Wright-Fisher evolution

Carsouw, M.F.J.

M.F.J. Carsouw

Wright-Fisher evolution

Bachelor thesis, July 1, 2012

Supervisor: Prof. Dr. W.Th.F. den Hollander

Mathematical Institute, Leiden University

Abstract

The *Wright-Fisher model* is a discrete-time model for the genetic evolution of a finite haploid population of constant size $2N$, where each individual is of *type* say A or a. Time starts at $n = 0$ and at each unit of time $n \in \mathbb{N}$, each individual randomly chooses an individual from the previous generation and adopts its type. The probability that type a goes extinct equals the initial fraction of As in the population. The *genetic variability* H_n at time n is the probability that two different individuals, randomly drawn from the population at time n, are of diferent type. The expectation of the time τ it takes for one of the types to go extinct equals $\mathbb{E}(\tau) = 2NH_0$.

The *Moran model* is a continuous-time version of the WF-model. Making a space-time rescaling and sending the population size to infinity leads to a convergence of both the WF-model and the Moran model to the diffusion limit $(Y_t)_{t\geq 0}$ known as the *WF-diffusion*. The latter is *dual* to a pure death process $(D_t)_{t\geq 0}$ from which the *the (Kingman) coalescent* $(R_t)_{t\geq 0}$, describing the genealogy of a large (haploid) population, can be constructed. The different states through which $(R_t)_{t\geq 0}$ evolves form the *jump chain* of which the transition probabilities can be calculated.

Consider a population at time t that has been evolving indefinitely in accordance with $(Y_t)_{t\geq 0}$, then all the individuals in the population a.s. have a *most recent common ancestor* (MRCA) that lived at time $A_t < t$. The expectation of the (rescaled) time between the population and its MRCA equals $\mathbb{E}(t - A_t) = 2$. At any time t, there are two oldest families in the population. The time F_t at which one of these families dies out, the other family *fixates* in the population, which causes a jump of the MRCA. Using [2] it is possible to conclude that the *MRCA-process* $(A_t)_{t\geq 0}$ and the *fixation process* $(F_t)_{t\geq 0}$ are rate-1 Poisson processes.

Contents

1 Introduction

Sewall Green Wright and Ronald Aylmer Fisher are considered to be two of the founders of modern theoretical population genetics. One specific stochastic model for genetic evolution is named after them: the *Wright-Fisher model*. This model captures an important aspect of evolution theory known as *resampling*.

The WF-model follows the evolution of two alleles A and a, within a *haploid* population over (discrete) time, i.e., every individual of the population carries only one copy of each of its chromosomes, and therefore only one copy of either allele A or allele a. Consequently, each individual can be identified with its *type* (A or a). Humans and most other higher organisms are *diploid*, carrying two copies of their genetic material. Doubling each (haploid) individual in the population of the WF-model, and therefore the whole population size, means identifying each individual and its duplicate with a new diploid individual. This way, the WF-model can also be applied to the human species.

In the WF-model, the population size is assumed to be finite and constant. At each time unit, each individual randomly chooses one of the individuals from the previous generation, and adopts its type. This is a form of random reproduction. There are several interesting questions about the WF-model to investigate, both *forward* and *backward* in time. Some forward questions are:

(1) Given the initial numbers of As and as, what is the probability that type a goes extinct?

(2) What is the expectation of the time it takes for one of the types to *fixate* in the population?

After a proper *space-time rescaling*, the WF-model converges to a *diffusion limit* known as the *Wright-Fisher diffusion*. This limiting process describes the evolution of the *genetic variability* of a large haploid population over time.

(3) How should this limiting process be identified, and convergence be proven?

The moments of the WF-diffusion can be determined through the *duality* between the WF-diffusion and a pure death process on the natural numbers. This death process can be used to investigate the depth of the "backward family tree" of a large haploid population. At the very root of this tree lies the *most recent common ancestor* (MRCA) of the whole population. Some backward questions are:

(4) How long ago did this MRCA live? (Can the expectation of the depth of the family tree be calculated?)

The MRCA "jumps" to keep up with the current population's evolution. This jump process is known as the MRCA process.

(5) What is the distribution of the MRCA process?

Searching for answers to (4) and (5) leads us to defining *the (Kingman) coalescent*, which gives a full description of the genealogy of a large haploid population. Coalescent theory is of significant influence in population genetics.

This bachelor thesis aims to provide answers to these questions, to describe the stochastic processes that arise along the way, and to illustrate the mathematical stucture that binds all these stochastic processes together.

2 Wright-Fisher model

2.1 Random reproduction

The WF-model is a stochastic model for reproduction in population genetics. It considers a population consisting of N diploid individuals. This can be interpreted as $2N$ haploid individuals. The population size $2N$ is fixed.

The WF-model keeps track of two alleles, A and a, referred to as *types* and each individual has one of the two types. We refer to this by saying that each individual is of *type* A or a. Time starts at $n = 0$ and at each time unit $n \in \mathbb{N}$, each individual randomly chooses one of the $2N$ individuals from the time-$(n-1)$ generation, and adopts its type.

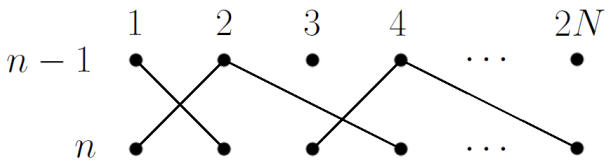

Figure 1: Example of the random reproduction in the WF-model for $n \in \mathbb{N}$. If at time $n-1$, individuals $1, 2$ and 4 are of type A, A and a, then at time n individuals $1, 2, 3, 4$ and $2N$ are of type A, A, a, A and a.

2.2 Fixation of a type

Let

$$X_n = \text{the number of } A\text{s at time } n. \tag{2.2.1}$$

The sequence $(X_n)_{n\in\mathbb{N}_0}$ is a discrete-time Markov chain on the state space $\Omega = \{0, 1, \ldots, 2N\}$, and gives us the opportunity to observe changes in the number of both As and as, until one of the types *fixates* in the population, i.e., until only one of the types remains due to extinction of the other. Since the WF-model describes random reproduction, the probability that an individual at time $n+1$ chooses an individual of type A, given that there are i individuals of type A at time n, equals $\frac{i}{2N}$. The complementary probability (the probability that the individual chooses an individual of type a) equals $\frac{2N-i}{2N}$. Taking into account the number of ways to pick j from $2N$ individuals, it follows that the transition probabilities are given by

$$p(i,j) = \mathbb{P}(X_{n+1} = j \mid X_n = i) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(\frac{2N-i}{2N}\right)^{2N-j}, \qquad i, j \in \Omega. \tag{2.2.2}$$

Note that once all individuals are of the same type, they will remain so forever, i.e., $p(0,0) = p(2N, 2N) = 1$. That is why we are interested in the time until fixation,

$$\tau = \inf\{n \in \mathbb{N}_0 : X_n = 0 \text{ or } X_n = 2N\}. \tag{2.2.3}$$

But before we say anything about τ, let us use its definition to calculate the probability that allele a goes extinct given the number of As at time $n = 0$.

Theorem 2.2.1 $\mathbb{P}(X_\tau = 2N \mid X_0 = i) = \frac{i}{2N}, \qquad i \in \Omega.$

Proof. First note that the transition probabilities $p(i, \cdot)$ are given by the probability distribution function of the binomial distribution with $2N$ trials and success probability $\frac{i}{2N}$, so that X is a *martingale*, i.e.,

$$\mathbb{E}(X_{n+1} \mid X_n, \ldots, X_0) = 2N\frac{X_n}{2N} = X_n. \tag{2.2.4}$$

Since the state space Ω is finite, we have $\mathbb{P}(\tau < \infty) = 1$. From this and the fact that $p(0,0) = p(2N, 2N) = 1$, it follows that

$$\lim_{n\to\infty} X_n = X_\tau. \tag{2.2.5}$$

Taking n iteration steps in (2.2.4), we get

$$i = \mathbb{E}(X_n \mid X_0 = i) = \mathbb{E}(X_\tau 1_{\{\tau\leq n\}} \mid X_0 = i) + \mathbb{E}(X_n 1_{\{\tau>n\}} \mid X_0 = i). \tag{2.2.6}$$

Now recall (2.2.5) and let $n \to \infty$, to obtain

$$\mathbb{E}(X_\tau \mid X_0 = i) = i. \tag{2.2.7}$$

However, a straightforward calculation of the conditional expectation of X_τ gives us

$$\mathbb{E}(X_\tau \mid X_0 = i) = 0 \times \mathbb{P}(X_\tau = 0 \mid X_0 = i) + 2N \times \mathbb{P}(X_\tau = 2N \mid X_0 = i). \tag{2.2.8}$$

Combining (2.2.7) and (2.2.8), we get

$$i = \mathbb{E}(X_\tau \mid X_0 = i) = 2N \times \mathbb{P}(X_\tau = 2N \mid X_0 = i), \tag{2.2.9}$$

from which it follows that

$$\mathbb{P}(X_\tau = 2N \mid X_0 = i) = \frac{i}{2N}. \tag{2.2.10}$$

□

The probability that two different individuals, randomly drawn from the population at time n, are of different type is called the *genetic variability* H_n of the population at time n, written

$$H_n = \frac{X_n}{2N}\frac{(2N - X_n)}{(2N-1)} + \frac{(2N - X_n)}{2N}\frac{X_n}{(2N-1)} = \frac{2X_n(2N - X_n)}{2N(2N-1)}. \tag{2.2.11}$$

We use H_n to calculate the expectation of τ, so that we have an indication of the time it takes for one of the alleles to fixate in the population. It is hereby useful to look at the population's *genealogical history*. For any individual k at time n, let its *backward ancestral path* π_k^n be given by

$$\pi_k^n = \{(\pi_m^n(k)) : 0 \leq m \leq n\}, \quad 1 \leq k \leq 2N, \tag{2.2.12}$$

where $\pi_m^n(k)$ is the label of the time-m ancestor of time-n individual k.

