

STOCHASTIC MODELS FOR GENETIC EVOLUTION

Luca Avena, Conrado da Costa, Frank den Hollander

Mathematical Institute, Leiden University,
P.O. Box 9512, 2300 RA Leiden, The Netherlands

January 2019

PREFACE

The goal of this course is to present a series of stochastic models from population dynamics capable of describing rudimentary aspects of DNA sequence evolution. Most of the course focuses on the Wright-Fisher model and its variations, describing a population of individuals (= genes) of different types (= alleles) organised into a single colony and subject to resampling, mutation and selection, as well as populations organised into multiple colonies subject to migration. These models capture core phenomena arising in population genetics, including the genealogy of populations.

The course assumes basic knowledge of probability theory, but not of population genetics.

Key words: Evolutionary forces (resampling, mutation, selection, migration, recombination), Wright-Fisher model, Moran-model, Wright-Fisher diffusion, duality, Kingman coalescent, most recent common ancestor, stepping stone model, hierarchically interacting Wright-Fisher diffusions, renormalization, universality.

Contents

1	Genetic background	5
1.1	DNA	5
1.2	Genetic evolution	7
1.3	What will be covered in the course?	7
I	Single-colony Wright-Fisher populations	9
2	Wright-Fisher: basic properties	10
2.1	Standard model	10
2.1.1	WF-model: parallel resampling	10
2.1.2	Moran-model: sequential resampling	14
2.2	Space-time scaling limit of WF-model and Moran-model	14
2.2.1	Scaling of WF-model	14
2.2.2	Scaling of Moran-model	17
2.2.3	WF-diffusion	18
2.3	Dual to WF-diffusion	19
3	Wright-Fisher: genealogy	22
3.1	Kingman coalescent	22
3.1.1	The finite coalescent	23
3.1.2	The infinite coalescent	26
3.1.3	Coming down from infinity	27
3.2	Most recent common ancestor	28
3.2.1	Depth of the coalescent tree	28
3.2.2	MRCA-process and F-process	29
4	Wright-Fisher with mutation	30
4.1	WF-model with mutation and two types	30
4.1.1	Mutation	30
4.1.2	Weak mutation	32
4.2	WF-model with mutation and infinitely many types	35
4.2.1	Infinite alleles model	35
4.2.2	Ewens sampling formula	38
4.2.3	Griffiths-Engen-McCloskey distribution	39
4.3	WF-model with mutation and infinitely many sites	43

5	Wright-Fisher with selection	45
5.1	Selection	45
5.2	Weak selection	48
II	Multi-colony Wright-Fisher populations	50
6	The stepping stone model	51
6.1	Model	51
6.2	Lineages	52
6.3	One-dimensional geographic space	53
6.3.1	One-dimensional lattice	53
6.3.2	One-dimensional lattice torus	54
6.4	Two-dimensional geographic space	55
6.4.1	Two-dimensional lattice	55
6.4.2	Two-dimensional lattice torus	56
6.4.3	Coalescing times	57
6.4.4	Without mutation	58
6.4.5	Panmictic behaviour for the two-dimensional torus	59
7	Hierarchical models	60
7.1	Hierarchically interacting diffusions	61
7.2	Hierarchy of space-time scales	63
7.3	Hierarchical mean-field limit	64
7.4	Renormalisation transformation	65
7.5	Conclusion	69
A	Appendix	69
A.1	Special probability distributions	69
A.2	Elementary stochastic processes	70
A.2.1	Markov chains	70
A.2.2	SDE and Martingale Problem	72
A.2.3	Poisson process	72
A.2.4	Random walk	72
A.2.5	Brownian motion	73
B	MRCA-process and F-process	73
B.1	Look-down construction	73
B.2	Look-down construction applied	75
B.2.1	Identification of the MRCA-process	75
B.2.2	Identification of the F-process	76

1 Genetic background

Section 1.1 gives a brief sketch of the role of DNA in genetic evolution, Section 1.2 describes the five main evolutionary forces, while Section 1.3 provides an outline of the course. For further background, see the website [www.dnaftb.org].

1.1 DNA

The hereditary information of most living organisms is carried by *DNA molecules*. DNA usually consists of two complementary chains twisted around each other to form a double helix (see Fig. 1). Each chain is a linear sequence of four different *nucleotides*:

A = adenine
C = cytosine
G = guanine
T = thymine

These nucleotides pair in the combinations $A - T$ and $C - G$ via hydrogen bonds. The binding energies in the two pairs are different. Each pair has a size of about $20\text{\AA}$, where $1\text{\AA} = 10^{-10}\text{m}$. The order in which the pairs occur is specific to each living organism and is referred to as the *genome* of that organism.

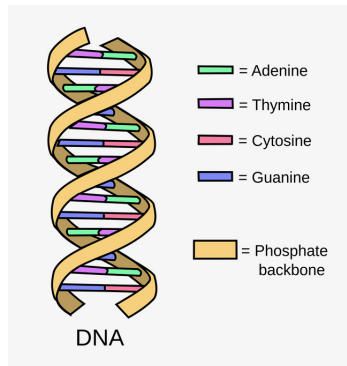


Figure 1: DNA double helix.

The yeast genome, for instance, consists of a sequence of 1.2×10^7 base pairs in which the nucleotides occur with approximate frequencies

A: 0.3090
C: 0.1917
G: 0.1913
T: 0.3078
——— +
0.9998

Note that the pairs $A - T$ and $C - G$ do not occur with equal frequency. The human genome consists of 3×10^9 base pairs, the longest polymer chain known to date. When folded out, the human DNA molecule is 6m long!

DNA is situated in the *chromosomes*, which reside in the *nucleus* of the cells (see Fig. 2). DNA plays the role of a *genetic codebook*. Embedded in the long string of base pairs that make up the genome there are so-called (protein-coding) *genes*. Each gene itself consists of a large number of base pairs. Genes come in different types, called *alleles*. In the human genome are embedded 6×10^4 different genes, pieces of DNA that play a specific genetic role. Apparently, some of the information encoded in DNA serves no purpose (and is to be interpreted as noise).



Figure 2: Copies of the DNA double helix in the cell.

Genes are *transcribed* into (messenger) RNA, which subsequently is *translated* into a multitude of different proteins. Both these mechanisms are highly complex. (RNA uses uracil *U* instead of thymine *T* and is single-stranded.) Proteins perform a multitude of tasks: they are the workhorses of life. *Amino acids* are the basic structural units of *proteins*. All proteins in all organisms, from bacteria to humans, are constructed from 20 different amino acids.

Alleles compete for a specific location on the chromosome, called *locus*. This competition is not direct, but rather takes place via their *phenotypical characteristics* (like eye sight, muscle strength, fur color): the manifestation of genes at the level of anatomy, physiology and behaviour. The fate of a gene is linked to the bodies in which it resides during successive generations, according to Darwin's *survival of the fittest* principle.

Remark:

- (1) The *genotype* of a cell or an organism is the specific makeup of its alleles. The genotype differs from the genomic sequence: while the genomic sequence is an absolute measure of the base pair composition, the genotype typically is a measure of how an organism differs within a group of organisms. The genotype of an organism is the inherited instructions it carries within its genetic code.
- (2) The *phenotype* is the composite of an organism's observable characteristics or traits, such as its morphology, development, biochemical and physiological properties, and behaviour. Phenotypes result from the *expression* of an organism's genes, as well as from the influence of *environmental factors*. Not all organisms with the same genotype look or act the same way, because appearance and behaviour are modified by environmental and developmental conditions. Likewise, not all organisms that look alike necessarily have the same genotype.

1.2 Genetic evolution

DNA is subject to different types of evolutionary forces. One type is *resampling*, in which genetic material is passed on from one generation to the next. Lower organisms (such as bacteria) are *haploid*, i.e., the chromosomes carry only one copy of the genetic material. Most higher organisms (such as humans) are *diploid*, i.e., the chromosomes carry two copies of the genetic material. (Some plants have more than two copies.) When haploid individuals reproduce, there is *one parent* that passes one copy of its genetic material to its offspring. When diploid individuals reproduce, there are *two parents* each of which passes one copy to its offspring. Before being passed on, the two copies of a chromosome may exchange genetic material, a phenomenon that is called *recombination*.

Another type of evolution is *mutation*, corresponding to a spontaneous local change in the content or the order of the base pairs. For instance, nucleotides may be substituted by others: the substitutions $A \leftrightarrow G$ and $C \leftrightarrow T$ are called *transitions*, all other substitutions are called *transversions*. Mutations occur all the time, but are mostly repaired by *enzymes*. The non-repaired mutations are called *deleterious mutations*, which may or may not be lethal. It turns out that, in the nucleus, transitions occur 10–20 times more frequently than transversions (in mitochondria, which are membrane-enclosed organelles found in most eukaryotic cells, they occur at more or less the same frequency). It is more probable for a transversion to be a deleterious mutation than for a transition.

Yet another type of evolution is *selection*. This means that different alleles may have different predispositions for *resampling*, for instance, different rates at which they resample. A further ingredient may be *migration*, i.e., genetic material is transmitted between different populations because the individuals carrying this material travel from one population to the next.

1.3 What will be covered in the course?

In this course we will focus on the so-called *Wright-Fisher model* for the evolution of a population of genes. In Part I (Chapters 2–5) we consider the single-colony Wright-Fisher model, describing a population of genes subject to *resampling*, *mutation* and *selection*. In Part II (Chapters 6–7) we look at the multiple-colony Wright-Fisher model, describing a population that additionally is subject to *migration*. We will not be including *recombination*, since this is much harder to analyse.

There are other interesting aspects of DNA evolution that we are also not addressing, e.g. the effects of damage in DNA, or the fact that DNA has denatured pieces, i.e., pieces where the two strings are locally detached as a result of the absorption of heat. The Wright-Fisher model is *extremely rudimentary*, yet it captures a number of key aspects of genetic evolution and serves as a jump board for much more sophisticated analysis. The literature is vast. Monographs that focus on the biological background are Hein, Schierup and Wiuf [15], Hartl and Clark [14] and Wakeley [21]. A monograph with a more mathematical flavour is Durrett [10].

Throughout the course we need some basic facts about special probability distribu-

tions (binomial, geometric, Poisson, exponential, beta, gamma) and elementary stochastic processes (discrete-time and continuous-time Markov chains, Poisson point processes, random walk, Brownian motion). A brief list is provided in Appendix A. The reader needs to familiarise him/herself with key probabilistic notions, which can be found in any textbook on probability theory and stochastic processes.

Appendix B looks at two key processes for the genealogy of populations described in Section refS3. Throughout the lecture notes there are exercises to help the reader grasp the material by working out the details of some of the steps in the presentation.

Part I

Single-colony Wright-Fisher populations

Part I is devoted to the evolution of a *single* population subject to *resampling*, *mutation* and *selection*.

In Chapter 2 we look at the standard Wright-Fisher model, where only resampling takes place. We first consider finite populations consisting of individuals of two types evolving under neutral resampling, and derive a few basic properties. After that we pass to the *scaling limit* where the size of the population is taken to infinity and time is scaled appropriately. This leads to a limiting process, called the Wright-Fisher diffusion, which we analyze in some detail. Next, we show that this limiting process has a *dual* that allows for easy computations.

In Chapter 3 we show that the dual process can be understood in terms of a coalescent process, called the *Kingman coalescent*, which captures the *genealogy* of finite samples drawn from a large population, i.e., it looks backwards in time and describes the ancestral relationships within the population. We also look at the *most recent common ancestor* of the population, and determine how far back in time this ancestor lived. The presentation of the material in Sections 3.1–3.2 is taken from the bachelor thesis of Maurits Carsouw [4].

In Chapters 4–5 we look at two variations of the standard Wright-Fisher model obtained by adding mutation, respectively, selection. In the case of weak mutation, respectively, weak selection, the large space-time scaling limit can again be analyzed in some detail. We also look at statistical methods that allow for an estimation of mutation and selection parameters.

2 Wright-Fisher: basic properties

In this chapter we introduce the Wright-Fisher model. The simplest version of this model dates back to the 1940's, i.e., before the discovery of the double-helix structure of DNA by Crick and Watson in 1950 (which provides the *molecular* basis of evolution). However, it was known much earlier that DNA is the carrier of hereditary information. The standard WF-model is used to describe the evolution of a population of individuals (= genes) of two different types (= alleles), called A and a . These types are *neutral*, i.e., their reproductive success does not depend on the type, and their reproduction is *random*. In Section 2.1 we describe the model and derive a few of its key properties. In Section 2.2 we look at a continuum limit of the standard WF-model that is obtained by scaling space and time. The limit is called the WF-diffusion, and is computationally more tractable. In Section 2.3 we show that the WF-diffusion is dual to a death process.

2.1 Standard model

2.1.1 WF-model: parallel resampling

Consider N *haploid* individuals, each carrying 1 copies of a specific *genetic locus* (a location of interest in the genome). Each individual (= gene) can be of two *types* (= alleles), A and a , which correspond to two different pieces of genetic information at the same locus.

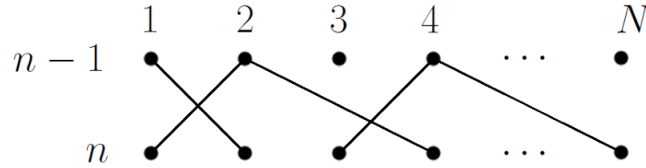


Figure 3: Resampling in the WF-model: the types at time $n - 1$ determine the types at time n .

Suppose that at each time unit each individual *randomly chooses another individual (possibly itself) from the population and adopts its type* (“parallel updating”). This is called *resampling*, and is a form of random reproduction (see Fig. 3). Suppose that all individuals update independently from each other and independently of how they updated at previous times. We are interested in the evolution of the following quantity:

$$X_n = \text{number of } A\text{'s at time } n. \quad (2.1)$$

Note that the total population size N is fixed. In what follows we will be mainly interested in the case $N \gg 1$. Later we will even pass to the limit $N \rightarrow \infty$.

Remark: For humans, one unit of time corresponds to one generation, so roughly 20 years. For plants, one unit of time typically is 1 year, for bacteria typically a few days or a few hours.

Remark: The WF-model is a model for haploid individuals. However, it may reasonably be used for diploid individuals as well, by considering a population of size $2N$. Some constraints need to be imposed on the resampling, because even in a genderless reproduction an individual cannot inherit two copies from the same ancestor.)

Remark: The biology behind the WF-model is the *gene pool approach*. Every individual produces a large number of *gametes* of the same type as the individual itself. (A gamete is a cell that fuses with another cell during fertilisation in organisms that reproduce sexually.) The offspring generation is then formed by sampling N times without replacement from this *gene pool*. This is basically the same as sampling with replacement from the parent population, so that effectively the offspring individuals choose their parents from the parent population with replacement and inherit their type.



Figure 4: Ω , the state space of the Wright-Fisher model.

Throughout the course we write $\mathbb{N}$ for the set of positive integers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ for the set of non-negative integers. The sequence $X = (X_n)_{n \in \mathbb{N}_0}$ is the discrete-time Markov chain on the state space

$$\Omega = \{0, 1, \dots, N\} \quad (2.2)$$

(see Fig. 4 and Appendix A.2.1) with transition kernel

$$p(i, j) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(\frac{N-i}{N}\right)^{N-j}, \quad i, j \in \Omega. \quad (2.3)$$

Indeed, given that at time n the number of individuals of type A equals i , in order to get j individuals of type A at time $n+1$, precisely j individuals have to choose an individual of type A and $N-j$ individuals have to choose an individual of type a . The latter two events occur with probabilities given by the second and the third factor. The number of ways to choose j individuals from N is given by the first factor. As initial condition we may pick any $X_0 \in \Omega$, e.g. $X_0 = \frac{N}{2}$ (= half of the population has type A , the other half has type a).

The states 0 and N are *traps*: $p(0, 0) = p(N, N) = 1$. Since all other states communicate, the process eventually gets trapped, in which case all individuals have the same type (all A or all a).

Eventually genetic variability is lost through chance.

This fact is an important consequence of Darwinian evolution based on chance!

Remark: Figures 7.4–7.5 in Hartl and Clark [14] show loss of genetic variability found in 107 populations of *Drosophila melanogaster* with $N = 16$ diploid entities and $n = 19$ generations.

We are interested in computing the time until fixation, i.e., the stopping time

$$\tau = \inf\{n \in \mathbb{N}_0 : X_n = 0 \text{ or } X_n = N\}, \quad (2.4)$$

as well as the probability of fixation at N , i.e., $X_\tau = N$. We also want to understand how the process behaves prior to τ . The answer will depend on N and X_0 , and is formulated in Lemmas 2.1–2.2 below.

Lemma 2.1 $\mathbb{P}(X_\tau = N \mid X_0 = i) = \frac{i}{N}$.

Proof. Abbreviate $p_i = \frac{i}{N}$. Then (2.3) says that $p(i, \cdot) = \text{BIN}(N, p_i)(\cdot)$, the binomial distribution with N trials and success probability p_i . Hence

$$\mathbb{E}(X_{n+1} \mid X_n = i) = Np_i = i = X_n. \quad (2.5)$$

What this says is that our Markov chain X is a *martingale*, i.e., a random process without bias (see Appendix A.2.2). Since the state space Ω is finite, we have $\mathbb{P}(\tau < \infty) = 1$ and

$$\lim_{n \rightarrow \infty} X_n = X_\tau \quad a.s. \quad (2.6)$$

Next, iterating (2.5), we may write

$$i = \mathbb{E}(X_n \mid X_0 = i) = \mathbb{E}(X_\tau 1\{\tau \leq n\} \mid X_0 = i) + \mathbb{E}(X_n 1\{\tau > n\} \mid X_0 = i), \quad (2.7)$$

where we use that $X_n = X_\tau$ when $\tau \leq n$. Now let $n \rightarrow \infty$ and use the dominated convergence theorem, to see that the first term tends to $\mathbb{E}(X_\tau \mid X_0 = i)$ and the second term tends to zero. Thus, we have

$$\mathbb{E}(X_\tau \mid X_0 = i) = i. \quad (2.8)$$

But

$$\mathbb{E}(X_\tau \mid X_0 = i) = 0 \times \mathbb{P}(X_\tau = 0 \mid X_0 = i) + N \times \mathbb{P}(X_\tau = N \mid X_0 = i). \quad (2.9)$$

Combine (2.8–2.9) to get the claim. ■

To investigate $\mathbb{E}(\tau)$, we consider the quantity

$$H_n = \frac{2X_n(N - X_n)}{N(N - 1)}. \quad (2.10)$$

This is called the *genetic variability* of the population at time n and equals the probability that two different individuals randomly drawn from the population at time n are of different type. Since $H_\tau = 0$, the quantity $1 - H_n$ is sometimes called the fixation index at time n .

Lemma 2.2 $\mathbb{E}(H_n \mid H_0) = (1 - \frac{1}{N})^n H_0$, $n \in \mathbb{N}_0$.

Proof. Randomly draw two individuals at time n . Draw the two *backward ancestral paths* of these two individuals, i.e., the paths labelling the individuals that were chosen as ancestors at all previous times (see Fig. 5). These paths, which are random, contain the full *genealogical history* of the two individuals. When traced backwards in time, these paths behave as two *coalescing random walks* on $\{1, \dots, N\}$, the labelling space of the population: they jump randomly and merge into one upon meeting. At each unit of time (in the backward sense), the two random walks have probability $\frac{1}{N}$ to meet, since they jump uniformly between the labels. The probability that they do not coalesce up to time n (which is time 0 in the forward sense) equals $(1 - \frac{1}{N})^n$. Clearly, the two individuals at time n are of different type if and only if their ancestral lineages do not coalesce and at time 0 are of different type. The latter has probability H_0 , the genetic variability of the initial state X_0 . ■

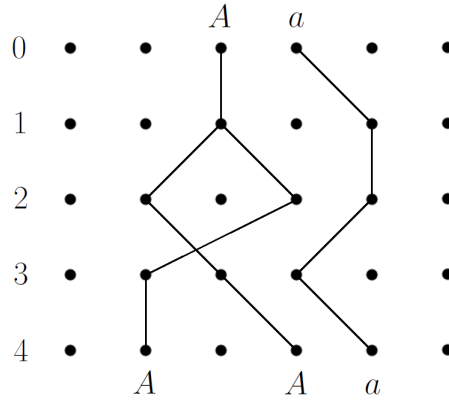


Figure 5: Example of backward ancestral paths for $N = 6$ and $n = 4$. Time runs downwards.

The proof of Lemma 2.2 shows that

$$\mathbb{P}(\tau > n \mid H_0) = \mathbb{P}(H_n \neq 0 \mid H_0) \geq \left(1 - \frac{1}{N}\right)^n H_0, \quad n \in \mathbb{N}_0. \quad (2.11)$$

Exercise 2.3 Check (2.11).

Hence $\mathbb{E}(\tau \mid H_0) = \sum_{n \in \mathbb{N}_0} \mathbb{P}(\tau > n \mid H_0) \geq NH_0$. Thus, fixation is slow in large populations. In Section 2.2 we will see that $\mathbb{E}(\tau \mid H_0)$ is proportional to NH_0 as $N \rightarrow \infty$, a property that goes under the name of *inbreeding*.

The time until genetic variability is lost is proportional to the size of the population.

2.1.2 Moran-model: sequential resampling

There is a *continuous-time* version of the WF-model, called the *Moran-model*, in which each individual chooses a random ancestor at rate 1 and adopts its type. In other words, the resampling is done *sequentially* rather than in parallel. The resulting process $X = (X_t)_{t \geq 0}$ is the *birth-death process* on the state space Ω with transition rates

$$\begin{aligned} i \rightarrow i+1 & \text{ at rate } b_i = (N-i) \frac{i}{N}, \\ i \rightarrow i-1 & \text{ at rate } d_i = i \frac{N-i}{N}. \end{aligned} \quad (2.12)$$

(See Appendix A.2.1.) Note that $b_i = d_i$, $i \in \Omega$, and $b_0 = d_0 = b_N = d_N = 0$.

The differences between the Moran-model and the WF-model are minor. In particular, in Section 2.2 we will see that after space-time scaling both models converge to the same limit, called the WF-diffusion, with the sole difference that the Moran-model runs at *twice* the speed of the WF-model.

Exercise 2.4 *Derive the analogues of Lemmas 2.1–2.2 for the Moran-model.*

Remark: The Moran-model can alternatively be viewed as follows: Each individual produces a new individual at rate 1. Each new individual inherits the type of its parent and replaces a randomly chosen individual (possibly its parent).

A straightforward extension of the Moran-model is obtained when the rate at which the individuals choose their ancestor is not 1 but $h(y)$, where y is the *fraction* of type A and $h: [0, 1] \rightarrow [0, \infty)$ is any Lipschitz function with $h(0) = h(1) = 0$. In other words, b_i and d_i are multiplied by $h(i/N)$. The function h plays the role of an *overall rate of resampling*, and models an effect the population size may have on the evolution rate. A classical example is:

$$h(y) = y(1-y) \quad \text{Ohta-Kimura model.} \quad (2.13)$$

This corresponds to an overall rate that is proportional to the genetic variability of the population. Thus, when the population is close to fixation it loses the incentive to reproduce.

2.2 Space-time scaling limit of WF-model and Moran-model

2.2.1 Scaling of WF-model

Lemma 2.2 suggests that it is of interest to consider the following *space-time rescaling* of our process:

$$Y_t^{(N)} = \frac{1}{N} X_{\lceil Nt \rceil}^{(N)}, \quad t \geq 0. \quad (2.14)$$

Here, $\lceil \cdot \rceil$ denotes the upper integer part, an upper index (N) is added to exhibit the underlying N -dependence, space is shrunk by a factor N , while time is blown up by a factor N . Note that (2.14) represents the *fraction* of individuals of type A in the

population at time t on time scale N , i.e., the time scale on which genetic variability is lost. We expect that, in the limit as $N \rightarrow \infty$, if the initial condition scales properly, i.e.,

$$w - \lim_{N \rightarrow \infty} Y_0^{(N)} = Y_0, \quad (2.15)$$

then the whole process scales properly, i.e.,

$$w - \lim_{N \rightarrow \infty} (Y_t^{(N)})_{t \geq 0} = (Y_t)_{t \geq 0}. \quad (2.16)$$

Here, $w - \lim_{N \rightarrow \infty}$ denotes weak limit, i.e., convergence in distribution on path space. In other words, we expect the rescaled process to converge to a limiting process, living on state space $[0, 1]$ and evolving in continuous time. This limiting process, which must also be a Markov process, turns out to be a *diffusion* (see Appendix A.2.5).

