights for the ψ-coalescent increase as our parameter ψ increases. We firstly notice that our results for $\psi \to 0$ are consistent with the tree height for Kingman's coalescent, as expected. Recall that in this underlying genealogical process we have rare reproductive events, with a greatly skewed offspring distribution. That is, ψ represents the proportion by which an individuals descendants replace the next (offspring) generation. As $\psi \to 1$, every sample of genes is likely to find its MRCA in the first sweepstake reproduction event. Further reading in to the ψ-coalescent, including the use of varying time scales can be seen in (Eldon and Wakeley, 2006 [6]). Similarly, we also calculate the expected total branch length for the ψ-coalescent, using rewards $\boldsymbol{r} = (n, n-1, \ldots, 2)$. Results are shown in Figure 4.2.

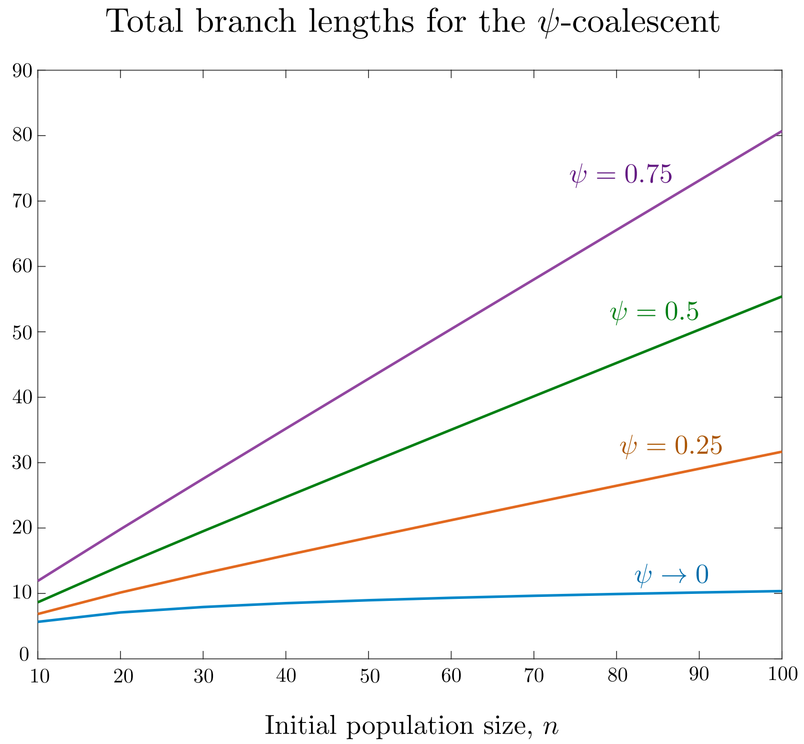

Figure 4.2: Expected total branch length, $\mathbb{E}[L_n]$, for the ψ-coalescent with varying values of $\psi \in (0,1)$. Branch lengths appear to increase linearly for initial population sizes less than $n = 100$, with the exception of $\psi \to 0$ (Kingman's coalescent), which approaches a limit.

As expected, the total branch lengths behaves in the same way as the previously seen tree height. As ψ increases, the total branch length also increases. As well as the expected tree height and total branch length, we can also compute higher moments. Here we make a correction to the paper (Hobolth et. al. [13]). It has been noticed that by a direct comparison for the case $\psi \to 0$ to results using Kingman's coalescent, that the computation for higher order moments is inconsistent. This should not occur, since as $\psi \to 0$ we should obtain the Kingman's coalescent results. The computations for higher order moments for the ψ-coalescent are missing a factor of 2 in their solutions. Recall that the second moment of a phase-type distributed random variable τ is denoted

$$\mathbb{E}[\tau^2] = 2\boldsymbol{\pi}A^{-2}\mathbf{1}. \tag{4.6}$$

It was noticed that the paper instead computes

$$\boldsymbol{\pi} A^{-2}\mathbf{1},$$

which has been corrected by plotting Equation 4.6, as shown in Figure 4.3.

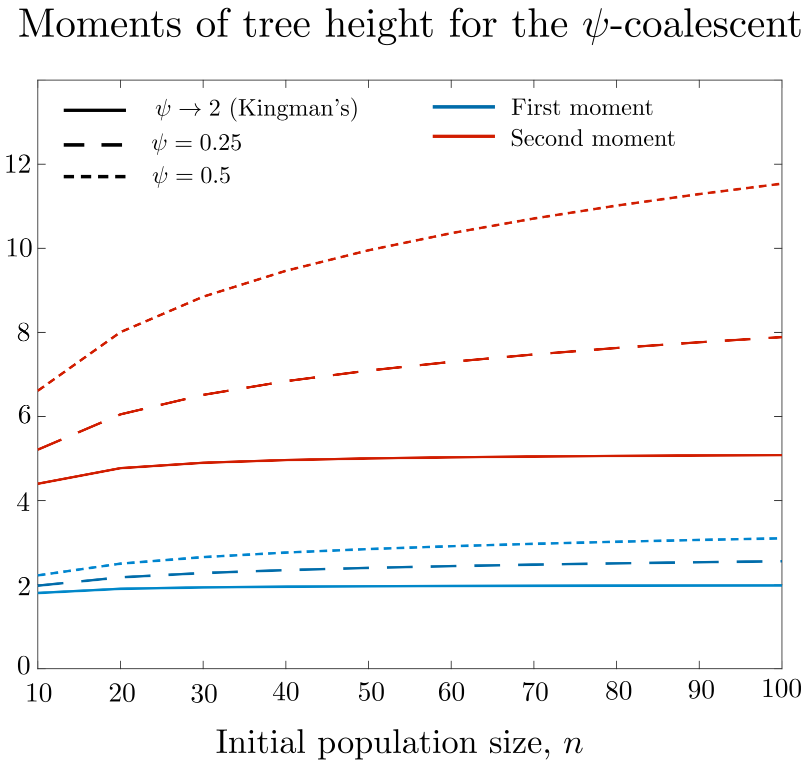

Figure 4.3: A corrected plot of the first and second moments of the ψ-coalescent for $\psi \in (0, 1)$, which was originally presented as a scaling error within Hobolth et.al. (2019) [13]

We may also verify now, by inspection, that the constraint $\mathbb{E}[\tau^2] \geq \mathbb{E}[\tau]^2$ holds, whereas previously it did not. We notice that the second moments appear to get steeper as $\psi \to 1$, whereas the first order moments flatten significantly. We therefore infer that as $\psi \to 1$, the genealogy displays greater variance. The inclusion of sweepstake reproduction events in the underlying genealogical process causes greater variance in the tree height. This result makes sense, since the number of genes involved in reproduction events as a larger range of potential values as $\psi \to 1$.

4.4 The β-coalescent

Recall that the β-coalescent is a special case of the Λ coalescent with intensities

$$g_{b,k} = \binom{b}{k}\lambda_{b,k}$$

for

$$\lambda_{b,k} = \frac{B(k-\alpha, b-k+\alpha)}{B(\alpha, 2-\alpha)}, \quad \text{where } B \text{ is the beta function, } 1 \leq \alpha < 2.$$

4.4.1 Tree height and total branch length

Tree height is demonstrated in Figure 4.4 and total branch length in Figure 4.5.

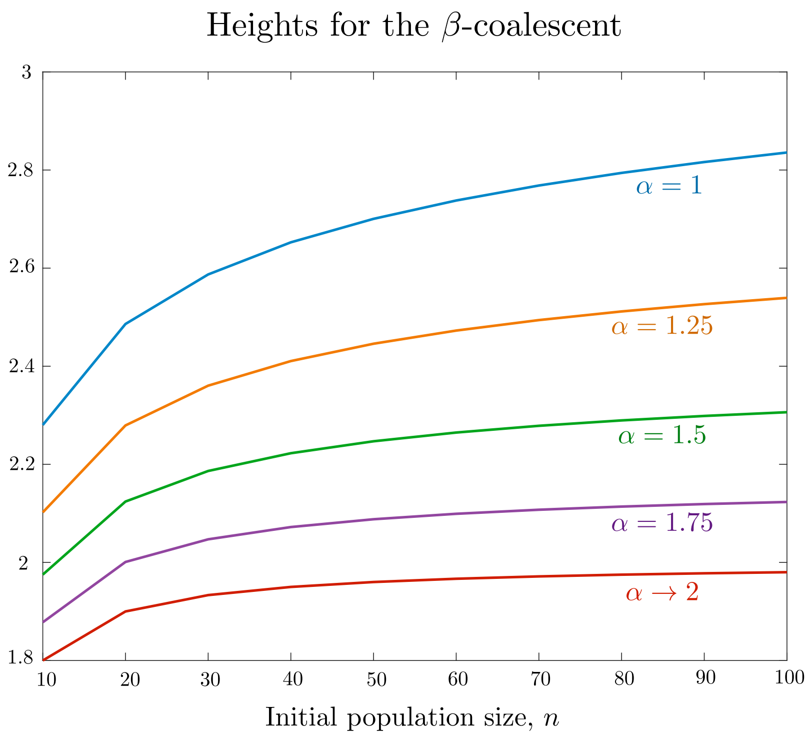

Figure 4.4: Expected tree height, $\mathbb{E}[\tau_n]$, for the β-coalescent with varying values of α. As α increases, the expected tree height decreases. As sample size n increases, then these expected tree heights appear to reach a limit, particularly as $\alpha \to 2$.

We observe that, as expected, as $\alpha \to 2$ we again arrive at the Kingman's coalescent. This is because as α increases, the probability of a large coalescent event (or, in turn, a large reproduction event) decreases. We reach this limit whereby only two genes are merging at a given time - resulting in the Kingman's result. In general, as α increases, the tree height decreases.

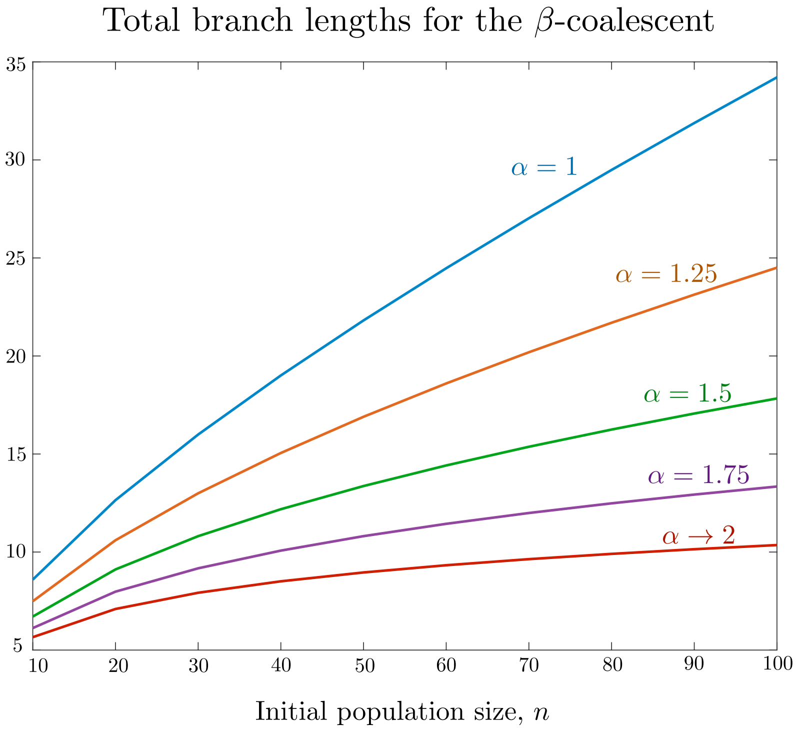

Figure 4.5: Expected total branch length, $\mathbb{E}[L_n]$, for the β-coalescent with varying values of α. Again, as expected, the expected total branch length increases as α decreases.

The expected branch length behaves in the same way as the expected tree height; increasing as α decreases. As in the case of the ψ-coalescent, we again compute second order moments. The same correction as mentioned in Section 4.3 is made to the paper by Hobolth et. al. (2019) [13], where we amend the previous scaling error in the calculation for the second moment for tree height. The corrected plot for the second moment is shown in Figure 4.6 below.