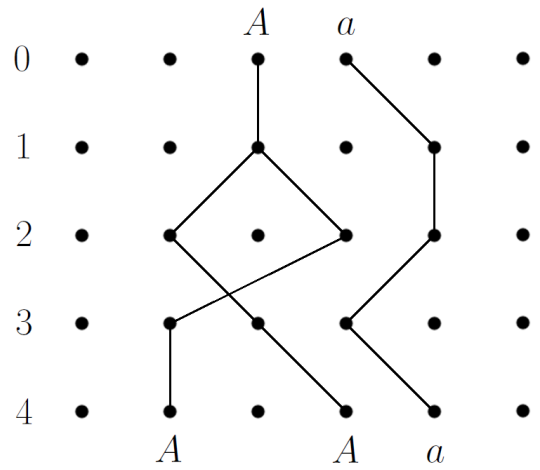

Figure 2: Example of three backward ancestral paths, for $N = 3$ and $n = 4$, where time runs downwards, starting at 0.

We are now ready to calculate the expectation of τ.

Theorem 2.2.2 $\mathbb{E}(\tau) = 2NH_0$.

Proof. For all time-n individuals i and j with $1 \leq i < j \leq 2N$, we have

$$\mathbb{P}(\pi_{m-1}^n(i) \neq \pi_{m-1}^n(j) \mid \pi_m^n(i) \neq \pi_m^n(j)) = 1 - \frac{1}{2N}, \quad 1 \leq m \leq n. \tag{2.2.13}$$

Labels $\pi_0^n(i)$ and $\pi_0^n(j)$ are equal if and only if the backward ancestral paths π_i^n and π_j^n have coalesced between times n and 0, so it follows that

$$\mathbb{P}(\pi_0^n(i) \neq \pi_0^n(j)) = \left(1 - \frac{1}{2N}\right)^n, \quad i \neq j. \tag{2.2.14}$$

Hence the probability H_n that two different random individuals i and j at time n are of different type equals the probability that they have two different ancestors at time 0 who also are of different type, so

$$\mathbb{E}(H_n \mid H_0) = \left(1 - \frac{1}{2N}\right)^n H_0. \tag{2.2.15}$$

Since the genetic variability equals $H_n = 0$ if and only if $n \geq \tau$, it now follows that

$$\mathbb{P}(\tau > n \mid H_0) = \mathbb{P}(H_n \neq 0 \mid H_0) = \left(1 - \frac{1}{2N}\right)^n H_0 \tag{2.2.16}$$

and therefore we conclude that

$$\mathbb{E}(\tau) = \sum_{n=0}^{2N} \left(1 - \frac{1}{2N}\right)^n H_0 = 2NH_0. \tag{2.2.17}$$

□

2.3 Moran model

The Moran model is a continuous-time version of the WF-model. In the Moran model, each individual randomly chooses an ancestor at rate 1 and adopts its type, i.e., each individual lives for an exponentially distributed amount of time, with expectation 1, and afterwards is randomly replaced (possibly by itself). This phenomenon is known as *sequential updating*, as opposed to the *parallel updating* in the WF-model, where all individuals randomly choose their ancestors at the same time.

As in the WF-model, we keep track of the number of individuals of one of the two types over time. So, let $(M_t)_{t\geq 0}$ be the continuous-time Markov process on Ω, defined as

$$M_t = \text{the number of As at time t}, \quad t \geq 0. \tag{2.3.1}$$

Then, as a consequence of the sequential updating in the Moran model, $(M_t)_{t\geq 0}$ is a birth-death process on Ω. A birth occurs when an individual of type a adopts the type of an individual of type A, and a death occurs in the complementary event. Thus, the transition rates are given by

$$\begin{array}{ll} k \to k+1 & \text{at rate } (2N-k)\frac{k}{2N}, \\ k \to k-1 & \text{at rate } k(1-\frac{k}{2N}), \end{array} \qquad k \in \Omega. \tag{2.3.2}$$

Since these rates are equal, we conclude that at any time $t \geq 0$, independently of the number of As, deaths and births are equally likely.

A similarity between the Moran model and the WF-model is that both models have the same fixation probabilities. In the Moran model, the probability that type A becomes fixed, given that there are initially i individuals of type A, equals $\frac{i}{2N}$. For the WF-model, we already proved this in Theorem 2.2.1. Furthermore, after space-time rescaling, both models have the same *diffusion limit*, known as the *Wright-Fisher diffusion*. The only difference is that the Moran model runs twice as fast as the WF-model, a result we will show in the next chapter.

3 Wright-Fisher diffusion

In nature it is often the case that a biological population is large in number. It can therefore be useful to consider a *limiting process* where the population size tends to infinity. This limiting process can then be used to approximate the behaviour of a large population. The WF-diffusion is such a limiting process. In particular, it is the *diffusion limit* of the WF-model (Redig [10], Section 2, contains a general approach to diffusion limits). The WF-diffusion describes changes in *allele frequency* within a large population in the WF-model. This also explains the role the WF-diffusion plays in population genetics; *genetic drift* can be read off from the WF-diffusion immediately.

3.1 Constructing the WF-diffusion

Recall the Markov process $(X_n)_{n\in\mathbb{N}_0} = (X_n^{(N)})_{n\in\mathbb{N}_0}$ with state space Ω and transition probabilities $p_N(i,\cdot) = \mathrm{BIN}(2N, \frac{i}{2N})(\cdot)$, as defined in (2.2.1), where we add the upper index (N) to emphasize the N-dependence. For $t \geq 0$ the following *space-time rescaling* is defined,

$$Y_t^{(N)} = \frac{1}{2N} X_{\lfloor 2Nt \rfloor}^{(N)}. \tag{3.1.1}$$

The process $(Y_t^{(N)})_{t\geq 0}$ is a continuous-time Markov process with state space $\{0, \frac{1}{2N}, \cdots, 1\}$ and *(infinitesimal) generator* L_N given by

$$(L_N f)\left(\frac{i}{2N}\right) = 2N \sum_{j=0}^{2N} p_N(i,j)\left(f\left(\frac{j}{2N}\right) - f\left(\frac{i}{2N}\right)\right). \tag{3.1.2}$$

The *Wright-Fisher diffusion* is constructed from the rescaled process $(Y_t^{(N)})_{t\geq 0}$ by letting $N \to \infty$. Define the WF-diffusion to be the continuous-time Markov process $(Y_t)_{t\geq 0}$ on $[0,1]$ with generator

$$(Lf)(y) = \frac{1}{2} y(1-y) f''(y). \tag{3.1.3}$$

Then the following theorem gives us the desired convergence, where $\mathcal{L}$ stands for law.

Theorem 3.1.1 *If there is convergence of initial conditions,*

$$\lim_{N\to\infty} \mathcal{L}\left(Y_0^{(N)}\right) = \mathcal{L}\left(Y_0\right), \tag{3.1.4}$$

then the whole process $(Y_t^{(N)})_{t\geq 0}$ converges (in distribution) to the WF-diffusion $(Y_t)_{t\geq 0}$, i.e.,

$$\lim_{N\to\infty} \mathcal{L}\left((Y_t^{(N)})_{t\geq 0}\right) = \mathcal{L}\left((Y_t)_{t\geq 0}\right). \tag{3.1.5}$$

Proof. It is enough to prove convergence of generators (see Ethier and Kurtz [5], Chapter 10, Theorem 1.1):

$$\lim_{N\to\infty} (L_N f)\left(\frac{i}{2N}\right) = (Lf)(y). \tag{3.1.6}$$

Here we have to specify which set of test functions f we consider. Let $C([0,1])$ be the set of real-valued, continuous functions on the unit interval, and define

$$C_0([0,1]) = \{f \in C([0,1]) : f(0) = f(1) = 0\}. \tag{3.1.7}$$

Since the local speed of the WF-diffusion is given by the *diffusion function* $d : [0,1] \to [0,\infty)$ with $d(y) = y(1-y)$, it is clear that the domain $D(L)$ of the generator L of the WF-diffusion must be a

subset of $C_0([0,1])$. In general, it is not easy to characterize $D(L)$, so the idea is to restrict $D(L)$ to a suitable subset $K(L) \subset D(L)$ that is large enough to maintain the generality of the arguments. We refer to $K(L)$ as a *core* of the generator. We require that $K(L)$ is dense in $C_0([0,1])$ with respect to the supremum norm. Then generality is obtained via continuous extension.

