The effect of latency in population genetics

vorgelegt von
Adrián González Casanova Soberón
geb. Cuernavaca, Mexiko

Von der Fakultät II - Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
Dr.rer.nat.

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. Volker Mehrmann
Berichter/Gutachter: Prof. Dr. Jochen Blath
Berichter/Gutachter: Prof. Dr. Noemi Kurt
Berichter/Gutachter: Prof. Dr. Anton Wakolbinger

Tag der wissenschaftlichen Aussprache: 9. October 2015

Berlin 2016

To the memory of those who died trying to cross the wall
between the USA and México

Acknowledgments

I can not imagine better PhD supervisors than Prof. Noemi Kurt and Prof. Jochen Blath. They were integral guides in this period of my life. I will always be grateful, not only for their time, their mathematical ideas and their hard work, but also for their patience, kindness and for always believing in me.

I also want to warmly thank three Professors from whom I have learned a lot, and with whom I had the opportunity to make research stays during my PhD: Prof. Julien Berestycki, Prof. Dario Spanò and Prof. Anton Wakolbinger. I am also very grateful with Prof. Anton Wakolbinger for being the external reviewer of this thesis.

Special thanks go to my coauthors, with whom I shared and enjoyed (and suffered) mathematics. Thank you Dr. Eneas Aguirre, Prof. Jochen Blath, Prof. Guadalupe Espín, Dr. Bjarki Eldon, Prof. Noemi Kurt, Prof. Luis Servin-Gonzalez, Prof. Gloria Soberón, Prof. Dario Spanò, Prof. Anton Wakolbinger, Maite Wilke-Berenguer, Dr. Linglong Yuan

I would like to thank Prof. Volker Mehrmann for agreeing to be the chairman in the defense of this thesis. I am grateful as well with Prof. Sylvie Roelly, my BMS Mentor, for all her support and with Prof. Dr. Wolfgang König, who gave me the opportunity to do a postdoc in his group, and allowed me to use some postdoc-time to finish writing my thesis.

Transatlantic support was received in big quantities. In particular, Prof. Maria Emilia Caballero, Prof. Juan Carlos Pardo and Prof. Jose Luis Pérez had been always available to hear my adventures and to generously advise me.

I am grateful with the following people for reading parts of my thesis and giving helpful comments: Eugenio Buzzoni, Alberto Chiarini, Giovanni Conforti, Matti Leimbach, Veronica Miro-Pina, Diana Nuñes and Dr. Pablo Soberón.

I was very lucky to share the office with 3 of my favorite Mathematicians: Alberto Channini, Giovanni Conforti and Atul Shekhar... it was like having (inside my head) a book of stochastic calculus opened in the right page, a good comedy movie and the *Radiotelevisione Italiana* (simultaneously and always). Without them or without the members of the almost-that-cool-office (Benedikt, Giuseppe and Matti), I wouldn't have enjoyed this so much.

I also want to thank Iliana who shared with me this period, and helped me to go through many difficult moments. We did it!

I want to thank my friends in both sides of the Atlantic Ocean and my family, specially my mother (who is also my co-author and invariably believes in me), my sister (who is always there for me) and my father and my grandfather (for inspiring me). Finally, I want to thank Ximena, who gave me the strength needed to write up this thesis.

To all of you I will always be grateful.

I acknowledge financial support from the DFG Research Training Group 1845 (scholarship) and from the Mexican Council of science (CONACyT) in cooperation with the DAAD (complementary support). I also had support to participate in conferences from the Berlin Mathematical School (BMS) and from the DFG priority program Probabilistic Structures in Evolution (DGF-SPP 1590).

Contents

Table 1: Notation: Chapter 1, Part 1

Symbol	Description	Page
N	Population size: number of individuals per generation	6
i, k	Discrete units of time	6
t	Continuous units of time	6
n, m	Sample size: number of individuals in a sample	6
$\{U^{(w)}(k)\}$	Family of independent uniform RV in $\{1, ..., N\}$	6
V_N	Set of vertexes of a Wright Fisher graph	6
E_N	Set of edges of a Wright Fisher graph	2
$v = (g, l) \in V_N$	The l-th individual in the $g-$th generation	6
(D_i^N)	Number of decedents at generation i of a sample at generation zero	6
(H_k^N)	Frequency process: $H_i^N = \frac{D_i^N}{N}$	6
$AL(v)$	Ancestral line of the individual $v \in V_N$	7
τ_N	Time to extinction or fixation	7
p_{jk}	the transition probability from the state j to the state k	7
(g_0, l_0)	Most recent common ancestor	7
T_{MRCA}^N	Time to the most recent common ancestor	7
$(\lvert A_g^N \rvert)$	Number of ancestors process	8
(A_g^N)	Ancestral process	9
$T_{MRCA}^N[n]$	T.M.R.C.A. of n individuals in generation zero	8
$\mathcal{S}_0$	A sample of n individuals in generation zero	9
$\mathcal{S}_{-g-1}$	A sample of m individuals in generation $-g-1$	9
$\overleftarrow{W}, \overrightarrow{W}, \mathcal{E}$	Useful events to prove moment duality	9
π_N	Probability of fixation starting from one individual	11
$[n]$	Set consisting of all the partitions of $\{1, 2, ..., n\}$	11
π	An element of $[n]$	11
$\lvert\pi\rvert$	Number of blocks in the partition π	11
$\succ$	Relation "follow" $\{\{1\}\{2\}\} \succ \{\{1, 2\}\}$	11
G_N	Geometric random variable with parameter $1/N$	11
(K_t)	The Kingman coalescent	11
ψ_i	Time of the i-th coalescence event	12
(X_t)	Generic symbol for a stochastic process (WF diffusion in page 13)	13
(B_t)	Brownian motion	13
(M_t)	Frequency process of the Moran model	13
(W_t)	Poisson process	13
(S_t)	Continuous time simple symmetric random walk	14
M	Set of *càdlàg* functions	14
d_M	Skorohod M_2 metric	15
Γ_1	Continuous graph of f_1	15
$\mathbb{S} = (M, \mathcal{M})$	Measurable (metric) space of *càdlàg* stochastic processes	15
$\mathcal{M}$	Borel σ algebra in (M, d_M)	15
$\Rightarrow$	Weak convergence	15
A, O	O is the set of open sets of $\mathcal{M}$. $A \in O$	16
A^ϵ	Open neighborhood of A	16
$\rho(\cdot, \cdot)$	Prohorov distance	16
$\rightrightarrows$	Convergence of the finite dimensional distributions	16

Table 2: Notation: Chapter 1, Part 2

Symbol	Description	Page
$\{P_t\}_{t\in I}$	Semigroup of operators. $P_t f(x) = \mathbb{E}_x[f(X_t)]$	17
$Af(x)$	The generator of a Markov process, applied to a function f at a point x	17
$D(A)$	Domain of the generator A	17
a	Drift of a diffusion	20
b^2	Diffusivity of a diffusion	20
τ_{coup}	Coupling time in the Doeblin coupling	21
$\|\| \cdot - \cdot \|\|_{TV}$	Total variation distance	22
ν	Stationary distribution	22
γ	Initial distribution	22
$\mathcal{P}(E)$	Set of probability measures over E	22
τ_{mix}	Mixing time	23
$H(x,n)$	Duality function	24
$H(x,n) = x^n$	Moment duality function	24
θ_1^N, θ_2^N	Mutation rates	25
$\mathcal{T}^n$	A "Kingman tree"	26
$\|\mathcal{T}^n\|$	Tree length of the tree $\mathcal{T}^n$	26
$\mathcal{S}$	Segregating sites	26
$\Delta_{i,j}$	Number of differences between the individuals i and j	27
Δ	Average pairwise differences	27
D	Tajima's D	27
s_N	Selective advantage	25
$N^{(i)}$	Population size in location i	29
$c^{1,2}, c^{2,1}$	Migration rate	29
$T_{MRCA}[m,n]$	T_{MRCA} with the starting configuration (n,m)	32

Table 3: Notation: Chapter 2

Symbol	Description	Page
μ_N	Seedbank age distribution	35
B	A μ_N distributed random variable	35
γ	Starting distribution	35
$(S_i^{(w)})$	The set of generations visited by an individual v_w is $\{S_i^{(w)} : i \in \mathbb{N}\}$	35
$\{U_k^{(w)}\}_{k\in\mathbb{N}}$	Independent uniform random variables with values in $\{1, ..., N\}$	36
T_r	Time of the r-th coalescence event	37
I_r	Blocks that take part in the r-th coalescence event	37
I_r	Blocks that take part in the r-th coalescence event, with $U_{T_r}^{(w)} = p$	37
J_r	Blocks that do not take part in the r-th coalescence event	37
q_i	The probability that an individual in generation zero has an ancestor in generation $-i$	39
L	Slowly varying function	39
Γ_α	Probability measures μ, such that $\mu(\{i, i+1, ...\}) = i^{-\alpha}L(i)$	39
$supp(\mu)$	Support of μ	41
(X_k)	Urn process	41
(M_i)	Successive visits of the urn process to the state zero	41
τ_i^{wj}	Times at which coalescence is possible between the blocks with smallest element w and j respectively.	43
τ_i	Times at which coalescence is possible	43
(X_k^N)	Urn process	43
(A_k^N)	Ancestral process constructed using (X_k^N) and $\{U_k^{(w)}\}$	43
(R_i^N)	Equal in distribution to $(X_{\tau_i}^N)$	43
$(\overline{R}_i^N)$	Sequence of independent ν_N distributed RV, $(\overline{R}_i^N, R_i^N)$ are constructed using optimal coupling	43
(Z_k^N)	$(Z_{\tau_i}^N) = R_i^N$ and constant in $k \notin \{\tau_i\}$	44
$(\overline{Z}_k^N)$	$(\overline{Z}_{\tau_i}^N) = \overline{R}_i^N$ and constant in $k \notin \{\tau_i\}$	44
$(\overline{L}_i^N)$	Ancestral process constructed using $(\overline{Z}_k^N)$ and $\{U_k^{(w)}\}$	44
(L_i^N)	Ancestral process constructed using (Z_k^N) and $\{U_k^{(w)}\}$	44
$\mathcal{L}$	Discrete generator of $(\overline{L}_k^N)$	45
Q	Absorption time of $(\overline{L}_k^N)$	45
$\underline{\mu}_N$	Truncated seedbank age distribution (SBAD)	45
$(\underline{X}_k^N)$	Urn process with SBAD $\underline{\mu}_N$	46
$(\underline{A}_k^N)$	Ancestral process induced by the Urn process $(\underline{X}_k^N)$	46
J_N	First time that (X_k^N) and $(\underline{X}_k^N)$ are different	46
G^α	Geometric random variable with parameter $N^{-\alpha}$	47

Table 4: Notation: Chapter 3

Symbol	Description	Page
N, M	Number of individuals per generation, N seeds, M plants	50
(X_k^N, Y_k^N)	Frequency process: (Frequency in the plants, Frequency in the seeds)	51
Z, U, V	Random variables that are used to describe the transitions of (X_k^N, Y_k^N)	51
(X_t, Y_t)	Solution to the Seedbank diffusion	53
c	Migration rate (from plant to seed and form seed to plant)	54
K	Relative seedbank size ($M = KN$)	54
(N_t, M_t)	Block counting process of the seed bank coalescent	54
(X_∞, Y_∞)	Limit in law when t goes to infinity of (X_t, Y_t)	55
$\mathcal{P}_k^{s,p}$	Space of marked partitions	56
(Π_t)	The seedbank coalescent	56
$\bowtie$	Relation "changing one flag" between marked partitions	56
(Π_t^N)	The ancestral process induced by a seedbank bank model	56
$t_{n,m}$	Expected time to the most recent common ancestor, with $(N_0, M_0) = (n, m)$	59
$l_{n,m}^{(a)}, l_{n,m}^{(d)}$	Expected tree length (active and inactive)	59
$\mathcal{P}_k^{\{p,s\}\times\{w,b\}}$	Space of coloured marked partitions	63
$(\underline{\Pi}_t)$	The coloured seedbank coalescent	63
$(\underline{N}_t, \underline{M}_t)$	The number of white blocks in the coloured seedbank coalescent	64
A_t^n	Number of deactivations until time t	64
τ_t^j	First time the number of active blocks is j	64
X_j^n	Indicator function of a deactivation at time τ_j^n	64
$\mathcal{B}_t$	Blocks with at least one white particle at time t	66
$(\overline{N}_t, \overline{M}_t)$	The block counting process of "another coloured coalescent"	68
H_k	First time that $\overline{N}_t + \overline{M}_t = k$	69
D_m	Time to reach $\sqrt{m}$ plants	69
J_m	First jump after D_m	69
$\mathcal{S}_r$	Lines that had visit the seed bank before time r	71
ρ^n	First time all lines, except maybe one, have visits the seedbank	71

Table 5: Notation: Chapter 4

Symbol	Description	Page
$F(B\|A)$	Fitness of a strain B relative to a strain A	75
i	Time measured in days	76
N	Number of bacterial cells at the beginning of each day	76
γN	Approximate number of bacterial cells at the end of each day	76
r	Reproduction rate	76
ϱ_N	Difference in the reproduction rate between the mutant and the basic population	77
b	$\varrho_N \sim N^{-b}$	77
r_o	Reproduction rate of the basic population (ancestor strain)	78
μ_N	Probability of a mutation occurring in a generation	78
a	$\mu_N \sim N^{-a}$	78
F_i	Average relative fitness at generation i	79
$R_{i,j}$	Reproduction rate of the j-th individual, at the begining of day i	79
$(Z_t^{(N)})$	Yule process	80
ς_N	Hitting time of $(Z_t^{(N)})$ to γN	80
σ_N	Hitting time of $(\mathbb{E}[Z_t^{(N)}])$ to γN (deterministic)	80
$(B_g^{(N,n)})$	Ancestral process	81
$(Y_t^{(N,k)})$	Population size at (interday) time t	82
$(M_t^{(k)})$	Number of mutants at (interday) time t	82
$(Z_t^{(N-k)})$	Number of non mutants at (interday) time t	82
(K_i)	Number of mutants at the begining of day i	82
τ_{fix}^N	Time (measured in days) until fixation	83
τ_{ext}^N	Time (measured in days) until extinction	83
τ^N	Time (measured in days) until either fixation or extinction	83
m_N	Time (measured in days) between the first mutation and the second mutation	84
H_i	Number of successful mutations until day i	85
$\overline{R}_i, \underline{R}_i$	Maximum (resp. minimum) reproduction rate of the individuals at day i	85
Φ_i	Approximate relative fitness	86
T_1^N	End of the first stage of a sweep	87
T_2^N	End of the second stage of a sweep	87
Γ	ΓN is the number of individuals at the end of a day	88
$(\underline{K}_i), (\overline{K}_i)$	Galton Watson processes that bound from above and below (K_i)	88
$\tilde{X}_j$	Indicator function that the j-th individual at the end of a day is sampled	88
$\overline{X}_j, \underline{X}_j$	Upper and lower (independent of j) stochastic bounds on $\tilde{X}_j$	89
J	Time at which the construction of $(\underline{K}_i), (K_i), (\overline{K}_i)$ stops working	89
$\tilde{J}$	Approximation of J	90
A_Γ	Event that γ and Γ are very similar	90
A	Event that $(\underline{K}_i) \leq (K_i) \leq (\overline{K}_i)$ until day $\varrho_N^{-1-\delta}$ or until day T_1^N	94
$(\underline{Q}_i), (Q_i), (\overline{Q}_i)$	Similar construction as $(\underline{K}_i), (K_i), (\overline{K}_i)$, for the third stage of the sweep	98
$I^{(j)}$	Time between the fixation of the j-th mutation and the $j+1$-th mutation	101
I_n	Sum of the first n-th variables $I^{(j)}$	101
D_i	No clonal interference until day i	102

Table 6: Notation: Appendix

Symbol	Description	Page
m_i	i-th arrival to $\{0\}$ of the process (X_n)	107
l_i	i-th arrival to $\{0\}$ of the process (Z_n)	107
σ_0	Times that belong to $\{m_i\}$ or to $\{l_i\}$	107
V_i	Length of the i-th visit to the state $\{0\}$ of the process (X_n)	107
W_i	Length of the i-th visit to the state $\{0\}$ of the process (Z_n)	108
R	An stochastic bound on the coupling time in terms of (V_i) and (W_i)	108
Z^r	Yule process with rate r	108
(E_i)	Independent exponential random variables	112
(G_i^N)	Galton Watson process	113
(H_i^N)	Generation size of individuals with an infinite line of descent	113

Motivation

Dormancy is defined by Lennon and Jones [38] as "any rest period or reversible interruption of the phenotypic development of an organism". It is a widespread evolutionary strategy and it has observable influence in ecology, adaptative evolution and genetic evolution. It is crucial for adaptation, for example it helps plants to survive the winter and bacteria to survive starvation. Further, Dormancy is also a useful tool in research laboratories. The aim of this thesis is to study dormancy by developing probabilistic-population-genetic models.

According to [38] up to 80% of the bacterial cells in the soil are in latent or dormant state. The soil is just an example, the number of active and inactive cells in bacterial populations tends to be of the same order of magnitude. Bacteria in dormant form constitute a genetic pool that challenges our intuition on the dynamics of genetic diversity. Some classic notions in population genetics become more complex in the presence of these reservoirs: What does it mean that a trait goes to fixation? What is the meaning of *generation*?

When we think about classic population genetics, the Wright Fisher model is the first object that comes to our mind. This is a probabilistic model for haploid populations in which there are numbered generations, each generation consists of exactly N individuals, where N is a natural number, and each individual *chooses* its parent from the previous generation uniformly at random (see Definition 1.1.1). The Wright Fisher model and its scaling limit, the Kingman coalescent (see Definition 1.1.26), have successfully been used to study the genealogy of many populations.

A seedbank, in the case of trees, is the set of seeds in the soil that can produce a newborn tree. The term seedbank has a more general connotation, it is used to denote a large group of individuals in latent state. The presence of seedbanks makes the Wright Fisher model an inadequate model, because the hypothesis that each individual *chooses* its parent form the previous generation becomes unrealistic. In this sense we say that the effect of seedbanks is not included in classical probabilistic modeling of population genetics.

To overcome these limitations, in 2001 Kaj, Krone and Lascoux [33] postulated an extension of the Wright Fisher model that includes seedbanks (see Section 2.1). Their model can be described as follows: fix the population size $N \in \mathbb{N}$ and a probability measure μ on the natural numbers. This measure determines the generation of the immediate ancestor of an individual backward in time, meaning that an individual living at generation $k \in \mathbb{Z}$ has its immediate ancestor in generation $k - l$ with probability $\mu(l)$.

In [33] the authors study the case where μ has finite support, and they conclude that on the evolutionary scale the ancestral process induced by their seedbank model converges to a constant time change of the Kingman coalescent.[1] Even though the Kaj Krone and Lascoux model (KKL model) and the Wright Fisher model are different, both are in the universality class of the Kingman coalescent. In this sense one can say that the effect of the seedbanks that can be effectively modelled using a bounded μ (weak seedbanks), is not drastic.

Chapter 2 has the goal of studying the KKL model beyond the assumption that μ has finite support. We extend the main result of [33] to cases in which μ has infinite support, but light tail (See Theorem 2.3.10). Further, we show that there exist choices of μ that change radically the behavior of the ancestral process, and that can't be modeled by manipulating the effective population size *i.e.* the limit cannot be a time changed Kingman coalescent (See Theorem 2.2.2). In particular, if μ has a very heavy tail, for

[1] The evolutionary scale is when time is measured in units of the population size *i.e.* the rescaling factor is N.

every fixed population size $N > 1$, the probability that two individuals do not have a common ancestor is positive.

In Chapter 2 we also extend the KKL model to let the measure μ depend on the population size N. We consider $\mu_N = (1-\epsilon)\delta_1 + \delta_{N^\beta}$, which is interpreted as follows: almost all individuals *choose* their parent from the previous generation, but few perform a very big jump. We prove that if $\beta < 1/5$ the time rescaled ancestral process, rescaled by a factor $N^{1+2\beta}$, converges to the Kingman coalescence. Interestingly, the relevant scale $N^{1+2\beta}$ is orders of magnitude bigger[2] than the evolutionary scale.[3]

The KKL model has the disadvantages that it is not Markovian and that it is hard to study it forward in time. In Chapter 3, we propose a seedbank model that allows a forward and a backward Markovian representation. The forward process converges in the evolutionary scale to a two dimensional diffusion, that we named **the seedbank diffusion**. The seedbank diffusion is characterized as the unique solution of the two dimensional SDE

$$\begin{aligned} \mathrm{d}X_t &= c(Y_t - X_t)\mathrm{d}t + \sqrt{X_t(1-X_t)}\mathrm{d}B_t, \\ \mathrm{d}Y_t &= cK(X_t - Y_t)\mathrm{d}t, \end{aligned}$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion. The ancestral process converges to a coalescent process, that we named **the seedbank coalescent**. The seedbank coalescent takes values in the space of marked partitions: each block has a label that can be either s or p (seed or plant), the label of each block changes at a certain rate and each pair of p-blocks coalesces at rate 1 (See definition 3.3.1). It turns out that the relation between the seedbank diffusion and the seedbank coalescent is similar to the classic relation between the Kingman coalescent and the Wright Fisher diffusion. In Theorem 3.2.7 we showed that the seedbank diffusion and the block counting process of the seedbank coalescent are moment dual. The rest of Chapter 3 deals with properties of the seedbank coalescent, for example in Theorem 3.4.8 we show that the expected time to the most recent common ancestor in a sample of size n is of order $\log\log(n)$.[4]

So far we have just discussed latency under natural conditions. However, latency is also used in research laboratories. It plays a crucial role in the area of biology known as "experimental evolution". The philosophy of this research area is to measure adaptative and genetic evolution of bacterial populations under laboratory conditions. To measure adaptation in practice it is imperative to compare an unevolved ancestral population with an evolved population. How can one compare two populations that exist at different times? The answer is: by using latency. To explain this better, we will describe "the long term experiment with *Escherichia coli*" [41], which is also known as *the Lenski experiment* in honor of Dr. Richard Lenski. This experiment is a cornerstone in experimental evolution and the main object of study of Chapter 4. The *Lenski experiment* investigates the long-term evolution of 12 initially identical populations of the bacteria *E. coli* in identical environments. One of the basic concepts of the Lenski experiment is that of *daily cycles*. Every day starts by sampling the same amount of cells from the bacteria available in the medium that was used the day before. This sample is propagated in an identical medium as that of the previous day. This procedure is repeated daily. Up to now, the experiment has been going on for more than 60000 generations (or 9000 days, see [39]). One important feature is that samples of ancestral populations were stored at low temperatures, forcing the bacteria to reach a latent state. Afterwards the bacteria can be made to reproduce under competition with later generations in order to experimentally determine the fitness of an evolved strain relative to the founder ancestor of the population by comparing their growth rates. It was observed, for example by Wiser et al.[74], that the relative fitness over time increases sublinearly, a behaviour which is commonly attributed to effects like clonal interference or epistasis.

In Chapter 4 we construct an individual based model that studies the adaptive evolution in the Lenski experiment, and that does neither include clonal interference nor epistasis. Our results show (Theorem 4.2.15) that in a suitable scale, the relative fitness increases approximately as the curve

$$f(t) = \sqrt{1 + \frac{4.04}{r_0^2}t}, \quad t \geq 0.$$

[2] These results were used to discuss the effect of seedbanks in evolution of bacteria, published in [23].
[3] Chapter 2 consists of results published in [7] and [6]. However, most of the proofs presented in this Chapter are new.
[4] Chapter 3 consists essentially of the paper [8] together with some results that can be found in [5].

where r_0 is the reproduction rate of the ancestral population. This is consistent with previous work on the topic (see [74]) and provides the new insight that the design of the experiment is a factor that shapes adaptive evolution in the Lenski experiment, and should be taken into account to make statements about the role of epsitasis and clonal interference.[5]

The rest of this thesis is organized as follows: In Chapter 1 we introduce several concepts and ideas, in order to make this work as self-contained as possible. In Chapter 2 we discuss generalizations of the KKL model, in particular when μ has an unbounded support and when $\mu := \mu_N$ is a function of the population size. In Chapter 3 we post our proposal for a seedbank model and we study its properties. Chapter 4 consists in a model that studies a particular case in which latency is used as a tool in an experiment: the Lenski experiment. Finally, Appendix A contains some technical results that are useful in different parts of the thesis.

[5] Chapter 4 is essentially the paper [24].

Chapter 1

Introduction

The goal of this section is to introduce some basic concepts and tools that will be useful during the rest of the thesis, and also to give a background so that the reader can put the main results into context. The aim is to make this thesis as self-contained as possible.[1]

1.1 Some basic models in population genetics

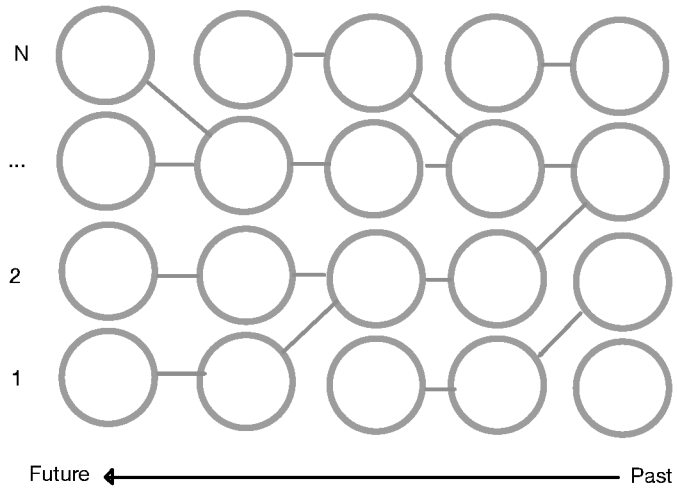

Figure 1.1: A realization of the Wright Fisher Graph. Columns are generations, in this figure the population size is $N = 4$. Each individual is the leftmost point of a line that connects it to its parent *i.e.* its parent is the rightmost point of the line.

The relation between mathematics and biology has been long and fruitful. Population Genetics is an area of knowledge that exists in the intersection of these two disciplines. According to [60] *Population genetics is the study of the frequency and interaction of alleles and genes in population.* There are different phenomena that shape the adaptive evolution of a population, known as evolutionary forces. Some examples are natural selection, mutation, gene flow and genetic drift. Genetic drift is a purely random force, which was introduced independently by Fisher [20] and Wright [75] and is better understood by means of a probabilistic model, the so called **Wright Fisher Model**. This model consists of numbered

[1] The content of this section is similar to Berestycki [4] and Etheridge [17].

generations, where each generation has exactly the same number of individuals and each individual has exactly one parent, which is a randomly chosen individual from the previous generation. The individuals and their relations form a graph. The following definition makes this precise. The graphical construction above of the Wright Fisher model is inspired in [7].

Definition 1.1.1. *Let* $V_N = \{v = (g,l) \in \mathbb{Z} \times \{1,2,...,N\}\}$, $\{U_v\}_{v \in V_N}$ *be a sequence of independent random variables, uniformly distributed in* $\{1,2,...,N\}$, *and*

$$E_N = \{\{(g-1, U_{(g,l)}), (g,l)\} \text{ for all } v = (g,l) \in V_N\}.$$

We define ***the N-Wright Fisher graph*** *to be the random graph with vertex set* V_N *and edge set* E_N.

Definition 1.1.2. *We define the* ***generation*** g, *for any* $g \in \mathbb{Z}$, *to be* $Gen(g) = \{v = (g,l') \in V_N : l' \in \{0,1,...,N\}\}$.

Remark 1.1.3. In the previous definition each vertex $v = (g,l) \in V_N$ should be understood as the l-th individual in the g-th generation. There are as many generations as integer numbers and each generation consists of N individuals.

Definition 1.1.4. *If* $v' = (g', l')$ *and* $v = (g,l)$ *are such that* $g - g' = k > 0$, *and there exist a sequence of exactly* $k-1$ *edges in* E_N *that connect* v' *and* v, *we say that* v' *is an ancestor of* v *and that* v *is a descendant of* v'. *If two individuals have a descendant/ancestor relation and* $g - g' = 1$, *then we say that they have an offspring/parent relation.*

The vertices of the graph are individuals, and the edges are the parental relations. If we assume that generation zero is the present, there are two natural ways to study the graph: going to the future or going to the past.

Let us first go in the direction of the future (forward in time). Let us fix one individual in generation zero, how many individuals in generation one are its offspring? Each individual in generation one will be its offspring with probability $1/N$, and there are N individuals, so the number of offspring in generation one of the individual in generation zero is Binomially distributed with parameters $(1/N, N)$. A similar reasoning can be applied if we sample more than one individual. Suppose that we sample n individuals in generation zero and we want to know how many individuals in generation one are descendants of some member of the sample. Each individual in generation one chooses a (fixed) member of the sample as its parent, with probability $1/N$. So the number of individuals in generation one that have a parent in the sample of size n in generation zero, is Binomially distributed with parameters $(n/N, N)$.

Let us call D_i^N the number of descendants at generation i of a sample of individuals in generation zero. Then $(D_i^N)_{i \in \mathbb{N}}$ is a Markov chain with values in $\{1,2,...,N\}$. Indeed, note that if $D_i^N = d$ then D_{i+1}^N is an independent Binomial random variable with parameters $(d/N, N)$ (this implies in particular that $(D_i^N)_{i \in \mathbb{N}}$ is a Markov chain). It is often useful to work with $H_i^N = \frac{D_i^N}{N}$.

Definition 1.1.5. *The* ***frequency process of the Wright Fisher model*** *is the Markov chain* $(H_i^N)_{i \in \mathbb{N}}$, *with state space* $\{0, 1/N, 2/N, ..., 1\}$ *and transition probabilities,*

$$p_{j,k} = \binom{N}{kN} j^{kN} (1-j)^{(1-k)N}$$

for any $j, k \in \{0, 1/N, 2/N, ..., 1\}$.

Remark 1.1.6. The name *frequency process* comes from the following interpretation: assume that in the Wright Fisher graph at generation zero there are two types of individuals, black and pink, so that n individuals are black and $N - n$ pink. The frequency of black individuals is a realization of (H_i^N).

Remark 1.1.7. Throughout the whole thesis we will denote a stochastic process by (Y_i^N), instead of writing $(Y_i^N)_{i \in I}$, whenever it is clear which is the index i and the index set I. For example, in the previous remark (H_i^N) stands for $(H_i^N)_{i \in \mathbb{N}}$.

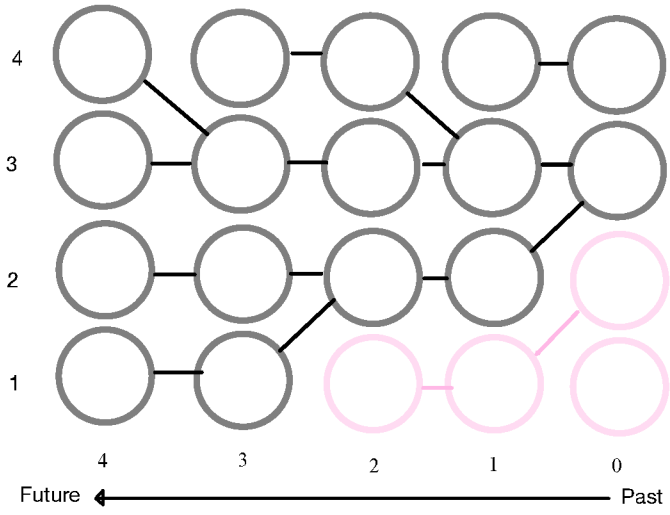

Figure 1.2: A realization of the Wright Fisher frequency process, with $H_0^N = 1/2$ and $N = 4$

Note that (H_i^N) has two absorbing states 0 and 1. Indeed, coming back to the types interpretation introduced in Remark 1.1.6, if a type gets extinguished it will be lost forever and if it manages to invade the whole population, then all the individuals will be of this type forever. A key observation is that one of these absorbing states will be reached in finite time almost surely. This means that, even if the process starts with two neutral types, one will go to fixation and the other will disappear. This phenomena of loss of variability caused by randomness is an important evolutionary force, which is called **Genetic Drift**.

Proposition 1.1.8. *Let $n \in \{0, 1, ..., N\}$. Assume that $H_0 = n/N$. Let τ^N be the time until there is only one type in the population, that is $\tau^N = \inf\{i : H_i^N(1 - H_i^N) = 0\}$, then τ^N is finite almost surely.*

Proof. Note that for all $j \in \{0, 1/N, ..., 1\}$, $p_{j0} + p_{j1} \geq (1/2)^N$ where p_{jk} is the transition probability from the state j to the state k, for any $j, k \in \{0, 1/N, ..., 1\}$. Then, for any starting point $s \in \{0, 1/N, ..., 1\}$, $\mathbb{P}_s(\tau^N > r) < (1 - (1/2)^N)^r$, which implies that $\lim_{r\to\infty} \mathbb{P}(\tau^N > r) = 0$. □

Let us now go back to Definition 1.1.1 and study the Wright Fisher graph in the direction of the past (backward in time). Let us consider again a sample of n individuals in generation zero. The question now is, do the sampled individuals have a common ancestor? If so how many generations ago did the common ancestor live?

If we sample one individual at generation $g \in \mathbb{Z}$, the Wright Fisher graph is constructed in such a way that there will be exactly one ancestor of this individual in each generation $g' < g$. The set of ancestors is called the ancestral line.

Definition 1.1.9. *Let $v = (g, l) \in V_N$. The ancestral line of v is the set $AL(v) \subseteq V_N$ of all ancestors of v. (Recall that the notion of ancestor is defined in 1.1.4)*

In the Wright Fisher model, once the ancestral lines of two individuals intersect, they follow the same trajectory (in the direction of the past). If we consider a sample of several individuals we can define a Markov process by counting the number of members of the ancestral lines of the sampled individuals in each generation. This is related to an important concept in population genetics, the time to the most recent common ancestor (T_{MRCA}).

Definition 1.1.10. *The most recent common ancestors of a sample $\{v_i\}_{i\in\{1,2,...,n\}} = \{(g_i, l_i)\}_{i\in\{1,2,...,n\}} \in V_N$ is defined as $v_0 = (g_0, l_0) \in \bigcap_{i=1}^n AL(v_i)$ such that if $v = (g, l) \in \bigcap_{i=1}^n AL(v_i)$, then $g_0 \geq g$.*

Suppose that $g_i = 0$ for all $i \in \{1, 2, ..., n\}$, this means that all the members of the sample belong to generation zero, then the time to the most recent common ancestor of a sample of size n is defined as $T_{MRCA}[n] := -g_0$, where (g_0, l_0) is the most recent common ancestor.

We usually write T_{MRCA} when the sample size is two, i.e. $T_{MRCA} := T_{MRCA}[2]$.

Remark 1.1.11. The time to the most recent common ancestor of two individuals in the Wright Fisher model is geometrically distributed with parameter $1/N$. This follows from the fact that two individuals at a certain generation have a common ancestor in the previous generation with probability $1/N$, and given that they do not have a common ancestor until generation $-g$ they will have a common ancestor at generation $-g-1$ with probability $1/N$.

To study the time to the most recent common ancestor of a bigger sample it is convenient to introduce a Markov process.

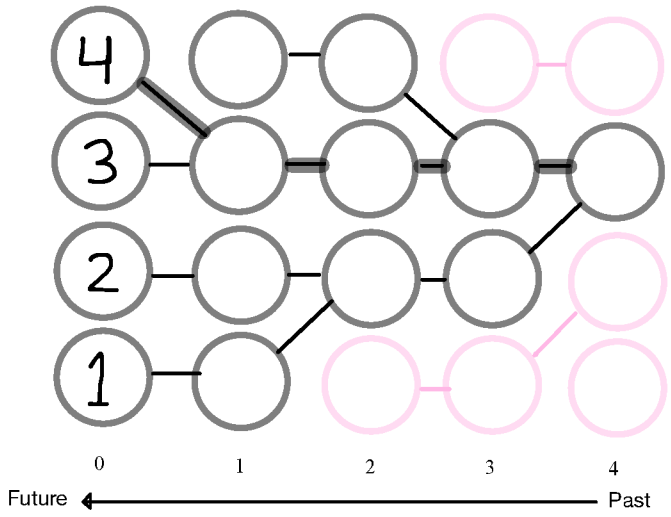

Figure 1.3: A realization of the Wright Fisher Ancestral process, with the ancestral line of individual $(0, 4)$ highlighted

Definition 1.1.12. *Let $\{(0, l_i)\}_{i\in\{1,2,...,n\}} \in V^N$ be a subset of generation zero. Then*

$$|A_g^{N,n}| = \Big|\bigcup_{i=1}^{n}\{AL(0, l_i)\} \cap \{Gen(-g)\}\Big|$$

*$(|A_g^{N,n}|)_{g\in\mathbb{N}}$ is **the number of ancestors process of the Wright Fisher model**. This process is also known as the **the block counting process of the Wright Fisher model**.*

Remark 1.1.13. To save notation, we will write $|A_g^N|$ instead $|A_g^{N,n}|$ whenever there is no risk of confusion.

Remark 1.1.14. Note that the process $(|A_g^N|)_{g\in\mathbb{N}}$ runs backwards in time: as g increases, $(|A_g^N|)_{g\in\mathbb{N}}$ counts the ancestors that existed further in the past.

The time to the most recent common ancestor of a sample of n individuals in generation zero, can be written in terms of $(|A_g^N|)$ as follows:

$$T_{MRCA}^N[n] = \inf\{g \geq 0 : |A_g^N| = 1 \text{ given that } A_0^N = n\}. \tag{1.1.1}$$

Here it is important to note that $(|A_g^N|)$ is a Markov process in the filtration $(\mathcal{F}_g)$, where $\mathcal{F}_g$ is generated by the generations posterior to $-g$. More precisely

$$\mathcal{F}_g = \langle\{U_{(g',l')}\}_{(g',l')\in V_N, g'>-g}\rangle,$$

where $\{U_{(g',l')}\}_{(g',l')\in V_N}$ is the sequence of uniform random variables introduced in Definition 1.1.1.

Remark 1.1.11 implies that $T^N_{MRCA}[2] = T_{MRCA}$ is finite almost surely and, more precisely, that $\mathbb{E}[T^N_{MRCA}[2]] = N$. It is interesting to note that for all $n \in \mathbb{N}$, $T^N_{MRCA}[n]$ is finite almost surely.

Proposition 1.1.15. *For every $n \in \mathbb{N}$, it holds that*

$$\mathbb{E}[T^N_{MRCA}[n]] < \infty.$$

Proof. The proof is immediate by induction. For $n = 2$ the claim follows. Assume the claim is true for $n-1$. If we denote $T^N_{MRCA}[n-1]$ the time to the most recent common ancestor of the individuals labeled $1, 2, ..., n-1$, then either the n labeled individual has a common ancestor with the rest of the sample at the random time $T^N_{MRCA}[n-1]$ or at time $T^N_{MRCA}[n-1]$ there are exactly two ancestors of the whole sample of size n. Then we have that by the strong Markov property

$$\mathbb{E}[T^N_{MRCA}[n]] \leq \mathbb{E}[T^N_{MRCA}[n-1]] + \mathbb{E}[T^N_{MRCA}[2]] < \infty.$$

□

Now we will introduce the **ancestral process**, which contains a bit more information than the number of ancestors process. The ancestral process takes values in the space of partitions. It is generated by the equivalence relations $\{\sim_g\}_{g\in\mathbb{N}}$, defined by the rule $v_1 \sim_g v_2$ if and only if v_1 and v_2 have a common ancestor in generation $-g$.

Definition 1.1.16. *Let $\{v_i\}_{i\in\{1,2,...,n\}} = \{(0, l_i)\}_{i\in\{1,2,...,n\}}$ be a sample of n individuals in generation zero. For every $g \in \mathbb{N}$ define $\sim_g$ to be the equivalent relation on $\{1, 2, ..., n\}$ characterized by the rule: For every $i, j \in \{1, 2, ..., n\}$, we say that i and j are **g-equivalent**, and we write $i \sim_g j$, if and only if*

$$|\{AL(0, l_j)\} \cap \{AL(0, l_i)\} \cap \{Gen(-g)\}| = 1.$$

*Let π_g be the equivalence classes generated by $\sim_g$. Then, **the ancestral process of the Wright Fisher model** is defined as*

$$(A^N_g)_{g\in\mathbb{N}} := (\pi_g)_{g\in\mathbb{N}}.$$

Let us denote $[n]$ the set of partitions of $\{1, 2, ..., n\}$ for any $n \in \mathbb{N}$.

Remark 1.1.17. Let π be a partition of $[n]$. Let $|\pi|$ be the number of blocks of the partition π. Note that the number of ancestors process is related to the ancestral process, as, in distribution,

$$(|A^N_g|)_{g\in\mathbb{N}} = (|\pi_g|)_{g\in\mathbb{N}}.$$

We have defined a forward and a backward process associated to the Wright Fisher graph. A natural question is: How do these two processes relate? We can follow the ideas of [50] to find an answer.

Fix $m, n \in \mathbb{N}$. Let $\mathcal{S}_{-g-1} = \{v_1, ..., v_m\} = \{(-g-1, l_1), ..., (-g-1, l_m)\}$ be a sample of size m of individuals at generation $-g-1 \in \mathbb{Z}$. For every $i \in \{0, 1, 2, ..., N\}$, define the event

$$\overrightarrow{W}(i) = \{\text{There are } i \text{ descendants of } \mathcal{S}_{-g-1} \text{ in generation } -1\}.$$

Now define $\mathcal{S}_0 = \{v'_1, ..., v'_n\} = \{(0, l'_1), ..., (-g-1, l'_n)\}$ to be a sample of size n of individuals at generation 0. For any $i \in \{1, 2, ..., n\}$, define the event

$$\overleftarrow{W}(i) = \{\text{There are } i \text{ ancestors of } \mathcal{S}_0 \text{ in generation } -g\}.$$

Finally, define the event

$$\mathcal{E} = \{\text{All the ancestors of } \mathcal{S}_0 \text{ in generation } -g-1 \text{ are contained in } \mathcal{S}_{g-1}\}. \tag{1.1.2}$$

Note that the events $\overrightarrow{W}(i)$, $\overleftarrow{W}(i)$ and $\mathcal{E}$ belong to the sigma algebra of the Wright Fisher graph. We can use the law of total probability in two different ways to calculate the probability of $\mathcal{E}$. On one hand we have

$$\mathbb{P}(\mathcal{E}) = \sum_{i=0}^{N} \mathbb{P}(\mathcal{E} \mid \overrightarrow{W}(i))\mathbb{P}(\overrightarrow{W}(i)) = \sum_{i=0}^{N} (i/N)^n \mathbb{P}_{m/N}(H^N_g = i/N) = \mathbb{E}_x[(H^N_g)^n] \tag{1.1.3}$$

Here, the crucial step was in the second equality, where we needed the following:

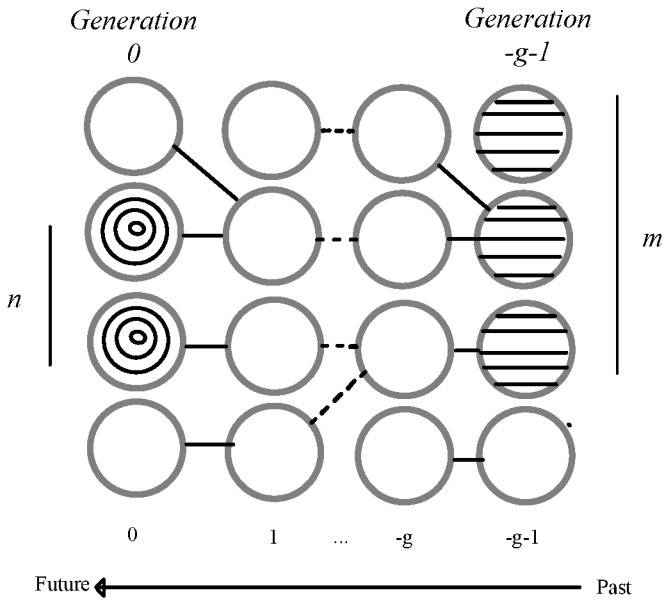

Figure 1.4: In this figure, $\mathcal{E}$ corresponds to the event that all the balls with circles in the interior have an ancestor with stripes in the interior.

1. In order for $\mathcal{E}$ to happen, all the members of S_0 must have a descendant of S_{-g-1} as its ancestor in generation -1. If there are i decedents of S_{-g-1} in generation -1, this happens with probability $(i/N)^n$, as each of the n members of S_0 chooses its parent independently at random.

2. We can construct a realization of the frequency process $(H_g^N)_{g\in\mathbb{N}}$, with initial condition $H_0^N = m/N$, by taking generation $-g-1$ to be our "zero" generation, and S_{-g-1} to be our set of "pink" individuals. Then,
$$\mathbb{P}(\overrightarrow{W}(i)) = \mathbb{P}_{m/N}(H_g^N = i/N).$$

Similarly, we can calculate the probability of $\mathcal{E}$ by conditioning on the number of ancestors of S_0 at generation g.
$$\mathbb{P}(\mathcal{E}) = \sum_{i=0}^{n} \mathbb{P}(E \mid \overleftarrow{W}(i))\mathbb{P}(\overleftarrow{W}(i)) = \sum_{i=0}^{N} x^i \mathbb{P}_n(|A_g^N| = i) = \mathbb{E}_x[x^{|A_g^N|}]. \tag{1.1.4}$$

Equations 1.1.3 and 1.1.4 lead to the relation between the backward and forward processes.

Theorem 1.1.18. *For every $g \in \mathbb{N}$, $n, m \in \{1, 2, ..., N\}$ and $x = m/N$, it is true that*
$$\mathbb{E}_x[(H_g^N)^n] = \mathbb{E}_n[x^{|A_g^N|}].$$

Proof. The proof follows by comparing Equation 1.1.3 and 1.1.4. □

Remark 1.1.19. Theorem 1.1.18 is Proposition 3.5 in [50].

The relation proved in Theorem 1.1.18 is called *moment duality* and it is very useful. For example, in Lemma 1.1.8 we saw that the time until the fraction of individuals of certain type reaches zero or one is finite almost surely. Using the moment duality one can deduce that if the frequency of this certain type at generation zero is x, then the probability that the frequency eventually reaches one is just x.

Remark 1.1.20. Moment duality is a very useful relation. It will be discussed in more detail in subsection 1.2.5, and several examples will be presented in Section 1.3. An important result of this Thesis (Theorem 3.2.7) is the moment duality between the seedbank diffusion and the block counting process of the seedbank coalescent.

Lemma 1.1.21. *For every $m \in \{0, 1, ..., N\}$, let $x = m/N$. As before $\tau^N = \inf\{i : H_i^N(1 - H_i^N) = 0\}$, then $\mathbb{P}_x(H_{\tau^N}^N = 1) = x$.*

Proof. As $H_{\tau^N}^N \in \{0, 1\}$ we have that $\mathbb{P}_x(H_{\tau^N}^N = 1) = \mathbb{E}_x[H_{\tau^N}^N]$. By Lemma 1.1.8 and the dominated convergence Theorem, we know that $\mathbb{E}_x[H_{\tau^N}^N] = \lim_{g\to\infty} \mathbb{E}_x[H_g^N]$. Finally, by Theorem 1.1.18 we know that $\lim_{g\to\infty} \mathbb{E}_x[(H_g^N)^n] = \lim_{g\to\infty} \mathbb{E}_n[x^{A_g^N}] = x$, where the last equality is a consequence of the dominated convergence Theorem and of Proposition 1.1.15. □

Remark 1.1.22. A similar technique is used in Corollary 3.2.3 to study the long term behavior of the seedbank diffusion.

Corollary 1.1.23. *Let $\pi_N := \mathbb{P}_{1/N}(H_{\tau^N}^N = 1)$. Then $\lim_{N\to\infty} N\pi_N = 1$.*

Proof. This is an immediate application of Lemma 1.1.21 for the case $x = 1/N$. □

Remark 1.1.24. In the presence of selection, the value of π_N changes. Compare this result with Haldane's formula (See Remark 1.3.7) and Theorem 4.2.10 in Chapter 4.

Let us now assume that the number of individuals in each generation is very big. It turns out that rescaling the backward and the forward processes suitably one obtains well defined limits, which turn out to be interesting and useful processes.

In Remark 1.1.11 we learned that the time to the most recent common ancestor of a sample of size two is Geometrically distributed with parameter $1/N$. For every $N \in \mathbb{N}$, let G_N be a geometric random variable with parameter $1/N$. Recall that for any $t > 0$,

$$\lim_{N\to\infty} \mathbb{P}(G_N > Nt) = \lim_{N\to\infty} (1 - 1/N)^{\lfloor Nt \rfloor} = e^{-t}. \tag{1.1.5}$$

This gives a hint of the right scale to look for a limit. Indeed, if we consider the sequence of processes $\{(A_{\lfloor Nt \rfloor}^N)_{t\in\mathbb{R}^+}\}_{N\in\mathbb{N}}$, it converges to a well defined object, the Kingman coalescent. We will denote this process by $(K_t)_{t\in\mathbb{R}^+}$. From (1.1.5) we grasp that this object should be such that each pair of ancestral lines coalesces after an exponential time with parameter 1. This is the first ingredient to define $(K_t)_{t\in\mathbb{R}^+}$, the second ingredient is that each pair of ancestral lines coalesces independently from the others. As we did with the ancestral process, we will first define a process with values in the natural numbers, and then a partition valued process.

Definition 1.1.25. *The block counting process of the Kingman coalescent $(|K_t|)_{t\in\mathbb{R}^+}$ is a continuous time Markov process, with values in $\mathbb{N}$ characterized by the transition rates: $\binom{n}{2}$ from n to $n - 1$. All other transition rates are zero.*

The Kingman n-coalescent ([34]), opposed to the block counting process of the Kingman coalescent, does not take values in the natural numbers. It is a process with values in the partitions of $[n]$. The intuition behind it is the same as in the discrete time case, which is that two individuals in a sample will be in the same block of the random partition at time t if and only if they have a common ancestor before time t. That is, a pair of blocks coalesces each time that the individuals inside the two blocks find a common ancestor.

We say that a partition $\pi_1 \in [n]$ follows a partition $\pi_0 \in [n]$ if π_1 can be constructed by merging exactly 2 blocks of π_0, in that case we write $\pi_0 \succ \pi_1$. For example, if $n = 3$, we can see that $\{\{1\}, \{2\}, \{3\}\} \succ \{\{1, 2\}, \{3\}\}$.

Definition 1.1.26. ***The Kingman n-coalescent**, (K_t), with initial distribution $K_0 = \pi_0 \in [n]$, is the continuous time Markov process with values in $[n]$, characterized by the transition rates: (K_t) goes from π_0 to π_1 at rate 1 if π_1 follows π_0. All other transition rates are zero.*

Remark 1.1.27. If we denote by $|\pi|$ the number of blocks of a partition $\pi \in [n]$, we note that $|K_t|$ is precisely the process defined in 1.1.25, as each partition of k blocks has $\binom{k}{2}$ partitions that follow it.

The Kingman coalescent is a key object in population genetics, not only because of its simplicity, but also because it is a universal limit for models in population genetics ([34], [51]). Now we will show the power of the Kingman coalescent by calculating the expected time to the most recent common ancestor of a sample of size n.

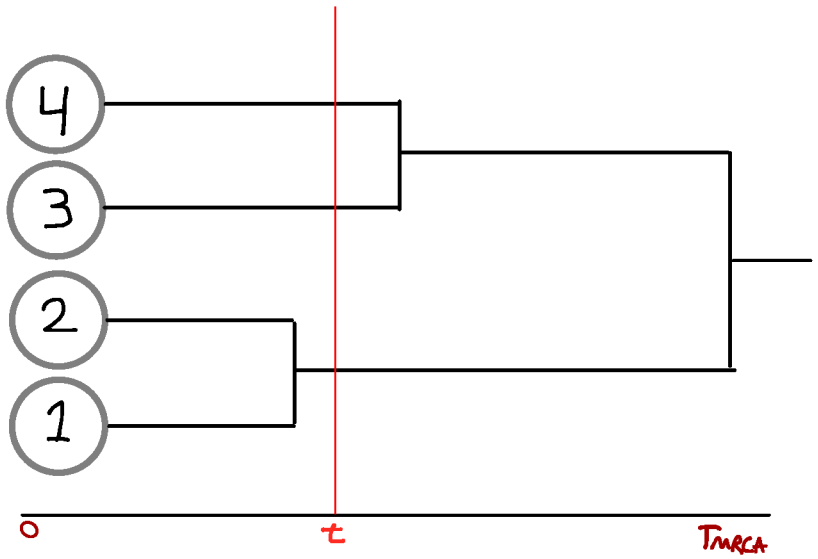

Figure 1.5: A realization of the Kingman coalescent. Each pair of blocks coalesce at rate 1. At time t the state of the process is the partition $\{\{1,2\},\{3\},\{4\}\}$.

Lemma 1.1.28. *(Kingman [34]) The expected time to the most recent common ancestor in the Kingman n-coalescent is*

$$\mathbb{E}[T_{MRCA}[n]] = 2\left(1-\frac{1}{n}\right),$$

where $T_{MRCA}[n] := \inf\{t>0 : K_t = \{\{1,2,...,n\}\}\}$ *and* (K_t) *is the Kingman n-coalescent.*

Proof. The main idea of the proof is to study the times of coalescence. Let ψ_1 be the time of the first coalescent event, ψ_2 the time of the second coalescent event and so on. That is, $\psi_i = \inf\{t > 0 : |K_t| = n-i\}$. Now note that, as each pair of blocks coalesce after an exponential time with parameter 1 and since when $|K_t| = n-i$ there are $\binom{n-i}{2}$ pairs, then for all $i \in \{1,2,...,n-1\}$, it holds that $\psi_{i+1} - \psi_i$ is the minimum of $\binom{n-i}{2}$ independent exponential random variables with parameter 1. This implies that $\psi_{i+1} - \psi_i$ is distributed exponential with parameter $\binom{n-i}{2}$. We finally observe that $T_{MRCA}[n] = \psi_{n-1} = \sum_{i=1}^{n-1}[\psi_i - \psi_{i-1}]$, where $\psi_0 = 0$. then

$$\begin{aligned}
\mathbb{E}[T_{MRCA}[n]] &= \mathbb{E}\left[\sum_{i=1}^{n-1}[\psi_i - \psi_{i-1}]\right] \\
&= \sum_{i=1}^{n-1}\mathbb{E}[\psi_i - \psi_{i-1}] \\
&= \sum_{i=1}^{n-1}\frac{1}{\binom{n-i+1}{2}} \\
&= \sum_{i=1}^{n-1}\frac{2}{(n-i+1)(n-i)} \\
&= 2\sum_{i=1}^{n-1}\left[\frac{1}{n-i} - \frac{1}{n-i+1}\right] \\
&= 2\left(1-\frac{1}{n}\right).
\end{aligned}$$

□

Remark 1.1.29. It is notable that $\mathbb{E}[T_{MRCA}[n]] < 2$ for all n. This is not the case for all coalescent process, for example in the Bolthausen-Sznitman coalescent and in the seedbank coalescent the expected time to the most recent common ancestor of a sample of n individuals is of order $\log\log(n)$ (See [22] and Theorem 3.4.8).

Now we discuss the scaling limit of the forward in time processes.

Definition 1.1.30. *The Wright Fisher diffusion is the pathwise unique solution to the stochastic differential equation $X_0 = x \in [0,1]$ and*

$$dX_t = \sqrt{X_t(1-X_t)}dB_t \tag{1.1.6}$$

where (B_t) is a Brownian motion.

Remark 1.1.31. The Wright Fisher diffusion is closely related to one of the main objects of study of Chapter 3, the seedbank diffusion (See Corollary 3.2.2).

Remark 1.1.32. Pathwise uniqueness of the solution of Equation (1.1.6) is a consequence of the Theorem of Yamada and Watanabe [76]. Existence of a solution can be prove using Theorem 3.2 of [30].

An introduction to Stochastic integration can be found in [15], and an introduction to diffusion theory can be found in [62]. Using the Dambis Dubins Schwarz Theorem (see [15]), an equivalent definition of the Wright Fisher diffusion is the following.

Proposition 1.1.33. *The Wright Fisher diffusion is the solution to the time change equation $X_0 = x \in [0,1]$ and*

$$X_t = B_{\int_0^t X_s(1-X_s)ds}$$

where B_t is a Brownian motion.

To give some intuition for (X_t) we will explain where the time change comes from. To do this let us introduce another classical model called the discrete Moran model. In this model there are N individuals per generation and at each time step an individual is chosen randomly to reproduce and an individual is chosen randomly to die. It can happen that the same individual is chosen to do both actions, in that case it does nothing. At generation zero individuals are assigned types (say black and pink) and each individual gets the same type of its parent. This model is very similar to the Wright Fisher model, with the difference that here only one individual performs an action at each time step, while in the Wright Fisher model all individuals in the population perform an action at each time step. If we denote (as in the Wright Fisher case) H_i^N the frequency of black individuals in the population at the i-th time step, we obtain a Markov chain. Note that if $H_i^N = x$, the frequency will increase to $x + 1/N$ in the next step if a black individual is chosen to reproduce and a pink one to die. This happens with probability $x(1-x)$ as x is the probability of choosing a black one to reproduce and $1-x$ is the probability of choosing a pink one to die. By similar arguments we see that $\mathbb{P}(H_{i+1}^N = x - 1/N | H_i^N = x) = (1-x)x$, so we can motivate the following formal definition:

Definition 1.1.34. *For any $N \in \mathbb{N}$, the N-frequency process of the discrete Moran model is the Markov chain, $(H_i^N)_{i\in\mathbb{N}}$ with state space $\{0, 1/N, 2/N, ..., 1\}$ and transition probabilities*

$$\mathbb{P}_x(H_1 = x + k) = \begin{cases} x(1-x) & \text{if } k = 1/N, \\ 1 - 2x(1-x) & \text{if } k = 0, \\ x(1-x) & \text{if } k = -1/N, \\ 0 & \text{in any other case.} \end{cases}$$

Now let us randomize the time to obtain a continuous time process. For this, we recall the Poisson process, which is a Markov process with values in the natural numbers that goes from the state n to the state $n+1$ at rate 1, and such that all other transitions are impossible. In the following definition we present a classic way to construct a Poisson process that will be useful for some examples later.

Definition 1.1.35. *Let $\{e_i\}_{i=1}^{\infty}$ be a sequence of independent identically distributed standard exponential random variables. Let*

$$W_t := \sup\{r \in \mathbb{N} : \sum_{i=1}^{r} e_i < t\}.$$

Then (W_t) is a standard Poisson process. For any $c \in \mathbb{R}$, we define a Poisson process with rate c, to be any Markov process equal in distribution to (W_{ct}).

Now we can use (H_i^N) and (W_t) to construct the frequency process of a Moran model.

