Properties of the Wright-Fisher diffusion with seed banks and multiple islands

vorgelegt von
M. Sc.
Eugenio Buzzoni

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:
Vorsitzende: Prof. Dr. Barbara Zwicknagl
Gutachter: Prof. Dr. Jochen Blath
Gutachter: Prof. Dr. Dario Spanò

Tag der wissenschaftlichen Aussprache: 19. September 2019

Berlin, 2019

Abstract

The main purpose of this thesis is the analysis under several viewpoints both of the *Wright-Fisher diffusion with seed bank*, introduced in [BGKWB16], and the *two-island diffusion*, investigated e.g. in [KZH08] and [NG93]. The former simulates a population in which some of the individuals can become inactive for long periods of time, like seeds or dormant bacteria, while the latter is used to investigate the behavior of a split population (e.g. geographically).

The main body of the thesis is composed of three parts. In the first one (Chapters 3 and 4), we make a comparison between the Wright-Fisher diffusion with seed bank and the two-island diffusion from several viewpoints, including stationary distribution, mixed moments and reversibility. In particular, we define the (strong) seed bank coalescent and the structured coalescent respectively as the moment dual processes to the Wright-Fisher diffusion with seed bank and the two-island diffusion. The main result of the first part regards boundary behavior. In fact, we provide a complete boundary classification of both processes, which is, as far as we know, a new result. The proof involves one of two martingale-based reasonings, that is, McKean's or Lyapunov's argument.

In the second part of the thesis (Chapter 5), we tackle the issue of scaling limits. The main result concerns the case of the seed bank diffusion under the additional assumption that reproduction occurs on a faster time-scale than both dormancy and resuscitation: if we speed up time in an appropriate way, we get a previously unknown scaling limit, describing the genealogy under the aforementioned regime, that we call the *ancient ancestral lines process*. This object is dual to a jump diffusion, and we make heavy use of duality to establish a remarkable convergence of the rescaled diffusion processes to the jump diffusion limit.

In the last part of this work (Chapter 6), we analyze several classical measures of population structure to distinguish the patterns of genetic variability produced by our models, with a focus on coalescent processes. In this case we are concerned not only with the seed bank and the structured coalescent, but also with the standard Kingman and with the so-called weak seed bank coalescent. Our main goal is to compare, with respect to the neutral Kingman case, the different ways in which our measures react to the presence of a seed bank and to the presence of population structure.

For this purpose, we first focus on the two-allele case, where we can derive exact likelihoods for the full sample probabilities by means of recursive formulas.

Then, we briefly introduce sample heterozygosity, Wright's F_{ST} and the expected site frequency spectrum (SFS), classical measures for population structure that can be easily computed under the aforementioned population models in the infinite sites case. Our main tool in this chapter is phase-type distribution theory in general and the formulae recently introduced by Hobolth et al. ([HSB19]) in particular.

2010 Mathematics Subject Classification: Primary, 92D10, secondary, 60K35.

Keywords: Wright-Fisher diffusion, seed bank, dormancy, two island model, coalescent process, boundary classification, scaling limits, duality, model selection.

Zusammenfassung der Arbeit

In der vorliegenden Dissertation werden verschiedene Aspekte der *Wright-Fisher-Diffusion mit Seed-Bank* und der *Two-Island-Diffusion* untersucht. Hierbei handelt es sich um zwei Modelle aus der mathematischen Populationsdynamik. Die Wright-Fisher-Diffusion mit Seed-Bank, eingeführt in der Arbeit von Blath, Gonzàlez Casanova, Kurt und Wilke-Berenguer ([BGKWB16]), beschreibt die Evolution einer Bevölkerung, in der einzelne Individuen langfristig inaktiv werden können. Klassische Beispiele hiervon sind samentragende Pflanzen oder Bakterien. Die Two-Island-Diffusion (siehe z.B. [KZH08] und [NG93]) modelliert hingegen die Genealogie einer räumlich geteilten Population.

Der Hauptteil dieser Dissertation besteht aus drei Teilen. Im ersten Teil (Kap. 3 und 4) wird die Wright-Fisher-Diffusion mit Seed-Bank mit der Two-Island-Diffusion in mehrerer Hinsicht verglichen, einschließlich stationärer Verteilungen, gemischter Momente und Reversibilität. Insbesondere wird der (starke) Seed-Bank-Koaleszent und der strukturierte Koaleszent als Momentenduale, respektiv der Wright-Fisher-Diffusion mit Seed-Bank und der Two-Island-Diffusion, definiert. Im wichtigsten Satz dieses ersten Teiles geht es um das Verhalten an den Rändern, d. h. um die Frage, unter welchen Bedingungen eine temporäre Extinktion eines Alleles vorliegt. Wir geben nämlich eine vollständige Klassifizierung der Ränder für diese beiden Diffusionsprozesse an. Der Beweis dafür benutzt ein von zwei Martingalargumenten, die das McKean-Argument und das Lyapunov-Argument genannt werden.