Theorem 2.1 *The scaling in (2.16) subject to (2.15) is true, with $Y = (Y_t)_{t \geq 0}$ the diffusion process on $[0, 1]$ given by the stochastic differential equation (SDE)*

$$dY_t = \sqrt{Y_t(1 - Y_t)} dW_t, \quad (2.17)$$

where $(W_t)_{t \geq 0}$ is standard Brownian motion. This SDE has a unique strong solution, i.e., there is a unique path $t \mapsto Y_t$ that is measurable w.r.t. the canonical filtration associated with the Brownian motion.

Proof. The process $Y^{(N)} = (Y_t^{(N)})_{t \geq 0}$ can be approximated by a continuous-time analogue: instead of waiting a deterministic time $1/N$ before making a transition, wait an exponential time of rate N . This gives us a Markov process with state space $\{0, \frac{1}{N}, \dots, 1 - \frac{1}{N}, 1\}$ and infinitesimal generator L_N given by

$$(L_N f) \left(\frac{i}{N} \right) = N \sum_{j=0}^N p_N(i, j) \left[f \left(\frac{j}{N} \right) - f \left(\frac{i}{N} \right) \right]. \quad (2.18)$$

The limiting process $Y = (Y_t)_{t \geq 0}$ is the diffusion with state space $[0, 1]$ and infinitesimal generator L given by

$$(Lf)(y) = \frac{1}{2}y(1 - y)f''(y). \quad (2.19)$$

Exercise 2.2 *Check that (2.18) and (2.19) are the correct infinitesimal generators (see Appendices A.2.1 and A.2.5).*

In (2.18)–(2.19), f is an appropriate test function. It is enough to prove convergence of generators (see Ethier and Kurtz [11], Chapter 10, Theorem 1.1), i.e.,

$$\lim_{N \rightarrow \infty} (L_N f) \left(\frac{i_N}{N} \right) = (Lf)(y) \quad \text{when} \quad \lim_{N \rightarrow \infty} \frac{i_N}{N} = y. \quad (2.20)$$

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

$$C_0([0, 1]) = \{f \in C([0, 1]): f(0) = f(1) = 0\}. \quad (2.21)$$

Since the “local speed” of the WF-diffusion is given by the *diffusion function* $g: [0, 1] \rightarrow [0, \infty)$ with $g(y) = y(1-y)$, it is possible to show that the domain $D(L)$ of the generator L of the WF-diffusion must be a subset of $C_0([0, 1])$. For general Markov processes it is not easy to characterise $D(L)$, but it suffices to consider the action of L on a subset $K(L) \subset D(L)$ that is large enough to maintain the generality of the argument. We refer to $K(L)$ as a *core* of the generator and require that $K(L)$ is dense in $C_0([0, 1])$ with respect to the supremum norm. Then the required generality is obtained via continuous extension.

In our case it suffices to consider the functions in $C_0([0, 1])$ that are infinitely differentiable:

$$K(L) = \{f \in C_0([0, 1]): f \text{ is infinitely differentiable}\}. \quad (2.22)$$

This choice 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

$$(L_N f)\left(\frac{i}{N}\right) = \sum_{j=0}^N p_N(i, j)(j-i) f'\left(\frac{i}{N}\right) + \frac{1}{2} \sum_{j=0}^N p_N(i, j) \frac{(j-i)^2}{N} f''\left(\frac{i}{N}\right) + R_N, \quad (2.23)$$

where R_N consists of third- and higher-order terms. Now put $Y_0 = y$ and $X_0^{(N)} = i = i_N$, so that (2.15) becomes

$$\lim_{N \rightarrow \infty} \frac{i_N}{N} = y \in [0, 1]. \quad (2.24)$$

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

$$\begin{aligned} F(y) &= \lim_{N \rightarrow \infty} \sum_{j=0}^N p_N(i, j)(j-i), \\ G(y) &= \lim_{N \rightarrow \infty} \sum_{j=0}^N p_N(i, j) \frac{(j-i)^2}{N}. \end{aligned} \quad (2.25)$$

Since R_N vanishes as $N \rightarrow \infty$, the limiting generator equals

$$\begin{aligned} &\lim_{N \rightarrow \infty} (L_N f)\left(\frac{i}{N}\right) \\ &= \lim_{N \rightarrow \infty} \left(\sum_{j=0}^N p_N(i, j)(j-i) f'\left(\frac{i}{N}\right) + \frac{1}{2} \sum_{j=0}^N p_N(i, j) \frac{(j-i)^2}{N} f''\left(\frac{i}{N}\right) \right) \\ &= F(y) f'(y) + \frac{1}{2} G(y) f''(y), \end{aligned} \quad (2.26)$$

where the last equality uses (2.24) 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 indeed is (2.19).

From the martingale property noted in the proof of Lemma 2.1, it follows that $\mathbb{E}(X_1^{(N)}) = \mathbb{E}(X_0^{(N)}) = i_N = i$. We have

$$\mathbb{E}(X_0^{(N)}) = i \sum_{j=0}^N p_N(i, j), \quad \mathbb{E}(X_1^{(N)}) = \sum_{j=0}^N j p_N(i, j). \quad (2.27)$$

Hence

$$F(y) = \lim_{N \rightarrow \infty} \sum_{j=0}^N p_N(i, j)(j - i) = \lim_{N \rightarrow \infty} 0 = 0. \quad (2.28)$$

Moreover, $\mathbb{E}(X_1^{(N)}) = i_N = i$ and so we have

$$\mathbb{V}\mathbb{A}\mathbb{R}(X_1^{(N)}) = \sum_{j=0}^N p_N(i, j)(j - i)^2. \quad (2.29)$$

Since $X_1^{(N)}$ has distribution $\text{BIN}(N, \frac{i}{N})$, it follows that

$$\mathbb{V}\mathbb{A}\mathbb{R}(X_1^{(N)}) = N \frac{i}{N} \left(1 - \frac{i}{N}\right). \quad (2.30)$$

Combining (2.29–2.30), we get

$$G(y) = \lim_{N \rightarrow \infty} \sum_{j=0}^N p_N(i, j) \frac{(j-i)^2}{N} = \lim_{N \rightarrow \infty} \frac{i_N}{N} \left(1 - \frac{i_N}{N}\right) = y(1 - y), \quad (2.31)$$

and so the desired result follows. ■

2.2.2 Scaling of Moran-model

A similar computation works for the Moran-model defined in Section 2.1.2. The rescaled Moran-model is the birth-death process with state space $\{0, \frac{1}{N}, \dots, 1 - \frac{1}{N}, 1\}$ and infinitesimal generator $\hat{L}_N$ given by

$$(\hat{L}_N f) \left(\frac{i}{N}\right) = N \frac{i}{N} (N - i) \left[f \left(\frac{i-1}{N}\right) + f \left(\frac{i+1}{N}\right) - 2f \left(\frac{i}{N}\right) \right]. \quad (2.32)$$

Exercise 2.3 Check that (2.32) is the correct infinitesimal generator. *Hint: The first factor N is a consequence of the rescaled time, the second factor $\frac{i}{N}(N - i)$ equals the birth rate b_i and the death rate d_i .*

Using the Taylor expansion of f around $\frac{i-1}{N}$ and $\frac{i+1}{N}$ up to second order, we obtain

$$(\hat{L}_N f) \left(\frac{i}{N}\right) = (N)^2 \frac{i}{N} \left(1 - \frac{i}{N}\right) \left[\frac{1}{(N)^2} f'' \left(\frac{i}{N}\right) + \hat{R}_N \right], \quad (2.33)$$

where $\hat{R}_N$ is an error term that vanishes as $N \rightarrow \infty$. Hence we have

$$\lim_{N \rightarrow \infty} (\hat{L}_N f) \left(\frac{i_N}{N}\right) = y(1 - y)f''(y) \quad \text{when} \quad \lim_{N \rightarrow \infty} \frac{i_N}{N} = y. \quad (2.34)$$

Thus, also the Moran-model converges to the WF-diffusion after space-time scaling, with the sole difference that the Moran-model runs at *twice* the speed of the WF-model. This is because in the WF-model *two* jumps give one chance of a coalescence, while in the Moran-model *one* jump does. (The intuitive reason why the WF-model and the Moran-model have the same scaling limit is that $\text{BIN}(N, 1/N)$ converges to $\text{POISSON}(1)$ as $N \rightarrow \infty$.)

The fact that the WF-model and the Moran-model have the same space-time scaling limit is an example of what is called *universality*. Biological systems that differ on a microscopic scale may have the same behaviour on a macroscopic scale.

2.2.3 WF-diffusion

The limiting process defined by (2.17) is called the *Wright-Fisher diffusion*. Think of this as a standard Brownian motion running at a “local speed” given by a *diffusion function* $g: [0, 1] \rightarrow [0, \infty)$, in our case $g(y) = y(1 - y)$ (see Fig. 6). Note that 0 and 1 are traps: the WF-diffusion stops when it reaches the boundary of $[0, 1]$ (a fact that is not entirely trivial to prove).

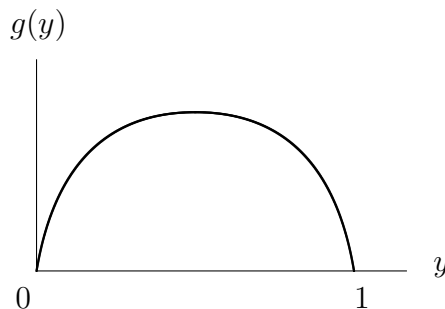


Figure 6: Wright-Fisher diffusion function.

Returning to the h -version of the Moran-model mentioned at the end of Section 2.1.2, we find that the space-time scaling in (2.14) leads to the following.

Theorem 2.4 *The h -version of the WF-diffusion satisfies the stochastic differential equation*

$$dY_t = \sqrt{Y_t(1 - Y_t)h(Y_t)} dW_t. \quad (2.35)$$

Exercise 2.5 *Give the proof of Theorem 2.4.*

It is expedient to define

$$g(y) = y(1 - y)h(y) \quad (2.36)$$

and to write (2.35) as

$$dY_t = \sqrt{g(Y_t)} dW_t. \quad (2.37)$$

Here, $g: [0, 1] \rightarrow [0, \infty)$ is the *local diffusion function*. In order for (2.37) to be properly defined and have a unique strong solution, some restrictions have to be placed on g . We will return to this in Chapter 7.

2.3 Dual to WF-diffusion

The WF-diffusion describes the WF-model on large space-time scales. Even though there is no easy explicit formula for Y_t in terms of $(W_s)_{0 \leq s \leq t}$, (2.17) has the advantage of being easier to manipulate in computations than the original Markov chain. To illustrate this advantage, we next turn to the notion of *duality*.

Theorem 2.1 *Let $D = (D_t)_{t \geq 0}$ be the death process on $\mathbb{N} = \{1, 2, \dots\}$ where transitions from n to $n - 1$ occur at rate $\binom{n}{2}$. Then*

$$\mathbb{E}([Y_t]^n \mid Y_0 = y) = \mathbb{E}(y^{D_t} \mid D_0 = n) \quad \forall y \in [0, 1], n \in \mathbb{N}, t \geq 0. \quad (2.38)$$

Proof. Abbreviate

$$\begin{aligned} a_t(y, n) &= \mathbb{E}([Y_t]^n \mid Y_0 = y), \\ b_t(y, n) &= \mathbb{E}(y^{D_t} \mid D_0 = n). \end{aligned} \quad (2.39)$$

Let $f(x) = x^n$. Then it follows from (2.19) that (see (A.6) in Appendix A.2.1)

$$\begin{aligned} a_t(y, n) - a_0(y, n) &= \mathbb{E}(f(Y_t) - f(Y_0) \mid Y_0 = y) \\ &= \mathbb{E} \left(\int_0^t (Lf)(Y_s) ds \mid Y_0 = y \right) \\ &= \mathbb{E} \left(\int_0^t Y_s(1 - Y_s) \frac{1}{2} f''(Y_s) ds \mid Y_0 = y \right) \\ &= \mathbb{E} \left(\int_0^t Y_s(1 - Y_s) \frac{1}{2} n(n-1) [Y_s]^{n-2} ds \mid Y_0 = y \right) \\ &= \binom{n}{2} \int_0^t [a_s(y, n-1) - a_s(y, n)] ds. \end{aligned} \quad (2.40)$$

Equation (2.40) is an integral recursion relation for $a_t(y, n)$, which in differential form reads

$$\frac{\partial}{\partial t} a_t(y, n) = \binom{n}{2} [a_t(y, n-1) - a_t(y, n)]. \quad (2.41)$$

However, it follows from the definition of D that $b_t(y, n)$ satisfies the same recursion:

$$\frac{\partial}{\partial t} b_t(y, n) = \binom{n}{2} [b_t(y, n-1) - b_t(y, n)]. \quad (2.42)$$

Exercise 2.2 *Prove (2.42). Hint: Use the backward Chapman-Kolmogorov equation for continuous-time Markov chains.*

Since

$$a_0(y, n) = y^n = b_0(y, n) \quad \forall y \in [0, 1], n \in \mathbb{N}, \quad (2.43)$$

it follows that

$$a_t(y, n) = b_t(y, n) \quad \forall y \in [0, 1], n \in \mathbb{N}, t \geq 0, \quad (2.44)$$

which proves the claim. ■

Remark: We may interpret the duality above as a description of the same event from two different perspectives: one looks forward in time, the other looks backward in time. More concretely,

$$\begin{aligned}
& \mathbb{E}([Y_t]^n \mid Y_0 = y) \\
&= \mathbb{P}(n \text{ individuals from an infinite population are of type } A \text{ at time } t) \\
&= \mathbb{P}(\text{the ancestors at time 0 of the } n \text{ individuals are of type } A) \\
&= \mathbb{E}(y^{D_t} \mid D_0 = n).
\end{aligned} \tag{2.45}$$

What Theorem 2.1 says is that the moments of Y_t can be computed from the distribution of D_t , and therefore so can the distribution of Y_t itself. Since the dual process D is *much simpler* than the WF-diffusion Y , this constitutes a considerable advantage. The following computation illustrates this advantage. We have

$$\mathbb{P}(Y_\infty = 1 \mid Y_0 = y) = \mathbb{E}(Y_\infty \mid Y_0 = y) = \mathbb{E}(y^{D_\infty} \mid D_0 = 1) = y, \tag{2.46}$$

where we use (2.38) and the fact that $Y_\infty \in \{0, 1\}$ and $D_\infty = 1$. This reflects the result in Lemma 2.1. Similarly,

$$\begin{aligned}
\mathbb{E}(Y_t(1 - Y_t) \mid Y_0 = y) &= \mathbb{E}(y^{D_t} \mid D_0 = 1) - \mathbb{E}(y^{D_t} \mid D_0 = 2) \\
&= y - [y \mathbb{P}(D_t = 1 \mid D_0 = 2) + y^2 \mathbb{P}(D_t = 2 \mid D_0 = 2)] \\
&= y(1 - y) \mathbb{P}(D_t = 2 \mid D_0 = 2) \\
&= y(1 - y)e^{-t},
\end{aligned} \tag{2.47}$$

where we use (2.38) together with the fact that $D_t \in \{1, 2\}$ when $D_0 = 2$ and $D_t = 1$ when $D_0 = 1$. This reflects the result in Lemma 2.2.

The dual process D can be traced back to the *backward* coalescing random walk encountered in the proof of Lemma 2.2. Indeed, Y describes the evolution of the type composition of a large population. Suppose that $Y_0 = y$ and we sample n individuals from this population at time t . The left-hand side of (2.38) is the probability to see only type A in the sample. But we can recompute this probability in a different manner. If we know that the n individuals in the sample are the descendants of D_t different ancestors at time 0, then by *averaging over the random genealogy* we obtain the right-hand side of (2.38) for the probability that all of the D_t ancestors are of type A .

More formally, consider the backward ancestral tree of a large population on the space-time scale given by (2.14). Then

$$D_t = \text{number of ancestral lineages at time } t_0 - t, \quad t \in [0, t_0], \tag{2.48}$$

where $t_0 \gg 1$ is any given observation time. In particular, $D_0 = N$ and $D_t = 1$ for $t \gg 1$.

Theorem 2.3 *The amount of time τ_k during which there are k lineages is equal to the amount of time during which the death process D equals k , and has distribution $\text{EXP}(\lambda_k)$ with $\lambda_k = \binom{k}{2}$, $k \in \mathbb{N} \setminus \{1\}$.*

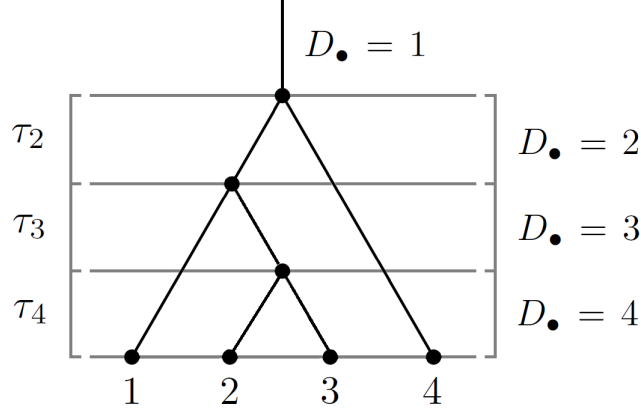


Figure 7: Part of a backward ancestral tree of a population in the WF-diffusion. (Each continuous unit of time corresponds to N discrete units of time.) In the death process time is running upwards, while biological time is running downwards. The expectations of the times τ_4 , τ_3 and τ_2 equal, respectively, $\frac{1}{6}$, $\frac{1}{3}$ and 1.

Proof. We first consider a sample of k individuals drawn from a WF-population of finite size $N \gg 1$. Afterwards we make the proper space-time rescaling and let the population size N tend to infinity.

The probability that two or more individuals from the sample have the same parent is

$$\lambda_k \frac{1}{N} + O(N^{-2}). \quad (2.49)$$

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

More precisely, the event that for k individuals in a population of size N two or more individuals have the same parent is the complement of the event that every individual has a different parent. This yields

$$\begin{aligned} & \mathbb{P}(\text{at least one parent in common}) \\ &= 1 - \mathbb{P}(\text{no parent in common}) \\ &= 1 - \prod_{j=1}^k \frac{N - (j-1)}{N} = 1 - \prod_{j=0}^{k-1} \left(1 - \frac{j}{N}\right) \\ &= 1 - \left(1 + \sum_{i=1}^{k-1} \sum_{0 < j_1, \dots, j_i < k} \prod_{l=1}^i \left(-\frac{j_l}{N}\right)\right) \\ &= 1 - \left(1 - \sum_{j=1}^{k-1} \frac{j_l}{N} + O(N^{-2})\right) = \binom{k}{2} N^{-1} + O(N^{-2}). \end{aligned} \quad (2.50)$$

When we follow the ancestral lineages of the k individuals backwards in time (see Fig. 7),

the probability that the k lineages do not coalesce during the first n generations equals

$$\left(1 - \lambda_k \frac{1}{N} + O(N^{-2})\right)^n = \exp\left[-\lambda_k \frac{n}{N} + O(nN^{-2})\right]. \quad (2.51)$$

Now rescale time by putting $t = n/N$. Then in the limit as $N \rightarrow \infty$ we have, for all $t \geq 0$,

$$\mathbb{P}(\tau_k > t) = e^{-\lambda_k t}. \quad (2.52)$$

Thus, when the population size tends to infinity, τ_k has distribution $\text{EXP}(\lambda_k)$, i.e., the exponential distribution with mean $1/\lambda_k$. ■

3 Wright-Fisher: genealogy

In this chapter we take a closer look at the WF-model and the WF-diffusion *backwards* in time. In Section 3.1 we make the duality encountered in Section 2.3 more transparent by introducing the Kingman coalescent, which is the *time-reversed* way of looking at the WF-diffusion and provides information on the *genealogy* of the WF-model. In Section 3.2 we look at the *most recent common ancestor* of the population and investigate how far back in time this ancestor lived. The presentation of the material in this chapter is taken from the bachelor thesis of Maurits Carsouw [4].

3.1 Kingman coalescent

Theorem 2.3 identifies the law of the random times of coalescence, but fails to capture the genealogy, i.e., the tree structure of the population. A full description of all the lineages of the n individuals in the sample is achieved with the help of a coalescent process called the *Kingman coalescent* (see Kingman [17, 18]). This is a random process taking values in the collection of partitions of $\{1, \dots, n\}$ such that, if the partition has j sets, then at rate $\frac{1}{2}j(j-1)$ two randomly chosen sets in the partition are joined together.

The Kingman coalescent, often simply called *the coalescent*, is a stochastic process that describes the family tree of a large WF-population ($N \gg 1$) backwards in time, up to the individual called *most recent common ancestor* (see Section 3.2). The coalescent is related to the WF-diffusion Y (see Section 2.2), and through duality to the death process D (see Section 2.3). It is a powerful tool in the study of the genealogy of the population up to the most recent common ancestor. Coalescent theory stands on itself as an important area in population genetics.

In this section, the coalescent is constructed from the continuous-time Markov process called *n-coalescent* (Section 3.1.1). The transition probabilities and the absolute probabilities of this process will be calculated using the death process D . Existence and uniqueness of the coalescent will be established by the Stone-Weierstrass theorem (Section 3.1.2). Finally, one of the coalescent's most striking features is shown: it “comes down from infinity” (Section 3.1.3).

3.1.1 The finite coalescent

For $n \in \mathbb{N}$, let E_n denote the set of *partitions* of the set $\{1, \dots, n\}$. The number of subsets in $R \in E_n$ is denoted by $|R|$.

The n -coalescent is the continuous-time Markov process $R^n = (R_t^n)_{t \geq 0}$ with state space E_n , initial state

$$R_0^n = \Delta = (\{1\}, \dots, \{n\}), \quad (3.1)$$

and transition rates

$$P_{RS} = \lim_{h \downarrow 0} h^{-1} \mathbb{P}(R_h^n = S \mid R_0^n = R) \quad (3.2)$$

given by

$$P_{RS} = \begin{cases} 1, & \text{if } S \triangleleft R, \\ 0, & \text{otherwise,} \end{cases} \quad R, S \in E_n, R \neq S, \quad (3.3)$$

where

$$S \triangleleft R \Leftrightarrow \{S \in E_n : |S| = |R| - 1, \forall x \in R \exists y \in S : x \subset y\}, \quad (3.4)$$

i.e., $S \triangleleft R$ means that the partition S is obtained from the partition R by merging together two of its subsets.

If we draw a sample of n individuals (from a large population) at time t_0 , then the n -coalescent gives us a description of *all the lineages* of the n individuals backwards in time, i.e., the full tree structure (= genealogy). To understand this, let R_t^n denote the partition such that i and j are in the same subset if and only if the individuals i and j have a common ancestor that is alive at time $t_0 - t$. With this definition, the process $R^n = (R_t^n)_{t \geq 0}$ indeed has the stochastic structure of an n -coalescent. Note that there is a one-to-one correspondence between the subsets in R_t^n and the common ancestors at $t_0 - t$ in the sample.

Fig. 8 shows an example of the backward ancestral tree of a sample of size $n = 4$, taken (from a large population) at time t_0 . One possible sequence of jumps of the 4-coalescent is:

$$\begin{aligned} R_0^4 &= (\{1\}, \{2\}, \{3\}, \{4\}) = \Delta, \\ R_t^4 &= (\{1\}, \{2\}, \{3, 4\}), \\ R_{t+s}^4 &= (\{1\}, \{2, 3, 4\}), \\ R_{t+s+v}^4 &= (\{1, 2, 3, 4\}) = \Theta. \end{aligned} \quad (3.5)$$

(Alternatively, as a first jump there could be a merging of $\{2\}$ and $\{4\}$, in which case R_t^4 would be $(\{1\}, \{2, 4\}, \{3\})$.)