Definition 1.1.36. *Let $(M_t^N)_{t\in\mathbb{R}^+}$ be a continuous time Markov process with state space $\{0, 1/N, 2/N, ..., 1\}$, defined by the formula*

$$M_t^N := H_{W_t}^N$$

*where (W_t) is a standard Poisson process and (H_i) is the N-frequency process of the discrete Moran model. $(M_{N^2t}^N)$ is the **N-frequency process of the Moran model.***

A reason for introducing this process is that we can construct it using a simple symmetric random walk in continuous time. Let (S_t) be a simple symmetric random walk in continuous time, and let $U_t^N = \frac{1}{N} S_{\int_0^t 2U_s(1-U_s)ds}$ then in distribution

$$(M_t) = (U_t^N).$$

Note that as $\frac{1}{N} S_{N^2t}^N \Rightarrow B_t$ weakly over the space of Skorohod, where (B_t) is a Brownian motion. It is intuitively clear that

$$(U_{N^2t}^N) \Rightarrow (X_{2t})$$

weakly over the space of Skorohod, where (X_t) is the Wright Fisher diffusion, introduced in Definition 1.1.30. The relevant scale here is N^2, in contrast to the relevant scale in the Wright Fisher case which is N. This follows from the fact that in the Wright Fisher model N (Reproduction/death) events occur each time step, while in the discrete Moran model only one event occurs. This convergence can be proved using generators (See Subsection 1.2.2).

1.2 Toolbox

In this subsection we introduce some notions that will be used during the rest of this work. We explain what convergence means for a sequence of stochastic processes. We talk about two tools to prove convergence of stochastic processes: the generator and couplings. Finally, we discuss the notion of duality of Markov processes.

1.2.1 Convergence of stochastic processes

Let us first give a general definition of a stochastic process

Definition 1.2.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let I be some index set. For each $i \in I$ let $X_i : \Omega \to E$ be a random variable. Then $\{X_i\}_{i\in I}$ is an E valued **stochastic process**. If $I = \mathbb{R}^+$ we say that $\{X_i\}_{i\in I}$ is a **continuous time stochastic process**, and if $I = \mathbb{N}$ we say that $\{X_i\}_{i\in I}$ is a **discrete time stochastic process**.*

A crucial concept in this thesis is convergence of stochastic processes. Our approach will be to think a stochastic process as a probability measure over a path space. The gain is that we will be able to understand convergence of stochastic processes, by regarding it as weak convergence of probability measures over a path space.

We should first precisely define what we mean by a path space.

Definition 1.2.2. *For any $T \in \mathbb{R}^+$ and for any complete and separable metric space E, we say that a function $f : [0, T] \to E$ is càdlàg, if f is continuous from the right with limit from the left. We define the set $M = M(T, E)$ to be the set of càdlàg functions, this is $M = \{f : [0, T] \to E : f \text{ is càdlàg}\}$.*

In order to define probability measures on M, we need to construct a suitable σ algebra. We will introduce a metric in the space M, which is known as the Skorohod M_2 distance [65]. We will make the definition for one dimensional path spaces ($f : [0,T] \to \mathbb{R}$), for a general definition the reader is invited to consult Chapter 13 of Whitt [73]. The analogous definition for discrete and countable spaces follows immediately. For example considering $f : [0,T] \to \mathbb{N}$ or $f : [0,T] \to [n]$ for some $n \in \mathbb{N}$. In the case of discrete spaces the Skorohod M_2 distance is equivalent to the uniform topology.

Definition 1.2.3 (M_2 Skorohod distance)**.** *Let $f_i : [0,T] \to \mathbb{R}$ be a càdlàg function, for $i = \{1,2\}$ and $T \in \mathbb{R}$. Let $\Gamma_i \subseteq \mathbb{R}^2$ be the continuous graph of f_i, defined by*

$$\Gamma_i := \{(x,y) \in \mathbb{R}^2 : (x,y) = (af_i(t^-) + (1-a)f_i(t), t) \text{ for some } t \in [0,T], a \in [0,1]\}$$

The M_2 Skorohod distance of f_1 and f_2 is the Hausdorff distance of Γ_1 and Γ_2, this is

$$d_M(f_1,f_2) = \max_{i,j\in\{1,2\}} \sup_{(x_i,y_i)\in\Gamma_i} \inf_{(x_j,y_j)\in\Gamma_j} \{d\big((x_i,y_i),(x_j,y_j)\big)\}$$

where $d\big(\cdot,\cdot\big)$ is its Euclidean distance in $\mathbb{R}^2$.

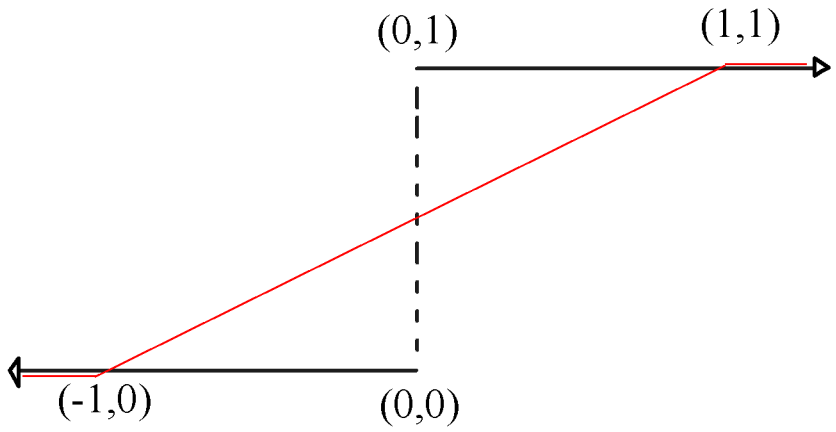

Figure 1.6: The black lines are the graph of the function $f(x) = 1_{\{x\geq 0\}}$. The black lines toghether with the dotted line are the continuous graph of $f(x) = 1_{\{x\geq 0\}}$. The thin line is the graph and the continuous graph of the function $g = 1_{\{x\in[-1,1]\}}(x/2 + 1/2) + 1_{\{x>1\}}$. One can see that $d_M(f,g) < 1/2$.

Remark 1.2.4. The M_2 Skorohod distance is not the most commonly used distance in path spaces. However, it has the advantage of having a nice relation to the Hausdorff metric, which in my opinion makes it simpler. The most popular distance is the J_1 Skorohod distance, also introduced in [65]. Most convergence results in the literature are stated for the J_1 distance. However, this is not a problem as convergence in J_1 implies convergence in M_2. One needs to keep in mind that convergence in M_2 allows that sequences of functions with arc length going to infinity converge to functions with finite arc length. In some applications one does not want this to happen, but in the examples discussed in this thesis this is not an issue.

We can now define our space of stochastic processes

Definition 1.2.5. *Let $\mathcal{M}$ be the Borel σ algebra, generated by the open sets of the metric space (M,d_M). We say that an E-valued stochastic process X_t is a càdlàg process if for all $x \in E$, $\mathbb{P}((X_t) \in \cdot | X_0 = x) =$*

$\mathbb{P}_x((X_t) \in \cdot)$ *is a probability measure on* $(M, \mathcal{M})$. *We say that a sequence of càdlàg stochastic processes* $\{(X_t^N)\}_{N\in\mathbb{N}}$ *converges, as* N *goes to infinity, weakly over the space of Skorohod to a stochastic process* (X_t), *and we write* $(X_t^N) \Rightarrow (X_t)$, *if for all continuous and bounded functions* $f : M \to \mathbb{R}$

$$\lim_{n\to\infty} \int_M f d\,\mathbb{P}((X_t^N) \in \cdot) = \int_M f d\,\mathbb{P}((X_t) \in \cdot).$$

We will denote by $\mathbb{S}$ *the space of càdlàg stochastic processes with the topology induced by the weak convergence.*

Remark 1.2.6. It turns out that $\mathbb{S}$ is a complete and separable metric space (See page 381 of [73]).

To really use this definition one would need to understand which functions $f : M \to \mathbb{R}$ are continuous with respect to the metric d_M. This looks like a hard task, but luckily Prohorov ([61]) found a metric that allow us to efficiently understand which sequences of stochastic processes converge weakly.

Definition 1.2.7. *Let* $A \in \mathcal{M}$, *for any* $\epsilon > 0$ *we define the open* ϵ*-neighborhood of* A *to be the set*

$$A^\epsilon = \{x \in M : d_M(x, y) < \epsilon, \text{ for some } y \in A\}.$$

Let O *be the set of all open sets in* $\mathcal{M}$. *Let* (X_t) *and* (Y_t) *be two càdlàg stochastic processes. We define the* ***Prohorov distance*** *between* (X_t) *and* (Y_t) *to be*

$$\rho\Big((X_t), (Y_t)\Big) = \inf\{\epsilon : \mathbb{P}\big((X_t) \in A\big) \leq \mathbb{P}\big((Y_t) \in A^\epsilon\big) + \epsilon, \forall A \in O\}.$$

It turns out that $\rho(\cdot, \cdot)$ is indeed a metric and convergence under $\rho(\cdot, \cdot)$ is equivalent to weak convergence. A proof of this can be found in section 11.3 of [73]. Now we write a precise statement of this fact.

Theorem 1.2.8. *If* (E, d) *is a complete and separable metric space, then the space of laws of càdlàg stochastic processes with values in* (E, d) *together with the Prohorov metric, that we denote* $(\mathbb{S}, \rho)$, *is a complete and separable metric space, and* $\rho\big((X_t^N), (X_t)\big) \to 0$ *if and only if* $(X_t^N) \Rightarrow (X_t)$ *weakly.*

The structure the space of *càdlàg* paths allows us to use a less strict notion of convergence, this is convergences of the finite dimensional distributions. Indeed, if we consider $(X_t) \in \mathbb{S}$, and we fix $t \in \mathbb{R}^+$, then X_t^N is just an E valued random variable and weak convergence of E valued random variables is the well known the convergence in distribution that is often taught in undergraduate courses (at least for $E = \mathbb{R}$). Moreover, if we consider a finite amount of fixed time points, $\{t_1, ..., t_d\}$ for some $d \in \mathbb{N}$, the random vector $(X_{t_1}^N, ..., X_{t_d}^N)$ is a random variable with values in E^d, and again we have a good grasp of what it means that a sequence of such random variables converges. The intuition suggests, that weak convergence of stochastic process is similar to convergence of random vectors $(X_{t_1}^N, ..., X_{t_d}^N)$ for a sufficiently large collection of time points.

Definition 1.2.9. *We say that a sequence of stochastic processes* $\{(X_t^N)\}_{N\in\mathbb{N}}$ *converges to* (X_t) *in the sense of convergence of the finite dimensional distributions, and we write* $(X_t^N) \rightrightarrows (X_t)$, *if for any finite subset* $\{t_1, ..., t_d\}$ *of a dense set* $D \in I$, *the following convergence in distribution holds*

$$(X_{t_1}^N, ..., X_{t_d}^N) \to (X_{t_1}, ..., X_{t_d}).$$

Convergence in the finite dimensional sense does not implies weak convergence in the space of Skorohod (for a counterexample see Example 11.6.1 of [73]). However, if a sequence of stochastic processes is relatively compact and converges in the finite dimensional sense, then it converges weakly over the space of Skorohod. The following Theorem makes this statement precise and its proof can be found in Chapter 4 of Whitt [73].

Theorem 1.2.10. *A sequence of processes* $\{(X_t^N)\}_{N\in\mathbb{N}}$ *converges weakly over the space of Skorohod to* (X_t) *if* $(X_t^N) \rightrightarrows (X_t)$ *and* $\{(X_t^N)\}_{N\in\mathbb{N}}$ *is relatively compact.*

So far we have developed a clear notion of convergence of stochastic processes, but we need to construct appropriate tools to be able to prove such convergence. This will be our task in the next subsections.

1.2.2 The generator of a stochastic process

In this subsection we define the semigroup of a Markov process and its generator. We discuss that in certain cases, convergence of a sequence of generators implies convergence of stochastic processes. We provide some interesting examples. The main reference for this subsection is [18].

A stochastic process is a very complex mathematical object. Our goal in this section will be to show how to construct an operator (which is a much simpler mathematical object) that captures the essence of a stochastic process. This technique works appropriately for Markov processes under mild conditions. Let (X_t) be a Markov process realized in a filtered probability space $(\Omega, \mathbb{F}, \mathbb{F}_t, \mathbb{P})$ with values is $(\mathbb{R}, \mathbb{B})$. We well write $\mathbb{P}_x(X_t \in \cdot) := \mathbb{P}(X_t \in \cdot | X_0 = x)$ and $\mathbb{E}_x(X_t) := \mathbb{E}(X_t | X_0 = x)$. We will start by associating a semigroup of operators to each Markov process.

Definition 1.2.11. *The semigroup of operators associated to a Markov process $(X_t)_{t\in\mathbb{R}^+}$, that we denote by $\{P_t\}_{t\in\mathbb{R}^+}$, is defined by the formula*

$$P_t f(x) = \mathbb{E}_x[f(X_t)]$$

for any $x \in \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$ measurable.

The generator of a Markov process is simply the generator of its semigroup of operators.

Definition 1.2.12. *The generator of a Markov process (X_t) is defined as the L_1 limit*

$$Af(x) = \lim_{t\to 0} \frac{\mathbb{E}_x[f(X_t) - f(x)]}{t}.$$

The set of functions such that Af exists is called the domain of the generator and we denote it by $D(A)$.

The generator is a very useful object to study Markov processes, as it is only one operator, but contains a summary of a Markov process. Heuristically, if one has a sequence of Markov processes $\{(X_t^N)\}_{N\in\mathbb{N}}$, with generators $\{A^N\}_{N\in\mathbb{N}}$, then if for every function $f : \mathbb{R} \to \mathbb{R}$ in a large enough class of functions $\mathcal{C}$, such that $\mathcal{C} \subseteq D(A^N)$ for all $N \in \mathbb{N}$, the sequence of functions $\{A^N f(x)\}_{N\in\mathbb{N}}$ converges uniformly to a function $Af(x)$, and A happens to be the generator of a Markov process (X_t), then $(X_t^N) \Rightarrow (X_t)$. For a precise statement of this claim and its proof see Chapter 2 and 4.8 of [18]. The intuition behind this useful result is that (under some mild conditions) convergence of the generators implies convergence of the finite dimensional distributions, and relative compactness as well. This allow us to apply Theorem 1.2.10 and obtain weak convergence over the space of Skorohod.

Let us start by calculating the generator of a Poisson process.

Proposition 1.2.13. *The generator A of a Poisson process (W_{ct}) (defined in 1.1.35) with rate c, applied to a polynomial $f : \mathbb{N} \mapsto \mathbb{R}$ is the operator such that for any $n \in \mathbb{N}$*

$$Af(n) = c(f(n+1) - f(n)).$$

Its domain $D(A)$ contains all bounded functions $f : \mathbb{N} \to \mathbb{R}$.

Proof. The proof consists of applying the definitions 1.2.12 and 1.1.35. Indeed,

$$\begin{aligned}
Af(n) &= \lim_{t\to 0} \frac{\mathbb{E}_n[f(W_{ct}) - f(n)]}{t} \\
&= \lim_{t\to 0} [f(n+1) - f(n)] \frac{1}{t} \mathbb{P}_n(W_{ct} = n+1) \\
&= \lim_{t\to 0} [f(n+1) - f(n)] \frac{1}{t} \mathbb{P}_0(W_{ct} = 1) \\
&= [f(n+1) - f(n)] \lim_{t\to 0} \frac{1 - e^{-ct}}{t} \\
&= [f(n+1) - f(n)]c.
\end{aligned}$$

L_1 convergence follows from the previous calculation given that f is bounded. □

Remark 1.2.14. It is very useful to *read* the generator: we have $Af(n) = c(f(n+1) - f(n))$, and we interpret it as: the process N_t goes from the state n to the state $n+1$ at rate c

We can go from the standard Poisson process (W_t) to a Poisson process with rate c by considering a constant time change (W_{ct}). Instead of using a constant to change the time of our Poisson process, we can use a function of the state of the process. Let $g : \mathbb{N} \to \mathbb{R}^+$ be measurable, we will consider the process $M_t = W_{\int_0^t g(M_s)ds}$. This is a well defined process, that can be simply constructed as the Markov process with values in $\mathbb{N}$, that goes from the state n to the state $n+1$ at rate $g(n)$.

Remark 1.2.15. The generator A of the process $M_t = W_{\int_0^t g(M_s)ds}$ is such that, applied to any polynomial $f : \mathbb{N} \to \mathbb{R}$,

$$Af(n) = g(n)[f(n+1) - f(n)].$$

The interpretation of this generator is simple, the process goes from the state n to the state $n+1$ at rate $g(n)$.

We can construct our favorite process using this time change technique. Indeed, let $n_0 \in \mathbb{N}$,

$$|\overline{K}_t| = n_0 - W_{\int_0^t \binom{|K_s|}{2} ds},$$

then $(|\overline{K}_t|)$ is equal in distribution to the block counting process of the Kingman coalescent $(|K_t|)$ started with n_0 blocks, introduced in Definition 1.1.25.

Example 1.2.1. The generator of the block counting process of the Kingman coalescent is such that for every function $f : \mathbb{N} \mapsto \mathbb{R}$ and every $n \in \mathbb{N}$

$$Af(n) = \binom{n}{2}[f(n-1) - f(n)].$$

Now we give a notion of generator for a discrete process.

Definition 1.2.16. *Let $(X_i)_{i\in\mathbb{N}}$ be a Markov chain with values on a discrete space E. The discrete generator of (X_i) is the generator of the continuous time process $(X_{W_t})_{t\in\mathbb{R}}$, where (W_t) is a standard Poisson process. The discrete generator of $(X_{\lfloor ct \rfloor})_{t\in\mathbb{R}^+}$ is defined to be the generator of $(X_{W_{ct}})$.*

Lemma 1.2.17. *Let $\{p_{n,j}\}_{n,j\in E}$ be the transition matrix of a Markov chain $(X_i)_{i\in\mathbb{N}}$, defined as in 1.2.16. The discrete generator of $(X_{\lfloor ct \rfloor})_{t\in\mathbb{R}^+}$ is the operator such that for any bounded and measurable function $f : E \to \mathbb{R}$ and any $n \in E$*

$$Af(n) = c\sum_{j\in E} p_{n,j}[f(j) - f(n)].$$

Proof. The proof consists of applying Definition 1.2.16.

$$\begin{aligned}
Af(n) &= \lim_{t\to 0} \frac{\mathbb{E}_n[f(X_{W_t}) - f(n)]}{t} \\
&= \lim_{t\to 0} \sum_{j\in E} (f(j) - f(n)) \frac{1}{t} \mathbb{P}_n(X_{W_t} = j) \\
&= \lim_{t\to 0} \sum_{j\in E} (f(j) - f(n)) \frac{1}{t} \mathbb{P}_n(X_1 = j)\mathbb{P}_0(W_t = 1) \\
&= \sum_{j\in E} (f(j) - f(n)) p_{n,j} \lim_{t\to 0} \frac{1 - e^{-ct}}{t} \\
&= c\sum_{j\in E} (f(j) - f(n)) p_{n,j}
\end{aligned}$$

Again, L_1 convergence follows from pointwise convergence as f is bounded.

□

Using the same technique we can characterize the generator of a general continuous-time discrete-space Markov process

Lemma 1.2.18. *Let $(Z_t)_{t\in\mathbb{R}^+}$ be a continuous time discrete space Markov process with countable state space E and such that $r_{nj} \in \mathbb{R}$ is the transition rate form the state $n \in E$ to the state $j \in E$. The generator of $(Z_t)_{t\in\mathbb{R}^+}$ is the operator such that for any $f : E \to \mathbb{R}$ and $n \in E$*

$$Af(n) = \sum_{j\in E} r_{nj}[f(j) - f(n)]. \tag{1.2.1}$$

Proof. Let

$$p_{nj} := \frac{r_{nj}}{\sum_{j\in E} r_{nj}}$$

and consider the function $g : E \to \mathbb{R}^+$, defined by the formula

$$g(n) = \sum_{j\in E} r_{nj}.$$

First note that $\{p_{nj}\}_{n,j\in E}$ is a stochastic matrix, because

$$\sum_{j\in E} p_{nj} = \sum_{j\in E} \frac{r_{nj}}{\sum_{j\in E} r_{nj}} = 1.$$

Then, there exists a discrete time Markov chain $(X_i)_{i\in\mathbb{N}}$ with state space E and transition matrix $\{p_{nj}\}_{n,j\in E}$. Let $Y_t = X_{W_{\int_0^t g(Y_s)ds}}$. Note that in distribution $(Y_t) = (Z_t)$, so they have the same generator.

The rest of the proof consist of applying Definition 1.2.12 to (Y_t).

$$\begin{aligned} Af(n) &= \lim_{t\to 0} \frac{\mathbb{E}_n[f(Y_t) - f(n)]}{t} \\ &= \sum_{j\in E} [f(j) - f(n)] p_{nj} g(n) \\ &= \sum_{j\in E} [f(j) - f(n)] r_{nj}. \end{aligned}$$

□

Our next task is to study the generator of some continuous time real valued processes with continuous trajectories. For this we follow Chapter 7.1 in [16]. We will define a class of processes called **diffusions** using the notion of generator and later we will make intuitive sense of these objects.

Definition 1.2.19. *A one dimensional diffusion process is a continuous Markov process with values in $\mathbb{R}$ and infinitesimal generator A, such that for every bounded and two times differentiable function $f : \mathbb{R} \mapsto \mathbb{R}$ and every point $x \in \mathbb{R}$,*

$$Af(x) = a(x)\frac{\partial}{\partial x} f(x) + \frac{1}{2} b(x) \frac{\partial^2}{\partial x^2} f(x),$$

where $a : \mathbb{R} \mapsto \mathbb{R}$ and $b : \mathbb{R} \mapsto \mathbb{R}$.

Remark 1.2.20. It is not the case that for every $a : \mathbb{R} \to \mathbb{R}$ and $b : \mathbb{R} \to \mathbb{R}$ there exists a diffusion process. For a deeper study of diffusion theory the reader is referred to [62].

The function a is known as the drift and function b is known as the diffusivity of the diffusion. Let us gain some intuition, by studying some examples.

Example 1.2.2. Let X_t be the solution of the ordinary differential equation

$$\frac{d}{dt} X_t = a(X_t) \tag{1.2.2}$$

Assume that a is continuous and such that the ODE has a unique solution. Let us calculate the generator of X_t using Definition 1.2.12. Let f be bounded and twice differentiable and $x \in \mathbb{R}$.

$$\begin{aligned} Af(x) &= \lim_{t\to 0} \frac{\mathbb{E}_x[f(X_t) - f(x)]}{t} \\ &= \lim_{t\to 0} \frac{\mathbb{E}_x[\int_0^t \frac{\partial}{\partial s} f(X_s) ds]}{t} \\ &= \lim_{t\to 0} \frac{\mathbb{E}_x[\int_0^t f'(X_s) \frac{\partial}{\partial s} X_s ds]}{t} \\ &= \lim_{t\to 0} \frac{\mathbb{E}_x[\int_0^t f'(X_s) a(X_s)]}{t} \\ &= f'(x)a(x). \end{aligned}$$

In the second equality we used the Fundamental Theorem of Calculus and in the third the chain rule. In the last equality we used the continuity at x of a and and of f' at the point x. A consequence of this example is that, provided that the ODE in Equation (1.2.2) has a solution, the diffusion with drift term $a(\cdot)$ and no diffusivity, is a deterministic trajectory.

Now we want to know which process has the generator $Af(x) = \frac{1}{2}f''(x)$. Let (S_t) be a simple symmetric random walk. For every fixed $N \in \mathbb{N}$ let $(S_t^N) := (\frac{1}{N} S_{\lfloor N^2 t \rfloor})$. Let A^N be the Generator of (S_t^N). Let f be a bounded a twice differentiable function and $x \in \mathbb{N}/N$. Using Lemma 1.2.17, and then applying Taylor expansion we observe

$$\begin{aligned} A^N f(x) &= N^2[1/2 f(x + 1/N) + 1/2 f(x - 1/N) - f(x)] \\ &= N^2 \left[\frac{1}{2} \sum_{i=0}^{2} (1/N)^i \frac{f^{(i)}(x)}{i!} + \frac{1}{2} \sum_{i=0}^{2} (-1/N)^i \frac{f^{(i)}(x)}{i!} - f(x) \right] + O(1/N) \\ &= \frac{1}{2} f''(x) + O(1/N), \end{aligned}$$

where $f^{(i)}$ is the i-th derivative of f. Defining the operator A to be such that $Af(x) = \frac{1}{2}f''(x)$, we conclude that for all bounded a twice differentiable $f : \mathbb{R} \mapsto \mathbb{R}$, and for every $x \in \mathbb{R}$

$$A^N f(x) \to Af(x).$$

in L_1. On one hand it can be checked that the sequence of generators $\{A^N\}_{N\in\mathbb{N}}$ fulfills the technical requirements to ensure that the convergence of the generator implies the weak convergence over the space of Skorohod $(S_t^N) \Rightarrow (B_t)$. For example, Theorem 4.8.2 in [18] can be applied. On the other hand it is known (using the Lévy construction of Brownian motion or the Donsker invariance principle) that $(S_t^N) \Rightarrow (B_t)$, where (B_t) is the Brownian motion. With a bit more work this intuitive argument can be turned into a proof that the Brownian motion has generator A characterized by the formula $Af(x) = \frac{1}{2}f''(x)$.

Indeed, one can think of a diffusion process as an ODE perturbed by a time-change of a Brownian motion. The most common way to define a diffusion is as the solution (if it exists) to the stochastic differential equation

$$X_t = X_0 + \int_0^t a(X_s) ds + \int_0^t b^2(X_s) dB_t. \tag{1.2.3}$$

To learn more about SDEs see [15]. One can verify that that the solution to Equation 1.2.3 (if it exists) is continuous and has generator

$$Af(x) = a(x) \frac{\partial}{\partial x} f(x) + \frac{1}{2} b(x) \frac{\partial^2}{\partial x^2} f(x)$$

as in Definition 1.2.19.

Now, the most important diffusion in Population Genetics is the Wright Fisher diffusion, which is the solution to the SDE, $X_0 \in [0, 1]$ and

$$dX_t = \sqrt{X_t(1 - X_t)} dB_t.$$

The generator A of (X_t) is such that for every continuous and twice differentiable function $f \in D(A)$ and every $x \in [0,1]$,

$$Af(x) = \frac{1}{2}x(1-x)\frac{\partial}{\partial x^2}f(x). \tag{1.2.4}$$

Remark 1.2.21. When we work with $\mathbb{R}$ valued Markov processes, it is convenient to consider functions f to be bounded and twice differentiable. Unbounded functions are frequently outside the domain of $\mathbb{R}$ valued Markov processes. When we work with Markov processes with values in some compact subset of $\mathbb{R}$, it is often convenient to take f to be a polynomial. In these cases. the domain includes all polynomials and they are a dense class of functions.

Recall the frequency process of the Moran model, $(M^N_{N^2t})_{t\in\mathbb{R}^+}$, defined in 1.1.36. Let A^N be the generator of $(M^N_{N^2t})_{t\in\mathbb{R}^+}$. Then, by Lemma 1.2.18, for any polynomial $f : [0,1] \mapsto \mathbb{R}$ and $x \in \mathbb{R}$,

$$\begin{aligned} A^N f(x) &= x(1-x)N^2[1/2f(x+1/N) + 1/2f(x-1/N) - f(x)] \\ &= x(1-x)\frac{1}{2}f''(x) + O(1/N) \end{aligned} \tag{1.2.5}$$

Note that the error term $O(1/N)$ is uniformly bounded in x. Comparing Equation (1.2.4) and Equation (1.2.5) (and verifying some technical conditions) allow us to apply Theorem 4.8.2 in [18] and conclude that, as $N \to \infty$

$$(M^N_t) \Rightarrow (X_t).$$

1.2.3 Couplings and weak convergence of stochastic processes

We define the notion of a coupling of random variables, and stochastic processes. We define convergence in probability of a sequence of stochastic processes. We show that convergence in probability implies weak convergence of stochastic processes. The main reference for this subsection is [43]

Definition 1.2.22. *A coupling of two random variables X, Y, with values on E and defined in the probability spaces $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ is a pair of random variables X' and Y' defined in the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that, in distribution, $X = X'$ and $Y = Y'$. The pair (X', Y') is a $E \times E$ valued random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$.*
A coupling of two probability measures μ_1, μ_2 is a coupling of two random variables X, Y such that the distribution of X is μ_1 and the distribution of Y is μ_2.

Let $(X_i)_{i\in I}$ and $(Y_i)_{i\in I}$ be two stochastic processes with values in the same metric space (E, d). What is $\mathbb{P}(d(X_i, Y_i) < \epsilon)$? The problem with this question is that in principle it does not make any sense: as the two processes might be realized in different probability spaces. $\mathbb{P}$ could be the product measure, but it could also be something else. This leads us to the definition of coupling for stochastic processes.

Definition 1.2.23. *A coupling of two stochastic processes $(X_i)_{i\in I}$ and $(Y_i)_{i\in I}$ is a pair of stochastic processes $(X'_i)_{i\in I}$ and $(Y'_i)_{i\in I}$ defined in the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that, in distribution, $(X_i) = (X'_i)$ and $(Y_i) = (Y'_i)$.*

We can use the concept of couplings to prove weak convergence of stochastic processes, without relying on the Markov property. Here we stress that the method discussed in the previous subsection (the generator) relies on the Markov property. Coupling will be a crucial tool in Chapter 2, where we deal with non-Markovian processes.

Example 1.2.3. Let $(X_i)_{i\in\mathbb{N}}$ and $(Y_i)_{i\in\mathbb{N}}$ be two Markov processes with values in E, realized on Ω_1 and Ω_2 respectively and with the same transition probabilities. Assume than $X_0 \sim \gamma_1$ and $Y_0 \sim \gamma_2$, where γ_1, γ_2 are probability measures in E. Let $\tau_{coup} = \inf\{i \in \mathbb{N} : X_i = Y_i\}$. Define

$$Y'_i = \begin{cases} Y_i & \text{for all } i \leq \tau_{coup} \\ X_i & \text{for all } i \geq \tau_{coup}. \end{cases} \tag{1.2.6}$$

The pair $(X_i)_{i\in\mathbb{N}}$ and $(Y'_i)_{i\in\mathbb{N}}$ is a coupling of $(X_i)_{i\in\mathbb{N}}$ and $(Y_i)_{i\in\mathbb{N}}$ realized on the product space $\Omega_1 \times \Omega_2$ (Note that τ_{coup} is a stopping time in the product σ-algebra). Note that $\mathbb{P}(X_i \neq Y'_i) = \mathbb{P}(\tau_{coup} > i)$. This coupling is very useful, and it is know as Doeblin coupling [13].

Definition 1.2.24. *We say that a sequence of stochastic processes, $\{(X_i^N)\}_{N\in\mathbb{N}}$, with values in a Metric space (E,d), converges in probability to a process (X_i) with values in the same metric space, if for every $N\in\mathbb{N}$ there exists a coupling $((X_i^{N\prime})_{i\in I},(X_i')_{i\in I})$ and it is such that for every $\epsilon>0$*

$$\lim_{N\to\infty}\mathbb{P}(d_M\big((X_i^{N\prime}),(X_i')\big)>\epsilon)=0.$$

where d_M is the Skorohod M_2 distance.

Lemma 1.2.25. *If $\{(X_i^N)\}_{N\in\mathbb{N}}$ converges in probability to (X_i), then $\{(X_i^N)\}_{N\in\mathbb{N}}$ converges weakly in the space of Skorohod equipped with the M_2 metric to (X_i).*

Proof. To prove this Lemma we need to apply Theorem 1.2.8. We will show that for any fixed $\delta>0$, $\lim_{N\to\infty}\rho\big(\mathbb{P}((X_i^N)\in\cdot),\mathbb{P}((X_i)\in\cdot)\big)<\delta$. Recall that $\rho(\cdot,\cdot)$ is the Prohorov distance (see Definition 1.2.7).

The convergence in probability of $\{(X_i^N)\}_{N\in\mathbb{N}}$ to (X_i) implies that there exists a coupling $((X_i^{N\prime}),(X_i'))$ of (X_i^N) and (X_i), such that for all $\delta>0$, there exists $N_\delta\in\mathbb{N}$ and for every $N>N_\delta\in\mathbb{N}$,

$$\mathbb{P}(d_M\big((X_i^{N\prime}),(X_i')\big)>\delta)<\delta.$$

Let $A\in\mathcal{M}$ be open ($\mathcal{M}$ was defined in 1.2.5) and let $N>N_\delta$ (recall the Definition of A^δ introduced in 1.2.7). Then,

$$\begin{aligned}\mathbb{P}((X_i^N)\in A)&=\mathbb{P}((X_i^{N\prime})\in A)\\&\le\mathbb{P}((X_i')\in A^\delta)+\delta\\&=\mathbb{P}((X_i)\in A^\delta)+\delta.\end{aligned}$$

This implies that

$$\lim_{N\to\infty}\rho\big(\mathbb{P}((X_i^N)\in\cdot),\mathbb{P}((X_i)\in\cdot)\big)=\inf\{\epsilon:\mathbb{P}\big((X_t)\in A\big)\le\mathbb{P}\big((Y_t)\in A^\epsilon\big)+\epsilon,\forall A\in O\}<\delta$$

where O is the set of open elements of $\mathcal{M}$. □

The following Lemma will be useful in the proof and the applications of Theorem 2.3.3.

Lemma 1.2.26. *Let $\{(\overline{L}_i^N)_{i\in I}\}_{N\in\mathbb{N}}$ and $\{(L_i^N)_{i\in I}\}_{N\in\mathbb{N}}$ be two sequences of stochastic processes realized in the same probability space Ω with values in the same metric space E. If $\{(L_i^N)_{i\in I}\}_{N\in\mathbb{N}}$ converges weakly over the space of Skorohod to $(X_i)_{i\in I}$ and $\lim_{N\to\infty}\mathbb{P}\big((L_i^N)=(\overline{L}_i^N),\forall i\in I\big)=1$, then $\{(\overline{L}_i^N)\}_{N\in\mathbb{N}}$ converges to (X_i) weakly over the space of Skorohod .*

Proof. Let $\epsilon>0$. Consider $N_\epsilon\in\mathbb{N}$ such that for all $N>N_\epsilon$, $\rho\big(\mathbb{P}((L_i^N)\in\cdot),\mathbb{P}((X_i)\in\cdot)\big)<\epsilon/2$ and $\mathbb{P}\big((L_i^N)=(\overline{L}_i^N),\forall i\in\mathbb{N}\big)>1-\epsilon/2$. Let $A\in\mathcal{M}$ be open. Then

$$P((X_i)\in A)\le\mathbb{P}((L_i^N)\in A^{\epsilon/2})+\epsilon/2\le\mathbb{P}((\overline{L}_i^N)\in A^{\epsilon/2})+\epsilon,$$

then we conclude that $\rho\big(\mathbb{P}((\overline{L}_i^N)\in\cdot),\mathbb{P}((X_i)\in\cdot)\big)<\epsilon$. □

1.2.4 Couplings, stationary distribution and mixing time

There are ways to quantify the distance between two probability measures, meaning how similar they are. For example, the Prohorov distance and the Total Variation distance. We will explain the second example, because it will be relevant in Chapter 2. The main reference for this subsection is [43]

Definition 1.2.27. *Let μ_1 and μ_2 be two probability measures on the same measurable space $(\Omega,\mathcal{F})$. We define their **total variation distance**, $\|\mu_1-\mu_2\|_{TV}$, by the formula*

$$\|\mu_1-\mu_2\|_{TV}=\sup_{A\in\mathcal{F}}|\mu_1(A)-\mu_2(A)|.$$

This notion of distance of probability measures allows us to understand better a very relevant object in the study of Markov processes, the stationary distribution.

Definition 1.2.28. *A stationary distribution of a stochastic process with values in E, $(X_i)_{i\in I}$ is a probability distribution ν on E, such that for every open set $A \subseteq E$, and for every $i \in \mathbb{N}$*

$$\nu(A) = \mathbb{P}_\nu(X_i \in A).$$

The stationary distribution is very useful, mostly due to the following theorem. We will denote $\mathcal{P}(E)$ the set of all probability measures in over E.

Theorem 1.2.29. *Suppose that $(X_i)_{i\in I}$ is an irreducible and aperiodic Markov chain with finite state space E, then there exist a unique stationary distribution $\nu \in \mathcal{P}(E)$, such that for any initial distribution $\gamma \in \mathcal{P}(E)$,*

$$\lim_{i\to\infty} ||\mathbb{P}_\gamma(X_i \in \cdot) - \nu||_{TV} = 0.$$

A proof of this Theorem can be found in Chapter 4.3 of [43]. A probabilistic proof of this claim was made by Doeblin, and it is based on the Doeblin coupling introduced in Example 1.2.3. This approach is related with the coupling interpretation of the total variation distance, and the optimal coupling, that we now discuss.

Lemma 1.2.30. *Let $\mu_1, \mu_2 \in \mathcal{P}(E)$ be two probability measures on a discrete space E. Then*

$$||\mu_1 - \mu_2||_{TV} = \inf_{(X_1',X_2') \text{ is a coupling of } \mu_1 \text{ and } \mu_2} \mathbb{P}(X_1' \neq X_2').$$

Furthermore, there exists a coupling (X_1, X_2), that we call ***the optimal coupling****, such that*

$$||\mu_1 - \mu_2||_{TV} = \mathbb{P}(X_1 \neq X_2).$$

Proof. First note that for any subset $A \subseteq E$, and any coupling (X_1', X_2') of μ_1 and μ_2, it holds that

$$\begin{aligned} |\mu_1(A) - \mu_2(A)| &= |\mathbb{P}(X_1' \in A) - \mathbb{P}(X_2' \in A)| \\ &\leq \mathbb{P}(X_1' \in A, X_2' \in E/A) + \mathbb{P}(X_2' \in A, X_1' \in E/A) \\ &\leq \mathbb{P}(X_1' \neq X_2') \end{aligned}$$

From this it is immediate that

$$||\mu_1 - \mu_2||_{TV} \leq \inf\{\mathbb{P}(X_1' \neq X_2' : (X_1', X_2')) \text{ is a coupling of } \mu_1 \text{ and } \mu_2\}.$$

To show that the equality can be realized, we construct the optimal coupling. Let $\mu_0 \in \mathcal{P}(E)$ be such that for every subset $A \subseteq E$,

$$\mu_0(A) = \frac{\min\{\mu_1(A), \mu_2(A)\}}{\sum_{x\in E} \min\{\mu_1(x), \mu_2(x)\}}.$$

Now, for $s = 1, 2$, let $\overline{\mu}_s$ be such that

$$\overline{\mu}_s(A) = \frac{\mu_s(A) - \min\{\mu_1(A), \mu_2(A)\}}{1 - \sum_{x\in E} \min\{\mu_1(x), \mu_2(x)\}}.$$

Note that $\overline{\mu}_s \in \mathcal{P}(E)$, because for every subset $A \subseteq E$, $\mu_s(A) - \min\{\mu_1(A), \mu_2(A)\} > 0$ and $\sum_{x\in E}(\mu_s(x) - \min\{\mu_1(x), \mu_2(x)\}) = 1 - \sum_{x\in E} \min\{\mu_1(x), \mu_2(x)\}$.

Consider the following independent random variables. Let Y_0 be a random variable with distribution μ_0, and for $s = 1, 2$ let Y_s be a random variable with distribution $\overline{\mu}_s$. Let W be a random variable with values on $\{0, 1\}$, such that $P(W = 1) = \sum_{x\in E} \min\{\mu_1(x), \mu_2(x)\}$. Note that, for $s = 1, 2$, the distribution of

$$X_s = WY_0 + (1 - W)Y_s$$

is μ_s. Then (X_1, X_2) is a coupling of μ_1 and μ_2, and

$$\mathbb{P}(X_1 \neq X_2) = \mathbb{P}(W = 0) = 1 - \sum_{x\in E} \min\{\mu_1(x), \mu_2(x)\} = ||\mu_1 - \mu_2||_{TV},$$

where the last equality follows by an algebraic manipulation (see Proposition 4.7 of [43].). □

Example 1.2.4. Let $(X_i)_{i\in\mathbb{N}}$ be an irreducible and aperiodic Markov process. If one uses the Doeblin coupling between the process (X_i) started in some point $x \in E$ and the same process started in the stationary distribution ν, (in the notation of Example 1.2.3) one observes that for every $i \in \mathbb{N}$

$$||\nu - \mathbb{P}_x(X_i \in \cdot)||_{TV} \leq \mathbb{P}(\tau_{coup} > i).$$

This is the observation that helped Doeblin to prove Theorem 1.2.29 using probabilistic arguments.

Theorem 1.2.29 implies that, after enough time, the distribution of any irreducible and aperiodic Markov chain with finite state space becomes similar to its stationary distribution. However, the time one needs to wait in order to be *close* to stationarity depends strongly on the Markov process. There is a very efficient way to quantify this.

Definition 1.2.31. *The **mixing time** of a Markov chain $(X_i)_{i\in\mathbb{N}}$ with values in E and unique invariant distribution $\nu \in \mathcal{P}(E)$ is defined as*

$$\tau_{mix} := \inf\left\{ i > 0 : \sup_{x\in E} \|\mathbb{P}_x(Y_i = \cdot\,) - \nu\|_{TV} \leq \frac{1}{4} \right\}. \tag{1.2.7}$$

The mixing time gives a very precise notion of how much time one needs to wait until the process is very close to stationarity.

Lemma 1.2.32. *For any $x \in E$, $l, s \in \mathbb{N}$ and any Markov chain $(X_i)_{i\in\mathbb{N}}$ that fulfills the assumptions of Theorem 1.2.29, it is true that*

$$\|\mathbb{P}_x(X_{l\tau_{mix}+s} \in \cdot) - \nu\|_{TV} \leq 2^{-l}.$$

A proof of this can be found in Section 4.5 of [43].

1.2.5 Duality of Markov processes

We define duality of Markov processes. We show how to use generators to prove duality. We introduce the moment duality, and explain its importance. The main reference of this subsection is [31].

Duality is a tool that allows us to obtain information about a Markov process using information that we have about an other Markov process. To be precise:

Definition 1.2.33. *Let $(X_t)_{t\in I}$ be a Markov process with values in E_1 and $(N_t)_{t\in I}$ be a Markov process with values in E_2. Let $H : E_1 \times E_2 \to \mathbb{R}$ be a function. We say that (X_t) **and** (N_t) **are dual with respect to** H if for all $t \in I$, $x \in E_1$ and $n \in E_2$, it is true that H is measurable and*

$$\mathbb{E}_x[H(X_t, n)] = \mathbb{E}_n[H(x, N_t)].$$

In practice an effective method to prove duality is by means of the generator. Indeed, we have the following useful result, its proof can be found in Proposition 1.2 in [31].

Proposition 1.2.34. *Let (X_t) and (Y_t) be Markov processes with state space E_1 and E_2 and generators A and $\overline{A}$ respectively. Let $H : E_1 \times E_2 \to \mathbb{R}$ be bounded and continuous. If $H(x,\cdot), P_t^1 H(x,\cdot) \in D(A)$ for all $x \in E_1$, $t > 0$ and $H(\cdot, n), P_t^2 H(\cdot, n) \in D(\overline{A})$ for all $n \in E_2$, $t > 0$. Then,*

$$AH(x, y) = \overline{A}H(x, y) \qquad \forall x \in E, y \in F, \tag{1.2.8}$$

if and only if (X_t) and (Y_t) are dual with respect to H.

A duality function H which is of great importance in population genetics is $H : [0,1] \times \mathbb{N} \to [0,1]$ defined by the formula

$$H(x, n) = x^n.$$

Definition 1.2.35. *Let $(X_t)_{t\in I}$ be a Markov process with values in $[0,1]$ and let $(N_t)_{t\in I}$ be a Markov process with values in $\mathbb{N}$. We say that (X_t) and (N_t) are **moment duals** if for every $t \in I$, $x \in [0,1]$ and $n \in \mathbb{N}$*

$$\mathbb{E}_x[X_t^n] = \mathbb{E}_n[x^{N_t}].$$

Remark 1.2.36. The function $H : [0,1] \times \mathbb{N} \mapsto \mathbb{R}$ defined by the formula $H(x,n) = x^n$, is bounded. Fixing n, it is an analytic function of x. Then, it can be verified for the examples that we will study in Section 1.3, that the conditions: $H(x,\cdot), P_t^1 H(x,\cdot) \in D(A)$ for all $x \in E_1$, $t > 0$ and $H(\cdot,n), P_t^2 H(\cdot,n) \in D(\overline{A})$ for all $n \in E_2$, $t > 0$ are satisfied. Later, we will only concentrate in proving Equation (1.2.8) to apply Proposition 1.2.34.

A classic example of moment duals are the Wright Fisher diffusion and the Kingman coalescent.

Theorem 1.2.37. *The Wright Fisher diffusion and the block counting process of the Kingman coalescent are moment duals.*

Proof. Let $H(x,n) = x^n$. Let $\overleftarrow{A}$ be the generator of the block counting process of the Kingman coalescent and $\overrightarrow{A}$ be the generator of the Wright Fisher diffusion. Using Example 1.2.1 we observe that

$$\begin{aligned}
\overleftarrow{A} H(x,n) &= \binom{n}{2}[H(x,n-1) - H(x,n)] \\
&= \frac{n(n-1)}{2}[x^{n-1} - x^n] \\
&= \frac{1}{2}x(1-x)n(n-1)x^{n-2} \\
&= \frac{1}{2}x(1-x)\frac{\partial^2}{\partial x \partial x}H(x,n).
\end{aligned}$$

The right hand side of the last equation is $\overrightarrow{A} H(x,n)$, the generator of the Wright Fisher diffusion applied to $H(\cdot,n) \in D(\overrightarrow{A})$ evaluated in $x \in [0,1]$ (see Equation (1.2.4)). □

1.3 Further evolutionary forces

We briefly discuss some generalizations of the Wright Fisher model. The main reference for this subsection is [17]

1.3.1 Mutation

We discuss a two type model with neutral mutation. We talk about duality. We also introduce the infinite sites model and the Tajima D. Mutation is a key ingredient in Chapter 4 and in [5].

There are many different approaches to generalize the Wright Fisher model to include mutation. In this introduction we will focus on two: the two types Wright Fisher frequency process with mutation, and the infinite site model.

In the two types model, a population consisting of two types evolves as in the Wright Fisher frequency process, but, contrary to the Wright Fisher model, not every individual has the same type as its parent. Most of the times this is the case, but with certain probability an offspring can be affected by a mutation. This causes that a parent and its offspring are of different types. In this model we study a diffusion approximation, as we did with the Wright Fisher frequency process.

The infinite sites model ([47],[72]) assumes that each mutation occurs in a different place. Then, in principle there can be more than two types. As time evolves the number of types increases. In many situations this model is appropriated to study real data.

Let us now study the two types Wright Fisher model with mutation. Assume that each generation consists of N individuals. Each individual chooses its parent from the previous generation uniformly at random. At generation zero, each individual is assigned a type, which can be either a or A. Let $\theta_1^N \in [0,1]$ and $\theta_2^N \in [0,1]$. The type of an individual which is the offspring of an individual of type a will be a with probability $1-\theta_2^N$ and will be A with probability θ_2^N (if affected by a mutation). Analogously, The type of an individual which is the offspring of an individual of type A will be A with probability $1-\theta_1^N$ and otherwise will be of type a. We observe that if at certain generation there are xN individuals of type a, the number of type a individuals at the following generation will be binomially distributed with parameters N and $x(1-\theta_2^N) + (1-x)\theta_1^N$. This leads to the following definition:

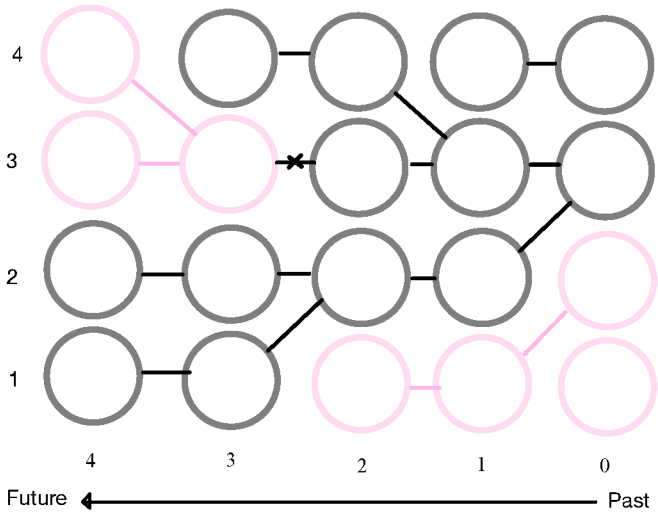

Figure 1.7: A realization of the two type Wright Fisher frequency process with mutation. In the figure, each individual inherits the type of its parent, except individual $(3,3)$ who has a different type than the type of its father, because it was affected by a mutation.

Definition 1.3.1. *The frequency process of type a individuals, in the two types Wright Fisher model with mutation, with population size $N \in \mathbb{N}$, mutation parameters $\theta_1^N \in [0,1]$ and $\theta_2^N \in [0,1]$, is the Markov chain $(X_i^N)_{i\in\mathbb{N}}$, with state space $\{0, 1/N, 2/N, ..., 1\}$ and transition probabilities*

$$p_{j,k} = \binom{N}{kN} \left(j(1-\theta_2^N) + (1-j)\theta_1^N\right)^{kN} \left(1 - j(1-\theta_2^N) - (1-j)\theta_1^N)\right)^{(1-k)N}$$

for any $j, k \in \{0, 1/N, 2/N, ..., 1\}$.

When the population is big, we can consider a diffusion approximation.

Proposition 1.3.2. *Assume that $\theta_1^N = \theta_1/N$ for some $\theta_1 \in \mathbb{R}^+$ and $\theta_2^N = \theta_2/N$ for some $\theta_2 \in \mathbb{R}^+$. Let (X_t) be the solution to the SDE*

$$dX_t = (\theta_1(1-X_t) - \theta_2 X_t)dt + \sqrt{X_t(1-X_t)}dB_t,$$

then $\{(X^N_{\lfloor Nt\rfloor})_{t\in\mathbb{R}^+}\}_{N\in\mathbb{N}}$ converges weakly over the space of Skorohod to $(X_t)_{t\in\mathbb{R}}$.

Proof. Let us calculate the discrete generator A^N of $(X^N_{\lfloor Nt\rfloor})$ applied to a polynomial $f : \mathbb{R} \mapsto \mathbb{R}$, at a point $x \in [0,1]$.

$$\begin{aligned} A^N f(x) &= N\mathbb{E}_x[f(X_1) - f(x)] \\ &= N\mathbb{E}_x[(X_1 - x)f'(x)] + N\mathbb{E}_x\left[\frac{(X_1-x)^2}{2}f''(x)\right] + O(1/N) \\ &= N(\mathbb{E}_x[X_1] - x)f'(x) + N\frac{1}{2}\mathrm{var}_x[X_1]f''(x) + O(1/N) \\ &= \theta_1^N(1-x)f'(x) - \theta_2^N x f'(x) + \frac{1}{2}x(1-x)f''(x) + O(1/N). \end{aligned}$$

The error term $O(1/N)$ is uniformly bounded on $x \in [0,1]$ as f is continuous. The third equality follows from the fact that all the higher moments are of smaller order (See Appendix A.2 for a detail manipulation of higher moments). □

In practice, the data that biologists extract from a population consist of DNA sequences. In mathematical language, an individual of a sample of a population which has r nucleotide in its DNA can be regarded as a point x in $\{g,a,t,c\}^r$. A sample of size n consists of n points $x_1,...,x_n \in \{g,a,t,c\}^r$. If we are given such sample, can we recover the underling coalescent structure? The answer is positive to some extent.

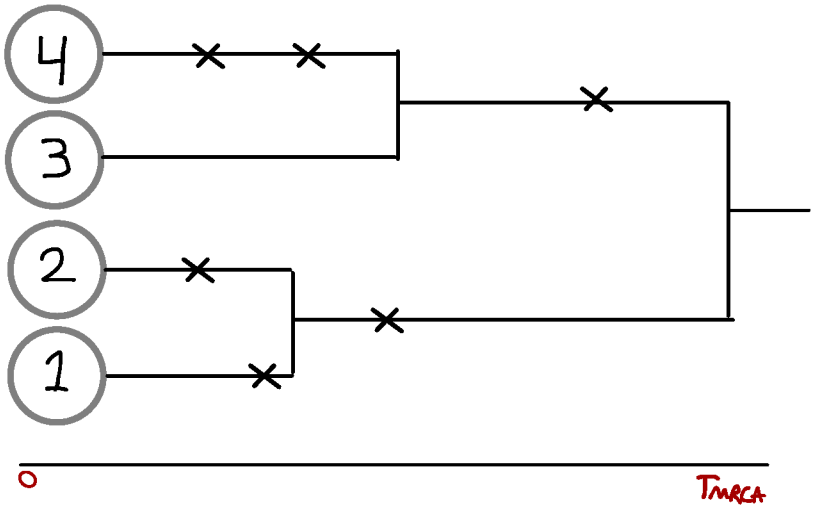

Figure 1.8: A realization of the infinite sites model. We can see that individual 1 and 2 have two differences, while individual 1 and 4 have five differences.

Probably, the most important model to process real data is the Infinite Sites Model (ISM). The model consists in a random tree $\mathcal{T}^n$ and a set of special points in the tree. To be precise $\mathcal{T}^n$ is a realization of the Kingman n-coalescent. After $\mathcal{T}^n$, a Poisson point process with intensity θ is realized over $\mathcal{T}^n$ to determine the special points, that we denote by $\mathcal{S}$. $(\mathcal{T}^n, \mathcal{S})$ should be understood as follows, $\mathcal{T}^n$ is the genealogical tree of a population of n-individuals, and each point in $\mathcal{S}$ corresponds to a mutation event. The most important assumption of this model is that each mutation affects a different position of the DNA, thus a mutation can't be erased.

The set of points in $\mathcal{S}$ are called segregating sites. Note that if we denote $|\mathcal{S}|$ the cardinality of $\mathcal{S}$ and $|\mathcal{T}^n|$ the length of the tree $\mathcal{T}^n$, then

$$\mathbb{E}[|\mathcal{S}|] = \theta \mathbb{E}[|\mathcal{T}^n|] = \theta 2 \sum_{i=1}^{n} 1/i \sim \theta 2 \log(n).$$

The symbol $\sim$ stands for the relation $a_n \sim b_n$ if and only if $\lim_{n\to\infty} a_n/b_n = 1$. The fact that $\mathbb{E}[|\mathcal{T}^n|] \sim \log(n)$ is an interesting result that can be consulted in [16]. Note that $2\theta = \mathbb{E}[|\mathcal{S}|]/\big(\sum_{i=1}^{n} 1/i\big)$.

Two individuals are considered to be very different if their ancestral lines, until their time to the most recent common ancestor, contain a lot of segregating sites. The number of segregating sites that are contained in exactly one of the ancestral lines of two individuals i,j in the sample, that we denote Δ_{ij}, is called the pairwise difference of the individuals i and j. Indeed, denoting $AL(i)$ the ancestral line of the individual i and $AL(j)$ the ancestral line of the individual j,

$$\Delta_{i,j} = |\{AL(i) \cap \mathcal{S}\} \cup \{AL(j) \cap \mathcal{S}\} - \{AL(j) \cap AL(i) \cap \mathcal{S}\}|.$$

Note that

$$\mathbb{E}[\Delta_{i,j}] = 2\theta \mathbb{E}[T_{MRCA}[2]] = 2\theta.$$

When studying a real sample of n DNA sequences, one can define a segregating site as follows: A nucleotide (a position in the DNA) is a segregation site if there are at least two elements of the sample having different letters in this nucleotide. Then, the number of segregation sites and the average pairwise differences can be obtained from the sampled sequences of DNA, assuming that each site can be hit by a mutation at most once. By our derivations on the theoretical model it follows that, if the sample was subject to neutral evolution, the experimentally obtained quantities $|\mathcal{S}|$ and $\Delta = \frac{2}{n(n-1)} \sum_{i \neq j} \Delta_{i,j}$ should be such that $\Delta - |\mathcal{S}| / \sum_{i=1}^{n} 1/i$ is close to zero. Close to zero here means that

$$D = \frac{\Delta - |\mathcal{S}| / \sum_{i=1}^{n} 1/i}{\mathrm{Var}[\Delta - |\mathcal{S}| / \sum_{i=1}^{n} 1/i]} \in [-2, 2]$$

If $|D| > 2$ this is a strong evidence that the underling tree is not Kingman (no neutral selection). A negative D can be the product of a selective sweep. A positive D can be the product of a decrease in the population size. The quantity D is known as Tajima's D, and it is one of the best known neutrality tests.

In [5] we discuss the effect of strong seedbanks in the Tajima D of a population. In Chapter 4 we study a model with mutation, in which the mutations arrive asymptotically as a Poisson process (see Theorem 4.2.14).

1.3.2 Selection

Selection will be important in Chapter 4, so we present here a short introduction.

Let us now study the two types Wright Fisher model with selection. Assume that each generation consists of N individuals. At generation zero, each individual is assigned a type, which can be either a or A. Let $s_N \in \mathbb{R}^+$ (s_N is known as the selective advantage). Each individual at each generation is assigned a weight which is 1 if the individual is of type a and $1 + s_N$ if the individual is of type A. Each individual chooses its parent from the previous generation uniformly on the weights, meaning that if the i-th individual has weight w_i, each individual in the following generation chooses it as its parent with probability $\frac{w_i}{\sum_{j=1}^{N} w_j}$. The type of each individual is the same as its parent's type. We observe that if in a certain generation there are xN individuals of type a, the number of type a individuals in the following generation will be binomially distributed with parameters N and $\frac{x}{1+s_N(1-x)} = x - \frac{s_N x(1-x)}{1+s_N(1-x)}$. This leads to the following definition:

Definition 1.3.3. *The frequency process of type a individuals, in the two types N-Wright Fisher model with selection, with selection parameter $s_N > 0$, is the Markov chain (X_t^N), with state space $\{0, 1/N, 2/N, ..., 1\}$ and transition probabilities*

$$p_{j,k} = \binom{N}{kN} \left(\frac{x}{1 + s_N(1-x)} \right)^{kN} \left(1 - \frac{x}{1 + s_N(1-x)} \right)^{(1-k)N}$$

for any $j, k \in \{0, 1/N, 2/N, ..., 1\}$.

Remark 1.3.4. The quantity s_N is known in the literature as the selective advantage of one type over the other. Selective advantage is a crucial concept in Chapter 4 (See Proposition 4.2.8).

When the population is big and the selective advantage is small, we can consider a diffusion approximation. This happens to be very useful in Subsection 4.3.6.

Proposition 1.3.5. *Assume that $s_N = s/N$ for some $s \in \mathbb{R}^+$. Let (X_t) be the solution to the SDE*

$$dX_t = -sX_t(1 - X_t)dt + \sqrt{X_t(1 - X_t)}dB_t,$$

then, if $x_0^N \to x_0$,

$$X_{Nt}^N \Rightarrow X_t$$

weakly over the space of Skorohod.