In our case it suffices to consider the functions that are infinitely differentiable and have appropriate boundary conditions,

$$K(L) = \{f \in C_0([0,1]) : f \text{ is infinitely differentiable}\}. \tag{3.1.8}$$

Test functions need only be twice continuously differentiable in the proof that follows below, but $K(L)$ in (3.1.8) is already dense in $C_0([0,1])$. This enables us to use the Taylor expansion of any test function up to any order.

Using the Taylor expansion of f around $\frac{i}{2N}$ up to second order, we obtain

$$(L_N f)\left(\frac{i}{2N}\right) = \sum_{j=0}^{2N} p_N(i,j)(j-i) f'\left(\frac{i}{2N}\right) + \frac{1}{2}\sum_{j=0}^{2N} p_N(i,j)\frac{(j-i)^2}{2N} f''\left(\frac{i}{2N}\right) + R_N, \tag{3.1.9}$$

where R_N consists of third- and higher-order terms. Now put $y = Y_0$ and $X_0^{(N)} = i = i_N$, so that (3.1.4) becomes

$$\lim_{N\to\infty} \frac{i_N}{2N} = y \in [0,1]. \tag{3.1.10}$$

Let $F, G : [0,1] \to \mathbb{R}$ be given by

$$F(y) = \lim_{N\to\infty} \sum_{j=0}^{2N} p_N(i,j)(j-i), \tag{3.1.11}$$

$$G(y) = \lim_{N\to\infty} \sum_{j=0}^{2N} p_N(i,j)\frac{(j-i)^2}{2N}. \tag{3.1.12}$$

Since third- and higher-order terms of (3.1.9) disappear when $N \to \infty$, the limiting generator equals

$$\begin{aligned} &\lim_{N\to\infty} (L_N f)\left(\frac{i}{2N}\right) \\ &= \lim_{N\to\infty} \left(\sum_{j=0}^{2N} p_N(i,j)(j-i) f'\left(\frac{i}{2N}\right) + \frac{1}{2}\sum_{j=0}^{2N} p_N(i,j)\frac{(j-i)^2}{2N} f''\left(\frac{i}{2N}\right)\right) \\ &= F(y)f'(y) + \frac{1}{2}G(y)f''(y), \end{aligned} \tag{3.1.13}$$

where the last equality uses (3.1.10) and the fact that all derivatives of f are continuous. We will prove that $F(y) = 0$ and $G(y) = y(1-y)$, so that the limiting generator is indeed equal to (3.1.3).

From the martingale property in (2.2.4) it follows that $\mathbb{E}(X_1^{(N)})$ equals $\mathbb{E}(X_0^{(N)})$, which is $i_N = i$. We have

$$\mathbb{E}(X_0^{(N)}) = i\sum_{j=0}^{2N} p_N(i,j), \tag{3.1.14}$$

while the expectation definition gives us

$$\mathbb{E}(X_1^{(N)}) = \sum_{j=0}^{2N} j\, p_N(i,j). \tag{3.1.15}$$

Combining (3.1.14) and (3.1.15), we get

$$F(y) = \lim_{N\to\infty} \sum_{j=0}^{2N} p_N(i,j)(j-i) = 0. \tag{3.1.16}$$

As stated earlier, $\mathbb{E}(X_1^{(N)})$ equals $i_N = i$, from which it follows that

$$\mathrm{Var}(X_1^{(N)}) = \sum_{j=0}^{2N} p_N(i,j)(j-i)^2. \tag{3.1.17}$$

Since $X_1^{(N)}$ follows the binomial distribution BIN$(2N, \frac{i}{2N})$, we also have

$$\mathrm{Var}(X_1^{(N)}) = 2N\,\frac{i}{2N}\left(1-\frac{i}{2N}\right). \tag{3.1.18}$$

Combining (3.1.17) and (3.1.18), we get

$$G(y) = \lim_{N\to\infty} \sum_{j=0}^{2N} p_N(i,j)\frac{(j-i)^2}{2N} = \lim_{N\to\infty} \frac{i_N}{2N}\left(1-\frac{i_N}{2N}\right) = y(1-y), \tag{3.1.19}$$

and thus the desired result follows. □

The WF-diffusion can also be constructed from the Moran model. Since the Moran model is a continuous-time version of the WF-model, we expect the Moran model to converge to the WF-diffusion as well. But first we have to make the proper space-time rescaling, as we did for the WF-model in (3.1.1). The rescaled Moran model is characterized by the birth-death Markov process on $\{0, \frac{1}{2N}, \cdots, 1\}$ with generator $\hat{L}_N$ given by

$$(\hat{L}_N f)\left(\frac{i}{2N}\right) = 2N\,\frac{i}{2N}(2N-i)\left(f\left(\frac{i-1}{2N}\right) + f\left(\frac{i+1}{2N}\right) - 2f\left(\frac{i}{2N}\right)\right). \tag{3.1.20}$$

The first factor $2N$ is a consequence of the rescaled time, the second factor $\frac{i}{2N}(2N-i)$ equals both the birth rate and the death rate in the Moran model. Using the Taylor expansion of f around $\frac{i-1}{2N}$ and $\frac{i+1}{2N}$ up to second order, we obtain

$$(\hat{L}_N f)\left(\frac{i}{2N}\right) = (2N)^2\frac{i}{2N}\left(1-\frac{i}{2N}\right)\left(\frac{1}{(2N)^2}f''(\frac{i}{2N}) + \hat{R}_N\right), \tag{3.1.21}$$

where $\hat{R}_N$ consists of third- and higher-order terms that will disappear when $N \to \infty$. Assuming convergence of initial conditions of the rescaled Moran model to the initial condition of the WF-diffusion $Y_0 = y$, we let

$$\lim_{N\to\infty} \frac{i_N}{2N} = y \in [0,1]. \tag{3.1.22}$$

Then the limiting generator equals

$$\lim_{N\to\infty} (\hat{L}_N f)\left(\frac{i}{2N}\right) = \lim_{N\to\infty} \frac{i}{2N}\left(1-\frac{i}{2N}\right) f''\left(\frac{i}{2N}\right) = y(1-y)f''(y). \tag{3.1.23}$$

This is equal to the generator of the WF-diffusion, apart from the factor $\frac{1}{2}$ in (3.1.3). From this we conclude that, after space-time rescaling, the Moran model converges to the WF-diffusion, but runs at twice the speed of the WF-model. Indeed, if we reverse time, then each collision of two different lineages in the Moran model is obtained by only *one* jump, while such a collision requires *two* jumps in the WF-model. (Events where three or more lineages collide at the same time can be discarded, since these events have probabilities of order N^{-3} and will therefore not appear in the WF-diffusion.)

3.2 Dual to the WF-diffusion

In this section we analyze the WF-diffusion through *duality* and explain how the dual process is constructed.

Consider the backward ancestral tree of a population taken from the WF-diffusion, and define

$$D_t = \text{number of ancestral lineages at time } t_0 - t, \quad t \geq 0, \tag{3.2.1}$$

where $t_0 \in \mathbb{R}$ is the observation time. Then it follows from the construction of the WF-model that $(D_t)_{t\geq 0}$ is a pure death process on the natural numbers, where ancestral lineages are killed one by one. The following theorem enables us to give a more formal definition of this death process.

Theorem 3.2.1 *The amount of time τ_k during which there are k lineages is equal to the amount of time during which $D_\bullet$ equals k, and has an exponential distribution with mean $2/k(k-1)$, $k = 2, 3, \ldots$.*

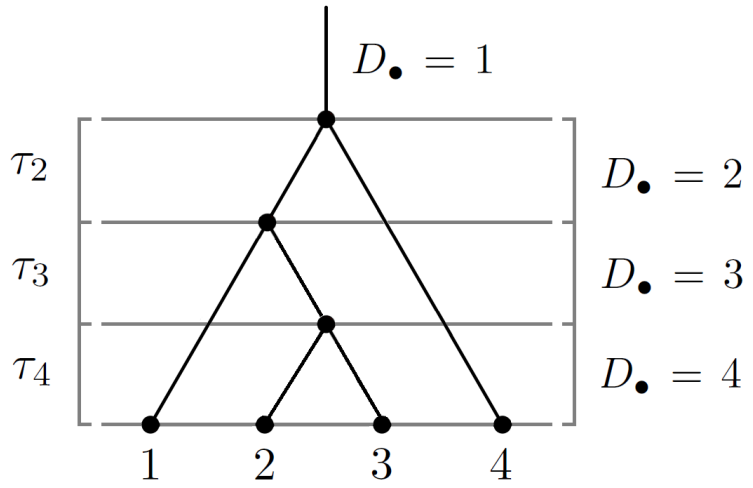

Figure 3: Part of a backward ancestral tree of a population in the WF-diffusion. In the death process $(D_t)_{t\geq 0}$ time is running upwards, while biological time is running downwards. The expectations of the times τ_4, τ_3 and τ_2 equal, respectively, $\frac{1}{6}, \frac{1}{3}$ and 1.