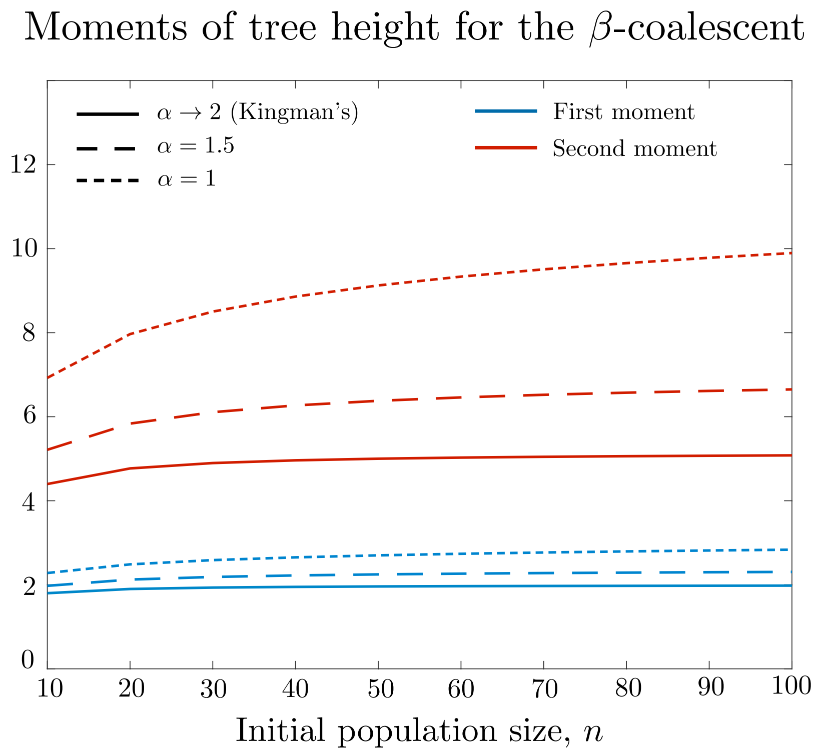

Figure 4.6: Corrected second moments for the β-coalescent, for $1 \leq \alpha < 2$, which was originally presented as a scaling error within Hobolth et.al. (2019) [13].

Again, results in Figure 4.6 were verified by comparing the Kingman's coalescent case ($\alpha \to 2$) to confirm that the same factor of 2 was indeed missing from the second order moment computation. It holds now that $\mathbb{E}[\tau^2] \leq \mathbb{E}[\tau]^2$, and from the figure we can interpret the behaviour of variance for the tree height. The variance for tree height is largest in the case of $\alpha = 1$, that is, when the underlying genealogical process is producing its largest offspring numbers. This coalescent process is known as the Bolthausen–Sznitman coalescent [26]. The variance of tree height in this case tends to a constant as population size increases, in comparison the ψ-coalescent case where variance appears to continue increasing for larger population sizes.

4.5 The seed-bank coalescent

Recall that within the seed-bank coalescent, genes can enter an in-active (dormant) state, and also re-activate. The active genes can coalesce according to the dynamics of the Kingman's coalescent. We will demonstrate that the tree height for a seed-bank coalescent is phase-type distributed as follows.

4.5.1 Tree height

The matrix for tree height for the seed-bank coalescent has a more complicated structure than those we have previously defined. In this case, we need to incorporate a block matrix structure. Recall that

$$\lambda_{i,j} = \binom{i-j}{2} + (i-j)c + jK,$$

where i denotes the total size of our population, and j denotes the number of dormant genes. We firstly define the $(i+1) \times (i+1)$ matrix

$$\Gamma(i) = \begin{pmatrix} -\lambda_{i,0} & ic & & & & & \\ K & -\lambda_{i,1} & (i-1)c & & & & \\ & 2K & -\lambda_{i,2} & (i-2)c & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & & & & \\ & & & & & -\lambda_{i,i-1} & c \\ & & & & & iK & -\lambda_{i,i} \end{pmatrix} \tag{4.7}$$

containing the rates at which genes become in-active and re-activate, such that the system starts and remains with total size i. The entries of this matrix are denoted

$$\Gamma(i) = \{\gamma_{\ell,j}(i)\}.$$

We will now explain the structure of this matrix by looking at rows 0, ℓ and i, denoting the first row as row 0.

Starting with row 0, corresponding to the state in which there are 0 in-active genes (all genes are active):

- The $(0,1)$ entry, $\gamma_{0,1}(i) = ic$, represents the rate at which any one of the $(i-0)$ active genes becomes in-active.
- The $(0,j)$ entry (for $j \neq 1$), $\gamma_{0,j}(i) = 0$ is null because it is not possible for more than 1 gene to become in-active simultaneously.

Now looking at row ℓ, which corresponds to the state in which there are $j = \ell$ in-active genes ($i - \ell$ active genes):

- The $(\ell, \ell-1)$ entry, $\gamma_{\ell,\ell-1}(i) = \ell K$ represents the rate at which any one of the l in-active lineages re-activates.

- The $(\ell, \ell+1)$ entry, $\gamma_{\ell,\ell+1}(i) = (i-\ell)c$ represents the rate at which any one of the $(i-l)$ active genes become in-active.

- The (ℓ, j) entry (for $j < \ell - 1$), $\gamma_{\ell,j}(i) = 0$ is null because it is not possible for more than one of the ℓ in-active genes to re-activate.

- The (ℓ, j) entry (for $j > \ell + 1$), $\gamma_{\ell,j}(i) = 0$ is null because it is not possible for more than one gene to become in-active at the same time.

Finally, row i corresponds to the state in which there are i in-active genes (all genes are in-active):

- The $(i, i-1)$ entry, $\gamma_{i,i-1}(i) = iK$ is the rate at which any of the in-active lineages re-activate (here, all i genes are in-active).

- The (i, j) entry for $j < i - 1$ is null because it is not possible for more than one of the i in-active genes to re-activate.

As well as the $\Gamma(i)$ matrix, we also define the following $(i+1) \times i$ matrix

$$D(i) = \begin{pmatrix} \binom{i}{2} & & & & \\ & \binom{i-1}{2} & & & \\ & & \binom{i-2}{2} & & \\ & & & \ddots & \\ & & & & \\ & & & & 0 \end{pmatrix}. \tag{4.8}$$

which contains transition rates for when the system loses an element through an event of coalescence, such that it started from total population size i. This matrix does not have a structure that can be interpreted by itself in the same way as matrix $\Gamma(i)$, but rather requires us to interpret each quantity as a jump to another set of states. We will also elaborate on the structure of this matrix by again providing an outline of rows 0, $\ell \in \{1, 2, \ldots, i-1\}$ and i:

- The $(0, 0)$ entry represents the rate at which any two of the i total active lineages coalesce.

- The (ℓ, j) entry, for $\ell = j,\ 0 < \ell, j < i$ represents the rate at which any two of the $i - j$ active genes coalesce

- The (i, i) entry represents the rate at which any two of the 0 active genes coalesce, which is of course null.

Putting together the two matrices $\Gamma(i)$ and $D(i)$ in to the following form, we define the sub-intensity matrix for the tree height of a population with n genes as

$$B = \begin{pmatrix} \Gamma(n) & D(n) & & & & \\ & \Gamma(n-1) & D(n-1) & & & \\ & & \Gamma(n-2) & D(n-2) & & \\ & & & \ddots & \ddots & \\ & & & & & \\ & & & & & \Gamma(2) \end{pmatrix}, \tag{4.9}$$

which, as mentioned previously, has block matrix structure. Before explaining how this system works in thorough detail, we will look at the expanded form of the tree height matrix (4.9). This block structure can be visualised in Figure 4.7 below, for the case where $i = 5$. We see in Figure 4.7 that the sub-matrices $\Gamma(i)$,which has size $(i+1) \times (i+1)$, and $D(i)$ which has size $(i+1) \times i$, fit together in such a way that the tree height matrix, B, is indeed square. The $\Gamma(i)$ matrices are shown in green, and the $D(i)$ matrices are coloured in orange. Note that this is a sub-intensity matrix, so the diagonal elements are

$$-\lambda_{i,j} = -\left(\binom{i-j}{2} + (i-j)c + jK\right) \quad \text{for } j = 0, 1, \ldots, i.$$

The first row of block matrices, which we denote

$$[\Gamma(5) \quad D(5)] = \begin{bmatrix} -\lambda_{5,0} & 5c & 0 & 0 & 0 & 0 & \binom{5}{2} & 0 & 0 & 0 & 0 \\ K & -\lambda_{4,1} & 4c & 0 & 0 & 0 & 0 & \binom{4}{2} & 0 & 0 & 0 \\ 0 & 2K & -\lambda_{3,2} & 3c & 0 & 0 & 0 & 0 & \binom{3}{2} & 0 & 0 \\ 0 & 0 & 3K & -\lambda_{2,3} & 2c & 0 & 0 & 0 & 0 & \binom{2}{2} & 0 \\ 0 & 0 & 0 & 4K & -\lambda_{1,4} & c & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5K & -\lambda_{0,5} & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

contain the transition rates for our system such that the total population starts and remains with size 5. We will explain the states of this sub-matrix row-by-row, starting at row 0.

0: Row 0 corresponds to the state where there are $j = 0$ in-active genes in the population (all genes are active). It contains the intensities

$$\begin{bmatrix} -\lambda_{5,0} & 5c & 0 & 0 & 0 & 0 & \binom{5}{2} & 0 & 0 & 0 & 0 \end{bmatrix}$$

where

- At rate $5c$, one of the 5 currently active genes will enter a dormant state, and

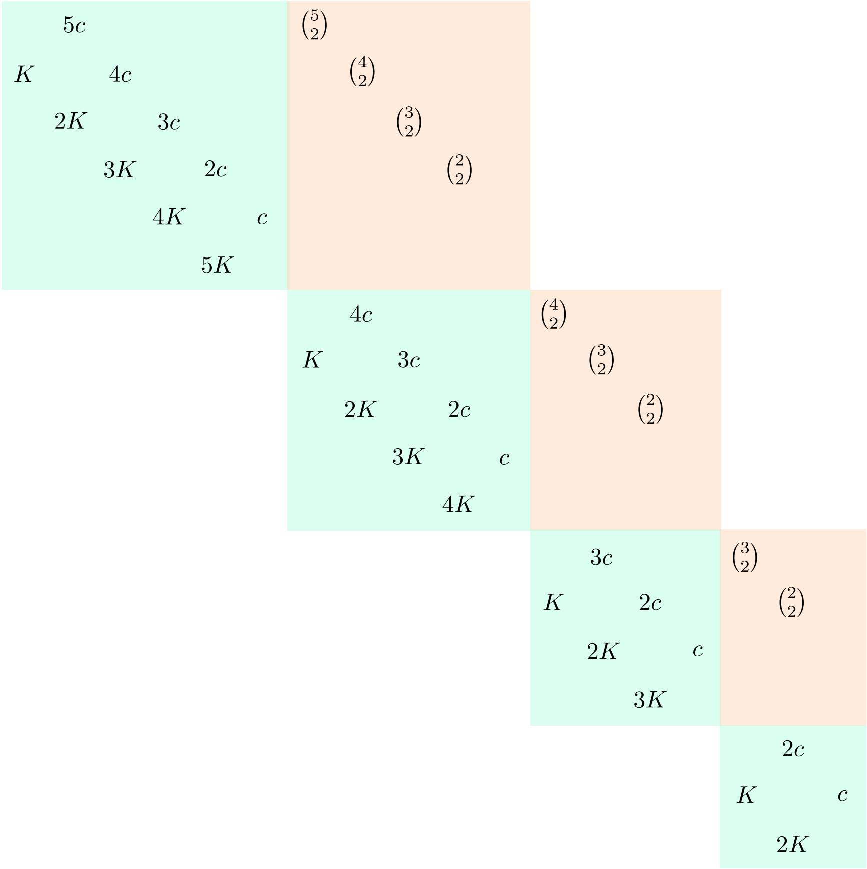

Figure 4.7: Block matrix structure for the seed-bank coalescent for the case $i = 5$. The green sub-matrices contain rates corresponding to genes entering or exiting a dormant state. The orange sub-matrices contain rates corresponding to coalescent events which follow the dynamics of Kingman's coalescent. Note that all empty elements (besides the diagonal entries) equal zero.