Proof. Let us calculate the generator A^N of (X^N_{Nt}) applied to a polynomial $f : \mathbb{R} \to \mathbb{R}$, at a point $x \in \mathbb{R}$.

$$\begin{aligned} A^N f(x) &= N\mathbb{E}_x[f(X_1) - f(x)] \\ &= N\mathbb{E}_x\left[(X_1 - x)f'(x)\right] + N\mathbb{E}_x\left[\frac{(X_1 - x)^2}{2}f''(x)\right] + O(1/N) \\ &= N(\mathbb{E}_x[X_1] - x)f'(x) + N\frac{1}{2}\mathrm{var}_x[X_1]f''(x) + O(1/N) \\ &= -sx(1-x)f'(x) + \frac{1}{2}x(1-x)f''(x) + O(1/N). \end{aligned}$$

As before, $O(1/N)$ is a converge uniformly to zero, as a function of $x \in [0, 1]$. □

The process (X_t) in the previous proposition is known as the Wright Fisher diffusion with selection, and admits a very nice dual that was first introduced in [53], and that we will study now.

Proposition 1.3.6. *Let (X_t) be as in Proposition 1.3.5 and let A be its generator. Let (N_t) be the $\mathbb{N}$-valued continuous time Markov chain with generator $\overline{A}$ such that, applied to any function $f : \mathbb{N} \to \mathbb{R}$ at any point $n \in \mathbb{N}$, fulfills*

$$\overline{A}f(n) = sn[f(n+1) - f(n)] + \binom{n}{2}[f(n-1) - f(n)].$$

Then, (X_t) and (N_t) are moment dual.

Proof. The proof consists in applying the generator to the function $H(x, n) = x^n$.

$$\begin{aligned} \overline{A}H(x,n) &= sn[x^{n+1} - x^n] + \binom{n}{2}[x^{n-1} - x^n] \\ &= -sx(1-x)nx^{n-1} + \frac{1}{2}x(1-x)n(n-1)x^{n-2} \\ &= -sx(1-x)\frac{\partial}{\partial x}H(x,n) + \frac{1}{2}x(1-x)\frac{\partial^2}{\partial x^2}H(x,n) \\ &= AH(x,n). \end{aligned}$$

□

Remark 1.3.7. When studying selection, a very important question is, if the number of individuals per generation is $N \in \mathbb{N}$, what is the probability that a single individual with selective advantage $s > 0$, is able to go to fixation? Fixation means that at certain generation all individuals are its descendants. A famous result in this direction is Haldane's formula, which states that in the Wright Fisher model with selection, in the case $1 \ll s > 0$ and $Ns \ll 1$, denoting
$\pi_N = \mathbb{P}$(fixation of genetic type of a single mutant with selective advantage s), we have that

$$\lim_{N\to\infty} \pi_N = 4s$$

A proof of Haldane's formula can be found in [16]. The constant 4 is model dependent, but the fact that it is a linear function of the selective parameter is a widespread property (see [56]). We will study a similar question for the case of moderate selection in Chapter 4. Moderate selection is the regime in which $s_N \to 0$ and $Ns_N \to \infty$.

Remark 1.3.8. Besides the probability of fixation, the time that a mutation takes to either get extinct or to go to fixation is also an interesting quantity to calculate. Let τ^N be the time until there is only one type in the population, that is $\tau^N = \inf\{i : X^N_i(1 - X^N_i) = 0\}$. In the proof of Theorem 1.1.8 there is a trivial bound that applies for all kinds of selection parameters s_N, which is $\lim_{N\to 0} \mathbb{P}(\tau^N > 2^N) = 0$. This is an extremely bad bound. Obtaining decent bounds is not always easy. In the case of $s_N = s > 0$ it can be proved that the time to fixation is of order $(2/s)\log(N)$. Theorem 4.2.10 in Chapter 4 deals with this problem in the case of moderate selection.

1.3.3 Structured coalescent

The structured coalescent ([28, 29]) is closely related to the seedbank coalescent, which is the main object of study of Chapter 3. Also, the auxiliary process constructed in Section 2.3.1 can be thought as a structured coalescent. For these reasons we present here a short introduction. We discuss a two type, two islands model. We also talk about duality.

It can be the case that the population we are interested in is affected by its geographic distribution. For example, we might want to study bacteria that live in different ponds. In this case most of the offspring of a bacterium in a certain pond will remain and reproduce in the same pond, but with certain probability they might emigrate (by the action of wind, for example). In this cases it is appropriate to use the so call structured coalescent.

Assume that each generation consists in $\{N^{(r)}\}_{r\in G}$ individuals. Here, r is the label of a location (locations can be thought as islands), in the system of locations G. To simplify our discussion we will assume that there are only two locations (in this case, the model is known as the *two islands model*). At generation zero, each individual, at each of the islands, is assigned a type, which can be either a or A. Let $c^{1,2}(N) \in [0,1]$ and $c^{2,1}(N) \in [0,1]$ (if there is no risk of confusion we will only write $c^{1,2}$ and $c^{2,1}$). At each generation each individual in the location 1 selects its parent independently of the others, by first choosing the location of its parent: the probability that its parent is in location 1 is $1-c^{1,2}$, and the probability that it is in location 2 is $c^{1,2}$. Once the location of the parent is fixed, the parent is chosen uniformly at random from the individuals of the selected location. Individuals at location 2 behave similarly, with the difference that they choose their parent in a different location with probability $c^{2,1}$. The type of each individual is the same as the type of its parent. We observe that if at certain generation there are $(xN^{(1)}, yN^{(2)})$ individuals of type a in each of the locations, the number of type a individuals at the following generation $(X_1N^{(1)}, Y_1N^{(2)})$ will be such that X_1 is binomially distributed with parameters $N^{(1)}$ and $x(1-c^{1,2})+yc^{1,2}$ and Y_1 is binomially distributed with parameters $N^{(2)}$ and $y(1-c^{2,1})+xc^{2,1}$. This lead to the following definition:

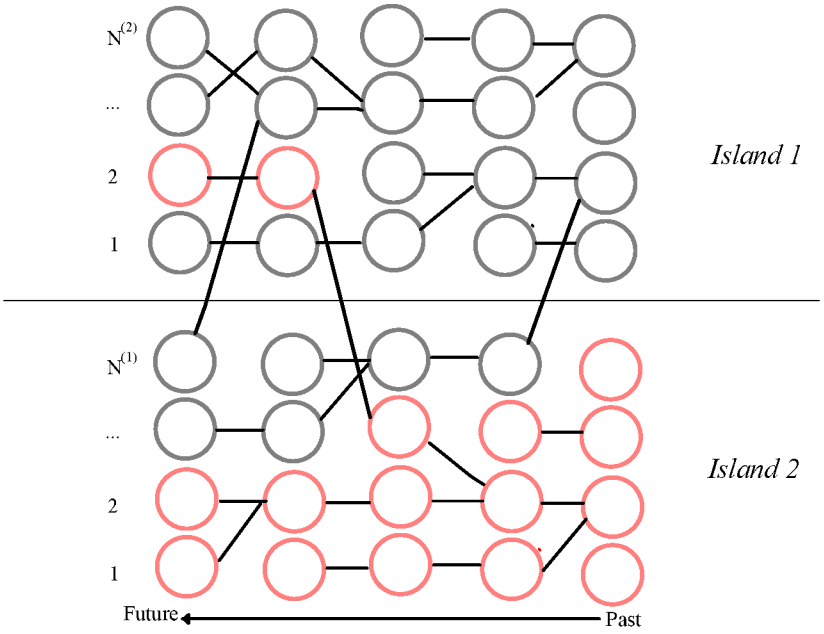

Figure 1.9: A realization of the two islands frequency process.

Definition 1.3.9. *The frequency process of type a individuals in the two types $N^{(1)}, N^{(2)}$-Wright Fisher two island model with migration parameters $c^{1,2} \in [0,1]$ and $c^{2,1} \in [0,1]$ is the two dimensional Markov chain $(X_i^N, Y_i^N)_{i\in\mathbb{N}}$ with state space $\{0, 1/N^{(1)}, 2/N^{(1)}, ..., 1\} \times \{0, 1/N^{(2)}, 2/N^{(2)}, ..., 1\}$ and transition*

probabilities

$$p_{(j_1,j_2),(k_1,k_2)} = \binom{N^{(1)}}{k_1 N^{(1)}} (x(1-c^{1,2}) + yc^{1,2}))^{k_1 N^{(1)}} (1 - (x(1-c^{1,2}) + yc^{1,2}))^{(1-k_1)N^{(1)}} \times \binom{N^{(2)}}{k_2 N^{(2)}} (y(1-c^{2,1}) + xc^{2,1}))^{k_2 N^{(2)}} (1 - (y(1-c^{2,1}) + xc^{2,1}))^{(1-k_2)N^{(2)}},$$

for any $j_i, k_i \in \{0, 1/N^{(i)}, 2/N^{(i)}, ..., 1\}$, $i \in \{1, 2\}$.

Again, when the population is large, under suitable conditions we can consider a diffusion approximation.

Proposition 1.3.10. *Assume that* $N^{(1)} = N$ *and* $N^{(2)} = \epsilon N$, *for some* $\epsilon > 0$. *Let* $c^{2,1}(N) = c^{2,1}/N$ *and* $c^{1,2}(N) = c^{1,2}/N$ *for some* $c^{2,1}, c^{1,2} \in \mathbb{R}^+$. *Let* (X_t, Y_t) *be the solution to the SDE*

$$\begin{aligned} dX_t &= c^{1,2}(Y_t - X_t)dt + \sqrt{X_t(1-X_t)}dB_t^1, \\ dY_t &= c^{2,1}(X_t - Y_t)dt + \epsilon\sqrt{Y_t(1-Y_t)}dB_t^2 \end{aligned}$$

where (B_t^1) *and* (B_t^2) *are two independent Brownian motions. Then, if* $(x_0^N, y_0^N) \to (x_0, y_0)$,

$$(X_{Nt}^N, Y_{Nt}^N) \Rightarrow (X_t, Y_t),$$

weakly over the space of Skorohod.

Proof. Let us calculate the generator A^N of (X_t^N, Y_t^N) applied to a polynomial $f : [0,1] \times [0,1] \to \mathbb{R}$, at a point $(x, y) \in [0,1] \times [0,1]$. We will use the two dimensional Taylor's formula

$$\begin{aligned} A^N f(x,y) =& N\mathbb{E}_x[f(X_1, Y_1) - f(x,y)] \\ =& N\mathbb{E}_x[(X_1 - x)\frac{\partial}{\partial x}f(x,y)] + N\mathbb{E}_x\left[\frac{(X_1-x)^2}{2}\frac{\partial^2}{\partial^2 x}f(x,y)\right] \\ &+ N\mathbb{E}_x[(Y_1 - y)\frac{\partial}{\partial y}f(x,y)] + N\mathbb{E}_x\left[\frac{(Y_1-y)^2}{2}\frac{\partial^2}{\partial y^2}f(x,y)\right] + O(1/N) \\ =& c^{1,2}(y-x)\frac{\partial}{\partial x}f(x,y) + \frac{x(1-x)}{2}\frac{\partial^2}{\partial x^2}f(x,y) \\ &+ c^{2,1}(x-y)\frac{\partial}{\partial y}f(x,y) + \epsilon\frac{y(1-y)}{2}\frac{\partial^2}{\partial y^2}f(x,y) + O(1/N). \end{aligned}$$

where $O(1/N)$ is uniformly bounded on (x, y). The proof is complete by standard arguments, as the generator of (X_t, Y_t) is exactly

$$Af(x,y) = c^{1,2}(y-x)\frac{\partial}{\partial x}f(x,y) + \frac{x(1-x)}{2}\frac{\partial^2}{\partial^2 x}f(x,y) + c^{2,1}(x-y)\frac{\partial}{\partial y}f(x,y) + \epsilon\frac{y(1-y)}{2}\frac{\partial^2}{\partial^2 y}f(x,y).$$

For a detailed manipulation of the smaller order terms in a similar calculation, see Appendix A.2. □

Now, we can deduce the moment dual.

Proposition 1.3.11. *Let* (X_t, Y_t) *be as in Proposition 1.3.10 and let* A *be its generator. Let* (N_t, M_t) *be the* $\mathbb{N} \times \mathbb{N}$ *valued continuous time Markov chain with generator* $\overline{A}$ *such that, applied to any function measurable and bounded* $f : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ *at any point* $(n, m) \in \mathbb{N} \times \mathbb{N}$, *fulfills*

$$\begin{aligned} \overline{A}f(n,m) &= c^{1,2}n[f(n-1, m+1) - f(n,m)] + \binom{n}{2}[f(n-1,m) - f(n,m)] \\ &+ c^{2,1}m[f(n+1, m-1) - f(n,m)] + \epsilon\binom{m}{2}[f(n,m-1) - f(n,m)]. \end{aligned}$$

Then, (X_t, Y_t) *and* (N_t, M_t) *are such that*

$$\mathbb{E}_{(x,y)}[X_t^n Y_t^m] = \mathbb{E}_{(n,m)}[x^{N_t} y^{M_t}].$$

Proof. The proof consists of applying the generator $\overline{A}$ to the function $H(x,y,n,m) = x^n y^m$.

$$\begin{aligned}
\overline{A}H(x,y,n,m) =& c^{1,2} n[x^{n-1}y^{m+1} - x^n y^m] + \binom{n}{2}[x^{n-1}y^m - x^n y^m] \\
&+ c^{2,1} m[x^{n+1}y^{m-1} - x^n y^m] + \epsilon\binom{m}{2}[x^n y^{m-1} - x^n y^m] \\
=& c^{1,2} n x^{n-1} y^m (y-x) + \binom{n}{2} x^{n-2} y^m x(1-x) \\
&+ c^{2,1} m x^n y^{m-1}(x-y) + \binom{m}{2} x^n y^{m-2} y(1-y) \\
=& c^{1,2}(y-x)\frac{\partial}{\partial x}H(x,y,n,m) + \frac{x(1-x)}{2}\frac{\partial^2}{\partial x^2}H(x,y,n,m) \\
&+ c^{2,1}(x-y)\frac{\partial}{\partial y}H(x,y,n,m) + \epsilon\frac{y(1-y)}{2}\frac{\partial^2}{\partial y^2}H(x,y,n,m) \\
=& AH(x,y,n,m).
\end{aligned}$$

□

Definition 1.3.12. *Define the* ***time to the most recent common ancestor*** *in the structured coalescent, of a sample of n individuals from island 1 and m from island 2, to be*

$$T_{MRCA}[n,m] := \inf\{t > 0 : N_t + M_t = 1 \text{ given that } N_0 = n, M_0 = m\},$$

where (N_t, M_t) is as in Proposition 1.3.10.

Two important properties of the structured coalescent are that it comes down from infinity and that $\mathbb{E}[T_{MRCA}[\cdot,\cdot]] : \mathbb{N}\times\mathbb{N} \mapsto \mathbb{R}^+$ is a bounded function (compare with Theorem 3.4.4 and Theorem 3.4.8 in Chapter 3). Now we prove that $\mathbb{E}[T_{MRCA}[\cdot,\cdot]]$ is a bounded function.

Lemma 1.3.13. *There exists a constant independent of n and m, $k > 0$ such that for every $n, m \in \mathbb{N}$,*

$$\mathbb{E}[T_{MRCA}[n,m]] < k.$$

Proof. The main idea is to imitate the calculation of the expected time to the most recent common ancestor for the Kingman coalescent (see Lemma 1.1.28). Let $k = n+m$. Let ψ_1 be the time of the first coalescent event, ψ_2 the time of the second coalescent event and so on: $\psi_i = \inf\{t > 0 : N_t + M_t = k - i\}$. Now note that if $N_t + M_t = k - i$, then at least $\lfloor \frac{k-i+1}{2} \rfloor$ individuals are in the same island. Without loss of generality assume that $\epsilon < 1$. This implies that

$$\mathbb{E}[\psi_{i+1} - \psi_i] \leq \left(\epsilon\binom{\frac{k-i+1}{2}}{2}\right)^{-1} \tag{1.3.1}$$

We finally observe that $T_{MRCA}[n,m] = \psi_{k-1} = \sum_{i=1}^{k-1}(\psi_i - \psi_{i-1})$, where $\psi_0 = 0$. Then for any $k > 2$,

$$\begin{aligned}
\mathbb{E}[T_{MRCA}[n,m]] =& \mathbb{E}\left[\sum_{i=1}^{k-1}(\psi_i - \psi_{i-1})\right] \\
=& \sum_{i=1}^{k-1}\mathbb{E}[\psi_i - \psi_{i-1}] \\
\leq& 2\sum_{s=1}^{\lfloor (k+1)/2 \rfloor}\left(\epsilon\binom{s}{2}\right)^{-1} + \sup_{n_2+m_2=2}\mathbb{E}[T_{MRCA}[n_2,m_2]] \\
=& 4\epsilon^{-1}\left(1 - \frac{1}{\lfloor (k+1)/2 \rfloor}\right) + \sup_{n_2+m_2=2}\mathbb{E}[T_{MRCA}[n_2,m_2]]
\end{aligned}$$

The first inequality follows from Equation (1.3.1). We finish the proof by observing that the right hand side is finite and independent of n, m. □

Part I

Seedbanks

Chapter 2

Generalizations of the KKL model

Now we present the first seedbank model. Some parts of this Chapter are based on [7] and [6]. The main Theorem of the chapter (Theorem 2.3.3) is based on discussions with J. Berestycki and N. Kurt.

Seedbanks can play an important role in the population genetics of a species, acting as a buffer against evolutionary forces such as random genetic drift and selection as well as environmental variability (see e.g. [70] for an overview). Their presence typically leads to significantly increased genetic variability resp. effective population size (see, e.g., [68], [42], [55], [69]) and could thus be considered as an important 'evolutionary force'. In particular, classical mechanisms such as fixation and extinction of genes become more complex: genetic types can in principle disappear completely from the *active* population at a certain time while returning later due to the germination of seeds or activation of dormant forms.

Seedbanks and dormant forms are known for many taxa. For example, they have been suggested to play an important role in microbial evolution [38], [23], where certain bacterial endospores can remain viable for (in principle arbitrarily) many generations.

In 2001, Kaj, Krone and Lascoux [33] postulated and studied an extension of the classical Wright Fisher model that includes seedbanks effects. In their model, each generation consists of a fixed amount of N individuals. Each individual chooses its parent a random amount of generations in the past and copies its genetic type. Here, the number of generations that separates each parent and offspring is understood as the time that the offspring spends as a seed or dormant form. Formally, a parent is assigned to each individual in generation g by first sampling a random number B, which is assumed to be independent and identically distributed for each individual, and then choosing a parent uniformly among the N individuals in generation $g-B$ (note that the case $B \equiv 1$ is just the classical Wright Fisher model). The distribution of B, that we denote $\mu \in \mathcal{P}(\mathbb{N})$, is called **the seedbank age distribution**.

The main result in [33] is that if μ is restricted to finitely many generations $\{1, 2, \dots, m\}$, where m is independent of N, then the ancestral process induced by the seedbank model converges, after the usual scaling of time by a factor N, to a time changed (delayed) Kingman coalescent, where the coalescent rates are multiplied by $1/\mathbb{E}[B]^2$. An increase of the expected value of the seedbank age distribution thus further decelerates the coalescent, leading to an increase in the effective population size. However, as observed by [70], since the overall coalescent tree structure is retained, this leaves the relative allele frequencies within a sample unchanged. In this scenario we thus speak of a 'weak' seedbank effect. In this chapter we will focus on 'weak' seedbank effect, but we will also show that seedbanks can cause stronger effects.

We are interested in the case where B is an unbounded random variable. In particular, we will assume that

$$\mathbb{P}(B > k) = \mu(\{k, k+1, ...\}) = L(k)k^{-\alpha},$$

for all $k \in \mathbb{N}$, where L is a slowly varying function. We will show that the genealogical process convergences to a time change of the Kingman coalescent in the case $\alpha > 1$. On the other hand, we will show that a *strong* seedbank effects can lead to a behaviour which is very different from the Kingman coalescent. In particular, if the seedbank age distribution is 'heavy-tailed', say, then if $\alpha < 1$ the expected time for the most recent common ancestor is infinite, and if $\alpha < 1/2$ two randomly sampled individuals do not have a common ancestor at all with positive probability. Hence this will not only delay, but actually completely alter the effect of random genetic drift.

Finally we will study a seedbank model in which the seedbank age distribution depends on the population size. We will study an example in which convergence to the Kingman coalescence holds after rescaling time with a scale function that goes to infinity orders of magnitude faster than N. This is a seedbank model with

$$\mu = \mu(N) = (1-\varepsilon)\delta_1 + \varepsilon\delta_{N^\beta}, \beta > 0, \varepsilon \in (0,1).$$

In particular, we will show that for $\beta < 1/5$ the ancestral process converges, after rescaling the time by the non-classical factor $N^{1+2\beta}$, to a time-changed Kingman coalescent, so that the expected time to the most recent common ancestor is highly elevated in this scenario. However, since the above model is highly non-Markovian, the results in other parameter regimes, in particular $\beta = 1$, are still elusive.

2.1 Construction of the model

The formal construction of our model follows [33], [7], [6]. Fix $\beta > 0, \varepsilon \in (0,1)$. For each $N \in \mathbb{N}$ let

$$\mu_N := \sum_{i=1}^{\infty} a_i^N \delta_i, \tag{2.1.1}$$

where $\sum_{i=1}^{\infty} a_i^N = 1$, $0 \le a_i^N \le 1$ and δ_i is the atomic Dirac measure with support $\{i\}$.

Fix once and for all a reference generation 0, from which time in discrete generations runs backwards. Fix a sample size $m \ge 2$ and a sampling measure γ for the generations of the original sample on the integers $\mathbb{N}$. We will usually assume that γ has finite support (independent of N), an important example being $\gamma = \delta_0$. Let $m \in \mathbb{N}$ be independent of N, and assume that $m < N$. The ancestral lineages of m sampled individuals indexed by $w \in \{1, ..., m\}$ in the seedbank process, who lived in generations sampled according to γ with respect to reference time 0, are constructed as follows. Let $\{(S_i^{(w)})_{i\in\mathbb{N}}\}_{w\in\{1,...,m\}}$ be a family of independent Markov chains, whose state space is the non-negative integers $\mathbb{N}_0$, with $S_0^{(w)} \sim \gamma$, and homogeneous transition probabilities,

$$\mathbb{P}\big(S_1^{(w)} = k' \,\big|\, S_0^{(w)} = k\big) = \mu_N(k'-k), \qquad 0 \le k < k',\ i = 1, ..., m.$$

The interpretation is that $S_0^{(w)}$ represents the generation of individual w, and $S_1^{(w)}$ the generation of its parent (backward in time), and so on. The set $\{S_0^{(w)}, S_1^{(w)}, ...\} \subseteq \mathbb{N}_0$ is thus the set of generations of all ancestors of individual w, including the individual itself.

This construction should be thought as a generalization of the Wright Fisher Graph introduced in Definition 1.1.1. Indeed, one can construct the Kaj Krone and Lascoux graph.

Definition 2.1.1. *Let $V_N = \big\{v = (g,l) \in \mathbb{Z} \times \{1, 2, ..., N\}\big\}$, $\{U_v\}_{v\in V_N}$ be a sequence of independent random variables, uniformly distributed in $\{1, 2, ..., N\}$, $\{l_v\}_{v\in V_N}$ be a sequence of independent μ_N distributed random variables and*

$$E_N = \big\{\{(g - l_{(g,l)}, U_{(g,l)}), (g,l)\} \text{ for all } v = (g,l) \in V_N\big\}.$$

*We define **the N-KKL graph** to be the random graph with vertex set V_N and edge set E_N.*

In order to construct the ancestral process of several individuals, we introduce interaction between ancestral lines as follows. Within the population of size N, in any fixed generation k, the individuals are labeled from 1 to N. Let $(U_i^{(w)})_{i\in\mathbb{N},w\in\{1,...,m\}}$ denote a family of independent random variables distributed uniformly on $\{1, ..., N\}$. We think of $U_{S_i^{(w)}}^{(w)}$ as the label within the population of size N of the i-th ancestor of individual w. This means that the label of each ancestor of each individual is picked uniformly at random in each generation that the ancestral line of this individual visits, exactly as it is done in the Wright Fisher model. The difference is that an ancestral line in the Wright Fisher model visits every generation, while in the Kaj Krone and Lascoux (KKL) model it does not. Note that of course all the random variables introduced up to now depend on the population size N.

Definition 2.1.2. *The* time to the most recent common ancestor *of two individuals w and j, denoted by $T_{MRCA}(2)$, is defined as*

$$T_{MRCA}(2) := \inf\left\{k > 0 : \exists i, r \in \mathbb{N}, k = S_i^{(w)} = S_r^{(j)} \text{ and } U_k^{(w)} = U_k^{(j)}\right\}. \tag{2.1.2}$$

In words, $T_{MRCA}(2)$ is the first generation back in time (counted from 0 on) in which two randomly sampled individuals ("initial" generations sampled according to γ) both have an ancestor, and both ancestors have the same label, hence, it is indeed the first generation back in time that w and j have the *same* ancestor.

It should be clear how to generalize this construction to lead to a full ancestral process of $m \geq 2$ individuals: Construct the process $(S_i^{(w)}, U_i^{(w)})_{i\in\mathbb{N}}$ independently for each individual, and couple the lines of individual w and individual j at the time of their most recent common ancestor by letting them evolve together from this time onward, as represented in Figure 1.

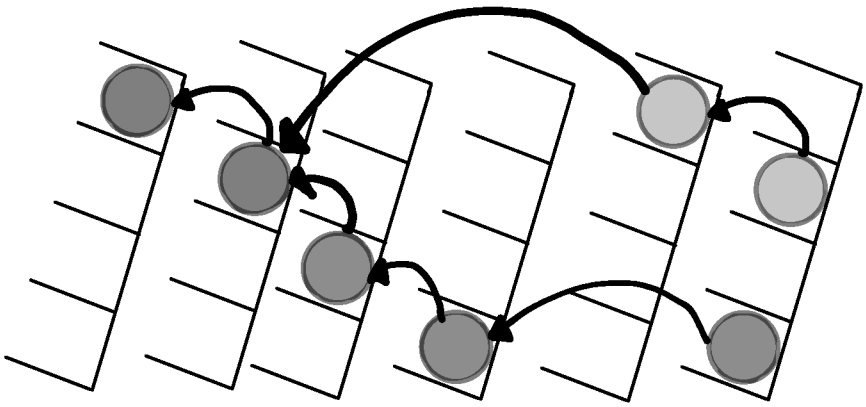

Figure 2.1: Coalescing ancestral lines of two individuals. The slots within each generation indicate the different individuals.

A precise construction is given in the following way. Let

$$T_1 := \inf\left\{k > 0 : \exists i, l \in \mathbb{N}, w \neq j \in \{1, ..., m\} : k = S_i^{(w)} = S_l^{(j)}, U_k^{(w)} = U_k^{(j)}\right\}, \tag{2.1.3}$$

be the time of the first coalescence of two (or more) lines, and let the set of individuals whose lines participate in a coalescence at time T_1 be denoted by

$$I_1 := \left\{w \in \{1, ..., m\} : \exists i, l \in \mathbb{N}, j \neq w : T_1 = S_i^{(w)} = S_l^{(j)}, U_{T_1}^{(w)} = U_{T_1}^{(j)}\right\}. \tag{2.1.4}$$

I_1 can be further divided into (possibly empty) pairwise disjoint sets

$$I_1^p := \left\{w \in I_1 : U_{T_1}^{(w)} = p\right\}, \quad p = 1, ..., N. \tag{2.1.5}$$

Note that by construction there is at least one p such that I_1^p is non-empty, which actually means that any such I_1^p contains at least two elements. Let

$$i_1^p := \min I_1^p,$$

and let

$$J_1 := \bigcup_{p: I_1^p \neq \emptyset} \{i_1^p\}.$$

After time T_1 we discard all $S^{(j)}$ for $j \in I_1^p, j \neq i_1^p$, and only keep $S^{(i_1^p)}$ for every $p = 1, ..., N$. We interpret this as merging the ancestral lineages of all individuals from I_1^p into one lineage at time T_1, separately for every p with $I_1^p \neq \emptyset$. In case there are several non-empty I_1^p, we observe simultaneous mergers. For $r \geq 2$ we define now recursively

$$T_r := \inf\big\{k > T_{r-1} : \exists i, l \in \mathbb{N}, \exists w, j \in I_{r-1}^c \cup J_{r-1}, i \neq j : k = S_i^{(w)} = S_l^{(j)}, U_k^{(w)} = U_k^{(j)}\big\}, \tag{2.1.6}$$

and

$$I_r := \big\{w \in I_{r-1}^c \cup J_{r-1} : \exists i, l \in \mathbb{N}, \exists j \neq w, j \in I_{r-1}^c \cup J_{r-1} : T_r = S_i^{(w)} = S_l^{(j)}, U_{T_r}^{(w)} = U_{T_r}^{(j)}\big\}, \tag{2.1.7}$$

and similarly $I_r^p := \{w \in I_r : U_{T_r}^{(w)} = p\}, i_r^p = \min I_r^p, p = 1, ..., N$, and $J_r = \cup_{p: I_r^p \neq \emptyset}\{i_r^p\}$. We stop the recursive construction as soon as $I_r^c = \emptyset$, which happens after finitely many r. Now we can finally define the main object of interest of this chapter.

Definition 2.1.3. *Fix $N \in \mathbb{N}, \beta > 0$, and $\varepsilon > 0$. Fix $m \ll N$ and an initial distribution γ on $\mathbb{N}_0$. Define a partition-valued process $(A_k^N)_{k\in\mathbb{N}_0}$, starting with $A_0^N = \big\{\{1\}, ..., \{m\}\big\}$, by setting $A_k^N = A_{k-1}^N$ if $k \notin \{T_1, T_2, ..\}$, and constructing the $A_{T_r}^N, r = 1, 2, ...$ in the following way: For each $p \in \{1, ..., N\}$ such that $I_r^p \neq \emptyset$, the blocks of $A_{T_r-1}^N$ that contain at least one element of I_r^p, are merged. Such merging is done separately for every p with $I_r^p \neq \emptyset$, and the other blocks are left unchanged. The resulting process $(A_k^N)_{k\in\mathbb{N}}$ is called the **ancestral process of m individuals in the Wright Fisher model with seedbank age distribution μ_N and initial distribution γ**. The time to the most recent common ancestor of the m individuals is defined as*

$$T_{MRCA}^N(m) := \inf\big\{k \in \mathbb{N} : A_k^N = \{1, ..., m\}\big\}. \tag{2.1.8}$$

Remark 2.1.4. This definition of ancestral process is slightly more involved than the equivalent definition for the Wright Fisher model, Definition 1.1.16. The reason for this is merely the technical difficulties, induced by the fact that in the KKL model ancestral lines do not visit all generations. However, the reader should remember than the concept behind the two definitions is the same.

It is important to note that (A_k^N) is not a Markov process: The probability that a coalescent event occurs at time k depends on more than just the configuration A_{k-1}^N. In fact, it depends on the values $\max\{S_n^{(w)} : S_n^{(w)} \leq k-1\}$, $w = 1, ..., m$, that is, on the generation of the last ancestor of each individual before generation k.

Now let us present an equivalent construction of (A_k^N) in terms of renewal processes.

Fix $N \in \mathbb{N}$ and a probability measure μ on the natural numbers. Let $v \in V_N := \mathbb{Z} \times \{1, ..., N\}$ denote an individual of our population. For $v \in V_N$ we write $v = (g_v, l_v)$ with $g_v \in \mathbb{Z}$, and $1 \leq l_v \leq N$, hence g_v indicating the generation of the individual in $\mathbb{Z}$, and l_v the label among the N individuals alive in this generation.

The ancestral line $AL(v) = \{v_0 = v, v_1, v_2, \dots\}$ of our individual v is a set of sites in V_N, as in definition 1.1.9, where $g_{v_0}, g_{v_1}, \dots \downarrow -\infty$ is a strictly decreasing sequence of generations, with independent decrements $g_{v_k} - g_{v_{k-1}} =: \eta_l, l \geq 1$ with distribution μ, and where the $l_{v_0}, l_{v_1}, \dots$ are i.i.d. uniform random variables with values in $\{1, \dots, N\}$, independent of $\{g_{v_k}\}_{k\in\mathbb{N}_0}$. Letting

$$S_i := \sum_{k=0}^{i} \eta_l,$$

where we assume $S_0 = \eta_0 = 0$, we obtain a discrete renewal process with interarrival law μ. In the language of [45], we say that a renewal takes place at each of the times $S_i, i \geq 0$, and we write $(q_i)_{i\in\mathbb{N}_0}$ for the renewal sequence, that is, q_i is the probability that i is a renewal time.

It is now straightforward to give a formal construction of the full ancestral process starting from $m \in \mathbb{N}$ individuals at time 0 in terms of a family of N independent renewal processes with interarrival law μ and a sequence of independent uniform random variables $U_k^w, k \in -\mathbb{N}, w \in \{1, \dots, m\}$, with values in $\{1, \dots, N\}$ (independent also of the renewal processes). Indeed, let the ancestral processes pick previous generations according to their respective renewal times, and then among the generations pick

labels according to their respective uniform random variables. As soon as at least two ancestral lineages hit a joint ancestor, their renewal processes couple, i.e. follow the same realization of one of their driving renewal processes (chosen arbitrarily, and discarding those remaining parts of the renewal processes and renewal times which aren't needed anymore). In other words, their ancestral lines merge.

Denote by P_N^μ the law of the above ancestral process. For $v = (g_v, l_v) \in V_N$ with $g_v = 0$, we have

$$q_i = P_N^\mu\Big(AL(v) \cap \big(\{-i\} \times \{1, ..., N\}\big) \neq \emptyset\Big), \tag{2.1.9}$$

and the probability that $w = (g_w, l_w) \in V_N$ is an ancestor of v, for $g_w < g_v$, is given by

$$P_N^\mu(w \in AL(v)) = \frac{1}{N} q_{g_v - g_w} = \frac{1}{N} q_{-g_w}.$$

For notational convenience, let us extend q_i to $i \in \mathbb{Z}$ by setting $q_i = 0$ if $i < 0$. Note that $q_0 = 1$.

In the rest of this chapter, we denote by P_γ the law of $(S_i^{(1)})$, indicating the initial distribution of the generations of the individuals. We write $P_{\otimes\gamma^m}$ for the law of the process (A_k^N) if the generations, of each of the m sampled individuals, are chosen independently according to γ. We abbreviate by slight abuse of notation both P_{δ_0} and $P_{\otimes\delta_0^m}$ by P_0.

2.2 Three different behaviors

In this subsection, a seedbank effect with unbounded seedbank age distribution μ is considered. To make this statement precise we need the following definition.

Definition 2.2.1. *For each $\alpha > 0$, let $\Gamma_\alpha := \{\mu_\alpha\}, \alpha \in (0, \infty)$ be the set of all measures μ such that*

$$\mu(\{i, i+1, ...\}) = i^{-\alpha} L(i), \quad n \in \mathbb{N},$$

for some slowly varying function $L : \mathbb{N} \mapsto \mathbb{N}$, i.e. for any $a \in \mathbb{R}^+$,

$$\lim_{i \to \infty} \frac{L(\lfloor ai \rfloor)}{L(i)} = 1$$

Below, we identify three regimes concerning the time to the most recent common ancestor: If $\alpha > 1$, then the expected time to the most recent common ancestor is of order N, and the ancestral process converges to a constant time change of Kingman's coalescent under classical rescaling by the population size. For $1/2 < \alpha < 1$, the time to the most recent common ancestor is finite almost surely, but the expectation is infinite for any N. If $\alpha < 1/2$, then there might be no common ancestor at all.

Theorem 2.2.2 (Existence and expectation of the time to the most recent common ancestor)**.** *Let $\mu \in \Gamma_\alpha$, $\gamma = \delta_0$, $v, w \in V_N, v \neq w$ and $N \in \mathbb{N}$.*

(a) If $\alpha \in (0, 1/2)$, then $\mathbb{P}(AL(v) \cap AL(w) \neq \emptyset) < 1$ for all $N \in \mathbb{N}$,

(b) If $\alpha \in (1/2, 1)$, then $\mathbb{P}(AL(v) \cap AL(w) \neq \emptyset) = 1$ and $\mathbb{E}[T_{MRCA}[2]] = \infty$ for all $N \in \mathbb{N}$.

(c) If $\alpha > 1$, then $\mathbb{P}(AL(v) \cap AL(w) \neq \emptyset) = 1$ for all $N \in \mathbb{N}$ and $\mathbb{E}[T_{MRCA}[2]] < \infty$ for all $N \in \mathbb{N}$.

In other words, for $\alpha > 1/2$ two individuals almost surely share a common ancestor, but the expected time to the most recent common ancestor is finite for $\alpha > 1$ and infinite if $\alpha \in (1/2, 1)$. Compare this Theorem with Lemma 1.1.15.

Remark 2.2.3. In the boundary case $\alpha = 1$, the choice of the slowly varying function L becomes relevant. If we choose $L = const.$, then it is easy to see from the proof that $\mathbb{E}[\tau] = \infty$. The case $\alpha = 1/2$ also depends on L and requires further investigation.

To prove Theorem 2.2.2 we will need some bounds on the q_i (defined in Equation(2.1.9)) that can be obtained via Tauberian theorems. **Proof of Theorem 2.2.2.** We first prove (c), which corresponds to

the case where we have convergence to Kingman's coalescent. Without loss of generality, assume $g_v =$

$g_w = 0$. Denote by (R_i) and (R'_i) the sequences of renewal times of the renewal processes corresponding to v and w respectively, that is, $R_i = \mathbf{1}_{\{i \in \{S_0, S_1, \ldots\}\}}$. In other words, $R_i = 1$ if and only if v has an ancestor in generation $-i$, and $q_i = P(R_i = 1)$. Let

$$T := \inf\{i : R_i = R'_i = 1\}$$

denote the coupling time of the two renewal processes. Since each time v and w have an ancestor in the same generation, these ancestors are the same with probability N, we get

$$\mathbb{E}[T_{MRCA}] = N\mathbb{E}[T].$$

But if $\alpha > 1$, we have that $E_\mu[\eta_1] < \infty$, and therefore by Proposition 2 of [44], $E[T] < \infty$.

(b) We will use Lemma 5.1.b. in [27], which is:

Lemma 2.2.4 ([27] 5.1.b)**.** *Let $\mu \in \Gamma_\alpha$. Let $\{q_i\}_{i\in\mathbb{N}}$ be as in Equation (2.1.9). The sum*

$$\sum_{i=0}^{\infty} q_i^2$$

is finite if $\alpha \in (0, 1/2)$ and infinite if $\alpha > 1/2$.

Now, for independent samples R and R', the expected number of generations where both individuals have an ancestor, is given by

$$E\big[\sum_{i=0}^{\infty} R_i R'_i\big] = \sum_{i=0}^{\infty} E[R_i]E[R'_i] = \sum_{i=0}^{\infty} q_i^2,$$

which is infinite if $\alpha > 1/2$ due to Lemma 2.2.4. Each of these times, the ancestors are the same with probability $1/N$, therefore with probability one $A(v)$ and $A(w)$ eventually meet. However, the expected time until this event is bounded from below by the expectation of the step size,

$$E[T_{MRCA}] \geq E[\eta] = \infty$$

if $\alpha < 1$.

(a) In this case, $E\big[\sum_{i=0}^{\infty} R_i R'_i\big] = \sum_{i=0}^{\infty} q_i^2 < \infty$, and therefore

$$P\big(\sum_{i=0}^{\infty} R_i R'_i = \infty\big) = 0,$$

which implies that the probability that $AL(v)$ and $AL(w)$ never meet is positive.

Remark 2.2.5. In [23] the observation that the existence of seedbanks can drastically delay the time to the most recent common ancestor was used to provide a plausible explanation for the peculiar genetic relation between *Azotobacter vinelandii* and *Pseudomonas*. It turns out that around 60% of the genome of *Azotobacter* is shared with *Pseudomonas*. This is rather large, but apparently not enough to consider *Azotobacter* as a *Pseudomonas*. *Azotobacter* is known for its ability to produce cysts. We proposed that *Azotobacter* has an ancester *Pseudomonas* in the very far past. We supported our hypothesis by comparing the genome of *Azotobacter* with the genome of different *Pseudomonas*. We concluded that the existence of a seedbank could have had a crucial effect in the evolution of *Azotobacter*.

2.3 Convergence to the Kingman coalescent

We will now study the *weak seed bank regime*, in which the ancestral process of the Kaj, Krone and Lascoux model converges to a time changed Kingman coalescent.

This section consists of three subsections: in the first we introduce an auxiliary process. In the second we state and prove a criterion for convergence to the Kingman coalescent. Finally in the last subsection we apply the criterion to some important particular cases.

2.3.1 An auxiliary process and its stationary distribution

In [33] and [7], an auxiliary Markov urn process plays a crucial role. We present this process and derive some of its properties.

Fix N and μ_N as in (2.1.1). Fix a probability measure γ on $\mathbb{N}$, and assume that $\mathrm{supp}(\gamma) \subseteq \mathrm{supp}(\mu_N)$. Let $\mathbb{P}_\gamma$ be the law of a Markov chain $(X_k)_{k\in\mathbb{N}_0}$ on $\mathbb{N}$ with initial distribution γ that moves according to the following rules: For $k > 0$, depending on the current state X_{k-1}, we have transitions

$$X_k = \begin{cases} i & \text{with probability } \mu_N(i-1)1_{\{X_{k-1}=0\}}, \text{ for any } i \in \mathbb{N}, i \neq 0, i \neq X_{k-1}-1 \\ 0 & \text{with probability } \mu_N(1)1_{\{X_{k-1}=0\}} + 1_{\{X_{k-1}=1\}}, \\ X_{k-1}-1 & \text{with probability } 1_{\{X_{k-1}>1\}}. \end{cases} \tag{2.3.1}$$

As in [33], we call this process an *urn process*, because we think of X_k as the position (urn) of a ball that is moved among countably many urns. Figure 2 shows the possible jumps of (X_k).

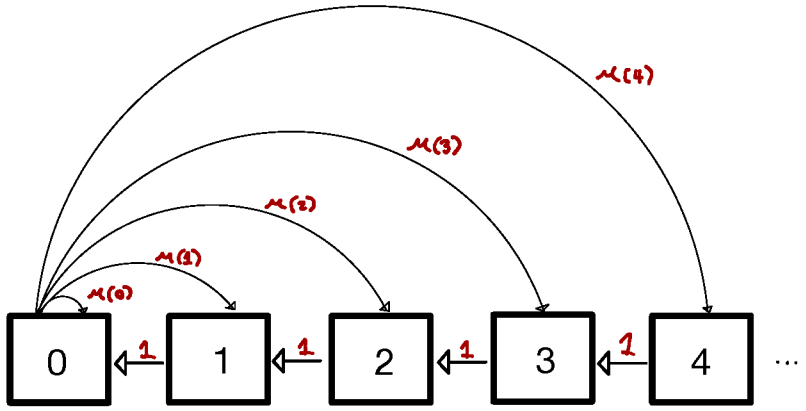

Figure 2.2: The possible jumps of X_k : From urn 0 it jumps acording to μ_N, from urn $i > 0$ it always goes urn $i - 1$.

How does this new process connect to the original processes $(S_k^{(w)})$ resp. our ancestral process (A_k^N)? We can couple (X_k) and $(S_k^{(w)})$ such that the successive times that (X_k) visits urn 0 are exactly the successive values visited by the process $(S_k^{(w)})$, that is, the generations in which individual i has an ancestor. This coupling is achieved as follows: Define

$$M_0 := \inf\{k \geq 0 : X_k = 0\}, \quad M_i = \inf\{k > M_{i-1} : X_k = 0\}, \ i \geq 1. \tag{2.3.2}$$

Then we have

Lemma 2.3.1. *Let $(X_i)_{i\in\mathbb{N}_0}$ be the above urn process with initial distribution γ on $\{0, ..., N^\beta - 1\}$, and let $(M_i)_{i\in\mathbb{N}_0}$ be defined as in (2.3.2). Then the process $(M_i)_{i\in\mathbb{N}_0}$ has the same distribution as $(S_i^{(w)})_{i\in\mathbb{N}_0}$ started in γ, and for all $k \in \mathbb{N}$,*

$$\mathbb{P}_\gamma(\exists i : S_i^{(w)} = k) = \mathbb{P}_\gamma(X_k = 0). \tag{2.3.3}$$

Proof. Immediate by construction. □

Note that, conversely, given $S_i^{(w)}, i \in \mathbb{N}_0$, equation (2.3.3) uniquely determines $M_i, i \in \mathbb{N}_0$. Taking γ as the distribution of S_0 and as the initial distribution for (X_i), since the urn process (X_i) is determined by the successive times it visits urn 1, Lemma 2.3.1 yields a one-to-one correspondence between ancestral lines in the KKL process, and the above urn process.

Let $\mathbb{E}_{\mu_N}$ denote the expectation of a μ_N-distributed random variable i.e.

$$\mathbb{E}_{\mu_N} = \mathbb{E}[B] = \sum_{i=1}^{\infty} \mu_N(\{k, k+1, ...\}).$$

Lemma 2.3.2 (Lemma 1 in [33]). *The probability measure ν_N on $\{0, ..., N^\beta - 1\}$ defined by*

$$\nu_N(k) := \frac{\mu_N\big(\{k, k+1, ...\}\big)}{\mathbb{E}_{\mu_N}}, \quad k = 0, 1, 2..., \tag{2.3.4}$$

is the unique stationary distribution of the urn process (X_k).

Proof. The proof follows from equation (2.3.1).

$$\begin{aligned}
\mathbb{P}_{\nu_N}(X_1 = k) =& \mathbb{P}_{\nu_N}(X_1 = k | X_0 = k+1)\mathbb{P}_{\nu_N}(X_0 = k+1) \\
&+ \mathbb{P}_{\nu_N}(X_1 = k | X_0 = 0)\mathbb{P}_{\nu_N}(X_0 = 0) \\
=& \nu_N(k+1) + \mu_N(k)\nu_N(0) \\
=& \nu_N(k)
\end{aligned}$$

□

In view of Lemma 2.3.1, an important quantity in this paper will be the probability that $X_k = 0$, which under stationarity is equal to

$$\nu_N(0) = \frac{1}{\mathbb{E}_{\mu_N}}. \tag{2.3.5}$$

This is particularly important because

$$\mathbb{P}_{\nu_N \otimes \nu_N}(X_k^{(i)} = X_k^{(j)} = 0) = \nu_N^2(0) = \left(\frac{1}{\mathbb{E}_{\mu_N}}\right)^2. \tag{2.3.6}$$

2.3.2 A mixing time criterion for convergence to the Kingman coalescent

Theorem 2.3.3. *Fix a sequence of seedbank age distributions $\{\mu_N\}_{N\in\mathbb{N}}$, such that for every $N \in \mathbb{N}$ $|supp(\mu_N)| < \infty$. Let $(X_k^N)_{k\in\mathbb{N}_0}$ be the urn process associated to μ_N, with $X_0^N = (0, ..., 0)$. Let (A_k^N) be the ancestral process induced by the urn process (X_k^N). If $(X_k^N)_{k\in\mathbb{N}_0}$ is irreducible and aperiodic with stationary distribution ν_N and a mixing time τ_{mix}^N, and they are such that*

$$\frac{\tau_{mix}^N \mathbb{E}_{\mu_N}^2}{N} \to 0 \text{ as } N \to \infty, \tag{2.3.7}$$

then for any sample size $m \in \mathbb{N}$, the rescaled ancestral processes $(A^N_{\lfloor N\mathbb{E}^2_{\mu_N} t\rfloor})_{t\in\mathbb{R}^+}$ converges weakly over the space of Skorohod to the Kingman m-coalescent.

Remark 2.3.4. For simplicity we will always assume $X_0^N = (0, ..., 0)$, but the results are true under more general initial conditions. In [7] and [6] particular cases of the previous Theorem are discussed for more general initial conditions (independent of N).

Remark 2.3.5. If one considers unbounded seedbank age distributions, in the sense of $|supp(\mu_N)| = \infty$, then $\tau_{mix}^N = \infty$. In that case the theorem is still true, but it is absolutely not useful. One can study *unbounded seedbanks* by truncating μ_N and applying a coupling argument. Unfortunately, this method still doesn't seem to work as well as one would like it to work (see Theorem 2.3.7 and compare it with the main Theorem of [7]).

Proof. Heuristically, if equation (2.3.7) holds, by the time until the first coalescent event "has an opportunity to happen", the urn process (X_k^N) gets very close to its stationarity distribution ν_N. Imagine an urn process which is always in stationarity, then the probability of observing a coalescence event in the ancestral process that it induces, depends only on the number of ancestors (the number of balls in the urn process or equivalently the number of blocks in the state of the ancestral process). In this case, the ancestral process is a Markov process. The philosophy of this proof is to compare the ancestral process (A_k^N) with an artificial ancestral process that comes form an artificial urn process which, in some sense, is always in stationarity. The gain of doing this is that we can use the generator of the artificial ancestral process to prove convergence to the Kingman coalescent and then use a coupling argument to extend the result to *the real* ancestral process. The main idea to construct the coupling is to observe the position of each pair of individuals $w, j \in \{1, 2, ..., m\}$ only when both members of the pair have the same label *i.e.* $U_k^{(w)} = U_k^{(j)}$ (when there is an opportunity to coalesce). This gives enough time to the urn processes $X_k^{(w)}, X_k^{(j)}$ to have a very similar distribution to the stationary distribution ν.

The proof is notationally heavy, so the reader is invited to consult Table 2.1.

Table 2.1: Notation

Symbol	Description	State space
$\{U_k^{(w)}\}$	Family of independent uniform on $\{0, 1, ..., N\}$ RV	$\{0, 1, ..., N\}$
τ_i^{wj}	Times at which coalescence is possible between the blocks with smallest element w and j respectively	$\mathbb{N}$
τ_i	Times at which coalescence is possible	$\mathbb{N}$
(X_k^N)	Urn process	$\mathbb{N}$
(A_k^N)	Ancestral process constructed using (X_k^N) and $\{U_k^{(w)}\}$	$[m]$
(R_i^N)	Equal in distribution to $(X_{\tau_i}^N)$	$\mathbb{N}$
$(\overline{R}_i^N)$	Sequence of independent ν_N distributed RV, $(\overline{R}_i^N, R_i^N)$ are constructed using optimal coupling	$\mathbb{N}$
(Z_k^N)	$(Z_{\tau_i}^N) = R_i^N$ and constant in $k \notin \{\tau_i\}$	$\mathbb{N}$
$(\overline{Z}_k^N)$	$(\overline{Z}_{\tau_i}^N) = \overline{R}_i^N$ and constant in $k \notin \{\tau_i\}$	$\mathbb{N}$
$(\overline{L}_i^N)$	Ancestral process constructed using $(\overline{Z}_k^N)$ and $\{U_k^{(w)}\}$	$[m]$
(L_i^N)	Ancestral process constructed using (Z_k^N) and $\{U_k^{(w)}\}$	$[m]$
Q	Absorption time of (L_i^N)	$\mathbb{N}$
k	Time measured in generation	$\mathbb{N}$
i	Time measured in opportunities to coalesce	$\mathbb{N}$
t	Time in the limiting scale	$\mathbb{R}^+$

Fix the seedbank age distributions μ_N and fix the sample size $m \in \mathbb{N}$. Denote $E_N = supp(\mu_N)$. The N will be dropped from the notation when there is no ambiguity. Recall the family of random variables $(U_k^{(w)})_{k \in \mathbb{N}_0}$, for each $w \in \{1, ..., m\}$, which is the sequence of IID uniform in $\{1, ..., N\}$ random variables, introduced in the construction of (A_k^N) in Section 2.1. For any pair $j, w \in \{1, 2, ..., m\}$, $j \neq w$, let $\tau_0^{jw} = 0$ and

$$\tau_i^{jw} = \inf\{k > \tau_{i-1}^{jw} : U_k^{(j)} = U_k^{(w)}\}.$$

Let $\tau_0 = 0$ and

$$\tau_i = \inf\{k > \tau_{i-1} : k = \tau_s^{jw} \text{ for some } s \in \mathbb{N},\ j, w \in \{1, 2, ..., m\}\}.$$

Observe that for all $w, j \in \{1, 2, ..., m\}$ it holds that almost surely $\tau_i \leq \tau_i^{jw}$. Note further that $\{\tau_i^{jw} - \tau_{i-1}^{jw}\}_{i \in \mathbb{N}}$ is a sequence of IID geometric random variables with parameter $1/N$. Indeed,

$$\mathbb{P}(\tau_i^{jw} - \tau_{i-1}^{jw} \geq k) = \mathbb{P}(\tau_1^{jw} \geq k) = \mathbb{P}(U_s^{(j)} \neq U_s^{(w)} \, \forall \, s < k) = (1 - 1/N)^{k-1}$$

This implies in particular that, for every $i \in \mathbb{N}$ and $j, w \in \{1, 2, ..., m\}$, τ_i^{jw} is almost surely finite. Then we conclude that for every $i \in \mathbb{N}$, τ_i is almost surely finite.

Let $\overline{R}_0^N = R_0^N = X_0^N$. Conditionally on $R_{i-1}^N = r = (r_1, ..., r_m)$ and $\overline{R}_{i-1}^N = \overline{r} = (\overline{r}_1, ..., \overline{r}_m)$, let $(\overline{R}_i^N, R_i^N)$ be the optimal coupling of the stationary distribution of (X_k^N), which is $\nu_N^{\otimes m}$, and the

probability measure

$$\xi_{i,r}(\cdot) := \mathbb{P}_r(X^N_{\tau_i-\tau_{i-1}} \in \cdot) = \mathbb{P}_{r_1,...,r_m}((X^{(1)}_{\tau_i-\tau_{i-1}}, ..., X^{(m)}_{\tau_i-\tau_{i-1}}) \in \cdot), \tag{2.3.8}$$

which are two probability measures on E_N^m. Applying the strong Markov property to the almost surely finite stopping time τ_{i-1} we observe that

$$\xi_{i,r}(\cdot) = \mathbb{P}_r(X^N_{\tau_i-\tau_{i-1}} \in \cdot) = \mathbb{P}_r(X^N_{\tau_1} \in \cdot) =: \xi_r(\cdot). \tag{2.3.9}$$

As we construct R_i^N and $\overline{R}_i^N$ using the optimal coupling, we know by Lemma 1.2.30 that

$$\mathbb{P}\big(R_i^N = \overline{R}_i^N | R_{i-1}^N = (r_1, ..., r_m)\big) = 1 - ||\nu_N^{\otimes m} - \xi_{r_1,...,r_m}||_{TV} \tag{2.3.10}$$

Lemma 2.3.6. *The above construction implies that:*

- $(R_i^N, \overline{R}_i^N)_{i\in\mathbb{N}}$ *is a Markov process.*
- *The stochastic process* $(\overline{R}_i^N)_{i\in\mathbb{N}}$ *is a sequence of independent* ν_N*-distributed random variables .*
- *In distribution* $(R_i^N)_{i\in\mathbb{N}} = (X^N_{\tau_i})_{i\in\mathbb{N}}$.

Proof. The first claim is immediate by construction. The second claim is verified inductively. It is clear by construction that $\overline{R}_1^N$ is ν_N-distributed and independent of $\overline{R}_0^N$. Assume that $(\overline{R}_i^N)_{i\in\{1,...,n-1\}}$ is a collection of independent ν_N-distributed random variables. Then,

$$\begin{aligned}
\mathbb{P}(\overline{R}_n^N = a_n, \overline{R}_{n-1}^N = a_{n-1}, ..., \overline{R}_1^N = a_1) &= \sum_{b\in E_N} \mathbb{P}(\overline{R}_n^N = a_n, \overline{R}_{n-1}^N = a_{n-1}, ..., \overline{R}_1^N = a_1, R_{n-1}^N = b) \\
&= \sum_{b\in E_N} \mathbb{P}(\overline{R}_n^N = a_n \big| \overline{R}_{n-1}^N = a_{n-1}, ..., \overline{R}_1^N = a_1, R_{n-1}^N = b) \\
&\quad \times \mathbb{P}(\overline{R}_{n-1}^N = a_{n-1}, ..., \overline{R}_1^N = a_1, R_{n-1}^N = b) \\
&= \sum_{b\in E_N} \mathbb{P}(\overline{R}_n^N = a_n \big| \overline{R}_{n-1}^N = a_{n-1}, R_{n-1}^N = b) \\
&\quad \times \mathbb{P}(\overline{R}_{n-1}^N = a_{n-1}, ..., \overline{R}_1^N = a_1, R_{n-1}^N = b) \\
&= \nu_N(a_n) \sum_{b\in E_N} \mathbb{P}(\overline{R}_{n-1}^N = a_{n-1}, ..., \overline{R}_1^N = a_1, R_{n-1}^N = b) \\
&= \nu_N(a_n)\mathbb{P}(\overline{R}_{n-1}^N = a_{n-1}, ..., \overline{R}_1^N = a_1)
\end{aligned}$$

This proofs the inductive step and proofs the claim.

To verify the last claim notice that by construction $(R_i^N)_{i\in\mathbb{N}}$ is a Markov chain, (the right hand side of equation (2.3.9) depends only on $(r_1, ..., r_m)$). As $X_0^N = R_0^N$, to prove the claim we just need to verify that the transition matrix of $(R_i^N)_{i\in\mathbb{N}}$ is the same as the transition matrix of $(X^N_{\tau_i})_{i\in\mathbb{N}}$.

For any fixed $(r_1, ..., r_m) \in E_N^{\otimes m}$ and $(x_1, ..., x_m) \in E_N^{\otimes m}$,

$$\begin{aligned}
\mathbb{P}_{r_1,...,r_m}(R_1^N = (x_1, ..., x_m)) &= \xi_{i,r_1,...,r_m}((x_1, ..., x_m)) \\
&= \mathbb{P}_{r_1,...,r_m}((X^{(1)}_{\tau_1}, ..., X^{(m)}_{\tau_1}) = (x_1, ..., x_m)).
\end{aligned} \tag{2.3.11}$$

Then the transition probabilitis are equal, and the equality in distribution holds.

□

For every $k \in \mathbb{N}$, let $i_k := \max\{i : \tau_i \le k\}$. We define

$$Z_k := R^N_{i_k}$$

and

$$\overline{Z}_k := \overline{R}^N_{i_k}$$

We construct (L_k^N) and $(\overline{L}_k^N)$ just as we constructed (A_k^N) in Section 2.1, where (Z_k) (resp. $(\overline{Z}_k)$) plays the role of (X_k). This means that j and w coalesce in (L_k^N) the first time $k>0$, such that $Z_k^{(w)} = Z_k^{(j)}$ and $U_k^{(w)} = U_k^{(j)}$. By Equation (2.3.11), in distribution $(A_k^N) = (L_k^N)$ as in either process coalescence can occur only on times $k = \tau_i^N$, for some $i \in \mathbb{N}$.