Proof. First we consider a sample of k individuals drawn from a WF-population of finite size $2N \gg 1$. Later on we make the proper space-time rescaling, after which we send the population size to infinity. The probability that two (or more) individuals from the sample have the same parent, is

$$\frac{k(k-1)}{2}\frac{1}{2N} + O\left(N^{-2}\right). \tag{3.2.2}$$

Indeed, there are $\binom{k}{2}$ different pairs of individuals in the sample, and in each pair the two individuals have the same parent with probability $\frac{1}{2N}$. The second term takes into account events where there are two pairs who have the same parents, or there are three or more individuals who all choose the same parent.

When we follow the ancestral lineages of the k individuals backwards in time, the probability that the k lineages do not coalesce during the first n generations equals

$$\left(1 - \frac{k(k-1)}{2}\frac{1}{2N} + O(N^{-2})\right)^n = \exp\left(-\frac{k(k-1)}{2}\frac{n}{2N} + O(N^{-2})\right). \tag{3.2.3}$$

Now let $\lambda_k = k(k-1)/2$, and rescale time by putting $t = n/2N$. Then in the limit as $N \to \infty$ we have, for all $t \geq 0$,

$$\mathbb{P}(\tau_k > t) = e^{-\lambda_k t}. \tag{3.2.4}$$

From this it follows that, when the population size tends to infinity, τ_k has an exponential distribution with mean

$$\lambda_k^{-1} = 2/k(k-1). \tag{3.2.5}$$

□

Theorem 3.2.1 says that we can equivalently define $(D_t)_{t\geq 0}$ as the pure death process on $\mathbb{N}$ where transitions from k to $k-1$ occur at rate $k(k-1)/2$. The process $(D_t)_{t\geq 0}$ is *dual* to the WF-diffusion $(Y_t)_{t\geq 0}$ in the following sense:

$$\mathbb{E}([Y_t]^n \mid Y_0 = y) = \mathbb{E}(y^{D_t} \mid D_0 = n) \quad \forall\, y \in [0,1],\ n \in \mathbb{N},\ t \geq 0. \tag{3.2.6}$$

See Den Hollander [6], Section 2.1.3. To illustrate the use of duality in studying the WF-diffusion, we will prove the *dual representations* of Theorem 2.2.1 and (2.2.15) by using (3.2.6). First write

$$Y_\infty = \lim_{t\to\infty} Y_t \quad \text{and} \quad D_\infty = \lim_{t\to\infty} D_t. \tag{3.2.7}$$

Then the dual representation of Theorem 2.2.1 becomes

$$\mathbb{P}(Y_\infty = 1 \mid Y_0 = y) = y, \qquad y \in [0,1]. \tag{3.2.8}$$

This can be easily proved as follows. First note that $Y_\infty \in \{0,1\}$, so that

$$\mathbb{P}(Y_\infty = 1 \mid Y_0 = y) = \mathbb{E}(Y_\infty \mid Y_0 = y). \tag{3.2.9}$$

By the definition of $(D_t)_{t\geq 0}$, we have that $D_\infty = 1$, and so it follows from (3.2.6) that the right-hand side of (3.2.9) equals

$$\mathbb{E}(y^{D_\infty} \mid D_0 = 1) = y. \tag{3.2.10}$$

(Without the notion of duality, this would be a more difficult result to accomplish.)

The dual representation of (2.2.15) is

$$\mathbb{E}(Y_t(1-Y_t) \mid Y_0 = y) = y(1-y)e^{-t}, \qquad t \geq 0,\ y \in [0,1]. \tag{3.2.11}$$

The proof is as follows:

$$\begin{aligned} \mathbb{E}(Y_t(1-Y_t) \mid Y_0 = y) &= \mathbb{E}(y^{D_t} \mid D_0 = 1) - \mathbb{E}(y^{D_t} \mid D_0 = 2) \\ &= y - \big[y\mathbb{P}(D_t = 1 \mid D_0 = 2) + y^2\mathbb{P}(D_t = 2 \mid D_0 = 2)\big] \\ &= y(1-y)\mathbb{P}(D_t = 2 \mid D_0 = 2) \\ &= y(1-y)e^{-t}. \end{aligned} \tag{3.2.12}$$

The first equality uses (3.2.6), the second equality holds since 1 is the absorbing state of $(D_t)_{t\geq 0}$, the third equality follows from the fact that $\mathbb{P}(D_t = 1 \mid D_0 = 2) = 1 - \mathbb{P}(D_t = 2 \mid D_0 = 2)$, while the last equality is a consequence of Theorem 3.2.1.

In most computations, $(D_t)_{t\geq 0}$ is an easier process to work with than $(Y_t)_{t\geq 0}$. The duality gives a relation between the distribution of D_t and the moments of Y_t, so that the distribution of Y_t can be determined.

4 The coalescent

Kingman's coalescent, often called *the coalescent*, is a stochastic process that describes the family tree of a large haploid population backwards in time, up to the individual called *most recent common ancestor* (see Chapter 5). The coalescent is related to the WF-diffusion $(Y_t)_{t\geq 0}$ (see Section 3.1), and therefore through duality to the death process $(D_t)_{t\geq 0}$ (see Section 3.2). It is a powerful tool in the study of the most recent common ancestor. Nevertheless, coalescent theory stands on itself as an important area in population genetics (see Kingman [8]).

In this chapter, the coalescent is constructed from the continuous-time Markov process called *n-coalescent*. The transition probabilities and the absolute probabilities of the corresponding *jump chain* will be calculated using the death process $(D_t)_{t\geq 0}$. Existence and uniqueness of the coalescent will be established by the Stone-Weierstrass theorem. Finally, one of the coalescent's most striking features is shown: it comes down from infinity.

4.1 Constructing the n-coalescent

For $n \in \mathbb{N}$, let E_n denote the set of equivalence relations on the set $\{1, 2, \ldots, n\}$. Note that every element of E_n is a set consisting of pairs (i, j) with $i, j \in \{1, 2, \ldots, n\}$ (such that reflexivity, symmetry, and transitivity hold). The number of equivalence classes of an element $R \in E_n$ is denoted by $|R|$.

The *n-coalescent* is a continuous-time Markov process $(R_t^n)_{t\geq 0}$ with state space E_n, such that

$$R_0^n = \Delta = \{(i, i) : 1, 2, \ldots, n\} \tag{4.1.1}$$

and

$$p_{RS} = \begin{cases} 1 & \text{if } R \lhd S \\ 0 & \text{otherwise} \end{cases}, \quad R, S \in E_n, R \neq S, \tag{4.1.2}$$

where p_{RS} is the transition rate given by

$$p_{RS} = \lim_{h\downarrow 0} h^{-1}\mathbb{P}(R_{t+h}^n = S \mid R_t^n = R) \tag{4.1.3}$$

and

$$R \lhd S \Leftrightarrow \{S \subset R \text{ and } |S| = |R| + 1\}. \tag{4.1.4}$$

Intuitively, $R \lhd S$ means that S can be obtained from R by merging together two of R's equivalence classes.

Now, if we draw a sample of n individuals from a large haploid population at time t_0, then the n-coalescent gives us a full description of all the lineages of the n individuals, backwards in time. To understand this, let R_t^n contain (i, j) with $i, j \in \{1, 2, \ldots, n\}$ if and only if the individuals i and j have a common ancestor that is alive at time $t_0 - t$. With this definition of R_t^n, it is easy to check that R_t^n is an equivalence relation for all $t \geq 0$. Furthermore, the process $(R_t^n)_{t\geq 0}$ indeed has the stochastic structure of an n-coalescent. Note also that there is a one-to-one correspondence between the equivalence classes of R_t^n and the time-$(t_0 - t)$ common ancestors involved in the sample.

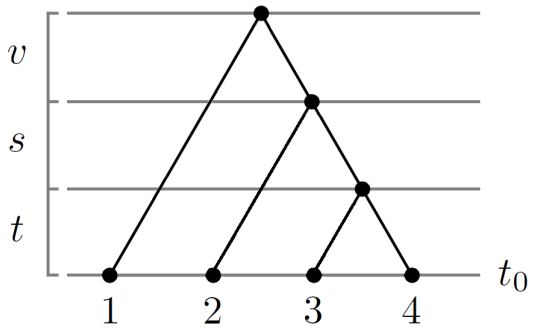

Figure 4: Example of the backward ancestral tree of a sample of size $n = 4$, taken from a large haploid population at time t_0. The jumps of the 4-coalescent are given by

$R^4_0 = \Delta = \{(i,i) : 1,2,3,4\}$,
$R^4_t = \Delta \cup \{(i,j) : i,j = 3,4\}$,
$R^4_{t+s} = R^4_t \cup \{(i,j) : i,j = 2,3,4\}$,
$R^4_{t+s+v} = \Theta = \{(i,j) : i,j = 1,2,3,4\}$.