Write $D^n = (D_t^n)_{t \geq 0}$ for the natural restriction of the death process $D = (D_t)_{t \geq 0}$ (defined in Section 2.3) to the set $\{1, \dots, n\}$, so that D^n is the death process on $\{1, \dots, n\}$ with initial value $D_0^n = n$ and transitions from k to $k - 1$ that occur at rate $\lambda_k = \binom{k}{2}$. If we define

$$D_t^n = |R_t^n|, \quad (3.6)$$

then we can show that $(D_t^n)_{t \geq 0}$ is a death process that evolves according to (2.48). It is clear that D_t^n equals the number of subsets in the partition R_t^n . Indeed, each merger of R_t^n corresponds to a reduction by one in the number of different lineages in the

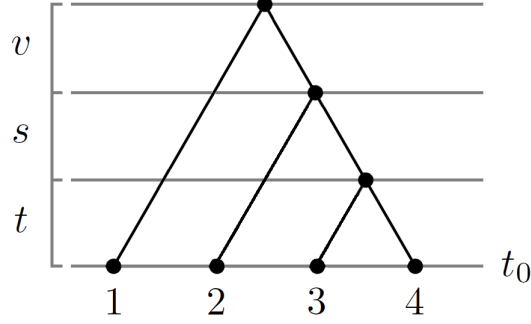


Figure 8: Example of the backwards ancestral tree.

genealogy of the sample. Since Theorem 2.3 gives us the transition rates of D (equal to $\lambda_k = \binom{k}{2}$, $k \in \mathbb{N} \setminus \{1\}$, for $(D_t^n)_{t \geq 0}$ to be a death process we need to see why these are the same as the rates at which one merger occurs in R_t^n . To that end it suffices to recall (3.3):

$$P_R = \sum_{S \in E_n} P_{RS} = \#\{S \mid S \triangleleft R\} = \binom{|R|}{2} = \lambda_{|R|}. \quad (3.7)$$

Thus, the n -coalescent jumps with transition rates λ_k through a sequence of partitions $\mathfrak{R}_k$ with $|\mathfrak{R}_k| = k$, $k = n, n-1, \dots, 1$, such that

$$\Theta = \mathfrak{R}_1 \triangleleft \mathfrak{R}_2 \triangleleft \dots \triangleleft \mathfrak{R}_n = \Delta. \quad (3.8)$$

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

$$(\mathfrak{R}_n, \mathfrak{R}_{n-1}, \dots, \mathfrak{R}_2, \mathfrak{R}_1), \quad (3.9)$$

which is directly related to its n -coalescent as

$$R_t^n = \mathfrak{R}_{D_t^n}. \quad (3.10)$$

From (3.3) and (3.7), we can now calculate the transition probabilities of the jump chain. For $R \in E_n$ with $|R| = k \in \{2, \dots, n\}$, we have

$$\mathbb{P}(\mathfrak{R}_{k-1} = S \mid \mathfrak{R}_k = R) = \frac{P_{RS}}{P_R} = \begin{cases} 1/\lambda_k, & \text{if } S \triangleleft R, \\ 0, & \text{otherwise.} \end{cases} \quad (3.11)$$

Indeed, exactly one of the $\lambda_k = \binom{k}{2}$ coalescence events turns R into S .

Finding the absolute (unconditioned) probabilities of the jump chain requires doing a little more work, which is the content of the next theorem.

Theorem 3.1 *For any $R \in E_n$ with $|R| = k$,*

$$\mathbb{P}(\mathfrak{R}_k = R) = \frac{k! (n-k)!(k-1)!}{n! (n-1)!} n_1! \times \dots \times n_k!, \quad (3.12)$$

where $n_1, \dots, n_k$ are the sizes of the k subsets in R , with $n_1 + \dots + n_k = n$.

Proof. The proof is based on Durrett [10, Section 1.2.2, Theorem 1.5]. We use induction on k , working backwards from $k = n$. For $k = n$ we have $\mathfrak{R}_k = \mathfrak{R}_n = \Delta$, so that all subsets are of size 1,

$$n_1 = \dots = n_k = 1. \quad (3.13)$$

Thus, we have

$$\frac{k!}{n!} \frac{(n-k)!(k-1)!}{(n-1)!} n_1! \times \dots \times n_k! = 1. \quad (3.14)$$

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

$$\mathbb{P}(\mathfrak{R}_n = R) = 1. \quad (3.15)$$

Now, let $k \in \{2, \dots, n\}$ be arbitrary, and assume that the theorem holds for k (induction hypothesis). We will prove that (3.12) also holds for $k-1$.

Let $S \in E_n$ with $|S| = k-1$ and $S \triangleleft R$. Then from (3.11) it follows that

$$\mathbb{P}(\mathfrak{R}_{k-1} = S) = \frac{2}{k(k-1)} \sum_{R: S \triangleleft R} \mathbb{P}(\mathfrak{R}_k = R). \quad (3.16)$$

Suppose that the subsets in S have sizes $n_1, \dots, n_{k-1}$. Then there exist $l \in \{1, \dots, k-1\}$ and $m \in \{1, \dots, n_l-1\}$ such that the subsets in R have sizes

$$n_1, \dots, n_{l-1}, m, n_l - m, n_{l+1}, \dots, n_{k-1}. \quad (3.17)$$

From the induction hypothesis it follows that the right-hand side of (3.16) equals

$$\begin{aligned} & \frac{2}{k(k-1)} \sum_{l=1}^{k-1} \sum_{m=1}^{n_l-1} \frac{k!}{n!} \frac{(n-k)!(k-1)!}{(n-1)!} \\ & n_1! \times \dots \times n_{l-1}! m! (n_l - m)! n_{l+1}! \times \dots \times n_{k-1}! \binom{n_l}{m} \frac{1}{2}. \end{aligned} \quad (3.18)$$

Here, $\binom{n_l}{m} \frac{1}{2}$ equals the number of ways to pick R with $S \triangleleft R$ so that the subsets in R with sizes m and $n_l - m$ coalesce to form the l -th subset in S with size n_l . (The factor $\frac{1}{2}$ appears because we are counting twice when we summing over m running from 1 to $n_l - 1$.) From the simple fact that

$$n_1! \times \dots \times n_{l-1}! m! (n_l - m)! n_{l+1}! \times \dots \times n_{k-1}! \binom{n_l}{m} = n_1! \times \dots \times n_{k-1}!, \quad (3.19)$$

we conclude that the right-hand side of (3.18) equals

$$\begin{aligned} & \frac{k!}{n!} \frac{(n-k)!(k-1)!}{(n-1)!} n_1! \times \dots \times n_{k-1}! \frac{1}{k(k-1)} \sum_{l=1}^{k-1} \sum_{m=1}^{n_l-1} 1 \\ & = \frac{k!}{n!} \frac{(n-k)!(k-1)!}{(n-1)!} \frac{(n-(k-1))}{k(k-1)} n_1! \times \dots \times n_{k-1}! \\ & = \frac{(k-1)!}{n!} \frac{(n-(k-1))!((k-1)-1)!}{(n-1)!} n_1! \times \dots \times n_{k-1}! \end{aligned} \quad (3.20)$$

and thus (3.12) indeed holds for $k-1$. ■

3.1.2 The infinite coalescent

We will now construct the coalescent from the n -coalescent (see also Kingman [17], Section 7). For $2 \leq m < n$, define the restriction $\rho_{nm}: E_n \rightarrow E_m$ by dropping all the labels $m+1, \dots, n$ from the subsets in the partitions making up E_n . If $R^n = (R_t^n)_{t \geq 0}$ is the n -coalescent, then $(\rho_{nm} R_t^n)_{t \geq 0}$ is the m -coalescent.

Let E denote the set of partitions of the set $\mathbb{N}$, and define the restriction $\rho_n: E \rightarrow E_n$ by dropping all the labels $> n$. We search for a process $R = (R_t)_{t \geq 0}$ on E such that $(\rho_n R_t)_{t \geq 0}$ is the n -coalescent for all $n \geq 2$. Such a process is called *a* coalescent. The following theorem justifies talking about *the* coalescent.

Theorem 3.2 *There exists a unique coalescent $R = (R_t)_{t \geq 0}$.*

Proof. Note that E is a subset of the powerset $2^{\mathbb{N}}$, and is a closed subset with respect to the (compact) product topology on $2^{\mathbb{N}}$. When E is equipped with the subspace topology, it becomes a compact Hausdorff space (a condition needed for the Stone-Weierstrass theorem below).

A coalescent exists when 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 , we consider the expectation

$$\mathbb{E}(f(R_{t_1}, R_{t_2}, \dots, R_{t_k})) \quad (3.21)$$

for bounded continuous functions $f: E^k \rightarrow \mathbb{R}$ and ordered times $0 \leq t_1 < t_2 < \dots < t_k$. The requirement that $(\rho_n R_t)_{t \geq 0}$ is the n -coalescent determines the value of (3.21) when there exists a function $g: (E_n)^k \rightarrow \mathbb{R}$ such that f is of the form

$$f(S_1, \dots, S_k) = g(\rho_n S_1, \dots, \rho_n S_k). \quad (3.22)$$

The Stone-Weierstrass theorem implies that the set of functions f of this form is dense in the set of bounded continuous functions $f: E^k \rightarrow \mathbb{R}$. Continuous extension then determines the value of (3.21) for *all* f , so that the consistent specification of the finite-dimensional distributions is established. Since this is done uniquely, it follows that any two coalescents have the same finite-dimensional distributions. ■

Exercise 3.3 *Check the details of the proof.*

With this definition, the coalescent $R = (R_t)_{t \geq 0}$ is the Markov process on E with initial condition $R_0 = (\{i\}: i \in \mathbb{N})$ such that each pair of subsets in R_t coalesces at rate 1 for all $t \geq 0$. This means that every two lineages in the coalescent coalesce at rate 1 as we go backwards in time.

Where the n -coalescent gives a description of the ancestral lineages of a sample of size n taken from a large population, the coalescent gives a description of the ancestral lineages of an *infinite* population, up to the single individual from which the population has descended. In the WF-diffusion the 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? The answer is provided in the next section: an *entrance law* can be used to describe how a Markov process “comes down from infinity”.

3.1.3 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 [2], Section 2.1.2).

As before, we let $D_t = |R_t|$ denote the number of subsets in 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 finitely many lineages. This is expressed by saying that the coalescent *comes down from infinity*.

Theorem 3.4 $\mathbb{P}(\forall t > 0: D_t < \infty) = 1$.

Proof. We must show that

$$\forall t > 0 \forall \epsilon > 0 \exists N_{t,\epsilon} \in \mathbb{N}: \mathbb{P}(D_t > N_{t,\epsilon}) < \epsilon. \quad (3.23)$$

Indeed, (3.23) implies that $\mathbb{P}(D_t < \infty) = 1$ for all $t \geq 0$, which settles the claim because $t \mapsto D_t$ is non-increasing.

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 subsets in R_t^n . Let τ_k be the random variable with distribution $\text{EXP}(\lambda_k)$, $k \in \mathbb{N} \setminus \{1\}$. Then Theorem 2.3 says that we can use τ_k to represent the amount of time during which $D_t = k$. Therefore, using the Markov inequality, we have

$$\begin{aligned} \mathbb{P}(D_t^n > N_{t,\epsilon}) &= \mathbb{P}\left(\sum_{k=N_{t,\epsilon}}^n \tau_k > t\right) \leq \frac{1}{t} \mathbb{E}\left(\sum_{k=N_{t,\epsilon}}^n \tau_k\right) \\ &\leq \frac{1}{t} \sum_{k=N_{t,\epsilon}}^{\infty} \mathbb{E}(\tau_k) = \frac{1}{t} \sum_{k=N_{t,\epsilon}}^{\infty} \frac{1}{\lambda_k}, \end{aligned} \quad (3.24)$$

where we note that the last sum is independent of n . Since $\sum_{k \in \mathbb{N}} 1/\lambda_k < \infty$, we can choose $N_{t,\epsilon}$ large enough to make sure that

$$\frac{1}{t} \left(\sum_{k=N_{t,\epsilon}}^{\infty} \frac{1}{\lambda_k} \right) < \epsilon, \quad (3.25)$$

from which it follows that

$$\mathbb{P}(D_t > N_{t,\epsilon}) = \lim_{n \rightarrow \infty} \sup_{m \geq n} \mathbb{P}(D_t^m > N_{t,\epsilon}) < \epsilon, \quad (3.26)$$

which is (3.23). ■

The Kingman coalescent describes the genealogy of the WF-population in the space-time scaling limit. It is a universal process, in the sense that other populations like the Moran-population have the same limit.

3.2 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 shrinks, 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 is a common ancestor of the whole population.

In this section 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 (Section 3.2.1). 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 in Appendix B by means of a “particle construction” introduced by Donnelly and Kurtz [9]. In particular, we will determine the distribution of the waiting time between the successive jumps in the MRCA-process (Section 3.2.2).

3.2.1 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 $t_0 \in \mathbb{R}$. Then this population has a MRCA a.s. (i.e., 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 3.1 *Let T be the time between the population in the WF-diffusion and its MRCA. Then $\mathbb{E}(T) = 2$.*

Proof. Let τ_k be the amount of time during which there are k lineages in the coalescent tree of the population, $k \in \mathbb{N} \setminus \{1\}$. Then Theorem 2.3 states that $\mathbb{E}(\tau_k) = 1/\lambda_k = 1/\binom{k}{2}$. Since the time between the population and its MRCA equals the depth of the coalescent tree, the desired expectation equals

$$\mathbb{E}(T) = \mathbb{E}\left(\sum_{k=2}^{\infty} \tau_k\right) = \sum_{k=2}^{\infty} \mathbb{E}(\tau_k) = \sum_{k=2}^{\infty} \frac{2}{k(k-1)} = 2. \quad (3.27)$$

■

Remark: Theorem 3.1 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 $N = 50,000$, then according to the space-time rescaling in (2.14), we expect that the MRCA lived $50,000 \times 2 = 100,000$ generations (discrete-time units) ago.

The average time to the most recent common ancestor is twice the population size.

3.2.2 MRCA-process and F-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 [9] 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 that 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. Each time F_t when 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 *F-process*.

Suppose that the points of $(A_t)_{t \in \mathbb{R}}$ are enumerated as $\{\alpha_i\}_{i \in \mathbb{Z}}$. Then the points $\{\beta_i\}_{i \in \mathbb{Z}}$ of $(F_t)_{t \in \mathbb{R}}$ are given by $\beta_i = \inf\{t \in \mathbb{R} : A_t = \alpha_i\}$ (see Fig. 9), and the path of $(A_t)_{t \in \mathbb{R}}$ is constant on each interval $[\beta_i, \beta_{i+1})$:

$$\forall t \in [\beta_i, \beta_{i+1}) : A_t = \alpha_i. \quad (3.28)$$

In this way, at each time β_i the MRCA, which lives at time α_i , is established.

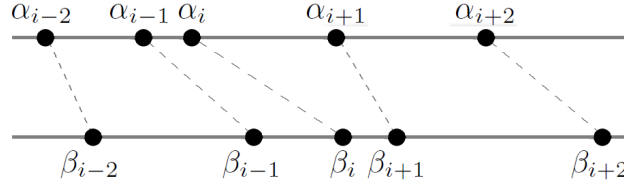


Figure 9: Time is running horizontally to the right. The time points α_j and β_j , $i - 2 \leq j \leq i + 2$, of the MRCA-process $(A_t)_{t \in \mathbb{R}}$ and the F-process $(F_t)_{t \in \mathbb{R}}$ are drawn for fixed $i \in \mathbb{Z}$. The dotted lines relate each time α_j at which a MRCA lives to the time β_j at which this MRCA is established.

In Appendix B we will prove the following two theorems (see Appendix A.2.3).

Theorem 3.2 *The MRCA-process $(A_t)_{t \in \mathbb{R}}$ is a rate-1 Poisson process on $\mathbb{R}$.*

Theorem 3.3 *The F-process $(F_t)_{t \in \mathbb{R}}$ is a rate-1 Poisson process on $\mathbb{R}$.*

In the WF-diffusion, successive MRCA's occur at Poisson times. The average time between two successive MRCA's is 1.

Exercise 3.4 *In Theorem 3.1 we saw that the average time between the population and its MRCA is 2. Is there a paradox because $1 \neq 2$?*

4 Wright-Fisher with mutation

In this chapter we add mutation. In Section 4.1 we look at the standard model with two types, in Sections 4.2–4.3 at the model with infinitely many types, respectively, with infinitely many sites (= base pairs).

4.1 WF-model with mutation and two types

4.1.1 Mutation

Suppose that we modify the WF-model in the following manner. At each time unit each individual, immediately after it has chosen its ancestor and adopted its type (resampling), suffers a *type mutation*: type a spontaneously mutates into type A with probability u , and type A spontaneously mutates into type a with probability v (see Fig. 10). Here, $0 < u, v < 1$, and mutations occur independently for different individuals.

An alternative way to include the mutation is to say that, after the resampling, each individual mutates into type A or a with probability u , respectively, v *irrespective of its type*. Since the transitions $A \rightarrow A$ and $a \rightarrow a$ do not effect the population, this type-independent mutation gives the same model, but is easier to work with (see below).

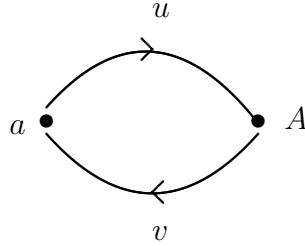


Figure 10: Two-type mutation.

Our goal is to investigate what effect the mutation has on the behavior of the model. With mutation $X = (X_n)_{n \in \mathbb{N}_0}$ is a Markov chain on the state space Ω with transition kernel

$$p(i, j) = \binom{N}{j} (p_i)^j (1 - p_i)^{N-j}, \quad i, j \in \Omega, \quad (4.1)$$

with

$$p_i = \left(\frac{i}{N} \right) (1 - v) + \left(\frac{N - i}{N} \right) u. \quad (4.2)$$

Indeed, either an A is drawn (probability $\frac{i}{N}$) and it does not mutate into an a (probability $1 - v$), or an a is drawn (probability $\frac{N-i}{N}$) and it does mutate into an A (probability u). Compare (4.1–4.2) with (2.3).

A first consequence of the presence of mutation is that the traps at 0 and N disappear: $p(i, j) > 0$ for all $i, j \in \Omega$.

Due to mutation there no longer is loss of genetic variability!

In fact, because N is finite, we have $w - \lim_{n \rightarrow \infty} X_n = X$ with

$$\pi(i) = \mathbb{P}(X = i), \quad i \in \Omega, \quad (4.3)$$

an equilibrium distribution solving the set of equations

$$\sum_{i \in \Omega} \pi(i) p(i, j) = \pi(j), \quad j \in \Omega, \quad (4.4)$$

normalized such that $\sum_{i \in \Omega} \pi(i) = 1$. The equilibrium has full support, i.e., $\pi(i) > 0$ for all $i \in \Omega$. In principle it is possible to compute π , but the formulas are not so pretty. We get some insight by computing the first two moments of X .

Lemma 4.1 *The first and second moment of X are given by*

$$\begin{aligned} \mathbb{E}\left(\frac{X}{N}\right) &= \rho, \\ \mathbb{E}\left(\frac{X(X-1)}{N(N-1)}\right) &= \chi\rho + (1-\chi)\rho^2, \end{aligned} \quad (4.5)$$

where

$$\rho = \frac{u}{u+v}, \quad \chi = \frac{(1-\mu)^2}{(1-\mu)^2 + N[1 - (1-\mu)^2]}, \quad \mu = 1 - (1-u)(1-v). \quad (4.6)$$

Proof. Suppose that the system is in equilibrium. Let $\eta_i = 1$ if the i -th individual is A and $\eta_i = 0$ otherwise. Then

$$X = \sum_{i=1}^N \eta_i, \quad (4.7)$$

and, by exchangeability,

$$\begin{aligned} \mathbb{E}(X) &= N\mathbb{P}(\eta_1 = 1), \\ \mathbb{E}(X(X-1)) &= N(N-1)\mathbb{P}(\eta_1 = \eta_2 = 1). \end{aligned} \quad (4.8)$$

To compute the first probability in (4.8), abbreviate $\rho = \mathbb{P}(\eta_1 = 1)$. In equilibrium we have

$$\rho = (1-v)\rho + u(1-\rho). \quad (4.9)$$

This gives $\rho = \frac{u}{u+v}$ and proves the first line of (4.5). To compute the second probability in (4.8), note that in equilibrium

$$\mathbb{P}(\eta_1 = \eta_2 = 1) = \chi\rho + (1-\chi)\rho^2, \quad (4.10)$$

where

$$\begin{aligned} \chi &= \text{the probability that two individuals are } \textit{identical by descent}, \\ &\text{i.e., their lineages coalesce before a mutation affects either lineage.} \end{aligned} \quad (4.11)$$

To compute χ we argue as follows. Let $\mu = 1 - (1 - u)(1 - v)$ be the probability of mutation per unit of time backwards on a lineage. Then the probability $1 - \chi$ that two individuals are not identical by descent satisfies the equation

$$1 - \chi = [1 - (1 - \mu)^2] + (1 - \mu)^2 \left(1 - \frac{1}{N}\right) (1 - \chi). \quad (4.12)$$

Indeed, if at the first time step backwards there is no mutation in either of the two lineages (probability $(1 - \mu)^2$) nor coalescence of the two lineages (probability $\frac{1}{N}$), then “the game starts all over again”. Solving (4.12) for χ and inserting into (4.8), we find the second line of (4.5). ■

An interesting consequence of Lemma 4.1 is that

$$\lim_{N \rightarrow \infty} \frac{X}{N} = \rho \quad \text{in probability.} \quad (4.13)$$

In large populations with mutation the fraction of types is close to a deterministic value.

In Section 4.1.2 we will see that the limit is random when u, v tend to zero as $N \rightarrow \infty$ in an appropriate manner.

Remark: In the presence of mutation, the coalescent (= the backward tree of lineages) modifies in the following manner. At each unit of backward time, a backward path: (i) is killed and labelled A with probability u ; (ii) is killed and labelled a with probability v ; (iii) jumps to a randomly chosen parent with probability $(1 - u)(1 - v)$; and (i)–(iii) occur independently). Killing a path determines the state of all its descendants (see Fig. 11). Note that if all the backward paths are killed, then the state of the system no longer depends on the initial configuration. This is why the system reaches equilibrium in the presence of mutation.

4.1.2 Weak mutation

Interesting behaviour shows up in the limit as $N \rightarrow \infty$, provided we take the mutation probabilities small:

$$u = u(N) = \frac{q}{2N}, \quad v = v(N) = \frac{r}{2N}, \quad q, r > 0. \quad (4.14)$$

The reason behind this choice is apparent from Lemma 4.1: for $u, v \downarrow 0$ we have $\mu \approx (u + v)$ and $\chi \approx 1/(1 + 2N(u + v))$, which shows that $2N(u + v)$ is the relevant quantity for the scaling behavior. Think of $\frac{1}{2}q, \frac{1}{2}r$ as the *mutation rates of the whole population*.

Remark: The reason why we divide by $2N$ rather than N in (4.14) is a matter of convention. In the Moran-model, which for large N runs at twice the speed of the WF-model (recall Section 2.2.2), we must divide by N to get the correct correspondence.

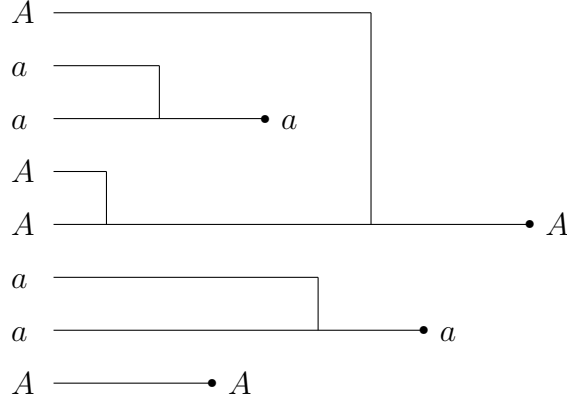


Figure 11: The coalescent with mutation in equilibrium. Time for the coalescent runs to the right (which means that biological time runs to the left). The black dots denote killing due to mutation, at which moment type A or type a is chosen with probability u , respectively, v and the rest of the coalescent is removed. After the A 's and the a 's have been assigned at the black dots, the states of the descendants are determined.