Now, notice that $(\overline{L}_k^N)$ is a Markov chain with values in $[m]$, the space of partitions of m elements. Let $\pi, \pi' \in [m]$. Recall that, we say that π' follows π, and we write $\pi \succ \pi'$, if π' can be constructed from π by merging exactly 2 blocks. Assume that $\pi \succ \pi'$. Assume further, without loss of generality, that π' is constructed by merging the blocks of π with smallest element w and j respectively. Then,

$$\begin{aligned}\mathbb{P}(\overline{L}_k^N = \pi' \,|\, \overline{L}_{k-1}^N = \pi) =& \mathbb{P}(\{\exists i : k = \tau_i^{wj}\}, \overline{Z}_k^{(w)} = \overline{Z}_k^{(j)} = 0) + o(1/N) \\ =& \mathbb{P}(\exists i : k = \tau_i^{wj}) \mathbb{P}(\overline{Z}_k^{(w)} = \overline{Z}_k^{(j)} = 0 | \exists i : k = \tau_i^{wj}) + o(1/N) \\ =& \frac{1}{N} \frac{1}{\mathbb{E}_{\mu_N}^2} + o(1/N)\end{aligned}$$

Now, let $\underline{\pi}$ be a partition of m elements, that is different from π and that can not be constructed by merging 2 blocks of π. Then, in order to go from π to $\underline{\pi}$ in one step, a necessary condition is that at least 3 elements of $\{U_k^{(w)}\}_{w\in\{1,\dots,m\}}$ are equal or at least 2 pairs of elements of $\{U_k^{(w)}\}_{w\in\{1,\dots,m\}}$ are equal. Either event happen with probability N^{-2}. Further, at least 3 elements of $\{\overline{Z}_k^{(w)}\}_{w\in\{1,\dots,m\}}$ must be at state zero. This happen with probability $\mathbb{E}_{\mu_N}^{-3}$

$$\mathbb{P}(\overline{L}_k^N = \underline{\pi} \,|\, \overline{L}_{k-1}^N = \pi) \leq \frac{1}{N^2 \mathbb{E}_{\mu_N}^3}.$$

From these two equations we can calculate the generator of $\overline{L}_k^N$. For any $\pi \in [m]$, let $[m]_\pi = \{\pi' \in [m] : \pi \succ \pi'\}$. Let $f : [m] \mapsto \mathbb{R}$.[1] Let $\overline{\mathcal{L}}$ be the discrete generator of $\overline{L}_k^N$. Then

$$\overline{\mathcal{L}} f(\pi) = \sum_{\pi' \in [m]_\pi} \frac{f(\pi') - f(\pi)}{N \mathbb{E}_{\mu_N}^2} + O(\frac{1}{N^2 \mathbb{E}_{\mu_N}^3}).$$

This suffices to conclude that $(\overline{L}^N_{\lfloor N\mathbb{E}^2_{\mu_N} t\rfloor})_{t>0} \Rightarrow (K_t)_{t>0}$, weakly over the space of Skorohod over $[m]$, where (K_t) is the Kingman coalescent (see, [18] Theorem 4.8.2 and Theorem 3.7.8).

It remains to show that with probability going to 1 as N goes to infinity, the processes (L_k^N) and the process $(\overline{L}_k^N)$ follow the same trajectory. Let $Q = \inf\left\{i \in \mathbb{N} : \overline{L}_{\tau_i}^N = \{\{1, 2, ..., m\}\}\right\}$ be the absorption time of $(\overline{L}_{\tau_i}^N)$. Note that $\{\{1, 2, ..., m\}\}$ is an absorbing state of $\overline{L}_{\tau_i}^N$ and $L_{\tau_i}^N$. Our aim for the rest of the proof will be to verify the hypothesis of Lemma 1.2.26.

$$\begin{aligned}\mathbb{P}\big((L_k^N) = (\overline{L}_k^N), \forall k \in \mathbb{N}\big) =& \mathbb{P}\big((L_{\tau_i}^N) = (\overline{L}_{\tau_i}^N), \forall i \leq Q\big) \\ \geq& \mathbb{P}(R_i^N = \overline{R}_i^N, \forall i \leq Q)\end{aligned}$$

Note further that for any function $f : \mathbb{N} \mapsto \mathbb{R}^+$, such that $\lim_{N\to\infty} f(N) = \infty$, it holds that

$$\lim_{N\to\infty} \mathbb{P}(Q > f(N)\mathbb{E}_{\mu_N}^2) \leq \lim_{N\to\infty} 1 - \big(1 - (1 - \mathbb{E}_{\mu_N}^{-2})^{f(N)\mathbb{E}_{\mu_N}^2/m}\big)^m = 0.$$

To obtain the inequality one can consider a modified coalescent in which only one pair has the possibility to coalesce at each time and then one can calculate the probability that non of the m coalescent events (neccesary to attain the absorbing sate) takes more that $f(N)\mathbb{E}_{\mu_N}^2/m$ units of time. Then we observe that the two last equations lead to

$$\begin{aligned}\lim_{N\to\infty} \mathbb{P}\big((L_k^N) = (\overline{L}_k^N), \forall k \in \mathbb{N}\big) \geq& \lim_{N\to\infty} \mathbb{P}(R_i^N = \overline{R}_i^N, \forall i \leq f(N)\mathbb{E}_{\mu_N}^2) \\ \geq& \lim_{N\to\infty} \left(\inf_{r\in E^m} (1 - ||\nu_N^{\otimes m} - \xi_{i,r}||_{TV})\right)^{\lfloor f(N)\mathbb{E}_{\mu_N}^2 \rfloor}\end{aligned} \tag{2.3.12}$$

[1] We do not require further properties of the function f, because $[m]$ is a finite (discrete) space.

Here we used Equation (2.3.10) in the second inequality. Now observe that, for all $r = r_1, ..., r_m \in E^m$,

$$\begin{aligned}
||\nu_N^{\otimes m} - \xi_{i,r}||_{TV} =& ||\nu_N^{\otimes m} - \sum_{s=1}^{\infty} \mathbb{P}_{r_1,...,r_m}((X_s^{(1)}, ..., X_s^{(m)}) \in \cdot)\mathbb{P}(\tau_i - \tau_{i-1} = s)||_{TV} \\
\leq & \sum_{s=1}^{\infty} ||\nu_N^{\otimes m} - \mathbb{P}_{r_1,...,r_m}((X_s^{(1)}, ..., X_s^{(m)}) \in \cdot)||_{TV}\mathbb{P}(\tau_1 = s) \\
\leq & \sum_{h=1}^{\infty}\sum_{k=1}^{\tau_{mix}} ||\nu_N^{\otimes m} - \mathbb{P}_{r_1,...,r_m}((X^{(1)}_{r\tau_{mix}+k}, ..., X^{(m)}_{h\tau_{mix}+k}) \in \cdot)||_{TV}\mathbb{P}(\tau_1 = \tau_{mix} + k) \\
\leq & \tau_{mix}\mathbb{P}(\tau_1 = 1)\sum_{h=1}^{\infty} 2^{-h} \\
\leq & \binom{m}{2}\frac{\tau_{mix}}{N},
\end{aligned}$$

where in the first inequality we used the triangle inequality and the fact that in distribution $\tau_i - \tau_{i-1} = \tau_1$. In the second inequality we used Lemma 1.2.32 and that τ_1 is a geometric random variable, so $\mathbb{P}(\tau_1 = i)$ is a decreasing function of i.

Substituting in equation (2.3.12), we conclude

$$\lim_{N\to\infty} \mathbb{P}\big((L_k^N) = (\overline{L}_k^N), \forall k \in \mathbb{N}\big) \quad \geq \quad \lim_{N\to\infty} \left(1 - \binom{m}{2}\frac{\tau_{mix}}{N}\right)^{\lfloor (f(N)\mathbb{E}_{\mu_N}^N)^2 \rfloor}. \tag{2.3.13}$$

If the assumption stated in Equation (2.3.7) is satisfied we can choose the function

$$f(N) = \ln(\frac{N}{\tau_{mix}\mathbb{E}_{\mu_N}^2}),$$

so we have that $f(N) \to \infty$ and, using the Bernoulli's inequality,

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{P}\big((L_k^N) = (\overline{L}_k^N), \forall k \in \mathbb{N}\big) \quad &\geq \quad \lim_{N\to\infty} \left(1 - \binom{m}{2}\ln(\frac{N}{\tau_{mix}(\mathbb{E}_{\mu_N}^N)^2})\frac{\tau_{mix}\mathbb{E}_{\mu_N}^2}{N}\right) \\
&= \quad 1.
\end{aligned}$$

We conclude by Lemma 1.2.26 that $(L^N_{\lfloor N\mathbb{E}^2_{\mu_N} t\rfloor})$ converges to (K_t), weakly over the space of Skorohod, which finally imply that $(A^N_{\lfloor N\mathbb{E}^2_{\mu_N} t\rfloor})$ converges to (K_t), weakly over the space of Skorohod. □

2.3.3 Applications of the criterion for convergence to the Kingman coalescent

We now present a proof of a special case of the main theorem of [7], that follows an application of 2.3.3. This is also a proof of the main Theorem of [33], as a particular case.

Theorem 2.3.7. *Assume that $\mu_N = \mu$ for all $N \in \mathbb{N}$, where μ is a seedbank age distribution such that*

$$\mu(\{n : n > i\}) = L(i)i^{-\alpha},$$

for some slowly varying function $L : \mathbb{N} \to \mathbb{R}$ and some $\alpha > 1$. Fix a sample of size $m \in \mathbb{N}$ and let $(X_0^N) = (0, ..., 0)$. Then, as $N \to \infty$, the sequence of stochastic processes $(A^N_{\lfloor N\mathbb{E}^2_{\mu} t\rfloor})_{t\in\mathbb{R}^+}$ converges weakly over the space of Skorohod to Kingman's m-coalescent.

Proof. We will use a coupling argument. Fix a constant $\beta \in (\alpha^{-1}, 1)$, so that $\alpha\beta > 1$. Let $(\underline{X}_k^N)$ be the urn process of a KKL model with seedbank age distribution $\underline{\mu}_N$ defined by

$$\underline{\mu}_N(i) = \begin{cases} \mu(i) & \text{for } i \in \{2, ..., N^\beta\} \\ \mu(1) + \mu(\{i > N^\beta\}) & \text{for } i = 1, \\ 0 & \text{for } i > N^\beta. \end{cases} \tag{2.3.14}$$

Let (X_k^N) be the urn process of a seedbank model with seedbank age distribution μ. Note that $\mathbb{E}_{(\underline{\mu})} \leq \mathbb{E}_\mu < \infty$.

First let us prove that $(\underline{A}^N_{\lfloor N\mathbb{E}_\mu^2 t\rfloor}) \Rightarrow (K_t)$, where $(\underline{A}^N_{\lfloor N\mathbb{E}_\mu^2 t\rfloor})$ is the ancestral process that correspond to $(\underline{X}^N_{\lfloor N\mathbb{E}_\mu^2 t\rfloor})$. Let τ^N_{mix} be the mixing time of $(\underline{X}^N_k)$. Note that for every $\epsilon > 0$, there exist $N_\epsilon \in \mathbb{N}$, such that for all $N > N_\epsilon$

$$\tau^N_{mix} \leq N^{\beta+\epsilon}. \tag{2.3.15}$$

This bound (and probably much better bounds) can be shown using renewal theory or the Doeblin coupling (see Example 1.2.3 and Example 1.2.4). A quick argument follows an easy proposition:

Let $V_k^N \in \{1, ..., 3N^\beta\}$ be the number of generations, among generations $\{-3N^\beta, ..., 0\}$, that are visited by the ancestral lines of two individuals v_1, v_2, such that individual v_1 belongs to generation zero and individual v_2 belongs to generation $-k \in \{-N^\beta, ..., 0\}$. Note that $\tau_{coup} \leq k$ if and only if $V_k^N > 0$.

Proposition 2.3.8. $\liminf_{N\to\infty} \mathbb{P}(V_k^N > 0) > 0$.

Proof. Note that $\mathbb{E}[V_k] < 3N^\beta \mathbb{P}(V_k > 0)$. We will show that $\liminf_{N\to\infty} \mathbb{P}(V_k > 0) > 0$, by showing that $\mathbb{E}[V_k] = O(N^\beta)$. For every $N \in \mathbb{N}$, by convergence to the stationary distribution of the urn process, $\lim_{i\to\infty} q_i > 1/(2\mathbb{E}_\mu)$, where q_i is as in Equation (2.1.9). So that for every $k \in \{1, 2, ..., N^\beta\}$

$$\lim_{N\to\infty} \frac{\sum_{i=1}^{3N^\beta} q_i q_{i-k}}{N^\beta} > \mathbb{E}_\mu^{-2}.$$

This implies that $V_k = o(N^\beta)$. □

The bound on the mixing time is finally obtained by applying the Doebling coupling in each coordinate of the urn process independently (as in Example 1.2.4), and noticing that, by Proposition 2.3.8, the coupling has positive probability of being successful (independent of N) each $3N^\beta$ generations. Clearly this implies that the probability that the coupling is not successful for at least one coordinate, in the first $N^{\beta+\epsilon}$ generations, is exponentially small. Then, Equation (2.3.15) follows and we observe that, for large enough N,

$$\frac{\tau^N_{mix}\mathbb{E}^2_{\underline{\mu}_N}}{N} \leq N^{\beta+\epsilon-1}.$$

If we take $0 < \epsilon < 1 - \beta$, the assumption of Theorem 2.3.3 (Equation (2.3.7)) is true and we can then conclude that $(\underline{A}^N_{\lfloor N\mathbb{E}_\mu^2 t\rfloor}) \Rightarrow (K_t)$.

To finish the proof, we construct a coupling of $(\underline{X}^N_k)$ and (X^N_k), by the rule $\underline{X}^{N\prime}_0 = X^N_0$ and if $\underline{X}^{N\prime}_k = X^N_k$ and $X^{N(w)}_{k+1} - X^{N(w)}_k < N^\beta$ for all $w \in \{1, 2, ..., m\}$, then $\underline{X}^{N\prime}_{k+1} = X^N_{k+1}$. Otherwise $\underline{X}^{N\prime}_{k+1}$ is independent of X^N_{k+1} and has the same transition law as $\underline{X}^N_{k+1}$.

Let $J_N = \inf\{k \in \mathbb{N} : \underline{X}^{N\prime}_k \neq X^N_k\}$. Let $u \in (0, \alpha\beta - 1)$, note that for any starting point $\underline{x} \in E^{\otimes m}$,

$$\begin{aligned}\mathbb{P}_{\underline{x}}(J_N < N^{1+u}) &< 1 - (\mu(\{i < N^\beta\}))^{mN^{1+u}} \\ &< 1 - (1 - N^{-\alpha\beta})^{mN^{1+u}} \to 0.\end{aligned}$$

Denote $\underline{T}^N_{MRCA}$ the time to to the most recent common ancestor of $(\underline{A}^N_k)$. Note that the convergence to the Kingman coalescent of $(\underline{A}^N_k)$, implies in particular that $\mathbb{P}_{\overline{x}}(\underline{T}^N_{MRCA} > N^{1+u}) \to 0$. Then

$$\mathbb{P}_{\overline{x}}(\underline{T}^N_{MRCA} > J_N) < \mathbb{P}_{\overline{x}}(\underline{T}^N_{MRCA} > N^{1+u}) + \mathbb{P}_{\overline{x}}(J_N < N^{1+u}) \to 0.$$

This implies that $\mathbb{P}(\underline{X}^{N\prime}_k = X^N_k, \forall k \in \mathbb{N}) \to 1$. Thus the statement of the Theorem follows by applying Lemma 1.2.26. □

Remark 2.3.9. We only proved the convergence to the Kingman coalescent for $\alpha > 1$ and not for the less restrictive condition $\mathbb{E}_\mu < \infty$, which can be proved using renewal theory (see [7]). This is due to our bound on the mixing time, which is bad (Equation (2.3.15)). The intuition suggests that the mixing time is of order one. I believe that mixing times are an excellent tool to study weak seedbanks with unbounded seedbank age distribution, and that these techniques will lead to new results (also involving further evolutionary forces).

Now we present a slightly weaker result than the main Theorem of [6].

Theorem 2.3.10. *Consider the ancestral process of a sample of size $m \in \mathbb{N}$ in a seedbank model with seedbank age distribution*

$$\mu_N = (1-\varepsilon)\delta_1 + \varepsilon\delta_{N^\beta} \tag{2.3.16}$$

and starting condition for the urn process $(X_0^N) = (0, ..., 0)$. If $0 < \beta < 1/5$ the sequence of processes $\{(A^N_{\lfloor \varepsilon^2 N^{1+2\beta} t \rfloor})_{t\geq 0}\}_{N\in\mathbb{N}}$ converges to Kingman's m-coalescent weakly on the space of Skorohod.

Proof. First note that $\mathbb{E}_{\mu_N} = 1-\varepsilon+\varepsilon N^\beta = O(N^\beta)$, this imply that $\frac{\tau^N_{mix} E^2_{\mu_N}}{N} = O(\frac{\tau^N_{mix}}{N^{1-2\beta}})$. It is proved in the Appendix that $\tau_{mix} < N^{3\beta+\delta}$ for every $\delta > 0$ (see Lemma A.1.1 in the Appendix). Then if $\beta < 1/5$, we can take $1-5\beta > \delta > 0$, so that the assumption of Theorem 2.3.3 (Equation (2.3.7)) is satisfied. The statement of the Theorem follows Theorem 2.3.3. □

Remark 2.3.11. The result of Theorem 2.3.10 was proved in [6] for $\beta < 1/4$. What happens for $\beta > 1/4$ is still open.

We will sketch a last example, which is related to the main object of the next chapter. We take

$$\mu^\alpha_N = (1 - \frac{1}{N^\alpha})\delta_1 + \frac{1}{N^\alpha}\sum_{i=1}^{\infty} \mathbb{P}(G^\alpha = i)\delta_i$$

where $\alpha > 0$ and G^α is a geometric random variable such that $\mathbb{E}[G^\alpha] = N^\alpha$. Note that $\mathbb{E}_{\mu_N} = 2 - \frac{1}{N^\alpha}$. It can be verified that the mixing time of is of order N^α. Then, by Theorem 2.3.3, if $\alpha < 1$, the ancestral process converges to the Kingman coalescent. However, if $\alpha = 1$, then $\frac{\tau_{mix}(E_\mu)^2}{N} = O(1)$, and Theorem 2.3.3 does not apply. It turns out that in this case the scaling limit is not the Kingman coalescent. In the next chapter we study an equivalent model to the case $\alpha = 1$, which happens to converge to *the seedbank coalescent.*

Chapter 3

The Seedbank Coalescent

3.1 Introduction

This chapter consists essentially of the paper [8] and contains parts of [5].

As we saw in the last chapter, while there are mathematical results in the weak seedbank regime, it appears as if the 'right' scaling regimes for stronger seedbank models, and the potentially new limiting coalescent structures, have not yet been identified. This is in contrast to many other population genetic models, where the interplay of suitably scaled evolutionary forces (such as mutation, genetic drift, selection and migration) often leads to elegant limiting objects, such as the ancestral selection graph [53], or the structured coalescent [29, 54]. A particular problem is the loss of the Markov property in Wright Fisher models with long genealogical 'jumps'.

In this chapter we thus propose a new Markovian Wright Fisher type seedbank model that allows for a clear forward and backward scaling limit interpretation. In particular, the forward limit in a bi-allelic setup will consist of a pair of (S)DEs describing the allele frequency process of our model, while the limiting genealogy, linked by a duality result, is given by a coalescent structure which we call *seedbank coalescent*. In fact, the seedbank coalescent can be thought of as a structured coalescent of a two island model in a 'weak migration regime', in which however coalescences are *completely blocked* in one island. Despite this simple description, the seedbank coalescent exhibits qualitatively altered genealogical features, both in comparison to the Kingman-coalescent and the structured coalescent. In particular, we prove in Theorem 3.4.4 that the seedbank coalescent *does not come down from infinity*, and in Theorem 3.4.8 that the expected time to the most recent common ancestor of an n sample is of asymptotic order $\log\log n$ as n gets large. Interestingly, this latter scale agrees with the one for the Bolthausen-Sznitman coalescent identified by Goldschmidt and Martin [22].

Summarizing, the seedbank coalescent seems to be an interesting and natural scaling limit for populations in the presence of a 'strong' seedbank effect. In contrast to previous genealogies incorporating (weak) seedbank effects, it is a new coalescent structure and not a time-change of Kingman's coalescent, capturing the essence of seedbank effects in many relevant situations.

The remainder of this chapter is organised as follows:

In Subsection 3.2, we discuss the Wright Fisher model with a seedbank component that has a geometric age structure, and show that its two bi-allelic frequency processes (for 'active' individuals and 'seeds') converge to a two-dimensional system of SDEs. We derive their *dual* process and employ this duality to compute the fixation probabilities as $t \to \infty$ (in law) of the system.

In Subsection 3.3, we define the *seedbank coalescent* corresponding to the previously derived dual block counting process and show how it describes the ancestry of the Wright Fisher geometric seedbank model.

In Subsection 3.4, we prove some interesting properties of the seedbank coalescent, such as '*not* coming down from infinity' and asymptotic bounds on the expected time to the most recent common ancestor, which show that genealogical properties of a population in the presence of strong seedbanks are altered qualitatively.

3.2 The seedbank model

3.2.1 The forward model and its scaling limit

Consider a haploid population of fixed size N reproducing in fixed discrete generations $k = 0, 1, ...$ Assume that individuals carry a genetic type from some type-space E (we will later pay special attention to the bi-allelic setup, say $E = \{a, A\}$, for the forward model).

Further, assume that the population also sustains a *seedbank* of constant size $M = M(N)$, which consists of the dormant individuals. For simplicity, we will frequently refer to the N 'active' individuals as 'plants' and to the M dormant individuals as 'seeds'.

Given $N, M \in \mathbb{N}$, let $\varepsilon \in [0, 1]$ such that $\varepsilon N \leq M$ and set $\delta := \varepsilon N/M$, and assume for convenience that $\varepsilon N = \delta M$ is a natural number (otherwise replace it by $\lfloor \varepsilon N \rfloor$ everywhere). Let $[N] := \{1, \dots, N\}$ and $[N]_0 := [N] \cup \{0\}$. The dynamics of our Wright Fisher model with strong seedbank component are then as follows:

- The N active individuals (plants) from generation 0 produce $(1 - \varepsilon)N$ active individuals in generation 1 by multinomial sampling with equal weights.
- Additionally, $\delta M = \varepsilon N$ uniformly (without replacement) sampled seeds from the seedbank of size M in generation 0 'germinate', that is, they turn into exactly one active individual in generation 1 each, and leave the seedbank.
- The active individuals from generation 0 are thus replaced by these $(1-\varepsilon)N+\delta M = N$ new active individuals, forming the population of plants in the next generation 1.
- Regarding the seedbank, the N active individuals from generation 0 produce $\delta M = \varepsilon N$ seeds by multinomial sampling with equal weights, filling the vacant slots of the seeds that were activated.
- The remaining $(1 - \delta)M$ seeds from generation 0 remain inactive and stay in the seedbank (or, equivalently, produce exactly one offspring each, replacing the parent).
- Throughout reproduction, offspring and seeds copy/resp. maintain the genetic type of the parent.

Thus, in generation 1, we have again N active individuals and M seeds. This probabilistic mechanism is then to be repeated independently to produce generations $k = 2, 3, ...$ Note that the offspring distribution of active individuals (both for the number of plants and for the number of seeds) is exchangeable within their respective sub-population. Further, one immediately sees that the time that a given seed stays in the seedbank before becoming active is geometric with success parameter δ, while the probability a given plant produces a dormant seed is ε.

Definition 3.2.1 (The seedbank model)**.** Fix population size $N \in \mathbb{N}$, seedbank size $M = M(N)$, genetic type space E and δ, ε as before. Given initial type configurations $\xi_0 \in E^N$ and $\eta_0 \in E^M$, denote by

$$\xi_k := \big(\xi_k(i)\big)_{i\in[N]}, \quad k \in \mathbb{N},$$

the random genetic type configuration in E^N of the plants in generation k (obtained from the above mechanism), and denote by

$$\eta_k := \big(\eta_k(j)\big)_{j\in[M]}, \quad k \in \mathbb{N},$$

correspondingly the genetic type configuration of the seeds in E^M. We call the discrete-time Markov chain $(\xi_k, \eta_k)_{k\in\mathbb{N}_0}$ with values in $E^N \times E^M$ the *type configuration process* of the *seedbank model*.

Remark 3.2.1. This way of introducing a type process is in the spirit of the classic definition of Cannings processes (see [9]). It is intuitively clear how to carry out a random-graph construction of the model, similar to our construction of the Wright Fisher graph (Definition 1.1.1). However, if we do that, notation would become unnecessarily heavy. In any case, it is important to note that the seedbank model allows a natural backwards in time representation.

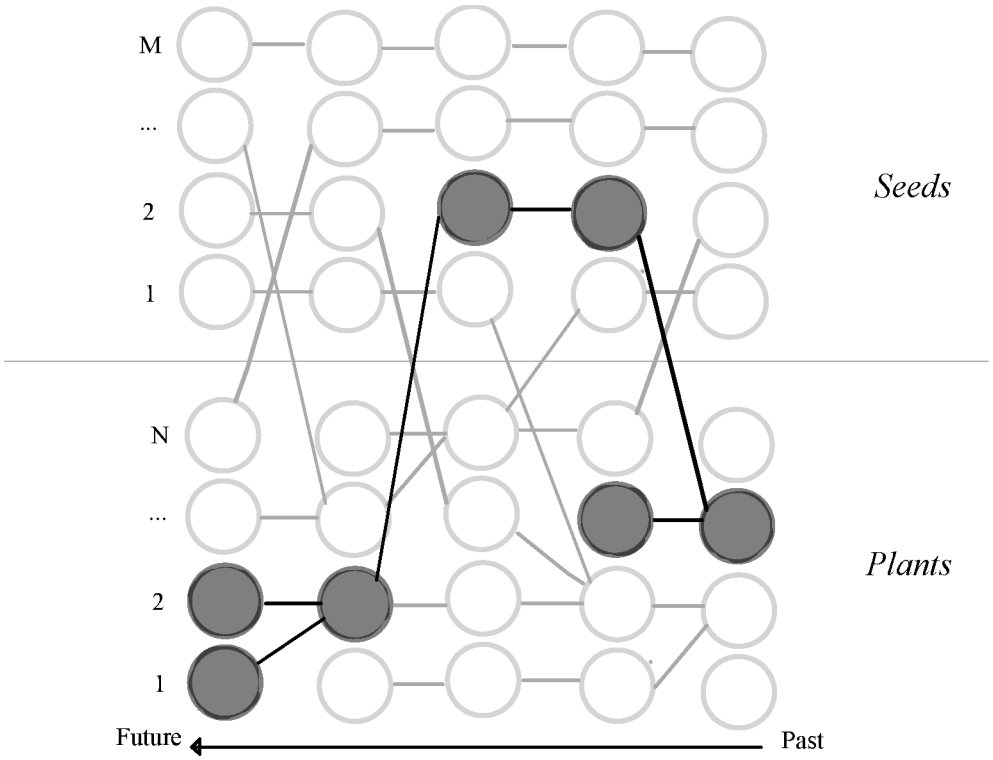

Figure 3.1: A realisation of the ancestral relationships in a seedbank model. Here, the genetic type of the third plant in generation 0 (highlighted) is lost after one generation, but returns in generation three via the seedbank, which acts as a buffer against genetic drift and maintains genetic variability.

We now specialise to the bi-allelic case $E = \{a, A\}$ and define the frequency chains of a alleles in the active population and in the seedbank. Define

$$X_k^N := \frac{1}{N} \sum_{i \in [N]} \mathbf{1}_{\{\xi_k(i)=a\}} \quad \text{and} \quad Y_k^M := \frac{1}{M} \sum_{j \in [M]} \mathbf{1}_{\{\eta_k(j)=a\}}, \quad k \in \mathbb{N}_0. \tag{3.2.1}$$

Both are discrete-time Markov chains taking values in

$$I^N = \Big\{0, \frac{1}{N}, \frac{2}{N}, \dots, 1\Big\} \subseteq [0,1] \quad \text{resp.} \quad I^M = \Big\{0, \frac{1}{M}, \frac{2}{M}, \dots, 1\Big\} \subseteq [0,1].$$

Denote by $\mathbb{P}_{x,y}$ the distribution for which (X^N, Y^M) starts in $(x,y) \in I^N \times I^M$ $\mathbb{P}_{x,y}$-a.s., i.e.

$$\mathbb{P}_{x,y}(\,\cdot\,) := \mathbb{P}\big(\,\cdot\,\big|X_0^N = x,\, Y_0^M = y\big) \quad \text{for} \quad (x,y) \in I^N \times I^M$$

(with analogous notation for the expectation, variance etc). The corresponding time-homogeneous transition probabilities can now be characterized.

Proposition 3.2.2. *Let $c := \varepsilon N = \delta M$ and assume $c \in [N]_0$. With the above notation we have for (x,y) resp. $(\bar{x}, \bar{y}) \in I^N \times I^M$,*

$$\begin{aligned} p_{x,y} :&= \mathbb{P}_{x,y}\big(X_1^N = \bar{x},\, Y_1^M = \bar{y}\big) \\ &= \sum_{i=0}^{c} \mathbb{P}_{x,y}(Z = i)\mathbb{P}_{x,y}(U = \bar{x}N - i)\mathbb{P}_{x,y}(V = M(\bar{y} - y) + i) \end{aligned}$$

where Z, U, V are independent under $\mathbb{P}_{x,y}$ with distributions

$$\mathcal{L}_{x,y}(Z) = \mathrm{Hyp}_{M,c,yM}, \qquad \mathcal{L}_{x,y}(U) = \mathrm{Bin}_{N-c,x}, \qquad \mathcal{L}_{x,y}(V) = \mathrm{Bin}_{c,x}\,.$$

Here, $\mathrm{Hyp}_{M,c,yM}$ denotes the hypergeometric distribution with parameters $M, c, y \cdot M$ and $\mathrm{Bin}_{c,x}$ is the binomial distribution with parameters c and x.

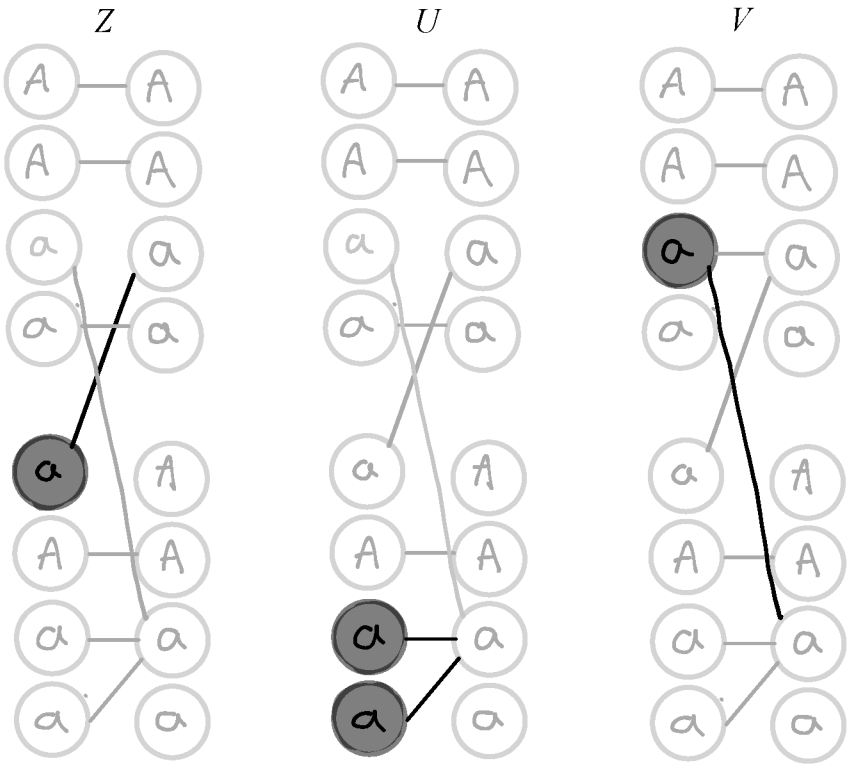

Figure 3.2: In this figure $Z = 1$, $U = 2$ and $V = 1$

Remark 3.2.3. The random variables introduced in Proposition 3.2.2 have a simple interpretation:

- Z is the number of plants in generation 1, that are offspring of a seed of type a in generation 0. This corresponds to the number of seeds of type a that germinate / become active in the next generation (noting that, in contrast to plants, the 'offspring' of a germinating seed is always precisely one plant and the seed vanishes).
- U is the number of plants in generation 1, that are offspring of plants of type a in generation 0.
- V is the number of seeds in generation 1, that are produced by plants of type a in generation 0.

Proof of Proposition 3.2.2. With the interpretation of Z, U and V given in Remark 3.2.3 their distributions are immediate as described in the Definition 3.2.1. By construction we then have $X_1^N = \frac{U+Z}{N}$ and $Y_1^M = y + \frac{V-Z}{M}$ and thus the claim follows. □

In many modeling scenarios in population genetics, parameters describing evolutionary forces such as mutation, selection and recombination are scaled in terms of the population size N in order to reveal a non-trivial limiting structure (see e.g. [17] for an overview). In our case, the interesting regime is reached by letting ε, δ (and M) scale with N. More precisely, assume that there exist $c, K \in (0, \infty)$ such that

$$\varepsilon = \varepsilon(N) = \frac{c}{N} \quad \text{and} \quad M = M(N) = \frac{N}{K}. \tag{3.2.2}$$

Without loss of generality $c \in [N]_0$ as $N \to \infty$. Under assumption (3.2.2), the seedbank age distribution is geometric with parameter

$$\delta = \delta(N) = \frac{c}{M(N)} = \frac{cK}{N}, \tag{3.2.3}$$

and c is the number of seeds that become active in each generation, resp. the number of individuals that move to the seedbank. The parameter K determines the relative size of the seedbank with respect to the active population.

Proposition 3.2.4. *Assume that (3.2.2) holds. Consider test functions $f \in C^{(2)}([0,1]^2)$. For any $(x,y) \in I^N \times I^M$, let $A^N = A^N_{(\varepsilon,\delta,M)}$ be the discrete generator of the frequency Markov chain $(X^N_{Nt}, Y^M_{Nt})_{t\in\mathbb{R}^+}$, which act on f by*

$$A^N f(x,y) := N\mathbb{E}_{x,y}\Big[f\big(X^N_1, Y^M_1\big) - f(x,y)\Big].$$

Then for all $(x,y) \in [0,1]^2$,

$$\lim_{N\to\infty} A^N f(x,y) = Af(x,y),$$

where A is defined by

$$Af(x,y) := c(y-x)\frac{\partial f}{\partial x}(x,y) + cK(x-y)\frac{\partial f}{\partial y}(x,y) + \frac{1}{2}x(1-x)\frac{\partial f^2}{\partial x^2}(x,y).$$

A proof can be found in the appendix, see Proposition A.2.1. Since the state space of our frequency chain can be embedded in the compact unit square $[0,1]^2$, we get tightness and convergence on path-space easily by standard argument (see, e.g. [18] Theorem 4.8.2 and Theorem 3.7.8) and can identify the limit of our frequency chains as a pair of the following SDEs:

Corollary 3.2.2 (Seedbank diffusion)**.** *Under the conditions of Proposition 3.2.4, if $X^N_0 \to x$ a. s. and $Y^M_0 \to y$ a.s., we have that*

$$(X^N_{\lfloor Nt\rfloor}, Y^N_{\lfloor Nt\rfloor})_{t\geq 0} \Rightarrow (X_t, Y_t)_{t\geq 0}$$

on $D_{[0,\infty)}([0,1]^2)$ as $N\to\infty$, where $(X_t,Y_t)_{t\geq 0}$ is a 2-dimensional diffusion solving

$$\begin{aligned} dX_t &= c(Y_t - X_t)\,dt + \sqrt{X_t(1-X_t)}\,dB_t,\\ dY_t &= cK(X_t - Y_t)\,dt, \end{aligned} \tag{3.2.4}$$

where $(B_t)_{t\geq 0}$ is standard Brownian motion.

The proof again follows from standard arguments, cf. e.g. [18], where in particular Proposition 2.4 in Chapter 8 shows that the operator A is indeed the generator of a Markov-Process.

Remark 3.2.5. If we abandon the assumption $N = KM$ there are situations in which we can still obtain meaningful scaling limits. If we assume $N/M \to 0$, and we rescale the generator as before by measuring the time in units of size N, we obtain (cf. Proposition A.2.1)

$$\lim_{N\to\infty} A^N f(x,y) = c(y-x)\frac{\partial f}{\partial x}(x,y) + \frac{1}{2}x(1-x)\frac{\partial f^2}{\partial x^2}(x,y).$$

This shows that the limiting process is purely one-dimensional, namely the seedbank frequency Y_t is constantly equal to y, and the process $(X_t)_{t\geq 0}$ is a Wright Fisher diffusion with migration (with migration rate c and reverting to the mean y). The seedbank, which in this scaling regime is much larger than the active population, thus acts as a reservoir with constant allele frequency y, with which the plant population interacts.

The case $M/N \to 0$ leads to a simpler limit: If we rescale the generator by measuring the time in units of size M we obtain

$$\lim_{M\to\infty} A^M f(x,y) = c(y-x)\frac{\partial f}{\partial y}(x,y)$$

and constant frequency $X \equiv x$ in the plant population, which tells us that if the seedbank is of smaller order than the active population, the genetic configuration of the seedbank will converge to the genetic configuration of the active population, in a deterministic way.

The above results can be extended to more general genetic types spaces E in a standard way using the theory of measure-valued resp. Fleming-Viot processes. This will be treated elsewhere. Before we investigate some properties of the limiting system, we first derive its *dual* process.

3.2.2 The dual of the seedbank frequency process

As we saw in Theorem 1.1.18, the Wright Fisher diffusion is known to be dual to the block counting process of the Kingman-coalescent, and similar duality relations hold for other models in population genetics (see Section 1.3). Such dual processes are often extremely useful for the analysis of the underlying system, and it is easy to see that our seedbank diffusion also has a nice dual.

Definition 3.2.6. We define the *block counting process of the seedbank coalescent* $(N_t, M_t)_{t\geq 0}$ to be the continuous time Markov chain taking values in $\mathbb{N}_0 \times \mathbb{N}_0$ with transitions

$$(n,m) \mapsto \begin{cases} (n-1, m+1) & \text{at rate } cn, \\ (n+1, m-1) & \text{at rate } cKm, \\ (n-1, m) & \text{at rate } \binom{n}{2}. \end{cases} \tag{3.2.5}$$

Note that the three possible transitions correspond respectively to the drift of the X-component, the drift of the Y-component, and the diffusion part of the system (3.2.4). This connection is exploited in the following result.

Denote by $\mathbb{P}^{n,m}$ the distribution for which $(N_0, M_0) = (n, m)$ holds $\mathbb{P}^{n,m}$-a.s., and denote the corresponding expected value by $\mathbb{E}^{n,m}$. It is easy to see that, eventually, $N_t + M_t = 1$ (as $t \to \infty$), $\mathbb{P}^{n,m}$-a.s. for all $n, m \in \mathbb{N}_0$. We now show that $(N_t, M_t)_{t\geq 0}$ is the *moment dual* of $(X_t, Y_t)_{t\geq 0}$.

Theorem 3.2.7. *For every $(x, y) \in [0, 1]^2$, every $n, m \in \mathbb{N}_0$ and every $t \geq 0$*

$$\mathbb{E}_{x,y}\left[X_t^n Y_t^m\right] = \mathbb{E}^{n,m}\left[x^{N_t} y^{M_t}\right]. \tag{3.2.6}$$

Proof. Let $f(x, y; n, m) := x^n y^m$. Applying for fixed $n, m \in \mathbb{N}_0$ the generator A of $(X_t, Y_t)_{t\geq 0}$ to f acting as a function of x and y gives

$$\begin{aligned} Af(x,y) =& c(y-x)\frac{df}{dx}f(x,y) + \frac{1}{2}x(1-x)\frac{d^2 f}{dx^2}f(x,y) + cK(x-y)\frac{df}{dy}f(x,y) \\ =& c(y-x)nx^{n-1}y^m + \frac{1}{2}x(1-x)n(n-1)x^{n-2}y^m \\ &+ cK(x-y)x^n m y^{m-1} \\ =& cn(x^{n-1}y^{m+1} - x^n y^m) + \binom{n}{2}(x^{n-1}y^m - x^n y^m) \\ &+ cKm(x^{n+1}y^{m-1} - x^n y^m). \end{aligned}$$

Note that the right-hand side is precisely the generator of $(N_t, M_t)_{t\geq 0}$ applied to f acting as a function of n and m, for fixed $x, y \in [0, 1]$. Hence the duality follows from standard arguments, see e.g. [31], Proposition 1.2. □

3.2.3 Long-term behaviour and fixation probabilities

The long-term behaviour of our system (3.2.4) is not obvious. While a classical Wright Fisher diffusion (introduced in Definition 1.1.30) will get absorbed at the boundaries after finite time a.s. (in fact with finite expectation), hitting 1 with probability $X_0 = x$ (as in Lemma 1.1.21), this is more involved for our frequency process in the presence of a strong seedbank. Nevertheless, one can still compute its fixation probabilities as $t \to \infty$, at least in law, using very similar arguments as in the proof of Lemma 1.1.21. Obviously, $(0, 0)$ and $(1, 1)$ are absorbing states for the system (3.2.4). They are also the only absorbing states, since absence of drift requires $x = y$, and for the fluctuations to disappear, it is necessary to have $x \in \{0, 1\}$.

Proposition 3.2.8. *All mixed moments of $(X_t, Y_t)_{t\geq 0}$ solving (3.2.4) converge to the* same *finite limit depending only on x, y, K. More precisely, for each fixed $n, m \in \mathbb{N}$, we have*

$$\lim_{t\to\infty} \mathbb{E}_{x,y}[X_t^n Y_t^m] = \frac{y + xK}{1 + K}. \tag{3.2.7}$$

Proof. Let $(N_t, M_t)_{t\geq 0}$ be as in Definition 3.2.6, started in $(n,m) \in \mathbb{N}_0 \times \mathbb{N}_0$. To save notation, we write:

$$T := T_{MRCA}[n,m] = \inf\{t > 0 : N_t + M_t = 1\}.$$

Note that for any finite initial configuration (n,m), the stopping time T has finite expectation. Now, by Theorem 3.2.7,

$$\begin{aligned}
\lim_{t\to\infty} \mathbb{E}_{x,y}\left[X_t^n Y_t^m\right] &= \lim_{n\to\infty} \mathbb{E}^{n,m}\left[x^{N_t} y^{M_t}\right] \\
&= \lim_{t\to\infty} \mathbb{E}^{n,m}\left[x^{N_t} y^{M_t} \mid T \leq t\right] \mathbb{P}^{n,m}(T \leq t) \\
&\quad + \lim_{t\to\infty} \underbrace{\mathbb{E}^{n,m}\left[x^{N_t} y^{M_t} \middle| T > t\right]}_{\leq 1} \mathbb{P}^{n,m}(T > t) \\
&= \lim_{t\to\infty} \left(x\mathbb{P}^{n,m}(N_t = 1, T \leq t) + y\mathbb{P}^{n,m}(M_t = 1, T \leq t)\right) \\
&= \lim_{t\to\infty} \left(x\mathbb{P}^{n,m}(N_t = 1) + y\mathbb{P}^{n,m}(M_t = 1)\right) \\
&= \frac{xK}{1+K} + \frac{y}{1+K},
\end{aligned}$$

where the last equality holds by convergence to the invariant distribution of a single particle, jumping between the two states 'plant' and 'seed' at rate c resp. cK, which is given by $(K/(1+K), 1/(1+K))$ and independent of the choice of n, m. □

Corollary 3.2.3 (Fixation in law)**.** *Given c, K, (X_t, Y_t) converges in distribution as $t \to \infty$ to a two-dimensional random variable (X_∞, Y_∞), whose distribution is given by*

$$\mathcal{L}_{(x,y)}(X_\infty, Y_\infty) = \frac{y + xK}{1+K}\delta_{(1,1)} + \frac{1 + (1-x)K - y}{1+K}\delta_{(0,0)}. \tag{3.2.8}$$

Note that this is in line with the classical results for the Wright Fisher diffusion: As $K \to \infty$ (that is, the seedbank becomes small compared to the plant population), the fixation probability of a alleles approaches x. Further, if K becomes small (so that the seedbank population dominates the plant population), the fixation probability is governed by the initial fraction y of a-alleles in the seedbank.

Proof. It is easy to see that the only two-dimensional distribution on $[0,1]^2$, for which all moments are constant equal to $\frac{xK+y}{1+K}$, is given by

$$\frac{y + xK}{1+K}\delta_{(1,1)} + \frac{1 + (1-x)K - y}{1+K}\delta_{(0,0)}.$$

Indeed, uniqueness follows from the moment problem, which is uniquely solvable on $[0,1]^2$. Convergence in law follows from convergence of all moments due to Theorem 3.3.1 in [18] and the Stone-Weierstraß Theorem. □

Remark 3.2.9 (Almost sure fixation)**.** Observing that $(KX_t + Y_t)_{t\geq 0}$ is a bounded martingale, and given the shape of the limiting law (3.2.8), one can also get almost sure convergence of (X_t, Y_t) to (X_∞, Y_∞) as $t \to \infty$. However, as we will see later, fixation will not happen in finite time, since the block counting process $(N_t, M_t)_{t\geq 0}$, started from an infinite initial state, *does not come down from infinity* (see Section 3.4), which means that the whole (infinite) population does not have a most recent common ancestor. Thus, in finite time, initial genetic variability should never be completely lost. We expect that with some extra work, this intuitive reasoning could be made rigorous in an almost sure sense with the help of a "lookdown construction", and will be treated in future work. The fact that fixation doesn't occur in finite time can also be understood from (3.2.4), where we can compare the seed-component $(Y_t)_{t\geq 0}$ to the solution of the deterministic equation

$$dy_t = -cKy_t dt,$$

corresponding to a situation where the drift towards 0 is maximal (or to $dy_t = cK(1-y_t)dt$ where the drift towards 1 is maximal). Since $(y_t)_{t\geq 0}$ doesn't reach 0 in finite time if $y_0 > 0$, neither does $(Y_t)_{t\geq 0}$.

3.3 The seedbank coalescent

3.3.1 Definition and genealogical interpretation

In view of the form of the block counting process, it is now easy to guess the stochastic process describing the limiting gene genealogy of a sample taken from the Wright Fisher model with seedbank component. Recall the notation, for $k \geq 1$, $[k]$ is the set of partitions of $\{1, 2, ..., k\}$. For $\pi \in [k]$, $|\pi|$ is the number of blocks of the partition π. We define the space of *marked* partitions to be

This enables us to attach to each partition block a flag which can be either 'plant' or 'seed' (p or s), so that we can trace whether an ancestral line is currently in the active or dormant part of the population. For example, for $k = 5$, an element π of $\mathcal{P}_k^{\{p,s\}}$ is the marked partition $\pi = \{\{1,3\}^p, \{2\}^s, \{4,5\}^p\}$.

Consider two marked partitions $\pi, \pi' \in \mathcal{P}_k^{\{p,s\}}$, we say $\pi \succ \pi'$ if π' can be constructed by merging exactly 2 blocks of π carrying the p-flag, and the resulting block in π' obtained from the merging both again carries a p-flag. For example

$$\{\{1,3\}^p\{2\}^s\{4,5\}^p\} \succ \{\{1,3,4,5\}^p\{2\}^s\}.$$

We use the notation $\pi \bowtie \pi'$ if π' can be constructed by changing the flag of precisely one block of π, for example

$$\{\{1,3\}^p\{2\}^s\{4,5\}^p\} \bowtie \{\{1,3\}^s\{2\}^s\{4,5\}^p\}.$$

Definition 3.3.1 (The seedbank k-coalescent)**.** For $k \geq 2$ and $c, K \in (0, \infty)$ we define the *seedbank k-coalescent* $(\Pi_t^{(k)})_{t\geq 0}$ with seedbank intensity c and relative seedbank size $1/K$ to be the continuous time Markov chain with values in $\mathcal{P}_k^{\{p,s\}}$, characterised by the following transitions:

$$\pi \mapsto \pi' \text{ at rate } \begin{cases} 1 & \text{if } \pi \succ \pi', \\ c & \text{if } \pi \bowtie \pi' \text{ and one } p \text{ is replaced by one } s, \\ cK & \text{if } \pi \bowtie \pi' \text{ and one } s \text{ is replaced by one } p. \end{cases} \tag{3.3.1}$$

If $c = K = 1$, we speak of the *standard seedbank k-coalescent*.

Comparing (3.3.1) to (3.2.5) it becomes evident that (N_t, M_t) introduced in Definition 3.2.6 is indeed the block counting process of the seedbank coalescent.

Definition 3.3.2 (The seedbank coalescent)**.** We may define the *seedbank coalescent*, $(\Pi_t)_{t\geq 0} = (\Pi_t^{(\infty)})_{t\geq 0}$ with seedbank intensity c and relative seedbank size $1/K$ as the unique Markov process distributed according to the projective limit as k goes to infinity of the laws of the seedbank k-coalescents (with seedbank intensity c and relative seedbank size $1/K$). In analogy to Definition 3.3.1 we call the case of $c = K = 1$ the *standard seedbank coalescent*.

Remark 3.3.3. Note that the seedbank coalescent is a well-defined object. Indeed, for the projective limiting procedure to make sense, we need to show *consistency* and then apply the Kolmogorov extension theorem. This can be roughly sketched as follows. Define the process $(\overrightarrow{\Pi}_t^{(k)})_{t\geq 0}$ as the projection of $(\Pi_t^{(k+1)})_{t\geq 0}$, the $k+1$ seedbank coalescent, to the space $\mathcal{P}_k^{\{p,s\}}$. Mergers and flag-flips involving the singleton $\{k+1\}$ are only visible in $(\Pi_t^{(k+1)})_{t\geq 0}$, but do not affect $(\overrightarrow{\Pi}_t^{(k)})_{t\geq 0}$. Indeed, by the Markov-property, a change involving the singleton $\{k+1\}$ does not affect any of the other transitions. Hence, if $\overrightarrow{\Pi}_0^{(k)} = \Pi_0^{(k)}$, then

$$(\overrightarrow{\Pi}_t^{(k)})_{t\geq 0} = (\Pi_t^{(k)})_{t\geq 0}.$$

holds in distribution. By the Kolmogorov extension theorem the projective limit exists and is unique.

Note that it is obvious that the distribution of the block counting process of the seedbank coalescent, counting the number of blocks carrying the p and s-flags, respectively, agrees with the distribution the process $(N_t, M_t)_{t\geq 0}$ from Definition 3.2.6 (with suitable initial conditions).

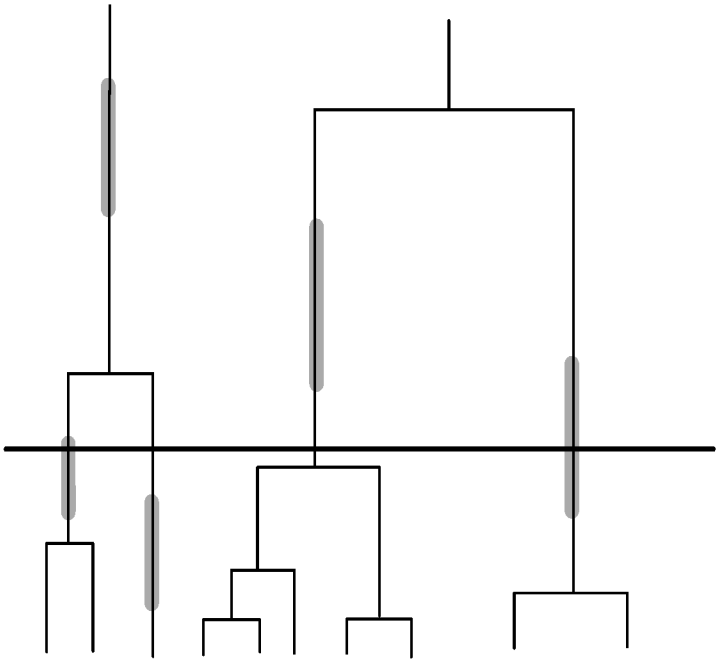

Figure 3.3: A possible realization of the standard 10-seedbank coalescent. Fat-gray lines indicate 'inactive lineages' (carrying an s-flag, which are prohibited from merging). At the time marked with the horizontal line the process is in the state $\{\{1,2\}^s\{3\}^p\{4,5,6,7,8\}^p\{9,10\}^s\}$.

Further, it is not hard to see that the seedbank coalescent appears as the limiting genealogy of a sample taken from the seedbank model in the same way as the Kingman coalescent describes the limiting genealogy of a sample taken from the classical Wright Fisher model (here, we merely sketch a proof, which is entirely standard).

Indeed, consider the genealogy of a sample of $k \ll N$ individuals, sampled from present generation 0. We proceed backward in time, keeping track in each generation of the ancestors of the original sample among the active individuals (plants) and among the seeds. To this end, denote by $\Pi_i^{(N,k)} \in \mathcal{P}_k^{\{p,s\}}$ the configuration of the genealogy at generation $-i$, where two individuals belong to the same block of the partition $\Pi_i^{(N,k)}$ if and only if their ancestral lines have met until generation $-i$, which means that all individuals of a block have exactly one common ancestor in this generation, and the flag s or p indicates whether said ancestor is a plant or a seed in generation $-i$. According to our forward in time population model, there are the following possible transitions from one generation to the previous one of this process:

- One (or several) plants become seeds in the previous generation.
- One (or several) seeds become plants in the previous generations.
- Two (or more) individuals have the same ancestor in the previous generation (which by construction is necessarily a plant), meaning that their ancestral lines merge.
- Any possible combination of these three events.

It turns out that only three of the possible transitions play a role in the limit as $N \to \infty$, whereas the others have a probability that is of smaller order.

Proposition 3.3.4. *In the setting of Proposition 3.2.4, additionally assume that*

$$\Pi_0^{(N,k)} = \{\{1\}^p, ..., \{k\}^p\},$$

$\mathbb{P}$-*a.s. for some fixed* $k \in \mathbb{N}$. *Then for* $\pi, \pi' \in \mathcal{P}_k^{\{p,s\}}$,

$$\mathbb{P}\big(\Pi_{i+1}^{(N,k)} = \pi' \,\big|\, \Pi_i^{(N,k)} = \pi\big) = \begin{cases} \frac{1}{N} + O(N^{-2}) & \text{if } \pi \succ \pi', \\ \frac{c}{N} + O(N^{-2}) & \text{if } \pi \bowtie \pi' \text{ and a } p \text{ is replaced by an } s, \\ \frac{cK}{N} + O(N^{-2}) & \text{if } \pi \bowtie \pi' \text{ and an } s \text{ is replaced by a } p, \\ O(N^{-2}) & \text{otherwise.} \end{cases} \tag{3.3.2}$$

for all $i \in \mathbb{N}_0$.

Proof. According to the definition of the forward in time population model, exactly c out of the N plants become seeds, and exactly c out of the $M = N/K$ seeds become plants. Thus whenever the current state $\Pi_i^{(N,k)}$ of the genealogical process contains at least one p-block, then the probability that a *given* p-block changes flag to s at the next time step is equal to $\frac{c}{N}$. If there is at least one s-block, then the probability that any given s-block changes flag to p is given by $\frac{cK}{N}$, and the probability that a given p-block chooses a *fixed* plant ancestor is equal to $\big(1 - \frac{c}{N}\big)\frac{1}{N}$ (where $1 - c/N$ is the probability that the ancestor of the block in question is a plant, and $1/N$ is the probability to choose one particular plant among the N).

From this we conclude that the probability of a coalescence of two given p-blocks in the next step is

$$\mathbb{P}\big(\text{two given } p\text{-blocks merge}\big) = \big(1 - \frac{c}{N}\big)^2 \frac{1}{N}.$$

Since we start with k blocks, and the blocks move independently, the probability that two or more blocks change flag at the same time is of order at most N^{-2}. Similarly, the probability of any combination of merger or block-flip events other than single blocks flipping or binary mergers is of order N^{-2} or smaller, since the number of possible events (coalescence or change of flag) involving at most k blocks is bounded by a constant depending on k but not on N. □

Corollary 3.3.1. *For any* $k \in \mathbb{N}$, *under the assumptions of Proposition 3.2.4,* $(\Pi_{\lfloor Nt \rfloor}^{(N,k)})_{t \geq 0}$ *converges weakly as* $N \to \infty$ *to the seedbank coalescent* $(\Pi_t^{(k)})_{t \geq 0}$ *started with* k *plants.*

Proof. From Proposition 3.3.4 it is easy to see that the discrete generator of $(\Pi_{\lfloor Nt \rfloor}^{(N,k)})$ converges to the generator of $(\Pi_t^{(k)})$, which is defined via the rates given in (3.3.1). Then standard results (see Theorem 3.7.8 in [18]) yield weak convergence of the process. □

3.3.2 Related coalescent models

The structured coalescent The seedbank coalescent is reminiscent of the *structured coalescent* arising from a two-island population model (see Subsection 1.3.3 or [75, 67, 54, 28, 29]). Indeed, consider two Wright Fisher type (sub-) populations of fixed relative size evolving on separate 'islands', where individuals (resp. ancestral lineages) may migrate between the two locations with a rate of order of the reciprocal of the total population size (the so-called 'weak migration regime'). Since offspring are placed on the same island as their parent, mergers between two ancestral lineages are only allowed if both are currently in the same island. This setup again gives rise to a coalescent process defined on 'marked partitions', with the marks indicating the location of the ancestral lines among the two islands. Coalescences are only allowed for lines carrying the same mark at the same time, and marks are switched according to the scaled migration rates. See [71] for an overview.

In our seedbank model, we consider a similar 'migration' regime between the two sub-populations, in our case called 'plants' and 'seeds'. However, in the resulting seedbank coalescent, coalescences can only happen while in the plant-population. This asymmetry leads to a behaviour that is qualitatively different to the usual two-island scenario (e.g. with respect to the time to the most recent common ancestor, whose expectation is always finite for the structured coalescent, even if the sample size goes to infinity, as we proved in Lemma 1.3.13).

The peripatric coalescent The seedbank coalescent was recently independently discovered, under the name *The peripatric coalescent* by Lambert and Ma (see [37]). It arises as the scaling limit of the ancestral process of populations with a central structure in which individuals get isolated for long

periods and then return to the main bulk of individuals. The seedbank coalescent is a simple and natural mathematical object, which will likely appear as scaling limit in different contexts. Thus, the properties of the seedbank coalescent have an interest that goes beyond the study of seedbanks. We will prove some of them in this chapter.