Write $(D^n_t)_{t\geq 0}$ for the natural restriction of the death process $(D_t)_{t\geq 0}$ (defined in Section 3.2) to the set $\{1,2,\ldots,n\}$, so that $(D^n_t)_{t\geq 0}$ is the death process on $\{1,2,\ldots,n\}$ with initial value $D^n_0 = n$ and transitions from k to $k-1$ that occur at rate $\lambda_k = k(k-1)/2$. From (3.2.1), it is clear that D^n_t equals the number of equivalence classes of R^n_t, i.e.,

$$D^n_t = |R^n_t|. \tag{4.1.5}$$

Indeed, each collision of two equivalence classes of R^n_t corresponds to the extinction of one of the lineages in the genealogy of the sample. Since Theorem 3.2.1 gives us the transition rates of $(D_t)_{t\geq 0}$, it follows from (4.1.5) that the total transition rate of R^n equals

$$p_R = \lim_{h\downarrow 0} h^{-1}\mathbb{P}(R^n_{t+h} \neq R \mid R^n_t = R) = \frac{|R|(|R|-1)}{2}. \tag{4.1.6}$$

So the n-coalescent jumps with transition rates $k(k-1)/2$ through a sequence of equivalence relations $\Re_k$, with $|\Re_k| = k$ and $k = n, n-1, \ldots, 1$, such that

$$\Delta = \Re_n \lhd \Re_{n-1} \lhd \cdots \lhd \Re_1 = \Theta, \tag{4.1.7}$$

where $\Theta = \{(i,j) : i,j = 1,2,\cdots,n\}$ is called the *absorbing state*.

The *jump chain* of an n-coalescent is defined as the Markov chain

$$(\Re_n, \Re_{n-1}, \ldots, \Re_2, \Re_1), \tag{4.1.8}$$

which is directly related to its n-coalescent as

$$R^n_t = \Re_{D^n_t}. \tag{4.1.9}$$

From (4.1.2) and (4.1.6), we can now calculate the transition probabilities of the jump chain. For $R \in E_n$ with $|R| = k \in \{2,\ldots,n\}$, we have

$$\mathbb{P}(\Re_{k-1} = S \mid \Re_k = R) = \frac{p_{RS}}{p_R} = \begin{cases} 2/k(k-1) & \text{if } R \lhd S, \\ 0 & \text{otherwise.} \end{cases} \tag{4.1.10}$$

Indeed, exactly one of the $\binom{k}{2}$ coalescence events turns R into S. Finding the absolute probabilities of the jump chain requires a little more work (see e.g. R. Durrett [4], Section 1.2.2, Theorem 1.5).

Theorem 4.1.1 *Let $R \in E_n$, and let the number of equivalence classes of R be $|R| = k$. Then the absolute probabilities of the jump chain are given by*

$$\mathbb{P}(\Re_k = R) = \frac{k!}{n!}\frac{(n-k)!(k-1)!}{(n-1)!}\lambda_1!\lambda_2! \times \cdots \times \lambda_k!,$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the sizes of the equivalence classes of R.

Proof. We use induction on k, working backwards from $k = n$.

For $k = n$ we have $\Re_k = \Re_n = \Delta$, so that all equivalence classes are of size 1,

$$\lambda_1 = \ldots = \lambda_k = 1. \tag{4.1.11}$$

Thus we have

$$\frac{k!}{n!}\frac{(n-k)!(k-1)!}{(n-1)!}\lambda_1!\lambda_2! \times \cdots \times \lambda_k! = 1. \tag{4.1.12}$$

Since $R \in E_n$ with $|R| = n$ implies that $R = \Delta$, this is indeed equal to the statement that

$$\mathbb{P}(\Re_n = R) = 1. \tag{4.1.13}$$

Now let $k \in \{2, \ldots, n\}$ be arbitrary, and assume that the theorem holds for k (induction hypothesis). We will prove that then the theorem also holds for $k - 1$.

Let $S \in E_n$ with $|S| = k - 1$ and $R \lhd S$. Then from (4.1.10) it follows that

$$\mathbb{P}(\Re_{k-1} = S) = \frac{2}{k(k-1)}\sum_{R \lhd S}\mathbb{P}(\Re_k = R). \tag{4.1.14}$$

Suppose that the equivalence classes of S have sizes $\lambda_1, \ldots, \lambda_{k-1}$. Then there exist $l \in \{1, \ldots, k-1\}$ and $\mu \in \{1, \ldots, \lambda - 1\}$ such that the equivalence classes of R have sizes

$$\lambda_1, \ldots, \lambda_{l-1}, \mu, \lambda_l - \mu, \lambda_{l+1}, \ldots, \lambda_{k-1}. \tag{4.1.15}$$

From the induction hypothesis it follows that the right-hand side of (4.1.14) equals

$$\frac{2}{k(k-1)}\sum_{l=1}^{k-1}\sum_{\mu=1}^{\lambda_l - 1}\frac{k!}{n!}\frac{(n-k)!(k-1)!}{(n-1)!}\lambda_1! \times \cdots \times \lambda_{l-1}!\mu!(\lambda_l - \mu)!\lambda_{l+1}! \times \cdots \times \lambda_{k-1}!\binom{\lambda_l}{\mu}\frac{1}{2}. \tag{4.1.16}$$

Here $\binom{\lambda_l}{\mu}\frac{1}{2}$ equals the number of ways of picking $R \lhd S$, so that the equivalence classes of R with sizes μ and $\lambda_l - \mu$ coalesce to form the l-th equivalence class of S with size λ_l. From the simple fact that

$$\lambda_1! \times \cdots \times \lambda_{l-1}!\mu!(\lambda_l - \mu)!\lambda_{l+1}! \times \cdots \times \lambda_{k-1}!\binom{\lambda_l}{\mu} = \lambda_1! \times \cdots \times \lambda_{k-1}!, \tag{4.1.17}$$

we conclude that the right-hand side of (4.1.16) equals

$$\begin{aligned}
&\frac{k!}{n!}\frac{(n-k)!(k-1)!}{(n-1)!}\lambda_1!\lambda_2! \times \cdots \times \lambda_{k-1}!\frac{1}{k(k-1)}\sum_{l=1}^{k-1}\sum_{\mu=1}^{\lambda_l - 1}1 \\
&= \frac{k!}{n!}\frac{(n-k)!(k-1)!}{(n-1)!}\frac{(n-(k-1))}{k(k-1)}\lambda_1!\lambda_2! \times \cdots \times \lambda_{k-1}! \\
&= \frac{(k-1)!}{n!}\frac{(n-(k-1))!((k-1)-1)!}{(n-1)!}\lambda_1!\lambda_2! \times \cdots \times \lambda_{k-1}!
\end{aligned} \tag{4.1.18}$$

and thus the theorem holds for $k - 1$. □

4.2 Constructing the coalescent

We will now construct the coalescent from the n-coalescent (see also Kingman [7], Section 7). For $2 \leq m < n$, define the restriction $\rho_{nm} : E_n \to E_m$ given by $\rho_{nm}R = \{(i,j) \in R : 1 \leq i,j \leq m\}$. It is clear that ρ_{nm} is well-defined. In fact, if $(R_t^n)_{t\geq 0}$ is an n-coalescent, then $(\rho_{nm}R_t^n)_{t\geq 0}$ is an m-coalescent. Let E be the set of all equivalence relations on $\mathbb{N}$, and let $\rho_n : E \to E_n$ be given by $\rho_n R = \{(i,j) \in R : 1 \leq i,j \leq n\}$. We search for a process $(R_t)_{t\geq 0}$ on E such that $(\rho_n R_t)_{t\geq 0}$ is an n-coalescent for all $n \geq 2$. Such a process is called a coalescent. The following theorem justifies talking about *the* coalescent.

Theorem 4.2.1 *There exists a unique coalescent* $(R_t)_{t\geq 0}$.

Proof. Note that E is a subset of the powerset of $\mathbb{N} \times \mathbb{N}$, so that we have the inclusion $E \subset 2^{\mathbb{N}\times\mathbb{N}}$, where E is a closed subset with respect to the (compact) product topology on $2^{\mathbb{N}\times\mathbb{N}}$. For E equipped with the subspace topology, it follows that E is a compact Hausdorff space (a condition needed for the Stone-Weierstrass theorem).