- At rate $\binom{5}{2}$, any two of the 5 currently active genes will coalesce.

1: Row 1 corresponds to the state where there is $j = 1$ in-active gene in the population, and the remaining genes are active. It contains the intensities

$$\begin{bmatrix} \color{red}{K} & \color{red}{-\lambda_{4,1}} & \color{red}{4c} & 0 & 0 & 0 & 0 & \color{blue}{\binom{4}{2}} & 0 & 0 & 0 \end{bmatrix}$$

where

- At rate K, the single in-active gene will re-activate,
- At rate $4c$, one of the 4 currently active genes will become in-active, and
- At rate $\binom{4}{2}$, any two of the 4 currently active genes will coalesce.

2: Row 2 corresponds to the state where there are $j = 2$ in-active genes in the population, and the remaining genes are active. It contains the intensities

$$\begin{bmatrix} 0 & 2K & -\lambda_{3,2} & 3c & 0 & 0 & 0 & 0 & \binom{3}{2} & 0 & 0 \end{bmatrix}$$

where

- At rate $2K$, one of the presently in-active genes will re-activate
- At rate $3c$, one of the presently active genes will become dormant, and
- At rate $\binom{3}{2}$, any two of the three active genes will coalesce.

3: Row 3 corresponds to the state where there are $j = 3$ in-active genes in the population, and the remaining genes are active. It contains the intensities

$$\begin{bmatrix} 0 & 0 & 3K & -\lambda_{2,3} & 2c & 0 & 0 & 0 & 0 & \binom{2}{2} & 0 \end{bmatrix}$$

where

- At rate $3K$, one of the three presently in-active genes will re-activate
- At rate $2c$, one of the two presently active genes will become dormant, and
- At rate $\binom{2}{2} = 1$, the two active genes will coalesce.

4: Row 4 corresponds to the state where there are $j = 4$ in-active genes in the population, and one gene remains active. It contains the intensities

$$\begin{bmatrix} 0 & 0 & 0 & 4K & -\lambda_{1,4} & c & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

where

- At rate $4K$, one of the four presently in-active genes will re-activate
- At rate c, the one active gene will become dormant.

5: Row 5 corresponds to the state where there are $j = 5$ in-active genes in the population (all genes are in-active). It contains the intensities

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 5K & -\lambda_{0,5} & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

where

- at rate $5K$, one of the five in-active genes will re-activate.

When an event of coalescence occurs the population size decreases by one, and so the Markov process transitions in to the second block of sub-matrices

$$[\Gamma(4) \quad D(4)],$$

where it behaves in the same way as it did within the first block. Similarly, after an event of coalescence occurs, the process transitions in to the third block of sub-matrices

$$[\Gamma(3) \quad D(3)].$$

Again, after an event of coalescence the Markov process then transitions in to the final block,

$$[\Gamma(2)],$$

which is the final matrix of sub-states until the process reaches its absorbing state. That is, until the genes have found their MRCA.

Now that we have described the $n = 5$ case, let us describe the matrix structure for a general n. It is important to note that we may consider the case where our initial population contains $j \geq 1$ in-active genes, according to our initial distribution vector, $\boldsymbol{\pi}$. Regardless of the size of our total initial population, however, we begin our jump process in the first block of sub-matrices,

$$[\Gamma(n) \quad D(n)]$$

. The number of active genes within the initial population, which is specified by $\boldsymbol{\pi}$, is what determines the row of this first block in which the jump process starts.

So long as our total population remains with size n, we remain jumping between the transient states of this first block,

$$[\Gamma(n)].$$

When an event of coalescence occurs, our population size decreases by one. This is when our Markov chain jumps to the next set of states, that is, the second row of block matrices,

$$[\Gamma(n-1) \quad D(n-1)].$$

We again remain transitioning between the transient states,

$$[\Gamma(n-1)].$$

until a coalescence event occurs. We will then transition to the next row of block matrices.

The process continues in this way, until we reach the final block matrix

$$D(2).$$

From this set of states, the Markov process is absorbed. That is, the n genes have found their MRCA.

The tree height, τ_n, has phase-type distribution $\text{PH}(\boldsymbol{\pi}, B)$ where $\boldsymbol{\pi} = (1, 0, \ldots, 0)$ and B is as defined above in Equation 4.9. This quantity is phase-type distributed because as we mentioned, the single absorbing state represents the generation at which the most recent common ancestor is found, and all other states are transient.

4.5.2 Theoretical results

Using our results from phase-type theory we can again compute descriptors for the tree height and total branch length for the seed-bank case. We have expected tree height,

$$\mathbb{E}[\tau_n] = \boldsymbol{\pi}(-B^{-1})\mathbf{1},$$

and total branch length

$$\mathbb{E}[L_n] = \boldsymbol{\pi}(-B^{-1})\boldsymbol{r},$$

where our reward vector is denoted

$$\boldsymbol{r} = (n, n-1, \ldots, 2, n-1, n-2, \ldots, 2, \ldots, \ldots, 3, 2)^{\intercal}.$$

Results for for tree height and total branch lengths such that we equal rates of dormancy and re-activation, $c = K = 1$ for varying initial population size n, can be seen in Figure 4.8.

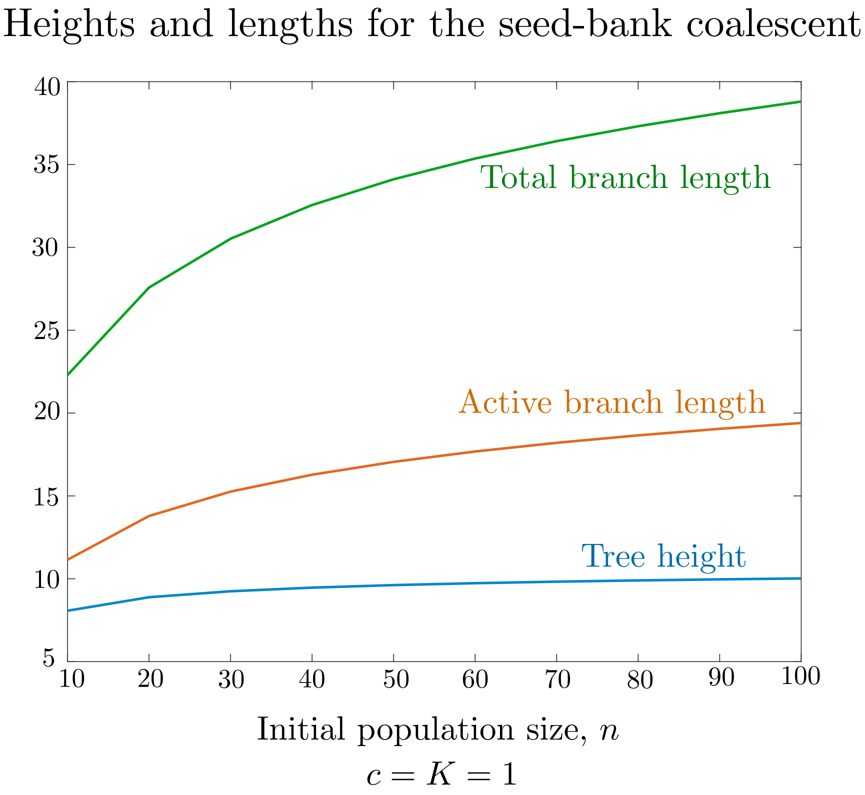

Figure 4.8: Expected tree height and total branch length for the seed-bank coalescent where $c = K = 1$. The active branch length contributes to approximately half of the total branch length, whilst tree height reaches a limit at 10.

Expected tree height demonstrates asymptotic behaviour, and tends to approximately 10. The active branch length makes up about a third of the total branch length, for all values of n. We also investigate heights and lengths for differing rate parameters c and K, which is shown in Figure 4.9.

Referring to Figure 4.9, we notice that when $c << K$, the branch length for the active genes is much closer to that of the total branch length. This is since a large proportion of the total branch length is contributed to by the active genes, and only a small proportion is from the dormant genes. This result is expected since for $c << K$, the number of dormant genes in the population will consistently be much less than the number of active ones. On the other hand, if $c >> K$, a large proportion of the population will consist of dormant genes. In this case, the active branch length is expected to be much shorter. This is because a greater proportion of our total branch length is due to the increasing length of the primarily dormant lineages.

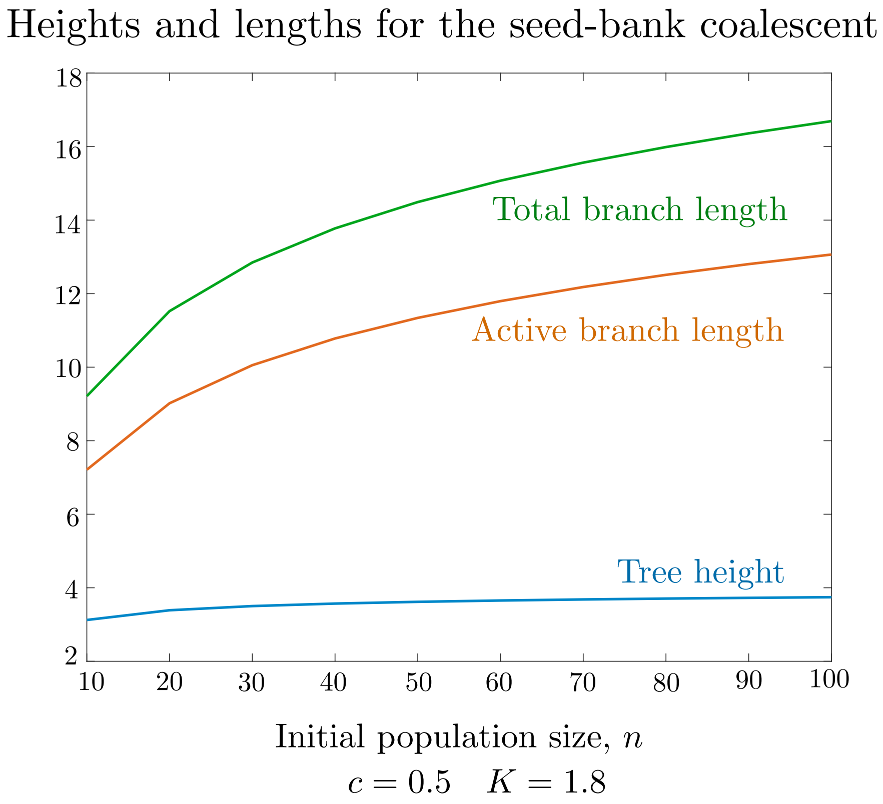

Figure 4.9: Seed-bank coalescent for the case where $c = 0.5$, $K = 1.8$. We notice that the active branch length accounts for most of the total branch length, and tree height remains approximately constant regardless of initial population size, n. The branch length increases as population size increases.

Chapter 5

Two-island model

A new phase-type structure has been developed to model the two-island model. An outline of this new formulation is seen in Section 5.2, where we introduce a sub-intensity matrix for tree height using a block-matrix structure, which describes the dynamics of a two-island model. New results for tree height and total branch length are presented in Section 5.3, along with calculations for the correlation coefficient between the two total branch lengths.

5.1 Motivation

The n-island model was developed by Sewall Wright [27] and models the interactions between n populations that have a social structure. An n-island structure means that a total population is divided in to n disjoint groups, and these groups of individuals (or genes) can interact with one another through migration. It has been studied by various researchers since its development, including A. Arredondo (2021) [1] where it has been used to infer demographic history of genes.

Since the island model assumes social structure, it is more accurately representative of a true population in comparison to the Wright-Fisher model which assumes no population structure. Most species display population structure to some extent, which is why the n-island model is more useful in such instances. That is, genes reproduce with other individuals that are in physical proximity. In humans, for example, the probability of mating within the same continent is larger than mating between them. On the other hand, for plant species, the probability of pollination is larger for individual plants close together than for individual plants at opposite ends of the field [11].