The coalescent with freeze Another related model is the *coalescent with freeze*, see [14], where blocks become completely inactive at some rate. This model is different from ours because once a block has become inactive, it cannot be activated again. Hence, it cannot coalesce at all, which clearly leads to a different long-time behaviour. In particular one will not expect to see a most recent common ancestor in such a coalescent.

3.4 Properties of the seedbank coalescent

3.4.1 Some interesting recursions

One can compute the expected time to most recent common ancestor recursively as follows. We will use the notation $(N_t^{(n,m)}, M_t^{(n,m)})$ to indicate the initial condition of the block counting process is (n, m).

Definition 3.4.1. We define the **time to the most recent common ancestor** of a sample of n plants and m seeds, to be

$$T_{MRCA}[(n,m)] = \inf\{t > 0 : (N_t^{(n,m)}, M_t^{(n,m)}) = (1,0)\}.$$

To shorten notation, we will write

$$t_{n,m} := \mathbb{E}[T_{MRCA}[(n,m)]], \tag{3.4.1}$$

Remark 3.4.1. This definition is completely analogous to Definition 1.3.12, where we were studying the structured coalescent. However, in the seedbank model coalescence is only possible in *the plant island*, and thus $T_{MRCA}[(n,m)] = \inf\{t > 0 : N_t^{(n,m)} + M_t^{(n,m)} = 1\}$.

Observe that we need to consider both types of lines in order to calculate $t_{n,m}$. Write

$$\lambda_{n,m} := \binom{n}{2} + cn + cKm, \tag{3.4.2}$$

and abbreviate

$$\alpha_{n,m} := \frac{\binom{n}{2}}{\lambda_{n,m}}, \quad \beta_{n,m} := \frac{cn}{\lambda_{n,m}}, \quad \gamma_{n,m} := \frac{cKm}{\lambda_{n,m}}. \tag{3.4.3}$$

Proposition 3.4.2. *Let $n, m \in \mathbb{N}_0$. Then we have the following recursive representations*

$$\mathbb{E}_{n,m}[T_{\text{MRCA}}] = t_{n,m} = \lambda_{n,m}^{-1} + \alpha_{n,m} t_{n-1,m} + \beta_{n,m} t_{n-1,m+1} + \gamma_{n,m} t_{n+1,m-1}, \tag{3.4.4}$$

with initial conditions $t_{1,0} = t_{0,1} = 0$.

Proof of Proposition 3.4.2. Let τ_1 denote the time of the first jump of the process $(N_t, M_t)_{t\geq 0}$. If started at (n,m), this is an exponential random variable with parameter $\lambda_{n,m}$. Applying the strong Markov property we obtain

$$\begin{aligned} t_{n,m} =& \mathbb{E}_{n,m}[\tau_1] + \mathbb{E}_{n,m}\big[\mathbb{E}_{N_{\tau_1}, M_{\tau_1}}[T_{MRCA}]\big] \\ =& \lambda_{n,m}^{-1} + \alpha_{n,m} t_{n-1,m} + \beta_{n,m} t_{n-1,m+1} + \gamma_{n,m} t_{n+1,m-1}. \end{aligned}$$

□

Remark 3.4.2. A recursion for the variance of T_{MRCA} can also be computed. It is not included in this thesis, but is given in the supplementary material of [5].

Since the process $N_t + M_t$ is non-increasing in t, these recursions can be solved iteratively. In fact,

$$t_{2,0} = 1 + \frac{2}{K} + \frac{1}{K^2} = \frac{(K+1)^2}{K^2}. \tag{3.4.5}$$

Notably, $t_{2,0}$ is constant as c varies (Table 3.1) and in particular does not converge for $c \to 0$ to the Kingman case. This effect is similar to the corresponding behaviour of the structured coalescent with two islands if the migration rate goes to 0, cf. [52]. However, the Kingman coalescent values are recovered as the seedbank size decreases (e.g. for $K = 100$ in Table 3.1).
The fact that $t_{2,0} = 4$ for $K = 1$ can be understood heuristically if c is large: In that situation, transitions between active and dormant states happen very fast, thus at any given time the probability that a line is active is about $1/2$, and therefore the probability that both lines of a given pair are active (and thus able to merge) is approximately $1/4$. We can therefore conjecture that for $K = 1$ and $c \to \infty$ the genealogy of a sample is given by a time change by a factor 4 of Kingman's coalescent.
Tables 3.1 shows values of $t_{n,0}$ obtained from (3.4.4) for various parameter choices and sample sizes. The relative size of the seedbank (K) has a significant effect on $\mathbb{E}_{n,0}[T_{MRCA}]$; a large seedbank ($K$ small) increases $\mathbb{E}_{n,0}[T_{MRCA}]$, while the effect of c is to dampen the increase in $\mathbb{E}_{n,0}[T_{MRCA}]$ with sample size (Table (3.1)).

In order to investigate the genetic variability of a sample, in terms e.g. of the number of segregating sites and the number of singletons, it is useful to have information about the total tree length and the total length of external branches. Let $L^{(\mathrm{a})}$ denote the total length of all branches while they are active, and $L^{(\mathrm{d})}$ the total lenght of all branches while they are dormant. Their expectations

$$l^{(a)}_{n,m} := \mathbb{E}_{n,m}[L^{(\mathrm{a})}], \quad l^{(\mathrm{d})}_{n,m} := \mathbb{E}_{n,m}[L^{(\mathrm{d})}]. \tag{3.4.6}$$

may be calculated using the following recursions for $n, m \in \mathbb{N}_0$, and with $\lambda_{n,m}$ given by (3.4.2),

Proposition 3.4.3 (Recursion: Total tree length)**.** *For $n, m \in \mathbb{N}$ we have*

$$l^{(a)}_{n,m} = n\lambda^{-1}_{n,m} + \alpha_{n,m} l^{(a)}_{n-1,m} + \beta_{n,m} l^{(a)}_{n-1,m+1} + \gamma_{n,m} l^{(a)}_{n+1,m-1} \tag{3.4.7}$$

$$l^{(\mathrm{d})}_{n,m} = m\lambda^{-1}_{n,m} + \alpha_{n,m} l^{(\mathrm{d})}_{n-1,m} + \beta_{n,m} l^{(\mathrm{d})}_{n-1,m+1} + \gamma_{n,m} l^{(\mathrm{d})}_{n+1,m-1}, \tag{3.4.8}$$

Proof of Proposition 3.4.3. The result can easily be obtained observing that each stretch of time of length τ in which we have a constant number of n active blocks and m dormant blocks contributes with $n\tau$ to the total active tree length, and with $m\tau$ to the total dormant tree length. Thus we have

$$l^{(a)}_{n,m} = n\mathbb{E}_{n,m}[\tau_1] + \mathbb{E}_{n,m}\left[\mathbb{E}_{N_{\tau_1}, M_{\tau_1}}[L^{(\mathrm{a})}]\right],$$

and we proceed as in the proof of Proposition 3.4.2. From these quantities we easily obtain the expected total tree length as $l^{(a)}_{n,m} + l^{(\mathrm{d})}_{n,m}$. □

Similar recursions hold for their variances as well as for the corresponding values of the total length of external branches, which can be found in the supplementary material of [5] together with the respective proofs. From (3.4.7) and (3.4.8) one readily obtains

$$l^{(a)}_{2,0} = \frac{2(1+K)}{K}, \quad l^{(d)}_{2,0} = \frac{2(1+K)}{K^2}. \tag{3.4.9}$$

We observe that $l^{(a)}_{2,0}$ and $l^{(d)}_{2,0}$ given in (3.4.9) are independent of c as also seen for $t_{2,0}$ cf. (3.4.5).

The numerical solutions of (3.4.7) and (3.4.8) indicate that for $n \geq 2$,

$$l^{(a)}_{n,0} = K l^{(d)}_{n,0}. \tag{3.4.10}$$

Hence the expected total length of the active and the dormant parts of the tree are proportional, and ratio is given by the relative seedbank size.

One can further investigate this relation by writing

$$\frac{2t_{n,0}}{l^a_{n,0}} = \frac{l^a_{n,0} + l^d_{n,0}}{l^a_{n,0}} = \frac{K+1}{K}$$

Further, using equation 3.4.5 one observes that

$$\frac{2t_{n,0}}{l^a_{n,0}} = \sqrt{\tau_{2,0}}$$

which, taking $n = 2$ reduces to

$$2\sqrt{t_{2,0}} = l^a_{2,0}$$

This equation tells an nice story: in the Kingman case "$K = \infty$", we know that $t_{2,0} = 1$, and $l^a_{2,0} = 2$, meaning that the 2 individuals are active the whole time (obviously, as in this case there is no seedbank). If $K = 1$, we know that $t_{2,0} = 4$, and we see that $l^a_{2,0} = 4$, which means that half of the time the individuals where active and half inactive. Finally, as the seedbank grows (K goes to 0) one can see that the proportion of active time decreases quickly (as the inverse of a square root).

Table 3.1: The expected time to most recent common ancestor ($\mathbb{E}_{n,0}[T_{\mathrm{MRCA}}]$) of the seedbank coalescent, obtained from (3.4.4), with seedbank size K, sample size n and dormancy initiation rates c as shown. All sampled lines are from the active population (sample configuration $(n,0)$). For comparison, $\mathbb{E}[T_{\mathrm{MRCA}}[n]] = 2(1-1/n)$ when associated with the Kingman coalescent ($K=\infty$). The multiplication $\times 10^4$ only applies to the first table with $K = 0.01$.

$K = 0.01$, $\times 10^4$			
	sample size n		
c	2	10	100
0.01	1.02	2.868	5.185
0.1	1.02	2.731	4.487
1	1.02	2.187	2.666
10	1.02	1.878	2.085
100	1.02	1.84	2.026

$K = 1$			
	sample size n		
c	2	10	100
0.01	4	10.21	17.18
0.1	4	9.671	14.97
1	4	8.071	10.02
10	4	7.317	8.221
100	4	7.212	7.954

$K = 100$			
	sample size n		
c	2	10	100
0.01	1.02	1.846	2.052
0.1	1.02	1.838	2.026
1	1.02	1.836	2.02
10	1.02	1.836	2.02
100	1.02	1.836	2.02
$K = \infty$	1	1.80	1.98

3.4.2 Coming down from infinity

The notion of *coming down from infinity* was discussed by Pitman [59] and Schweinsberg [63]. They say that an exchangeable coalescent process *comes down from infinity* if the corresponding block counting process (of an infinite sample) has finitely many blocks immediately after time 0 (i.e. the number of blocks is finite almost surely for each $t > 0$). Further, the coalescent is said to *stay infinite* if the number of blocks is infinite a. s. for all $t \geq 0$. Schweinsberg also gives a necessary and sufficient criterion for so-called "Lambda-coalescents" to come down from infinity. In particular, the Kingman coalescent does come down from infinity. However, note that the seedbank coalescent does not belong to the class of Lambda-coalescents, so that Schweinsberg's result does not immediately apply. For an overview of the properties of general exchangeable coalescent processes see e.g. [4] (the reader is invited to compare the following Theorem with Lemma 1.3.13).

Theorem 3.4.4. *The seedbank coalescent does* not *come down from infinity. In fact, its block counting process $(N_t, M_t)_{t\geq 0}$ stays infinite for every $t \geq 0$, $\mathbb{P}$-a.s. To be precise, for each starting configuration (n, m) where $n + m$ is (countably) infinite,*

$$\mathbb{P}\big(\forall t \geq 0 : \ M_t^{(n,m)} = \infty\big) = 1.$$

The proof of this theorem is based on a coupling with a dominated simplified *coloured seedbank coalescent* process introduced below. In essence, the coloured seedbank coalescent behaves like the normal seedbank coalescent, except we mark the individuals with a colour to indicate whether they have (entered and) left the seedbank at least once. This will be useful in order to obtain a process where the number of plant-blocks is non-increasing. We will then prove that even if we consider only those individuals that have never made a transition from seed to plant (but possibly from plant to seed), the corresponding block counting process will stay infinite. This will be achieved by proving that infinitely many particles enter the seedbank before any positive time. Since they subsequently leave the seedbank at a linear rate, this will take an infinite amount of time.

Definition 3.4.5 (A coloured seedbank coalescent)**.** In analogy to the construction of the seedbank coalescent, we first define the set of *coloured*, marked partitions as

$$\begin{aligned}\mathcal{P}_k^{\{p,s\}\times\{w,b\}} &:= \big\{(\pi, \vec{u}, \vec{v}) \mid (\pi, \vec{u}) \in \mathcal{P}_k^{\{p,s\}}, \vec{v} \in \{w, b\}^k\big\}, \quad k \in \mathbb{N},\\ \mathcal{P}^{\{p,s\}\times\{w,b\}} &:= \big\{(\pi, \vec{u}, \vec{v}) \mid (\pi, \vec{u}) \in \mathcal{P}^{\{p,s\}}, \vec{v} \in \{w, b\}^{\mathbb{N}}\big\}.\end{aligned}$$

It corresponds to the marked partitions introduced earlier, where now each element of $\{1, 2, ..., k\}$, resp. $\mathbb{N}$, has an additional flag indicating its colour: w for *white* and b for *blue*. We write $\pi \succ \pi'$, if π' can be constructed from π by merging two blocks with a p-flag in π that result into a block with a p-flag in π', while each individual retains its colour. It is important to note that the p- or s-flags are assigned to *blocks*, the colour-flags to *individuals*, i. e. elements of $[k]$ resp. $\mathbb{N}$. We use $\pi \ltimes \pi'$, to denote that π' results from π by changing the flag of a block from p to s and leaving the colours of all individuals unchanged and $\pi \rtimes \pi'$, if π' is obtained from π, by changing the flag of a block from s to p *and colouring all the individuals in this block blue*, i.e. setting their individual flags to b. In other words, after leaving the seedbank, individuals are always coloured blue.

For $k \in \mathbb{N}$ and $c, K \in (0, \infty)$ we now define the *coloured seedbank k-coalescent with seedbank intensity c and seedbank size $1/K$*, denoted by $(\underline{\Pi}_t)_{t\geq 0}$, as the continuous time Markov chain with values in $\mathcal{P}_k^{\{p,s\}\times\{w,b\}}$ and transition rates given by

$$\pi \mapsto \pi' \text{ at rate } \begin{cases} 1, & \text{if } \pi \succ \pi', \\ c, & \text{if } \pi \ltimes \pi', \\ cK, & \text{if } \pi \rtimes \pi'. \end{cases} \tag{3.4.11}$$

The *coloured seedbank coalescent with seedbank intensity c and seedbank size $1/K$* is then the unique Markov process on $\mathcal{P}^{\{p,s\}\times\{w,b\}}$ given by the projective limit of the distributions of the k-coloured seedbank coalescents, as k goes to infinity.

Remark 3.4.6. 1. Note that the coloured seedbank coalescent is well-defined. Since the colour of an individual only depends on its own path and does not depend on the colour of other individuals (not

even those that belong to the same block), the consistency of the laws of the k-coloured seedbank coalescents boils down to the consistency of the seedbank k-coalescents discussed in Remark 3.3.3. In much the same way we then obtain the existence and uniqueness of the coloured seedbank coalescent from Kolmogorov's Extension Theorem.

2. The normal seedbank (k-)coalescent can be obtained from the coloured seedbank (k-)coalescent by omitting the flags indicating the colouring of the individuals. However, if we only consider those blocks containing *at least* one white individual, we obtain a coalescent similar to the seedbank coalescent, where lineages are discarded once they leave the seedbank.

For $t \geq 0$ define $\underline{N}_t$ to be the number of *white plants* and $\underline{M}_t$ the number of *white seeds* in $\underline{\Pi}_t$. We will use a superscript (n, m) to denote the processes started with n plants and m seeds $\mathbb{P}$-a.s., where $n, m = \infty$ means we start with a countably infinite number of plants, resp. seeds. We will always start in a configuration were all individual labels are set to w, i.e. with only white particles. Note that our construction is such that $(\underline{N}_t)_{t\geq 0}$ is non-increasing.

Proposition 3.4.7. *For any $n, m \in \mathbb{N} \cup \{\infty\}$, there exist a coupling of $(N_t^{(n,m)}, M_t^{(n,m)})_{t\geq 0}$ and $(\underline{N}_t^{(n,m)}, \underline{M}_t^{(n,m)})_{t\geq 0}$ such that*

$$\mathbb{P}\left(\forall t \geq 0 : \ N_t^{(n,m)} \geq \underline{N}_t^{(n,m)} \text{and } M_t^{(n,m)} \geq \underline{M}_t^{(n,m)}\right) = 1.$$

Proof. This result is immediate if we consider the coupling through the coloured seedbank coalescent and the remarks in 3.4.6. □

Proof of Theorem 3.4.4. Proposition 3.4.7 implies that it suffices to prove the statement for $(\underline{M}_t)_{t\geq 0}$ instead of $(M_t)_{t\geq 0}$. In addition, we will only have to consider the case of $m = 0$, since starting with more (possibly infinitely many) seeds will only contribute towards our desired result.

For $n \in \mathbb{N} \cup \{\infty\}$ let

$$\tau_j^n := \inf\{t \geq 0 : \underline{N}_t^{(n,0)} = j\}, \quad 1 \leq j \leq n-1,$$

be the first time that the number of active blocks of an n-sample reaches k. Note that $(\underline{N}_t)_{t\geq 0}$ behaves like the block counting process of a Kingman coalescent where in addition to the coalescence events, particles may "disappear" at a rate proportional to the number of particles alive. Since the corresponding values for a Kingman coalescent are finite $\mathbb{P}$-a.s., it is easy to see that the τ_j^n are, too. Clearly, for any n, $\tau_j^n - \tau_{j-1}^n$ has an exponential distribution with parameter

$$\lambda_j := \binom{j}{2} + cj.$$

At each time of a transition τ_j^n, we distinguish between two events: *coalescence* and *deactivation* of an active block, where by deactivation we mean a transition of $(\underline{N}_t^n, \underline{M}_t^n)_{t\geq 0}$ of type $(j+1, l) \mapsto (j, l+1)$ (for suitable $l \in [n]$), i.e. the transition of a plant to a seed.

Then

$$\mathbb{P}\big(\text{deactivation at } \tau_{j-1}^n\big) = \frac{cj}{\binom{j}{2} + cj} = \frac{2c}{j + 2c - 1}, \tag{3.4.12}$$

independently of the number of inactive blocks. Thus

$$X_j^n := \mathbf{1}_{\left\{\text{deactivation at } \tau_{j-1}^n\right\}}, \quad j = 2, ..., n, j < \infty,$$

are independent Bernoulli random variables with respective parameters $2c/(j+2c-1), j = 2, ..., n$. Note that X_j^n depends on j, but the random variable is independent of the random variable τ_{j-1} due to the memorylessness of the exponential distribution. Now define A_t^n as the (random) number of deactivations up to time $t \geq 0$ that is, for $n \in \mathbb{N} \cup \{\infty\}$,

$$A_t^n := \sum_{j=2}^{n} X_j^n \mathbf{1}_{\{\tau_{j-1}^n < t\}}. \tag{3.4.13}$$

For $n \in \mathbb{N}$, since $\lambda_j \geq \binom{j}{2}$, it follows from a comparison with the block counting process of the Kingman coalescent, denoted by $(|K_t^n|)_{t\geq 0}$ (if started in n blocks), that for all $t \geq 0$,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\big(\tau^n_{\lfloor \log n - 1\rfloor} \leq t\big) &\geq \lim_{n\to\infty} \mathbb{P}\big(|K_t^n| \leq \lfloor \log n - 1\rfloor\big)\\ &\geq \lim_{n\to\infty} \mathbb{P}\big(|K_t| \leq \log n - 1\big) = 1.\end{aligned}$$

where the last equality follows from the fact that the Kingman-coalescent $(K_t)_{t\geq 0}$ comes down from infinity, cf. [63, 59]. For $t \geq 0$,

$$\mathbb{P}\Big(A_t^n \geq \sum_{j=\log n}^{n} X_j^n\Big) \geq \mathbb{P}\Big(\mathbf{1}_{\{\tau^n_{\log n-1}<t\}} \sum_{j=\log n}^{n} X_j^n \geq \sum_{j=\log n}^{n} X_j^n\Big) \tag{3.4.14}$$

$$\geq \mathbb{P}\big(\tau^n_{\log n-1} < t\big). \tag{3.4.15}$$

and hence, by (3.4.14)

$$\lim_{n\to\infty} \mathbb{P}\Big(A_t^n \geq \sum_{j=\log n}^{n} X_j^n\Big) = 1. \tag{3.4.16}$$

Note that due to (3.4.12),

$$\mathbb{E}\Big[\sum_{j=\log n}^{n} X_j^n\Big] = \sum_{j=\log n}^{n} \frac{2c}{j+2c-1} = 2c(\log n - \log\log n) + R(c,n), \tag{3.4.17}$$

where $R(c,n)$ converges to a finite value depending on the seedbank intensity c as $n \to \infty$. Since the X_j^n are independent Bernoulli random variables, we obtain for the variance

$$\begin{aligned}\mathbb{V}\Big[\sum_{j=\log n}^{n} X_j^n\Big] = \sum_{j=\log n}^{n} \mathbb{V}[X_j^n] &= \sum_{j=\log n}^{n} \frac{2c}{j+2c-1}\Big(1 - \frac{2c}{j+2c-1}\Big)\\ &\leq 2c\log n \text{ as } n \to \infty.\end{aligned} \tag{3.4.18}$$

For any $\epsilon > 0$ we can choose n large enough such that, $\mathbb{E}[\sum_{j=\log n}^{n} X_k] \geq (2c-\epsilon)\log n$ holds, which yields

$$\begin{aligned}\mathbb{P}\Big(\sum_{j=\log n}^{n} X_j^n < c\log n\Big) &\leq \mathbb{P}\Big(\sum_{j=\log n}^{n} X_j^n - \mathbb{E}\Big[\sum_{j=\log n}^{n} X_j^n\Big] < -(c-\epsilon)\log n\Big)\\ &\leq \mathbb{P}\Big(\Big|\sum_{j=\log n}^{n} X_j^n - \mathbb{E}\Big[\sum_{j=\log n}^{n} X_j^n\Big]\Big| > (c-\epsilon)\log n\Big)\\ &\leq \frac{2c}{(c-\epsilon)^2\log n},\end{aligned} \tag{3.4.19}$$

by Chebyshev's inequality. In particular, for any $\kappa \in \mathbb{N}$,

$$\lim_{n\to\infty} \mathbb{P}\Big(\sum_{j=\log n}^{n} X_j^n < \kappa\Big) = 0,$$

and together with (3.4.16) we obtain for any $t > 0$

$$\lim_{n\to\infty} \mathbb{P}(A_t^n < \kappa) = 0. \tag{3.4.20}$$

Since the $(A_t^n)_{t\geq 0}$ are coupled by construction for any $n \in \mathbb{N}\cup\{\infty\}$, we know in particular that $\mathbb{P}(A_t^\infty < \kappa) \leq \mathbb{P}(A_t^n < \kappa)$, for any $n \in \mathbb{N}, t \geq 0, \kappa \geq 0$ and therefore $\mathbb{P}(A_t^\infty < \kappa) = 0$, which yields

$$\forall t \geq 0 : \ \mathbb{P}(A_t^\infty = \infty) = 1. \tag{3.4.21}$$

Since in addition, $(A_t^\infty)_{t\geq 0}$ is non-decreasing in t, we can even conclude

$$\mathbb{P}(\forall t \geq 0 : \ A_t^\infty = \infty) = 1. \tag{3.4.22}$$

Thus we have proven that, for any time $t \geq 0$, there have been an infinite amount of movements to the seedbank $\mathbb{P}$-a.s. Now we are left to show that this also implies the presence of an infinite amount of lineages in the seedbank, i.e. that a sufficiently large proportion is saved from moving back to the plants where it would be "instantaneously" reduced to a finite number by the coalescence mechanism.

Define $\mathcal{B}_t$ to be the blocks of a partition that visited the seedbank at some point before a fixed time $t \geq 0$ and were visible in the "white" seedbank coalescent, i.e.

$$\mathcal{B}_t := \{B \subseteq \mathbb{N} \mid \exists 0 \leq r \leq t : \ B^{\{s\}} \in \underline{\Pi}_r^{(\infty,0)} \text{ and contains at least one } \textit{white} \text{ particle } \}.$$

Since we started our coloured coalescent in $(\infty, 0)$, the cardinality of $\mathcal{B}_t$ is at least equal to A_t^∞ and therefore we know $\mathbb{P}(|\mathcal{B}_t| = \infty) = 1$. Since $\mathcal{B}_t$ is countable, we can enumerate its elements as $\mathcal{B}_t = \bigcup_{n\in\mathbb{N}}\{B_t^n\}$ and use this to define the sets $\mathcal{B}_t^n := \{B_t^1, \dots, B_t^n\}$, for all $n \in \mathbb{N}$. Since $\mathcal{B}_t$ is infinite $\mathbb{P}$-a.s., these $\mathcal{B}_t^n$ exist for any n, $\mathbb{P}$-a.s. Now observe that the following inequalities hold even pathwise by construction:

$$\underline{M}_t^{(\infty,0)} \geq \sum_{B\in\mathcal{B}_t} \mathbf{1}_{\{B^{\{s\}}\in\underline{\Pi}_t^{(\infty,0)}\}} \geq \sum_{B\in\mathcal{B}_t^n} \mathbf{1}_{\{B^{\{s\}}\in\underline{\Pi}_t^{(\infty,0)}\}}$$

and therefore the following holds for any $\kappa \in \mathbb{N}$:

$$\begin{aligned}\mathbb{P}(\underline{M}_t^{(\infty,0)} \leq \kappa) &\leq \mathbb{P}\Big(\sum_{B\in B_t^n} \mathbf{1}_{\{B^s\in\underline{\Pi}_t^{(\infty,0)}\}} \leq \kappa\Big) \\ &\overset{*}{\leq} \sum_{i=1}^{\kappa} \binom{n}{i} (e^{-ct})^i (1-e^{-ct})^{n-i} \xrightarrow{n\to\infty} 0\end{aligned}$$

which in turn implies $\mathbb{P}(\underline{M}_t^{(\infty,0)} = \infty) = 1$. In * we used that for each of the n blocks in $\mathcal{B}_t^n$ we know $\mathbb{P}(B \in \underline{\Pi}_t^{(\infty,0)}) \geq e^{-ct}$ and they leave the seedbank independently of each other, which implies that the sum is dominated by a Binomial random variable with parameters n and e^{-ct}.

Since the probability on the left does not depend on n, and the above holds for any $\kappa \in \mathbb{N}$, we obtain $\mathbb{P}(\underline{M}_t^{(\infty,0)} = \infty) = 1$ for all $t > 0$. Note that this also implies $\mathbb{P}(\underline{M}_t^{(\infty,0)} + \underline{N}_t^{(\infty,0)} = \infty) = 1$ for all $t > 0$, from which, through the monotonicity of the sum, we can immediately deduce the stronger statement

$$\mathbb{P}\big(\forall t > 0 : \ \underline{M}_t^{(\infty,0)} + \underline{N}_t^{(\infty,0)} = \infty\big) = 1.$$

On the other hand, we have seen that $\mathbb{P}(\underline{N}_t^{(\infty,0)} < \infty) = 1$, for all $t > 0$, which , again using its monotonicity, yields $\mathbb{P}(\forall t > 0 : \ \underline{N}_t^{(\infty,0)} < \infty) = 1$. Putting these two results together we obtain $\mathbb{P}(\forall t > 0 : \ \underline{M}_t^{(\infty,0)} = \infty) = 1$ □

3.4.3 Bounds on the time to the most recent common ancestor

In view of the previous subsection it is now quite obvious that the seedbank causes a relevant delay in the time to the most recent common ancestor of finite samples. We will mostly be interested in the case where the sample is drawn from plants only, and write $T_{MRCA}[n] := T_{MRCA}[(n, 0)]$. The main results of this section are asymptotic logarithmic bounds on the expectation of $T_{MRCA}[n]$. (The reader is invited to compare the following Theorem with Lemma 1.3.13)

Theorem 3.4.8. *For all $c, K \in (0, \infty)$, the seedbank coalescent satisfies*

$$\mathbb{E}\big[T_{MRCA}[n]\big] \asymp \log\log n. \tag{3.4.23}$$

Here, the symbol $\asymp$ denotes weak asymptotic equivalence of sequences, meaning that we have

$$\liminf_{n\to\infty} \frac{\mathbb{E}\big[T_{MRCA}[n]\big]}{\log\log n} > 0, \tag{3.4.24}$$

and

$$\limsup_{n\to\infty} \frac{\mathbb{E}\big[T_{MRCA}[n]\big]}{\log\log n} < \infty. \tag{3.4.25}$$

The proof of Theorem 3.4.8 will be given in Proposition 3.4.9 and Proposition 3.4.11. The intuition behind this result is the following. The time until a seed gets involved in a coalescence event is much longer than the time it takes for a plant to be involved in a coalescence, since a seed has to become a plant first. Thus the time to the most recent common ancestor of a sample of n plants is governed by the number of individuals that become seeds before coalescence, and by the time of coalescence of a sample of seeds.

Due to the quadratic coalescence rates, it is clear that the time until the ancestral lines of all sampled plants have either coalesced into one, or have entered the seedbank at least once, is finite almost surely. The number of lines that enter the seedbank until that time is a random variable that is asymptotically of order $\log n$, due to similar considerations as in (3.4.17). Thus we need to control the time to the most recent common ancestor of a sample of $O(\log n)$ seeds. The linear rate of migration then leads to the second log.

Turning this reasoning into bounds requires some more work, in particular for an upper bound. As in the proof of Theorem 3.4.4, let $X_k, k = 1, ..., n$ denote independent Bernoulli random variables with parameters $2c/(k + 2c - 1)$. Similar to (3.4.13) define

$$A^n := \sum_{k=2}^{n} X_k. \tag{3.4.26}$$

Lemma 3.4.3. *Under our assumptions, for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(A^n \geq (2c + \epsilon)\log n) = 0$$

and

$$\lim_{n\to\infty} \mathbb{P}(A^n \leq (2c - \epsilon)\log n) = 0.$$

Proof. As in the proof of Theorem 3.4.4 before we have

$$\mathbb{E}[A^n] = \sum_{k=2}^{n} \frac{2c}{k + 2c - 1} = 2c\log n + R'(c, n),$$

where $R'(c, n)$ converges to a finite value depending on c as $n \to \infty$, and

$$\mathbb{V}(A^n) \sim 2c\log n \text{ as } n \to \infty.$$

Thus again by Chebyshev's inequality, for sufficiently large n (and recalling that c is our model parameter)

$$\begin{aligned}\mathbb{P}(A^n \geq (2c + \epsilon)\log n) &\leq \mathbb{P}(A^n - \mathbb{E}[A^n] \geq \epsilon\log n)\\ &\leq \mathbb{P}(|A^n - \mathbb{E}[A^n]| \geq \epsilon\log n)\\ &\leq \frac{2c}{\epsilon^2\log n}.\end{aligned}$$

This proves the first claim. The second statement follows similarly, cf. (3.4.19). □

Recall the process $(\underline{N}_t, \underline{M}_t)_{t\geq 0}$ from the previous subsection. The coupling of Proposition 3.4.7 leads to the lower bound in Theorem 3.4.8.

Proposition 3.4.9. *For all $c, K \in (0, \infty)$, the seedbank coalescent satisfies*

$$\liminf_{n\to\infty} \frac{\mathbb{E}\big[T_{MRCA}[n]\big]}{\log\log n} > 0. \tag{3.4.27}$$

Proof. The coupling with $(\underline{N}_t, \underline{M}_t)_{t\geq 0}$ yields

$$T_{MRCA}[n] \geq \underline{T}_{MRCA}[n],$$

where $\underline{T}_{MRCA}[n]$ denotes the time until $(\underline{N}_t, \underline{M}_t)$ started at $(n, 0)$ has reached a state with only one block left. By definition, A^n of the previous lemma gives the number of individuals that at some point become seeds in the process $(\underline{N}_t, \underline{M}_t)_{t\geq 0}$. Thus $\underline{T}_{MRCA}[n]$ is bounded from below by the time it takes until these A^n seeds migrate to plants (and then disappear). Since the seeds disappear independently of each other, we can bound $\underline{T}_{MRCA}[n]$ stochastically from below by the extinction time of a pure death process with death rate cK started with A^n individuals. For such a process started at $A^n = l \in \mathbb{N}$ individuals, the expected extinction time as $l \to \infty$ is of order $\log l$. Thus we have for $\epsilon > 0$ that there exists $C > 0$ such that

$$\begin{aligned}\mathbb{E}\big[T_{MRCA}[n]\big] \geq& \mathbb{E}\big[T_{MRCA}[n]\mathbf{1}_{\{A^n \geq (2c-\epsilon)\log n\}}\big] \\ \geq& C \log\log n \mathbb{P}\big(A^n \geq (2c-\epsilon)\log n\big),\end{aligned}$$

and the claim follows from the fact that by Lemma 3.4.3, $A^n \geq (c-\epsilon)\log n$ almost surely as $n \to \infty$. □

To prove the corresponding upper bound, we couple (N_t, M_t) to a functional of another type of coloured process.

Definition 3.4.10. Let $(\overline{N}_t, \overline{M}_t)_{t\geq 0}$ be the continuous-time Markov process with state space $E \subseteq \mathbb{N}\times\mathbb{N}$, characterised by the transition rates:

$$(n,m) \mapsto \begin{cases} (n-1, m+1) & \text{at rate } cn, \\ (n+1, m-1) & \text{at rate } cKm, \\ (n-1, m) & \text{at rate } \binom{n}{2} \cdot \mathbf{1}_{\{n \geq \sqrt{n+m}\}}. \end{cases}$$

This means that $(\overline{N}_t, \overline{M}_t)_{t\geq 0}$ has the same transitions as (N_t, M_t), but coalescence is suppressed if there are too few plants relative to the number of seeds. The effect of this choice of rates is that for $(\overline{N}_t, \overline{M}_t)_{t\geq 0}$, if $n \gtrsim \sqrt{m}$, then coalescence happens at a rate which is of the same order as the rate of migration from seed to plant.

Lemma 3.4.4. *The processes $(\overline{N}_t, \overline{M}_t)_{t\geq 0}$ and $(N_t, M_t)_{t\geq 0}$ can be coupled such that*

$$\mathbb{P}\left(\forall t \geq 0 :\ N_t^{(n,m)} \leq \overline{N}_t^{(n,m)} \text{ and } M_t^{(n,m)} \leq \overline{M}_t^{(n,m)}\right) = 1.$$

Proof. We construct both processes from the same system of blocks. Start with $n+m$ blocks labelled from $\{1, ..., n+m\}$, and with n of them carrying an s-flag, the others a p-flag. Let $S^i, P^i, i = 1, ..., n+m$ and $V^{i,j}, i, j = 1, ..., n+m, i < j$ be independent Poisson processes, S^i with parameter cK, P^i with parameter c, and $V^{i,j}$ with parameter 1. Moreover, let each block carry a colour flag, blue or white. At the beginning, all blocks are supposed to be blue. The blocks evolve as follows: At an arrival of S^i, if block i carries an s-flag, this flag is changed to p irrespective of the colour and the state of any other block. Similarly, at an arrival of P^i, if block i carries a p-flag, this is changed to an s-flag. At an arrival of $V^{i,j}$, and if blocks i and j both carry a p-flag, one observes the whole system, and proceeds as follows:

(i) If the total number of p-flags in the system is greater or equal to the square root of the total number of blocks, then blocks i and j coalesce, which we encode by saying that the block with the higher label (i or j) is discarded. If the coalescing blocks have the same colour, this colour is kept. Note that here the *blocks* carry the color, unlike in the coloured process of the previous sections, where the individuals were coloured. If the coalescing blocks have different colours, then the colour after the coalescence is blue.

(ii) If the condition on the number of flags in (i) is not satisfied, then there is no coalescence, but if both blocks were coloured blue, then the block (i or j) with the higher label is coloured white (this can be seen as a "hidden coalescence" in the process where colours are disregarded).

It is then clear by observing the rates that (N_t, M_t) is equal in distribution to the process which counts at any time t the number of blue blocks with p-flags and with s-flags respectively, and $(\overline{N}_t, \overline{M}_t)$ is obtained by counting the number of p-flags and s-flags of any colour. By construction we obivously have $\overline{N}_t \geq N_t$ and $\overline{M}_t \geq M_t$ for all t. □

Define now

$$\overline{T}_{MRCA}[m] := \inf\big\{t \geq 0 : (\overline{N}_t^{(0,m)}, \overline{M}_t^{(0,m)}) = (1,0)\big\}.$$

Lemma 3.4.5. *There exists a finite constant C independent of m such that*

$$\overline{T}_{MRCA}[m] \leq C \log m.$$

Proof. Define for every $k \in 1, 2, ..., m-1$ the hitting times

$$H_k := \inf\{t > 0 : \overline{N}_t + \overline{M}_t = k\}. \tag{3.4.28}$$

We aim at proving that $\mathbb{E}^{0,m}[H_{m-1}] \leq \frac{C}{\sqrt{m}}$ and $\mathbb{E}^{0,m}[H_{j-1} - H_j] \leq \frac{C}{j-1}$ for $j \leq m-1$, for some $0 < C < \infty$. Here and throughout the proof, C denotes a generic positive constant (independent of m) which may change from instance to instance. To simplify notation, we will identify $\sqrt{j}$ with $\lceil \sqrt{j} \rceil$, or equivalently assume that all occurring square roots are natural numbers. Moreover, we will only provide the calculations in the case of the standard seedbank-coalescent, that is, $c = K = 1$. The reader is invited to convince himself (or herself) that the argument can also be carried out in the general case.

We write $\overline{\lambda}_t$ for the total jump rate of the process $(\overline{N}, \overline{M})$ at time t, that is,

$$\overline{\lambda}_t = \binom{\overline{N}_t}{2} 1_{\{\overline{N}_t \geq \sqrt{\overline{N}_t + \overline{M}_t}\}} + \overline{N}_t + \overline{M}_t,$$

and set

$$\overline{\alpha}_t := \frac{\binom{\overline{N}_t}{2} 1_{\{\overline{N}_t \geq \sqrt{\overline{N}_t + \overline{M}_t}\}}}{\overline{\lambda}_t}, \quad \overline{\beta}_t := \frac{\overline{N}_t}{\overline{\lambda}_t}, \quad \overline{\gamma}_t := \frac{\overline{M}_t}{\overline{\lambda}_t}$$

for the probabilities that the first jump after time t is a coalescence, a migration from plant to seed or a migration from seed to plant, respectively. Even though all these rates are now random, they are well-defined conditional on the state of the process. The proof will be carried out in three steps.

Step 1: Bound on the time to reach $\sqrt{m}$ plants. Let

$$D_m := \inf\{t > 0 : \overline{N}_t^{(0,m)} \geq \sqrt{m}\} \tag{3.4.29}$$

denote the first time the number of plants is at least $\sqrt{m}$. Due to the restriction in the coalescence rate, the process $(\overline{N}_t^{(0,m)}, \overline{M}_t^{(0,m)})_{t \geq 0}$ has to first reach a state with at least $\sqrt{m}$ plants before being able to coalesce, hence $D_m < H_{m-1}$ a.s. Hence for any $t \geq 0$, conditional on $t \leq D_m$ we have $\overline{\lambda}_t = m$ and $\overline{N}_t < \sqrt{m}$. Thus $\overline{M}_t > m - \sqrt{m}$ a.s. and we note that at each jump time of $(\overline{N}_t, \overline{M}_t)$ for $t \leq D_m$

$$\overline{\gamma}_s \geq \frac{m - \sqrt{m}}{m} = 1 - \frac{1}{\sqrt{m}} \text{ a.s. } \forall s \leq t$$

and

$$\overline{\beta}_s \leq \frac{1}{\sqrt{m}} \text{ a.s. } \forall s \leq t.$$

The expected number of jumps of the process $(\overline{N}_t, \overline{M}_t)$ until D_m is therefore bounded from above by the expected time it takes a discrete time asymmetric simple random walk started at 0 with probability $1 - 1/\sqrt{m}$ for an upward jump and $1/\sqrt{m}$ for a downward jump to reach level $\sqrt{m} - 1$. It is a well-known fact (see for example [19], Ch. XIV.3) that this expectation is bounded by $C\sqrt{m}$ for some $C \in (0, \infty)$. Since the time between each of the jumps of $(\overline{N}_t, \overline{M}_t)$, for $t < D_m$, is exponential with rate $\overline{\lambda}_t = m$, we get

$$\mathbb{E}^{0,m}[D_m] \leq C\sqrt{m} \cdot \frac{1}{m} = \frac{C}{\sqrt{m}}\,. \tag{3.4.30}$$

Step 2: Bound on the time to the first coalescence after reaching $\sqrt{m}$ plants. At time $t = D_m$, we have $\overline{\lambda} = \binom{\sqrt{m}}{2} + \sqrt{m} + m - \sqrt{m}$, and thus

$$\overline{\beta}_t = \frac{\sqrt{m}}{\binom{\sqrt{m}}{2} + m} = \frac{2\sqrt{m}}{3m - \sqrt{m}} \leq \frac{C}{\sqrt{m}} \text{ a.s.},$$

and

$$\overline{\alpha}_t = \frac{m-\sqrt{m}}{3m-\sqrt{m}} \geq \frac{1}{3}\big(1-\frac{1}{\sqrt{m}}\big) \text{ a.s.}$$

Denote by J_m the time of the first jump after time D_m. At J_m there is either a coalescence taking place (thus reaching a state with $m-1$ individuals and hence in that case $H_{m-1} = J_m$), or a migration. In order to obtain an upper bound on H_{m-1}, as a "worst-case scenario", we can assume that if there is no coalescence at J_m, the process is restarted from state $(0,m)$, and then run again until the next time that there are at least $\sqrt{m}$ plants (hence after J_m, the time until this happens is again equal in distribution to D_m). If we proceed like this, we have that the number of times that the process is restarted is stochastically dominated by a geometric random variable with parameter $\frac{1}{3}(1-\frac{1}{\sqrt{m}})$, and since

$$\mathbb{E}^{0,m}[J_m - D_m] = \lambda_{D_m}^{-1} = \frac{1}{\binom{\sqrt{m}}{2}+m} \leq \frac{C}{m},$$

we can conclude (using (3.4.30)) that

$$\begin{aligned}\mathbb{E}^{0,m}[H_{m-1}] &\leq \mathbb{E}^{0,m}[J_m]\frac{3\sqrt{m}}{\sqrt{m}-1} \\ &= \big(\mathbb{E}^{0,m}[D_m] + \mathbb{E}^{0,m}[J_m - D_m]\big)\frac{3\sqrt{m}}{\sqrt{m}-1} \\ &\leq \frac{C}{\sqrt{m}}. \end{aligned} \tag{3.4.31}$$

Step 3: Bound on the time between two coalescences. Now we want to estimate $\mathbb{E}^{0,m}[H_{j-1} - H_j]$ for $j \leq m-1$. Obviously at time H_j-, for $j \leq m-1$, there are at least $\sqrt{j+1}$ plants, since $\overline{N}_t + \overline{M}_t$ can decrease only through a coalescence. Therefore[1] $\overline{N}_{H_j} \geq \sqrt{j}-1$. Hence either we have $\overline{N}_{H_j} \geq \sqrt{j}$ and coalescence is possible in the first jump after H_j, or $\overline{N}_{H_j} = \sqrt{j}-1$, in which case $\overline{\gamma}_{H_j} \geq \frac{j-\sqrt{j}}{j} = 1-\frac{1}{\sqrt{j}}$, meaning that if coalescence is not allowed at H_j, with probability at least $1-\frac{1}{\sqrt{j}}$ it will be possible after the first jump after reaching H_j. Thus the probability that coalescence is allowed either at the first or the second jump after time H_j is bounded from below by $1-\frac{1}{\sqrt{j}}$.

Assuming that coalescence is possible at the first or second jump after H_j, denote by L_j the time to either the first jump after H_j if $\overline{N}_{H_j} \geq \sqrt{j}$, or the time of the second jump after H_j otherwise. Then in the same way as before, we see that $\overline{\alpha}_{L_j} \geq 1-\frac{C}{\sqrt{j}}$. Thus the probability that H_{j-1} is reached no later than two jumps after H_j is at least $\big(1-\frac{C}{\sqrt{j}}\big)^2$. Otherwise, in the case where there was no coalescence at either the first or the second jump after H_j, we can obtain an upper bound on H_{j-1} by restarting the process from state $(0,j)$. The probability that the process is restarted is thus bounded from above by $\frac{C}{\sqrt{j}}$. We know from equation (3.4.31) that if started in $(0,j)$, there is $\mathbb{E}^{0,j}[H_{j-1}] \leq \frac{C}{\sqrt{j}}$. Noting that $\overline{\lambda}_{H_j} \geq j$, and we need to make at most two jumps, we have that $\mathbb{E}^{0,m}[L_j] \leq 2/j$. Thus we conclude

$$\begin{aligned}\mathbb{E}^{0,m}[H_{j-1} - H_j] &\leq \mathbb{E}^{0,m}[L_j]\big(1-\frac{C}{\sqrt{j}}\big)^2 + \frac{C}{\sqrt{j}}\mathbb{E}^{0,j}[H_{j-1}] \\ &\leq \frac{2}{j-1}\big(1-\frac{C}{\sqrt{j}}\big) + \Big(\frac{C}{\sqrt{j}}\Big)^2 \\ &\leq \frac{C}{j-1}. \end{aligned} \tag{3.4.32}$$

These three bounds allow us to finish the proof, since when starting $(\overline{N}_t, \overline{M}_t)$ in state $(0,m)$ our calculations show that

$$\begin{aligned}\mathbb{E}\big[\overline{T}_{MRCA}[m]\big] &= \mathbb{E}^{0,m}[H_1] = \mathbb{E}[H_m] + \sum_{j=2}^{m-1}\mathbb{E}[H_{j-1} - H_j] \\ &\leq \frac{C}{\sqrt{m}} + C\sum_{j=2}^{m-1}\frac{1}{j-1} \sim C\log m \end{aligned} \tag{3.4.33}$$

[1] keeping in mind our convention that Gauß-brackets are applied if necessary, and hence $\overline{N}_{H_j} \geq \sqrt{j+1}-1 \geq \sqrt{j}-1$.

as $m \to \infty$. □

This allows us to prove the upper bound corresponding (qualitatively) to the lower bound in (3.4.27).

Proposition 3.4.11. *For $c, K \in (0, \infty)$, the seedbank coalescent satisfies*

$$\limsup_{n\to\infty} \frac{\mathbb{E}\big[T_{MRCA}[n]\big]}{\log\log n} < \infty. \tag{3.4.34}$$

Proof. Assume that the initial n individuals in the sample of the process $(\Pi_t^{(n)})_{t\geq 0}$ are labelled $1, ..., n$. Let

$$\mathcal{S}_r := \{k \in [n] : \exists 0 \leq t \leq r : k \text{ belongs to an } s\text{-block at time } t\}$$

denote those lines that visit the seedbank at some time up to t. Let

$$\varrho^n := \inf\{r \geq 0 : |\mathcal{S}_r^c| = 1\}$$

be the first time that all those individuals which so far had not entered the seedbank have coalesced. Note that ϱ^n is a stopping time for the process $(\Pi_t^{(n)})_{t\geq 0}$, and $N_{\varrho^n}^{(n,0)}$ and $M_{\varrho^n}^{(n,0)}$ are well-defined as the number of plant blocks, resp. seed blocks of $\Pi_{\varrho^n}^{(n)}$. By a comparison of ϱ^n to the time to the most recent common ancestor of Kingman's coalescent cf. [71], $\mathbb{E}[\varrho^n] \leq 2$ for any $n \in \mathbb{N}$, and thus

$$\begin{aligned}\mathbb{E}\big[T_{MRCA}[(n,0)]\big] &\leq 2 + \mathbb{E}\big[T_{MRCA}[(N_{\varrho^n}^{(n,0)}, M_{\varrho^n}^{(n,0)})]\big] \\ &\leq 2 + \mathbb{E}\big[T_{MRCA}[(0, N_{\varrho^n}^{(n,0)} + M_{\varrho^n}^{(n,0)})]\big],\end{aligned} \tag{3.4.35}$$

where the last inequality follows from the fact that every seed has to become a plant before coalescing. Recall A^n from (3.4.26) and observe that

$$N_{\varrho^n}^{(n,0)} + M_{\varrho^n}^{(n,0)} \leq A^n + 1 \quad \text{stochastically.} \tag{3.4.36}$$

This follows from the fact that for every individual, the rate at which it is involved in a coalescence is increased by the presence of other individuals, while the rate of migration is not affected. Thus by coupling $(N_t, M_t)_{t\geq 0}$ to a system where individuals, once having jumped to the seedbank, remain there forever, we see that $N_{\varrho^n} + M_{\varrho^n}$ is at most $A^n + 1$.

By the monotonicity of the coupling with $(\overline{N}_t, \overline{M}_t)$, we thus see from (3.4.35), for $\epsilon > 0$,

$$\begin{aligned}\mathbb{E}\big[T_{MRCA}[n]\big] &\leq 2 + \mathbb{E}\big[\overline{T}_{MRCA}[A^n + 1]\big] \\ &= 2 + \mathbb{E}\big[\overline{T}_{MRCA}[A^n + 1]\mathbf{1}_{\{A^n \leq (2c+\epsilon)\log n\}}\big] \\ &\quad + \mathbb{E}\big[\overline{T}_{MRCA}[A^n + 1]\mathbf{1}_{\{A^n > (2c+\epsilon)\log n\}}\big].\end{aligned} \tag{3.4.37}$$

From Lemma 3.4.5 we obtain

$$\mathbb{E}\big[\overline{T}_{MRCA}[A^n + 1]\mathbf{1}_{\{A^n \leq (2c+\epsilon)\log n\}}\big] \leq C\log(2c - \epsilon)\log n \leq C\log\log n,$$

and since $A^n \leq n$ in any case, we get

$$\mathbb{E}\big[\overline{T}_{MRCA}[A^n + 1]\mathbf{1}_{\{A^n > (2c+\epsilon)\log n\}}\big] \leq C\log n \cdot \mathbb{P}(A^n > (2c + \epsilon)\log n) \leq C.$$

This completes the proof. □

Remark 3.4.12. In the same manner as in the proof of Theorem 3.4.8, one can show that for any $a, b \geq 0$,

$$\mathbb{E}\big[T_{MRCA}[an, bn]\big] \asymp \log\big(\log(an) + bn\big).$$

Part II

Modeling the Lenski experiment

Chapter 4

An individual based model for the Lenski experiment, and the deceleration of the relative fitness

4.1 Introduction

This chapter consists essentially on the paper [24].

The *Lenski experiment* (see [40, 41, 39] for a detailed description) is a cornerstone in experimental evolution. It investigates the long-term evolution of 12 initially identical populations of the bacterium E. coli in identical environments. One of the basic concepts of the Lenski experiment is that of the *daily cycles*. Every day starts by sampling approximately $5 \cdot 10^6$ cells from the bacteria available in the medium that was used the day before. This sample is propagated in a minimal glucose medium. The bacteria then reproduce (by binary splitting) with an exponential population growth. The reproduction continues until the medium is deployed, i.e., when there is no more glucose available. Then the reproduction stops and a phase of starvation starts. This phase lasts until the beginning of the next day, when the new sample is transferred to fresh medium. Around $5 \cdot 10^8$ cells are present at the end of each day.

Up to now the experiment has been going on for more than 60'000 generations (or 9000 days, see [39]). One important feature is that samples of ancestral populations were stored in low temperature, forcing the bacteria to go to a dormant state. Afterwards, the bacteria could be made to reproduce under competition with later generations in order to experimentally determine the fitness of an evolved strain relative to the founder ancestor of the population by comparing their growth rates in the following manner [40]: A population of size A_0 of the unevolved strain and a population of size B_0 of the evolved strain perform a direct competition in the minimal glucose medium. The respective population sizes at the end of the day, that is, after the glucose is consumed, are denoted by A_1 and B_1. The (empirical) *relative fitness* $F(B|A)$ of strain B with respect to strain A is then given by the ratio of the exponential growth rates, calculated as

$$F(B|A) \quad = \quad \frac{\log(B_1/B_0)}{\log(A_1/A_0)}. \tag{4.1.1}$$

Considerable changes of the relative fitness have been observed in the more than 25 years of the experiment ([41, 3, 74]). As expected, the relative fitness of the population increases over time, but one of the features that have been observed is a pronounced deceleration in the increase of the relative fitness, see Figure 2 in [74]. In particular it has been observed that it increases sublinearly over time. Several questions have arisen in this context ([3, 74]): How can the change of relative fitness be explained or approximated? Which factors account for the deceleration in the increase of the relative fitness?

In [3], the authors perform an analysis on the change of the relative fitness for the first 20'000 generations of the experiment, and of the mutations that go to fixation during the same period. They conjecture that effects of dependence between mutations, like *clonal interference* and *epistasis*, contribute crucially to the deceleration of the gain of relative fitness.

In [74], the authors analyse the change of the relative fitness for the first 50'000 generations of the experiment, and fit the observations to a power law function. They also conjecture that clonal interference and epistasis contribute crucially to the quantitative behavior of relative fitness, and support this conjecture by sketching a mathematical model which predicts a power law function for the relative fitness.

In this paper, we propose a basic mathematical model for a population that captures essential features of the Lenski experiment, in particular the daily cycles. It models an asexually reproducing population whose growth in each cycle is stopped after a certain time, and a new cycle is started with a sample of the original population. We include (beneficial) mutations into the model by assuming that an individual may mutate with a certain (small) probability and draws a certain (small) reproductory benefit from the mutation. We then calculate the probability of fixation of a beneficial mutation, and its time to fixation. Using this, we can prove that under some conditions on the parameters of mutation and selection, with high probability there will be no clonal interference in the population, which means in our situation that, with high probability, beneficial mutations only arrive when the population is homogeneous (in the sense that all its individuals have the same reproduction rate). This result implies that we are essentially dealing with a model of adaptive evolution, which allows a thorough mathematical analysis. In particular, using convergence results for Markov chains in the spirit of [35], we are able to prove that the relative fitness of the population, on a suitable timescale in terms of the population size, converges locally uniformly to a deterministic curve (see Figure 4.2).

In this way we arrive at an explanation of a power law behavior (with a deceleration in the increase) of the relative fitness. This explanation is in terms of the experiment's design, and does not invoke clonal interference, nor a direct epistatic effect of the befeficial mutations.

More specifically, in our model every beneficial mutation which is succesful in the sense that it goes to fixation, will increases the individual reproduction rate by the same amount (ρ, say), irrespective of the current value r of the individual reproduction rate. In this sense the model is "non-epistatic". However, there will be an indirect epistatic effect caused by the design of the experiment: since the amount of glucose, which the bacteria get for their population growth, remains the same from day to day, a population with a high individual reproduction rate will consume this amount more quickly than a population with a low individual reproduction rate. In other words, the daily duration of the experiment (that is the time $t = t_i$ during which the population grows at day i) will depend on the current level $r = r_i$ of the individual reproduction rate, and will become shorter as r increases. Indeed, the ratio of the two expected growth factors in one day is $\exp((r+\rho)t)/\exp(rt) = \exp(\rho t)$. Even though ρ does not depend on r by our assumption, this ratio does depend on r, because, as stated above, the duration $t = t_i$ of the daily cycles becomes smaller as r increases. We are well aware that clonal interference as well as direct epistatic effects will also be at work in the Lenski experiment, and should be modelled. On the other hand, we are convinced that our results will help to separate these effects from the indirect epistatic effect caused by the constant daily nutrition of the population.

In the remainder of this introductory section we discuss our mathematical approach and main results, and put our methods into the context of related work. The formal statement of the model and the main results will be given in Section 4.2, and the proofs in Section 4.3. The most intricate proof is that of Theorem 4.2.10 which relies on a coupling of the daily sampling scheme with near-critical Galton Watson processes that is successful over a sufficiently long time period. Some tools from the theory of branching processes (Yule and Galton Watson processes) are presented in the Appendix.

4.1.1 A neutral model for the daily cycles

We build our model on few basic assumptions: Every individual reproduces independently by binary splitting at a given rate until the end of a growth cycle, which corresponds to one *day* (in the sense of [39]). Our daily cycle model is determined by specifying the reproduction rate of each individual, and a stopping rule to end the growth of the population. To illustrate this we assume for the moment a *neutral situation*, i.e. all individuals have the same reproduction rate. The experiment is laid out such that the total number of bacteria at the end of one day is roughly the same for every day. This suggests the following mathematical assumptions: Each day starts with a population of N individuals. These individuals reproduce by binary splitting at some fixed rate r until the maximum capacity is reached. We assume that this happens (and that the "Lenski day" is over) as soon as the total number of cells in the medium is close to γN for some constant $\gamma > 1$ (a precise definition and a discussion of the

corresponding stopping rules will be given in Section 4.2.1). The description of the experiment suggests to think of $N = 5 \cdot 10^6$, and $\gamma \approx 100$, since at the end of each day, one gets around $5 \cdot 10^8$ bacteria, see supplementary material of [40]. The subsequent day is started by sampling N individuals from the approximately γN total amount available, and the procedure is repeated.

This setting induces a genealogical process, which we study on the evolutionary time scale, that is with one unit of time corresponding to $N = 5 \cdot 10^6$ days. On this time scale, the genealogical process turns out to be approximately a constant time change of the Kingman coalescent, where the constant is $c_\gamma := 2(1 - \frac{1}{\gamma})$. In this sense, N/c_γ plays the role of an *effective population size*. With the stated numbers, this is much larger than the number (≈ 9000) of "Lenski days" that have passed so far. In other words, in the neutral model so far only a small fraction of one unit of the evolutionary timescale has passed. Still, this model provides a good basis to introduce mutation and selection. In fact, we will see that the design of the experiment (via the stopping rule that defines the end of each day) affects the selective advantage provided by a beneficial mutation and in this way has an influence that goes well beyond the determination of the effective population size in the neutral model.

Our genealogical model arises naturally from the daily cycle setting, see Figure 4.1. Schweinsberg [64] obtained a Cannings dynamics by sampling generation-wise N individuals from a supercritical Galton Watson forest, and analysed the arising coalescents as $N \to \infty$. Our model is similar in spirit, with the binary splitting leading to Yule processes. We will introduce the additional feature that some individuals reproduce at a faster rate; in this sense Schweinsberg's sampling approach to neutral coalescents is naturally extended to a case with selection.