A coalescent exists if we can consistently specify its finite-dimensional distributions. The consistency between different values of n is a consequence of the fact that $\rho_{nm}(\rho_n S) = \rho_m S$ for all $S \in E$ and $2 \leq m < n$. For fixed n, however, we observe the value of $\mathbb{E}(f(R_{t_1}, R_{t_2}, \ldots, R_{t_k}))$, for bounded continuous functions $f : (E)^k \to \mathbb{R}$ and times $0 \leq t_1 < t_2 < \ldots < t_k$. The requirement that $(\rho_n R_t)_{t\geq 0}$ is an n-coalescent determines the value of $\mathbb{E}(f(R_{t_1}, R_{t_2}, \ldots, R_{t_k}))$ when there exists a function $g : (E_n)^k \to \mathbb{R}$ such that f is of the form $f(S_1, \ldots, S_k) = g(\rho_n S_1, \ldots, \rho_n S_k)$. The Stone-Weierstrass theorem implies that the collection of functions f of this form lies dense in the set of bounded continuous functions $f : (E)^k \to \mathbb{R}$. Continuous extension then determines the value of $\mathbb{E}(f(R_{t_1}, R_{t_2}, \ldots, R_{t_k}))$ for *all* f, so that the consistent specification of the finite-dimensional distributions is established. Since this is done uniquely, it follows that every two coalescents have the same finite-dimensional distributions. □

With the above definition, the coalescent $(R_t)_{t\geq 0}$ is a Markov process with values in E and trivial initial condition $R_0 = \{(i,i) : i \in \mathbb{N}\}$, such that each pair of equivalence classes of R_t coalesces at rate 1, for all $t \geq 0$. The latter is obtained from (4.1.6) by taking $|R| = 2$. This means that every two lineages in the coalescent tree coalesce at rate 1, as we go backwards in time.

Where an n-coalescent gives a discription of the ancestral lineages of a sample of size n, taken from a large haploid population, the coalescent gives a discription of *all* the lineages of the whole population, up to the single individual from which the population has descended. In the WF-diffusion this population is assumed to be infinite. Therefore we have to ask ourselves the question: How is it possible that the infinite number of lineages in the coalescent tree decreases to a finite number at positive times? An *entrance law* can be used to describe how a Markov process "comes from infinity". The following section provides an answer to the question.

4.3 Coming down from infinity

Now that we have familiarized ourselves with the construction of both the n-coalescent and *the* coalescent, we are ready to understand *and* prove an interesting phenomenon (see Berestycki [1], Section 2.1.2). As before, we let $D_t = |R_t|$ denote the number of equivalence classes of R_t. Then we have $D_0 = \infty$, which corresponds to the statement that there are infinitely many lineages at the beginning of our coalescent tree. The following theorem says that after any positive time we are left with only a finite number of lineages. This is expressed by saying that the coalescent *comes down from infinity*.

Theorem 4.3.1 $\mathbb{P}(\forall\ t > 0 : D_t < \infty) = 1$.

Proof. We must show that for every $t > 0$,

$$\forall\ \epsilon > 0\ \exists\ N_\epsilon \in \mathbb{N} :\ \mathbb{P}(D_t > N_\epsilon) < \epsilon. \tag{4.3.1}$$

Let $t > 0$ and $\epsilon > 0$ be arbitrary. Furthermore, let $R_t^n = \rho_n(R_t)$ be the natural restriction of the coalescent to E_n, and let $D_t^n = |R_t^n|$ denote the number of equivalence classes of R_t^n. Let τ_k be an exponentially distributed random variable with mean λ_k^{-1}, where $\lambda_k = k(k-1)/2$, $k = 2, 3, \ldots$. Then Theorem 3.2.1 says that we can use τ_k to represent the amount of time during which $D_t = k$. Therefore, using Markov's inequality, we have that

$$\begin{aligned}\mathbb{P}(D_t^n > N_\epsilon) &= \mathbb{P}\left(\sum_{k=N_\epsilon}^{n} \tau_k > t\right) \leq \frac{1}{t}\mathbb{E}\left(\sum_{k=N_\epsilon}^{n} \tau_k\right) \leq \frac{1}{t}\mathbb{E}\left(\sum_{k=N_\epsilon}^{\infty} \tau_k\right) \\ &= \frac{1}{t}\left(\sum_{k=N_\epsilon}^{\infty} \mathbb{E}(\tau_k)\right) = \frac{1}{t}\left(\sum_{k=N_\epsilon}^{\infty} \frac{2}{k(k-1)}\right).\end{aligned} \tag{4.3.2}$$

Now note that the last sum implies that, independently of n, we can choose N_ϵ large enough to make sure that

$$\frac{1}{t}\left(\sum_{k=N_\epsilon}^{\infty} \frac{2}{k(k-1)}\right) < \epsilon, \tag{4.3.3}$$

from which it follows that

$$\limsup_{n\to\infty} \mathbb{P}(D_t^n > N_\epsilon) = \lim_{n\to\infty}\left(\sup_{m\geq n} \mathbb{P}(D_t^m > N_\epsilon)\right) < \epsilon. \tag{4.3.4}$$

□

5 Most recent common ancestor

An object of interest in population genetics, in particular, in coalescent theory, is the most recent common ancestor (MRCA). Consider a population where all individuals have a common allele (or locus, or gene). We can trace this allele in the ancestry of the population. As we move backwards in time, the collection of ancestors starts to shrink, until we are left with the single individual from which the whole population has descended. This individual is found at the root of the coalescent tree, and is called the MRCA. It is the most recent individual that has the allele such that it is a common ancestor of the whole population.

In this chapter we investigate the MRCA of a population in the WF-diffusion. The coalescent enables us to calculate the expectation of the time between the population and its MRCA. As time runs forward, the MRCA jumps forward in order to keep up with the population. This jump process is called the *MRCA-process*, which we will analyze by means of a particle construction introduced by Donnelly and Kurtz [2]. This leads to the corresponding *fixation process*. In particular, we will determine the distribution of the waiting time between jumps in this process.

5.1 Depth of the coalescent tree

Consider a population in the WF-diffusion where every individual is of type A or of type a. If the population has a MRCA, then this MRCA is either of type A or of type a, from which it follows that the whole population is either of type A or of type a. The reverse does not necessarily hold: if a population descends for example from two individuals instead of one, and these two individuals are both of the same type, then the whole population is of one type, but does not have a MRCA.

Suppose that the WF-diffusion has been running indefinitely, and observe a population in the WF-diffusion at some reference time $t \in \mathbb{R}$. Then this population has a MRCA a.s. (the MRCA did not live an infinite amount of time ago). In fact, given the time at which the population lives, we can predict the time of its MRCA.

Theorem 5.1.1 *The expectation of the time T_t between a time-t population in the WF-diffusion and its MRCA equals $\mathbb{E}(T_t) = 2$.*

Note: this theorem is stated in terms of a continuous *rescaled* time. For example, if we use the WF-diffusion to approximate the behaviour of a population of size $2N = 50,000$, then according to the space-time rescaling in (3.1.1), we expect that the MRCA lives $50,000 \times 2 = 100,000$ generations (time units) ago.

Proof. Let τ_k be the amount of time during which there are k lineages in the coalescent tree of the time-t population, $k = 2, 3, \ldots$. Then Theorem 3.2.1 states that the expectation of τ_k equals $\mathbb{E}(\tau_k) = \binom{k}{2}^{-1}$. Since the time T_t between the time-t population and its MRCA equals the depth of the coalescent tree, the desired expectation equals

$$\mathbb{E}(T_t) = \mathbb{E}\left(\sum_{k=2}^{\infty} \tau_k\right) = \sum_{k=2}^{\infty} \mathbb{E}(\tau_k) = \sum_{k=2}^{\infty} \frac{2}{k(k-1)} = 2. \tag{5.1.1}$$

□

5.2 MRCA-process and fixation process

Let $t \in \mathbb{R}$ be the observation time of a population in the WF-diffusion, and let T_t be the depth of the corresponding coalescent tree. Then $A_t = t - T_t$ is the time at which the MRCA of the time-t

population lives. We define the *MRCA-process* as the process $(A_t)_{t\in\mathbb{R}}$ on state space $\mathbb{R}$. (Donnelly and Kurtz [3] refer to $(A_t)_{t\in\mathbb{R}}$ as the Eve process.)

Consider the MRCA that lived at time A_t, and the two individuals directly descended from this MRCA. The time-t population is then divided into two disjoint parts, where each part consists of the time-t offspring of one of the two individuals. We refer to these parts as the two oldest families in the population. The time F_t at which one of these families fixates in the population, a new MRCA is established. The process $(F_t)_{t\in\mathbb{R}}$ on state space $\mathbb{R}$ is called the *fixation process*.

Suppose that the points of $(A_t)_{t\in\mathbb{R}}$ are enumerated as $\{\alpha_i\}_{i=-\infty}^{\infty}$. Then the points $\{\beta_i\}_{i=-\infty}^{\infty}$ of $(F_t)_{t\in\mathbb{R}}$ are given by $\beta_i = \inf\{t : A_t = \alpha_i\}$, and the path of $(A_t)_{t\in\mathbb{R}}$ is constant on each interval $[\beta_i, \beta_{i+1})$:

$$\forall\ t \in [\beta_i, \beta_{i+1}) : A_t = \alpha_i. \tag{5.2.1}$$

At each time β_i the MRCA, which lives at time α_i, is thus established.