Throughout the following section will consider a coalescent process such that the underlying genealogy follows the two-island model. By looking backwards-in-time we therefore consider coalescent events as well as migration events. We will denote our populations as Island 1 and Island 2. We define the following rate parameters:

- M, rate at which a single gene migrates from Island 1 to Island 2,
- R, the rate at which a single gene migrates from Island 2 to Island 1,
- α, the rate at which a single gene can coalesce on Island 1,
- β, the rate at which a single gene can coalesce on Island 2,

and we suppose $i = I_1 + I_2$ represents total population size, where I_1 is the population size of Island 1, and I_2 is the population size of Island 2. Figure 5.1 demonstrates the dynamics of such a population.

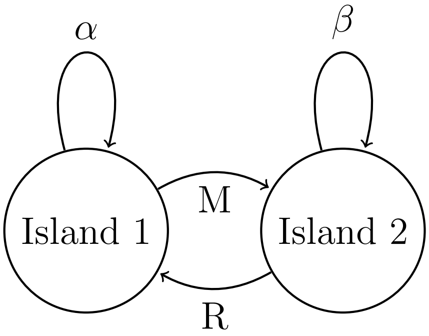

Figure 5.1: State diagram for the two-island population structure. Rates M and R correspond to migration, and rates α and β correspond to coalescence.

We will discuss the structure of a two-island coalescent model by firstly looking at an example. Figure 5.2 demonstrates the backwards-in-time genealogy of such a process such that we begin with an initial population of $i = 6$, assuming that 3 genes are initially present on each island. Island 1 (left) initially holds genes $1, 2$ and 3. Island 2 (right) initially holds genes $4, 5$ and 6. The islands are separated visually by the blue vertical dotted line through the centre of the figure. The coalescent process for this genealogy is as follows:

1. The population begins with $i = 6$ genes such that $I_1 = 3$ and $I_2 = 3$.
2. Gene 3 migrates to Island 2
3. Gene 1 migrates to Island 2
4. At time T_6, genes 5 and 6 coalesce
5. Gene $(5, 6)$ migrates to Island 1
6. At time $T_6 + T_5$, genes 3 and 4 coalesce
7. Gene $(3, 4)$ migrates to Island 1

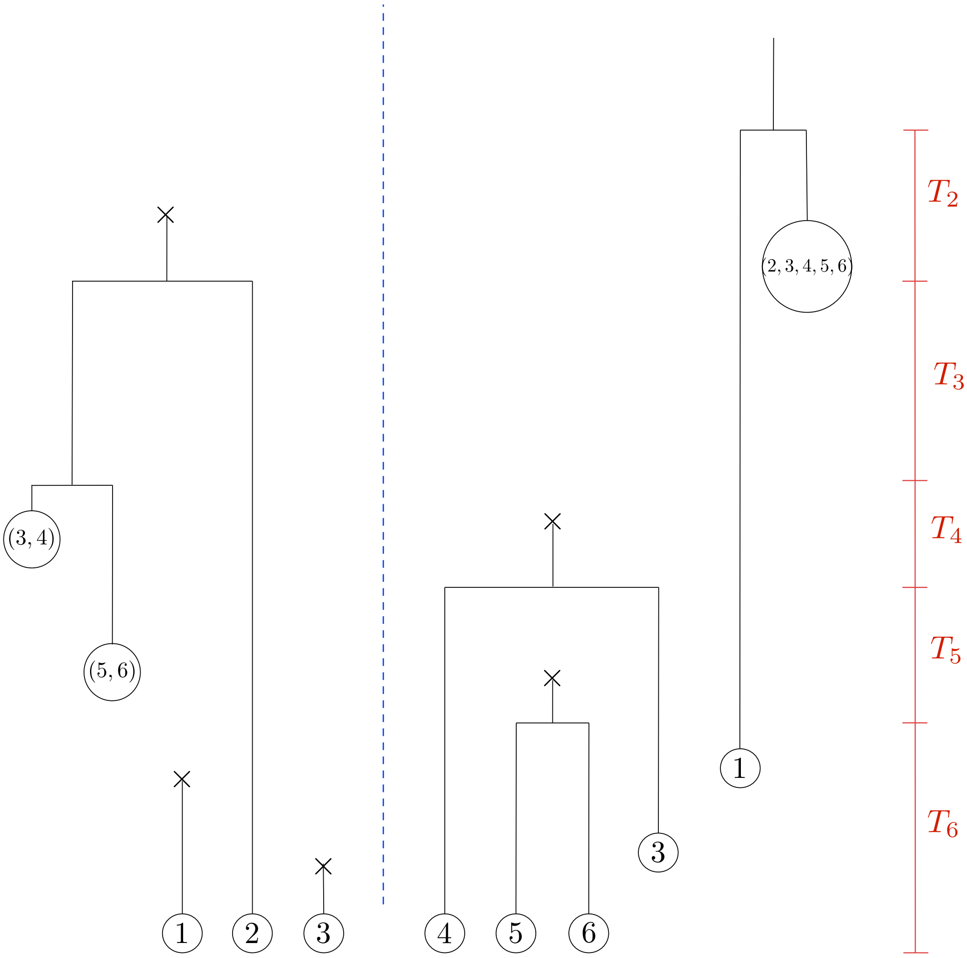

Figure 5.2: A sample construction for a population of $i = 6$ genes total where $I_1 = 3$ and $I_2 = 3$. The × symbols represent the time of a migration event.

8. At time $T_6 + T_5 + T_4$, genes $(3, 4)$ and $(5, 6)$ coalesce
9. At time $T_6 + T_5 + T_4 + T_3$, genes 2 and $(3, 4, 5, 6)$ coalesce
10. Gene $(2, 3, 4, 5, 6)$ migrates to Island 2
11. Finally, at time $\tau = T_6 + T_5 + T_4 + T_3 + T_2$, genes 1 and $(2, 3, 4, 5, 5)$ coalesce and the MRCA as been found

The tree height is a quantity that is shared over both islands. For an initial population of size $i = n$, this quantity is again defined as $\tau_n = T_n + T_{n-1} + \ldots + T_2$. Total branch length, on the other hand, is calculated separately for each island. We re-introduce the reward concept, which will be required to compute this quantity later.

We can visualise Figure 5.2 in a more straight forward way using colours, and our newly constructed version is shown in Figure 5.3. The orange lineages represent the genes on Island 1, and the blue lineages represent those on Island 2. This means only genes of the same colour can coalesce. The moment at which a lineage changes colour represents the time at which the gene migrates.

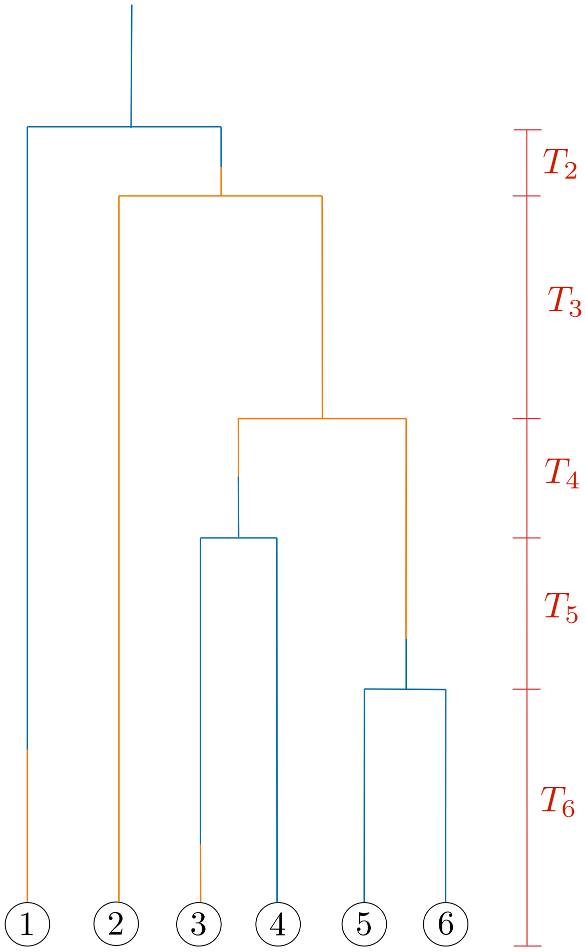

Figure 5.3: A coalescent tree representing the same genealogical process from Figure 5.2 where initially, $i = 6$, $I_1 = 3$ and $I_2 = 3$.

5.2 Phase-type representation

A new phase-type representation has been constructed for the tree height (time until the MRCA) of the two-island model. Throughout this chapter we will denote the states of our underlying CTMC as

$$(I_1, I_2)$$

where I_1 and I_2 represent the number of genes present on Island 1 and Island 2, respectively. We denote total population size as i, where $i = I_1 + I_2$. We again, as in the case for the seed-bank coalescent model, require the use of block matrices to define the sub-intensity matrix for tree height. Using appropriate rewards, which we will introduce next, we extend our calculations to include total branch lengths. Recall that within the seed-bank coalescent, the reward corresponds to the number of active genes in the population. Similarly, within the two-island case, the reward value corresponds to the number of genes present on the island. For example, if the system is in state $(I_1, I_2) = (a, b)$ then the corresponding rewards for each island, r_1 and r_2, are $(r_1, r_2) = (a, b)$. We consider a reward vector for each island, as defined in Table 5.1.

State	Island 1 rewards	Island 2 rewards
$(n, 0)$	n	0
$(n-1, 1)$	$n-1$	1
$(n-2, 2)$	$n-2$	2
$\vdots$	$\vdots$	$\vdots$
$(1, n-1)$	1	$n-1$
$(0, n)$	0	n
$(n-1, 1)$	$n-1$	1
$(n-2, 2)$	$n-2$	2
$(n-3, 3)$	$n-3$	3
$\vdots$	$\vdots$	$\vdots$
$(2, n-2)$	2	$n-2$
$(1, n-1)$	1	$n-1$
$(n-2, 2)$	$n-2$	2
$(n-3, 3)$	$n-3$	3
$(n-4, 4)$	$n-4$	4
$\vdots$	$\vdots$	$\vdots$
$(3, n-3)$	3	$n-3$
$(2, n-2)$	2	$n-2$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
$(2, 0)$	2	0
$(1, 1)$	1	1
$(0, 2)$	0	2

Table 5.1: States of the two-island model and their corresponding rewards.

Now, to construct our sub-intensity matrix for tree height, we firstly define the $(i+1) \times (i+1)$

matrix

$$\Gamma(i) = \begin{pmatrix} -\lambda_{i,0} & iM & & & & \\ R & -\lambda_{i,1} & (i-1)M & & & \\ & 2R & -\lambda_{i,2} & (i-2)M & & \\ & & \ddots & \ddots & \ddots & \\ & & & & & \\ & & & & & \\ & & & & -\lambda_{i,i-1} & M \\ & & & & iR & -\lambda_{i,i} \end{pmatrix} \tag{5.1}$$

along with the $(i+1) \times i$ matrix

$$D(i) = \begin{pmatrix} \alpha\binom{i}{2} & & & & \\ & \alpha\binom{i-1}{2} & & & \\ & \beta\binom{2}{2} & \alpha\binom{i-2}{2} & & \\ & & \beta\binom{3}{2} & \alpha\binom{i-3}{2} & \\ & & & \beta\binom{4}{2} & \ddots \\ & & & & \ddots \\ & & & & \\ & & & & 0 \end{pmatrix} \tag{5.2}$$

where the parameters M, R, α and β are as defined in Section 5.1. We fit our $\Gamma(i)$ and $D(i)$ matrices together to represent the sub-intensity matrix for tree height in the same way as for the seed-bank coalescent. We thus have that

$$B(i) = \begin{pmatrix} \Gamma(i) & D(i) & & & \\ & \Gamma(i-1) & D(i-1) & & \\ & & \Gamma(i-2) & D(i-2) & \\ & & & \ddots & \ddots \\ & & & & \\ & & & & D(3) \\ & & & & \Gamma(2) \end{pmatrix}$$