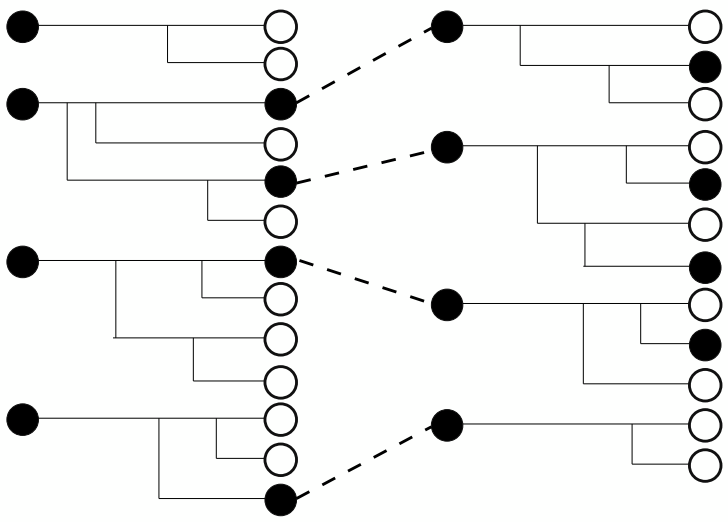

Figure 4.1: Two of the daily cycles (or "days"), with $N = 4$ and $\gamma = 3$. The N-sample at the end of day 1 constitutes the parental population at the beginning of day 2.

4.1.2 Mutants versus standing population

Next we consider a modification of the previous model, supposing that at a certain day a fraction of the population reproduces at rate r, while the complementary fraction (founded by some beneficial mutant in the past) reproduces faster, say at rate $r + \varrho_N$, with $\varrho_N > 0$. Our assumptions will be that the increment of the reproduction rate ϱ_N is small, but not too small, more precisely we will assume that $\varrho_N \sim N^{-b}$ for some $0 < b < 1/2$ ($\sim$ denoting asymptotic equivalence, i.e. the convergence of the ratio to 1 as $N \to \infty$). We assume that the reproduction rate is heritable. Based on the observation that with the stopping procedure indicated above a "Lenski day" lasts approximately $\frac{\log \gamma}{r}$ units of time of the Yule process, we will prove in Proposition 4.2.8 that the expected number of offspring at the beginning of the next day of an individual with reproduction rate $r + \varrho_N$ is increased for large N by approximately $\varrho_N \frac{\log \gamma}{r}$ compared

to an individual with reproduction rate r. In this sense the *effective selective advantage* of a beneficial mutation is approximately $\varrho_N \frac{\log \gamma}{r}$.

Let us emphasize that here one obtains a dependence on the reproduction rate r of the standing population due to the relation between r and the "length of a day", i.e. the time span it takes the total population to reach the maximum capacity. The implication of this result is that the selective advantage provided by reproducing ϱ_N units faster is comparatively large if the standing population is not well adapted and thus reproduces at a low rate, and is comparatively small if the population is well adapted in the sense that it already reproduces fast.

4.1.3 Genetic and adaptive evolution

In order to study the genetic and adaptive evolution of a population under the conditions of the Lenski experiment, we consider a model with *moderately strong selection – weak mutation* and constant additive fitness effect of the mutations. We assume that the population reproduces in daily cycles as described above, and that at each day with probability μ_N a beneficial mutation occurs within the ancestral population of that day, where $\mu_N \to 0$ as $N \to \infty$. Following the ansatz described above, we assume that an individual affected by such a beneficial mutation increases its reproduction rate and that of its offspring by ϱ_N. Some of these mutations will go to fixation (in which case they will be called "successful"), while the others are lost from the population. Calculating the probability of fixation of a beneficial mutation is a classical problem, studied already at the beginning of the last century by Haldane in the Wright Fisher model. These questions still have a major interest in modern times, and have recently been studied in different contexts (see for example [36] or [56]).

Assume now that the initial population on day i consists of $N-1$ individuals that reproduce at rate r and one mutant that reproduces at rate $r+\varrho_N$. We will see in Theorem 4.2.10 that the probability of fixation of such a mutant is asymptotically

$$\frac{\rho_N \log \gamma}{r} \frac{\gamma}{\gamma - 1} \tag{4.1.2}$$

as $N \to \infty$. A crucial role in the proof of our result is played by an intricate approximation of the number of the mutants' descendants by near-critical Galton Watson process, as long as their number is relatively small compared to the total population.

In Proposition 4.2.13, we prove that in a certain regime of the model parameters, namely if $\varrho_N \sim N^{-b}, \mu_N \sim N^{-a}$, with $b \in (0, 1/2)$ and $a > 3b$, the time it takes for a mutation to go to fixation or extinction is with high probability shorter than the time between two mutation events which is of order μ_N^{-1}. This result allows us to exclude clonal interference even on the time scale $\lfloor t\varrho_N^{-2}\mu_N^{-1} \rfloor$, and to approximate the reproduction rate process of our original model by a simple Markov chain which can be interpreted as an idealized process where successful mutations fixate immediately on the scale of their arrival rate, and unsuccessful ones are neglected.

In this respect, the analysis presented in this paper can be seen in the framework of the theory of stochastic adaptive dynamics, as studied by Champagnat, Méléard and others, see [10, 11] and references therein. Let us emphasize, however, that we prove the validity of our approximation by taking *simultaneous limits* of the population size $N \to \infty$, the rate of mutation $\mu_N \to 0$, and the increment of the reproduction rate $\varrho_N \to 0$, which requires some care, and is carried out by taking the specifics of our model into account.

4.1.4 Deterministic approximation on longer time scales

The calculation of the fixation probability in Theorem 4.2.10 and the exclusion of clonal interference in Proposition 4.2.13, as well as the resulting Markov chain approximation of the reproduction rate process are the key steps in the analysis of the long-term behaviour of the population in the Lenski experiment. This allows to derive the process counting the number of eventually successful beneficial mutations until a certain day, and the process of the relative fitness of the evolved population compared to the initial fitness.

It turns out, as we prove in Theorem 4.2.14 that for large N, on the time scale $\lfloor t\varrho_N^{-1}\mu_N^{-1} \rfloor$, the number of successful mutations is approximately a Poisson process with constant rate $\frac{\gamma \log \gamma}{(\gamma-1)r_0}$, if the observation of the population starts on some day where the reproduction rate is constant and equal to $r_0 > 0$.

In order to define the fitness of an evolved strain relative to the unevolved one, we assume that the unevolved population, taken from the first day of the experiment, is homogeneous and evolves at rate r_0. In view of (4.1.1) we define the fitness of the population at the beginning of day i with respect to that at the beginning of day 0 as

$$F_i := \frac{\log \frac{1}{N} \sum_{j=1}^{N} e^{R_{i,j} t}}{\log e^{r_0 t}} \tag{4.1.3}$$

where $R_{i,j}, j = 1, \dots, N$ are the reproduction rates of the individuals present at the beginning of day i, and t is a given time for which the two populations are allowed to grow together. (This time could also depend on i, which would not affect our results.) For brevity we call F_i the *relative fitness at day i*.

We prove in Theorem 4.2.15 that the time-rescaled process $(F_{\lfloor t \varrho_N^{-2} \mu_N^{-1} \rfloor})_{t \geq 0}$ converges locally uniformly as $N \to \infty$ to the parabola

$$f(t) = \sqrt{1 + \frac{2\gamma \log(\gamma)}{(\gamma - 1) r_0^2} t}, \quad t \geq 0. \tag{4.1.4}$$

Hence our model, which should be regarded as idealized and basic, still succeeds to describe the observed sublinear increase of relative fitness quite well on a qualitative level, even without incorporating the effects of clonal interference or epistasis.

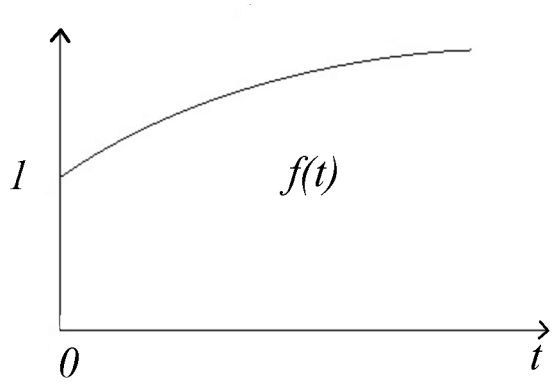

Figure 4.2: The limiting relative fitness curve for $N \to \infty$, if time is rescaled by $\lfloor t \varrho_N^{-2} \mu_N^{-1} \rfloor, t \geq 0$. The curve $f(t)$ is given by (4.1.4).

4.1.5 Diminishing returns and epistasis.

In this subsection we summarize the heuristics which leads to the formula (4.1.2) for the fixation probability in our individual-based model, and compare it with the ansatz of Wiser et al. [74].

Our basic assumption is that every beneficial mutation adds a fixed amount ϱ_N to the reproduction rate r of the individual that undergoes the mutation. When all (or nearly all) individuals that are present at day i have reproduction rate r, then this day ends (approximately) at time $\sigma := \frac{\log \gamma}{r}$, because then $e^{r\sigma} = \gamma$. Consequently, over this day the growth factor of a mutant population whose reproduction rate is $r + \varrho_N$ is $e^{(r+\varrho_N)\sigma}$, and the ratio of these two growth factors is $e^{\varrho_N \sigma} \approx 1 + \frac{\varrho_N \log \gamma}{r}$, revealing that the *selective advantage* of the mutant is $s_N := \frac{\varrho_N \log \gamma}{r}$. In the branching process approximation for the onset of the mutant, $1 + s_N$ is the offspring mean, while the quantity $c_\gamma = 2(1 - \frac{1}{\gamma})$ that appeared already in Sec.4.1.1 converges for $N \to \infty$ to the offspring variance, see the discussion after Theorem 4.2.5. In view of Lemma A.3.6 in the Appendix, this explains the form (4.1.2) of the fixation probability.

It is interesting to note that our model leads to quite similar conclusions as the one proposed in [74], although the basic hypotheses are somewhat different. Motivated by [21] the authors of [74] *assume* that the $(n+1)$-st successful mutation increases the individual reproduction rate by a factor $1 + \hat{S}_{n+1}$, where S_{n+1} is distributed exponentially with some parameter α_n, and the distribution of $\hat{S}_{n+1}$ is that of S_{n+1} conditioned to the event that the mutation goes to fixation (surviving also clonal interference). They make the following assumption in order to model *diminishing returns*: The sequence $\alpha_n, n \in \mathbb{N}_0$, satisfies

$$\alpha_{n+1} = \alpha_n (1 + g\langle S_{n+1} \rangle), \tag{4.1.5}$$

where g is a positive constant and $\langle S_{n+1} \rangle$ is the expected value of $\hat{S}_{n+1}$. According to [74], the parameter g serves to model the phenomenon of *epistasis*, which corresponds to a non-linearity in the fitness effects. Through (4.1.5), it is a priori assumed that the expected value of the beneficial effect of a mutation decreases as the number of successful mutations increases. Arguing heuristically by a branching process approximation, the authors of [74] obtain an approximation of the relative fitness by the function

$$\overline{w} = (ct + 1)^{1/2g}. \tag{4.1.6}$$

Here c depends on clonal interference and epistasis. In [74] the approximation is compared to real data, taking different pairs (g, c) and proving that the power law approximation in equation (4.1.6) fits better to data than the hyperbolic curve proposed in [3].

Our Theorem 4.2.15 is consistent with (4.1.6), as we prove that, under the assumptions of our model,

$$\overline{w} = (c't + 1)^{1/2}. \tag{4.1.7}$$

Notably, the "diminishing returns" for the case $g = 1$ emerge in our model under the assumption that every beneficial mutation adds a constant amount ϱ_N to the intraday individual reproduction rate, which corresponds to the absence of epistasis in this part of the model. This shows that the observed power law behaviour of the relative fitness can to some extent be explained by the mere design of the experiment, based on a simple non-epistatic intraday model – a fact which may also be seen as a strengthening of the argument of Wiser et al [74] that a power law is an appropriate approximation to the evolution of relative fitness.

In order to arrive at a power law (4.1.6) for more general g, we have to extend our model slightly. Indeed, in Corollary 4.2.16 we prove that a gain in the reproduction rate of $x^{-q}\varrho_N$, for some $q > -1$, if the present relative fitness is x, leads to a power law fitness curve with exponent $1/(2(1 + q))$, which compares to (4.1.6) by taking $q = g - 1$.

For a recent study that proposes a general framework for quantifying patterns of macroscopic epistasis from observed differences in adaptability, including a discussion of fitness and mutation trajectories in the Lenski experiment, see [25]. We refer also to the discussion in [12] of various epistatic models that would explain a declining adaptability in microbial evolution experiments, and to the discussion in [49] concerning the evolutionary dynamics on epistatic versus non-epistatic fitness landscapes with finitely many genotypes.

4.2 Models and main results

4.2.1 Mathematical model of daily population cycles

In this section, we construct a mathematical model for the daily reproduction and growth cycle of a bacterial population in the Lenski experiment, and state some first results, in particular on fixation probabilities of beneficial mutations. These are the foundations for our main results to be presented in Section 4.2.5.

4.2.2 Neutral model

We start by introducing the neutral model, where all individuals in the population reproduce at the same rate. The model consists of a continuous time intraday dynamics, and a discrete time interday dynamics, the latter is governed by a stopping- and a sampling rule. We number the daily cycles, or "days" as we call them for simplicity, by $i \in \mathbb{N}_0$. Fix $N \in \mathbb{N}$, and $r > 0$. We assume that every daily cycle starts with exactly N individuals that reproduce at rate r, the *basic reproduction rate*. More precisely, we decree that, independently for every day $i \in \mathbb{N}_0$, the (neutral) intraday *population size process* has the distribution of a *Yule process*, denoted by $(Z_t^{(N)})_{t \geq 0}$, with reproduction parameter r, started with $Z_0^{(N)} = N$ individuals. Consequently, for every $t > 0$, the random variable $Z_t^{(N)}$ follows a negative binomial distribution with parameters N and e^{-rt} (see Corollary A.4 in Appendix A). In Appendix A.3.1, we collect the properties of Yule process that are relevant for this paper.

Fix now $\gamma > 1$, and define stopping times

$$\varsigma_N := \inf\{t > 0 : Z_t^{(N)} \geq \gamma N\} \tag{4.2.1}$$

and

$$\sigma^{(N)} := \inf\{t > 0 : \mathbb{E}[Z_t^{(N)}] \geq \gamma N\}. \tag{4.2.2}$$

Note that ς_N is a random variable, while $\sigma^{(N)}$ is deterministic. In fact, since $\mathbb{E}[Z_t^{(N)}] = Ne^{rt}$, we see immediately that $\sigma^{(N)}$ does not depend on N and equals

$$\sigma := \frac{\log \gamma}{r}. \tag{4.2.3}$$

Definition 4.2.1 (Neutral model). *Fix $N \in \mathbb{N}$, $r > 0$, $\gamma > 1$. In the neutral model, independently for every $i \in \mathbb{N}_0$, the population size at the end of day i is given by a copy of the random variable $Z_\sigma^{(N)}$, where $(Z_t^{(N)})_{t\geq 0}$ is defined above.*

In other words, at every day the neutral population is started with N individuals that reproduce by binary splitting at rate r (which leads to the above Yule process), with the population growth stopped at time σ that depends on γ and r.

Remark 4.2.2 (Stopping rules). The two stopping times ς_N and σ give rise to two different stopping rules for the population: The *stopping rule* 1 stops the population growth at time ς_N, that is the time when population size has reached exactly $\lceil \gamma N \rceil$. On the other hand, *stopping rule 2* uses σ instead, which implies that the size of the stopped population, given by $Z_\sigma^{(N)}$, has a negative binomial distribution with parameters N and $\frac{1}{\gamma}$. While ς_N might be a more natural choice for the stopping time of the population growth, σ is easier to deal with. In this paper we will work under stopping rule 2, but we expect the essentials of our results to be true for ς_N as well. In fact, as we show in Lemma A.3.5, ς_N converges to σ in distribution.

4.2.3 The genealogy

Before turning our attention to the model with selection, we briefly discuss the neutral genealogy. If we label the individuals within this process, we can keep track of their ancestral relationship by specifying a sampling rule.

Definition 4.2.3 (Sampling rule). *The parent population of day $i+1$ is a uniform sample of size N taken from the population at the end of day i.*

Let $\nu^i = (\nu_1^i, \cdots, \nu_N^i)$, $i = 0, 1, 2, \ldots$, be a sequence of vectors such that ν_j^i is the number of offspring in the population at the beginning of day i of individual j from the population at the beginning of day $i-1$. Since $(\nu^i)_{i\in\mathbb{N}_0}$ are independent and identically distributed, and for each i the components of ν^i are exchangeable and sum to N, we are facing a Cannings model, where the "days" play the role of generations (see [71] for more background on Cannings models and coalescents). We can now fix a generation i and consider the genealogy of a sample of $n(\leq N)$ individuals. Here, for conceptual and notational convenience, we shift the "present generation" to the time origin and extend the Cannings dynamics (which is time-homogeneous) to all the preceding generations as well.

Definition 4.2.4 (Ancestral process). *Sample n individuals at generation 0 and denote them by $l_1, \cdots, l_n$. Let $[n]$ be the set of partitions of $\{1, 2, \cdots, n\}$ and $B^{(N,n)} = (B_g^{(N,n)})_{g\in\mathbb{N}_0}$ be the process taking values in $[n]$ such that any j, k being in the same block in $B_g^{(N,n)}$ if and only if there is a common ancestor at generation $-g$ for individuals l_j, l_k. Then $B^{(N,n)}$ is the ancestral process of the chosen sample.*

It turns out that the genealogical process converges after a suitable time-scaling to the classical Kingman coalescent (see [71] for a definition and more details on the relevance of Kingman's coalescent in population genetics). The time-rescaling depends on the population size N and is determined by a constant depending on γ.

Theorem 4.2.5 (Convergence to Kingman's coalescent). *For all $n \in \mathbb{N}$, the sequence of ancestral processes $\left(B^{(N,n)}_{\lfloor Nt/2(1-\frac{1}{\gamma})\rfloor}\right)_{t\geq 0}$ converges weakly on the space of càdlàg paths as $N \to \infty$ to Kingman's n-coalescent.*

The proof of Theorem 4.2.5 is given in Appendix A. Here we give a brief heuristic explanation of the time change factor $2(1-1/\gamma)/N$. This factor is asymptotically equal to $c_{\gamma,N}$, the *pair coalescence probability in one generation*, which in turn equals the probability that the second of two sampled individuals belongs to the same (one generation) offspring as the first one. Hence, in the limit $N \to \infty$, $c_{\gamma,N}$ is asymptotically equal to the ratio $(\mathbb{E}\hat{G}-1)/(N\mathbb{E}G)$, where G is the one-generation offspring number of a single individual, and $\hat{G}$ is a size-biased version of G. If G has a geometric distribution with expectation γ (which is the case in our setting, as can be seen from Lemma A.3.3 in the Appendix), then $\mathbb{E}\hat{G} = \mathbb{E}G^2/\mathbb{E}G = 2\gamma - 1$, and hence $c_{\gamma,N} \sim 2(1-\frac{1}{\gamma})/N$. (In particular, for large γ, G/γ is asymptotically exponential, $\mathbb{E}\hat{G} \sim 2\mathbb{E}G$, and $c_{\gamma,N} \sim \frac{2}{N}$.)

4.2.4 Including selective advantage

We now drop the assumption that the relative fitness is constant over the whole population, and include some selective advantage. Fix $r > 0, \gamma > 1$ as before. For $N \in \mathbb{N}$ let $\varrho_N \geq 0$. Throughout this chapter, we will assume that the sequence $(\varrho_N)_{N\in\mathbb{N}}$ satisfies the condition

$$\exists b \in (0, 1/2) : \varrho_N \sim N^{-b} \text{ as } N \to \infty. \tag{4.2.4}$$

We extend our basic population model in the following way. Assume that at day i a number k among the N individuals of the initial population have a selective advantage in the sense that they reproduce at rate $r+\varrho_N$, and the remaining $N-k$ individuals reproduce at rate r. We call the selectively advantageous individuals the *mutants*, and the others the wild-type individuals. We assume that fitness is heritable, meaning that offspring (unless affected by a mutation) retain the fitness of their parent. The intraday population size process at day i is then of the form

$$Y_t := Y_t^{(N,k)} = M_t^{(k)} + Z_t^{(N-k)}, \quad t \geq 0, \tag{4.2.5}$$

where $(Z_t^{(N-k)})_{t\geq 0}$ is a Yule process with reproduction rate r, started with $Z_0^{(N-k)} = N-k$ individuals, while $(M_t^{(k)})_{t\geq 0}$ is a Yule process with reproduction rate $r+\varrho_N$, started with $M_0^k = k$ individuals, and independent of $(Z_t^{(N-k)})_{t\geq 0}$. Note that for fixed r and ϱ_N the distribution of $(Y_t)_{t\geq 0}$ is uniquely determined by the initial number $M_0^{(k)} = k$ of mutants.

We apply stopping rule 2 to this model, which translates into stopping population growth at a deterministic time depending on k (and N), namely at

$$\sigma_k := \sigma_k^{(N)} = \inf\{t \geq 0 : \mathbb{E}[Y_t] \geq \gamma N\}. \tag{4.2.6}$$

This is still a deterministic time, though somewhat harder to calculate than σ, which equals σ_0 in this notation. Due to our construction, at the end of day i the total population has size Y_{σ_k}, among which there are $M_{\sigma_k}^{(k)}$ mutants, and $Z_{\sigma_k}^{(N-k)}$ wild-type individuals.

One of the main tasks of this paper will be to calculate the *number of mutants* at the beginning of day i, for $i \in \mathbb{N}_0$. Assuming that we know the population $Y_{\sigma_k} = M_{\sigma_k}^{(k)} + Z_{\sigma_k}^{(N-k)}$ at the end of day $i-1$, we apply Definition 4.2.3, which means that given $M_{\sigma_k}^{(k)} = M$, and $Z_{\sigma_k}^{(N-k)} = Z$, we sample uniformly N out of the $M+Z$ individuals. Denote by K_i the number of mutants contained in this sample. Fixing K_0 and repeating this independently for $i \in \mathbb{N}$ defines the interday process $(K_i)_{i\in\mathbb{N}_0}$ counting the number of mutants in the model with selection at the beginning of each day. Summarizing, this process can be described as follows:

Proposition 4.2.6 (Model with selection)**.** *Fix $\gamma > 1, r > 0$ and $\varrho_N, N \in \mathbb{N}$ satisfying (4.2.4). Fix $K_0 \in \{1, ..., N\}$. Assume K_{i-1} has been constructed, and takes the value k. Let M follow a negative binomial distribution with parameters k and $e^{-(r+\varrho_N)\sigma_k}$, and let Z follow a negative binomial distribution with parameters $N-k$ and $e^{-r\sigma_k}$ independent of M. Conditional on M and Z, the number K_i is determined by sampling from the hypergeometric distribution with parameters N, M and $M+Z$.*

Proof. This follows from the construction, noting that $(M_t)_{t\geq 0}$ and $(Z_t)_{t\geq 0}$ evolve independently until the deterministic time σ_k, and recalling that sampling N individuals without replacement out of M of one type and Z of another type is described by the hypergeometric distribution. □

Remark 4.2.7 (More than two types). The definition of the model with selection generalizes in an obvious way to situations where there are more than two different types of individuals in the population. If there are ℓ different types reproducing at ℓ different (fixed) rates, the population within one day grows like ℓ independent Yule processes with suitable initial values and reproduction rates, the stopping time is defined accordingly, and the sampling remains uniform over the whole population.

Since the mutants reproduce faster, their proportion will increase (stochastically) during the day. Hence, sampling uniformly at random from the population at the end of day i we expect to sample more than the initial number of mutants, meaning that the fitness of the population will increase over time.

Proposition 4.2.8 (Selective advantage). *Under assumption* (4.2.4),

$$\mathbb{E}[K_1|K_0 = 1] - 1 \sim \varrho_N \frac{\log \gamma}{r} \quad \text{as } N \to \infty. \tag{4.2.7}$$

Remark 4.2.9. Recall Subsection 1.3.2 where we introduced the notion of selective advantage, in the context of generalized Wright Fisher models.

Under the condition $\{K_0 = 1\}$ the $N - K_1$ wild-type individuals that are sampled at the end of day 0 are exchangeably distributed upon the $N - 1$ wild-type ancestors that were present at the beginning of day 0. Hence, the expected (sampled) offspring of each of these wild-type ancestors is ~ 1 as $N \to \infty$, and thus, in view of Proposition 4.2.8, we can say that the *selective advantage* of a single mutant, resulting from the increase of its reproduction rate from r to $r + \varrho_N$, is given by $\varrho_N \frac{\log \gamma}{r}$.

The main result of this section concerns the fixation probability of a beneficial mutation affecting one individual at the beginning of day 0, and an estimate of the time that it takes for a successful mutation to go to fixation (or for an unsuccessful mutation to go extinct). Let

$$\pi_N := \mathbb{P}\big(\exists i \in \mathbb{N} : K_i = N \,|\, K_0 = 1\big) \tag{4.2.8}$$

denote the probability of fixation if the population size process is started with one mutant at day 0 and write

$$\tau^N_{\text{fix}} := \inf\{i \geq 1 : K_i = N\} \in [0, \infty] \tag{4.2.9}$$

for the time of fixation, and

$$\tau^N_{\text{ext}} := \inf\{i \geq 1 : K_i = 0\} \in [0, \infty] \tag{4.2.10}$$

for the time until the mutation has been lost from the population, with the usual convention that $\inf \emptyset = \infty$. Let

$$\tau^N := \tau^N_{\text{fix}} \wedge \tau^N_{\text{ext}}$$

be the first day at which either the whole population carries the mutation, or there are no more individuals in the population carrying the mutation. Let

$$C(\gamma) := \frac{\gamma \log \gamma}{\gamma - 1}. \tag{4.2.11}$$

Theorem 4.2.10 (Probability and speed of fixation). *Assume* (4.2.4), *and assume that a mutation affects exactly one individual at day 0, and that no further mutations happen after the first one. Then as* $N \to \infty$,

$$\pi_N \sim \varrho_N \frac{C(\gamma)}{r}. \tag{4.2.12}$$

Moreover, for any $\delta > 0$ *there exists* $N_\delta \in \mathbb{N}$ *such that for all* $N \geq N_\delta$

$$\mathbb{P}(\tau^N > \varrho_N^{-1-3\delta}) \leq (7/8)^{\varrho_N^{-\delta}}. \tag{4.2.13}$$

The proof, which will be given in Section 4.3, relies on a comparison with a supercritical (near-critical) Galton Watson process in the "early phase of the sweep". While the basic idea is classical (dating back to work of Fisher from the 1920's), the scaling (4.2.4) of the supercriticality and the specific nature of our Cannings dynamics required new arguments and a delicate analysis. For related results on near-critical Galton Watson processes (which in some parts inspired our reasoning) see the recent work of Parsons [56].

4.2.5 Genetic and adaptive evolution

Our ultimate goal is to understand the deceleration of the increase in the relative fitness observed in [74], in particular as compared to the linearly increasing number of successful mutations ("adaptive versus genetic evolution"). In our model the relevant scales for the two processes turn out to be different, since the assumptions are such that many successful mutations are needed in order to have a change of approximately one unit in the relative fitness.

This section is divided into two parts. First, we study the model on a short time scale, which is the relevant one for the arrivals of successful mutations. We prove that under some assumptions on the model parameters the number of successful mutations converges on a suitable time scale to a standard Poisson process. Afterwards, we introduce the process of relative fitness of the population, and we show that this process converges on a longer time scale to a deterministic function.

4.2.6 Genetic and adaptive evolution on a short scale

The assertion of Theorem 4.2.10 can be rephrased as follows: *In a background of wild-type individuals that reproduce at rate r, a beneficial mutation that leads to a reproduction rate $r+\varrho_N$ has a probability of fixation obeying* (4.2.12). Besides recalling condition (4.2.4) on the selection, in the following assumption we require that the mutation rate is small enough to exclude "effective clonal interference" between beneficial mutations.

Assumption (Additive, moderately strong selection-weak mutation)**.**

i) Beneficial mutations add ϱ_N to the reproduction rate of the individual that suffers the mutation.

ii) In each generation, with probability μ_N there occurs a beneficial mutation. The mutation affects only one (uniformly chosen) individual, and every offspring of this individual also carries the mutation.

iii) There exists $0 < b < 1/2$, and $a > 3b$, such that $\mu_N \sim N^{-a}$ and $\varrho_N \sim N^{-b}$ as $N \to \infty$.

We use the term *moderately strong selection* in order to indicate that the strength of selection in our model is between what is generally called *strong selection*, where $\varrho_N = O(1)$, and *weak selection* where $\varrho_N = O(N^{-1})$ as $N \to \infty$. Models with such types of selection were recently considered in the context of density dependent birth-death-mutation processes by Parsons [56, 57]. The term *weak mutation* is used to indicate that the mutation rate is small enough to guarantee the absence of clonal interference as $N \to \infty$, which we will prove in Proposition 4.2.13.

Definition 4.2.11 (Interfering mutations, clonal interference)**.** *Consider a pair of successive mutations. Recall that τ^N denotes the first time after the first mutation at which the individual reproduction rate is constant within the population. Denote by m_N the time of the second mutation. We say that the two* mutations interfere *if $m_N < \tau^N$. We say that* clonal interference *occurs if there exists a pair of interfering mutations. In particular, there is* no clonal interference until day i *if there is no mutation starting until day i that interferes with any other mutation.*

Remark 4.2.12. (i) As we will see in Proposition 4.2.13 below, Assumption A iii) guarantees that the probability of clonal interference of any pair of successive mutations is of order at most $\mu_N \varrho_N^{-1}$. In particular, this ensures that the probability of not observing any event of clonal interference on a time scale of order $\mu_N^{-1}\varrho_N^{-2}$ (which we will see to be relevant for our model) tends to 1 as $N \to \infty$.
(ii) Our assumption A iii) is somewhat stronger than requiring $\mu_N \ll \varrho_N$, which is a standard assumption in adaptive dynamics excluding clonal interference, see e.g. [11]. In view of Theorem 4.2.10 and of our detailed calculations in Section 3 we think that replacing $a > 3b$ by $a > b$ in Assumption A iii) should still lead to the same results. However, there are substantial technical difficulties to consider in this case, since $a > b$ only excludes clonal interference of two successive mutations, but not on the longer time scales that are relevant for our results.
(iii) While there is little doubt that there is clonal interference (of successive beneficial mutations) in the Lenski experiment [48], it is noticeable that, as will be seen in Theorem 4.2.15, in order to qualitatively explain certain features of the experimental results on the relative fitness of the population, it is not mandatory to include clonal interference as a model assumption. Including clonal interference into the model will be one goal of our future research in this topic.

Proposition 4.2.13 (Probability of clonal interference). *In our model, for any $\delta > 0$ there exists $N_\delta \in \mathbb{N}$ such that for all $N \geq N_\delta$,*

$$\mathbb{P}(m_N < \tau^N) \leq \mu_N \varrho_N^{-1-\delta}.$$

In particular, under Assumption A iii), for any $T > 0$,

$$\lim_{N\to\infty} \mathbb{P}(\textit{no clonal interference until day } \lfloor \varrho_N^{-2}\mu_N^{-1}T \rfloor) = 1. \tag{4.2.14}$$

A quantity of interest is the *number of successful mutations* up to a given day. Let H_i denote the number of eventually successful mutations that have started until day i, with $H_0 = 0$. Since mutations arrive independently at rate μ_N, and fixate with probability $\sim \frac{C(\gamma)\varrho_N}{r_0}$ (at least in the absence of clonal interference), we expect that successful mutations arrive at rate $\frac{C(\gamma)\mu_N\varrho_N}{r_0}$. Indeed, Proposition 4.2.13 allows us to make this rigorous.

Theorem 4.2.14 (Process of successful mutations). *Let $H_i, i \in \mathbb{N}$, be the number of successful mutations initiated until day i, with $H_0 = 0$. Let $r_0 > 0$ be the reproduction rate of the population at day 0, and let $(W(t))_{t\geq 0}$ be a standard Poisson process. Under Assumption A, for any $T > 0$, the process $(H_{\lfloor(\varrho_N\mu_N)^{-1}t\rfloor})_{0\leq t\leq T}$ converges in distribution (with respect to the Skorokhod topology on the space of càdlàg paths) to $\left(W(\frac{C(\gamma)}{r_0}t)\right)_{0\leq t\leq T}$.*

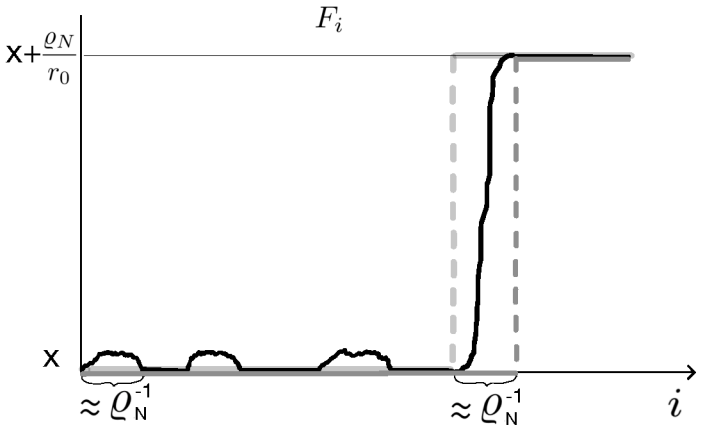

Figure 4.3: The fitness process F_i (solid black line), started at fitness x, depicted until the time of fixation of the next successful mutation, in the absence of clonal interference. The light grey line represents the approximation Φ_i defined in (4.2.15).

4.2.7 Genetic and adaptive evolution on a long time scale

Our next goal is to investigate the process describing the fitness of the evolved population relative to the ancestral population at day 0. Let $R_{i,j}$, for $i \in \mathbb{N}_0$ and $1 \leq j \leq N$, denote the reproduction rate of individual j at the beginning of day i. Assume that at day 0 every individual has reproduction rate r_0, that is, $R_{0,j} = r_0$ for all $j = 1, ..., N$. Recall from (4.1.3) the definition of the *relative fitness* at day i with respect to day 0. We can connect the relative fitness with the number of successful mutations in the following way. Let $\underline{R}_i := \min_{1\leq j\leq N} R_{i,j}$ and $\overline{R}_i := \max_{1\leq j\leq N} R_{i,j}$ denote the minimal and maximal reproduction rate at day i, respectively. Then we have

$$\frac{\underline{R}_i}{r_0} \leq F_i \leq \frac{\overline{R}_i}{r_0}, \quad i \in \mathbb{N}_0.$$

Moreover, on the event that there is no clonal interference up to day i, one has

$$r_0 + \varrho_N(H_i - 1) \leq \underline{R}_i \leq \overline{R}_i \leq r_0 + \varrho_n H_i.$$

Let

$$\Phi_i := 1 + \frac{\varrho_N}{r_0} H_i. \tag{4.2.15}$$

Thus on the event that there is no clonal interference we have

$$\Phi_i - \frac{\varrho_N}{r_0} \leq F_i \leq \Phi_i. \tag{4.2.16}$$

From Theorem 4.2.14 we see that the relevant time scale for the successful mutations is given by $\mu_N^{-1}\varrho_N^{-1}$. Since the selective advantage of a single mutation is of order ϱ_N (cf. (4.2.8)), in view of (4.2.15) it seems plausible that the time scale on which to expect a non-trivial limit of the fitness process is $\varrho_N^{-2}\mu_N^{-1}$. This suggests that the relative fitness has to be considered on a time scale different from that of the number of successful mutations.

Indeed our next theorem shows that the process $F := (F_{\lfloor \mu_N^{-1}\varrho_N^{-2} t \rfloor})_{t\geq 0}$ has a non-trivial scaling limit, which turns out to be a deterministic parabola.

Theorem 4.2.15 (Convergence of the relative fitness process)**.** *Assume $R_{0,j} = r_0$ for $j = 1, ..., N$, and let $(F_i)_{i\in\mathbb{N}_0}$ be the process of relative fitness. Then under Assumption A, the sequence of processes $(F_{\lfloor (\varrho_N^2\mu_N)^{-1}t \rfloor})_{t\geq 0}$ converges in distribution as $N \to \infty$ locally uniformly to the deterministic function*

$$f(t) = \sqrt{1 + \frac{2C(\gamma)t}{r_0^2}},\ t \geq 0\,.$$

The proof of this theorem will be given in Section 4.3.11. It relies on the fact that due to Proposition 4.2.13 the relative fitness process $(F_i)_{i\in\mathbb{N}_0}$ can be approximated by the process $(\Phi_i)_{i\in\mathbb{N}_0}$ defined in (4.2.15).

A similar result can be obtained if a beneficial mutation provides a slightly different advantage, for example due to epistasis. In particular, assume that a mutation that goes to fixation when the relative fitness is x, for any $x \geq 1$ provides an increment

$$\varrho_N^{(x)} = \psi(x)\varrho_N \tag{4.2.17}$$

to the reproduction rate, for some continuous function $\psi : \mathbb{R}^+ \to \mathbb{R}^+$.

Corollary 4.2.16. *Under Assumption A and (4.2.17), let F_i^ψ be the relative fitness of the population at day i with respect to the ancestral population at time 0. Then the process $(F^\psi_{\lfloor (\varrho_N^2\mu_N)^{-1}t \rfloor})_{t\geq 0}$ converges in distribution and locally uniformly as $N \to \infty$ to the deterministic function h which is the solution of the differential equation*

$$\dot{h}(t) = \frac{\psi(h(t))^2 C(\gamma)}{r_0^2 h(t)},\ h(0) = 1,\ t \geq 0.$$

In particular, if $\psi(x) = x^{-q}$ for some $q > -1$, then

$$h(t) = \Big(1 + \frac{2(1+q)C(\gamma)}{r_0^2} t\Big)^{\frac{1}{2(1+q)}}, t \geq 0. \tag{4.2.18}$$

This should be compared to [74], see also the discussion in Section 4.1.5.

4.3 Proof of the main results

In this section, we provide the proofs of the results that we stated in Section 4.2, in particular Theorem 4.2.10, which is technically the most involved and requires several preparatory steps, which are carried out first. After these preparations, the proof of Theorem 4.2.10 will be carried out in Section 4.3.8. The proofs of the other main results will be given in Sections 4.3.9 through 4.3.11.

It turns out that if the number of mutants reaches at least εN, for some $\varepsilon \in (0, 1)$, then the mutation will fixate with probability tending to one as $N \to \infty$. Our strategy for proving Theorem 4.2.10 is thus to divide the time between the occurrence of a mutation and its eventual fixation into three stages. For the case of a successful mutation this is depicted in Figure 4.4.

The first stage starts at the day of the mutation, and ends at the first day $i \in \mathbb{N}$ that the number K_i of mutants has reached a level εN, for some $\varepsilon \in (0, 1/2)$. The second stage starts upon reaching εN, and

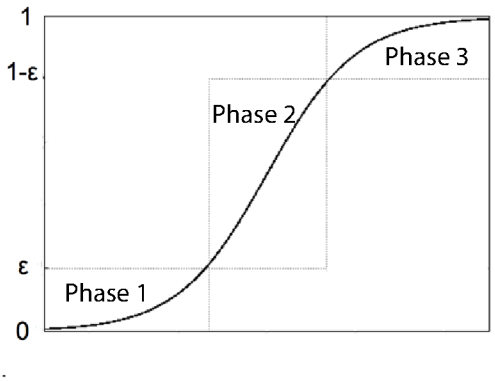

Figure 4.4: A sketch of the frequency of mutants during a selective sweep going to fixation. One distinguishes 3 parts: when the number of mutants is at most εN (phase 1), when the number of wild type individuals is at most εN (phase 3) and the intermediate stage (phase 2).

ends when the process $(K_i)_{i\in\mathbb{N}_0}$ reaches $(1-\varepsilon)N$. The last stage is between $(1-\varepsilon)N$ and N. We will use different methods to analyze the behaviour of the process during these three stages. The first stage is the most difficult to deal with, and we use a coupling to suitable Galton Watson processes to show that the probability that $(K_i)_{i\in\mathbb{N}_0}$ with $K_0 = 1$ ever reaches εN is approximated by (4.2.12). The second stage can be treated by a simple ODE approximation, from which one sees that if $K_i \geq \varepsilon N$ at some time i, then with probability tending to 1 (as $N \to \infty$) the process will eventually reach level $(1-\varepsilon)N$. The third stage will be dealt with in a manner that has some similarities to the first stage, observing that starting from at least $(1-\varepsilon)N$ mutants, there is always a positive probability to reach fixation in the next step. Moreover, our methods of proof will also show that with high probability each of these stages will not last longer than $\varrho_N^{-1-\delta}$, for any $\delta > 0$.
To be more specific, fix $0 < \varepsilon < 1/2$. Assume $K_0 = 1$. Let

$$T_1^N := \inf\{i : K_i \geq \varepsilon N\},$$

and

$$T_2^N := \inf\{i : K_i \geq (1-\varepsilon)N\}.$$

Then we can write τ_{fix}^N as the sum

$$\tau_{\text{fix}}^N = T_1^N + (T_2^N - T_1^N) + (\tau_{\text{fix}}^N - T_2^N). \tag{4.3.1}$$

The important intermediate steps of the proof, dealing with T_1^N, $(T_2^N - T_1^N)$, and $(\tau_{\text{fix}}^N - T_2^N)$, respectively, are given below in Sections 4.3.5, 4.3.6, and 4.3.7, after some preparatory steps in Sections 4.3.1 through 4.3.4. The proof of Theorem 4.2.10 is completed in Section 4.3.8.

Assumption and notation. Throughout all of Section 4.3 we fix $r > 0, \gamma > 1$, and work under the assumption (4.2.4), fixing $b \in (0, 1/2)$ accordingly. We a priori assume $\varepsilon \in (0, 1/2)$, but note that in some places we will impose further conditions. Unless stated otherwise, $\mathbb{P}_k, \mathbb{E}_k$, and $\text{var}_k, k \in \mathbb{N}$, refer to the law, expectation and variance of $(K_i)_{i\in\mathbb{N}_0}$, started at $K_0 = k$, or any random variables defined on the same probability space. We use $c, c', \tilde{c}, ...$ to denote generic constants which are independent of N, with possibly different values at different occurrences.

4.3.1 A simplified sampling and construction of the auxiliary Galton Watson processes

The construction of our model (as explained in Section 4.2.4) was such that K_{i+1} was obtained from K_i by letting two independent Yule populations with initial sizes K_i and $N - K_i$ and respective growth rates $r + \varrho_N$ and r evolve until time σ_{K_i} (defined in (4.2.6)) and then sampling uniformly N individuals from the total of those two populations, which amounts to a mixed hypergeometric sampling of the number of individuals (Proposition 4.2.6). In order to simplify the picture, we would like to use binomial rather than hypergeometric sampling, i.e. sampling individuals independently of each other with equal probability. In this way we will manage to construct two Galton Watson processes $(\underline{K}_i)_{i\in\mathbb{N}_0}$ and $(\overline{K}_i)_{i\in\mathbb{N}_0}$ that will serve as upper and lower bounds for our true process $(K_i)_{i\in\mathbb{N}_0}$ in the first stage of the sweep. We prepare this construction by first giving an alternative description for the sampling of mutants.

Consider the population at the end of a given day (day 0, say). Assume $K_0 = k$, hence by construction at the end of day 0 there are M_{σ_k} mutant individuals for which we want to determine whether or not they will be sampled for the next day. (Recall the definition of M_t from (4.2.5).) Label these mutant individuals with numbers $1, \dots, M_{\sigma_k}$. Define

$$X_j := 1_{\{\text{individual } j \text{ is selected}\}}, \quad j = 1, ..., M_{\sigma_k}.$$

Define a random variable

$$\Gamma := \frac{Y_{\sigma_k}}{N}. \tag{4.3.2}$$

Thus Γ is the ratio between the number of individuals at the end of day 0 and the number of individuals at the beginning of day 1, and by (4.2.6), $\mathbb{E}[\Gamma] = \gamma$. Moreover, $\Gamma \geq 1$, and $\mathbb{P}(\Gamma > 1)$ is exponentially close to 1 as $N \to \infty$. Conditional on Γ, for every $j = 1, ..., M_{\sigma_k}$,

$$\mathbb{P}(X_j = 1) = \frac{1}{\Gamma},$$

but due to our sampling mechanism, the $X_j, j = 1, ..., M_{\sigma_k}$, are not independent. Their joint law conditional on Γ and M_{σ_k} can be described as follows. Let $(U_j)_{j\in\mathbb{N}}$ be i.i.d uniform random variables on $[0, 1]$. Let $\tilde{X}_1 := 1_{\{U_1<1/\Gamma\}}$, and define recursively for $j \geq 2$

$$\tilde{X}_j := 1_{\left\{U_j < \frac{N - \sum_{l=1}^{j-1} \tilde{X}_l}{\Gamma N - (j-1)}\right\}}. \tag{4.3.3}$$

For later convenience we define U_j and $\tilde{X}_j$ for $j \in \mathbb{N}$, even though X_j is defined only for $j = 1, ..., M_{\sigma_k}$.

Lemma 4.3.1. *Conditional on* Γ, $(\tilde{X}_j)_{j=1,...,M_{\sigma_k}}$ *is equal in distribution to* $(X_j)_{j=1,...,M_{\sigma_k}}$.

Proof. Conditional on Γ, we can represent the sampling procedure as follows: Individual 1 has probability $1/\Gamma$ of being selected. For individual 2, the probability of being sampled depends on whether or not individual 1 was selected, in fact

$$\mathbb{P}(X_2 = 1) = \frac{N-1}{\Gamma N - 1}\mathbb{P}(X_1 = 1) + \frac{N}{\Gamma N - 1}\mathbb{P}(X_1 = 0), \tag{4.3.4}$$

or equivalently

$$\mathbb{P}(X_2 = 1|X_1) = \frac{N - X_1}{\Gamma N - 1}. \tag{4.3.5}$$

Proceeding thus recursively, we find that the probability that the jth individual is selected, conditional on knowing $X_1, ..., X_{j-1}$, is

$$\mathbb{P}(X_j = 1|X_1, ..., X_{j-1}) = \frac{N - \sum_{l=1}^{j-1} X_l}{\Gamma N - (j-1)} = \mathbb{P}(\tilde{X}_j = 1|\tilde{X}_1, ..., \tilde{X}_{j-1}). \tag{4.3.6}$$

This completes the proof. □

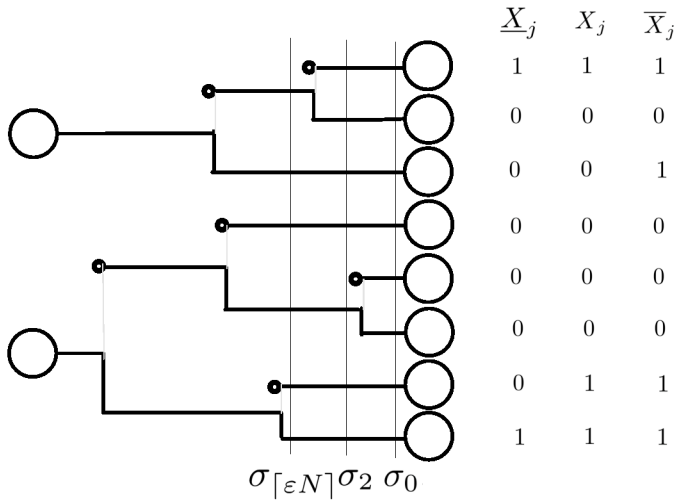

Figure 4.5: Construction of $\underline{X}, \tilde{X}$ and $\overline{X}$ from a Yule forest, starting from $\underline{k} = k = \overline{k} = 2$ mutants. The table shows the values of $\overline{X}$, $\tilde{X}$ and $\underline{X}$ for each of the individuals. E.g. the uppermost of the 8 final individuals is counted in K_1 and $\overline{K}_1$, but not in $\underline{K}_1$, since it has $\overline{X}_2 = X_2 = \underline{X}_2 = 1$, and it was born before time σ_2 but after time $\sigma_{\lceil \varepsilon N \rceil}$.

We can now construct the auxiliary Galton Watson processes. Fix $\alpha > 0$. We are going to specify a joint transition mechanism for $(K_i)_{i\in\mathbb{N}_0}$ and the auxiliary processes $(\underline{K}_i)_{i\in\mathbb{N}_0}$ and $(\overline{K}_i)_{i\in\mathbb{N}_0}$. To this purpose, let $\underline{k}, k, \overline{k}$ be natural numbers. Grow independent Yule trees at rate $r + \varrho_N$ up to time σ_0, and number these trees by $\ell = 1, 2 \ldots$. Number all the individuals in this forest at time σ_0 by $j = 1, 2, \ldots$ and denote the j-th individual by $\mathcal{I}_j$. Let $(U_j)_{j\in\mathbb{N}}$ be a sequence of independent uniformly on $[0, 1]$ distributed random variables, independent of the Yule processes. For $j \in \mathbb{N}$ define

$$\overline{X}_j := 1_{\{U_j < 1/\gamma + N^{-\alpha}\}}$$

and

$$\underline{X}_j := 1_{\{U_j < 1/\gamma - N^{-\alpha}\}}.$$

Also, define Γ as in (4.3.2), and $\tilde{X}_j$ by (4.3.3). We put

$$\begin{aligned} \underline{L} &:= |\{j : \mathcal{I}_j \text{ belongs to the first } \underline{k} \text{ trees and is born before time } \sigma_{\lceil \varepsilon N \rceil}, \text{ and } \underline{X}_j = 1\}|, \\ L &:= |\{j : \mathcal{I}_j \text{ belongs to the first } k \text{ trees and is born before time } \sigma_k, \text{ and } \tilde{X}_j = 1\}| \\ \overline{L} &:= |\{j : \mathcal{I}_j \text{ belongs to the first } \overline{k} \text{ trees and is born before time } \sigma_0, \text{ and } \overline{X}_j = 1\}|. \end{aligned} \tag{4.3.7}$$

Definition 4.3.2. *Let $(\underline{K}_i, K_i, \overline{K}_i)_{i\in\mathbb{N}_0}$ be a Markov chain whose transition probability from $(\underline{k}, k, \overline{k})$ is the joint distribution of $(\underline{L}, L, \overline{L})$ given by (4.3.7).*

By construction, the coordinate processes $(\underline{K}_i)_{i\in\mathbb{N}_0}, (K_i)_{i\in\mathbb{N}_0}$ and $(\overline{K}_i)_{i\in\mathbb{N}_0}$ are also Markov chains. We note in particular that because of Lemma 4.3.1 the dynamics of $(K_i)_{i\in\mathbb{N}_0}$ is the same as that described in Proposition 4.2.6.
Let

$$J := \inf\Big\{j : \frac{N - \sum_{l=1}^{j-1} \tilde{X}_l}{\Gamma N - (j-1)} \in \mathbb{R} \setminus \big[\frac{1}{\gamma} - N^{-\alpha}, \frac{1}{\gamma} + N^{-\alpha}\big]\Big\}. \tag{4.3.8}$$

By construction it is clear that for every $j \le J$

$$\sum_{l=1}^{j} \overline{X}_l \ge \sum_{l=1}^{j} \tilde{X}_l \ge \sum_{l=1}^{j} \underline{X}_l. \tag{4.3.9}$$

Thus if $\underline{K}_0 \le K_0 \le \overline{K}_0$, on the event $\{J \ge M_{\sigma_k}\}$ we have

$$\underline{L} \le L \le \overline{L}.$$

We will show in the next section that if $\alpha \in (b, 1/2)$, then $\mathbb{P}(J > M_{\sigma_k})$ is exponentially close to one for any $k \leq \varepsilon N$. From this we will deduce that with high probability, for $\underline{K}_0 = K_0 = \overline{K_0} = 1$, we have

$$\underline{K}_i \leq K_i \leq \overline{K}_i \quad \forall i \leq T_1^N. \tag{4.3.10}$$

Note that by definition

$$\underline{K}_i \leq \overline{K}_i, \quad \forall i \in \mathbb{N}_0 \tag{4.3.11}$$

always holds. The following characterization of $(\underline{K}_i)_{i\in\mathbb{N}_0}$ and $(\overline{K}_i)_{i\in\mathbb{N}_0}$ is immediate from the construction:

Proposition 4.3.3. *Let $\alpha > 0$ as before, and $(\underline{K}_i, K_i, \overline{K}_i)_{i\in\mathbb{N}_0}$ as in Definition 4.3.2. Then $(\overline{K}_i)_{i\in\mathbb{N}_0}$ is a Galton Watson process whose offspring distribution is mixed binomial with parameters $\overline{M}$ and $\frac{1}{\gamma} + N^{-\alpha}$, where $\overline{M}$ is geometric with parameter $e^{-(r+\varrho_N)\sigma_0}$. Similarly, $(\underline{K}_i)_{i\in\mathbb{N}_0}$ is a Galton Watson process whose offspring distribution is mixed binomial with parameters $\underline{M}$ and $\frac{1}{\gamma} - N^{-\alpha}$, where $\underline{M}$ is geometric with parameter $e^{-(r+\varrho_N)\sigma_{\lceil \varepsilon N\rceil}}$.*

4.3.2 A Galton Watson approximation

A crucial role in our analysis of stage 1 of the sweep will be played by equation (4.3.10), which we are now going to prove. Let b be such that (4.2.4) holds, and assume $K_0 = k$ for some $k \leq \varepsilon N$. We will show that if $\alpha > b$, then with sufficiently large probability $J > N$, and $M_{\sigma_k} < N$. The first part will require some work. To start with, we will work with a slight modification of J. Let

$$\tilde{J} := \inf\Big\{j : \frac{N - \sum_{l=1}^{j-1} \tilde{X}_l}{\Gamma N - (j-1)} \in \mathbb{R} \setminus \big[\frac{1}{\Gamma} - \frac{1}{2}N^{-\alpha}, \frac{1}{\Gamma} + \frac{1}{2}N^{-\alpha}\big]\Big\}. \tag{4.3.12}$$

Lemma 4.3.4. *Let $\alpha \in (b, 1/2)$. There exists a constant $\tilde{c}$ independent of N such that for N large enough,*

$$\mathbb{P}\Big(\tilde{J} > N \,\Big|\, |\gamma - \Gamma| \leq \frac{1}{2}N^{-\alpha}\Big) \geq 1 - 2e^{-\tilde{c}N^{1-2\alpha}}. \tag{4.3.13}$$

Proof. Let $A_\Gamma := \{|\gamma - \Gamma| \leq \frac{1}{2}N^{-\alpha}\}$. By the construction and the definition of $\tilde{X}_j$, equation (4.3.13) is equivalent to

$$\mathbb{P}\Big(\frac{N - \sum_{l=1}^{j-1} \tilde{X}_l}{\Gamma N - (j-1)} \in \big[\frac{1}{\Gamma} - \frac{1}{2}N^{-\alpha}, \frac{1}{\Gamma} + \frac{1}{2}N^{-\alpha}\big] \forall j \in \{1, ..., N\} \,\Big|\, A_\Gamma\Big) \geq 1 - 2e^{-\tilde{c}N^{1-2\alpha}}. \tag{4.3.14}$$

Now rearranging the terms one gets that for $0 \leq j \leq N-1$

$$\frac{1}{\Gamma} - \frac{1}{2}N^{-\alpha} \leq \frac{N - \sum_{l=1}^{j} \tilde{X}_l}{\Gamma N - j} \leq \frac{1}{\Gamma} + \frac{1}{2}N^{-\alpha} \tag{4.3.15}$$

is equivalent to

$$-\frac{1}{2}N^{1/2-\alpha}\big(\Gamma - \frac{j}{N}\big) \leq \frac{1}{\sqrt{N}}\sum_{l=1}^{j}\big(\tilde{X}_l - \frac{1}{\Gamma}\big) \leq \frac{1}{2}N^{1/2-\alpha}\big(\Gamma - \frac{j}{N}\big). \tag{4.3.16}$$

So our aim will be to show that with sufficiently large probability on the event A_Γ

$$\sup_{j\in\{0,1,2,...,N-1\}}\{\frac{1}{\sqrt{N}}\sum_{l=1}^{j}(\tilde{X}_l - \frac{1}{\Gamma})\} \leq \frac{1}{2}N^{1/2-\alpha}(\Gamma - 1)$$

and

$$\inf_{j\in\{0,1,2,...,N-1\}}\{\frac{1}{\sqrt{N}}\sum_{l=1}^{j}(\tilde{X}_l - \frac{1}{\Gamma})\} \geq -\frac{1}{2}N^{1/2-\alpha}(\Gamma - 1).$$

Due to our assumptions, we can consider $(\overline{X}_j)_{j=0,...,N-1}$ resp. $(\underline{X}_j)_{j=0,...,N-1}$ instead of $(\tilde{X}_j)_{j=0,...,N-1}$. Indeed, since $\gamma, \Gamma \geq 1$ we have on the event A_Γ

$$|\frac{1}{\gamma} - \frac{1}{\Gamma}| \leq |\gamma - \Gamma| \leq \frac{1}{2}N^{-\alpha}. \tag{4.3.17}$$

Then

$$[\frac{1}{\Gamma}-\frac{1}{2}N^{-\alpha},\frac{1}{\Gamma}+\frac{1}{2}N^{-\alpha}]\subseteq[\frac{1}{\gamma}-N^{-\alpha},\frac{1}{\gamma}+N^{-\alpha}],$$

which implies that on the event A_Γ, (4.3.9) is valid for every $i\leq\tilde{J}$. We recall the independence between $A_\Gamma,\bar{X},\underline{X}$. Thus we are done if we show

$$\mathbb{P}\Big(\sup_{j\in\{0,1,2,\dots,N-1\}}\{\frac{1}{\sqrt{N}}\sum_{l=1}^{j}(\overline{X}_l-\frac{1}{\Gamma})\}\leq\frac{1}{2}N^{1/2-\alpha}(\Gamma-1)\,\Big|\,A_\Gamma\Big)\geq 1-e^{-\tilde{c}N^{1-2\alpha}}\tag{4.3.18}$$

and

$$\mathbb{P}\Big(\inf_{j\in\{0,1,2,\dots,N-1\}}\{\frac{1}{\sqrt{N}}\sum_{l=1}^{j}(\underline{X}_l-\frac{1}{\Gamma})\}\geq\frac{1}{2}N^{1/2-\alpha}(-\Gamma+1)\,\Big|\,A_\Gamma\Big)\geq 1-e^{-\tilde{c}N^{1-2\alpha}}.\tag{4.3.19}$$

This is an application of large deviations for maxima of sums of independent random variables, see for example [1]. Observing that $\mathbb{E}[\overline{X}_j]=\frac{1}{\gamma}+N^{-\alpha}$ and $\mathrm{var}(\overline{X}_j)=\gamma^{-1}(1-\gamma^{-1}+O(N^{-\alpha}))$, we obtain by a direct application of Theorem 1 of [1] that for any $A>0$ there exists $\tilde{c}_1=\tilde{c}_1(A,\gamma)\in(0,\infty)$ such that

$$\mathbb{P}\Big(\sup_{j\in\{0,1,2,\dots,N-1\}}\{\frac{1}{\sqrt{N}}\sum_{l=1}^{j}(\overline{X}_l-\frac{1}{\gamma}-N^{-\alpha})\}>AN^{1/2-\alpha}\Big)\leq e^{-\tilde{c}_1N^{1-2\alpha}}.\tag{4.3.20}$$

Then (4.3.18) follows with (4.3.17). Similarly we obtain (4.3.19). □

Corollary 4.3.5. *Let $\alpha\in(b,1/2)$. There exists a constant c independent of N such that for N large enough*

$$\mathbb{P}\big(J>N\big)\geq 1-e^{-cN^{1-2\alpha}}.\tag{4.3.21}$$

Proof. Recall from the proof of the previous lemma that if $|\frac{1}{\gamma}-\frac{1}{\Gamma}|\leq\frac{1}{2}N^{-\alpha}$ then

$$[\frac{1}{\Gamma}-\frac{1}{2}N^{-\alpha},\frac{1}{\Gamma}+\frac{1}{2}N^{-\alpha}]\subseteq[\frac{1}{\gamma}-N^{-\alpha},\frac{1}{\gamma}+N^{-\alpha}],$$

which implies that in this case $\tilde{J}<J$. We already observed that $|\frac{1}{\gamma}-\frac{1}{\Gamma}|\leq|\gamma-\Gamma|$, so it remains to show that $|\gamma-\Gamma|<\frac{1}{2}N^{-\alpha}$ with large probability. Indeed, for $l=1,2,\dots,N$ and N large enough

$$\begin{aligned}\mathbb{P}\big(|\Gamma-\gamma|\leq\frac{1}{2}N^{-\alpha}\big)&=\mathbb{P}\big(|\frac{Y_{\sigma_l}}{N}-\gamma|\leq\frac{1}{2}N^{-\alpha}\big)\\&=\mathbb{P}\big(|\frac{Y_{\sigma_l}-N\gamma}{\sqrt{N}}|\leq\frac{1}{2}N^{1/2-\alpha}\big)\\&\geq 1-e^{-c'N^{1-2\alpha}}\end{aligned}$$

for some constant in c' independent of N, where the last inequality follows from a generalisation of Cramér's theorem, see Theorem 2 of [58] (note that σ_l is a sum of independent but not identically distributed random variables). Let c be a constant independent of N such that $c>\max(c',\tilde{c})$, where $\tilde{c}$ is the constant from Lemma 4.3.4. For N large enough

$$\begin{aligned}\mathbb{P}\Big(J>N\Big)&\geq\mathbb{P}\Big(J>N,|\frac{1}{\gamma}-\frac{1}{\Gamma}|\leq\frac{1}{2}N^{-\alpha}\Big)\\&\geq\mathbb{P}\Big(\tilde{J}>N,|\frac{1}{\gamma}-\frac{1}{\Gamma}|\leq\frac{1}{2}N^{-\alpha}\Big)\\&\geq\mathbb{P}\Big(\tilde{J}>N\Big)-\mathbb{P}\Big(|\frac{1}{\gamma}-\frac{1}{\Gamma}|>\frac{1}{2}N^{-\alpha}\Big)\\&\geq 1-2e^{-\tilde{c}N^{1-2\alpha}}-e^{-c'N^{1-2\alpha}}\\&\geq 1-e^{-cN^{1-2\alpha}}.\end{aligned}$$

□

Lemma 4.3.6. *Let $\alpha \in (b, 1/2)$, and $0 < \varepsilon < 1/\gamma$. Assume $\underline{K}_0 \le K_0 \le \overline{K}_0$ and $K_0 = k$, for some $k \le \varepsilon N$. There exists $c > 0$ independent of N such that for all N large enough,*

$$\mathbb{P}(M_{\sigma_k} < N) \ge 1 - e^{-cN}.$$

Proof. Let G_j be the number of offspring of the mutant number $j \le k \le \varepsilon N$ at the end of the day, namely at time σ_k. By construction they are i.i.d. with finite second moment. Let $(G'_j)_{j\in\mathbb{N}}$ be i.i.d random variables equal in distribution to G_1. Note that $\mathbb{E}[G_1] \le e^{(r+\varrho_N)\sigma_0} = \gamma(1 + o(1))$. Since $\varepsilon < 1/\gamma$ we can choose N large enough such that $\mathbb{E}[G_1] \le 1/\varepsilon$. Then

$$\mathbb{P}\big(M_{\sigma_k} < N\big) = \mathbb{P}\big(\sum_{j=1}^{k} G_j < N\big) \quad \ge \quad \mathbb{P}\big(\sum_{j=1}^{\varepsilon N} G'_j < N\big) \ge 1 - e^{-cN}$$

for a suitable $c > 0$. The last inequality follows from Cramer's Theorem, since $\varepsilon\mathbb{E}[G_1] < 1$. □

Recall that $T_1^N = \inf\{i \ge 1 : K_i \ge \varepsilon N\}$.