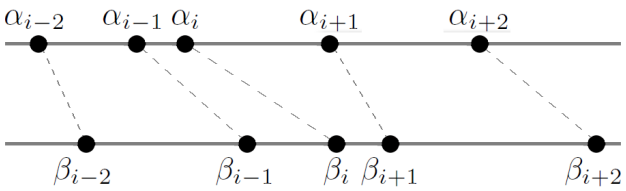

Figure 5: Time is running horizontally to the right. The points α_j and β_j, $i-2 \le j \le i+2$, of the MRCA-process $(A_t)_{t\in\mathbb{R}}$ and the fixation process $(F_t)_{t\in\mathbb{R}}$ are drawn for some $i \in \mathbb{Z}$. The dotted lines relate each time α_j, at which a MRCA lives, to the time β_j, at which the MRCA is established.

5.3 Particle construction

To investigate $(A_t)_{t\in\mathbb{R}}$ and $(F_t)_{t\in\mathbb{R}}$, we use Pfaffelhuber and Wakolbinger [9] and the particle construction of Donnelly and Kurtz [2], which we will now describe.

Identify each element (t, i) of the set $\mathbb{R}\times\mathbb{N}$ with the individual at time t at *level* i. To describe the births of new individuals in the population, define for all levels $i < j$ the *look-down process* $P_{ij} = (P_{ij}(t))_{t\in\mathbb{R}}$ as a rate-1 Poisson process on $\mathbb{R}$. At each point of P_{ij}, level j *looks down* to level i, which means that the individual at level i produces a copy of itself, which is inserted at level j. For each time t_0, when level j looks down to level i, the individuals at levels $j, j+1, \ldots$ are *pushed* one level up to make room for the new individual (t_0, j). As time runs forward, the new individual is in turn pushed one level up each time there is a birth at one of the lower levels. To be precise, the new individual born at level j at time t_0 is pushed one level up at all times $t_1 < t_2 < \ldots$, where t_n is a point of P_{ik} for some $1 \le i < k < j+n$, for all $n \in \mathbb{N}$. The evolution over time of an individual born at level j at time t_0 can thus be described by a *line* of the form

$$L = ([t_0, t_1)\times\{j\}) \cup ([t_1, t_2)\times\{j+1\}) \cup ([t_2, t_3)\times\{j+2\}) \cup \cdots \subset \mathbb{R}\times\mathbb{N}. \tag{5.3.1}$$

See Figure 6 for a graphical illustration. For each individual $(t, i) \in \mathbb{R}\times\mathbb{N}$ there is a unique line L such that $(t, i) \in L$. An individual at level j is pushed one level up at time t if and only if level k looks down to level i at time t for some $1 \le i < k \le j$. This leaves $j(j-1)/2$ possible values for i and k, and therefore pushing rates increase quadratically in j. Hence, $t_\infty = \lim_{n\to\infty} t_n$ is finite with probability 1. We say a line L *exits* at time t_∞. Note that each individual in L is "pushed to infinity" as L exits.

Thus, for each time t_0 when level j looks down to level i an individual (t_0, j) is born. As time runs forward, the individual is pushed up one level at a time until it reaches infinity, where the individual dies and its (unique) line L exits at some finite time t_∞. The random tree consisting of all lines $L \subset \mathbb{R}\times\mathbb{N}$ is known as the *look-down graph*. The term *particle construction* for the construction just described is justified by identifying each individual and its corresponding line with a particle and its trajectory in $\mathbb{R}\times\mathbb{N}$.

Any line L in the look-down graph backwards in time leads to a *coalescence* with another line L'. Indeed, the individual (t_0, j) in L that lives at the earliest time of all the individuals in L lives at a time when level j looks down to level i for some $i < j$. Since individual (t_0, i) has a unique line L' such that $(t_0, i) \in L'$, the lines L and L' coalesce in (t_0, i). Thus, the tracing of the evolution of an individual backwards in time does not stop at the end (or at the beginning in the forward sense) of its line, but leads to an unambiguous and endless trajectory through different lines in the look-down graph. We will call this trajectory the *ancestral lineage* of the individual, which we will now formally construct.

First note that the look-down graph always contains the *immortal line* $\omega = \mathbb{R} \times \{1\}$. Fix $s_0 \in \mathbb{R}$. For each $n \in \mathbb{N}$ there is a unique line $L = L_n$ of the form (5.3.1) such that $(s_0, n) \in L_n$, and a function $\Lambda_{L_n} : [t_0, t_\infty) \to \{j, j+1, \dots\}$, $t \mapsto i$ such that $(t, i) \in L_n$. Next, reverse time and let $X_t^n = \Lambda_{L_n}(s_0 - t)$, $t \in [0, s_0 - t_0]$. As time runs backwards, there comes a time where $t > s_0 - t_0$, so that X_t^n is no longer defined. However, since t_0 is a point of P_{ij} for some $i < j$, there exists an $m < n$ such that $(t_0, i) \in L_m$ and such that X_t^m (in particular) is defined for all $t \in (s_0 - t_0, s_0 - t_0']$, where $t_0' = \inf\{t : \exists j' \in \mathbb{N} : (t, j') \in L_m\}$. Now t_0' is a point of $P_{i'j'}$ for some $i' < j'$. Repeating this argument, we get a sequence $n > m > \dots > 1$ and a sequence of intervals $I_n = [0, s_0 - t_0]$, $I_m = (s_0 - t_0, s_0 - t_0']$, ... such that $\bigcup_i I_i = [0, \infty)$. (Indeed, $((-\infty, s_0] \times \{1\}) \subset L_1$.) Thus, for any individual $(s_0, n) \in \mathbb{R} \times \mathbb{N}$ we can define its *backward level process* $(\Phi_n(t))_{t \geq 0}$ by letting

$$\Phi_n(t) : [0, \infty) \to \mathbb{N},\ t \mapsto X_t^i \quad \text{if } t \in I_i. \tag{5.3.2}$$

The *ancestral lineage* of an individual (s_0, n) equals $(t, \Phi_n(t))_{t \geq 0}$. Note that the ancestral lineage of any individual will eventually coalesce with the immortal line, i.e.:

$$\forall\ n \in \mathbb{N}\ \exists\ t_n^*\ \forall\ t > t_n^* : \Phi_n(t) = 1. \tag{5.3.3}$$

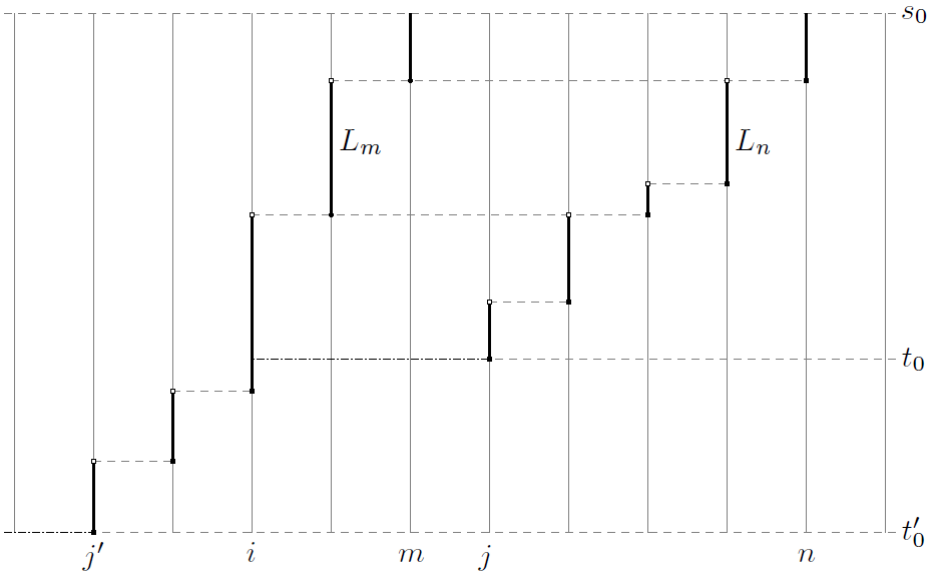

Figure 6: Part of a look-down graph. Biological time is running upwards and level numbers are running to the right. Only (parts of) two lines L_m and L_n are drawn. At time t_0 level j looks down to level i and individual (t_0, j) is born. Each time an individual in line L_m is pushed one level up, the individual in line L_n that lives at the same time is pushed one level up. The backward level process of individual (s_0, n) jumps from level j to level i at time t_0.

The *look-down tree* $\mathcal{T}_s$ consists of all the ancestral lineages $(t, \Phi_n(t))_{t \geq 0}$ of the individuals (s, n), $n = 1, 2, \dots$, for some reference time $s = s_0$.