Define, in analogy with (2.14),

$$Y^{(N)} = \frac{1}{N} X^{(N)}. \quad (4.15)$$

Then we expect that

$$w - \lim_{N \rightarrow \infty} Y^{(N)} = Y. \quad (4.16)$$

Theorem 4.2 *The convergence in (4.16) is true, with Y the random variable on $[0, 1]$ with distribution $\text{BETA}(q, r)$, i.e., with probability density*

$$f(x) = C_{q,r} x^{q-1} (1-x)^{r-1}, \quad x \in [0, 1], \quad (4.17)$$

where $C_{q,r} = \Gamma(q+r)/\Gamma(q)\Gamma(r)$ is the normalizing constant (Γ is the Gamma-function).

Proof. The proof is by brute force: we show that the moments of $Y^{(N)}$ converge to the moments of the beta function Y as in (4.17), which implies that the family $Y^{(N)}$ converges to Y . We note that

$$\mathbb{E}(Y^k) = m_k, \quad m_k = \frac{\prod_{j=0}^{k-1} (q+j)}{\prod_{j=0}^{k-1} (q+r+j)}, \quad k \in \mathbb{N}, \quad (4.18)$$

which are the moments of $\text{BETA}(q, r)$. The values of m_k satisfy the following recursive relation:

$$m_k = \frac{q+k-1}{q+r+k-1} m_{k-1}, \quad k \in \mathbb{N}, \quad m_0 = 1. \quad (4.19)$$

Using this relation and inductive reasoning, we will prove convergence of the moments.

Given that $X_{n-1}^{(N)} = i$, the distribution of $X_n^{(N)}$ is $\text{BIN}(N, p_i)$. Since the k -th factorial moment of a binomial distribution $\text{BIN}(N, p)$ equals $p^k \prod_{j=0}^{k-1} (N - j)$, we

$$\mathbb{E} \left(\prod_{j=0}^{k-1} (X_n^{(N)} - j) \mid X_{n-1}^{(N)} = i \right) = (p_i)^k \left(\prod_{j=0}^{k-1} (N - j) \right). \quad (4.20)$$

Dividing by $(N)^k$, expanding each side of (4.20) to leading order in $\frac{1}{N}$, and putting $Y_n^{(N)} = \frac{1}{N} X_n^{(N)}$, we obtain

$$\begin{aligned} & \mathbb{E} \left([Y_n^{(N)}]^k \mid Y_{n-1}^{(N)} = \frac{i}{N} \right) - \frac{k(k-1)}{2N} \mathbb{E} \left([Y_n^{(N)}]^{k-1} \mid Y_{n-1}^{(N)} = \frac{i}{N} \right) + O \left(\frac{1}{N^2} \right) \\ &= (p_i)^k \left[1 - \frac{k(k-1)}{2N} + O \left(\frac{1}{N^2} \right) \right], \end{aligned} \quad (4.21)$$

where we use that $\sum_{j=0}^{k-1} j = \frac{1}{2}k(k-1)$. Next, we expand

$$\begin{aligned} (p_i)^k &= \left(\frac{i}{N} (1 - u - v) + u \right)^k \\ &= \left(\frac{i}{N} \right)^k + k \left(\frac{i}{N} \right)^{k-1} \left[\frac{q}{2N} - \left(\frac{i}{N} \right) \frac{q+r}{2N} + O \left(\frac{1}{N^2} \right) \right], \end{aligned} \quad (4.22)$$

where we use (4.2) and (4.14). Combining (4.21–4.22), we get

$$\begin{aligned} & \mathbb{E} \left([Y_n^{(N)}]^k \mid Y_{n-1}^{(N)} \right) - \frac{k(k-1)}{2N} \mathbb{E} \left([Y_n^{(N)}]^{k-1} \mid Y_{n-1}^{(N)} \right) + O \left(\frac{1}{N^2} \right) \\ &= \left[1 - \frac{k(k-1)}{2N} + O \left(\frac{1}{N^2} \right) \right] \\ & \quad \times \left\{ \left[Y_{n-1}^{(N)} \right]^k + k \left[Y_{n-1}^{(N)} \right]^{k-1} \left[\frac{q}{2N} - \left[Y_{n-1}^{(N)} \right] \frac{q+r}{2N} + O \left(\frac{1}{N^2} \right) \right] \right\}. \end{aligned} \quad (4.23)$$

Taking the expectation over $Y_{n-1}^{(N)}$, letting $n \rightarrow \infty$, using that $w - \lim_{n \rightarrow \infty} Y_n^{(N)} = Y^{(N)}$, and carefully isolating the distinct moments and the error terms $O \left(\frac{1}{N^2} \right)$, we obtain the relation

$$\frac{k(k+q+r-1)}{2N} \mathbb{E}([Y^{(N)}]^k) = \frac{k(k+q-1)}{2N} \mathbb{E}([Y^{(N)}]^{k-1}) + O \left(\frac{1}{N^2} \right). \quad (4.24)$$

Finally, multiplying by $2N/k$, letting $N \rightarrow \infty$ and using the induction hypothesis, we conclude that the moments of $Y^{(N)}$ satisfy (4.19) and therefore $\lim_{N \rightarrow \infty} Y^{(N)} = Y$. ■

The qualitative behaviour of the probability density in (4.17) near the boundaries 0 and 1 is different for $q, r < 1$ and $q, r > 1$ (see Fig. 12).

The shape of the limiting distribution for the fractions of the types in a large population provides information on the mutation rates.

In the limit as $q, r \downarrow 0$, f concentrates on the boundaries, in accordance with the eventual loss of genetic variability in the WF-model without mutation (or rather in the WF-diffusion, because we took the limit $N \rightarrow \infty$). In the limit as $q, r \rightarrow \infty$, f concentrates around the value $\frac{q}{q+r}$, in accordance with what we found in (4.13).

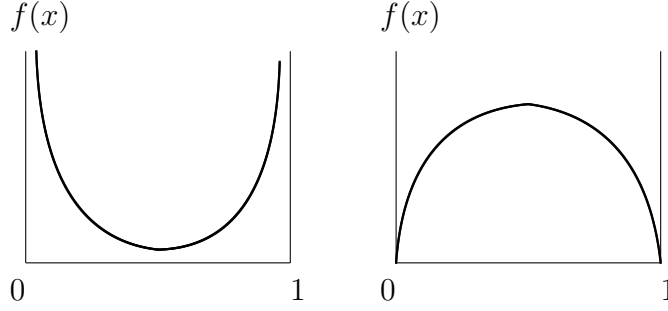


Figure 12: Qualitative picture of the probability density in (4.17) for $q, r < 1$ and $q, r > 1$, respectively.

4.2 WF-model with mutation and infinitely many types

In this section we consider the modification of the WF-model in which, instead of two types, there are *infinitely many types* and *each time a mutation occurs it brings in a new type*. The motivation for this model is the following. If a gene consists of, say, 500 nucleotides, then there are $3 \times 500 = 1500$ sequences that can be reached by a single base pair change. Therefore the probability of returning to the same sequence after two mutations is $1/1500$, which is small. Neglecting this return amounts to considering what is called the *infinite alleles model*.

In Section 4.2.1 we define the model, show that it is dual to a birth process called the Hoppe urn model, and use the latter to study the distribution of the number of different types in an n -sample of individuals drawn randomly from the population. In Section 4.2.2 we use the Hoppe urn model to derive the Ewen sampling formula, which describes the sizes of the families of different types in the n -sample. In Section 4.2.3 we look at the sizes of these families in the limit as $n \rightarrow \infty$.

4.2.1 Infinite alleles model

Types are labelled $0, 1, 2, \dots$. We consider a population with N individuals, *all* starting with type 0. At each unit of time, each individual with probability $1 - \mu$ chooses a random ancestor and adopts its type and with probability μ spontaneously mutates into a *new type*. All individuals update independently from each other and independently of how they updated at previous times. The first mutation in the population brings in an individual of type 1, the second an individual of type 2, etc. As time proceeds, *new types enter the population and old types die out*. However, we may expect that after a long time the distribution of the *number* of different types in the population settles down to a limiting distribution. The question we want to address is: “What is this distribution?”.

Remark: The infinite alleles model was proposed by Kimura in 1968. At that time it was not yet possible to sequence the genome, so the precise order of the nucleotides in an allele, which determines its type, could not yet be determined. However, with the help of e.g. electrophoresis of enzymes it was possible to *distinguish* between different types, and so the number of different types could be measured.

We will consider this question in the limit as $N \rightarrow \infty$, with

$$\mu = \mu(N) = \frac{\theta}{2N}, \quad \theta > 0, \quad (4.25)$$

where $\frac{1}{2}\theta$ may be thought of as the *mutation rate for the whole population*. In this limit, after time is scaled by a factor N as well (recall (2.14)), we have that

$$k \text{ lineages coalesce at rate } \lambda_k = \binom{k}{2} \text{ and mutate at rate } \frac{1}{2}\theta k. \quad (4.26)$$

(Use that $\text{BIN}(N, \theta/2N)$ converges to $\text{POISSON}(\frac{1}{2}\theta)$ as $N \rightarrow \infty$.)

Draw a random sample of n individuals from the population in equilibrium ($n \ll N \rightarrow \infty$) and ask for

$$K_n = \text{the number of different types in the } n \text{ sample.} \quad (4.27)$$

Lemma 4.1 *As $n \rightarrow \infty$,*

$$\begin{aligned} \mathbb{E}(K_n) &\sim \theta \log n, \\ \text{VAR}(K_n) &\sim \theta \log n. \end{aligned} \quad (4.28)$$

Moreover, the central limit holds, i.e., $w\text{-}\lim_{n \rightarrow \infty} [K_n - \mathbb{E}(K_n)] / \sqrt{\text{VAR}(K_n)} = N(0, 1)$, the standard normal distribution.

Proof. The proof uses duality. The dual process is called the *Hoppe urn model*, which is defined as follows:

- An urn contains 1 black ball (of mass θ) and any number of colored balls (of mass 1 each). A ball is selected with a probability *proportional to its mass*. If a *colored* ball is drawn, then an extra ball with the *same color* is put into the urn. If the *black* ball is drawn, then an extra ball with a *new color* is put into the urn. The ball that was drawn is put back into the urn also.

We start with 1 black ball at time 0. At time n the urn contains $n+1$ balls, 1 black and n colored. The state of the urn is the number of different colors and their multiplicity. The key observation is the following duality :

- The genealogy of n individuals in the infinite alleles model can be simulated by running the Hoppe urn model for n time steps.

Indeed, the genealogy of the infinite alleles model is described by a *coalescent with killing* similar in spirit as the one in Fig. 11 (recall (4.26)): (i) lineages change randomly and coalesce at rate 1; (ii) on each lineage there is an independent Poisson process with rate $\frac{1}{2}\theta$ of mutations that kill the lineage and determine the type of all the descendants. See Durrett [10, Figure 1.7]. A draw of the black ball signals a mutation in the backwards genealogy.

Exercise 4.2 Check the above duality. Hint: the probability to mutate in the infinite alleles model when there are k lineages equals $k\frac{1}{2}\theta/[k\frac{1}{2}\theta + \frac{1}{2}k(k-1)] = \theta/[\theta + k - 1]$, which is precisely the probability in the Hoppe urn model to bring in a new color at the k -th draw.

What is nice about the Hoppe urn model is that it is a simple process, controlled by the parameter θ , and that it simulates all sample sizes *at once*: the sample size n in the infinite alleles model is the number of draws in the Hoppe urn model.

With the above duality the proof of (4.28) is easy. Write

$$K_n = \sum_{i=1}^n \eta_i \quad (4.29)$$

with $\eta_i = 1$ if the i -th ball added in the Hoppe urn model has a new color and zero otherwise. Clearly, the successive η_i are independent (because at time i the urn always contains $i + 1$ balls, one with mass θ and i with mass 1), and

$$\mathbb{P}(\eta_i = 1) = \frac{\theta}{\theta + i - 1}. \quad (4.30)$$

Consequently,

$$\begin{aligned} \mathbb{E}(K_n) &= \sum_{i=1}^n \mathbb{P}(\eta_i = 1) = \sum_{i=1}^n \frac{\theta}{\theta + i - 1} \\ &\sim \theta \int_{\theta}^{\theta+n} \frac{dx}{x} = \theta[\log(\theta + n) - \log \theta] \sim \theta \log n. \end{aligned} \quad (4.31)$$

Similarly,

$$\text{VAR}(K_n) = \sum_{i=1}^n \text{VAR}(\eta_i) = \sum_{i=1}^n \frac{\theta}{\theta + i - 1} \left(1 - \frac{\theta}{\theta + i - 1}\right) \sim \mathbb{E}(K_n). \quad (4.32)$$

This proves (4.28). The central limit theorem follows from standard arguments. ■

Exercise 4.3 What are the conditions in the central limit theorem for sums of independent random variables (not necessarily identically distributed)? Do these conditions apply to (4.29)? Hint: Consult a standard probability textbook.

Lemma 4.1 says that the number of different types in an n sample is rather small, namely, of order $\log n$. Apparently, *new types rapidly eradicate old types*, so that it is hard for different types to coexist in large numbers.

The number of different types is proportional to θ . Hence θ can be estimated from genetic sample data, which is interesting because typically θ is not known.

Lemma 4.1 shows that $\hat{\theta}_n = K_n/\log n$ is an asymptotically sharp estimator of θ (“a sufficient statistic”). Unfortunately, the standard deviation of $\hat{\theta}_n$ decays like $1/\sqrt{\log n}$, which is too slow to get sharp estimates unless n is very large. Still, $\hat{\theta}$ can be used to estimate θ with some confidence interval.

4.2.2 Ewens sampling formula

The next result, which is referred to as *Ewens sampling formula*, describes the full family size distribution of the types in the n -sample. Let

$$\begin{aligned} A_i &= \text{the number of types that are present precisely } i \text{ times in the } n\text{-sample,} \\ i &= 1, \dots, n, \end{aligned} \quad (4.33)$$

and note that $\sum_{i=1}^n iA_i = n$. This is the “type frequency spectrum” of the n -sample. E.g. $(n, 0, \dots, 0)$ corresponds to n types each being present once, while $(0, \dots, 0, 1)$ corresponds to one type being present n times. In particular, $K_n = \sum_{i=1}^n A_i$, which establishes the link with Section 4.2.1.

Write $A = (A_i)_{i=1}^n$, let

$$\mathcal{A}_n = \left\{ a = (a_i)_{i=1}^n : \sum_{i=1}^n ia_i = n \right\}, \quad n \in \mathbb{N}, \quad (4.34)$$

and abbreviate $P_n(a) = \mathbb{P}(A = a)$, $a \in \mathcal{A}_n$.

Theorem 4.4 *For all $n \in \mathbb{N}$ and $a \in \mathcal{A}_n$,*

$$P_n(a) = n! \prod_{i=1}^n \frac{(\theta/i)^{a_i}}{(\theta + i - 1)a_i!}. \quad (4.35)$$

Proof. We will again make use of the duality with the Hoppe urn model. The proof is by induction on n . If $n = 1$, then $a = (a_1) = (1)$, and the claim in (4.35) is true with $P_1(1) = 1$.

Pick $n \geq 2$. For $a' \in \mathcal{A}_{n-1}$ and $a \in \mathcal{A}_n$, let $p(a', a)$ denote the transition probability from state a' at time $n - 1$ to state a at time n in the Hoppe urn model. Let $\bar{P}_n(a)$ denote the right-hand side of (4.35). We will show that

$$\bar{P}_n(a) = \sum_{a' \in \mathcal{A}_{n-1}} \bar{P}_{n-1}(a') p(a', a) \quad \forall a \in \mathcal{A}_n, \quad (4.36)$$

which will imply that $\bar{P}_n$ is the distribution of the Hoppe urn model at time n , i.e., $P_n = \bar{P}_n$, and will complete the proof by induction.

Suppose that the state at time $n - 1$ is a' and the state at time n is a . Then there are two possibilities:

- (1) $a_1 = a'_1 + 1$, $a_i = a'_i$ for $i > 1$, i.e., a new color is added at time n . For this case

$$\frac{\bar{P}_n(a)}{\bar{P}_{n-1}(a')} = \frac{n}{\theta + n - 1} \frac{\theta}{a'_1}, \quad p(a', a) = \frac{\theta}{\theta + n - 1}. \quad (4.37)$$

- (2) $a_j = a'_j - 1$, $a_{j+1} = a'_{j+1} + 1$, $a_i = a'_i$ for $1 \leq j \leq n - 1$ and $i \neq j, j + 1$, i.e., a color is added at time n that is present j times at time $n - 1$. For this case

$$\frac{\bar{P}_n(a)}{\bar{P}_{n-1}(a')} = \frac{n}{\theta + n - 1} \frac{ja'_j}{(j + 1)a_{j+1}}, \quad p(a', a) = \frac{ja'_j}{\theta + n - 1}. \quad (4.38)$$

No other transitions are possible. From (4.37–4.38) we get

$$\begin{aligned}
& \sum_{a' \in \mathcal{A}_{n-1}} \frac{\bar{P}_{n-1}(a')}{\bar{P}_n(a)} p(a', a) \\
&= \frac{\theta + n - 1}{n} \frac{a_1}{\theta} \frac{\theta}{\theta + n - 1} + \sum_{j=1}^{n-1} \frac{\theta + n - 1}{n} \frac{(j+1)a_{j+1}}{ja'_j} \frac{ja'_j}{\theta + n - 1} \\
&= \frac{a_1}{n} + \sum_{j=1}^{n-1} \frac{(j+1)a_{j+1}}{n} \\
&= \frac{1}{n} \sum_{j=0}^{n-1} (j+1)a_{j+1} = 1,
\end{aligned} \tag{4.39}$$

which shows that indeed (4.36) is true. ■

Remark: A more intuitive proof can be found in Griffiths and Lessard [13].

The Ewens sampling formula not only provides insight into the family size distribution of the types in an n -sample, it also allows for a more sophisticated way of statistically estimating θ .

By rewriting (4.35) as

$$P_n(a) = \frac{1}{N_{\theta,n}} \prod_{i=1}^n e^{-\theta/i} \frac{(\theta/i)^{a_i}}{a_i!}, \quad a \in \mathcal{A}_n, \tag{4.40}$$

with $N_{\theta,n}$ the normalizing constant, we see that P_n can be interpreted as the distribution of a random vector $(A_1, \dots, A_n)$ whose components A_i are independent and Poisson distributed with mean θ/i *conditioned* on $\sum_{i=1}^n iA_i = n$.

Small frequencies are more likely than large frequencies. The reason is that if a type does not appear in the n -sample so often, then it is less likely to reduce in number by mutation.

4.2.3 Griffiths-Engen-McCloskey distribution

The next result describes the behavior of the Ewens sampling formula in the limit as $n \rightarrow \infty$. Let us call the descendants of the k -th new color in the Hoppe urn model the k -th *family*. This is the collection of all the balls in the urn that carry color k . Let

$$S_k(n) = \frac{1}{n} \times \text{the size of the } k\text{-th family at time } n. \tag{4.41}$$

Consider the random vector

$$S(n) = (S_k(n) : k \in \mathbb{N}). \tag{4.42}$$

Note that the components of this vector sum up to 1 (and are 0 in the tail).

Theorem 4.5 $w - \lim_{n \rightarrow \infty} S(n) = B$ with $B = (B_k: k \in \mathbb{N})$ the random vector given by

$$B_k = \left[\prod_{j=1}^{k-1} (1 - Z_j) \right] Z_k, \quad k \in \mathbb{N}, \quad (4.43)$$

where $(Z_k)_{k \in \mathbb{N}}$ are i.i.d. with distribution $\text{BETA}(1, \theta)$, i.e., with probability density function $f(z) = \theta(1 - z)^{\theta-1}$, $z \in [0, 1]$.

Proof. We make use of the following observation, relating the Hoppe urn model to a branching process with immigration:

- Immigrants enter the population at the times of a Poisson process with rate θ . Each individual performs a binary branching process, i.e., splits into two at rate 1 (= gives birth to a new individual at rate 1 and does not die).

The key observation is:

- Start with 1 immigrant entering at time 0. If each immigrant is a *new* type and offspring are the *same* type as their parents, then the successive states of the branching process with immigration have the same distribution as the successive states of the Hoppe urn model.

What is done here is that the Hoppe urn model (discrete in time) is *embedded* into the branching process with immigration (continuous in time). (In the Hoppe urn model, n is the number of colored balls in the urn and $S_k(n)$ is the fraction of colored balls with color k .)

Exercise 4.6 *Prove the above embedding.*

Armed with these observations we argue as follows. Let $X(t)$ be the number of individuals at time t in the branching process *without* immigration. Then it is well known that $\mathbb{E}(X(t)) = e^t$, $t \geq 0$, and

$$w - \lim_{t \rightarrow \infty} e^{-t} X(t) = \mathcal{E} = \text{EXP}(1). \quad (4.44)$$

Exercise 4.7 *Look up the proof of (4.44) (see Durrett [10, Section 1.3.3]).*

Let $X_k(t)$ be the number of individuals in the k -th family at time t in the branching process *with* immigration. Then (4.44) implies that

$$w - \lim_{t \rightarrow \infty} (e^{-t} X_k(t): k \in \mathbb{N}) = (e^{-T_k} \mathcal{E}_k: k \in \mathbb{N}), \quad (4.45)$$

where $\mathcal{E}_k$, $k \in \mathbb{N}$, are i.i.d. copies of $\mathcal{E}$ and T_k , $k \in \mathbb{N}$ are the successive arrival times of an independent Poisson process with rate θ (= the immigration times). Here we use that the branching process *with* immigration is the independent sum of the branching

processes *without* immigration starting at the successive arrival times of the immigrants. Let

$$I(t) = \sum_{k \in \mathbb{N}} X_k(t) \quad (4.46)$$

be the total number of individuals at time t , and let

$$\sigma = \sum_{k \in \mathbb{N}} e^{-T_k} \mathcal{E}_k. \quad (4.47)$$

Then (4.45) gives

$$w - \lim_{t \rightarrow \infty} e^{-t} I(t) = \sigma. \quad (4.48)$$

It follows from (4.45–4.48) that the limit B_k we are after in (4.43) arises as

$$B_k = w - \lim_{t \rightarrow \infty} \frac{X_k(t)}{I(t)} = \frac{e^{-T_k} \mathcal{E}_k}{\sigma}, \quad k \in \mathbb{N}, \quad (4.49)$$

where we use the *embedding* of the Hoppe urn model into the branching process with immigration. To identify the distribution of $(B_k: k \in \mathbb{N})$, we define $(T_0 = 0)$

$$\begin{aligned} \sigma_j &= \sum_{k=j}^{\infty} e^{-(T_k - T_{j-1})} \mathcal{E}_k, \quad j \in \mathbb{N}, \\ Z_j &= \frac{\mathcal{E}_j}{\mathcal{E}_j + \sigma_{j+1}}, \quad j \in \mathbb{N}, \end{aligned} \quad (4.50)$$

and note that $\sigma_1 = \sigma$. For $k = 1$, write

$$B_1 = \frac{e^{-T_1} \mathcal{E}_1}{\sigma_1} = \frac{e^{-T_1} \mathcal{E}_1}{e^{-T_1} (\mathcal{E}_1 + \sigma_2)} = \frac{\mathcal{E}_1}{\mathcal{E}_1 + \sigma_2} = Z_1. \quad (4.51)$$

For $k = 2$, write

$$\begin{aligned} B_2 &= \frac{e^{-T_2} \mathcal{E}_2}{\sigma_1} = \frac{e^{-T_2} \mathcal{E}_2}{e^{-T_1} (\mathcal{E}_1 + \sigma_2)} = \frac{\sigma_2}{\mathcal{E}_1 + \sigma_2} \frac{e^{-(T_2 - T_1)} \mathcal{E}_2}{\sigma_2} \\ &= \frac{\sigma_2}{\mathcal{E}_1 + \sigma_2} \frac{e^{-(T_2 - T_1)} \mathcal{E}_2}{e^{-(T_2 - T_1)} (\mathcal{E}_2 + \sigma_3)} = \frac{\sigma_2}{\mathcal{E}_1 + \sigma_2} \frac{\mathcal{E}_2}{\mathcal{E}_2 + \sigma_3} = (1 - Z_1) Z_2. \end{aligned} \quad (4.52)$$

Exercise 4.8 Give the proof of (4.43) by induction on k . Use (4.51–4.52).