Proposition 4.3.7. *Let $\alpha \in (b, 1/2)$ and $0 < \varepsilon < 1/\gamma$. Assume $\underline{K}_0 \le K_0 \le \overline{K}_0$ and $K_0 = k \le \varepsilon N$. Then there exists c independent of N such that for N large enough*

$$\mathbb{P}\big(\overline{K}_{\min(i,T_1^N)} \ge K_{\min(i,T_1^N)} \ge \underline{K}_{\min(i,T_1^N)}, \forall i \le g\big) \ge (1 - 2e^{-cN^{1-2\alpha}})^g \quad \text{for all } g \in \mathbb{N}_0. \tag{4.3.22}$$

Proof. Corollary 4.3.5 implies that $\mathbb{P}(\underline{K}_1 \le K_1 \le \overline{K}_1 \mid M_{\sigma_k} < N) \ge 1 - e^{-cN^{1-2\alpha}}$. Thus by Lemma 4.3.6 we have

$$\mathbb{P}(\underline{K}_1 \le K_1 \le \overline{K}_1) \ge 1 - 2e^{-cN^{1-2\alpha}}, \tag{4.3.23}$$

which implies

$$\mathbb{P}\big(\overline{K}_{\min(g,T_1^N)} \ge K_{\min(g,T_1^N)} \ge \underline{K}_{\min(g,T_1^N)} \,|\, \overline{K}_{\min(i,T_1^N)} \ge K_{\min(i,T_1^N)} \ge \underline{K}_{\min(i,T_1^N)}, \forall i \le g-1\big) \ge 1 - 2e^{-cN^{1-2\alpha}}. \tag{4.3.24}$$

From (4.3.24) the result follows easily by induction: Assume that (4.3.22) is true for $g - 1$. Then

$$\begin{aligned}
&\mathbb{P}\big(\overline{K}_{\min(i,T_1^N)} \ge K_{\min(i,T_1^N)} \ge \underline{K}_{\min(i,T_1^N)}, \forall i \le g\big) \\
&= \mathbb{P}\big(\overline{K}_{\min(g,T_1^N)} \ge K_{\min(g,T_1^N)} \ge \underline{K}_{\min(g,T_1^N)} | \overline{K}_{\min(i,T_1^N)} \ge K_{\min(i,T_1^N)} \ge \underline{K}_{\min(i,T_1^N)}, \forall i \le g-1\big) \\
&\quad \times \mathbb{P}\big(\overline{K}_{\min(i,T_1^N)} \ge K_{\min(i,T_1^N)} \ge \underline{K}_{\min(i,T_1^N)}, \forall i \le g-1\big) \\
&\ge (1 - 2e^{-cN^{1-2\alpha}})(1 - 2e^{-cN^{1-2\alpha}})^{g-1}.
\end{aligned}$$

□

4.3.3 Asymptotics of the stopping rule

In order to put the Galton Watson bounds to use, we need some control on σ_k.

Lemma 4.3.8. *Under the assumptions of this section, for any $k = 1, 2, \dots, N$,*

$$\sigma_k = \frac{\log\gamma}{r + k\varrho_N/N} + \frac{k}{N}O(\varrho_N^2) + \frac{k^2}{N^2}O(\varrho_N^2). \tag{4.3.25}$$

where $|O(\varrho_N^2)|/\varrho_N^2$ is bounded uniformly in N and k.

Proof. Note that $\frac{\log\gamma}{r+\varrho_N} = \sigma_N \le \sigma_k \le \sigma_0 = \frac{\log\gamma}{r}$ for all $k = 0, ..., N$. Hence $\lim_{N\to\infty}\sigma_k = \frac{\log\gamma}{r}$ for all k. We assume that N is large enough such that $\frac{\log\gamma}{2r} \le \sigma_k \le \frac{\log\gamma}{r}$.
By (4.2.5) and (4.2.6) we have

$$\gamma N = \mathbb{E}[M_{\sigma_k}^{(k)}] + \mathbb{E}[Z_{\sigma_k}^{(N-k)}] = ke^{(r+\varrho_N)\sigma_k} + (N-k)e^{r\sigma_k}. \tag{4.3.26}$$

Hence σ_k satisfies the equation

$$\gamma N = e^{r\sigma_k}\big(ke^{\varrho_N\sigma_k} + N - k\big). \tag{4.3.27}$$

Dividing by N, taking logarithms on both sides, and using Taylor expansion first on the exponential and then on the logarithm leads to

$$\begin{aligned}\log\gamma =& r\sigma_k + \log\big(1 + \frac{k}{N}(e^{\varrho_N\sigma_k} - 1)\big)\\ =& r\sigma_k + \log\big(1 + \frac{k}{N}\varrho_N\sigma_k + \frac{k}{N}O(\varrho_N^2)\big)\\ =& r\sigma_k + \frac{k}{N}\varrho_N\sigma_k + \frac{k}{N}O(\varrho_N^2) + \frac{k^2}{N^2}O(\varrho_N^2).\end{aligned} \tag{4.3.28}$$

Here we use the fact that $\frac{\log\gamma}{2r} \le \sigma_k \le \frac{\log\gamma}{r}$ for all k if N is sufficiently large. Rewriting, we get the desired expression of σ_k. □

We will use this mostly in the following form, which is an immediate application of Lemma 4.3.8.

Corollary 4.3.9. *For any $k = 1, 2, \dots, N$, as $N \to \infty$*

$$e^{(r+\varrho_N)\sigma_k} = \gamma\big(1 + (1 - \frac{k}{N})\frac{\varrho_N}{r}\log\gamma + O(\varrho_N^2)\big)$$

where $|O(\varrho_N^2)|/\varrho_N^2$ is bounded uniformly in N and k.

4.3.4 Asymptotics of the approximating Galton Watson processes and Proof of Prop. 4.2.8

We can now calculate the asymptotic expectation and variance of our auxiliary Galton Watson processes.

Lemma 4.3.10. *Let $\alpha \in (b, 1/2)$. Let $(\underline{K}_i)_{i\in\mathbb{N}_0}$ and $(\overline{K}_i)_{i\in\mathbb{N}_0}$ be as defined in Section 4.3.1 with $\underline{K}_0 = K_0 = \overline{K}_0 = 1$. We have*

$$\mathbb{E}_1[\overline{K}_1] = 1 + \frac{\log\gamma}{r}\varrho_N + o(\varrho_N) \qquad \mathbb{E}_1[\underline{K}_1] = 1 + \frac{\log\gamma}{r}(1-\varepsilon)\varrho_N + o(\varrho_N), \tag{4.3.29}$$

and

$$\mathrm{var}_1[\overline{K}_1] = \frac{2(\gamma-1)}{\gamma}(1 + O(\varrho_N)) \qquad \mathrm{var}_1[\underline{K}_1] = \frac{2(\gamma-1)}{\gamma}(1 + O(\varrho_N)). \tag{4.3.30}$$

Proof. Recall $\underline{M}, \overline{M}$ from Proposition 4.3.3. By construction, and from Corollary 4.3.9

$$\begin{aligned}\mathbb{E}_1[\underline{K}_1] &= (1/\gamma - N^{-\alpha})\mathbb{E}[\underline{M}]\\ &= (1/\gamma - N^{-\alpha})e^{(r+\varrho_N)\sigma_{\lceil\varepsilon N\rceil}}\\ &= 1 + \frac{\log\gamma}{r}(1-\varepsilon)\varrho_N - \gamma N^{-\alpha} + o(\varrho_N)\\ &= 1 + \frac{\log\gamma}{r}(1-\varepsilon)\varrho_N + o(\varrho_N)\end{aligned}$$

where the last equality follows from the fact that our assumptions imply that $N^{-\alpha} = o(\varrho_N)$. In the same way we obtain

$$\mathbb{E}_1[\overline{K}_1] = 1 + \frac{\log\gamma}{r}\varrho_N + o(\varrho_N).$$

It remains to calculate the variance

$$\begin{aligned}\mathrm{var}_1[\underline{K}_1] &= \mathbb{E}_1[\mathrm{var}_1[\underline{K}_1|\underline{M}]] + \mathrm{var}_1[\mathbb{E}_1[\underline{K}_1|\underline{M}]]\\ &= \mathbb{E}_1\big([\underline{M}\big(\frac{1}{\gamma} - N^{-\alpha}\big)\big(1 - \frac{1}{\gamma} + N^{-\alpha}\big)\big] + \mathrm{var}_1\big[\underline{M}\big(\frac{1}{\gamma} - N^{-\alpha}\big)\big]\\ &= \big(\frac{1}{\gamma} - N^{-\alpha}\big)\big(1 - \frac{1}{\gamma} + N^{-\alpha}\big)e^{(r+\varrho_N)\sigma_{\lceil\varepsilon N\rceil}} + \big(\frac{1}{\gamma} - N^{-\alpha}\big)^2\big(e^{2(r+\varrho_N)\sigma_{\lceil\varepsilon N\rceil}} - e^{(r+\varrho_N)\sigma_{\lceil\varepsilon N\rceil}}\big).\end{aligned}$$

Plugging in Corollary 4.3.9, simplifying and taking into account that $N^{-\alpha} = o(\varrho_N)$ for $\alpha > b$ leads to

$$\mathrm{var}_1[\underline{K}_1] = \frac{2(\gamma-1)}{\gamma}\big(1 + (1-\varepsilon)\varrho_N \frac{\log\gamma}{r} + o(\varrho_N)\big) = \frac{2(\gamma-1)}{\gamma} + O(\varrho_N).$$

The same steps lead to $\mathrm{var}_1[\overline{K}_1] = 2(\gamma-1)/\gamma + O(\varrho_N)$. □

Remark 4.3.11. (i) This result together with Lemma 4.3.6 proves Proposition 4.2.8. (ii) Applying Lemma A.3.6 from the Appendix shows

$$\mathbb{P}((\overline{K}_i) \text{ survives}) \sim \frac{C(\gamma)}{r}\varrho_N$$

and

$$\mathbb{P}((\underline{K}_i) \text{ survives}) \sim \frac{(1-\varepsilon)C(\gamma)}{r}\varrho_N.$$

Corollary 4.3.12. *Under the assumptions of Lemma 4.3.10, for $k \le \varepsilon N$, as $N \to \infty$,*

$$\mathbb{P}_k((\overline{K}_i) \textit{ survives } | \, (\underline{K}_i) \textit{ survives }) = \mathbb{P}_k((\underline{K}_i) \textit{ dies out} | \, (\overline{K}_i) \textit{ dies out}) = 1. \tag{4.3.31}$$

Further,

$$\mathbb{P}_k((\underline{K}_i) \textit{ dies out } | \, (\overline{K}_i) \textit{ survives }) \le \varepsilon(1+o(1)), \tag{4.3.32}$$

and

$$\mathbb{P}_k((\overline{K}_i) \textit{ survives } | \, (\underline{K}_i) \textit{ dies out }) \le \varepsilon(1+o(1)). \tag{4.3.33}$$

Proof. The first equation follows immediately from (4.3.11). We prove (4.3.32), (4.3.33) follows similarly. Let $c(\gamma, r) := \frac{\gamma\log\gamma}{(\gamma-1)r}$. Note that

$$\begin{aligned}\mathbb{P}_k((\underline{K}_i) \text{ dies out} \, | \, (\overline{K}_i) \text{ survives }) =& \frac{\mathbb{P}_k((\underline{K}_i) \text{ dies out}) - \mathbb{P}_k((\overline{K}_i) \text{ dies out})}{\mathbb{P}_k((\overline{K}_i) \text{ survives})} \\ \sim& \frac{(1-c(\gamma,r)(1-\varepsilon)\varrho_N)^k - (1-c(\gamma,r)\varrho_N)^k}{1-(1-c(\gamma,r)\varrho_N)^k}.\end{aligned} \tag{4.3.34}$$

Let $g(k)$ be the r.h.s of (4.3.34). We will show below that g is decreasing in k if N is large, from which the statement follows, observing

$$g(k) \le g(1) \le \varepsilon(1+o(1)).$$

To prove the monotonicity of $g(k)$, let $a = c(\gamma, r)\varrho_N$. Let N large enough such that $0 < a < 1$. Assume that $k \ge 1$ and $k \in \mathbb{R}^+$. Then we can differentiate $\log(1-g(k))$ in k which yields

$$\frac{\partial}{\partial k}\log(1-g(k)) = \frac{(1-a)^k\log(1-a)}{1-(1-a)^k} - \frac{(1-a+a\varepsilon)^k\log(1-a+a\varepsilon)}{1-(1-a+a\varepsilon)^k}. \tag{4.3.35}$$

The function $\frac{x^k log(x)}{1-x^k}$ is a decreasing function in x, for $0 < x < 1$, as can be seen by differentiation. Apply this to the r.h.s of (4.3.35), we obtain $\frac{d\log(1-g(k))}{dk} \ge 0$ for all $k \ge 1$. This implies $\frac{dg(k)}{dk} \le 0$. So $g(k)$ is decreasing in k. □

4.3.5 First stage of the sweep

With these preparations we can now address the first stage of the sweep, cf. Figure 4.4. We are going to calculate the probability that the number of mutants reaches εN for some $\varepsilon > 0$, and determine the time it takes to reach εN. We achieve this by using the supercritical Galton Watson processes provided by Lemma 4.3.7. Recall $T_1^N = \inf\{i \ge 0 : K_i \ge \varepsilon N\}$.

Lemma 4.3.13. *Let $0 < \varepsilon < 1/\gamma$. Then we have as $N \to \infty$*

$$\frac{\varrho_N\log\gamma}{r}\frac{\gamma}{\gamma-1}(1-\varepsilon)(1+o(1)) \le \mathbb{P}_1(\exists i : K_i \ge \varepsilon N) \le \frac{\varrho_N\log\gamma}{r}\frac{\gamma}{\gamma-1}(1+o(1)), \tag{4.3.36}$$

and for any $\delta > 0$

$$\limsup_{N\to\infty} \mathbb{P}_1(T_1^N > \varrho_N^{-1-\delta} \, | \, T_1^N < \infty) \le \frac{\varepsilon}{1-\varepsilon}.$$

Proof. Let $\alpha \in (b, 1/2)$ and let $(\underline{K}_i)_{i\in\mathbb{N}_0}$ and $(\overline{K}_i)_{i\in\mathbb{N}_0}$ be defined as in Section 4.3.1, with $\underline{K}_0 = K_0 = \overline{K}_0 = 1$. We write (K_i) *reaches* εN for the event that there exists $i > 0$ such that $K_i \geq \varepsilon N$, and analogously for $(\underline{K}_i), (\overline{K}_i)$. By Remark 4.3.11, Lemma A.3.6, and Lemma A.3.7,

$$\mathbb{P}_1((\overline{K}_i) \text{ reaches } \varepsilon N) \sim \mathbb{P}_1((\overline{K}_i) \text{ survives}) \sim \frac{\varrho_N \log\gamma}{r} \frac{\gamma}{\gamma - 1} \tag{4.3.37}$$

and

$$\mathbb{P}_1((\underline{K}_i) \text{ reaches } \varepsilon N) \sim \mathbb{P}_1((\underline{K}_i) \text{ survives}) \sim \frac{\varrho_N \log\gamma}{r} \frac{\gamma}{\gamma - 1}(1 - \varepsilon). \tag{4.3.38}$$

Let

$$A := A(\gamma, \alpha, \varepsilon, \delta, N) := \{\underline{K}_i \leq K_i \leq \overline{K}_i \ \forall i \leq \min(T_1^N, \varrho_N^{-1-\delta})\}.$$

Setting $g := \varrho_N^{-1-\delta}$ in Proposition 4.3.7 and applying the Bernoulli inequality we have

$$\mathbb{P}_1(A^c) \leq 1 - (1 - 2e^{-cN^{1-2\alpha}})^{\varrho_N^{-1-\delta}} \leq \varrho_N^{-1-\delta} 2e^{-cN^{1-2\alpha}}, \tag{4.3.39}$$

implying $\mathbb{P}_1(A) \to 1$ exponentially fast as $N \to \infty$. Let $\underline{T}_1^N := \inf\{i > 0 : \underline{K}_i \geq \varepsilon N\}$. Then

$$\begin{aligned}
\mathbb{P}_1((K_i) \text{ reaches } \varepsilon N) &\geq \mathbb{P}_1((K_i) \text{ reaches } \varepsilon N, (\underline{K}_i) \text{ reaches } \varepsilon N, A, \underline{T}_1^N \leq \varrho_N^{-1-\delta}) \\
&= \mathbb{P}_1((\underline{K}_i) \text{ reaches } \varepsilon N, A, \underline{T}_1^N \leq \varrho_N^{-1-\delta}) \\
&\geq \mathbb{P}_1((\underline{K}_i) \text{ reaches } \varepsilon N, \underline{T}_1^N \leq \varrho_N^{-1-\delta}) - \mathbb{P}(A^c) \\
&\sim \mathbb{P}_1((\underline{K}_i) \text{ reaches } \varepsilon N)
\end{aligned} \tag{4.3.40}$$

using (4.3.39) and Lemma B.3 in the last inequality. Together with (4.3.38) this proves the lower bound in (4.3.36). For the upper bound, let $\overline{T}_0^N := \inf\{i : \overline{K}_i = 0\}$. Note that

$$\mathbb{P}_1((K_i) \text{ reaches } \varepsilon N) = \mathbb{P}((K_{i\wedge T_1^N}) \text{ reaches } \varepsilon N)$$

and

$$\mathbb{P}_1((K_{i\wedge T_1^N}) \text{ reaches } \varepsilon N) = 1 - \mathbb{P}((K_{i\wedge T_1^N}) \text{ dies out}).$$

Thus we have

$$\begin{aligned}
1 - \mathbb{P}_1((K_i) \text{ reaches } \varepsilon N) &\geq \mathbb{P}_1((K_{i\wedge T_1^N}) \text{ dies out}) \\
&\geq \mathbb{P}_1((K_{i\wedge T_1^N}) \text{ dies out}; (\overline{K}_i) \text{ dies out}; A; \overline{T}_0^N \leq \varrho_N^{-1-\delta}) \\
&= \mathbb{P}_1((\overline{K}_i) \text{ dies out}; A; \overline{T}_0^N \leq \varrho_N^{-1-\delta}) \\
&\sim \mathbb{P}_1((\overline{K}_i) \text{ dies out}) \\
&\sim 1 - \mathbb{P}_1((\overline{K}_i) \text{ reaches } \varepsilon N),
\end{aligned} \tag{4.3.41}$$

where we have used (A.3.3) from the Appendix and Lemma A.3.7. This implies the upper bound. We are thus left with proving the last statement of the Lemma. Fix $\delta > 0$. We have

$$\begin{aligned}
\mathbb{P}_1(T_1^N > \varrho_N^{-1-\delta} \,|\, (K_i) \text{ reaches } \varepsilon N) = &\frac{\mathbb{P}_1(T_1^N > \varrho_N^{-1-\delta}, (K_i) \text{ reaches } \varepsilon N, (\underline{K}_i) \text{ survives})}{\mathbb{P}_1((K_i) \text{ reaches } \varepsilon N)} \\
&+ \frac{\mathbb{P}_1(T_1^N > \varrho_N^{-1-\delta}, (K_i) \text{ reaches } \varepsilon N, (\underline{K}_i) \text{ dies out})}{\mathbb{P}_1((K_i) \text{ reaches } \varepsilon N)}.
\end{aligned} \tag{4.3.42}$$

By (4.3.40) and Lemma A.3.7 we have for large enough N the inequality

$$\mathbb{P}_1((K_i) \text{ reaches } \varepsilon N) \geq \mathbb{P}_1((\underline{K}_i) \text{ survives}),$$

and thus the first term on the right-hand side of (4.3.42) can be bounded from above by

$$\begin{aligned}
\mathbb{P}_1(T_1^N > \varrho_N^{-1-\delta} \,|\, (\underline{K}_i) \text{ survives}) &\leq \mathbb{P}_1(T_1^N > \varrho_N^{-1-\delta}, A \,|\, (\underline{K}_i) \text{ survives}) + \mathbb{P}_1(A^c \,|(\underline{K}_i) \text{ survives}) \\
&\leq \mathbb{P}_1(\underline{T}_1^N > \varrho_N^{-1-\delta} \,|\, (\underline{K}_i) \text{ survives}) + \frac{\mathbb{P}_1(A^c)}{\mathbb{P}((\underline{K}_i) \text{ survives})}.
\end{aligned} \tag{4.3.43}$$

The first term on the right-hand side converges to 0 due to Lemma A.3.8. By Lemma A.3.6 we have $\mathbb{P}_1((\underline{K}_i) \text{ survives}) \sim c\varrho_N$, therefore by (4.3.39) the second term on the right-hand side converges to 0 as well. Thus we have shown that the first summand in (4.3.42) converges to 0. To deal with the second term, we observe

$$\begin{aligned}
&\mathbb{P}_1(T_1^N > \varrho_N^{-1-\delta}, (K_i) \text{ reaches } \varepsilon N, (\underline{K}_i) \text{ dies out}) \\
&\qquad \leq \mathbb{P}_1((K_i) \text{ reaches } \varepsilon N, (\underline{K}_i) \text{ dies out}) \\
&\qquad = \mathbb{P}_1((K_i) \text{ reaches } \varepsilon N, (\underline{K}_i) \text{ dies out}, (\overline{K}_i) \text{ dies out}) \\
&\qquad\quad + \mathbb{P}_1((K_i) \text{ reaches } \varepsilon N, (\underline{K}_i) \text{ dies out}, (\overline{K}_i) \text{ survives}) \\
&\qquad \leq \mathbb{P}_1((K_i) \text{ reaches } \varepsilon N, (\overline{K}_i) \text{ dies out}) + \mathbb{P}_1((\underline{K}_i) \text{ dies out}, (\overline{K}_i) \text{ survives}) \\
&\qquad \leq \mathbb{P}_1((K_i) \text{ reaches } \varepsilon N, (\overline{K}_i) \text{ dies out}, \overline{K}_{\lfloor \varrho_N^{-1} \rfloor} > 0) \\
&\qquad\quad + \mathbb{P}_1((K_i) \text{ reaches } \varepsilon N, (\overline{K}_i) \text{ dies out}, \overline{K}_{\lfloor \varrho_N^{-1} \rfloor} = 0) \\
&\qquad\quad + \mathbb{P}_1((\underline{K}_i) \text{ dies out}, (\overline{K}_i) \text{ survives}).
\end{aligned} \tag{4.3.44}$$

We have

$$\mathbb{P}_1((K_i) \text{ reaches } \varepsilon N, (\overline{K}_i) \text{ dies out}, \overline{K}_{\lfloor \varrho_N^{-1} \rfloor} > 0) \leq \mathbb{P}((\overline{K}_i) \text{ dies out}, \overline{K}_{\lfloor \varrho_N^{-1} \rfloor} > 0)$$

which goes to 0 exponentially fast due to (A.3.3) in the Appendix, and using Lemma A.3.7 we get

$$\mathbb{P}_1((K_i) \text{ reaches } \varepsilon N, (\overline{K}_i) \text{ dies out}, \overline{K}_{\lfloor \varrho_N^{-1} \rfloor} = 0) \leq \mathbb{P}_1(A^c)$$

which goes to 0 exponentially fast due to (4.3.39). Finally we have

$$\begin{aligned}
\mathbb{P}_1((\underline{K}_i) \text{ dies out}, (\overline{K}_i) \text{ survives}) &= \mathbb{P}_1((\underline{K}_i) \text{ dies out} \,|\, (\overline{K}_i) \text{ survives}) \mathbb{P}_1((\overline{K}_i) \text{ survives}) \\
&\leq \varepsilon(1 + o(1)) \mathbb{P}_1((\overline{K}_i) \text{ survives}) \\
&= \frac{\varepsilon}{1-\varepsilon}(1 + o(1)) \mathbb{P}_1((\underline{K}_i) \text{ survives}),
\end{aligned}$$

see Corollary 4.3.12. Thus the second summand in (4.3.42) is bounded from above by $\frac{\varepsilon}{1-\varepsilon}(1+o(1))$, and the claim follows.

□

Corollary 4.3.14. *Let $T_0^N := \inf\{i : K_i = 0\}$. For $0 < \varepsilon < 1/\gamma \wedge 1/16$ there exists $N_\varepsilon^{(1)}$ such that for any $k \leq \varepsilon N$,*

$$\mathbb{P}_k(T_1^N \wedge T_0^N > \varrho_N^{-1-\delta}) \leq 1/2. \tag{4.3.45}$$

Proof. Fix $k \leq \varepsilon N$. We have

$$\begin{aligned}
\mathbb{P}_k(T_1^N \wedge T_0^N > \varrho_N^{-1-\delta}) \quad = \quad & \mathbb{P}_k(T_1^N > \varrho_N^{-1-\delta} | T_1^N \wedge T_0^N = T_1^N) \mathbb{P}_k(T_1^N \wedge T_0^N = T_1^N) \\
& + \mathbb{P}_k(T_0^N > \varrho_N^{-1-\delta} | T_1^N \wedge T_0^N = T_0^N) \mathbb{P}_k(T_1^N \wedge T_0^N = T_0^N).
\end{aligned}$$

Due to (4.3.32) we can see that all the steps leading to the last statement in Lemma 4.3.13 hold if the processes are started in $k \leq \varepsilon N$ instead of 1. Hence we have that for all $1 \leq k \leq \varepsilon N$

$$\limsup_{N\to\infty} \mathbb{P}_k(T_1^N > \varrho_N^{-1-\delta} | T_1^N < \infty) \leq \frac{\varepsilon}{1-\varepsilon}. \tag{4.3.46}$$

Moreover, if we stop (K_i) with $K_0 = k \leq \varepsilon N$ when the Markov chain is larger than εN, then (K_i) is an absorbing Markov chain with absorbing states 0 and any number larger than εN. That implies $\mathbb{P}_k(T_1^N \wedge T_0^N < \infty) = 1$. Notice that under event $\{T_1^N \wedge T_0^N < \infty\}$, we have $\{T_1^N < \infty\} = \{T_1^N \wedge T_0^N = T_1^N\}$. Altogether we obtain

$$\limsup_{N\to\infty} \mathbb{P}_k(T_1^N > \varrho_N^{-1-\delta} | T_1^N \wedge T_0^N = T_1^N) \leq \frac{\varepsilon}{1-\varepsilon}(1 + o(1)), \tag{4.3.47}$$

which is smaller than $1/4$ for our choice of ε. Therefore (4.3.45) holds for any $k \leq \varepsilon N$ such that $\mathbb{P}_k(T_0^N > \varrho_N^{-1-\delta} | T_1^N \wedge T_0^N = T_0^N) \leq 1/4$. Assume therefore that $\mathbb{P}_k(T_0^N > \varrho_N^{-1-\delta} | T_1^N \wedge T_0^N = T_0^N) > 1/4$.

Due to Proposition 4.3.7 and Lemma A.3.8 we then have that $\mathbb{P}_k(T_1^N \wedge T_0^N = T_0^N) \geq 1/4$ for N large enough. For such k

$$\begin{aligned}\mathbb{P}_k(T_0^N > \varrho_N^{-1-\delta} \,|\, T_1^N \wedge T_0^N = T_0^N) &\leq \mathbb{P}_k(K_{\lfloor \varrho_N^{-1-\delta} \rfloor} > 0, A \,|\, T_1^N \wedge T_0^N = T_0^N) + \mathbb{P}_k(A^c \,|\, T_1^N \wedge T_0^N = T_0^N) \\ &\leq 4\mathbb{P}_k(K_{\lfloor \varrho_N^{-1-\delta} \rfloor} > 0, A, (K_i)_{i\in\mathbb{N}} \text{ dies out}) + 4\mathbb{P}_k(A^c) \\ &\leq 4\mathbb{P}_k(\overline{K}_{\lfloor \varrho_N^{-1-\delta} \rfloor} > 0, (\underline{K}_i)_{i\in\mathbb{N}} \text{ dies out}) + 4\mathbb{P}_k(A^c).\end{aligned}$$

Equation (4.3.33) implies

$$4\mathbb{P}_k(\overline{K}_{\lfloor \varrho_N^{-1-\delta} \rfloor} > 0, (\underline{K}_i)_{i\in\mathbb{N}} \text{ dies out}) \leq 4\mathbb{P}_k(\overline{K}_{\lfloor \varrho_N^{-1-\delta} \rfloor} > 0, (\overline{K}_i)_{i\in\mathbb{N}} \text{ dies out}) + 4\varepsilon(1+o(1)).$$

By (4.3.39), $\mathbb{P}_k(A^c)$ goes to 0 exponentially fast, and $\mathbb{P}_k(\overline{K}_{\varrho_N^{-1-\delta}} > 0 \,|\, (\overline{K}_i)_{i\in\mathbb{N}_0} \text{ dies out})$ goes to 0 by (A.3.3). Thus if $\varepsilon < 1/16$ the right-hand side of the above inequality is bounded above by $1/4$, and we have completed the proof. □

4.3.6 Second stage of the sweep

In the second stage of the sweep we will make an approximation, just like the one we did in Proposition 1.3.5.

Lemma 4.3.15. *For $\varepsilon \in (0, 1/2)$ let $1-\varepsilon' \in (\varepsilon, 1)$. Then we have for any $k \geq \varepsilon N$*

$$\lim_{N\to\infty} \mathbb{P}_k(\exists i : K_i \geq \lfloor (1-\varepsilon')N \rfloor) = 1.$$

Moreover, $\lim_{N\to\infty} \mathbb{P}(T_2^N - T_1^N > \varrho_N^{-1-\delta}) = 0$ for any $\delta > 0$.

Proof. We use an ODE approximation. Recall that K_i denotes the number of mutants at the beginning of day i. Let $x \in [\varepsilon, 1)$. From Corollary 4.3.9, we obtain that the expected number of offspring at the end of day i of a *single* mutant, given that there are $\lfloor xN \rfloor$ mutants at the beginning of the day, is given by $e^{(r+\varrho_N)\sigma_{\lfloor xN \rfloor}}$. Using Corollary 4.3.9, we obtain

$$\mathbb{E}[K_i \,|\, K_{i-1} = \lfloor xN \rfloor] = \frac{\lfloor xN \rfloor}{\gamma}\big(e^{(r+\varrho_N)\sigma_{\lfloor xN \rfloor}}\big) = \lfloor xN \rfloor\big(1 + \varrho_N \frac{\log\gamma}{r}(1-xN) + O(\varrho_N^2)\big). \tag{4.3.48}$$

From Corollary A.3.4 and Corollary 4.3.9 we see that there exists $c = c(\gamma, r) < \infty$ such that

$$\operatorname{var}(K_i \,|\, K_{i-1} = k) \leq cN, k = 1, 2, \ldots, N.$$

For $f \in C^2[0,1]$ we define the rescaled discrete generator of $(K_i)_{i\in\mathbb{N}_0}$

$$A_N f(\frac{k}{N}) = \varrho_N^{-1}\mathbb{E}[f(K_i/N) - f(k/N) \,|\, K_{i-1} = k], \quad x \in [0,1].$$

Using Taylor approximation on f we infer that, for some $y \in [0,1]$,

$$A_N f(\frac{k}{N}) = \varrho_N^{-1}\Big(\mathbb{E}\big[(\frac{K_i}{N} - \frac{k}{N})f'(\frac{k}{N}) + \frac{1}{2}\big(\frac{K_i}{N} - \frac{k}{N}\big)^2 f''(y) \,\big|\, K_{i-1} = k\big]\Big).$$

We have,

$$\begin{aligned}\mathbb{E}_k\big[\big(\frac{K_1}{N} - \frac{k}{N}\big)^2\big] &= \frac{1}{N^2}\mathbb{E}_k[K_1^2 - \mathbb{E}_k[K_1]^2] + \frac{1}{N^2}\mathbb{E}_k[K_1]^2 - 2\frac{k}{N^2}\mathbb{E}_k[K_1] + (\frac{k}{N})^2 \\ &= \frac{1}{N^2}\operatorname{var}_k(K_1) + \big(\mathbb{E}_k[K_1]/N - x\big)^2 \\ &\leq \frac{c}{N} + O(\varrho_N^2),\end{aligned} \tag{4.3.49}$$

where $|O(\varrho_N^2)|/\varrho_N^2$ is bounded uniformly in N and k. Hence recalling $\varrho_N^{-1}N^{-\alpha} \to 0$ for $\alpha > b$ and the continuity of f'' on $[0,1]$, we obtain the following convergence which is uniform in k and y:

$$\sup_{y\in[0,1],k=0,1,\dots,N} |\varrho_N^{-1}\Big(\mathbb{E}_k\big[(\frac{K_1}{N}-\frac{k}{N})^2\big]f''(y)\Big)| \to 0, N\to\infty.$$

Since $\mathbb{E}_k[\frac{K_1}{N}-\frac{k}{N}] = \frac{k}{\gamma}e^{(r+\varrho_N)\sigma_k}-\frac{k}{N}$, one can apply Corollary 4.3.9. Together with the above display, we obtain

$$\sup_{k=0,1,\dots,N} |A_N f(\frac{k}{N}) - \frac{k}{N}(1-\frac{k}{N})f'(\frac{k}{N})| \to 0, N\to\infty.$$

Applying Theorem 1.6.5 and Theorem 4.2.6 of [18] we infer that for every $x\in[0,1]$, the sequence of processes $(\frac{1}{N}K_{\lfloor\varrho_N^{-1}t\rfloor})_{t\geq 0}$, $N=1,2,\dots$, $K_0=\lfloor xN\rfloor$ converges locally uniformly in distribution to the deterministic (increasing) function $g(t)$ which is defined by the initial value problem

$$g'(t) = g(t)(1-g(t))\frac{\log\gamma}{r}, \quad g(0)=x\in[0,1].$$

Now choose t_* such that $g(t_*)>1-\varepsilon'$, provided $g(0)=\varepsilon>0$. This implies

$$\lim_{N\to\infty}\mathbb{P}(K_{\lfloor\varrho_N^{-1}t_*\rfloor}\geq\lfloor(1-\varepsilon')N\rfloor|K_0\geq\lfloor\varepsilon N\rfloor)=1, \tag{4.3.50}$$

and *a fortiori,* $\lim_{N\to\infty}\mathbb{P}(T_1^N - T_2^N > \varrho_N^{-1-\delta}) = 0$ *for any positive* δ. □

Corollary 4.3.16. *For any $\varepsilon\in(0,1/2)$, there exist $N_\varepsilon^{(2)}\in\mathbb{N}$ such that for every $N>N_\varepsilon^{(2)}$, for every $k\geq\varepsilon N$,*

$$\mathbb{P}_k(\exists i\leq\varrho_N^{1-\delta} : K_i\geq\lfloor(1-\varepsilon)N\rfloor)\geq 1/2.$$

Proof. The proof follows immediately from (4.3.50). □

4.3.7 Third stage of the sweep

For the last stage of the sweep, after the number of mutants has reached at least $(1-\varepsilon)N$, we use a Galton Watson coupling similar in spirit to the coupling at the first stage. The difference is that this time we will be working with the process of wild type individuals rather than the mutants. Fix again $\alpha\in(b,1/2)$. Let $Q_i := N-K_i$ be the number of wild-type individuals at the beginning of day i. We proceed similarly as in Section 4.3.1 to define approximating Galton Watson processes $(\underline{Q}_i)_{i\in\mathbb{N}_0}$ and $(\overline{Q}_i)_{i\in\mathbb{N}_0}$, for $i\in\mathbb{N}$ constructing $\underline{Q}_i$ and $\overline{Q}_i$ recursively from the same Yule forest as Q_i : Recall that the wild type individuals reproduce at rate r. Assume that $\underline{Q}_{i-1}$ and $\overline{Q}_{i-1}$ are constructed, and start independent Yule trees growing at rate r for each individual as we did in Section 4.3.1 to construct $\overline{K}_i$ and $\underline{K}_i$. Assume $Q_{i-1}=q\in(0,\varepsilon N)$. Grow the Yule trees until time $\sigma_{\lceil(1-2\varepsilon)N\rceil}$ and distinguish the individuals according to whether they were born before σ_N, before σ_{N-q}, or before $\sigma_{\lceil(1-2\varepsilon)N\rceil}$. Taking the time of birth into consideration, the individuals born before σ_N will be sampled independently with probability $\gamma^{-1}-N^{-\alpha}$ to form $\underline{Q}_i$, born before σ_{N-q} will be chosen according to (4.3.3) to form Q_i, and those before $\sigma_{\lceil(1-2\varepsilon)N\rceil}$ with probability $\gamma^{-1}+N^{-\alpha}$ to form $\overline{Q}_i$.

It is clear that Lemma 4.3.4 and Corollary 4.3.5 still hold, and thus we can prove the equivalent to Proposition 4.3.7. Define

$$T_w^N(m) := \inf\{i : Q_i > m\varepsilon N \text{ or } Q_i = 0\},\ m\geq 1.$$

Lemma 4.3.17. *Let $\alpha\in(b,1/2)$. Let $m\geq 1$, and $0<\varepsilon<1/(m\gamma)$. Assume $\underline{Q}_0 = Q_0 = \overline{Q}_0 \leq \varepsilon N$. Then there exists c large enough such that for N large enough,*

$$\mathbb{P}\Big(\overline{Q}_{\min\{i,T_w^N(m)\}}\geq Q_{\min\{i,T_w^N(m)\}}\geq\underline{Q}_{\min\{i,T_w^N(m)\}},\forall i\leq g\Big)\geq(1-2e^{-cN})^g \quad \textit{for all } g\in\mathbb{N}_0.$$

for some constant c independent of N.

Proof. This follows from a straightforward adaptation of the proof of Proposition 4.3.7, since the condition $\varepsilon \leq 1/(m\gamma)$ allows us to prove the analog of Lemma 4.3.6, observing that the definition of $T_w^N(m)$ ensures that we stop the procedure if Q_i reaches $m\varepsilon N$ individuals (and not εN as in Proposition 4.3.7). □

We have the alternative description corresponding to Proposition 4.3.3: $(\overline{Q}_i)_{i\in\mathbb{N}_0}$ is the Galton Watson process whose offspring distribution is mixed binomial with parameters $\overline{W}$ and $\frac{1}{\gamma} + N^{-\alpha}$, where $\overline{W}$ is geometric with parameter $e^{-r\sigma_{\lceil(1-2\varepsilon)N\rceil}}$. Similarly, $(\underline{Q}_i)_{i\in\mathbb{N}_0}$ is the Galton Watson process whose offspring distribution is mixed binomial with parameters $\underline{W}$ and $\frac{1}{\gamma} - N^{-\alpha}$, where $\underline{W}$ is geometric with parameter $e^{-r\sigma_N}$. From this we obtain the analogue of Lemma 4.3.10.

Lemma 4.3.18. *For $(\underline{Q}_i)_{i\in\mathbb{N}_0}$ and $(\overline{Q}_i)_{i\in\mathbb{N}_0}$ defined above there exist $\overline{c}, \underline{c}$ independent of N such that for N large enough,*

$$\mathbb{E}_1[\overline{Q}_1] = 1 - \overline{c}\varrho_N + o(\varrho_N) \text{ and } \mathbb{E}_1[\underline{Q}_1] = 1 - \underline{c}\varrho_N + o(\varrho_N) \tag{4.3.51}$$

Proof. By construction, and from Corollary 4.3.9

$$\begin{aligned}\mathbb{E}_1[\underline{Q}_1] &= (1/\gamma - N^{-\alpha})\mathbb{E}[\underline{W}] = (1/\gamma - N^{-\alpha})e^{r\sigma_N} \\ &= (1/\gamma - N^{-\alpha})(\gamma - \varrho_N \frac{\log\gamma}{r}) + o(\varrho_N) \\ &= 1 - \frac{\log\gamma}{\gamma r}\varrho_N + o(\varrho_N),\end{aligned}$$

where the last equality follows from the fact that our assumptions imply that $N^{-\alpha} = o(\varrho_N)$. This is the first assertion in (4.3.51). In the same way we obtain $\mathbb{E}_1[\overline{Q}_1] = 1 - \overline{c}\varrho_N + o(\varrho_N)$, for some positive constant $\underline{c}$ independent of N. □

Lemma 4.3.19. *Let $m \geq 1$ and $0 < \varepsilon < 1/(m\gamma)$. For any $k \geq (1-\varepsilon)N$,*

$$\limsup_{N\to\infty} \mathbb{P}_k(\tau_{\text{fix}}^N > \varrho_N^{-1-\delta}) \leq 2/m$$

for any $\delta > 0$. In particular, $\mathbb{P}_k(\exists i : K_i = N) \geq 1 - 2/m$.

Proof. Under $\mathbb{P}_k$ we have by assumption that $K_0 = k \geq (1-\varepsilon)N$, and thus $Q_0 = N - k \leq \varepsilon N$. We consider $(\overline{Q}_i)_{i\in\mathbb{N}_0}, (\underline{Q}_i)_{i\in\mathbb{N}_0}$ as constructed at the beginning of this section, with $\alpha \in (b, 1/2)$. Let

$$A := A(\gamma,\alpha,\varepsilon,N,m) := \left\{\overline{Q}_{\min\{i,T_w^N(m)\}} \geq Q_{\min\{i,T_w^N(m)\}} \geq \underline{Q}_{\min\{i,T_w^N(m)\}}, \forall i \leq \varrho_N^{-1-\delta}\right\}.$$

Then Lemma 4.3.17 shows

$$\mathbb{P}(A) \to 1 \text{ as } N \to \infty.$$

Note that

$$\mathbb{E}_k[\overline{Q}_{\lfloor\varrho_N^{-1-\delta}\rfloor}] \sim (N-k)(1-\bar{c}\varrho_N)^{\varrho_N^{-(1+\delta)}} \leq (N-k)e^{-\bar{c}\varrho_N^{-\delta}} \leq \varepsilon N e^{-\bar{c}\varrho_N^{-\delta}} \to 0$$

as $N \to \infty$. Consequently, since on the event $\{T_\omega^N(m) > \varrho_N^{-1-\delta}\} \cap A$ we have $Q_{\lfloor\varrho_N^{-1-\delta}\rfloor} \geq 1$,

$$\begin{aligned}\mathbb{P}_k(T_w^N(m) > \varrho_N^{-1-\delta}) &\leq \mathbb{P}_k(T_w^N(m) > \varrho_N^{-1-\delta}, A) + \mathbb{P}_k(A^c) \leq \mathbb{E}_k[Q_{\lfloor\varrho_N^{-1-\delta}\rfloor}1_{\{T_w^N(m)>\varrho_N^{-1-\delta}\}}1_A] + \mathbb{P}_k(A^c) \\ &\leq \mathbb{E}_k[\overline{Q}_{\lfloor\varrho_N^{-1-\delta}\rfloor}] + \mathbb{P}_k(A^c) \to 0 \text{ as } N \to \infty.\end{aligned}$$

Since

$$\begin{aligned}\mathbb{P}_k(\tau_{\text{fix}}^N > \varrho_N^{-1-\delta}) =& \mathbb{P}_k(\tau_{\text{fix}}^N > \varrho_N^{-1-\delta}, T_w^N(m) > \varrho_N^{-1-\delta}) + \mathbb{P}_k(\tau_{\text{fix}}^N > \varrho_N^{-1-\delta}, T_w^N(m) \leq \varrho_N^{-1-\delta}) \\ \leq& \mathbb{P}_k(T_w^N(m) > \varrho_N^{-1-\delta}) + \mathbb{P}_k(Q_{T_w^N(m)} \geq \varepsilon m N),\end{aligned}$$

we are left with proving

$$\limsup_{N\to\infty} \mathbb{P}_k(Q_{T_w^N(m)} \geq \varepsilon m N) \leq 2/m. \tag{4.3.52}$$

Let κ be the first time when $(\overline{Q}_i)_{i\geq 0}$ is not less than εmN or equal to 0. Note that under $A\cap\{T_w^N(m)\leq \varrho_N^{-1-\delta}\}$, if $Q_{T_w^N(k)}\geq \varepsilon mN$, then necessarily, $\overline{Q}_{T_w^N(m)}\geq \varepsilon mN$. So in conclusion:

$$\mathbb{P}_k(Q_{T_w^N(m)}\geq \varepsilon mN, A, T_w^N(m)\leq \varrho_N^{-1-\delta})\leq \mathbb{P}_k(\overline{Q}_\kappa\geq \varepsilon mN, A, T_w^N(m)\leq \varrho_N^{-1-\delta}). \tag{4.3.53}$$

Notice that $(\overline{Q}_i)_{i\geq 0}$ is, as a sub-critical Galton Watson process, a supermartingale. Then $(\overline{Q}_i\wedge \varepsilon mN)_{i\geq 0}$ is a bounded supermartingale and, for any time strictly before κ, these two supermartingales are the same. Now we have

$$\varepsilon N\geq \mathbb{E}_k[\overline{Q}_0]=\mathbb{E}_k[Q_0\wedge \varepsilon mN]\geq \mathbb{E}_k[\overline{Q}_\kappa\wedge \varepsilon mN]=\mathbb{P}_k(\overline{Q}_\kappa\geq \varepsilon mN)\varepsilon mN.$$

So

$$\mathbb{P}_k(\overline{Q}_\kappa\geq \varepsilon mN)\leq 1/m.$$

Therefore using (4.3.53) we have for N large enough

$$\mathbb{P}_k(Q_{T_w^N(m)}\geq \varepsilon mN)\leq \mathbb{P}_k(\overline{Q}_\kappa\geq \varepsilon mN)+\mathbb{P}_k(T_w^N(m)>\varrho_N^{-1-\delta})+\mathbb{P}(A^c)\leq 2/m.$$

This implies (4.3.52), and moreover $\mathbb{P}_k(\exists i: K_i=N)=\mathbb{P}_k(Q_{T_w^N(k)}=0)\geq 1-2/m$.

□

This result will be useful in the following simple form:

Corollary 4.3.20. *For every $0<\varepsilon<1/(4\gamma)$ there exist $N_\varepsilon^{(3)}\in\mathbb{N}$ such that for all $N\geq N_\varepsilon^{(3)}, \delta>0$ and $k\geq(1-\varepsilon)N$*

$$\mathbb{P}_k(\tau_{\text{fix}}^N>\varrho_N^{-1-\delta})\leq 1/2.$$

Proof. Take $m\geq 4$ in Lemma 4.3.19. □

4.3.8 Proof of Theorem 4.2.10

We are now finally able to prove Theorem 4.2.10. Let $m\geq 4$ and $0<\varepsilon<1/(m\gamma)\wedge 1/16$. By Lemma 4.3.13 we have

$$\pi_N=\mathbb{P}_1(\exists i: K_i=N)\leq \mathbb{P}(K_i \text{ reaches } \varepsilon N)\leq \frac{\gamma\log\gamma}{\gamma-1}\frac{\varrho_N}{r}(1+o(1)).$$

Further, observe that for $1\leq k\leq k'\leq l\leq N$, by definition of the model,

$$\mathbb{P}_k(K_1\geq l)\leq \mathbb{P}_{k'}(K_1\geq l)$$

and therefore by induction $\mathbb{P}_k(K_i\geq l)\leq \mathbb{P}_{k'}(K_i\geq l), i\in\mathbb{N}$. Thus

$$\mathbb{P}_k((K_i) \text{ reaches } l)\leq \mathbb{P}_{k'}((K_i) \text{ reaches } l).$$

Therefore, for every $\varepsilon\in(0,1/(m\gamma)\wedge 1/16)$, by the strong Markov property and Lemma 4.3.13,

$$\begin{aligned}\pi_N\geq&\mathbb{P}_{\lfloor\varepsilon N\rfloor}(\exists i: K_i=N)\cdot\mathbb{P}_1(K_i \text{ reaches } \varepsilon N)\\ \geq&\mathbb{P}_{\lfloor\varepsilon N\rfloor}(\exists i: K_i=N)\cdot\frac{\gamma\log\gamma}{\gamma-1}\frac{\varrho_N}{r}(1-\varepsilon)(1+o(1)).\end{aligned}$$

From Lemmas 4.3.19 and 4.3.15 we obtain $\liminf_{N\to\infty}\mathbb{P}_{\lfloor\varepsilon N\rfloor}(\exists i: K_i=N)\geq 1-2/m$ for any $m\geq 2$. Thus

$$(1-\varepsilon)(1-2/m)\leq \liminf_{N\to\infty}\frac{\gamma-1}{\gamma\log\gamma}\frac{r}{\varrho_N}\pi_N\leq \limsup_{N\to\infty}\frac{\gamma-1}{\gamma\log\gamma}\frac{r}{\varrho_N}\pi_N\leq 1.$$

Sending $m\to\infty$ (and $\varepsilon\to 0$) gives (4.2.12).