where

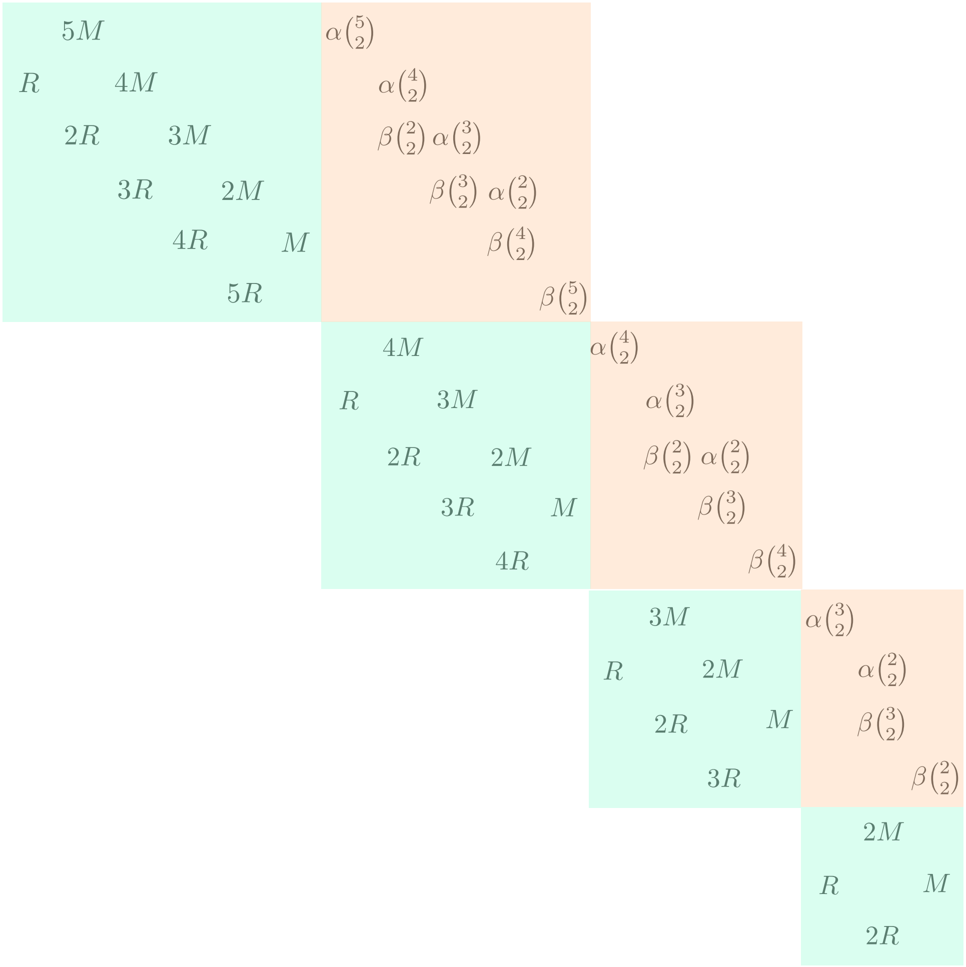

Figure 5.4: The sub-intensity matrix B for tree height for the two-island model. The green blocks represent the Γ sub-matrices and the orange represents the D sub-matrices.

(1) Row 1 corresponds to the states whereby there are $i = n$ genes in the total population,

(2) Row 2 corresponds to the states whereby there are $i = n-1$ genes in the total population, and

⋮

(k) Row k corresponds to the state whereby there are $i = n-k$ genes in the total population.

Figure 5.4 demonstrates the expanded version of this matrix for the case where there are $i = 5$ genes in the total population initially.

We elaborate on this matrix structure by looking at the expanded form such that $i = 5$. Note that B is a sub-intensity matrix, so the diagonal elements are

$$-\lambda_{i,I_2} = -\left(\binom{i-I_2}{2}\alpha + \binom{I_2}{2}\beta + (i-I_2)M + I_2R\right)$$

for $I_2 = 0, 1, \dots, i$. Note that these diagonal elements are not pictured in 4.7. The first row of sub-matrices which we denote

$$[\Gamma(5) \quad D(5)] = \begin{bmatrix} -\lambda_{5,0} & 5M & 0 & 0 & 0 & 0 & \alpha\binom{5}{2} & 0 & 0 & 0 & 0 \\ R & -\lambda_{4,1} & 4M & 0 & 0 & 0 & 0 & \alpha\binom{4}{2} & 0 & 0 & 0 \\ 0 & 2R & -\lambda_{3,2} & 3M & 0 & 0 & 0 & \beta\binom{2}{2} & \alpha\binom{3}{2} & 0 & 0 \\ 0 & 0 & 3R & -\lambda_{2,3} & 2M & 0 & 0 & 0 & \beta\binom{3}{2} & \alpha\binom{2}{2} & 0 \\ 0 & 0 & 0 & 4R & -\lambda_{1,4} & M & 0 & 0 & 0 & \beta\binom{4}{2} & 0 \\ 0 & 0 & 0 & 0 & 5R & -\lambda_{0,5} & 0 & 0 & 0 & 0 & \beta\binom{5}{2} \end{bmatrix}$$

contains the transition rates for our system such that the total population starts and remains with size 5. We will explain the states of this sub-matrix row-by-row, again starting from row 0.

0: Row 0 corresponds to the state where there are $I_2 = 0$ genes on Island 2 (all genes are in Island 1). It contains the intensities

$$\begin{bmatrix} -\lambda_{5,0} & 5M & 0 & 0 & 0 & 0 & \alpha\binom{5}{2} & 0 & 0 & 0 & 0 \end{bmatrix}$$

where

- At rate $5M$, one of the 5 genes on Island 1 will migrate to Island 2, and
- At rate $\alpha\binom{5}{2}$, any two of the 5 genes on Island 1 will coalesce.

1: Row 1 corresponds to the state where there is $I_2 = 1$ gene on Island 2, and the remaining genes are on Island 1. It contains the intensities

$$\begin{bmatrix} R & -\lambda_{4,1} & 4M & 0 & 0 & 0 & 0 & \alpha\binom{4}{2} & 0 & 0 & 0 \end{bmatrix}$$

where

- At rate R, the single on Island 2 will migrate to Island 1,
- At rate $4M$, one of the 4 genes on Island 1 will migrate to Island 2, and
- At rate $\alpha\binom{4}{2}$, any two of the 4 genes on Island 1 will coalesce.

2: Row 2 corresponds to the state where there are $I_2 = 2$ genes on Island 2, and the remaining genes are on Island 1. It contains the intensities

$$\begin{bmatrix} 0 & 2R & -\lambda_{3,2} & 3M & 0 & 0 & 0 & \beta\binom{2}{2} & \alpha\binom{3}{2} & 0 & 0 \end{bmatrix}$$

where

- At rate $2R$, one of the two genes on Island 2 will migrate to Island 1,
- At rate $3M$, one of the three genes on Island 1 will migrate to Island 2,
- At rate $\beta\binom{2}{2}$, the two genes on Island 2 will coalesce, and
- At rate $\alpha\binom{3}{2}$, any two of the three active genes on Island 1 will coalesce.

3: Row 3 corresponds to the state where there are $I_2 = 3$ genes on Island 2. It contains the intensities

$$\begin{bmatrix} 0 & 0 & 3R & -\lambda_{2,3} & 2M & 0 & 0 & 0 & \beta\binom{3}{2} & \alpha\binom{2}{2} & 0 \end{bmatrix}$$

where

- At rate $3R$, one of the three genes in Island 2 will migrate to Island 1,
- At rate $2M$, one of the two genes on Island 1 will migrate to Island 2,
- At rate $\beta\binom{3}{2}$, two of the three genes on Island 2 will coalesce, and
- At rate $\alpha\binom{2}{2} = \alpha$, the two genes on Island 1 will coalesce.

4: Row 4 corresponds to the state where there are $I_2 = 4$ genes on Island 2, and one gene on Island 1. It contains the intensities

$$\begin{bmatrix} 0 & 0 & 0 & 4R & -\lambda_{1,4} & M & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

where

- At rate $4R$, one of the four genes on Island 2 will migrate to Island 1,
- At rate M, the one gene on Island 1 will migrate to Island 2, and
- At rate $\beta\binom{4}{2}$, any two of the four genes on Island 2 will coalesce.

5: Row 5 corresponds to the state where there are $I_2 = 5$ genes on Island 2 (all genes are on Island 2). It contains the intensities

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 5M & -\lambda_{0,5} & 0 & 0 & 0 & 0 & \beta\binom{5}{2} \end{bmatrix}$$

where

- at rate $5R$, one of the five genes on Island 2 will migrate to Island 1in-active genes will re-activate, and
- at rate $\beta\binom{5}{2}$, any two of the five genes on Island 2 will coalesce

When an event of coalescence occurs the population size decreases by one, and so the Markov process transitions in to the second block of sub-matrices

$$[\Gamma(4) \quad D(4)],$$

where it behaves in the same way as it did within the first block. Similarly, after an event of coalescence occurs, the process transitions in to the third block of sub-matrices

$$[\Gamma(3) \quad D(3)].$$

Again, after an event of coalescence the Markov process then transitions in to the final block,

$$[\Gamma(2)],$$

which is the final matrix of sub-states until the process reaches its absorbing state. That is, until the genes have found their MRCA.

Now that we have described the case such that $i = 5$, let us describe the matrix structure for a general initial population size, $i = n$. Note that we may study at this model for any initial population sizes for Island 1 and Island 2. From now on we assume that $I_1 = I_2$ initially. This condition is specified by our initial distribution vector, $\boldsymbol{\pi}$, which in this case is a vector *of zero vectors*. We define $\boldsymbol{\pi} = (\boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \dots, \boldsymbol{\pi_p})$, where we have that $\boldsymbol{\pi_1}$ be a vector of zeros besides the $\left(\frac{n}{2}+1\right)^{th}$ element, which will equal 1. For example, in the case of $i = 6$, we have that $\boldsymbol{\pi} = (\boldsymbol{\pi_1}, ..., \boldsymbol{\pi_5})$. To specify that $I_1 = I_2$ initially, we let $\boldsymbol{\pi_1} = (0, 0, 0, 1, 0, 0, 0)$ and $\boldsymbol{\pi_i} = \mathbf{0}$ for $i \in \{2, 3, 4, 5\}$.

The Markov jump process always starts in the first block of sub-matrices,

$$[\Gamma(n) \quad D(n)].$$

. It is our initial distribution vector $\boldsymbol{\pi}$ that specifies which row of this first block that the jump process starts. As long as the total population remains with size $i = n$ (whilst no coalescent events have occurred), we remain jumping between the transient states of this first block,

$$[\Gamma(n)].$$

When an event of coalescence occurs, whether it be on Island 1 or Island 2, our population size decreases by one. This is when our Markov chain jumps to the next set of states (the second row of block matrices),

$$[\Gamma(n-1) \quad D(n-1)].$$

We again remain transitioning between the transient states,

$$[\Gamma(n-1)].$$

until a coalescence event occurs on either island. We will then transition to the next row of block matrices. The process continues in this way, until we reach the final block matrix

$$D(2).$$

It is from this set of states that the Markov process is absorbed, that is, when the total $i = I_1 + I_2$ genes have found their MRCA. By the diagonal structure of this matrix we can see that our jump process only transitions forward. The Markov chain will eventually be absorbed. The state in which a common ancestor of the two populations, I_1 and I_2, is found corresponds to the absorbing state.

5.2.1 Theoretical results

We can again study *tree height* and *total branch length* of this genealogy. In addition, with the use of multivariate phase-type theory, we are able to look at the dependence between these islands by calculating the *correlation* of their total branch lengths. This quantity is given by

$$\mathrm{corr}(L_1, L_2) = \frac{\mathrm{cov}(L_1, L_2)}{\sqrt{\mathrm{var}(L_1)\mathrm{var}(L_2)}},$$

where

- $\mathrm{cov}(L_1, L_2) = \mathbb{E}[L_1 L_2] - \mathbb{E}[L_1]\mathbb{E}[L_2]$,
- $\mathbb{E}[L_1 L_2] = \boldsymbol{\pi}(-A^{-1})\Delta(R_{.i})(-A^{-1})(R_{.j}) + \boldsymbol{\pi}(-A^{-1})\Delta(R_{.j})(-A^{-1})R_{.i}$, and
- $\mathrm{var}(L_1, L_2) = \mathbb{E}[L_i^2] - \mathbb{E}[L_i]^2$ for $i = 1, 2$.