It remains to show that $(Z_j)_{j \in \mathbb{N}}$ has the desired distribution. Recall the notation in Appendix A.2.1 and abbreviate $\text{GAMMA}(\theta, 1) = \text{GAMMA}(\theta)$, $\theta \in (0, \infty)$.

Lemma 4.9 σ has distribution $\text{GAMMA}(\theta)$.

Proof. Since $(\mathcal{E}_j, T_j - T_{j-1})$, $j \in \mathbb{N}$, are i.i.d. with distribution $\text{EXP}(1) \times \text{EXP}(\theta)$, we have that $\sum_{j \in \mathbb{N}} \delta_{\mathcal{E}_j, T_j}$ is the Poisson point process on $[0, \infty) \times [0, \infty)$ with intensity measure $\theta dx \otimes e^{-y} dy$. Consequently, $\sum_{j \in \mathbb{N}} \delta_{e^{-T_j} \mathcal{E}_j}$ is the Poisson point process on $[0, \infty)$ with

intensity measure $e^{-z\frac{\theta}{z}}dz$. Indeed, for any test function $f: (0, \infty) \rightarrow [0, \infty)$ that is continuous and has compact support, we have

$$\begin{aligned} \int_0^\infty dx \int_0^\infty dy \theta e^{-y} f(e^{-x}y) &= \int_0^\infty dy e^{-y} \int_0^y dz \frac{\theta}{z} f(z) \\ &= \int_0^\infty dz f(z) \frac{\theta}{z} \int_z^\infty dy e^{-y} = \int_0^\infty dz f(z) e^{-z} \frac{\theta}{z}. \end{aligned} \quad (4.53)$$

Since $e^{-z\frac{\theta}{z}}dz$ is the Lévy measure of the Gamma-process (see Bertoin [3]), the claim follows. ■

Lemma 4.10 (Z_j) , $j \in \mathbb{N}$, are i.i.d. with distribution $\text{BETA}(1, \theta)$.

Proof. We will need the following property of the Gamma-distribution ($\mathcal{L}$ denotes law):

$$\begin{aligned} \mathcal{L}[(G_1, G_2)] &= \text{GAMMA}(\theta_1) \otimes \text{GAMMA}(\theta_2) \\ \implies \mathcal{L}\left[\left(G_1 + G_2, \frac{G_1}{G_1 + G_2}\right)\right] &= \text{GAMMA}(\theta_1 + \theta_2) \otimes \text{BETA}(\theta_1, \theta_2). \end{aligned} \quad (4.54)$$

Exercise 4.11 Write out the proof.

Abbreviate $W_j = \mathcal{E}_j + \sigma_{j+1}$. We use (4.54) to show that

$$\mathcal{L}[(W_1, Z_1, \dots, Z_n)] = \text{GAMMA}(1 + \theta) \otimes \text{BETA}(1, \theta)^{\otimes n}, \quad n \in \mathbb{N}, \quad (4.55)$$

which will prove the claim.

First, we check that (4.55) is true for $n = 1$. Indeed,

$$W_1 = \mathcal{E}_1 + \sigma_2, \quad Z_1 = \frac{\mathcal{E}_1}{\mathcal{E}_1 + \sigma_2}. \quad (4.56)$$

By Lemma 4.9, $\mathcal{L}[\sigma_2] = \mathcal{L}[\sigma] = \text{GAMMA}(\theta)$, because the distribution of $(\sigma_j)_{j \in \mathbb{N}}$ is shift-invariant. Since $\mathcal{L}[\mathcal{E}_1] = \text{EXP}(1) = \text{GAMMA}(1)$ and $\mathcal{E}_1$ is independent of σ_2 (recall (4.50)), (4.54) and (4.56) yield

$$\mathcal{L}[(W_1, Z_1)] = \text{GAMMA}(1 + \theta) \otimes \text{BETA}(1, \theta). \quad (4.57)$$

Next, we suppose that (4.55) is true for n . Then, since the distribution of $(\mathcal{E}_j, T_j - T_{j-1})_{j \in \mathbb{N}}$ is shift-invariant, we have (recall (4.50))

$$\mathcal{L}[(W_2, Z_2, \dots, Z_{n+1})] = \text{GAMMA}(1 + \theta) \otimes \text{BETA}(1, \theta)^{\otimes n}. \quad (4.58)$$

But $(W_2, Z_2, \dots, Z_{n+1})$ is independent of $(\mathcal{E}_1, T_2 - T_1)$ and, since $W_1 = \mathcal{E}_1 + e^{-(T_2 - T_1)}W_2$, also independent of (W_1, Z_1) conditional on W_2 . Therefore, combining (4.57–4.58), we conclude that (4.55) is true for $n + 1$ aslo. Hence, by induction on n , it follows that (4.55) holds for all $n \in \mathbb{N}$. ■

Exercise 4.8 and Lemma 4.10 complete the proof of Theorem 4.5. ■

The limiting distribution in Theorem 4.5 is called the Griffiths, Engen, McCloskey distribution (GEM).

The GEM distribution describes the frequency spectrum of the types in a very large sample in the *age order*, i.e., the types are ordered according to their first appearance. It has a “random stick breaking” structure.

4.3 WF-model with mutation and infinitely many sites

In the *infinite alleles model* in Section 3.2 we only kept track of when sequences became different through mutation and referred to these differences as a change of type. We did not keep track of the *number* of differences between the sequences. In the present section we will introduce a model, called the *infinite sites model*, that keeps track of where in the sequence the mutations occur and we will count these mutations.

Given are two genes of length L . Think of these as a sample of size $n = 2$ drawn from a population with N genes (= individuals), each having a type that is given by its sequence of nucleotides of length L . Let Δ_2 be the number of *pairwise differences* in the two sequences. An example with $L = 14$ is

AATCGCTTGATACC
ACTCGCCTGATAAC

which has pairwise differences at positions 2, 7 and 13, so that $\Delta_2 = 3$. Let $\mu = \mu(N) = \theta/4N$ be the mutation rate as in (4.25). We are interested in the distribution of Δ_2 in the limit as $N \rightarrow \infty$ and $L \rightarrow \infty$. The latter limit is the reason why the model is called the *infinite sites model*.

Lemma 4.1 Δ_2 has distribution $\text{GEO}(1/(\theta + 1))$, i.e.,

$$\mathbb{P}(\Delta_2 = k) = \frac{1}{\theta + 1} \left(\frac{\theta}{\theta + 1} \right)^k, \quad k \in \mathbb{N}_0. \quad (4.59)$$

Proof. In the limit as $L \rightarrow \infty$, mutations always occur at different sites and so we may use the duality with the Hoppe urn model. As we saw in Section 4.2.1, in the limit as $N \rightarrow \infty$, with time multiplied by N , two lineages mutate at rate $2\mu(N)N = \theta$ and coalesce at rate $\frac{1}{N}N = 1$. Therefore the probability of *coalescence before mutation* is $1/(\theta + 1)$. If mutation occurs before coalescence, which happens with probability $\theta/(\theta + 1)$, then there is an equal probability of having another mutation before coalescence, i.e., the system starts from scratch. Hence, the probability to have k mutations followed by coalescence is given by the right-hand side of (4.59). ■

It follows from Lemma 4.1 that

$$\mathbb{E}(\Delta_2) = \theta, \quad \text{VAR}(\Delta_2) = \theta(1 + \theta). \quad (4.60)$$

As mentioned at the end of Section 3.2.1, it is of interest to estimate θ from data. Therefore we ask the same question for a sample of size n ($\ll N \rightarrow \infty$) drawn from the

population. To that end, given n genes, let Δ_{ij} be the number of pairwise differences between the i -th and the j -th sequence. Put

$$\Delta_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \Delta_{ij}, \quad (4.61)$$

which is the empirical mean of the pairwise differences in the n -sample.

Lemma 4.2 *For $n \geq 2$,*

$$\mathbb{E}(\Delta_n) = \theta, \quad \mathbb{V}\mathbb{A}\mathbb{R}(\Delta_n) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2. \quad (4.62)$$

Proof. The first claim is immediate from the fact that $\mathbb{E}(\Delta_{ij}) = \theta$ for all $1 \leq i < j \leq n$, as is obvious from (4.61) and the first half of (4.60). The second claim cannot be deduced from the second half of (4.60) because the Δ_{ij} are *not* independent. In fact, we need to compute three different variances, as we show next.

Square (4.61) to get

$$[\Delta_n]^2 = \binom{n}{2}^{-2} \sum_{1 \leq i < j \leq n} \Delta_{ij} \sum_{1 \leq k < l \leq n} \Delta_{kl}. \quad (4.63)$$

There are three types of terms in this sum:

- (1) $\binom{n}{2}$ terms with $i = k, j = l$.
- (2) $\binom{n}{2} 2(n-2)$ terms with $i = k, j \neq l$ or $i \neq k, j = l$.
- (3) $\binom{n}{2} \binom{n-2}{2}$ terms with $i \neq k, j \neq l$.

Hence, using exchangeability, we have

$$\begin{aligned} \mathbb{V}\mathbb{A}\mathbb{R}(\Delta_n) &= \binom{n}{2}^{-2} \sum_{1 \leq i < j \leq n} \sum_{1 \leq k < l \leq n} \mathbb{C}\mathbb{O}\mathbb{V}(\Delta_{ij}, \Delta_{kl}) \\ &= \binom{n}{2}^{-1} \left[U_2 + 2(n-2)U_3 + \binom{n-2}{2} U_4 \right] \end{aligned} \quad (4.64)$$

with

$$\begin{aligned} U_2 &= \mathbb{C}\mathbb{O}\mathbb{V}(\Delta_{12}, \Delta_{12}) = \mathbb{V}\mathbb{A}\mathbb{R}(\Delta_{12}), \\ U_3 &= \mathbb{C}\mathbb{O}\mathbb{V}(\Delta_{12}, \Delta_{13}), \\ U_4 &= \mathbb{C}\mathbb{O}\mathbb{V}(\Delta_{12}, \Delta_{34}). \end{aligned} \quad (4.65)$$

Thus, it remains to compute U_2, U_3, U_4 . In other words, general n reduces to $n = 4$.

We already know that $U_2 = \theta + \theta^2$. A straightforward but somewhat lengthy calculation (see Durrett [10], Section 1.4.3) yields

$$\begin{aligned} U_3 &= \frac{1}{2}\theta + \frac{1}{3}\theta^2, \\ U_4 &= \frac{1}{3}\theta + \frac{2}{9}\theta^2. \end{aligned} \quad (4.66)$$

After substituting this into (4.64), we get

$$\mathbb{V}\mathbb{A}\mathbb{R}(\Delta_n) = \binom{n}{2}^{-1} \left[\frac{1}{6}n(n+1)\theta + \frac{1}{9}(n^2 + n + 3)\theta^2 \right], \quad (4.67)$$

which proves the second claim in (4.62). ■

The first half of Lemma 4.2 says that Δ_n is an *unbiased estimator* of θ . However, the second half says that Δ_n is not asymptotically sharp, because $\lim_{n \rightarrow \infty} \mathbb{V}\mathbb{A}\mathbb{R}(\Delta_n) = U_4 \neq 0$. The fact that $U_4 \neq 0$ is due to coalescence: even though mutations occur independently in the sequences labelled 1, 2, 3, 4, coalescence between these sequences creates a dependence between Δ_{12} and Δ_{34} (the pairwise differences between 1, 2, respectively, 3, 4). This is precisely why the computation of U_3, U_4 is somewhat delicate.

Despite the lack of asymptotic sharpness of Δ_n as $n \rightarrow \infty$, Lemma 4.2 still allows for an estimate of θ with some confidence interval. This provides an alternative estimate of θ compared to what was mentioned in Section 4.2.1.

The term with θ in $\mathbb{V}\mathbb{A}\mathbb{R}(\Delta_n)$ is referred to as the *mutational variance*, because it is due to the occurrence of mutations on the genealogical tree, while the term with θ^2 is referred to as the *evolutional variance*, because it is due to fluctuations in the shape of the genealogical tree.

5 Wright-Fisher with selection

In all the models considered so far the evolution of the population was *neutral*: all types were equally fit for reproduction. In this section we look at the two-type WF-model with *selection*: one type reproduces faster than the other. We will pass to continuous time right away, so we start from the Moran-model introduced in Section 2.1.2. We will see that the slower type has an advantage over the faster type: the former is less likely to become extinct than the latter.

In Section 5.1 we look at two main effects of selection, while in Section 5.2 we look at weak selection, i.e., the selection parameter tends to zero as $N \rightarrow \infty$ in an appropriate manner.

5.1 Selection

Our model is the birth-death process $X = (X_t)_{t \geq 0}$ on state space $\Omega = \{0, 1, \dots, N\}$ with transition rates

$$\begin{aligned} i \rightarrow i+1 & \text{ at rate } b_i = (N-i) \frac{i}{N}, \\ i \rightarrow i-1 & \text{ at rate } d_i = i \frac{N-i}{N} (1-s), \end{aligned} \quad (5.1)$$

where X_t is the number of individuals of type A at time t in the population of size N , and $s \in (0, 1)$ is a *selection parameter*. Here, as before, individuals randomly choose an

ancestor and adopt its type, but type a resamples at rate 1 while type A resamples at rate $1 - s$. Note that

$$\frac{d_i}{b_i} = 1 - s, \quad \forall i \in \Omega, \quad (5.2)$$

i.e., the *relative fitness* of types A and a is $1 - s$.

Remark: The above selection is referred to as *viability selection*. An alternative way of defining the model is as follows: a 's reproduce at rate 1, A 's reproduce at rate $1 - s$, and their offspring replaces a randomly chosen individual. The latter is known as *fertility selection*.

As before, the states $i = 0$ (all a) and $i = N$ (all A) are traps, so that eventually type fixation will occur. We are interested in how the parameter s influences the probability of fixation at these two traps as well as the fixation time.

We saw in Lemma 2.1 that in the neutral case $s = 0$ the probability of fixation at N given $X_0 = i$ equals $\frac{i}{N}$. This now changes as follows (see Fig. 13).

Lemma 5.1 *Let $s \in (0, 1)$. Let $\tau = \inf\{t \geq 0: X_t = 0 \text{ or } X_t = N\}$. Then*

$$\mathbb{P}(X_\tau = N \mid X_0 = i) = \frac{1 - (1 - s)^i}{1 - (1 - s)^N}, \quad i \in \Omega. \quad (5.3)$$

Proof. Abbreviate $g(i) = \mathbb{P}(X_\tau = N \mid X_0 = i)$. By considering what happens at the first transition away from i , we have

$$g(i) = \frac{b_i}{b_i + d_i} g(i + 1) + \frac{d_i}{b_i + d_i} g(i - 1), \quad i \in \Omega \setminus \{0, N\}. \quad (5.4)$$

This recursion is to be solved subject to the boundary conditions $g(0) = 0$ and $g(N) = 1$. Rearranging the terms in (5.4), we may write

$$[g(i + 1) - g(i)] = \frac{d_i}{b_i} [g(i) - g(i - 1)] = (1 - s) [g(i) - g(i - 1)], \quad i \in \Omega \setminus \{0, N\}. \quad (5.5)$$

Use $g(0) = 0$, put $c = g(1)$ and iterate (5.5), to get

$$g(i + 1) - g(i) = c(1 - s)^i, \quad i \in \Omega \setminus \{N\}. \quad (5.6)$$

Summing this relation over $i = 0, \dots, j - 1$, we obtain

$$g(j) = \sum_{i=0}^{j-1} c(1 - s)^i = \frac{c}{s} [1 - (1 - s)^j], \quad j \in \Omega. \quad (5.7)$$

Since $g(N) = 1$, we have

$$\frac{c}{s} = [1 - (1 - s)^N]^{-1}. \quad (5.8)$$

Combining (5.7–5.8), we get the claim. ■

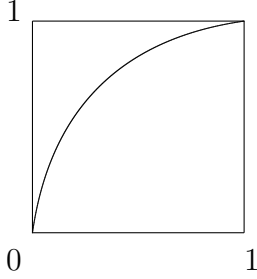


Figure 13: Qualitative picture of (5.3).

Fig. 13 shows the effect of selection: the slower type A has a selective advantage over the faster type a , resulting in a bias of the survival probability towards the slower type. Slow reproduction causes type A to be “less prone to wiping itself out”.

The average fixation time $\mathbb{E}(\tau)$ is not so easy to compute, unlike what we saw in Section 2.1 for the neutral WF-model ($s = 0$), where we found that $E(\tau \mid H_0) \geq NH_0$ with H_0 the genetic variability at time 0 (compare Durrett [10, Theorems 1.32 and 6.1]). The following lemma identifies the average of τ *conditional* on $X_0 = 1$ and $X_\tau = N$.

Lemma 5.2 *Let $s \in (0, 1)$. Then*

$$\mathbb{E}(\tau \mid X_0 = 1, X_\tau = N) \sim \frac{2}{s} \log N \quad \text{as } N \rightarrow \infty. \quad (5.9)$$

Proof. The proof is given in Durrett [10], Section 6.1. It is based on a somewhat lengthy calculation showing that it takes time

- (1) $\sim \frac{1}{s} \log N$ to go from $i = 1$ to $i = \frac{N}{\log N}$;
- (2) $\sim 2 \log \log N$ to go from $i = \frac{N}{\log N}$ to $i = N - \frac{N}{\log N}$;
- (3) $\sim \frac{1}{s} \log N$ to go from $i = N - \frac{N}{\log N}$ to $i = N$.

Thus, after the start at $X_0 = 1$ the system spends a long time near 0 (mostly a 's), then rapidly moves accross and spends a long time near N (mostly A 's), until it hits N (because of the condition $X_\tau = N$). ■

The fast crossing from mostly a 's to mostly A 's is called a *selective sweep*.

The computations in Section 2.1 show that, in the absence of selection ($s = 0$), $\mathbb{E}(\tau \mid X_0 = 1, X_\tau = N)$ is of order N , which is much larger than the order $\log N$ found in Lemma 5.2.

5.2 Weak selection

Interesting behavior occurs when there is *weak selection*, in particular, when $N \rightarrow \infty$ and

$$s = s(N) = \frac{\sigma}{2N}, \quad \sigma \in (0, \infty). \quad (5.10)$$

Think of $\frac{1}{2}\sigma$ as the *fitness advantage of A w.r.t. a in the population*. (The choice to divide by $2N$ rather than N is a matter of convention.)

Write $X_t^{(N)}$ to exhibit the N -dependence and define, in analogy with (2.14),

$$Y_t^{(N)} = \frac{1}{N} X_{Nt}^{(N)}, \quad t \geq 0. \quad (5.11)$$

The following result is the analogue of Theorem 2.1. We assume that

$$w - \lim_{N \rightarrow \infty} Y_0^{(N)} = Y_0 \quad (5.12)$$

and ask whether

$$w - \lim_{N \rightarrow \infty} (Y_t^{(N)})_{t \geq 0} = (Y_t)_{t \geq 0}. \quad (5.13)$$

Theorem 5.1 *The scaling in (5.13) subject to (5.12) is true, with $(Y_t)_{t \geq 0}$ the diffusion process on $[0, 1]$ given by the stochastic differential equation*

$$dY_t = \frac{1}{2}\sigma Y_t(1 - Y_t) dt + \sqrt{Y_t(1 - Y_t)} dW_t. \quad (5.14)$$

Exercise 5.2 *Give the proof of Theorem 5.1 along the same lines as the proof of Theorem 2.1. Use that*

$$b_{[Ny]} - d_{[Ny]} \sim Ny(1 - y)[1 - (1 - s)] = Nsy(1 - y) = \frac{1}{2}\sigma y(1 - y) \quad N \rightarrow \infty. \quad (5.15)$$

The stochastic differential equation in (5.14) has two parts: a *deterministic* part, given by a logistic drift term pushing the system towards all A , and a *random* part, given by the WF-diffusion term. Like (2.17), it has a unique strong solution. A dual process exists, but this is more complex than the death process we encountered in Theorem 2.1 acting as the dual process for the WF-diffusion ($\sigma = 0$). The dual process in the presence of selection requires a graphical construction (see Durrett [10], pp. 127–128). We see from (5.11–5.13) and Theorem 5.1 that under weak selection the fixation time is again of order N , so that on time scale N the fixation time is of order 1 (just as for the standard WF-model studied in Section 2.2).

Let us next see what happens when we add *mutation*. Then the states $i = 0$ and $i = N$ are no longer traps and the system has a non-trivial equilibrium $Y^{(N)}$. We consider the limit $N \rightarrow \infty$ with

$$u = u(N) = \frac{q}{2N}, \quad v = v(N) = \frac{r}{2N}, \quad s = s(N) = \frac{\sigma}{2N}. \quad (5.16)$$

The following result generalizes Theorem 4.2.

Theorem 5.3 *Subject to (5.16), the convergence in (4.15–4.16) is true, with Y the random variable on $[0, 1]$ with probability density*

$$f(x) = C_{q,r,\sigma} x^{q-1} (1-x)^{r-1} e^{\sigma x/2}, \quad x \in [0, 1], \quad (5.17)$$

with $C_{q,r,\sigma}$ the normalizing constant.

Proof. The proof for $\sigma = 0$ was based on an asymptotic expansion (recall (4.21–4.24)). The proof for $\sigma > 0$ can be done quickly by *comparison* with the case $\sigma = 0$. To that end, let $\pi(i)$, $i \in \Omega$, be the stationary distribution before passing to the limit $N \rightarrow \infty$. Then

$$\pi(i)d_i = \pi(i-1)b_{i-1}, \quad i \in \Omega \setminus \{0\}. \quad (5.18)$$

Hence, we have

$$\frac{\pi(i)}{\pi(0)} = \left[\prod_{j=1}^i \frac{b_{j-1}}{d_j} \right] = b_0 \left[\prod_{j=1}^{i-1} \frac{b_j}{d_j} \right] \frac{1}{d_i}, \quad i \in \Omega \setminus \{0\}. \quad (5.19)$$

Let b_i^*, d_i^* be the birth and death rates for the model *without* selection, given by (2.12), and let $\pi^*(i)$, $i \in \Omega$, be the associated stationary distribution. For this model the same recursion as in (5.18) applies, namely,

$$\pi^*(i)d_i^* = \pi^*(i-1)b_{i-1}^*, \quad i \in \Omega \setminus \{0\}, \quad (5.20)$$

and so the analogue of (5.19) reads

$$\frac{\pi^*(i)}{\pi^*(0)} = \left[\prod_{j=1}^i \frac{b_{j-1}^*}{d_j^*} \right] = b_0^* \left[\prod_{j=1}^{i-1} \frac{b_j^*}{d_j^*} \right] \frac{1}{d_i^*}, \quad i \in \Omega \setminus \{0\}. \quad (5.21)$$

Combining (5.19) and (5.21) with the observation that $b_i = b_i^*$ and $d_i = d_i^*(1-s)$, we obtain

$$\frac{\pi(i)}{\pi(0)} = (1-s)^{-i} \frac{\pi^*(i)}{\pi^*(0)}, \quad i \in \Omega \setminus \{0\}. \quad (5.22)$$

Now pick $i = \lceil Nx \rceil$ with $x \in (0, 1)$ and let $N \rightarrow \infty$. Then, since

$$(1-s)^{-i} \sim \left(1 - \frac{\sigma}{2N}\right)^{-Nx} \sim e^{\sigma x/2}, \quad n \rightarrow \infty, \quad (5.23)$$

it follows that the equilibrium distribution with selection carries an extra factor $e^{\sigma x/2}$ compared to the equilibrium distribution without selection, explaining the difference between (5.17) and (4.17). The two equilibrium distributions have different normalizing constants, which are cancelled out because we divide by $\pi(0)$, respectively, $\pi^*(0)$ before taking the limit $N \rightarrow \infty$. Qualitatively the same pictures as in Fig. 12 apply. ■

Weak selection tilts the equilibrium achieved under weak mutation.

Part II

Multi-colony Wright-Fisher populations

In Part I (Chapters 2–5) we considered models where a *single* population evolves under *resampling*, *mutation* and *selection*. In Part II (Chapters 6–7) we move on to include *migration*, i.e., we include the effect of spatial motion. This leads to models with a large number of *interacting populations*.