5.4 Particle construction applied

By the definition of the look-down processes P_{ij}, any two ancestral lineages in $\mathcal{T}_s$ coalesce with rate 1. This means that $\mathcal{T}_s$ has the same distribution as the coalescent (defined in Chapter 4). The particle

construction can therefore be used to describe the evolution of a population in the WF-diffusion. To that end, let the levels of the individuals at any time be *ordered by persistence*, i.e., $i < j$ if and only if the offspring of individual (t, i) outlives the offspring of individual (t, j). Note that with the ordering by persistence, the look-down processes P_{ij}, $i < j$, suffice to describe each birth in the population in the WF-diffusion.

Theorem 5.4.1 *The MRCA-process $(A_t)_{t\in\mathbb{R}}$ is a rate-1 Poisson process on $\mathbb{R}$.*

Proof. It is enough to show that $(A_t)_{t\in\mathbb{R}}$ coincides with P_{12}. First note that a MRCA in the particle construction must always be at level 1. We will prove that $(t, 1)$ is a MRCA if and only if t is a point of P_{12}.

"$\Rightarrow$": Observe a MRCA at time t in the coalescent tree and the two individuals that directly descended from the MRCA. (We can identify these two time-t individuals as the MRCA and its copy.) Then $t = A_s$ for some $s > t$ and the two oldest families in the time-s population have each descended from one of the two individuals. This means that the other individuals in the population at time $t = A_s$ do not have living offspring at time s. The ordering by persistence in the particle construction implies that these other individuals all have higher levels than the MRCA and its copy. Hence, at time A_s the copy of the MRCA is inserted at level 2. i.e., t is a point of P_{12}.

"$\Leftarrow$": Let t be a point of P_{12}. Then individual $(t, 2)$ is a copy of individual $(t, 1)$. The ordering by persistence implies that there exists an $s > t$ such that the time-s population has descended from the individual $(t, 1)$ and its copy $(t, 2)$, i.e., $(t, 1)$ is the MRCA of the time-s population. □

The proof of Theorem 5.4.1 implies that each point t_0 of P_{12} initiates a line L_F^i of the form (5.3.1) where $j = 2$ and $t_0 = \alpha_i$ for a certain $i \in \mathbb{Z}$. Such a line L_F^i is called a *fixation line*. Let ξ_i denote the exit time of L_F^i. Note that at time ξ_i the offspring of individuals (α_i, k), $k = 3, 4, \ldots$, goes extinct. Therefore we have that ξ_i equals the infimum of times t at which $(\alpha_i, 1)$ is the MRCA of the time-t population, thus it follows that the exit times of the fixation lines coincide with the points of $(F_t)_{t\in\mathbb{R}}$, i.e., $\xi_i = \beta_i$ for all $i \in \mathbb{Z}$.

Let A_t^n be the time at which the MRCA of the individuals (t, k), $k = 1, \ldots, n$, lives. Since the MRCA of the first $n + 1$ time-t individuals is a common ancestor of the first n time-t individuals, we have $A_t^n \geq A_t^{n+1}$ for all $n \in \mathbb{N}$. Define $\Theta_n = (\Theta_n(t))_{t\in\mathbb{R}}$ by letting

$$\Theta_n(t) = \begin{cases} 1 & \text{if } A_t^n > A_t^{n+1} \\ 0 & \text{if } A_t^n = A_t^{n+1} \end{cases} \tag{5.4.1}$$

and let $N_n(t)$ denote the number of times that Θ_{n-1} jumps from 1 to 0 during the time interval $[0, t]$, $n > 1$. Then $N_n = (N_n(t))_{t\in\mathbb{R}}$ is a rate-1 Poisson process (see Donnelly and Kurtz [3], Lemma 3.5). We will use the particle construction to show that this implies the following theorem.

Theorem 5.4.2 *The fixation process $(F_t)_{t\in\mathbb{R}}$ is a rate-1 Poisson process on $\mathbb{R}$.*

Proof. Define $\beta_i^n = \inf\{t : A_t^n = \alpha_i\}$. First we prove that the times $\{\beta_i^n\}_{i=-\infty}^{\infty}$ coincide with the jump times of N_n. Note that β_i^n is the time at which the fixation line L_F^i reaches level n: $\beta_i^n = \inf\{t : (t, n) \in L_F^i\}$. At this time, the MRCA of the first $n - 1$ individuals equals the MRCA of the first n individuals, so that $\Theta_{n-1}(t)$ jumps from 1 to 0 at time $t = \beta_i^n$, i.e., N_n jumps at time β_i^n. On the other hand, if N_n jumps at time t, then Θ_{n-1} jumps from 1 to 0 at time t, so t must be the infimum of times at which the MRCA of the first $n - 1$ individuals equals the MRCA of the first n individuals. This implies that a fixation line L_F^i reaches level n at time $t = \beta_i^n$. It follows that the times $\{\beta_i^n\}_{i=-\infty}^{\infty}$ are the jump times of the rate-1 Poisson process N_n. Next, since the points of $(F_t)_{t\in\mathbb{R}}$ are given by $\beta_i = \lim_{n\to\infty} \beta_i^n$, the theorem follows. □

A more intuitive, but less formal argument for the fact that $(F_t)_{t\in\mathbb{R}}$ is a rate-1 Poisson process is the following. First note that *if* $(F_t)_{t\in\mathbb{R}}$ is a Poisson process, then its rate equals the rate of $(A_t)_{t\in\mathbb{R}}$.

This is a consequence of Theorem 5.1.1: the expected time between a population and its MRCA is constant. It remains to show that the time between jumps of the MRCA is exponential. Let X_t and $1 - X_t$ denote the sizes of the two oldest families in the population at time t. Fix $t_0 \in \mathbb{R}$. There exists an $i \in \mathbb{Z}$ such that $t_0 \in [\beta_i, \beta_{i+1})$. The random reproduction in the WF-model implies that X_{t_0} is uniformly distributed on $[0, 1]$. This also holds conditioned on $A_{t_0} = \alpha_i$. So we have that X_t is uniformly distributed on $[0, 1]$ for all $t \in [\beta_i, \beta_{i+1})$. At time β_{i+1}, when the MRCA jumps from α_i to α_{i+1}, one of the oldest families dies out and two new oldest families come into existence. These two families are again of random size: for $t_1 \in [\beta_{i+1}, \beta_{i+2})$ we have that X_{t_1} (conditioned on $A_{t_1} = \alpha_{i+1}$) is (standard) uniformly distributed. This implies a memoryless waiting time between jumps of the MRCA, so that the time between jumps of the MRCA is (standard) exponentially distributed.

To derive Theorem 5.1.1 from the particle construction, let $i \in \mathbb{Z}$ and consider the fixation line L_F^i starting at time α_i. The infimum T_i of times t for which individual $(\alpha_i, 1)$ is a common ancestor of the population at time $\alpha_i + t$ equals the time it takes for L_F^i to exit. At any level $k \geq 2$ of L_F^i there are $\binom{k}{2}$ rate-1 Poisson processes to initiate a push to level $k + 1$, so it follows that $\mathbb{E}(T_i) = \sum_{k=2}^{\infty} \binom{k}{2}^{-1} = 2$. Now let $t \in \mathbb{R}$. Since $T_i = \beta_i - \alpha_i$, Theorem 5.1.1 holds if and only if $\mathbb{E}(\beta_i - \alpha_i) = \mathbb{E}(t - A_t)$. This may seem counter-intuitive because $A_t = \alpha_i$ if and only if $t \in [\beta_i, \beta_{i+1})$, but it can be explained via the so- called waiting-time paradox: a time point is more likely to fall in a long interval than in a short interval between Poisson events.

References

[1] N. Berestycki, Recent process in coalescent theory, Ens. Mat. 16 (2009) 1–193.

[2] P. Donnelly and T.G. Kurtz, Particle representations for measure-valued population models, Ann. Prob. 27 (1999) 166–205.

[3] P. Donnelly and T.G. Kurtz, The Eve process, unpublished manuscript.

[4] R. Durrett, *Probability Models for DNA Sequence Evolution* (2nd. ed.), Springer, Berlin, 2005.

[5] S.N. Ethier and T.G. Kurtz, *Markov Processes; characterization and convergence*, John Wiley Sons, New York, 1986.

[6] W.Th.F. den Hollander, *Stochastic Models for Genetic Evolution*, Lecture Notes, 2010.

[7] J.F.C. Kingman, On the genealogy of large populations, J. Appl. Prob. 19 (1982) 27–43.

[8] J.F.C. Kingman, The coalescent, Stoch. Proc. Appl. 13 (1982) 235–248.

[9] P. Pfaffelhuber and A. Wakolbinger, The process of most recent common ancestors in an evolving coalescent, Stoch. Proc. Appl. 116 (2006) 1836–1859.

[10] F. Redig, *Diffusion Limits in Population Models*, Lecture Notes, 2006.