We are interested in the tree height, total branch length and correlation for varying parameter values, α, β, M and R. The first case we look at is such that the migration rates and coalescent rates are equal, $M = R = \alpha = \beta = 1$. Tree height, total branch length and correlation are shown in Figure 5.5.

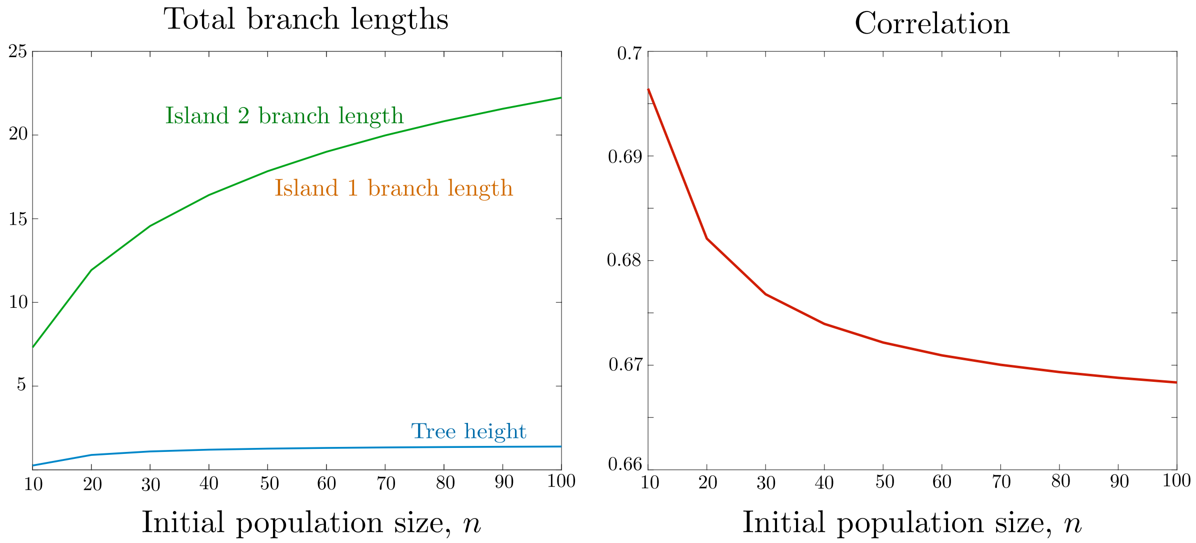

Figure 5.5: Left: Expected tree height and expected total branch lengths for Island 1 and Island 2. The total branch lengths for each island are equal, and the tree height approaches a limit. Right: Correlation between total branch lengths for $M = R = \alpha = \beta = 1$. Correlation decreases slightly as population size increases.

The total branch lengths in the case where $M = R = \alpha = \beta$ are equal for each island. This result is expected, since here the dynamics on Island 1 and Island 2 are exactly the same. There is nothing differing their genealogies and so their branch lengths do not vary from one another. The correlation between the total branch lengths of each island is moderate, as it tends to approximately 0.65.

We can also look at the scenario in which one of the islands has a higher rate of coalescence. Suppose that Island 1 has a higher rate of coalescence, that is, $\alpha >> \beta$, and the migration parameters are fixed. We set $\alpha = 2.8$, $\beta = 0.3$ and $M = R = 1$ and plot these results in Figure 5.6.

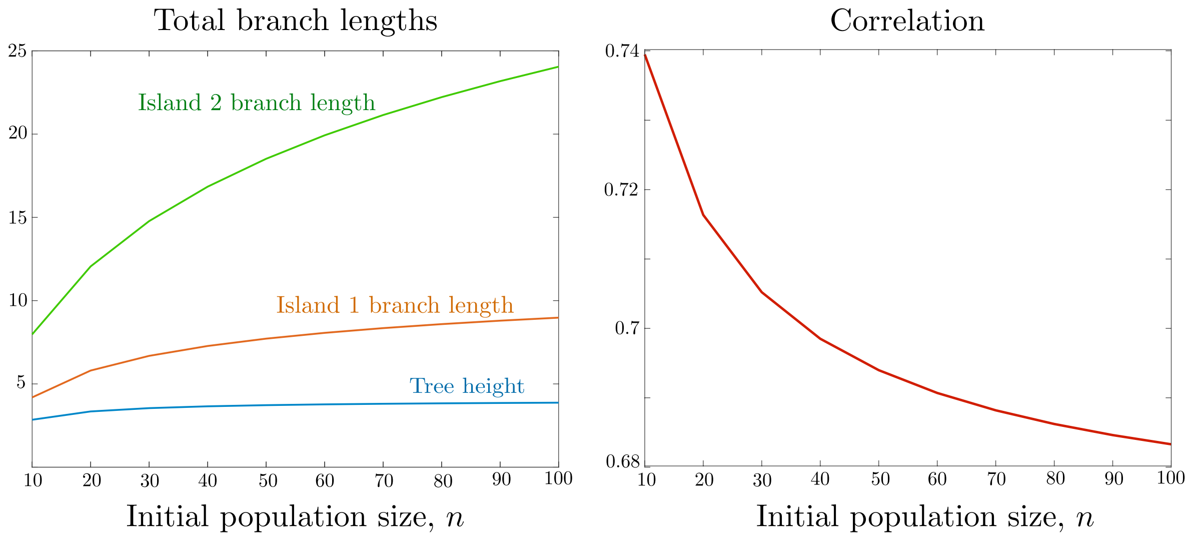

Figure 5.6: Final plots comparing branch lengths and tree height, and correlation for branch lengths for $M = R$, $\alpha >> \beta$. We have $M = R = 1$ and $\alpha = 2.8$, $\beta = 0.3$.

and the case where $M >> R$, shown in Figure 5.7. We notice a difference in branch lengths for this case. Island 2 has a branch length that is consistently more than twice the length of the Island 1 branch length. More specifically, we see that the higher rate of coalescence leads to a much shorter branch length. From a genealogical perspective, this is because the genes on Island 1 are locating their ancestors at a much higher rate, and on Island 2 it takes genes much longer to locate their common ancestors. We can also look at the case where the coalescent parameters are fixed and migration rates are changed, as shown in Figure 5.7.

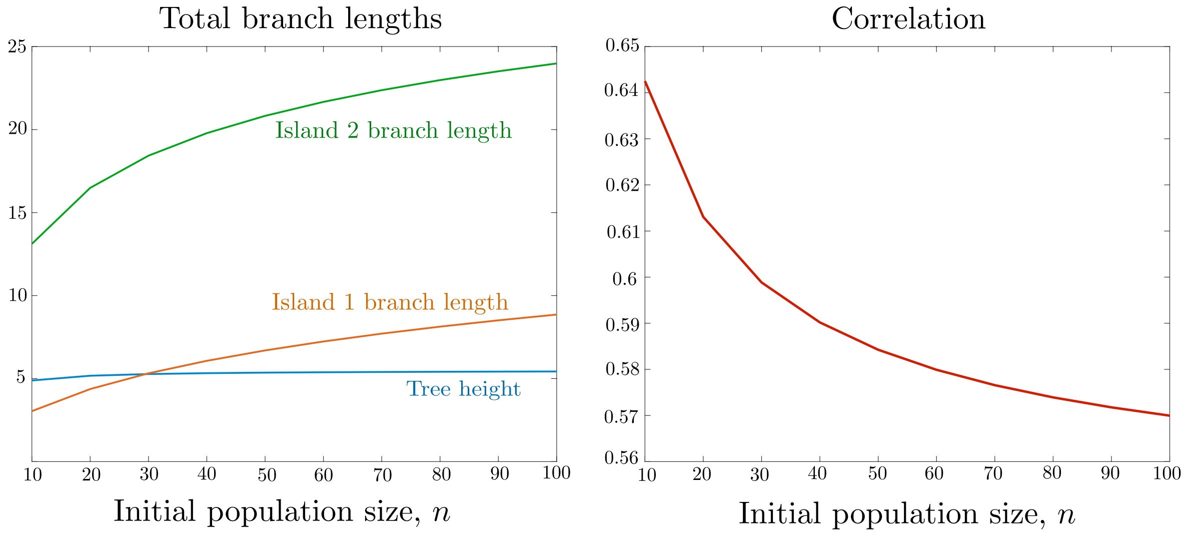

Figure 5.7: Final plots for correlation and branch length comparisons for $M >> R$, $\alpha = \beta$. Here, $\alpha = \beta = 0.5$ and $M = 3.5$,$R = 0.5$

Correlation between the total branch lengths has slightly more variation for the case where $M >> R$, in comparison to when $\alpha << \beta$. More specifically, the correlation decreases at a faster rate as the initial population size increases. It seems as if the migration parameters, have a greater impact on whether or not the branch lengths are correlated in contrast to the coalescent parameters. We are interested in investigating this result further, and do so by plotting the correlation coefficient for varying parameter values.

We firstly look at the correlation between total branch length such that M and R are varied for fixed $\alpha = \beta = 1$. Correlation between total branch lengths on each island is maximised when the two migration rates are equal. This can be seen in Figure 5.8, where notice that the correlation curves are at their peak when $M = R$. The mixing of the two islands is the most efficient when their migration rates are equal, and so this finding makes intuitive sense. The systems are essentially 'in balance', and the two-island structure is the least prominent at this instance. The population is most closely representative of a single island model.

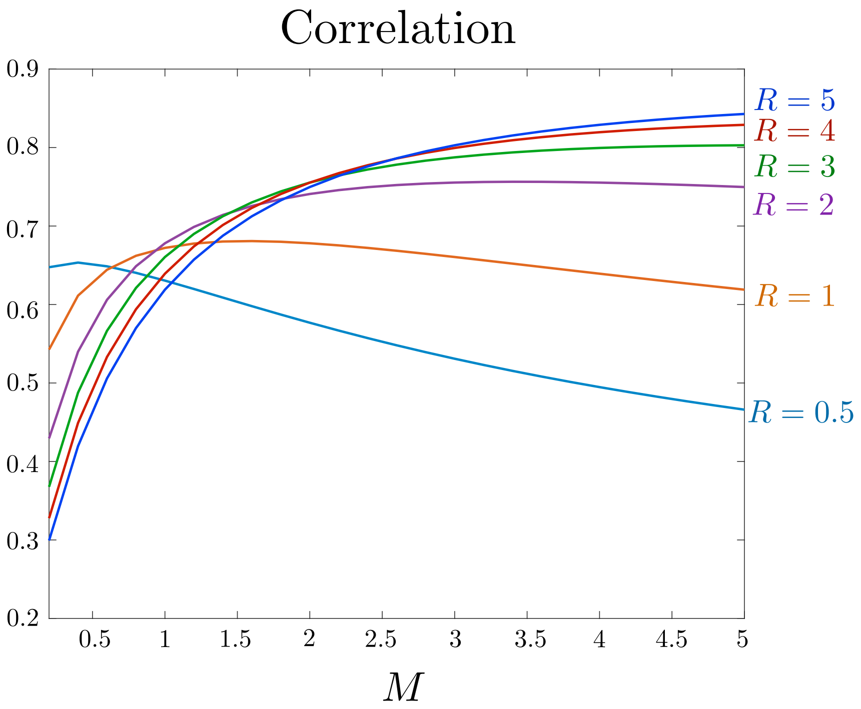

Figure 5.8: Correlation between L_1 and L_2 for varying migration parameters M and R and fixed $\alpha = \beta = 1$, such that $I_1 = I_2 = 25$ initially.

We can also look at the correlation coefficient of total branch lengths such that α and β are both varied, as shown in Figure 5.9.