We imagine that each population lives in a *colony* and that different colonies are organized into a *lattice* $\mathcal{L}$, which labels the colonies. In what follows we will consider two choices:

1. $\mathcal{L} = \mathbb{Z}^d$, the Euclidean lattice in dimension $d \in \mathbb{N}$ (Chapter 6),
2. $\mathcal{L} = \Omega_M$, the hierarchical lattice of order $M \in \mathbb{N}$ (Chapter 7).

We allow the individuals to choose their ancestors not only from their *own* colony, but also from *other* colonies, according to a prescribed random walk transition kernel $p(x, y)$, $x, y \in \mathcal{L}$. This is called *migration*, because it causes the types to migrate from one colony to another. Note that the individuals themselves do not migrate: in each colony the population remains fixed at N individuals, just as in the single-colony Wright-Fisher model.

Migration of types rather than individuals is appropriate e.g. for flowers and plants, where pollen or seeds travel by wind and via insects. It is less appropriate for animals.

6 The stepping stone model

In Section 6.1 we introduce the model. In Section 6.2 we derive a formula for the probability that two individuals, drawn randomly from the population, are identical by descent. In Sections 6.3.1–6.4.1 we analyse this probability in some specific setups.

6.1 Model

In this section we consider interacting Wright-Fisher populations *with* mutation but *without* selection. The colonies are labelled by the points of the lattice

$$\mathcal{L} = \mathbb{Z}^d, \quad (6.1)$$

which we refer to as the *geographic space* (see Fig. 14 for the special case $d = 2$) and are subject to *migration* according to a random walk transition kernel (see Appendix A.2.4)

$$p_\nu(x, y) = (1 - \nu)\delta_{x,y} + \nu Q(x, y), \quad x, y \in \mathbb{Z}^d, \quad (6.2)$$

where $\nu \in (0, 1]$ is a parameter and $Q: \mathcal{L} \times \mathcal{L} \rightarrow [0, 1]$ is a translation invariant transition kernel (w.l.o.g. we may assume that $Q(x, x) = 0$). We define a probability distribution $q: \mathbb{Z}^d \rightarrow [0, 1]$ by putting $Q(x, y) = q(y - x)$ (with $q(0) = 0$). We further assume that q has a finite third moment (i.e., $\sum_{y \in \mathbb{Z}^d} y^3 q(y) < \infty$) is symmetric:

$$\forall i = 1, \dots, d: \quad q(x_1, \dots, x_i, \dots, x_d) = q(x_1, \dots, -x_i, \dots, x_d).$$

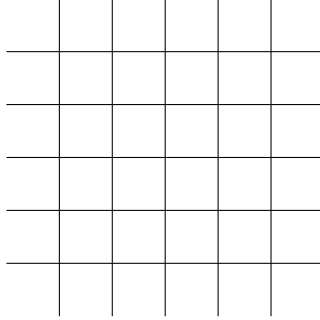


Figure 14: $\mathbb{Z}^2$: the square lattice with sites and nearest-neighbour bonds.

The evolution of the system is as follows:

- (1) At each unit of time, each individual mutates to a new type with probability $\mu \in (0, 1)$.
- (2) At each unit of time, each individual conditional on no mutation chooses a colony and adopts the type of a randomly chosen individual from that colony. Individuals in colony x choose colony y with probability $p_\nu(x, y)$.

- (3) Each colony contains N individuals. All individuals mutate and migrate independently (“parallel dynamics”).

The migration under (6.2) amounts to individuals choosing their own colony with probability $1 - \nu$ and a different colony with probability ν . Given that they choose a different colony, the choice is made with a probability that only depends on the distance to this colony. The parameter ν is the *migration probability*. Note that the mutation is like in the infinitely many alleles model studied in Section 3.2: each mutation introduces a new type.

6.2 Lineages

We are interested in the lineage of two individuals randomly drawn from colonies x and y when the system is in equilibrium. When $x = y$, we assume that two *distinct* individuals are drawn from the colony. Due to the presence of mutations, there are no traps and the system has an *ergodic equilibrium*.

Let (compare with (4.11))

$$\psi(x, y) = \text{probability in equilibrium that two individuals randomly drawn from colonies } x \text{ and } y \text{ are } \textit{identical by descent}. \quad (6.3)$$

Clearly, for the analysis of the genealogy of the populations this is a key quantity. We will compute $\psi(x, y)$ and study its dependence on x and y . By translation invariance, $\psi(x, y)$ is a function of $y - x$ only. We therefore write

$$\phi(x) = \psi(0, x). \quad (6.4)$$

We begin by expressing $\phi(y - x)$ in terms of $\phi(0)$ and the iterates of the transition kernel $p_\nu(x, y)$.

Theorem 6.1 *For all $x, y \in \mathbb{Z}^d$,*

$$\phi(y - x) = \psi(x, y) = \frac{1 - \phi(0)}{N} G_{\mu, \nu}(x, y) = \frac{G_{\mu, \nu}(x, y)}{N + G_{\mu, \nu}(0, 0)}, \quad (6.5)$$

where $G_{\mu, \nu}(x, y) = \sum_{n=1}^{\infty} (1 - \mu)^{2n} p_\nu^{2n}(x, y)$.

Proof. Let $\phi_n(y - x)$ be the probability in the right-hand side of (6.3) at time n when the system starts from a given initial configuration (e.g. all types 0). Then we have the recursion relation

$$\begin{aligned} \psi_{n+1}(x, y) &= (1 - \mu)^2 \sum_{z \in \mathbb{Z}^d} p(x, z) p(y, z) \left[\frac{1}{N} + \left(1 - \frac{1}{N}\right) \psi_n(z, z) \right] \\ &\quad + (1 - \mu)^2 \sum_{\substack{x', y' \in \mathbb{Z}^d \\ x' \neq y'}} p(x, x') p(y, y') \psi_n(x', y'). \end{aligned} \quad (6.6)$$

Here, we sum over the two steps $x \rightarrow x'$ and $y \rightarrow y'$ in the lineages of the two individuals between time n and time $n + 1$, and we require that no mutation occurs during either of these steps, so that the lineages continue from x' and y' . The first sum deals with the case $x' = y' = z$: with probability $\frac{1}{N}$ the same individual is chosen in colony z and coalescence occurs, while with probability $1 - \frac{1}{N}$ different individuals are chosen in colony z and the lineages continue from z . (Note that each step in the random walk causes a type to move in the opposition direction.)

By *ergodicity*, we have

$$\lim_{n \rightarrow \infty} \psi_n(x, y) = \psi(x, y), \quad x, y \in \mathbb{Z}^d. \quad (6.7)$$

Hence (6.6) yields

$$\psi(x, y) = (1 - \mu)^2 \sum_{z \in \mathbb{Z}^2} p(x, z) p(y, z) \frac{1 - \psi(z, z)}{N} + (1 - \mu)^2 \sum_{x', y' \in \mathbb{Z}^2} p(x, x') p(y, y') \psi(x', y'). \quad (6.8)$$

The second sum in (6.8) is the limit of the second sum in (6.6) with the diagonal part added. The first sum in (6.8) is the difference of the limit of the first sum in (6.6) and the diagonal part of the second sum in (6.8). The first sum in (6.8) gives the first summand in (6.5), i.e., the term with $n = 1$, because $\psi(z, z) = \phi(0)$ and $\sum_z p_\nu(x, z) p_\nu(y, z) = \sum_z p_\nu(x, z) p_\nu(z, y) = p_\nu^2(x, y)$ (use the symmetry of p_ν). The second sum in (6.8) can be iterated by resubstitution of (6.8). This generates all the higher summands in (6.5), i.e., $\sum_{n=1}^{\infty} (1 - \mu)^{2n} p_\nu^{2n}(x, y) = G_{\mu, \nu}(x, y)$.

By picking $x = y = 0$ in (6.5) and solving for $\phi(0)$, we get

$$\phi(0) = \frac{G_{\mu, \nu}(0, 0)}{N + G_{\mu, \nu}(0, 0)}. \quad (6.9)$$

Substituting this into (6.5), we get the claim. ■

Exercise 6.2 *Explain why the infinite series in (6.5) converges (so that the iterative argument in the proof indeed produces the correct formula).*

With Theorem 6.1 we have identified the N -dependence of $\psi(x, y)$, and all that we need to do is analyze how $G_{\mu, \nu}(x, y)$ behaves as a function of $y - x$ and μ, ν . This is a basic calculation involving simple random walk on $\mathbb{Z}^d$ with pausing (recall (6.2–6.12)). It is easy to analyse $G_{\mu, \nu}(x, y)$ *numerically*. However, let us see what can be done *analytically*. In what follows we focus on the cases $d = 1$ and $d = 2$.

6.3 One-dimensional geographic space

6.3.1 One-dimensional lattice

For $d = 1$ Theorem 6.1 yields the following asymptotic results:

Theorem 6.1 Consider a stepping stone model on $\mathbb{Z}$ with a symmetric translation invariant transition kernel with a finite third moment. Let $\sigma^2 = \sum_{y \in \mathbb{Z}} y^2 q(y)$. Define a characteristic length scale by putting $\ell = (\nu\sigma^2/2\mu)^{1/2}$. Then

$$\phi(0) \sim \frac{1}{1 + 2N(2\mu\nu\sigma^2)^{1/2}}, \quad (6.10)$$

$$\phi(x) \sim \phi(0) \exp(-|x|/\ell), \quad (6.11)$$

where $\sim$ means asymptotic equivalence as $\frac{\nu}{\mu} \rightarrow 0$.

See Durrett [10, Theorem 5.5] for a proof.

If we choose a symmetric simple random walk, i.e.,

$$q(z) = \begin{cases} \frac{1}{2} & \text{if } \|z\| = 1, \\ 0 & \text{otherwise,} \end{cases} \quad (6.12)$$

then we can even obtain an exact result:

Theorem 6.2 In the symmetric nearest-neighbour stepping stone model on $\mathbb{Z}$,

$$\phi(x) = \frac{\lambda_2^{|x|}}{1 + 2N\mu + 2N\nu(1 - \lambda_2)}, \quad (6.13)$$

where

$$\lambda_2 = 1 + \frac{\mu}{\nu} - \sqrt{\frac{2\mu}{\nu} + \frac{\mu^2}{\nu^2}}. \quad (6.14)$$

See Durrett [10, Theorem 5.1] for a proof.

Note that if $\frac{\mu}{\nu} \ll 1$, then $\lambda_2 \sim 1 - \sqrt{\frac{2\mu}{\nu}}$, $\lambda_2^{|x|} \sim \exp(-\sqrt{\frac{2\mu}{\nu}}|x|)$, and so

$$\phi(0) \sim \frac{1}{1 + 2N\sqrt{2\mu\nu}}, \quad (6.15)$$

$$\phi(x) = \frac{\lambda_2^{|x|}}{1 + 2N\mu + 2N\nu(1 - \lambda_2)} \sim \phi(0) \exp(-\sqrt{\frac{2\mu}{\nu}}|x|),$$

since $\sigma = 1$, $\ell = \nu/2\mu$. We see that the results from Theorem 6.1 in the case $\frac{\mu}{\nu} \ll 1$ are consistent with the exact results of Theorem 6.2.

6.3.2 One-dimensional lattice torus

We next investigate how the probability of being identical by descent behaves on a finite one-dimensional torus of size L , i.e., a ring of L -colonies. More precisely, we consider

$$\mathcal{L} = \mathbb{T}_L = \mathbb{Z} \cap [0, L - 1] = \{0, 1, \dots, L - 1\} \quad (6.16)$$

with *periodic boundary conditions* (which turns $\mathbb{T}_L$ into a torus and preserves the translation invariance). Accordingly, we write $\phi_L(x)$ for the probability for two lineages, sampled from colonies that differ by x modulo L , to be identical by descent (recall (6.3)).

We have the following theorem (see Durrett [10, Theorem 5.2] for a proof).

Theorem 6.3 *For the symmetric nearest-neighbour stepping stone model on $\mathbb{Z}/L\mathbb{Z}$,*

$$\phi(x) = C(\lambda_+^{xL/2} + \lambda_-^{xL/2}), \quad (6.17)$$

where $\lambda_+ > \lambda_-$ are given by

$$\lambda_{\pm} = 1 + \frac{\mu}{\nu} \pm \sqrt{2\frac{\mu}{\nu} + \frac{\mu^2}{\nu^2}} \quad (6.18)$$

and

$$1/C = (1 + 2N\mu)(\lambda_1^{-L/2} + \lambda_2^{-L/2}) + 2N\nu[(1 - \lambda_1)\lambda_1^{-L/2} + (1 - \lambda_2)\lambda_2^{-L/2}]. \quad (6.19)$$

It is reasonable to expect a difference in behaviour for $\phi_L(x)$ depending on whether

$$\begin{aligned} \text{I. } & \frac{1}{\mu} \ll \frac{1}{\nu} L^2, \\ \text{II. } & \frac{1}{\mu} \gg \frac{1}{\nu} L^2. \end{aligned} \quad (6.20)$$

Regime I corresponds to the system *not feeling the boundary of the box*. Indeed, $\frac{1}{\mu}$ is the average time required for a mutation to occur in one lineage. Consequently, $\frac{\nu}{\mu}$ is the variance of the displacement of a lineage before a mutation occurs (the variance per step is ν by (6.12)). If this ratio is much smaller than L^2 , then a mutation will occur before the lineages note the boundary of the box, and the behavior of $\phi_L(x)$ will be essentially independent of L . Regime II, on the other hand, corresponds to the situation where the lineages *wrap themselves around the torus many times* before a mutation occurs. In that case, the behavior of $\phi_L(x)$ will depend on L and will be more delicate.

The above heuristics can be actually turned into a rigorous statement. In fact, suppose that $\mu \ll \nu$, i.e., the mutation rate is much smaller than the migration rate. As a consequence of Theorem 6.3 we get the following qualitative results:

- If $\frac{1}{\mu} \ll L^2 \frac{1}{\nu}$, then $\phi_L(x) \sim \phi(x)$, i.e., the behaviour of ϕ_L obtained in Theorem 6.3 is similar to that of ϕ found in Theorem 6.2. Hence, there is little difference between the ring and the line.
- If $\frac{1}{\mu} \gg L^2 \frac{1}{\nu}$, then $\phi_L(x) \sim \frac{1}{1+2N\mu}$, i.e., the behaviour of ϕ_L obtained in Theorem 6.3 is similar to that of a single-site population in the Wright-Fisher model (see the asymptotic behaviour of (4.6) as $\mu \downarrow 0$). Hence, the stepping stone model behaves like a homogeneously mixing population (also called ‘panmictic’ behaviour).

See Durrett [10, Theorem 5.3] for further details.

6.4 Two-dimensional geographic space

6.4.1 Two-dimensional lattice

We next move to $d = 2$. From the general statement in Theorem 6.1 the following estimates can be derived.

Theorem 6.1 *Consider a stepping stone model on $\mathbb{Z}^2$ with symmetric translation invariant transition kernel with a finite third moment. Define $\ell = (\nu\sigma^2/2\mu)^{1/2}$. Then*

$$\phi(z) \sim \frac{1}{2\pi\nu N + \log(\ell)} \left[K_0\left(\frac{1}{\ell}\|z\|\right) - K_0(\|z\|) \right], \quad (6.21)$$

where $\sim$ means asymptotic equivalence as $\mu \downarrow 0$ and

$$K_0(u) = \int_0^\infty dv (1+v^2)^{-1/2} \cos(vu), \quad u \in (0, \infty), \quad (6.22)$$

is the modified Bessel function of the second kind of order 0, which satisfies

$$K_0(u) \sim \begin{cases} \log(1/u) & \text{for } u \downarrow 0, \\ \sqrt{\pi/2u} e^{-u} & \text{for } u \uparrow \infty. \end{cases} \quad (6.23)$$

See Durrett [10, Theorem 5.7] for a proof.

The result in Theorem 6.1 shows how the probability for two individuals to be identical by descent depends on their distance. In what follows we take a closer look at several subregimes

6.4.2 Two-dimensional lattice torus

As in Section 6.3.2, we may expect different regimes (see (6.20)) to hold when we restrict to the torus of size L , i.e.,

$$\mathcal{L} = \mathbb{T}_L^2 = \mathbb{Z}^2 \cap [0, L-1]^2 = \{0, 1, \dots, L-1\}^2$$

with periodic boundary conditions.

In regime I ($L^2 \gg \nu/\mu$), as for $d = 1$, the type does not have time to feel the boundary because mutation typically occurs much earlier. Therefore the behaviour is asymptotically equivalent to that described in Theorem 6.1.

Regime II, however, is harder to come by. Our first intuition is that the system behaves like a “homogeneously mixing population”, because the lineages wrap themselves around the torus many times before a mutation occurs and therefore sample the full torus. In other words, the system behaves like a Wright-Fisher model with a single colony containing NL^2 individuals: the ‘panmictic’ behaviour mentioned before. However, it turns out that this intuition is not quite correct. In fact, we will see that panmictic behaviour only shows up when

$$\mu \ll 1 \quad \text{and} \quad N\nu \gg \log L, \quad (6.24)$$

which requires that $N = N(L)$ and $\nu = \nu(L)$, i.e., N and ν depend on L .

In Section 6.4.3 we first look at Regime II without mutation. In Section 6.4.5 we add mutation and compute $\phi(y-x)$.

6.4.3 Coalescing times

To investigate Regime II, we first consider the case $\mu = 0$. Since mutations occur independently of resampling and migration, it will be easy to incorporate them later.

Getting the two lineages to the same colony is necessary to achieve coalescence. Afterwards, either coalescence occurs at that colony (when the lineages choose the same ancestor), or no coalescence occurs (when the lineages choose different ancestors). In the latter case the lineages must proceed further to achieve coalescence. The *time until coalescence* equals

$$\tau = T_0 + T_R \quad (6.25)$$

with

$$\begin{aligned} T_0 &= \text{the time until the two lineages meet in the same colony,} \\ T_R &= \text{the subsequent time lapse until coalescence.} \end{aligned} \quad (6.26)$$

Moreover,

$$T_R = T_1 + T_2 + \cdots + T_\sigma, \quad (6.27)$$

where $(T_i)_{i \in \mathbb{N}}$ are i.i.d. copies of the *return times* of the two lineages (which are independent of T_0), and σ is an independent geometrically distributed random variable with mean N . The reason for σ is that upon each collision the two lineages have probability $\frac{1}{N}$ to coalesce.

We first focus on T_0 . We write $\mathbb{P}_\pi$ for the law of the two lineages when they start from two *randomly* chosen locations in the L -torus.

Lemma 6.2 *For $\mu = 0$ and $\nu \in (0, 1]$,*

$$w - \lim_{L \rightarrow \infty} \frac{T_0}{(L^2 \log L)/2\pi\nu} = \text{EXP}(1) \quad \text{under } \mathbb{P}_\pi. \quad (6.28)$$

Proof. The proof is given in Cox and Durrett [6]. The intuition is that it takes order $L^2 \log L$ steps for the two lineages to visit the L -torus well enough so that they have an appreciable probability to meet somewhere. This corresponds to $(L^2 \log L)/\nu$ time units, because each lineage steps with probability ν at each time unit. The exponential limiting law is typical for “many trials with small success”: the probability for the two lineages to meet in a given time interval is small and therefore they will meet only after many trials. ■

A more refined result is obtained by looking at two lineages that start not randomly but at fixed colonies, say x and 0 . We write $\mathbb{P}_{x,0}$ for their law, and δ_0 for the unit measure at 0 .

Lemma 6.3 *For $\mu = 0$, $\nu \in (0, 1]$ and $\beta \in [0, 1]$,*

$$w - \lim_{L \rightarrow \infty} \frac{T_0}{(L^2 \log L)/2\pi\nu} = (1 - \beta) \delta_0 + \beta \text{EXP}(1) \quad \text{under } \mathbb{P}_{x(L),0} \quad (6.29)$$

for any $x(L)$ such that

$$\lim_{L \rightarrow \infty} \frac{\log \|x(L)\|}{\log L} = \beta. \quad (6.30)$$

Proof. See Durrett [10], Section 5.3. ■

Roughly speaking, Lemma 6.3 says that, for all $\beta \in [0, 1]$, if the lineages start at distance $\|x(L)\| = L^{\beta+o(1)}$ from each other, then there is sharp control over the time when they first meet. Note that all scales $\beta \in [0, 1]$ are relevant. Lemma 6.2 corresponds to $\beta = 1$, because if x is drawn from π , then $\|x\| = L^{1+o(1)}$ with a probability tending to 1 as $L \rightarrow \infty$.

We next focus on the collision times.

Lemma 6.4 *For $\mu = 0$ and $\nu \in (0, 1]$,*

$$\mathbb{E}_0(T_R) = NL^2 \tag{6.31}$$

and

$$w - \lim_{L \rightarrow \infty} \frac{T_R}{NL^2} = \text{EXP}(1) \quad \text{under } P_{x,0} \text{ uniformly in } x \text{ on the } L\text{-torus.} \tag{6.32}$$

Proof. We write $\mathbb{E}_0$ because the two lineages start from the same site. The claim in (6.31) is a well-known mean recurrence time property of Markov chains, for which we refer to standard textbooks. The claim in (6.32) is in the same spirit as before: success only occurs after many trials. ■

Exercise 6.5 *Note that $\mathbb{E}_0(T_R) = NL^2$ is independent of ν . What is the intuition behind this independence?*

6.4.4 Without mutation

A comparison of Lemmas 6.2–6.4 shows that within Regime II there are two subregimes:

$$\begin{aligned} \text{II.A.} \quad & \frac{1}{\nu} \log L \ll N : \quad \mathbb{E}(T_0) \ll \mathbb{E}_0(T_R), \\ \text{II.B.} \quad & \frac{1}{\nu} \log L \gg N : \quad \mathbb{E}(T_0) \gg \mathbb{E}_0(T_R). \end{aligned} \tag{6.33}$$

In regime II.A, the starting locations of the two individuals do not matter, while in regime II.B they do matter.

Let us summarize the above observations and formulate the scaling behavior in a more general form that also includes an interpolation between Regime II.A and Regime II.B. (See Cox, Durrett and Zähle [7] and Cox [5].)

Theorem 6.6 *If $\lim_{L \rightarrow \infty} \frac{4\pi N\nu}{\log L} = \alpha \in [0, \infty]$, $\lim_{L \rightarrow \infty} \frac{\log \|x(L)\|}{\log L} = \beta \in [0, 1]$. Let*

$$C_L = \begin{cases} NL^2 & \text{if } \alpha = \infty \\ (1 + \alpha)^{\frac{L^2 \log L}{2\pi\nu}} & \text{if } \alpha \in [0, \infty). \end{cases} \tag{6.34}$$

Then, under $\mathbb{P}_\pi$,

$$w - \lim_{L \rightarrow \infty} \frac{\tau}{C_L} = \text{EXP}(1), \tag{6.35}$$

while, under $\mathbb{P}_{x(L),0}$,

$$\begin{aligned} \alpha = \infty : \quad & w - \lim_{L \rightarrow \infty} \frac{\tau}{c_L} = \text{EXP}(1), \\ \alpha \in [0, \infty) : \quad & w - \lim_{L \rightarrow \infty} \frac{\tau}{c_L} = (1 - \gamma)\delta_0 + \gamma \text{EXP}(1), \end{aligned} \quad (6.36)$$

where

$$\gamma = \gamma(\alpha, \beta) = \beta + (1 - \beta) \frac{\alpha}{1 + \alpha}. \quad (6.37)$$

Note that the scaling for $\mathbb{P}_\pi$ is the same as the scaling for $\mathbb{P}_{x(L),0}$ with $\beta = 1$ (in which case $\gamma = 1$).