Now we will prove that $\mathbb{P}_1(\tau^N>\varrho_N^{-1-2\delta})\leq(7/8)^{\varrho_N^{-\delta}}$. Let $N_\varepsilon=\sup\{N_\varepsilon^{(1)},N_\varepsilon^{(2)},N_\varepsilon^{(3)}\}$ where $N_\varepsilon^{(1)},N_\varepsilon^{(2)},N_\varepsilon^{(3)}$ can be found respectively in Corollary 4.3.14, 4.3.16 and 4.3.20. Using the three corollaries and the strong Markov property of the process $(K_i)_{i\in\mathbb{N}_0}$ we know that for all $N>N_\varepsilon$, and for any $k\in\{1,2,...,N\}$

$$\mathbb{P}_k(\tau^N\leq 3\varrho_N^{-1-\delta})\geq(1/2)^3. \tag{4.3.54}$$

Using the Markov property at time $\lceil 3\varrho_N^{-1-\delta}\rceil$, we see that for any $n \in \mathbb{N}$

$$\begin{aligned}\mathbb{P}_1(\tau^N > 3n\varrho_N^{-1-\delta}) &\leq \mathbb{P}_1(\tau^N > \lceil 3\varrho_N^{-1-\delta}\rceil) \sum_{k=1}^{N-1} \mathbb{P}_k(\tau^N > 3(n-1)\varrho_N^{-1-\delta})\mathbb{P}_1(K_{\lceil \varrho_N^{-1-\delta}\rceil} = k) \\ &\leq (1-(1/2)^3) \sum_{k=1}^{N-1} \mathbb{P}_k(\tau^N > 3(n-1)\varrho_N^{-1-\delta})\mathbb{P}_1(K_{\lceil \varrho_N^{-1-\delta}\rceil} = k).\end{aligned}$$

Thus, proceeding iteratively, and using the fact that (4.3.54) is uniform in $k \in \{1, ..., N-1\}$, we obtain

$$\mathbb{P}_1(\tau^N > 3n\varrho_N^{-1-\delta}) \leq (1-(1/2)^3)^n.$$

In particular, choosing $n = \lceil \varrho_N^{-\delta}\rceil$ we obtain for $\delta > 0$

$$\mathbb{P}_1(\tau^N > \varrho_N^{-1-3\delta})\mathbb{P}_1(\tau^N > 3\varrho_N^{-1-2\delta}) \leq (7/8)^{\varrho_N^{-\delta}}.$$

□

4.3.9 Proof of Proposition 4.2.13

Due to Theorem 4.2.10, and due to the Assumption that the mutations arrive independently of each other at geometric times with parameter μ_N, we have that for any $\delta' > 0$

$$\begin{aligned}\mathbb{P}(m_N < \tau^N) \leq& 1 - \mathbb{P}(m_N > \varrho_N^{-1-\delta'} \,|\, \tau^N < \varrho_N^{-1-\delta'})\mathbb{P}(\tau^N < \varrho_N^{-1-\delta'}) \\ \leq& 1 - (1-\mu_N)^{\lfloor \varrho_N^{-1-\delta'}\rfloor}(1-(7/8)^{\lfloor \varrho_N^{-\delta'/3}\rfloor}).\end{aligned}$$

Now the Bernoulli inequality yields

$$\begin{aligned}\mathbb{P}(m_N < \tau^N) &\leq 1 - (1-\mu_N\lfloor\varrho_N^{-1-\delta'}\rfloor)(1-(7/8)^{\lfloor \varrho_N^{-\delta'/3}\rfloor}) \\ &= \mu_N\lfloor\varrho_N^{-1-\delta'}\rfloor + (7/8)^{\lfloor \varrho_N^{-\delta'/3}\rfloor} - \mu_N\lfloor\varrho_N^{-1-\delta'}\rfloor(7/8)^{\lfloor \varrho_N^{-\delta'/3}\rfloor}.\end{aligned}$$

From this we obtain

$$\mathbb{P}(m_N < \tau^N) \leq \mu_N\varrho_N^{-1-\delta}$$

for any $\delta > \delta'$, provided N is large enough. This proves the first claim. Now, let E_j be the event that there is no clonal interference until the day that the j-th successful mutation starts. Observe that $\mathbb{P}(E_1)$ is given by the probability that any unsuccessful mutation started before the first successful one has disappeared before the next mutation (successful or unsuccessful) starts. By the first part of this theorem, for any given mutation this is the case with probability $\mathbb{P}(m_N \geq \tau^N) \geq 1 - \mu_N\varrho_N^{-1-\delta}$, for $\delta > 0$. Denote by L the number of mutations until the first successful one. Since the mutations arrive independently of each other, we see by induction that for $l \in \mathbb{N}_0$

$$\mathbb{P}(\text{ no clonal interference in the first } l \text{ mutations} \,|\, L = l+1) \geq (1-\mu_N\varrho_N^{-1-\delta})^l.$$

By Theorem 4.2.10, L is (asymptotically) geometric with success parameter $C(\gamma)\varrho_N/r_0$. Thus summing over all possible values of L we obtain by Theorem 4.2.10 and the first part of this proof, for $\delta > 0$,

$$\begin{aligned}\mathbb{P}(E_1) &\geq \sum_{l=0}^{\infty} \mathbb{P}(L = l+1)(1-\mu_N\varrho_N^{-1-\delta})^l \\ &\geq \sum_{l=0}^{\infty} (1-\frac{C(\gamma)\varrho_N}{r_0})^l \frac{C(\gamma)\varrho_N}{r_0}(1-\mu_N\varrho_N^{-1-\delta})^l \\ &= \frac{C(\gamma)\varrho_N}{r_0}\sum_{i=0}^{\infty}(1-\frac{C(\gamma)\varrho_N}{r_0} - \mu_N\varrho_N^{-1-\delta} + 3\frac{C(\gamma)}{r_0}\varrho_N^{-\delta})^l \\ &= \frac{C(\gamma)\varrho_N}{r_0}\frac{1}{C(\gamma)\varrho_N r_0^{-1} + \mu_N\varrho_N^{-1-\delta} - C(\gamma)r_0^{-1}\mu_N\varrho_N^{-\delta}} \\ &= \frac{1}{1+\mu_N\varrho_N^{-2-\delta}r_0C(\gamma)^{-1} - \mu_N\varrho_N^{-1-\delta}} \\ &\geq 1 - \mu_N\varrho_N^{-2-\delta''} + o(\mu_N\varrho_N^{-2-\delta''})\end{aligned}$$

for N large enough and $\delta'' > \delta$. Fix $n \in \mathbb{N}$. Similar to the previous calculation, for $j \leq n\varrho_N^{-1}$, we have $\mathbb{P}(E_{j+1}|E_j) \geq 1 - \mu_N \varrho_N^{-2-\delta''}/3 + o(\mu_N \varrho_N^{-2-\delta''})$. Proceeding iteratively one thus observes that for any fixed $n \in \mathbb{N}$

$$\begin{aligned}\mathbb{P}(E_{\lfloor \varrho_N^{-1} n \rfloor}) &\geq (1 - \mu_N \varrho_N^{-2-\delta''} + o(\mu_N \varrho_N^{-2-\delta}))^{\lfloor n\varrho_N^{-1} \rfloor} \\ &\geq 1 - n\mu_N \varrho_N^{-3-3\delta''}(1 + o(1)). \end{aligned} \tag{4.3.55}$$

By Assumption A iii) this tends to 1 for $\delta'' > 0$ small enough. Let I_n be the day at which the $\lfloor \varrho_N^{-1} n \rfloor$-th successful mutation starts. We can write

$$I_n = \sum_{j=1}^{n} I^{(j)},$$

if $I^{(j)}$ denotes the time between the fixation of the $j-1$th and the initiation of the jth successful mutation (and $I^{(1)} = I_1$). Let $L^{(j)}$ denote the number of unsuccessful mutations that happen during time $I^{(j)}$. The success probability of a mutation that happens during $I^{(j)}$ is according to Theorem 4.2.10 given by $C(\gamma)\frac{\varrho_N}{r_0+(j-1)\varrho_N}$. Therefore, conditional on E_j, $L^{(j)}$ is geometrically distributed with success parameter $C(\gamma)\frac{\varrho_N}{r_0+(j-1)\varrho_N}$. Moreover, conditional on E_j, the time between two of the $L^{(j)}$ unsuccessful mutations is stochastically larger than a geometric random variable with parameter μ_N, since this is the rate at which mutations arrive, and the geometric distribution is memoryless. Thus we see that the time $I^{(j)}$ is stochastically larger than a geometric random variable with parameter $\frac{C(\gamma)\mu_N\varrho_N}{r_0+(j-1)\varrho_N}$ and a fortiori stochastically larger than G_j^N, if $(G_j^N)_{j\in\mathbb{N}_0}$ is a sequence of independent geometric random variables with parameter $C(\gamma)\mu_N\varrho_N/r_0$. Thus conditionally on $E_{\lfloor n\varrho_N^{-1}\rfloor}$, stochastically $I_n \geq \sum_{j=1}^{\lfloor \varrho_N^{-1} n\rfloor} G_j^N$. Let $n = \lceil 2Tr_0/C(\gamma)\rceil$. Then

$$\begin{aligned}\lim_{N\to\infty} \mathbb{P}(\text{no clonal interference until } \varrho_N^{-2}\mu_N^{-1}T) &\geq \mathbb{P}(E_{\lfloor \varrho_N^{-1} n\rfloor},\ I_n > \lceil \varrho_N^{-2}\mu_N^{-1}T\rceil) \\ &= \mathbb{P}(E_{\lfloor \varrho_N^{-1} n\rfloor})\mathbb{P}(I_n > \lceil \varrho_N^{-2}\mu_N^{-1}T\rceil | E_{\lfloor \varrho_N^{-1} n\rfloor}) \\ &\geq \mathbb{P}(E_{\lfloor \varrho_N^{-1} n\rfloor})(1 - 2\mathbb{P}(\sum_{j=1}^{\lfloor \varrho_N^{-1} n\rfloor} G_j^N < \lfloor \varrho_N^{-2}\mu_N^{-1}T\rfloor)) \end{aligned}$$

By Cramér's large deviation principle the second factor tends to 1. Thus the statement follows from (4.3.55). □

4.3.10 Proof of Theorem 4.2.14

Denote by D_i the event that there is no clonal interference up to day i, that is, any mutation that starts until or including day i happens in a homogeneous population. Define

$$\tilde{H}_i := H_i 1_{D_i} - \infty 1_{D_i^c}.$$

Then we have for any $T > 0$ that the two processes $(H_i)_{1\leq i\leq \varrho_N^{-2}\mu_N^{-1}T}$ and $(\tilde{H}_i)_{1\leq i\leq \varrho_N^{-2}\mu_N^{-1}T}$ coincide on the event $(D^c_{\lceil \varrho_N^{-2}\mu_N^{-1}T\rceil})$, whose probability converges to 0 as $N \to \infty$, by Proposition 4.2.13. Thus it is sufficient to show that $(\tilde{H}_{\lfloor t\varrho_N^{-1}\mu_N^{-1}\rfloor})_{0\leq t\leq T}$ converges in distribution to $(M(C(\gamma)t/r_0))_{0\leq t\leq T}$ w. r. to the Skorokhod topology, cf. Theorem 3.3.1 in [18]. This will be achieved by a standard generator calculation. The process $(H_i)_{i\in\mathbb{N}_0}$ is a Markov chain on $\mathbb{N}_0 \cup \{-\infty\}$ with the following transition probabilities: If $n \geq 0$, then

$$\begin{aligned}\mathbb{P}(\tilde{H}_{i+1} = n+1 \,|\, \tilde{H}_i = n) &= \frac{C(\gamma)\mu_N\varrho_N}{r_0 + n\varrho_N}\mathbb{P}(D_{i+1} \,|\, D_i), \\ \mathbb{P}(\tilde{H}_{i+1} = n \,|\, \tilde{H}_i = n) &= \Big(1 - \frac{C(\gamma)\mu_N\varrho_N}{r_0 + n\varrho_N}\Big)\mathbb{P}(D_{i+1} \,|\, D_i), \\ \mathbb{P}(\tilde{H}_{i+1} = -\infty \,|\, \tilde{H}_i = n) &= \mathbb{P}(D_{i+1}^c \,|\, D_i), \end{aligned}$$

and

$$\mathbb{P}(\tilde{H}_{i+1} = -\infty \,|\, \tilde{H}_i = -\infty) = 1.$$

Observe first that for any $\delta > 0$ we have

$$\mathbb{P}(D_{i+1}^c \,|\, D_i) \leq \mu_N^2 \varrho_N^{-1-\delta}. \tag{4.3.56}$$

This follows since conditional on the event D_i, the event D_{i+1}^c can only happen if at day $i+1$ a new mutation happens, and interferes with the previous mutation. The probability that a new mutation happens is given by μ_N, and the probability of interference of a pair of mutations is $\mathbb{P}(m_N < \tau^N)$. Thus (4.3.56) follows from Proposition 4.2.13.
For bounded functions g on $\mathbb{N}_0 \cup \{-\infty\}$, the discrete generator of $(\tilde{H}_i)_{i\in\mathbb{N}_0}$ on the time scale $i = \varrho_N^{-1}\mu_N^{-1}t$ is given by (cf. Theorem 1.6.5 of [18])

$$\begin{aligned}
B_N g(n) :=& \frac{1}{\varrho_N \mu_N} \mathbb{E}\big[g(\tilde{H}_{i+1}) - g(n) \,\big|\, \tilde{H}_i = n\big] \\
=& \frac{1}{\varrho_N \mu_N} \Big(\frac{C(\gamma)\mu_N \varrho_N}{r_0 + n\varrho_N} \mathbb{P}(D_{i+1} \,|\, D_i)(g(n+1) - g(n)) + \mathbb{P}(D_{i+1}^c \,|\, D_i)(g(-\infty) - g(n)) \Big) \\
=& \frac{C(\gamma)}{r_0 + n\varrho_N} \mathbb{P}(D_{i+1} \,|\, D_i)(g(n+1) - g(n)) + \frac{\mathbb{P}(D_{i+1}^c \,|\, D_i)}{\varrho_N \mu_N}(g(-\infty) - g(n)).
\end{aligned}$$

Due to (4.3.56) and Assumption A iii), the r.h.s. converges as $N \to \infty$ to

$$\frac{C(\gamma)}{r_0}(g(n+1) - g(n)),$$

which is the generator of the Poisson process $(W(C(\gamma)t/r_0))_{t\geq 0}$. By Theorem 4.2.6 of [18] this implies convergence of the corresponding processes. □

4.3.11 Convergence of the fitness process

Proof of Theorem 4.2.15. We proceed analogously to the proof of Theorem 4.2.14. Define

$$\tilde{\Phi}_i := 1 + \frac{\varrho_N}{r_0}\tilde{H}_i,$$

and recall $\Phi_i = 1 + \frac{\varrho_N}{r_0} H_i$. As above, observe that the two processes $(\Phi_i)_{1\leq i\leq \varrho_N^{-2}\mu_N^{-1}T}$ and $(\tilde{\Phi}_i)_{1\leq i\leq \varrho_N^{-2}\mu_N^{-1}T}$ coincide on the event $D^c_{\lceil \varrho_N^{-2}\mu_N^{-1}T\rceil}$, whose probability converges to 0 as as $N \to \infty$, and that $(\tilde{\Phi}_i)_{i\in\mathbb{N}_0}$ is a Markov chain with transition probabilities

$$\begin{aligned}
\mathbb{P}(\tilde{\Phi}_{i+1} = x + \frac{\varrho_N}{r_0} \,|\, \tilde{\Phi}_i = x) &= \frac{C(\gamma)\mu_N\varrho_N}{x r_0}\mathbb{P}(D_{i+1} \,|\, D_i), \\
\mathbb{P}(\tilde{\Phi}_{i+1} = x \,|\, \tilde{\Phi}_i = x) &= \Big(1 - \frac{C(\gamma)\mu_N\varrho_N}{x r_0}\Big)\mathbb{P}(D_{i+1} \,|\, D_i), \\
\mathbb{P}(\tilde{\Phi}_{i+1} = -\infty \,|\, \tilde{\Phi}_i = x) &= \mathbb{P}(D_{i+1}^c \,|\, D_i),
\end{aligned}$$

for $x > 0$ and

$$\mathbb{P}(\tilde{\Phi}_{i+1} = -\infty \,|\, \tilde{\Phi}_i = -\infty) = 1.$$

Thus the discrete generator of $(\tilde{\Phi}_i)_{i\in\mathbb{N}_0}$ on the time scale $i = \varrho_N^{-2}\mu_N^{-1}t$ is given by

$$\begin{aligned}
A_N g(n) :=& \frac{1}{\varrho_N^2 \mu_N} \mathbb{E}\big[g(\tilde{\Phi}_{i+1}) - g(x) \,\big|\, \tilde{\Phi}_i = x\big] \\
=& \frac{1}{\varrho_N^2 \mu_N} \Big(\frac{C(\gamma)\mu_N \varrho_N}{x r_0} \mathbb{P}(D_{i+1} \,|\, D_i)(g(x + \frac{\varrho_N}{r_0}) - g(x)) + \mathbb{P}(D_{i+1}^c \,|\, D_i)(g(-\infty) - g(x)) \Big) \\
=& \frac{C(\gamma)}{\varrho_N r_0 x} \mathbb{P}(D_{i+1} \,|\, D_i)(g(x + \frac{\varrho_N}{r_0}) - g(x)) + \frac{\mathbb{P}(D_{i+1}^c \,|\, D_i)}{\varrho_N^2 \mu_N}(g(-\infty) - g(x)).
\end{aligned}$$

Due to (4.3.56) and Assumption A iii), the r.h.s. converges for a continuously differentiable function $g : \mathbb{R} \to \mathbb{R}$ that vanishes at ∞, as $N \to \infty$ to

$$Ag(x) := \frac{C(\gamma)}{r_0^2 x} g'(x)$$

as $N \to \infty$ (as can be seen from Taylor's expansion, compare the proof of Lemma 4.3.15). This, in turn, is the generator of the solution to the (deterministic) differential equation

$$\dot{h}(t) = \frac{1}{h(t)} \frac{C(\gamma)}{r_0^2}, \ t \geq 0,$$

whose solution (for the initial value $h(0) = 1$) is f. So we can apply Theorem 4.2.6 in [18] to conclude that $\tilde{\Phi}$ and then Φ converges in distribution to $(f(t))_{t \geq 0}$ in the Skorokhod topology. Convergence of F follows from the relation (4.2.16). Since f is continuous, this amounts to locally uniform convergence in distribution. □

Proof of Corollary 4.2.16. The proof is as for Theorem 4.2.15, with the only difference that now we replace $\tilde{\Phi}$ by $\tilde{\Phi}^{\psi}$, with transition probabilities for $x \geq 1$

$$\begin{aligned}
\mathbb{P}(\tilde{\Phi}^{\psi}_{i+1} = x + \frac{\psi(x)\varrho_N}{r_0} \,|\, \tilde{\Phi}^{\psi}_i = x) &= \frac{C(\gamma)\mu_N \varrho_N \psi(x)}{x r_0} \mathbb{P}(D_{i+1} \,|\, D_i), \\
\mathbb{P}(\tilde{\Phi}^{\psi}_{i+1} = x \,|\, \tilde{\Phi}^{\psi}_i = x) &= \Big(1 - \frac{C(\gamma)\mu_N \varrho_N \psi(x)}{x r_0}\Big) \mathbb{P}(D_{i+1} \,|\, D_i), \\
\mathbb{P}(\tilde{\Phi}^{\psi}_{i+1} = -\infty \,|\, \tilde{\Phi}^{\psi}_i = x) &= \mathbb{P}(D^c_{i+1} \,|\, D_i),
\end{aligned}$$

for $x > 0$ and

$$\mathbb{P}(\tilde{\Phi}^{\psi}_{i+1} = -\infty \,|\, \tilde{\Phi}^{\psi}_i = -\infty) = 1.$$

which leads to a slightly different discrete generator

$$A^{\psi}_N g(x) \quad = \quad \frac{C(\gamma)\psi(x)}{\varrho_N r_0 x} \mathbb{P}(D_{i+1} \,|\, D_i) \big(g(x + \frac{\psi(x)\varrho_N}{r_0}) - g(x)\big) + \frac{\mathbb{P}(D^c_{i+1} \,|\, D_i)}{\varrho_N^2 \mu_N} (g(-\infty) - g(x)).$$

Thus we get

$$\lim_{N \to \infty} A^{\psi}_N g(x) = \frac{\psi(x)^2 C(\gamma)}{r_0^2 x} g'(x)$$

and we conclude as above. In particular, solving

$$\dot{h}(t) = \frac{C(\gamma)}{r_0^2} \frac{1}{h(t)^{2q+1}}$$

yields (4.2.18).

□

Appendix A

Some calculations and technical remarks

A.1 Bound on a mixing time

We will prove a bound on the mixing time, that we used in the proof of Theorem 2.3.10.

Lemma A.1.1. *Let $\epsilon > 0$ and $\beta > 0$. Let $\mu_N = (1-\epsilon)\delta_1 + \epsilon\delta_{N^\beta}$. Let (X_k) be the urn process and ν_N its stationary distribution. For $\beta > 0$ let $\mathcal{P}_{N^\beta}$ denote the set of probability measures on $\{0, ..., N^\beta - 1\}$. For all $\lambda > 3\beta > 0$, there exist $\delta > 0$ and $N_0 \in \mathbb{N}$, such that for all $N \geq N_0$*

$$\sup_{\mu \in \mathcal{P}_{N^\beta}} \|\mathbb{P}_\mu(X_{N^\lambda} \in \cdot\,) - \nu_N\|_{TV} \leq e^{-N^\delta}.$$

In particular the mixing time of (X_k), τ_{mix}, fulfills

$$\tau_{mix} \leq N^\lambda.$$

Proof. Let $(Z_n)_{n\in\mathbb{N}_0}$ be a realization of the urn process started in the invariant distribution ν_N independent of $(X_n)_{n\in\mathbb{N}_0}$. We couple (X_n) and (Z_n) by the Doeblin coupling in the following way: let $\sigma_0 := \inf\{n \in \mathbb{N}_0 : X_n = Z_n\}$. Define

$$\tilde{X}_n := \begin{cases} X_n & \text{if } n \leq \sigma_0, \\ Z_n & \text{if } n > \sigma_0. \end{cases}$$

Write $\mathbb{P} := \mathbb{P}_{\gamma\otimes\nu_N}$. Then $\mathbb{P}(\tilde{X}_n = k) = \mathbb{P}_\gamma(X_n = k)$ for all $n \in \mathbb{N}_0, k \in \{0, ..., N^\beta - 1\}$. By Example 1.2.4, we have

$$\|\mathbb{P}_\mu(X_n \in \cdot\,) - \nu_N\|_{TV} \leq \mathbb{P}(\tilde{X}_n \neq Z_n) = \mathbb{P}(\sigma_0 > n). \tag{A.1.1}$$

Our aim is therefore to bound $\mathbb{P}(\sigma_0 > n)$. To this end we consider the difference of the two process at particular times. Define $m_0 := \inf\{n \geq 0 : X_n = 0\}, l_0 := \inf\{n \geq 0 : Z_n = 0\}$, and let recursively, for $i \geq 1$,

$$m_i := \inf\{n > m_{i-1} : X_n = 0, X_{n-1} = 1\}$$

and

$$l_i := \inf\{n > l_{i-1} : Z_n = 0, Z_{n-1} = 1\}.$$

Note that for all $i \geq 0$ we have $Z_{m_i} - X_{m_i} \geq 0$ and $X_{l_i} - Z_{l_i} \geq 0$. Without loss of generality we can assume that $Z_0 - X_0 > 0$, which implies $m_0 < l_0$. Since the difference of the two processes remains constant as long as none of the two processes is in urn 0, we see that

$$\sigma_0 \in \{m_i : i \geq 2\} \cup \{l_i : i \geq 1\}, \tag{A.1.2}$$

i.e. the coupling always happens in urn 0, and it happens if either process (Z_n) jumps from 1 to 0 while (X_n) is in 0 or vice versa.

Define for $i \geq 0$

$$V_i := \big|\{n \in \{m_i, ..., m_{i+1} - 1\} : X_n = 0\}\big|,$$

and

$$W_i := \big|\{n \in \{l_i, ..., l_{i+1} - 1\} : Z_n = 0\}\big|,$$

the number of visits in urn 0 of either of the process during one 'cycle' (note that between m_i and m_{i+1} the process (X_k) has exactly one jump of lenght N^β. By construction, $(V_i)_{i\geq 0}$ and $(W_i)_{i\geq 0}$ are independent sequences of iid geometric random variables with parameter ε, and

$$(Z_{m_i} - X_{m_i}) - (Z_{m_{i-1}} - X_{m_{i-1}}) = W_{i-1} - V_{i-1}, \tag{A.1.3}$$

$$(X_{l_i} - Z_{l_i}) - (X_{l_{i-1}} - Z_{l_{i-1}}) = V_{i-1} - W_{i-1}, \tag{A.1.4}$$

$i \geq 1$. Moreover we note that

$$m_{i+1} - m_i = V_i + N^\beta, \quad l_{i+1} - l_i = W_i + N^\beta. \tag{A.1.5}$$

The random sequence $(\sum_{i=0}^k (V_i - W_i))_{k\geq 0}$ is a random walk with centered increments whose variance (depending on ε but not on N) is finite. Moreover, σ_0 can be controlled by the first time this random walk exits the set $\{-N^\beta + 1, ..., N^\beta - 1\}$, since this event corresponds to either (Z_n) 'catching up' with (X_n), or vice versa. More precisely, defining

$$R := \inf\big\{k \geq 0 : |\sum_{i=0}^{k} (V_i - W_i)| \geq N^\beta\big\},$$

we see from (A.1.3) and (A.1.4) that

$$\sigma_0 \leq m_R. \tag{A.1.6}$$

Equation (A.1.6) implies that for any $\lambda > 0$,

$$\begin{aligned}
\mathbb{P}\big(\sigma_0 > N^\lambda\big) &\leq \mathbb{P}\big(\sum_{i=1}^{R} (m_i - m_{i-1}) > N^\lambda\big) \\
&= 1 - \mathbb{P}\big(\sum_{i=1}^{R} (m_i - m_{i-1}) \leq N^\lambda\big) \\
&\leq 1 - \mathbb{P}\big(\{R < \frac{1}{2} N^{\lambda-\beta}\} \cap \{m_i - m_{i-1} \leq 2N^\beta \ \forall i = 1, ..., \frac{1}{2} N^{\lambda-\beta}\}\big) \\
&\leq \mathbb{P}\big(\{R > \frac{1}{2} N^{\lambda-\beta}\} \cup \{\exists 1 \leq i \leq \frac{1}{2} N^{\lambda-\beta} : m_i - m_{i-1} > 2N^\beta\}\big) \\
&\leq \mathbb{P}\big(R > \frac{1}{2} N^{\lambda-\beta}\big) + \mathbb{P}\big(\exists 1 \leq i \leq N^{\lambda-\beta} : m_i - m_{i-1} > 2N^\beta\big).
\end{aligned} \tag{A.1.7}$$

To control the first term on the rhs, we use classical bounds on the exit time from an interval of symmetric random walks with finite variance, see e.g. Theorem 23.2 of [66]. This provides that for every $\delta' > 0$ there exists $\delta > 0$ such that

$$\mathbb{P}\big(R > N^{2\beta+\delta'}\big) \leq e^{-N^\delta}. \tag{A.1.8}$$

For $\lambda > 3\beta$, we can choose $\delta' > 0$ such that $2\beta + \delta' < \lambda - \beta$, hence we find the bound

$$\mathbb{P}\big(R > \frac{1}{2} N^{\lambda-\beta}\big) \leq e^{-N^\delta}. \tag{A.1.9}$$

To bound the second term in (A.1.7), by (A.1.5) and a union bound we find, for N large enough,

$$\begin{aligned}
\mathbb{P}\big(\exists 1 \leq i \leq N^{\lambda-\beta} : m_i - m_{i-1} > 2N^\beta\big) &\leq N^{\lambda-\beta} \mathbb{P}\big(V_1 > N^\beta\big) \\
&= N^{\lambda-\beta}(1-\varepsilon)^{N^\beta} \leq N^{\lambda-\beta} e^{-\varepsilon N^\beta} \leq e^{-N^{\beta/2}}.
\end{aligned} \tag{A.1.10}$$

In view of (A.1.1), together the bounds (A.1.7), (A.1.9) and (A.1.10) prove the Lemma. □

A.2 Convergence to the seedbank diffusion

Proposition A.2.1. *Assume $c = \varepsilon N = \delta M$ and $M \to \infty$, $N \to \infty$. Let $(D_{N,M})_{N,M\in\mathbb{N}}$ be an array of positive real numbers. Then the discrete generator of the allele frequency process $(X^N_{\lfloor D_{N,M}t\rfloor}, Y^M_{\lfloor D_{N,M}t\rfloor})_{t\in\mathbb{R}^+}$ on time-scale $D_{N,M}$ is given by*

$$(A^N f)(x,y) = D_{N,M}\Big[\frac{c}{N}(y-x)\frac{\partial f}{\partial x}(x,y) + \frac{c}{M}(x-y)\frac{\partial f}{\partial y}(x,y) + \frac{1}{N}\frac{1}{2}x(1-x)\frac{\partial^2 f}{\partial x^2}(x,y) + R(N,M)\Big],$$

where the remainder term $R(N,M)$ satisfies that there exists a constant $C_1(c,f) \in (0,\infty)$, independent of N and M, such that

$$|R(N,Ma)| \le C_1(N^{-3/2} + M^{-2} + N^{-1}M^{-1} + NM^{-3}).$$

In particular, in the situation where $M = O(N)$ as $N \to \infty$ and $D_{N,M} = N$ we immediately obtain Proposition 3.2.4.

Proof. We calculate the generator of $(X^N_k, Y^M_k)_{k\ge 0}$ depending on the scaling $(D_{N,M})_{N,M\in\mathbb{N}}$. For $f \in \mathcal{C}^3([0,1]^2)$ we use Taylor expansion in 2 dimensions to obtain

$$\begin{aligned}(A^N f)(x,y) = \frac{1}{D_{N,M}}\Big[&\frac{\partial f}{\partial x}(x,y)\mathbb{E}_{x,y}\left[X^N_1 - x\right] + \frac{\partial f}{\partial y}(x,y)\mathbb{E}_{x,y}\left[Y^M_1 - y\right]\\ &+ \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(x,y)\mathbb{E}_{x,y}\left[\left(X^N_1 - x\right)^2\right]\\ &+ \frac{1}{2}\frac{\partial^2 f}{\partial y^2}(x,y)\mathbb{E}_{x,y}\left[\left(Y^M_1 - y\right)^2\right]\\ &+ \frac{\partial^2 f}{\partial x\partial y}(x,y)\mathbb{E}_{x,y}\left[\left(X^N - x\right)\left(Y^M_1 - y\right)\right]\\ &+ \mathbb{E}_{x,y}\Big[\sum_{\substack{\alpha,\beta\in\mathbb{N}_0\\ \alpha+\beta=3}} R^{\alpha,\beta}(X^N_1, Y^N_1)(X^M_1 - x)^\alpha (Y^M_1 - y)^\beta\Big]\Big]\end{aligned}$$

where the remainder is given by

$$R^{\alpha,\beta}(\bar{x},\bar{y}) := \frac{\alpha+\beta}{\alpha!\beta!}\int_0^1 (1-t)^{\alpha+\beta-1}\frac{\partial^3 f}{\partial x^\alpha \partial y^\beta}\big(x - t(\bar{x}-x), y - t(\bar{y}-y)\big)dt$$

for any $\bar{x},\bar{y} \in [0,1]$. In order to prove the convergence, we thus need to calculate or bound all the moments involved in this representation.

Given $\mathbb{P}_{x,y}$ the following holds: By Proposition 3.2.2

$$\begin{aligned}X^N_1 &= \frac{1}{N}(U+Z),\\ Y^M_1 &= \frac{1}{M}(yM - Z + V),\end{aligned}$$

in distribution where U, V and Z are independent random variables such that

$$\begin{aligned}U &\sim \mathrm{Bin}(N-c,x),\\ V &\sim \mathrm{Bin}(c,x),\\ Z &\sim \mathrm{Hyp}(M,c,yM).\end{aligned}$$

Thus we have

$$\mathbb{E}_{x,y}[U] = Nx - cx, \quad \mathbb{E}_{x,y}[V] = cx, \quad \mathbb{E}^{[}_{x,y}Z] = cy,$$

and moreover

$$\mathbb{V}_{x,y}(U) = (N-c)x(1-x).$$

One more observation is that as $0 \leq V \leq c$ and $0 \leq Z \leq c$, it follows that $|Z - cX| \leq c$ and $|V - Z| \leq c$, which implies that for every $\alpha \in \mathbb{N}$

$$|\mathbb{E}_{x,y}[(Z - cX)^\alpha]| \leq c^\alpha,$$
$$|\mathbb{E}_{x,y}[(Z - V)^\alpha]| \leq c^\alpha,$$

and for every $\alpha, \beta \in \mathbb{N}$

$$|\mathbb{E}_{x,y}[(Z - cX)^\alpha (V - Z)^\beta]| \leq c^{\alpha+\beta} \tag{A.2.1}$$

We are now prepared to calculate all the mixed moments needed.

$$\begin{aligned}\mathbb{E}_{x,y}[X_1^N - x] &= \frac{1}{N}\mathbb{E}_{x,y}[U + Z - Nx] \\ &= \frac{1}{N}\mathbb{E}_{x,y}[U - Nx + cx] + \frac{1}{N}\mathbb{E}_{x,y}[Z - cx] \\ &= \frac{c}{N}(y - x)\end{aligned}$$

Here we used (A.2), in particular $\mathbb{E}_{x,y}[U - Nx + cx] = \mathbb{E}_{x,y}[U - \mathbb{E}_{x,y}[U]] = 0$. Similarly,

$$\begin{aligned}\mathbb{E}_{x,y}[Y_1^M - y] &= \frac{1}{M}\mathbb{E}_{x,y}[My + V - Z - My] \\ &= \frac{1}{M}\mathbb{E}_{x,y}[V - Z] \\ &= \frac{c}{M}(x - y).\end{aligned}$$

Noting $X_1^N - x = \frac{1}{N}(U - Nx + cx) + \frac{1}{N}(Z - cx)$ leads to

$$\begin{aligned}\mathbb{E}_{x,y}[(X_1^N - x)^2] =& \frac{1}{N^2}\mathbb{E}_{x,y}[(U - Nx + cx)^2] \\ &+ \frac{2}{N^2}\mathbb{E}_{x,y}[U - Nx + cx]\mathbb{E}_{x,y}[Z - cx] \\ &+ \frac{1}{N^2}\mathbb{E}_{x,y}[(Z - cx)^2] \\ =& \frac{1}{N^2}\mathbb{V}_{x,y}[U] + \frac{1}{N^2}\mathbb{E}_{x,y}[(Z - cx)^2] \\ =& \frac{1}{N}x(1 - x) - \frac{c}{N^2}x(1 - x) + \frac{1}{N^2}\mathbb{E}_{x,y}[(Z - cx)^2],\end{aligned}$$

where

$$\Big| - \frac{c}{N^2}x(1 - x) + \frac{1}{N^2}\mathbb{E}_{x,y}[(Z - cx)^2]\Big| \leq \frac{c^2}{N^2}.$$

Moreover we have

$$\big|\mathbb{E}_{x,y}[(Y_1^M - y)^2]\big| = \big|\frac{1}{M^2}\mathbb{E}_{x,y}[(V - Z)^2\big| \leq \frac{c^2}{M^2}.$$

Using Equation (A.2.1) we get

$$\begin{aligned}\big|\mathbb{E}_{x,y}[(X_1^N - x)(Y_1^M - y)]\big| \leq& \big|\frac{1}{NM}\mathbb{E}_{x,y}[U - xN + cx]\mathbb{E}_{x,y}[V - Z]\big| \\ &+ \big|\frac{1}{NM}\mathbb{E}_{x,y}[(Z - cx)(V - Z)]\big| \\ \leq& \frac{c^2}{NM}.\end{aligned}$$

We are thus left with the task of bounding the remainder term in the Taylor expansion. Since $f \in \mathcal{C}^3([0,1]^2)$, we can define

$$\tilde{C}^f := \max\{\frac{\partial^3 f}{\partial x^\alpha \partial y^\beta}(\bar{x}, \bar{y}) \mid \alpha, \beta \in \mathbb{N}_0, \alpha + \beta = 3, \bar{x}, \bar{y} \in [0,1]\}$$

which yields a uniform estimate for the remainder in the form of

$$|R^{\alpha,\beta}(\bar{x},\bar{y})| \leq \frac{1}{\alpha!\beta!\bar{C}^f}$$

which in turn allows us to estimate

$$\begin{aligned}\Big|\mathbb{E}_{x,y}\Big[\sum_{\substack{\alpha,\beta\in\mathbb{N}_0\\ \alpha+\beta=3}} R^{\alpha,\beta}(X_1^N,Y_1^N)(X_1^N-x)^\alpha(Y_1^M-y)^\beta\Big]\Big| \\ \leq \frac{1}{\alpha!\beta!\bar{C}^f}\sum_{\substack{\alpha,\beta\in\mathbb{N}_0\\ \alpha+\beta=3}} \mathbb{E}_{x,y}\big[|(X_1^N-x)^\alpha(Y_1^M-y)^\beta|\big].\end{aligned}$$

Thus the claim follows if we show that the third moments are all of small enough order in N and M. Observe that for $\alpha \in \{0,1,2\}$ we have

$$\mathbb{E}_{x,y}[|(U-Nx+cx)|^\alpha] \leq N. \tag{A.2.2}$$

For $\alpha = 0$ this is trivially true, for $\alpha = 1$ it is due to the fact that the binomial random variable U is supported on $0,...,N-c$ and $Nx-cx$ is its expectation, and for $\alpha = 2$ it follows from the fact that $(U-Nx+cx)^2 = |(U-Nx+cx)^2|$ and the formula for the variance of a binomial random variable. For $\alpha = 3$ it follows e.g. from Lemma 3.1 in [32] that

$$\mathbb{E}_{x,y}[|(U-Nx+cx)|^3] = O(N^{3/2}). \tag{A.2.3}$$

Thus we get for any $0 \leq \alpha,\beta \leq 3$ such that $\alpha+\beta=3$ that

$$\begin{aligned}&\mathbb{E}_{x,y}\big[|(X_1-x)^\alpha(Y_1^M-y)^\beta|\big] \\ &\quad= \frac{1}{N^\alpha M^\beta}\sum_{i=0}^{\alpha}\binom{\alpha}{i}\mathbb{E}_{x,y}[|(U-Nx+cx)^i(Z-cx)^{\alpha-i}(V-Z)^\beta|] \\ &\quad\leq \frac{1}{N^\alpha M^\beta}\sum_{i=0}^{\alpha}\binom{\alpha}{i}\mathbb{E}_{x,y}[|(U-Nx+cx)^i|]\mathbb{E}_{x,y}[|(Z-cx)^{\alpha-i}(V-Z)^\beta|] \\ &\quad\leq \frac{1}{N^\alpha M^\beta}\sum_{i=0}^{\alpha}\binom{\alpha}{i}N(2c)^{\alpha-i+\beta}1_{\{1,2,3\}}(\alpha)+\frac{3(2c)^3}{N^{3/2}}1_{\{3\}}(\alpha) \\ &\quad\leq C\big(\frac{1}{NM}+\frac{1}{M^2}+\frac{1}{N^{3/2}}\frac{N}{M^3}\big),\end{aligned}$$

from (A.2.1), (A.2.2) and (A.2.3), where the constant C depends only on c. This completes the proof. □

A.3 Basics on Yule processes

A.3.1 Basics on Yule processes and proof of Theorem 4.2.5

Definition A.3.1 (Yule process)**.** *A Yule process with rate r is a continuous-time Markov process taking values in $\mathbb{N}$ such that the transition rates are given by:*

$$\begin{cases} n \to n+1 & \text{at rate} \quad rn \\ n \to \text{others} & \text{at rate} \quad 0. \end{cases}$$

Remark A.3.2. Consider a population model starting with n_0 individuals, where each individual reproduces independently at rate r by splitting into two individuals. Then counting the total number of individuals, one gets a Yule process. This is the population model which we consider in the Lenski experiment during one day, with starting population size $n_0 = N$.

Lemma A.3.3. *Let Z^r be a Yule process with rate r and $Z_0^r = 1$. Then, for $t > 0$, $Z^r(t)$ follows a geometric distribution with parameter e^{-rt}.*

Proof. Let $(E_i)_{i\geq 1}$ be independent exponential random variables with parameters i. Then it follows immediately from Definition A.3.1 that

$$\mathbb{P}(Z^r(t) > k) = \mathbb{P}(\sum_{i=1}^{k} E_i < rt). \tag{A.3.1}$$

Let $\{(W_t^{(i)})\}_{1\leq i\leq k}$ be k i.i.d unit Poisson processes. For each $i = 1, \ldots, k$, let T_i be the first jumping time of (W_i). Then we see that

$$\sup_{1\leq i\leq k}\{T_i\} \overset{(d)}{=} \sum_{i=1}^{k} E_i.$$

Consequently, (A.3.1) equals $(1 - e^{-rt})^k$. □

Corollary A.3.4. *If $Z^r(0) = n_0 \in \mathbb{N}$, then $Z^r(t)$ follows a negative binomial distribution with parameters n_0 and e^{-rt}. In particular,*

$$\mathbb{E}[Z^r(t)] = n_0 e^{rt}, \quad \text{and} \quad \operatorname{var}(Z^r(t)) = e^{rt}(e^{rt} - 1)n_0.$$

Proof. This just follows from the fact that the individuals reproduce independently and the fact that the negative binomial distribution is obtained by summing independent geometric random variables. □

The next lemma shows that ς_N is asymptotically equal to σ.

Lemma A.3.5. *Let ς_N and $\sigma = \sigma_0$ be as defined in (4.2.1) and (4.2.6). Then*

$$\varsigma_N \overset{(d)}{\to} \sigma.$$

Proof. During one day in the Lenski experiment, consider the population consisting of N subpopulations each of whose sizes follows an independent Yule process with parameter r. Let $Z_N^r(t)$ denote the size of total population at time t. Then $Z_N^r(t)$ is the sum of N i.i.d geometric variables with parameter e^{-rt}. Let $\varepsilon > 0$. Then due to the law of large numbers

$$\mathbb{P}\Big(\frac{Z_N^r(\sigma-\varepsilon)}{\gamma N} < 1\Big) \overset{N\to\infty}{\to} 1; \;\; \mathbb{P}\Big(\frac{Z_N^r(\sigma+\varepsilon)}{\gamma N} > 1\Big) \overset{N\to\infty}{\to} 1.$$

Therefore $\mathbb{P}(\sigma - \varepsilon \leq \varsigma_N \leq \sigma + \varepsilon) \overset{N\to\infty}{\to} 1$. Since ε can be arbitrarily small, the lemma follows. □

Proof of Theorem 4.2.5. This is a direct application of Theorem 2.1 in [51]. Fix a generation in the Cannings model and let c_N be the probability for a pair of individuals to be coalesced in the previous generation and d_N the probability for a triple of individuals to be coalesced in the previous generation. Then it suffices to prove that

$$c_N \overset{N\to\infty}{\to} 0, \;\; d_N/c_N \overset{N\to\infty}{\to} 0. \tag{A.3.2}$$

Notice that c_N, d_N do not depend on the generation since the reproduction, sampling and labeling in each day do not depend on the past and on the future. Therefore we can consider a typical day (the population at the beginning of a day constitutes a generation) and take the notations at the beginning of Section 2.1.1. Let Y_t^i be the size of the family of individual i at time t. Then

$$Z_t^N = Y_t^1 + Y_t^2 + \cdots Y_t^N,$$

with $(Y_t^i)_{1\leq i\leq N}$ identically and independently distributed as a geometric distribution with parameter e^{-rt}. The day ends at time $\sigma = \frac{\log\gamma}{r}$ and notice that the population for the next day will be chosen uniformly, hence one can express c_N, d_N as follows:

$$c_N = \mathbb{E}\Big[\frac{\sum_{i=1}^N \binom{Y_\sigma^i}{2}}{\binom{Z_\sigma^N}{2}}\Big] \sim \frac{2(1-\frac{1}{\gamma})}{N}, \;\; d_N = \mathbb{E}\Big[\frac{\sum_{i=1}^N \binom{Y_\sigma^i}{3}}{\binom{Z_\sigma^N}{3}}\Big] = O(N^{-2}),$$

which gives (A.3.2), and thus completes the proof. □

A.3.2 Properties of near-critical Galton Watson processes

The following lemma (Theorem 3 of [2], and see also Theorem 5.5 in [26] under weaker conditions) provides the survival probability for certain near-critical Galton Watson trees.

Lemma A.3.6. *Consider a sequence of supercritical Galton Watson processes $(G_i^N)_{i\in\mathbb{N}_0}$, $N = 1, 2, \ldots$, with offspring mean $1+\beta_N$ (with $\beta_N \to 0$) and offspring variance $\sigma^2 + v_N$ (with $v_N \to 0$) and uniformly bounded third moment, starting from one ancestor in generation 0. Then the survival probability ϕ_N obeys $\phi_N \sim \frac{2\beta_N}{\sigma^2}$.*

Lemma A.3.7. *Let $(G_i^N)_{i\in\mathbb{N}_0}, N = 1, 2, \ldots$ be as in Lemma A.3.6. Assume that $\beta_N N \to \infty$ as $N \to \infty$. Then, for every $\varepsilon > 0$, $\mathbb{P}(\exists i : G_i^N \geq \varepsilon N) \sim \mathbb{P}(\lim_{i\to\infty} G_i^N = \infty)$.*

Proof. Again let ϕ_N be the survival probability of G^N started in one individual. Then

$$\mathbb{P}(\lim_{i\to\infty} G_i = \infty | \exists i : G_i \geq \varepsilon N) \geq 1 - (1-\phi_N)^{\varepsilon N} \sim 1 - (1 - \frac{2\beta_N}{\sigma^2})^{\varepsilon N} \to 1,\ N \to \infty.$$

□

Lemma A.3.8. *Let $(G_i^N)_{i\in\mathbb{N}_0}, N = 1, 2, \ldots$ be as in Lemma A.3.6. Assume that $\beta_N \sim cN^{-b}, N = 1, 2, \ldots$, for some $c > 0$ and $b \in (0, 1)$. For fixed $\varepsilon \in (0, 1)$, let $\omega_N := \inf\{i \geq 0 : G_i^N \geq \varepsilon N\}$. Then we have for any $\delta > 0$*

$$\lim_{N\to\infty} \mathbb{P}_1(\omega_N > \beta_N^{-1-\delta} \,|\, \omega_N < \infty) = 0.$$

Further, let $\upsilon_N := \inf\{i \geq 0 : G_i^N = 0\}$. Then for any $\delta > 0$, for N large enough,

$$\mathbb{P}_1(\upsilon_N > \beta_N^{-1-\delta} \,|\, \upsilon_N < \infty) \leq e^{-N^{b\delta}}. \tag{A.3.3}$$

Proof. First we consider the difference between conditioning G^N on survival (forever) and on reaching εN, respectively. Since we know (from Lemma B1) that

$$\mathbb{P}_1(G^N \text{ survives}) \sim \frac{2\beta_N}{\sigma^2} \sim c' N^{-b}, \tag{A.3.4}$$

we can infer, using the strong Markov property, that

$$\mathbb{P}_1(G^N \text{ reaches } \varepsilon N \text{ and } G^N \text{ does not survive }) \leq \mathbb{P}_{\lfloor \varepsilon N \rfloor}(G^N \text{ does not survive })$$

$$= (1-\phi_N)^{\lfloor \varepsilon N \rfloor} \leq \left(1 - c' N^{-b}\right)^{\lfloor \varepsilon N \rfloor} \leq \exp(-c(\varepsilon) N^{1-b}). \tag{A.3.5}$$

Thus we can estimate

$$\mathbb{P}_1(\omega_N > \beta_N^{-1-\delta} | G^N \text{ reaches } \varepsilon N) = \frac{1}{\mathbb{P}_1(G_N \text{ reaches } \varepsilon N)} \mathbb{P}_1(\omega_N > \beta_N^{-1-\delta}, G^N \text{ reaches } \varepsilon N)$$

$$\leq \frac{1}{\mathbb{P}_1(G^N \text{ reaches } \varepsilon N)} \mathbb{P}_1(G^N \text{ reaches } \varepsilon N \text{ and does not survive})$$

$$+ \frac{1}{\mathbb{P}_1(G^N \text{ survives })} \mathbb{P}_1(\omega_N > \beta_N^{-1-\delta}, G^N \text{ survives}).$$

The first summand on the r.h.s tends to 0 as $N \to \infty$ because of (A.3.4) and (A.3.5). Thus, for proving the lemma it suffices to show that

$$\lim_{N\to\infty} \mathbb{P}_1(\omega_N > \beta_N^{-1-\delta} | G^N \text{ survives }) = 0. \tag{A.3.6}$$

Let ϕ_N be the survival probability of G^N, and denote by H_i^N, $i = 0, 1, \ldots$, the generation sizes of those individuals that have an infinite line of descent, conditioned on survival of G^N. Then we have (cf. Proposition 5.28 in [46])

$$f^*(s) := \sum_{k\geq 0} s^k \mathbb{P}_1(H_1^N = k) = \mathbb{E}_1[s^{H_1^N}] = \frac{\mathbb{E}[(1-\phi_N+\phi_N s)^{G_1^N}] - (1-\phi_N)}{\phi_N}, s \geq 0.$$

Obviously, $\mathbb{P}_1(H_1^N = 0) = f^*(0) = 0$ and $\mathbb{P}_1(H_1^N = 1) = (f^*)'(0) = \mathbb{E}[G_1^N(1-\phi_N)^{G_1^N-1}]$, which, using Taylor expansion, is transformed to

$$\begin{aligned}\mathbb{E}[G_1^N\Big(1-(G_1^N-1)\phi_N+\frac{(G_1^N-1)(G_1^N-2)\phi_N^2}{2}(1-t\phi_N)^{G_1^N-3}\Big)\\ =\mathbb{E}_1[G_1^N(1-(G_1^N-1)\phi_N)]+O(\phi_N^2)=1-\beta_N+o(\beta_N),\end{aligned} \tag{A.3.7}$$

where $t = t(G_1^N) \in (0,1)$. The first equality is due to the assumption in Lemma A.3.6 that the third order moment of G_1^N is uniformly bounded. We can thus infer that, for any fixed $\eta \in (0,1)$,

$$\mathbb{P}_1(H_1^N \geq 2) \geq \eta\beta_N, \text{when } N \text{ is large enough.}$$

We can now give a lower bound for G_i^N, conditioned on survival of G^N, in two steps: first by H_i^N, and then by a (discrete time) Galton Watson process with offspring distribution $(1-\eta\beta_N)\delta_1 + \eta\beta_N\delta_2$. Call this process B^N. With $\frac{1}{\lfloor \eta\beta_N \rfloor}$ generations as a new time unit, the sequence of processes B^N converges, as $N \to \infty$, to a standard Yule process. This means that, for every fixed $t > 0$, at a time of $\lfloor t\eta\beta_N \rfloor^{-1}$ generations, B^N has an approximate geometric distribution with parameter e^{-t}. Thus we conclude after $\lfloor \beta_N \rfloor^{-1-\delta}$ generations, B^N (and a fortiori also G^N when conditioned to survival) is larger than εN with probability tending to 1 as $N \to \infty$. This shows (A.3.6), and concludes the proof of the first statement. For the last statement, observe that by Theorem 5.28 of [46] the distribution of (G_i^N) conditioned on extinction is equal to the distribution of a Galton Watson process with probability generating function

$$\overline{f}(s) := (1-\phi_N)^{-1}\sum_{k\geq 0}((1-\phi_N)s)^k\mathbb{P}_1(G_1^N = k).$$

Thus we have

$$\mathbb{E}_1[G_1^N | G^N \text{ dies out}] = \overline{f}'(1) = \mathbb{E}[G_1^N(1-\phi_N)^{G_1^N-1}] = 1-\beta_N+o(\beta_N),$$

where the last equality follows from equation (A.3.7). Then, by Proposition 5.2 in [46] we observe that

$$\mathbb{E}_1[G^N_{\lceil \beta_N^{-1-\delta} \rceil} | G^N \text{ dies out}] = (1-\beta_N+o(\beta_N))^{\beta_N^{-1-\delta}} \leq e^{-N^{b\delta}} \tag{A.3.8}$$

so we conclude

$$\mathbb{P}_1(v_N > \beta_N^{-1-\delta} | v_N < \infty) = \mathbb{P}_1(G^N_{\lceil \beta_N^{-1-\delta} \rceil} > 0 | G^N \text{ dies out}) \leq \mathbb{E}_1[G^N_{\lceil \beta_N^{-1-\delta} \rceil} | G^N \text{ dies out}] \leq e^{-N^{b\delta}}.$$

□

Bibliography

[1] Aleshkyavichene, A. K. On the probabilities of large deviations for the maximum of sums of independent random variables. *Theory Probab. App. 24*, 1 (1979), 16–33.

[2] Athreya, K. B. Rates of decay for the survival probability of a mutant gene. *J. Math. Biol. 30* (1992), 577–581.

[3] Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., Lenski, R. E., and Kim, J. F. Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature 461*, 7268 (2009), 1243–1247.

[4] Berestycki, N. Recent progress in coalescent theory. In *Ensaios Matematicos*, vol. 16. 2009.

[5] Blath, J., Eldon, B., González Casanova, A., Kurt, N., and Wilke-Berenguer, M. Genetic variability under the seed bank coalescent. *Genetics*, 200(3) (2015), 921–34.

[6] Blath, J., González Casanova, A., Eldon, B., and Kurt, N. Genealogy of a wright fisher model with strong seedbank component. *Birkhauser progress in probability, Special issue of the XI Symposium of Probability and Stochastic Processes* (2015).

[7] Blath, J., González Casanova, A., Kurt, and Spanò, D. The ancestral process of long-range seed bank models. *J. Appl. Probab. 50*, 3 (09 2013), 741–759.

[8] Blath, J., González Casanova, A., Kurt, N., and Wilke-Berenguer, M. A new coalescent for seedbank models. *Accepted for publication: Annals of Applied Probability* (2015).

[9] Cannings, C. The latent roots of certain markov chains arising in genetics: A new approach, 1. haploid models. *Advances in Applied Probability 6*, 2 (1974), pp. 260–290.

[10] Champagnat, N. A microscopic interpretation for adaptive dynamics trait substitution sequence models. *Stochastic Process. Appl. 116* (2006), 1127–1160.

[11] Champagnat, N., Jabin, P., and Méléard, S. Adaptation in a stochastic multi-resources chemostat modell. *J. Math. Pures Appl. (9) 101*, 6 (2014), 755–788.

[12] Couce, A., and Tenaillon, O. The rule of declining adaptability in microbial evolution experiments. *Front. Genet. 6:99* (2015).

[13] Doeblin, W. Expose de la theorie des chaines simples constantes de markov a un nombre fini d etats. *Mathematique de l Union Interbalkanique 2* (1938), 77–105.

[14] Dong, R., Gnedin, A., and Pitman, J. Exchangeable partitions derived from markovian coalescents. *Ann. Appl. Prob. 17* (2007), 1172–1201.

[15] Durrett, R. *Stochastic Calculus: A Practical Introduction*. Probability and Stochastics Series. Taylor & Francis, 1996.

[16] Durrett, R. *Probability Models for DNA Sequence Evolution*. Probability and Its Applications. Springer, 2008.

[17] Etheridge, A. Some mathematical models from population genetics. *Lectures from the 39th Probability Summer School held in Saint-Flour, 2009* (2011).

[18] Ethier, S. N., and Kurtz, T. G. *Markov processes: characterization and convergence*. Wiley series in probability and mathematical statistics. Wiley, New York, 1986.

[19] Feller, W. *An Introduction to Probability Theory and Its Applications*, vol. 1. Wiley, 1968.

[20] Fisher, R. A. *The Genetical Theory of Natural Selection*, 1 ed. Oxford University Press, USA, Apr. 2000.

[21] Gerrish, P. J., and Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica 102-103*, 1-6 (1998), 127–144.

[22] Goldschmidt, C., and Martin, J. Random recursive trees and the bolthausen-sznitman coalescent. *Electron. J. Probab. 10* (2005), 718–745.

[23] González Casanova, A., Aguirre-von Wobeser, E., Espín, G., Servín-González, L., Kurt, N., Spanò, D., Blath, J., and Soberón-Chávez, G. Strong seedbank effects in bacterial evolution. *J. Theor. Biol.*, 356 (2014), 62–70.

[24] González Casanova, A., Kurt, N., Wakolbinger, A., and Yuan, L. A basic mathematical model for the lenski experiment, and the deceleration of the relative fitness. *http://arxiv.org/abs/1505.01751* (2015).

[25] Good, B. H., and Desai, M. M. The impact of macroscopic epistasis on long-term evolutionary dynamics. *Genetics 199*, 1 (2015), 177–190.

[26] Haccou, P., Jagers, P., and Vatutin, V. A. *Branching processes: variation, growth, and extinction of populations*, vol. 5. Cambridge University Press, 2005.

[27] Hammond, A., and Sheffield, S. Power law pólya's urn and fractional brownian motion. *Probability Theory and Related Fields 157*, 3-4 (2013), 691–719.

[28] Herbots, H. M. *Stochastic models in population genetics: genealogical and genetic differentiation in structured populations. PhD Dissertations*. University of London., 1994.

[29] Herbots, H. M. The structured coalescent. *Progress in Population Genetics and Human Evolution IMA Vol. Math. Appl. 87* (1997), 231–255.

[30] Ikeda, N., and Watanabe, S. *Stochastic differential equations and diffusion processes*. North-Holland mathematical library. North-Holland Pub. Co. Tokyo, Amsterdam, New York, 1989.

[31] Jansen, S., and Kurt, N. On the notion(s) of duality for markov processes. *Probab. Surveys 11* (2014), 59–120.

[32] Jenkins, P. A., Fearnhead, P., and Song, Y. S. Tractable stochastic models of evolution for loosely linked loci. *ArXiv:1405.6863* (2014).

[33] Kaj, I., Krone, S. M., and Lascoux, M. Coalescent theory for seed bank models. *J. Appl. Probab. 38*, 2 (06 2001), 285–300.

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

[35] Kurtz, T. G. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *J. Appl. Probab. 8*, 2 (1971), 344–356.

[36] Lambert, A. Probability of fixation under weak selection: A branching process unifying approach. *Theor. Popul. Biol. 69*, 4 (2006), 419–441.

[37] Lambert, A., and Ma, C. The peripatric coalescent. *J. Appl. Prob. (in press)* (2015).

[38] Lennon, J. T., and Jones, S. E. Microbial seedbanks: the ecological and evolutionary implications of dormancy. *Nature Reviews Microbiology 9* (2011), 119–130.

[39] Lenski, R. E. The e. coli long-term experimental evolution project site, `http://myxo.css.msu.edu/ecoli`., 2015.

[40] Lenski, R. E., Rose, M. R., Simpson, S. C., and Tadler, S. Long-term experimental evolution in escherichia coli. i. adaptation and divergence during 2,000 generations. *Am. Nat. 138*, 6 (1991), 1315–1341.

[41] Lenski, R. E., and Travisano, M. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc. Natl. Acad. Sci. USA 91*, 15 (1994), 6808–6814.

[42] Levin, D. A. The seedbank as a source of genetic novelty in plants. *American Naturalist 135* (1990), 563–572.

[43] Levin, D. A., Peres, Y., and Wilmer, E. L. *Markov chains and mixing times*. Providence, R.I. American Mathematical Society, 2009. With a chapter on coupling from the past by James G. Propp and David B. Wilson.

[44] Lindvall, T. On coupling of renewal processes with use of failure rates. *Stochastic Processes and their Applications 22*, 1 (1986), 1 – 15.

[45] Lindvall, T. *Lectures on the coupling method*. Wiley series in probability and mathematical statistics. Wiley, New York, 1992. A Wiley-Interscience publication.

[46] Lyons, R., and Peres, Y. Probability on Trees and Networks. In preparation: Cambridge University Press available at `http://mypage.iu.edu/~rdlyons/`, 2014.

[47] M., K. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *J. Appl. Prob. 38* (1975), 285–300.

[48] Maddamsetti, R., Lenski, R. E., and Barrick, J. E. Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with escherichia coli. *biorxiv.org,* *http://biorxiv.org/content/early/2015/03/25/017020* (2015).

[49] McCandlish, D. M., Otwinowski, J., and Plotkin, J. B. The diversity of evolutionary dynamics on epistatic versus non-epistatic fitness landscapes. *arXiv:1410.2508v3 [q-bio.PE],* *http://arxiv.org/abs/1410.2508* (2015).

[50] Möhle, M. The concept of duality and applications to markov processes arising in neutral population genetics models. *Bernoulli 5*, 5 (10 1999), 761–777.

[51] Möhle, M., and Sagitov, S. A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab. 29*, 4 (10 2001), 1547–1562.

[52] Nath, H., and Griffiths, R. The coalescent in two colonies with symmetric migration. *J. Math. Biol. 31* (1993), 841–852.

[53] Neuhauser, C., and Krone, S. The genealogy of samples in models with selection. *Genetics 145* (1997), 519–534.

[54] Notohara, M. The coalescent and the genealogical process in geographically structured population. *J. Math. Biol. 29* (1990), 59–75.

[55] Nunney, L. The effective size of annual plant populations: the interaction of a seedbank with fluctuating population size in maintaining genetic variation. *American Naturalist 160* (2002), 195–204.

[56] Parsons, T. *Asymptotic Analysis of Some Stochastic Models from Population Dynamics and Population Genetics, (PhD thesis)*. University of Toronto, available at `http://www.math.toronto.edu/%7Eparsons/pdf/ToddParsons_Thesis.pdf`, 2012.

[57] Parsons, T. L., Quince, C., and Plotkin, J. B. Absorption and fixation times for neutral and quasi-neutral populations with density dependence. *Theor. Popul. Biol. 74*, 4 (2008), 302–310.

[58] Petrov, V. V., and Robinson, J. Large deviations for sums of independent non identically distributed random variables. *Comm. Statist. Theory Methods 37*, 18 (2008), 2984–2990.

[59] Pitman, J. Coalescents with multiple collisions. *Ann. Probab. 27*, 4 (10 1999), 1870–1902.

[60] Postlethwalt, J. Modern biology. *Holt, Rinehart and Winston. 2* (2009), 317.

[61] Prokhorov, Y. Convergence of random processes and limit theorems in probability theory. *Theory of Prob. And Appl. 2* (1956), 157–214.

[62] Rogers, L. C. G., and Williams, D. *Diffusions, Markov processes, and martingales. Vol. 2*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000. Itô calculus, Reprint of the second (1994) edition.

[63] Schweinsberg, J. A necessary and sufficient condition for the λ-coalescent to come down from infinity. *Electron. Comm. Probab. 5* (2000), 1–11.

[64] Schweinsberg, J. Coalescent processes obtained from supercritical galton-watson processes. *Stochastic Process. Appl. 106*, 1 (2003), 107–139.

[65] Skorokhod, A. Limit theorems for stochastic processes. *Th. Probab. Appl 3* (1956), 261–290.

[66] Spitzer, F. *Principles of Random Walk*. Springer, 2000.

[67] Takahata, N. The coalescent in two partially isolated diffusion populations. *Genetical Research 52* (1988), 213–222.

[68] Templeton, A. R., and Levin, D. A. Evolutionary consequences of seed pools. *American Naturalist 114* (1979), 232–249.

[69] Vitalis, R., Glémin, S., and Oliviere, I. When genes got to sleep: The population genetic consequences of seed dormancy and monocarpic perenniality. *American Naturalist 163* (2004), 295–311.

[70] Živković, D., and Tellier, A. Germ banks affect the inference of past demographic events. *Molecular Ecology 21* (2012), 5434–5446.

[71] Wakeley, J. *Coalescent Theory: An Introduction*, 1st edition ed. Roberts and Company Publishers, 2008.

[72] Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theoretical population biology 7*, 2 (Apr. 1975), 256–276.

[73] Whitt, W. *Stochastic-Process Limits : An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, 2002.

[74] Wiser, M. J., Ribeck, N., and Lenski, R. E. Long-Term Dynamics of Adaptation in Asexual Populations. *Science 342*, 6164 (2013), 1364–1367.

[75] Wright, S. Evolution in mendelian populations. *University of Chicago, Chicago, Illinois* (1931).

[76] Yamada, T., and Watanabe, S. On the uniqueness of solutions of stochastic differential equations. *J. Math. Kyoto Univ. 11*, 1 (1971), 155–167.