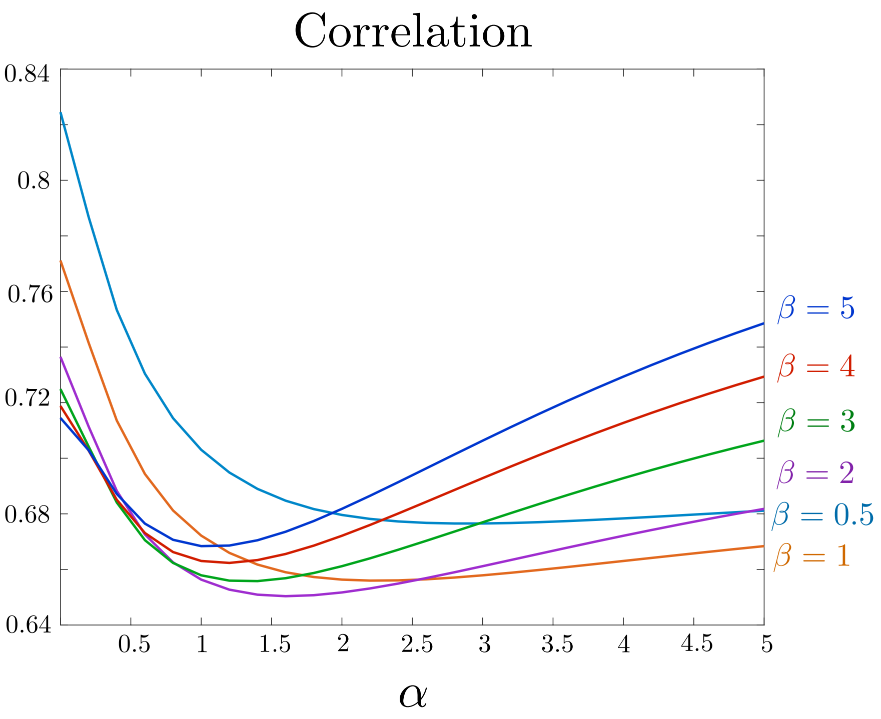

Figure 5.9: Correlation for total branch lengths of each island such that we vary the rate parameters α and β. Here, $M = R = 1$ such that there are $I_1 = I_2 = 25$.

When $\alpha = 0$, coalescent events will only occur on Island 2. In this case, once a coalescent event occurs, there will be a relatively increased rate of migration into Island 2. At this point, migration will somewhat append the disparity in population sizes between the two islands. We also see that the minimum correlation depends on the value of β. Interestingly, as α passes this minimum, we see an increase in correlation which continues for all observed values up to $\alpha = 5$. On the other hand, for the case where $\beta = 0.5$, the correlation appears to reach a limit rather than increase.

We have observed that correlation between total branch lengths has a greater range when the migration parameters vary, in contrast to when the coalescent rates vary. That is, correlation is more sensitive to migration events than it is to coalescent events. This is demonstrated by the wide span of correlations seen in Figure 5.8, in comparison to the smaller range of correlation values in Figure 5.9.

Chapter 6

Future work and summary

Some further extensions into phase-type theory and coalescent models are mentioned by Hobolth et. al. (2019) [13], including the suggestion of more sophisticated statistical inference methods for genetic data using the phase-type framework. Another mentioned research topic was the isolation-with-migration model with multiple populations (Hey, (2010) [12]), in which statistical inference is known to be challenging. Phase-type theory is proposed as a tool to overcome such statistical difficulties. Since we have constructed the phase-type representation for the two-island case, we will discuss some potential extensions to this representation including

1. the incorporation of a larger island structure,

2. more complex coalescent dynamics, which could follow the ψ or β coalescent dynamics, and finally

3. block migration structure, whereby large migration events can occur.

The phase-type representation for the tree height of the island model can be generalised further to include a population n islands, for $n > 2$. To account for each additional island, we will need to add another dimension to the sub-intensity matrix. That is, our block matrix structure would contain block matrices in itself. It is evident that this would be a much more complex construction than the two-island case, and due to time constraints, could not be investigated throughout this manuscript. Once the phase-type representation is constructed, however, descriptors of genealogical quantities can be computed as they have been within this thesis, due to their explicit expressions in the phase-type framework.

Alternatively, or as well as adding more islands in to the population structure, we may consider coalescent events that do not necessarily follow the dynamics of the Kingman's coalescent. The possibility of large coalescent events could be incorporated in to the island model, perhaps following the dynamics of a ψ or β-coalescent, for example. In fact, we could also consider each island to follow different coalescent dynamics. That is,

- some islands have a non-zero probability of recurring sweepstake reproduction events whilst others do not, or
- all islands may follow a ψ or β-coalescent, but with varying parameters of ψ and β, respectively.

A third extension suggestion focuses on varying the migration dynamics. Within this framework, we may consider large migration events in the same way as we previously considered large coalescent events. In other words, since we can consider sweepstake reproduction events, we could also consider sweepstake migration. What if the structure of the ψ and β-coalescence was not applied to only coalescent events, but migration events as well?

6.1 Summary

We have demonstrated how phase-type theory provides a useful framework for calculating distributions and summary statistics for models within population genetics. As well as presenting a comprehensive self-contained description of phase-type theory, we have provided a well-rounded theory and analysis for a variety of well known coalescent models. We have developed a new phase-type representation for the two-island model, and are therefore able to infer characteristics of a two island population structure without the requirement of simulations. The ancestry, and thus the evolution of such a population can be better understood by computing not only the tree height and total branch length of these genealogies, but also the dependence between them. We find that using multivariate phase-type theory techniques, the dependence of branch lengths of two populations is particularly sensitive to migration rates, in comparison to rates of coalescence. This research has formed a basis for further development within the n-island coalescent model, as well as a group migration model. With the application of such phase-type techniques it has been demonstrated that, despite the present prominence of coalescent theory, it also belongs to the future.

Appendix A

The following code was produced using MATLAB, and was required throughout the development of this manuscript for simulation purposes, as well as for the computation of theoretical results. All figures throughout the manuscript were produced using output from this code. Section A.1 contains code which simulates the genealogy of a Kingman's coalescent, by implementing the structure of the underlying CTMC using the Gillespie Algorithm [9], as outlined in Chapter 3. The remaining code is used throughout chapters 4 and 5. Section A.2 is used to theoretically compute descriptors tree height and total branch length for the Kingman's coalescent by construction of appropriate sub-intensity matrices. Similarly, sections A.3 and A.4 contain code which again populates our sub-intensity matrix for tree heights for the Λ and seed-bank coalescent models. Here, we incorporate the concept of rewards to compute total branch length. Note that the code to compute quantities for the Λ-coalescent used in a general sense, meaning that the same code is used for the ψ and β-coalescent cases.

A.1 Simulating Kingman's coalescent

```
%Algorithm for Kingman's coalescent

%Initialise tree heights and branch lengths
meantreeheight = [];
meanbranchlength = [];

%Inititalising generations
generations = [];
new_generation = [];
treeheight = [];
branchlength = [];

%Iterate through population sizes
for n = 10:2:100
```

```

 %Average 1000 samples for each population size
for r = 1:1000

%Populate the original generation (generation k)
for i = 1:n
    for j = 1:n
    generations(j,i) = i;
    end
end


time_counter = 0;
T = []; %time counter
S = []; %store the times of the jumps

k = n;
index = 1;

while k > 1


    %generate waiting time until coalescence
    T_k = exprnd(1/nchoosek(k,2));

    %T vector holds times BETWEEN coalescence
    T = [T T_k];

    time_counter = time_counter + T_k;

    %S vector holds times OF coalescent events
    S = [S time_counter];

    l = datasample(generations(index,:),1);
    temp = setdiff(generations(index,:),l);
    m = datasample(temp,1); %since 1 <= l < m <= k. This line chooses random number st. its Neq to l


    locate = find(generations(index,:) == min(l,m));


    index = index + 1;

    for i = index:n
    generations(i,locate) = max(l,m);
    end

    k = k-1;



end

%Calculate tree height
height = sum(T(1:n-1));

L=0;
%for i = n:-1:2
%     L = L + i*T(i);
%end

%Calculate total branch length
L = [n:-1:2]*T';


treeheight(r) = height; %for n=10, this is approx 2*(1-1/n)
branchlength(r) = L; %for n=10,

end

%Averaging the tree heights and branch lengths
meantreeheight = [meantreeheight mean(treeheight)];
meanbranchlength = [meanbranchlength mean(branchlength)];

end

exptreeheight = [];
expbranchlength = [];

```

```
%Comparing to the theoretical
for n=10:2:100
    counter = 0;
    exptreeheight = [exptreeheight 2*(1-1/n)]

    for j=1:n-1
        counter = counter + 1/j ;
    end
    expbranchlength = [expbranchlength 2*counter]
end
```

A.2 Theoretical results for Kingman's coalescent

```
%Code to compute theoretical results for Kingman's coalescent

%Initialising
A = [];
S = [];
hts=[];
lengths=[];
Pi=[];

nn=20;
%for trials = 10:2:100
%for nn=1:trials

%A(nn-1,nn) = -1;

%Constructing our tree height and total branch length matrices, A and S
for i = 1:nn-1

    for j = 1:nn-1

        if i==j
            A(i,j) = -nchoosek(nn-i+1,2);
            S(i,j) = -nchoosek(nn-i+1,2)/(nn-i+1);
        end

        if j == i+1
            A(i,j) = nchoosek(nn-i+1,2);
```

```
                S(i,j) = nchoosek(nn-i+1,2)/(nn-i+1);
            end


        end
        Pi=[1 zeros(1, length(A)-1)];
end
%end
%hts(trials)=Pi*(-inv(A))*ones(1,length(Pi))';
%lengths(trials) = Pi*(-inv(S))*ones(1,length(Pi))';
%end
%final_heights = nonzeros(hts)
%final_branchlengths = nonzeros(lengths)
```

A.3 Theoretical results for the Λ-coalescent

```
%Code to calculate (and plot) theoretical results for Lambda coalescent
%Can be adapted for both psi and beta coalescent

%Initialising
hts = []; %tree heights
lengths = []; %total branch lengths
hts_second_moment = []; %second moment for tree height
lengths_second_moment = []; %second moment for branch length
Pi = []; %Initial distribution vector

%Parameters: psi and beta (depending which coalescent we're looking at)
%alpha=1;
psi = 0.5;

%Initial population size (number of genes)
%nn = 20;


for trials = 100:50:200 %Repeat for initial population sizes 10 to 100
for nn=1:trials

%Initialising sub-intensity matrix
```

```
A = zeros(nn-1,nn-1);

%Reward vector
R = [nn:-1:2]';
%R = diag(Rewards);

%Constructing the sub-intensity matrix given our parameters
for i = 1:nn-1

    for j = 1:nn-1-i

        A(i,i+j) = g(nn-i+1,j+1,psi);

    end
    A(i,i) = -sum(A(i,:)); %diagonal elements
    Pi=[1 zeros(1, length(A)-1)]; %initial distribution vector

end
A(i,i) = -g(2,2,psi); %bottom right element (always =-1 in Lambda case)

end

%Storing tree heights
hts(trials)=Pi*(-inv(A))*ones(1,length(Pi))';

%Storing total branch lengths
lengths(trials) = Pi*(-inv(A))*R;

%Calculating second moment for the tree height
hts_second_moment(trials) = 2*Pi*(-inv(A))^2*ones(1,length(Pi))';

end

%Final tree heights, total branch lengths, second moments for pop. sizes 10:10:100
final_heights = nonzeros(hts);
final_branchlengths = nonzeros(lengths);
final_second_moments = nonzeros(hts_second_moment);

```

```
%This function determines the value of each element (for our matrix A)
function g_ki = g(k,i,psi)

 %Rates for psi-coalescent
 g_ki = nchoosek(k,i)*psi^(i-2)*(1-psi)^(k-i); %Psi coalescent

 %Rates for beta-coalescent
 %g_ki = (nchoosek(k,i))*(beta(i-alpha,k-i+alpha))/(beta(alpha,2-alpha));
end
```