6.4.5 Panmictic behaviour for the two-dimensional torus

Having thus obtained the asymptotics of the time to coalescence, we reintroduce mutation and return to the probability of the two lineages being equal by descent, defined in (6.3). Writing $\psi(x, y) = \phi(y - x) (= \phi(x - y))$, we have

$$\begin{aligned} \phi_L(x) &= \mathbb{E}_{0,x}([1 - \mu]^{2\tau}), \\ \phi_L &= \frac{1}{L^2} \sum_{x \in [0,L)^2 \cap \mathbb{Z}^2} \phi(x) = \mathbb{E}_\pi([1 - \mu]^{2\tau}). \end{aligned} \quad (6.38)$$

Inserting $w - \lim_{L \rightarrow \infty} \tau/C_L = (1 - \gamma)\delta_0 + \gamma \text{EXP}(1)$ into (6.38), we compute (with $\mathbb{E}$ standing for either $\mathbb{E}_\pi$ or $\mathbb{E}_{x,0}$)

$$\begin{aligned} \mathbb{E}([1 - \mu]^{2\tau}) &= \int_0^\infty [1 - \mu]^{2t} \mathbb{P}(\tau \in dt) = \int_0^\infty [1 - \mu]^{2C_L s} \mathbb{P}\left(\frac{\tau}{C_L} \in ds\right) \\ &\sim (1 - \gamma) + \gamma \int_0^\infty e^{-2C_L s \log(\frac{1}{1-\mu})} e^{-s} ds = 1 - \gamma + \gamma \frac{1}{1 + 2C_L \log(\frac{1}{1-\mu})}, \end{aligned} \quad (6.39)$$

where $\sim$ refers to $L \rightarrow \infty$. For $\mu \downarrow 0$ we have $\log(\frac{1}{1-\mu}) \sim \mu$, and so we arrive at

$$\mathbb{E}([1 - \mu]^{2\tau}) \sim (1 - \gamma) + \gamma \frac{1}{1 + 2C_L \mu}, \quad (6.40)$$

where $\sim$ refers to $L \rightarrow \infty$ and $\mu \downarrow 0$.

We can now insert the various choices for C_L and γ appearing in Theorem 6.6, to obtain the asymptotic behavior of ϕ_L and $\phi_L(x)$ in the various cases. For instance, for ϕ_L (corresponding to $\gamma = 1$) we find that if $\alpha = \infty$, then

$$\phi_L \sim \frac{1}{1 + 2N_{\text{eff}}\mu} \quad \text{with } N_{\text{eff}} = NL^2. \quad (6.41)$$

This is the same as the probability for two individuals to be identical by descent in a *single colony* Wright-Fisher model with N_{eff} individuals (recall the calculation of χ in (4.6) in the proof of Lemma 4.1). Thus, in this limit where $\alpha = \infty$, the stepping model behaves as if all individuals are thrown into a single colony and ancestors are chosen in

a uniform manner from the entire population, which is the *panmictic* behavior alluded to at the beginning of this section. In this case, apparently the spatial structure of the stepping stone model is irrelevant and it exhibits so-called *mean-field behavior*, i.e., all individuals influence each other in the same manner. Similarly, if $\alpha \in (0, \infty)$, then

$$\phi_L \sim \frac{1}{1 + 2N_{\text{eff}}\mu} \quad \text{with } N_{\text{eff}} = (1 + \alpha) \frac{L^2 \log L}{2\pi\nu} = \frac{1 + \alpha}{\alpha} NL^2. \quad (6.42)$$

Here, the effective population size N_{eff} is the total population NL^2 times a moderation factor $\frac{1+\alpha}{\alpha}$ that encapsulates the effect of the spatial structure.

Exercise 6.7 Note that $\frac{1+\alpha}{\alpha} > 1$. Explain why it is plausible that $N_{\text{eff}} > NL^2$.

7 Hierarchical models

In this section we make a different choice for the lattice $\mathcal{L}$ labelling the colonies, namely, the so-called *hierarchical lattice of order M* . We will show that in the limit as $M \rightarrow \infty$ this model displays *universal behavior on large space-time scales*. This universal behavior will turn out to be the result of repeated application of a *renormalization transformation* connecting successive hierarchical levels.

Remark: Renormalization is a key tool in *statistical physics*, an area that aims to describe the large space-time behavior of interacting particle systems. The idea is that on large space-time scales an effective law of large numbers is in force, allowing us to describe the macroscopic behavior of such systems in terms of partial differential equations that capture the flow of macroscopic quantities like density, momentum and energy. An example is the set of *hydrodynamic equations* describing the flow of macroscopic quantities in fluids and gases. The shape of the hydrodynamic equations is the same for all fluids and gases, but they contain certain *effective parameters* (like compressibility and viscosity) that depend on the fluid or the gas in question. These effective parameters are the “finger prints” of the specific microscopic interactions and dynamics in the system after the limit to macroscopic space-time scales has been taken. We will see in this section that a similar behavior occurs for the stepping stone model on the hierarchical lattice.

In Section 7.1 we define the model. In Section 7.2 we introduce the block averages on successive hierarchical space-time scales. In Section 7.3 we show that these block averages behave like autonomous diffusions in the limit as $M \rightarrow \infty$, with a diffusion function that depends on the scale. In Section 7.4 we analyze the renormalization transformation that links the diffusion functions on successive scales and analyze its orbit. It turns out that the WF-diffusion arises as the *universal attractor* of the renormalization transformation, a fact that is emphasized in Section 7.5.

7.1 Hierarchically interacting diffusions

The hierarchical lattice of order M is the set

$$\Omega_M = \left\{ x = (x_k)_{k \in \mathbb{N}} : x_k \in \{0, 1, \dots, M-1\}, \sum_{k \in \mathbb{N}} x_k < \infty \right\}. \quad (7.1)$$

Think of x as the *genetic address* of colony x :

x_1 is the house, x_2 is the street, x_3 is the town, x_4 is the province, x_5 is the country, etc.

(See Sawyer and Felsenstein [20] for the genetic background of this choice.) The restriction $\sum_{k \in \mathbb{N}} x_k < \infty$ (i.e., all x end with only zeroes) makes Ω_M countable. With componentwise addition modulo M , Ω_M becomes a group.

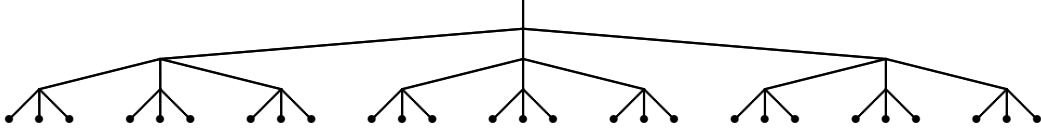


Figure 15: The hierarchical lattice Ω_3 (dots) and its hierarchical distance structure (lines).

On Ω_M there is a natural distance, called the *hierarchical distance*:

$$\|x - y\| = \min\{k \in \mathbb{N}_0 : x_l = y_l \ \forall l \geq k\}. \quad (7.2)$$

In words, the height of the first common ancestor of x and y in the tree structure given in Fig. 15 represents the distance between two points. The hierarchical distance is in fact an *ultrametric*:

$$\|x - y\| \leq \max\{\|x - z\|, \|z - y\|\} \quad \forall x, y, z \in \Omega_M. \quad (7.3)$$

In particular, all elements in any two disjoint balls around any two sites have the same distance to each other.

Exercise 7.1 *Explain why the last two statements are true.*

We will be interested in the following *system of coupled stochastic differential equations*:

$$dY_x(t) = c \sum_{y \in \Omega_M} r(x, y) [Y_y(t) - Y_x(t)] dt + \sqrt{g(Y_x(t))} dW_x(t), \quad x \in \Omega_M, t \geq 0, \quad (7.4)$$

where:

- (1) $c > 0$ is a constant;

- (2) $r: \Omega_N \times \Omega_N$ is a non-degenerate transition rate kernel;
- (3) $g: [0, 1] \rightarrow [0, \infty)$ is a non-degenerate diffusion function;
- (4) $\{W_x\}_{x \in \Omega_M}$ is an i.i.d. collection of standard Brownian motions.

As initial condition we take

$$Y_x(0) = \theta \in (0, 1) \quad \forall x \in \Omega_M. \quad (7.5)$$

The system in (7.4) models a collection of interacting Wright-Fisher diffusions with migration rate c , migration kernel $r(x, y)$, and diffusion function g (recall (2.37)). It is a continuous space-time version of the stepping stone model studied in Section 3.1, with c and $r(x, y)$ taking over the role of ν and $q(y - x)$ in (6.2). There is *no mutation*, and the diffusion function g is more general than the diffusion function g^* , given by $g^*(x) = x(1 - x)$ encountered for the Wright-Fisher diffusion in Section 2.1.2. Note that the first term captures the *interaction* between the colonies, while the second term is *autonomous* for each colony.

For each $x \in \Omega_M$, $Y_x(t)$ denotes the fraction of individuals of type A in colony x at time t (in the scaling limit given by (2.14–2.16)). The colonies evolve via *resampling*, described by the second term in (7.4), and interact via *migration*, described by the first term in (7.4). In order for the second term to make sense, we need to place some restrictions on g . In what follows we will assume that:

$$\begin{aligned} (i) \quad & g(0) = g(1) = 0; \\ (ii) \quad & g(u) > 0 \quad \forall u \in (0, 1); \\ (iii) \quad & g \text{ is Lipschitz on } [0, 1]. \end{aligned} \quad (7.6)$$

Under these restrictions, (7.4) has a *unique strong solution*, i.e., $\{Y_x(t): x \in \Omega_M\}$ is measurable w.r.t. the sigma-algebra generated by $\{W_x(s): x \in \Omega_M, 0 \leq s \leq t\}$ for all $t \geq 0$. The *class* of functions satisfying (7.6) will be denoted by $\mathcal{H}$.

In what follows we will make a *special choice for the migration kernel* (just as we did in (6.12) for the stepping stone model on $\mathbb{Z}^2$), namely,

$$r(x, y) = \frac{1}{N_M} \sum_{k \geq \|x-y\|} \frac{1}{M^{2k-1}}, \quad (7.7)$$

where N_M is the normalizing constant. This choice, which will turn out to be particularly *well adapted to the hierarchical structure* of Ω_M , amounts to a migration mechanism in which an individual, for each $k \in \mathbb{N}$, with a probability proportional to $1/M^{k-1}$ chooses “space horizon” k , and randomly chooses a colony from the k -block around its own colony.

Exercise 7.2 Show that $N_M = M^2/(M^2 - 1)$.

7.2 Hierarchy of space-time scales

We will look at the system along an *increasing sequence of space-time scales*. To that end, we define

$$Y_x^{[k]}(t) = \frac{1}{M^k} \sum_{\substack{y \in \Omega_M \\ \|y-x\| \leq k}} Y_y(M^k t), \quad x \in \Omega_M, k \in \mathbb{N}. \quad (7.8)$$

This is the average of the components taken over a block of radius k centered at x , with time speeded up proportional to the volume of the block. Note that $M^k = |\{y \in \Omega_M : \|y-x\| \leq k\}|$ for all $x \in \Omega_M$. We will refer to $Y_x^{[k]}(\cdot)$ as the *block average around x on space-time scale k* (see Fig. 16).

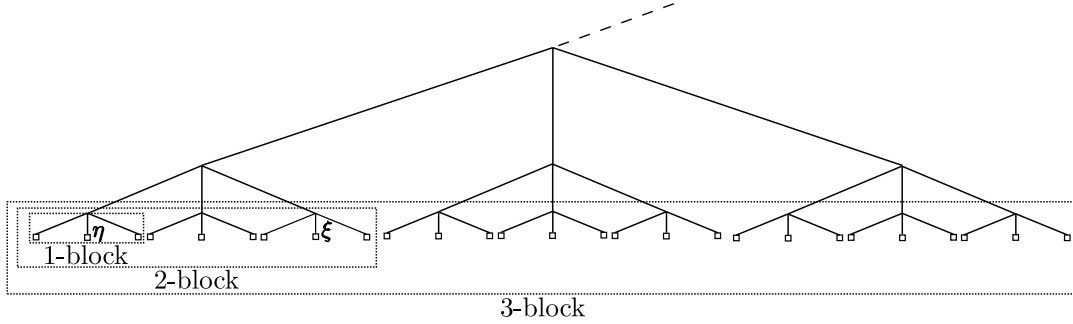


Figure 16: Picture of a k -block around the site $\eta \in \Omega_3$ for $k = 1, 2, 3$. The site $\xi \in \Omega_3$ falls in the k -block for $k \geq 2$, but not for $k = 1$.

The block averages themselves satisfy a system of stochastic differential equations:

$$\begin{aligned} dY_x^{[k]}(t) = & c \sum_{l \in \mathbb{N}} \frac{1}{M^{l-1}} [Y_x^{[k+l]}(M^{-l}t) - Y_x^{[k]}(t)] dt \\ & + \sqrt{\frac{1}{M^k} \sum_{\substack{y \in \Omega_M \\ \|y-x\| \leq k}} g(Y_y(M^k t))} dW_x^{[k]}(t), \quad x \in \Omega_M, t \geq 0, k \in \mathbb{N}, \end{aligned} \quad (7.9)$$

where $\{W_x^{[k]}\}_{x \in \Omega}$ are i.i.d. standard Brownian motions for every $k \in \mathbb{N}$.

Exercise 7.1 Check (7.9). *Hints: To find the term multiplying $dW_x^{[k]}(t)$, use the scaling properties of the independent Brownian motions involved, namely, $aW_x(t) \stackrel{\text{dis}}{=} dW_x(a^2 t)$ and $\sqrt{a}W + \sqrt{b}W' \stackrel{\text{dis}}{=} \sqrt{a+b}W''$. To find the drift term:*

- (1) Write Ω_M as a sum of points that are equally distant from the base point x of (7.4).
- (2) Invert the order of summation, to arrive at an expression that depends on the block averages around x .

- (3) Split the sum into terms that are at distance $\leq k$ from x and terms that are at distance $> k$ of x (use the ultrametric property).
- (4) Note that the first part is zero (by symmetry).
- (5) Change the summation indices in the second part to find the desired expression.

7.3 Hierarchical mean-field limit

In this section we look at the successive block averages and identify their behaviour in the limit as $M \rightarrow \infty$.

$k = 0$: In this case (7.9) reads

$$dY_x^{[0]}(t) = c \sum_{l \in \mathbb{N}} \frac{1}{M^{l-1}} [Y_x^{[l]}(M^{-l}t) - Y_x^{[0]}(t)] dt + \sqrt{g(Y_x^{[0]}(t))} dW_x^{[0]}(t). \quad (7.10)$$

As $M \rightarrow \infty$, only the term with $l = 1$ survives. Moreover, $w - \lim_{M \rightarrow \infty} Y_x^{[1]}(M^{-1}t) = Y_x^{[1]}(0) = \theta$ for all $t \geq 0$, by (7.5). Therefore we obtain that

$$w - \lim_{M \rightarrow \infty} Y_x^{[0]}(t) = Z^{[0]}(t) \quad (7.11)$$

with $Z^{[0]}(t)$ the solution of the stochastic differential equation

$$dZ(t) = c[\theta - Z(t)] dt + \sqrt{g(Z(t))} dW(t). \quad (7.12)$$

In other words, in the limit as $M \rightarrow \infty$ the components *decouple* and each component performs an *autonomous diffusion* with migration rate c , drift towards θ and diffusion function g .

$k = 1$: As $M \rightarrow \infty$, again only the term with $l = 1$ in (7.9) survives. Moreover, $w - \lim_{M \rightarrow \infty} Y_x^{[2]}(M^{-1}t) = Y_x^{[2]}(0) = \theta$ for all $t \geq 0$, by (7.5). Furthermore, for fixed t the family

$$\{Y_y(Mt)\}_{\substack{y \in \Omega_M : \\ \|y-x\| \leq 1}} \quad (7.13)$$

decouples and each member *converges almost instantly to the equilibrium distribution associated with* (7.12), with the drift towards θ replaced by a drift towards $Y_x^{[1]}(t)$, the *instantaneous* value of the 1-block average. Thus, as $M \rightarrow \infty$,

$$w - \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{\substack{y \in \Omega_M : \\ \|y-x\| \leq 1}} g(Y_y(Mt)) = (Fg)(Y^{[1]}(t)), \quad (7.14)$$

where

$$(Fg)(v) = \int_0^1 g(u) \nu_{c,v,g}(du) \quad (7.15)$$

with $\nu_{c,v,g}$ the *equilibrium distribution* associated with the diffusion

$$dZ(t) = c[v - Z(t)] dt + \sqrt{g(Z(t))} dW(t), \quad (7.16)$$

which is the same as (7.12) but with θ replaced by v . Therefore we conclude that

$$w - \lim_{M \rightarrow \infty} Y_x^{[1]}(t) = Z^{[1]}(t) \quad (7.17)$$

with $Z^{[1]}(t)$ the solution of the stochastic differential equation

$$dZ(t) = c[\theta - Z(t)] dt + \sqrt{(Fg)(Z(t))} dW(t). \quad (7.18)$$

Note that this is again an *autonomous diffusion*, but with a *different* diffusion function than in (7.12), namely, Fg instead of g . In Section 7.4 we will give an explicit formula for the map F defined by (7.15–7.16).

$k \geq 2$: By moving up further in the hierarchy we find, iterating the above argument, that

$$w - \lim_{M \rightarrow \infty} Y_x^{[k]}(t) = Z^{[k]}(t) \quad (7.19)$$

with $Z^{[k]}(t)$ the solution of the stochastic differential equation

$$dZ(t) = c[\theta - Z(t)] dt + \sqrt{(F^k g)(Z(t))} dW(t) \quad (7.20)$$

with diffusion function $F^k g$, the k -th iterate of F applied to g .

Theorem 7.1 *The above convergence scheme holds for all $c > 0$, $\theta \in (0, 1)$ and $g \in \mathcal{H}$, the class defined by (7.6).*

Proof. The proof is beyond the scope of this course. We refer to Dawson and Greven [8]. ■

The limit $M \rightarrow \infty$ is referred to as the *hierarchical mean-field limit*. In this limit each colony has a large number of neighbours with which it interacts equally strongly. For each k , the k -block feels a drift towards the value of the $(k+1)$ -block around it and equilibrates fast w.r.t. the slower motion of that block, reaching what is called a *quasi-equilibrium*. The diffusion function of the $(k+1)$ -block is the average of the diffusion function of the constituent k -blocks w.r.t. that quasi-equilibrium (see Fig. 17).

7.4 Renormalisation transformation

We next move on to studying the renormalisation transformation F in detail. To simplify the calculations somewhat, we henceforth pick $c = 1$.

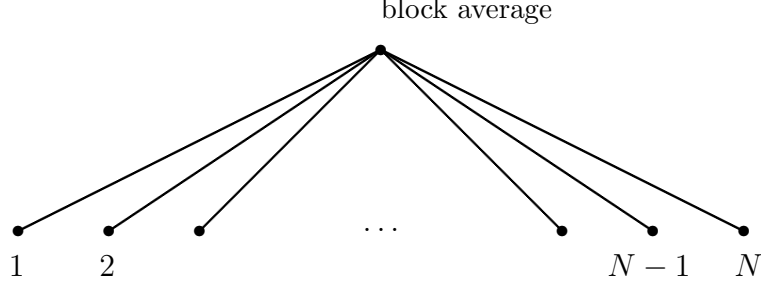


Figure 17: Given the value of a block average, the $N \gg 1$ constituent components equilibrate on a time scale that is fast w.r.t. the time scale on which the block average fluctuates. Consequently, the volatility of the block average is the expectation of the volatility of the constituent components under the conditional quasi-equilibrium.

Theorem 7.1 *For every $v \in (0, 1)$ and $g \in \mathcal{H}$, the stochastic differential equation in (7.16) has a unique ergodic equilibrium $\nu_{v,g}$ given by*

$$\nu_{v,g}(du) = \frac{1}{N_{v,g}} \mu_{v,g}(u) du, \quad u \in [0, 1], \quad (7.21)$$

where

$$\mu_{v,g}(u) = \frac{1}{g(u)} \exp \left[- \int_u^v \frac{w-v}{g(w)} dw \right], \quad \mu_{v,g}(0) = \mu_{v,g}(1) = 0, \quad (7.22)$$

and $N_{v,g}$ is the normalizing constant.

Proof. The equilibrium $\nu_{v,g}$ is defined by the requirement that

$$\int_0^1 (\mathcal{L}_{v,g} f)(u) \nu_{v,g}(du) = 0 \quad \forall f \in \mathcal{D}(\mathcal{L}_{v,g}), \quad (7.23)$$

with

$$\mathcal{L}_{v,g} = (v-u) \frac{\partial}{\partial u} + g(u) \frac{\partial^2}{\partial u^2} \quad (7.24)$$

the generator of the diffusion in (7.16) and $\mathcal{D}(\mathcal{L}_{v,g})$ its domain. The latter domain is dense in the set of continuous functions on $[0, 1]$. After inserting (7.21) into (7.23), ignoring the normalising constant and partially integrating the resulting equation (the first term once and the second term twice), we get

$$\int_0^1 f(u) \left[- \frac{\partial}{\partial u} \{ (v-u) \mu_{v,g}(u) \} + \frac{\partial^2}{\partial u^2} \{ g(u) \mu_{v,g}(u) \} \right] du = 0 \quad \forall f \in \mathcal{D}(\mathcal{L}_{v,g}). \quad (7.25)$$

The boundary terms that arise from the partial integration vanish because $\nu_{v,g}$ puts no mass at 0 and 1. We conclude that the term between square brackets must be zero. Removing one differentiation, we end up with the equation

$$-(v-u) \mu_{v,g}(u) + \frac{\partial}{\partial u} \{ g(u) \mu_{v,g}(u) \} = C_{v,g}, \quad (7.26)$$

where $C_{v,g}$ is some integration constant. Solving (7.26) we find (7.22), after we absorb $C_{v,g}$ into the normalising constant $N_{g,v}$. ■

Exercise 7.2 *Check the above computations. For the fine details, in particular, issues of integrability near the boundaries of $[0, 1]$, see Baillon et al. [1].*

In view of Theorem 7.1, F takes on the form

$$(Fg)(v) = \frac{\int_0^1 g(u) \mu_{v,g}(u) du}{\int_0^1 \mu_{v,g}(u) du}, \quad v \in (0, 1). \quad (7.27)$$

It can further be shown that, for all $g \in \mathcal{H}$,

$$w - \lim_{v \downarrow 0} \nu_{v,g} = \delta_0, \quad w - \lim_{v \uparrow 1} \nu_{v,g} = \delta_1, \quad (7.28)$$

so that

$$(Fg)(0) = g(0) = 0, \quad (Fg)(1) = g(1) = 0. \quad (7.29)$$

Note that F in (7.27) is a *non-linear integral transform* acting on the class $\mathcal{H}$. Now that we have an explicit form for F , we can study its *iterations*.

Theorem 7.3 $F\mathcal{H} \subset \mathcal{H}$.

Proof. We must show that if $g \in \mathcal{H}$, then also $Fg \in \mathcal{H}$. It is clear from (7.29) that Fg satisfies condition (i) in (7.6). From (7.21–7.22) we see that, for all $v \in (0, 1)$, $\nu_{v,g}$ puts mass everywhere in $(0, 1)$. Therefore Fg satisfies condition (ii) in (7.6). Finally, it can be shown (see Baillon et al. [1]) that F lowers the Lipschitz constant, i.e., if

$$L[g] \stackrel{\text{def}}{=} \sup_{\substack{a, b \in [0, 1]: \\ a \neq b}} \left| \frac{g(a) - g(b)}{a - b} \right|, \quad (7.30)$$

then

$$L[Fg] \leq L[g] \quad \forall g \in \mathcal{H}. \quad (7.31)$$

Hence, also condition (iii) in (7.6) carries over. ■

According to Theorem 7.3, we can iterate F indefinitely and study its *orbit*

$$\{F^k g\}_{k \in \mathbb{N}_0}. \quad (7.32)$$

Our key result is the following *universal scaling behavior*:

Theorem 7.4 *For all $g \in \mathcal{H}$,*

$$\lim_{k \rightarrow \infty} k F^k g = g^* \quad \text{uniformly on } [0, 1], \quad (7.33)$$

with

$$g^*(u) = u(1 - u) \quad (7.34)$$

the Wright-Fisher diffusion function.

Proof. The proof uses the iterates of an associated sequence of Markov kernels. Fix $g \in \mathcal{H}$. Let $(\circ)$ denotes composition

$$\begin{aligned} K_g(v, du) &= \nu_{v,g}(du), \\ K_g^{[k]} &= K_{F^{k-1}g} \circ K_{F^{k-2}g} \circ \cdots \circ K_g, \quad k \in \mathbb{N}. \end{aligned} \quad (7.35)$$

Our equilibrium $\nu_{v,g}$ satisfies four relations:

$$\begin{aligned} (a) \quad & \int_0^1 \nu_{v,g}(du) = 1, \\ (b) \quad & \int_0^1 u \nu_{v,g}(du) = v, \\ (c) \quad & \int_0^1 u^2 \nu_{v,g}(du) = v^2 + (Fg)(v), \\ (d) \quad & \int_0^1 g(u) \nu_{v,g}(du) = (Fg)(v). \end{aligned} \quad (7.36)$$

Relations (a) and (d) are trivial.