A.4 Seed-bank theoretical results

```
%Computing the theoretical results for the seed-bank coalescent
%% Parameters
results3 = [];
lambda = 1;

%Activation and dormancy rates
c = 1;
K = 1;

%Only active population can coalesce
alpha = 1;
beta = 0;

%Initialising vectors (tree heights, branch lengths, correlations)
tree_height=[];
branch_length=[];


%for N=10:10:50
N=10;
Rewards = [];
Gamma = sparse(sum(N+1:-1:3),sum(N+1:-1:3));
p = 0;

for n = N:-1:2

    %% Constructing Lambda (Gamma) matrix

    Lambda = sparse(n+1,n+1);

    for j = 1:n
        Lambda(j+1,j) = j*K;
        Lambda(j,j+1) = (n-j+1)*c;
    end

    %% Constructing D matrix (coalescent rate matrix)
    if n == 2
        Gamma(1+p:n+1+p,1+p:n+1+p) = Lambda;
    else
        D = sparse(n+1,n);

        for k = 1:n
            if k<n
            D(k,k) = alpha*nchoosek(n-k+1,2);
            end

            if k > 1
            D(k+1,k) = beta*nchoosek(k,2); %Problem here
            end

        end
        Gamma(1+p:n+1+p,1+p:n+1+p) = Lambda;
        Gamma(1+p:n+1+p,n+2+p:2*n+1+p) = D;
        p = p + n+1;

    end

end
full(Gamma);
    len = size(Gamma);

    %Populating our full sub-intensity matrix B (joining the Gamma and D
    %matrices)
    for i = 1:len(1)

        if i == len-2
```

```

                    Gamma(i,i) = -sum(Gamma(i,:))-alpha;

            elseif i == len

                    Gamma(i,i) = -sum(Gamma(i,:))-beta;

            else Gamma(i,i) = -sum(Gamma(i,:));

            end
      end


full(Gamma);

sum(Gamma');

inverseGamma = inv(full(Gamma));

%Initial distribution vector
pi = zeros(1,length(inverseGamma));
for y=2:length(pi)
            pi(y)=0;
end

% If all genes are active at time 0
pi(1)=1;


%Storing rewards in a matrix
%Rewards = [I1_rewards' I2_rewards'];
Rewards1=[];

for i = N:-1:2
Rewards1 = [Rewards i:-1:0];
end

%Tree height and branch lengths
tree_height = [tree_height pi*(-inverseGamma)*ones(1,length(pi))'];
branch_length = [branch_length, pi*(-inverseGamma)*Rewards'];
```

```

%end
%end
```

A.5 Theoretical results for the two-island coalescent

```
%Computing the theoretical results for two island branch length
%clear all
%% Parameters
results3 = [];


lambda = 1;

%Migration rates
M = 1;
R = 1;

%Coalescent rates
%alpha = 1;
beta = 5;

%Initialising vectors (tree heights, branch lengths, correlations)
tree_height=[];
branch_length=[];
corrL1L2 = [];
corrL1Tau = [];
corrL2Tau = [];
%for M = 0.2:0.2:5

for alpha=0:0.2:5

N = 50; % number of genes

%for N=10:10:100

%N=10:10:50;
Rewards = [];
Gamma = sparse(sum(N+1:-1:3),sum(N+1:-1:3));

p = 0;
```

```

for n = N:-1:2

    %% Constructing Lambda (Gamma) matrix

    Lambda = sparse(n+1,n+1);

    for j = 1:n
       Lambda(j+1,j) = j*R;
       Lambda(j,j+1) = (n-j+1)*M;
    end



    %% Constructing D matrix (coalescent rate matrix)
    if n == 2
        Gamma(1+p:n+1+p,1+p:n+1+p) = Lambda;
    else
        D = sparse(n+1,n);

        for k = 1:n
            if k<n
            D(k,k) = alpha*nchoosek(n-k+1,2);
            end

            if k > 1
            D(k+1,k) = beta*nchoosek(k,2); %Problem here
            end

        end
        Gamma(1+p:n+1+p,1+p:n+1+p) = Lambda;
        Gamma(1+p:n+1+p,n+2+p:2*n+1+p) = D;
        p = p + n+1;

    end

end
full(Gamma);
    len = size(Gamma);

    %Populating our full sub-intensity matrix B (joining the Gamma and D
    %matrices)
    for i = 1:len(1)

        if i == len-2

            Gamma(i,i) = -sum(Gamma(i,:))-alpha;

        elseif i == len

            Gamma(i,i) = -sum(Gamma(i,:))-beta;

        else Gamma(i,i) = -sum(Gamma(i,:));

        end
    end


full(Gamma);

sum(Gamma');

inverseGamma = inv(full(Gamma));

%Initial distribution vector
pi = zeros(1,length(inverseGamma));
%pi(1)=1;
for y=2:length(pi)
        pi(y)=0;
end

%Half of the population start in I1, half start in I2
pi(N/2+1)=1;

I1_rewards=[];
I2_rewards=[];
count=0;

%Calculating rewards for each island for each time step
```

```
for i=N:-1:2
I1_rewards = [I1_rewards i:-1:0];
end

for i=N:-1:2
I2_rewards = [I2_rewards 0:i];
end

%Storing rewards in a matrix
Rewards = [I1_rewards' I2_rewards'];

%Tree height and branch length
tree_height = [tree_height pi*(-inverseGamma)*ones(1,length(pi))'];
branch_length = [branch_length ;pi*(-inverseGamma)*Rewards];

%Cross moments
EL1L2 = pi*(-inverseGamma)*diag(Rewards(:,1))*(-inverseGamma)*Rewards(:,2)+pi*(-inverseGamma)*diag(Rewards(:,2))*(-inverseGamma)*Rewards(:,1);
EL1Tau = pi*(-inverseGamma)*diag(Rewards(:,1))*(-inverseGamma)*ones(1,length(pi))'+pi*(-inverseGamma)*diag(ones(1,length(pi))')*(-inverseGamma)*Rewards(:,1);
EL2Tau = pi*(-inverseGamma)*diag(Rewards(:,2))*(-inverseGamma)*ones(1,length(pi))'+pi*(-inverseGamma)*diag(ones(1,length(pi))')*(-inverseGamma)*Rewards(:,2);

%Expectations (total branch lengths, tree height)
EL1 = pi*(-inverseGamma)*Rewards(:,1);
EL2 = pi*(-inverseGamma)*Rewards(:,2);
ETau = pi*(-inverseGamma)*ones(1,length(pi))';

%Covariances
CovL1L2 = EL1L2-EL1*EL2;
CovL1Tau = EL1Tau-EL1*ETau;
CovL2Tau = EL2Tau-EL2*ETau;

%EL1Sq = 2*pi*(-inverseGamma)*diag(Rewards(:,1))*(-inverseGamma)*Rewards(:,1);%2*pi*(-inverseGamma)*diag(Rewards(:,1))*(-inverseGamma)*Rewards(:,1);
%EL2Sq = 2*pi*(-inverseGamma)*diag(Rewards(:,2))*(-inverseGamma)*
```

```
    Rewards(:,2);%2*pi*(-inverseGamma)*diag(Rewards(:,2))*(-
    inverseGamma)*Rewards(:,2);

%Second moments for total branch lengths
EL1Sq = 2*pi*(-inverseGamma)*diag(Rewards(:,1))*(-inverseGamma)*
    Rewards(:,1);
EL2Sq = 2*pi*(-inverseGamma)*diag(Rewards(:,2))*(-inverseGamma)*
    Rewards(:,2);

%ETauSq = pi*((-inverseGamma)^(2))*ones(1,length(pi))';
%ETauSq = 2*pi*(-inverseGamma)*diag(ones(1,length(pi))')*(-
    inverseGamma)*ones(1,length(pi))' %+ pi*(-inverseGamma)*diag(ones
    (1,length(pi))')*(-inverseGamma)*ones(1,length(pi))'
%ETauSq = 2*pi*(-inverseGamma)*ones(1,length(pi))';

%Variances of total branch lengths
varL1 = EL1Sq-EL1^2;
varL2 = EL2Sq-EL2^2;
%varL1 = 2*pi*((full(Gamma))^(-2))*Rewards(:,1)-(pi*(inverseGamma)*
    Rewards(:,1))^2;
%varL2 = 2*pi*((full(Gamma))^(-2))*Rewards(:,2)-(pi*(inverseGamma)*
    Rewards(:,2))^2;
%varTau = ETauSq-ETau^2;
varTau = 2*pi*((full(Gamma))^(-2))*ones(1,length(pi))'-(pi*(
    inverseGamma)*ones(1,length(pi))')^2;
%varTau = 2*pi*(full(Gamma))^(-2)*ones(1,length(pi))'-(pi*
    inverseGamma*ones(1,length(pi))')^2

%Storing correlations for each iteration
corrL1L2 = [corrL1L2 (CovL1L2)/(sqrt(varL1*varL2))];
corrL1Tau = [corrL1Tau (CovL1Tau)/sqrt(varL1*varTau)];
corrL2Tau = [corrL2Tau (CovL2Tau)/sqrt(varL2*varTau)];

end
%end
```

Bibliography

[1] Armando Arredondo, Beatriz Mourato, Khoa Nguyen, Simon Boitard, Willy Rodríguez, Camille Noûs, Olivier Mazet, and Lounès Chikhi. Inferring number of populations and changes in connectivity under the n-island model. *Heredity*, pages 1–17, 2021.

[2] Søren Asmussen. *Applied probability and queues*, volume 51. Springer Science & Business Media, 2008.

[3] Søren Asmussen, Patrick J Laub, and Hailiang Yang. Phase-type models in life insurance: Fitting and valuation of equity-linked benefits. *Risks*, 7(1):17, 2019.

[4] Mogens Bladt and Bo Friis Nielsen. *Matrix-exponential distributions in applied probability*, volume 81. Springer, 2017.

[5] Jochen Blath, Adrián González Casanova, Noemi Kurt, and Dario Spanò. The ancestral process of long-range seed bank models. *Journal of Applied Probability*, 50(3):741–759, 2013.

[6] Bjarki Eldon and John Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172(4):2621–2633, 2006.

[7] A.K. Erlang. The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik B*, 20:33–39, 01 1909.

[8] Ronald A Fisher. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

[9] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

[10] John Burdon Haldane. *The causes of evolution*, volume 5. Princeton University Press, 1990.

[11] Jotun Hein, Mikkel Schierup, and Carsten Wiuf. *Gene genealogies, variation and evolution: a primer in coalescent theory.* Oxford University Press, USA, 2004.

[12] Jody Hey. Isolation with migration models for more than two populations. *Molecular biology and evolution*, 27(4):905–920, 2010.

[13] Asger Hobolth, Arno Siri-Jegousse, and Mogens Bladt. Phase-type distributions in population genetics. *Theoretical population biology*, 127:16–32, 2019.

[14] A. Jensen. *A Distribution Model, Applicable to Economics.* Copenhagen, 1954.

[15] John FC Kingman. Origins of the coalescent: 1974-1982. *Genetics*, 156(4):1461–1463, 2000.

[16] Paul Marjoram and Paul Joyce. Practical implications of coalescent theory. In *Problem Solving Handbook in Computational Biology and Bioinformatics*, pages 63–84. Springer, 2010.

[17] Fabrizio Menardo, Sébastien Gagneux, and Fabian Freund. Multiple merger genealogies in outbreaks of mycobacterium tuberculosis. *Molecular Biology and Evolution*, 38(1):290–306, 2021.

[18] Patrick Alfred Pierce Moran. Random processes in genetics. In *Mathematical proceedings of the cambridge philosophical society*, volume 54, pages 60–71. Cambridge University Press, 1958.

[19] Thomas Nagylaki. The island model with stochastic migration. *Genetics*, 91(1):163–176, 1979.

[20] Marcel Neuts. Probability distributions of phase type. *Probability Distributions of Phase Type*, 01 1975.

[21] Jim Pitman. Coalescents with multiple collisions. *Annals of Probability*, pages 1870–1902, 1999.

[22] Lluis Quintana-Murci. Understanding rare and common diseases in the context of human evolution. *Genome biology*, 17(1):1–14, 2016.

[23] Bruce Rannala and JA Hartigan. Estimating gene flow in island populations. *Genetics Research*, 67(2):147–158, 1996.

[24] Serik Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability*, pages 1116–1125, 1999.

[25] Jason Schweinsberg. Coalescent processes obtained from supercritical galton–watson processes. *Stochastic processes and their Applications*, 106(1):107–139, 2003.

[26] Jason Schweinsberg. Rigorous results for a population model with selection ii: Genealogy of the population. *Electronic Journal of Probability*, 22, 07 2015.

[27] M Slarkin. Gene flow in natural populations. *Annual review of ecology and systematics*, 16(1):393–430, 1985.

[28] Sewall Wright. Coefficients of inbreeding and relationship. *The American Naturalist*, 56(645):330–338, 1922.

[29] Sewall Wright. Isolation by distance. *Genetics*, 28(2):114, 1943.