Exercise 7.5 Deduce (b) and (c) from (7.21).

Iteration of (7.36) leads to four relations for the kernel $K_g^{[k]}$:

$$\begin{aligned} (a') \quad & \int_0^1 K_g^{[k]}(v, du) = 1, \\ (b') \quad & \int_0^1 u K_g^{[k]}(v, du) = v, \\ (c') \quad & \int_0^1 u^2 K_g^{[k]}(v, du) = v^2 + k(F^k g)(v), \\ (d') \quad & \int_0^1 g(u) K_g^{[k]}(v, du) = (F^k g)(v). \end{aligned} \quad (7.37)$$

Exercise 7.6 Derive these relations.

Now comes a clever trick. Subtract (c') from (b'), to get

$$\int_0^1 u(1-u) K_g^{[k]}(v, du) = v(1-v) - k(F^k g)(v). \quad (7.38)$$

Since the left-hand side is ≥ 0 , we see that

$$kF^k g \leq g^* \text{ uniformly on } [0, 1] \text{ for all } k \in \mathbb{N}_0. \quad (7.39)$$

This is already half of (7.33). To get the other half, we proceed as follows. It follows from (7.39) that $\lim_{k \rightarrow \infty} F^k g = 0$ uniformly on $[0, 1]$. Hence, (b') and (d') give that

$$w - \lim_{k \rightarrow \infty} K_g^{[k]}(\cdot) = (1-v)\delta_0(\cdot) + v\delta_1(\cdot) \quad \forall v \in (0, 1), \quad (7.40)$$

where we use that g vanishes at 0 and 1 only. Inserting the above identity into (7.38), we arrive at

$$\lim_{k \rightarrow \infty} [v(1-v) - k(F^k g)(v)] = 0 \quad \forall v \in (0, 1). \quad (7.41)$$

Thus we have proved (7.33) with pointwise convergence. To get uniform convergence, we use that the family $\{kF^k g\}_{k \in \mathbb{N}_0}$ is uniformly equicontinuous on $\mathcal{H}$, a fact that can be deduced from the analogue of (7.31):

$$L[Kf] \leq L[f] \quad \forall f, g \in \mathcal{H}. \quad (7.42)$$

The proof of (7.31) is technical (see Baillon et al. [1]). ■

7.5 Conclusion

The beauty of Theorem 7.4 is that it expresses *full universality*: no matter what $g \in \mathcal{H}$ we pick as the diffusion function for the *single components* of our system (7.4), the diffusion function $F^k g$ for the k -blocks (defined by (7.8)) with k large are *close to a multiple of the Wright-Fisher diffusion function*, namely,

$$F^k g \sim \frac{1}{k} g^* \text{ as } k \rightarrow \infty, \quad (7.43)$$

In other words, Wright-Fisher is a *global attractor on large space-time scales*.

In retrospect, this universality justifies the fact that we dedicated an entire course on the Wright-Fisher model!

A Appendix

In this appendix we recall a few basic facts about special probability distributions (binomial, geometric, Poisson, exponential, beta, gamma) and elementary stochastic processes (discrete-time and continuous-time Markov chains, Poisson processes, random walk, Brownian motion).

A.1 Special probability distributions

I. Discrete:

Binomial distribution $\text{BIN}(n, p)$:

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \dots, n, \text{ with } n \in \mathbb{N}, p \in (0, 1),$$

$$\mathbb{E}(X) = np, \quad \mathbb{V}\mathbb{A}\mathbb{R}(X) = np(1-p).$$

Geometric distribution $\text{GEO}(p)$:

$$f(k) = p(1-p)^k, \quad k \in \mathbb{N}_0, \text{ with } p \in (0, 1),$$

$$\mathbb{E}(X) = (1-p)/p, \quad \mathbb{V}\mathbb{A}\mathbb{R}(X) = (1-p)/p^2.$$

Poisson distribution $\text{POISSON}(\lambda)$:

$$f(k) = e^{-\lambda} \lambda^k / k!, \quad k \in \mathbb{N}_0, \text{ with } \lambda \in (0, \infty),$$

$$\mathbb{E}(X) = \lambda, \quad \mathbb{V}\mathbb{A}\mathbb{R}(X) = \lambda.$$

II. Continuous:

Exponential distribution $\text{EXP}(\lambda)$:

$$f(x) = \lambda e^{-\lambda x}, \quad x \in [0, \infty), \text{ with } \lambda \in (0, \infty),$$

$$\mathbb{E}(X) = 1/\lambda, \quad \mathbb{V}\mathbb{A}\mathbb{R}(X) = 1/\lambda^2.$$

Beta distribution $\text{BETA}(\alpha, \beta)$:

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad x \in [0, 1], \text{ with } \alpha, \beta \in (0, \infty),$$

$$\mathbb{E}(X) = \alpha/(\alpha + \beta), \quad \mathbb{V}\mathbb{A}\mathbb{R}(X) = \alpha\beta/(\alpha + \beta)^2(\alpha + \beta + 1).$$

Gamma distribution $\text{GAMMA}(\alpha, \beta)$:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x \in [0, \infty), \text{ with } \alpha, \beta \in (0, \infty),$$

$$\mathbb{E}(X) = \alpha/\beta, \quad \mathbb{V}\text{AR}(X) = \alpha/\beta^2.$$

A.2 Elementary stochastic processes

A.2.1 Markov chains

A Markov process is a random process without memory: its future depends on its present, not on its past. We focus on time-homogeneous Markov processes, for which the transition probabilities or transition rates do not depend on time.

Markov processes in discrete time $X = (X_n)_{n \in \mathbb{N}_0}$ on a countable state space S are characterized by an initial distribution $\mu = (\mu_i)_{i \in S}$ and a transition matrix $P = (P_{ij})_{i,j \in S}$, with the interpretation

$$\mu_i = \mathbb{P}(X_0 = i), \quad P_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i). \quad (\text{A.1})$$

The distribution at time n is given by the formula

$$\mathbb{P}(X_n = j) = (\mu P^n)_j = \sum_{i \in S} \mu_i (P^n)_{ij}. \quad (\text{A.2})$$

Markov processes in continuous time $X = (X_t)_{t \geq 0}$ on a countable state space S are characterised by an initial distribution $\mu = (\mu_i)_{i \in S}$ and an infinitesimal generator $L = (L_{ij})_{i,j \in S}$, with the interpretation

$$\mu_i = \mathbb{P}(X_0 = i), \quad L_{ij} = \lim_{t \downarrow 0} \frac{1}{t} [\mathbb{P}(X_t = j \mid X_0 = i) - \delta_{ij}]. \quad (\text{A.3})$$

Note that $L_{ij} \geq 0$ for $i \neq j$ and $\sum_{j \in S} L_{ij} = 0$: L_{ij} , $i \neq j$, has the interpretation of the rate to jump from i to j , while L_{ii} is minus the rate to jump away from i . The distribution at time n is given by the formula

$$\mathbb{P}(X_t = j) = (P_t f)_j = \sum_{i \in S} \mu_i (P_t)_{ij}, \quad (\text{A.4})$$

where $P_t = e^{tL}$ defines a semigroup $(P_t)_{t \geq 0}$.

When S is uncountable, sums are replaced by integrals, and some extra care is needed with the definitions of P and L , P_t , which become operators acting on appropriate sets of test functions f , e.g.

$$(Lf)(y) = \left[\frac{d}{dt} \mathbb{E}(f(X_t) \mid X_0 = y) \right] \Big|_{t=0}, \quad (P_t f)(y) = \mathbb{E}(f(X_t) \mid X_0 = y). \quad (\text{A.5})$$

The above implies that we can write the Dynkin formula (using the convention that $\mathbb{E}^y[f(X_t)] = \mathbb{E}[f(X_t) \mid X_0 = y]$):

$$\mathbb{E}^y[f(X_t)] = f(y) + \mathbb{E}^y \left[\int_0^t (Lf)X_s ds \right]. \quad (\text{A.6})$$

Equivalently, in terms of the semigroup $(S(t))_{t \geq 0}$ corresponding to the Markov process, given by $\mathbb{E}^y[f(X_t)] = (S(t)f)(y)$, we get

$$(S(t)f)(y) = f(y) + \int_0^t ds S(s)(Lf)(y). \quad (\text{A.7})$$

We can rewrite the above relation in a path-wise expression by defining a residue term:

$$M_t^f = f(X_t) - f(X_0) - \int_0^t ds (Lf)(X_s), \quad (\text{A.8})$$

where we note that $(M_t^f)_{t \geq 0}$ is a martingale with respect to $\mathbb{P}^y$, with $(\mathcal{F}_t)_{t \geq 0}$ the relevant filtration for every initial condition y . To see why, we compute for $t = u + s$

$$\begin{aligned} \mathbb{E}^y[M_t^f - M_s^f \mid \mathcal{F}_s] &= \mathbb{E}^y[f(Y_t) - f(Y_s) - \int_s^t (Lf)(X_r) dr \mid \mathcal{F}_s] \\ &= S(t-s)f(X_s) - f(X_s) - \int_s^t S(r-s)(Lf)(X_s) dr \\ &= S(u)f(X_s) - f(X_s) - \int_0^u S(r)(Lf)(X_s) dr = \varphi(u). \end{aligned} \quad (\text{A.9})$$

To conclude, it suffices to prove that $\varphi(u) = 0$ for every $u \geq 0$. To do so, note first that

$$\varphi(0) = 0 \quad (\text{A.10})$$

and

$$\begin{aligned} \varphi'(u) &= \lim_{h \rightarrow 0} \frac{1}{h} [(S(u+h)f)(X_s) - (S(u)f)(X_s) - \int_u^{u+h} S(r)(Lf)(X_s) dr] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} [(S(h)(S(u)f))(X_s) - (S(u)f)(X_s) - S(u)(Lf)(X_s)] \\ &= S(u)(Lf)(X_s) - S(u)(Lf)(X_s) = 0. \end{aligned} \quad (\text{A.11})$$

To understand the last passage, note that we can rewrite (A.5) as

$$\lim_{h \rightarrow 0} \frac{1}{h} [(S(h)g)(y) - g(y)] = (Lg)(y). \quad (\text{A.12})$$

So taking $g(y) = (S(u)f)(y)$, we see that

$$\lim_{h \rightarrow 0} \frac{1}{h} [(S(h)S(u)f)(X_s) - (S(u)f)(X_s)] = (LS(u)f)(X_s) = (S(u)Lf)(X_s). \quad (\text{A.13})$$

(See Ethier and Kurtz [11, Proposition 1.5].) This proves that $\varphi(u) = 0$ for every u , which implies that

$$\mathbb{E}^y[M_t^f - M_s^f \mid \mathcal{F}_s] = 0 \quad (\text{A.14})$$

A.2.2 SDE and Martingale Problem

There is a correspondence between an SDE and a Martingale Problem. The rule of thumb is the following. If $X = (X_t)_{t \geq 0}$ is a solution to the SDE

$$dX_t = b(X_t) dt + \sigma(X_t) dB_t \quad (\text{A.15})$$

with $B = (B_t)_{t \geq 0}$ a Brownian motion, then X is a solution to the martingale problem with generator L given by

$$(Lf)(y) = b(y)f'(y) + \frac{1}{2}\sigma^2(y)f''(y). \quad (\text{A.16})$$

The converse is true also: If we have a solution of to martingale problem, then we can construct a probability space and a random variable X such that

$$dX_t = b(X_t) dt + \sigma(X_t) dB_t. \quad (\text{A.17})$$

For more details on the correspondence between the coefficients of an SDE and the generator L , see Karatzas and Sheve [16, Propositions 4.2 and 4.6]

A.2.3 Poisson process

The Poisson process with rate $\lambda \in (0, \infty)$ is the continuous-time Markov process $N = (N_t)_{t \geq 0}$ on state space $\mathbb{N}_0$, starting at $N_0 = 0$ and making unit increments at random times $(T_k)_{k \in \mathbb{N}}$, with $0 = T_0 < T_1 < T_2 < \dots$, such that $(T_{k+1} - T_k)$, $k \in \mathbb{N}_0$, are i.i.d. with distribution $\text{EXP}(\lambda)$. These random times form a random subset of $[0, \infty)$, which is referred to as the Poisson point process with intensity λ . This random subset has the property that for any collection of disjoint subsets $U_l \subset [0, \infty)$, $l \in I$, with I some finite index set, the numbers $|\{k \in \mathbb{N}: T_k \in U_l\}|$, $l \in I$, are independent and have marginal distributions $\text{POISSON}(|U_l|)$, $l \in I$.

A.2.4 Random walk

A random walk on $\mathbb{Z}^d$, $d \geq 1$, in discrete time is any Markov process $X = (X_n)_{n \in \mathbb{N}_0}$ on state space $\mathbb{Z}^d$ whose transition matrix is given by

$$P_{ij} = p_{j-i}, \quad (\text{A.18})$$

where $i \mapsto p_i$ is some probability distribution on $\mathbb{Z}^d$, i.e., the transition probabilities are invariant under shifts. Due to this property, X has the representation

$$X_n = X_0 + (Y_1 + \dots + Y_n), \quad (\text{A.19})$$

where $(Y_k)_{k \in \mathbb{N}}$ are i.i.d. with $\mathbb{P}(Y_k = i) = p_i$.

The same formulas hold in continuous time with $L_{ij} = l_{j-i}$ replacing (A.18), where $i \mapsto l_i$ is any function satisfying $l_i \geq 0$, $i \neq 0$, and $\sum_{i \in \mathbb{Z}^d} l_i = 0$. The same definitions apply when $\mathbb{Z}^d$ is replaced by any set that is a group under the operation $+$.

A.2.5 Brownian motion

Brownian motion $W = (W_t)_{t \geq 0}$ is the continuous-time Markov process on $\mathbb{R}$ characterized by four properties: (1) its increments over disjoint time intervals are independent; (2) its increments over time intervals of length s have distribution $N(0, s)$, the normal distribution with mean 0 and variance s ; (3) it has continuous paths, i.e., $t \mapsto W_t$ is continuous; (4) $W_0 = 0$.

Brownian motion arises as the space-time scaling limit of simple random walk $S = (S_n)_{n \in \mathbb{N}_0}$ on $\mathbb{Z}$, the random walk whose i.i.d. increments are ± 1 with probability $\frac{1}{2}$ each. Namely,

$$w - \lim_{n \rightarrow \infty} \left(\frac{S_{[nt]}}{\sqrt{n}} \right)_{t \geq 0} = (W_t)_{t \geq 0}, \quad (\text{A.20})$$

where $w - \lim$ denotes weak convergence on path space.

Brownian motion has generator $(Lf)(y) = \frac{1}{2}f''(y)$, i.e.,

$$\mathbb{E}(f(W_t)) - \mathbb{E}(f(W_0)) = \mathbb{E} \left(\int_0^t \frac{1}{2} f''(W_s) ds \right), \quad (\text{A.21})$$

where $f: \mathbb{R} \rightarrow \mathbb{R}$ is any test function that is twice continuously differentiable.

Diffusion processes are obtained from Brownian motions by changing their local characteristics: add a *local drift* and add a *local volatility*. Durrett [10], Chapter 7, contains a quick introduction to diffusion processes. The WF-diffusion has generator $(Lf)(y) = y(1-y)\frac{1}{2}f''(y)$, i.e.,

$$\mathbb{E}(f(Y_t)) - \mathbb{E}(f(Y_0)) = \mathbb{E} \left(\int_0^t Y_s(1-Y_s) \frac{1}{2} f''(Y_s) ds \right). \quad (\text{A.22})$$

The analogue of (A.22) for the Brownian motion with local diffusion function $g: \mathbb{R} \rightarrow [0, \infty)$ reads

$$\mathbb{E}(f(Y_t)) - \mathbb{E}(f(Y_0)) = \mathbb{E} \left(\int_0^t g(Y_s) \frac{1}{2} f''(Y_s) ds \right). \quad (\text{A.23})$$

B MRCA-process and F-process

B.1 Look-down construction

To investigate $(A_t)_{t \in \mathbb{R}}$ and $(F_t)_{t \in \mathbb{R}}$, we use Pfaffelhuber and Wakolbinger [19] and the “particle construction” of Donnelly and Kurtz [9], called look-down construction, which we will now describe (see Fig. 18).

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

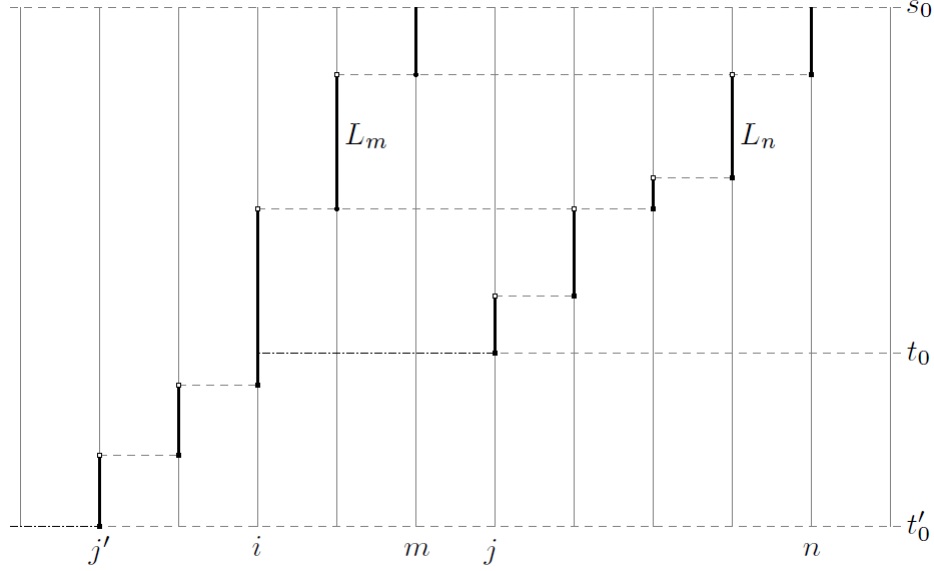


Figure 18: Part of a look-down graph. Biological time is running upwards and level numbers are running to the right. Only (parts of) two lines L_m and L_n are drawn. At time t_0 level j looks down to level i and individual (t_0, j) is born. Each time an individual in line L_m is pushed one level up, 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 .

to level i , the individuals at levels $j, j+1, \dots$ 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 $t_1 < t_2 < \dots$, where t_n is a point of P_{ik} for some $1 \leq i < k < j+n$, 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

$$L = ([t_0, t_1) \times \{j\}) \cup ([t_1, t_2) \times \{j+1\}) \cup ([t_2, t_3) \times \{j+2\}) \cup \dots \subset \mathbb{R} \times \mathbb{N}. \quad (\text{B.1})$$

See Fig. 18 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 $1 \leq i < k \leq j$. This leaves $j(j-1)/2$ possible values for i and k , and therefore pushing rates increase quadratically in j . Hence, $t_\infty = \lim_{n \rightarrow \infty} t_n$ is finite with probability 1. We say that 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 (B.1) such that $(s_0, n) \in L_n$, and a function $\Lambda_{L_n}: [t_0, t_\infty) \rightarrow \{j, j+1, \dots\}$, $t \mapsto i$ 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 $I_n = [0, s_0 - t_0]$, $I_m = (s_0 - t_0, s_0 - t'_0]$, $\dots$ such that $\cup_i I_i = [0, \infty)$. (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

$$\Phi_n(t): [0, \infty) \rightarrow \mathbb{N}, \quad t \mapsto X_t^i \quad \text{if } t \in I_i. \quad (\text{B.2})$$

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.,

$$\forall n \in \mathbb{N} \exists t_n^* \forall t > t_n^*: \Phi_n(t) = 1. \quad (\text{B.3})$$

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) , $n \in \mathbb{N}$, for some reference time $s = s_0$ (see Fig. 18).

B.2 Look-down 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 Section 3.1). 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.

B.2.1 Identification of the MRCA-process

In this section we prove Theorem 3.2.

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 3.2 implies that each point t_0 of P_{12} initiates a line L_F^i of the form (B.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) , $k = 3, 4, \dots$, 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) , $k = 1, \dots, n$, 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

$$\Theta_n(t) = \begin{cases} 1, & \text{if } A_t^n > A_t^{n+1}, \\ 0, & \text{if } A_t^n = A_t^{n+1}. \end{cases} \quad (\text{B.4})$$

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 [9], Lemma 3.5). We will use the particle construction to show that this implies the following theorem.

B.2.2 Identification of the F-process

In this section we prove Theorem 3.3.

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 $\beta_i = \lim_{n \rightarrow \infty} \beta_i^n$, 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 3.1: 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 3.1 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 $k \in \mathbb{N} \setminus \{1\}$ 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 3.1 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] J.-B. Baillon, Ph. Clément, A. Greven and F. den Hollander, On the attracting orbit of a non-linear transformation arising from renormalization of hierarchically interacting diffusions, Part I: The compact case, *Can. J. Math.* 47 (1995) 3–27.
- [2] N. Berestycki, Recent progress in coalescent theory, *Ens. Mat.* 16 (2009) 1–193.
- [3] J. Bertoin, *Lévy Processes*, Cambridge Tracts in Mathematics, Vol. 121, Cambridge University Press, Cambridge, 1996.
- [4] M.F.J. Carsouw, Wright-Fisher evolution, Bachelor thesis, July 2012, Mathematical Institute, Leiden University.
- [5] J.T. Cox, Intermediate range migration in the two-dimensional stepping stone model, *Ann. Appl. Prob.* 20 (2010) 785–805.
- [6] J.T. Cox and R. Durrett, The stepping stone model: new formulas expose old myths, *Ann. Appl. Prob.* 12 (2002) 1348–1377.
- [7] J.T. Cox, R. Durrett and I. Zähle, The stepping stone model. II: Genealogies of the infinite sites model, *Ann. Appl. Prob.* 15 (2005) 671–699.
- [8] D.A. Dawson and A. Greven, Multiple scale analysis of interacting diffusions, *Probab. Theory Relat. Fields* 95 (1993) 467–508.
- [9] P. Donnelly and T.G. Kurtz, Particle representations for measure-valued population models, *Ann. Probab.* 27 (1999) 166–205.
- [10] R. Durrett, *Probability Models for DNA Sequence Evolution* (2nd. ed.), Springer, New York, 2008.
- [11] S.N. Ethier and T.G. Kurtz, *Markov Processes; characterization and convergence*, John Wiley & Sons, New York, 1986.
- [12] W. Ewens, *Mathematical Population Genetics* (2nd. ed.), Springer, Berlin, 2005.
- [13] R.C. Griffiths and S. Lessard, Ewens’ sampling formula and related formulae: Combinatorial proofs, extensions to variable population size and applications to ages of alleles, *Theor. Popul. Biol.* 68 (2005) 167–177.
- [14] D. Hartl and A.G. Clark, *Principles of Populations Genetics*, (4th ed.), Sinauer Associates Inc., Sunderland MA, USA, 2007.
- [15] J. Hein, M.H. Schierup and C. Wiuf, *Gene Genealogies, Variation and Evolution*, Oxford University Press, Oxford, 2005.
- [16] I. Karatzas and S. Shreve *Brownian motion and stochastic calculus* (2nd. ed.), Graduate texts in mathematics, Springer-Verlag, 1988.

- [17] J.F.C. Kingman, On the genealogy of large polulations, J. Appl. Prob. 19 (1982) 27–43.
- [18] J.F.C. Kingman, The coalescent, Stoch. Proc. Appl. 13 (1982) 235–248.
- [19] P. Pfaffelhuber and A. Wakolbinger, The process of most recent common ancestors in an evolving coalescent, Stoch. Proc. Appl. 116 (2006) 1836–1859.
- [20] S. Sawyer and J. Felsenstein, Isolation by distance in a hierarchically clustered population, J. Appl. Probab. 20 (1983) 1–10.
- [21] J. Wakeley, *Coalescent Theory*, Ben Roberts Publications, Oxford, 2009.