Der zweite Teil der Dissertation (Kapitel 5) setzt sich mit Skalierungslimiten auseinander. Das wichtigste Ergebnis betrifft den Fall einer Seed-Bank-Diffusion, in der die Fortpflanzung auf einer schnelleren Zeitskala geschieht als sowohl Dormanz als auch Wiederbelebung. Falls wir in dem Fall auf eine beschleunigte Zeitskala übergehen, bekommen wir im Limes den sogenannten *Ancient Ancestral Lines Process*, der die Genealogie unter der gegebenen Skalierung beschreibt. Dieses Objekt ist dual zu einer Diffusion mit Sprüngen, und wir machen regen Gebrauch von Dualitätsargumenten, um eine beachtenswerte Koaleszenz der reeskalierten Diffusionsprozesse gegen dem Diffusionslimes mit Sprüngen aufzubauen.

Im letzten Teil dieser Arbeit (Kapitel 6) werden mehrere klassische Populationsstrukturmaße untersucht, wie Sample Heterozygosity, Wrights F_{ST} sowie das erwartete Allelen-Frequenzspektrum (SFS), allesamt klassische Indikatoren, um die Struktur einer Population zu messen. Unser Ziel ist es dabei, die Muster genetischer Variabilität, die durch die besagten Modelle erzeugt werden, voneinander zu unterscheiden. Wir konzentrieren uns dabei hauptsächlich auf die Koaleszentenprozesse. Neben dem Seed-Bank- und dem strukturierten Koaleszenten beschäftigen wir uns mit dem klassischen Kingman- sowie mit dem sogenannten schwachen Seed-Bank-Koaleszenten. Unser Hauptziel ist hier, den Einfluss dieser Populationsstrukturmaße auf eine Seed-Bank bzw. eine Population mit räumlicher Struktur zu untersuchen und mit dem neutralen Kingman-Fall zu vergleichen.

Für diesen Zweck konzentrieren wir uns zuerst auf Populationsmodelle mit einer endlichen Anzahl von Allelen, da in diesem Fall mit Hilfe von Rekursionsformeln exakte Likelihoods für die Stichproben-Wahrscheinlichkeiten angegeben werden können. Die Populationsstrukturmaße kann man hingegen auch im Falle unendlich vieler Allele für die gegebenen Populationsmodelle berechnen. Unser wichtigstes Werkzeug besteht dabei aus Phase-Type-Verteilungen und insbesondere aus den Formeln, die im Artikel von Hobolth et al. ([HSB19]) gegeben sind.

Acknowledgements and Co-Authorship

First of all, I would like to thank my supervisor, Jochen Blath, for his patient guidance, support and encouragement, together with everyone else I worked with during my PhD, especially Adrián González Casanova, Jere Koskela, Marcel Ortgiese, Martin Slowik, Maite Wilke Berenguer and Linglong Yuan. A special thanks to Dario Spanò and Barbara Zwicknagl for agreeing to be in the committee for my defense.

Many thanks to Jochen Blath, Matthias Hammer and Martin Slowik for their useful feedback on my thesis.

Moreover, thanks to all the people at TU Berlin for creating a pleasant and stimulating atmosphere in which it was a pleasure to work. Thanks to all my students, whom I hope to have infected with my passion for probability theory. Also thanks to Ernesto De Vito for all his advice.

And last but not least, a big thanks to all of my friends and my family for helping me during this hard work.

This thesis is based on the following papers and joint work (denoted by “key references" from now on):

[K1] **Structural properties of the seed bank and the two-island diffusion.** Journal of Mathematical Biology, (2019) 79: 369-392. DOI: 10.1007/s00285-019-01360-5. Joint work with: J. Blath, A. González Casanova and M. Wilke Berenguer. Chapters 1.2.1, 3, 4.1-4.3.

[K2] **Separation of time-scales for the seed bank diffusion and its jump-diffusion limit**. arXiv preprint, arXiv:1903.11795. Joint work with: J. Blath, A. González Casanova and M. Wilke Berenguer. Chapter 1.1, 5.

[K3] **Statistical tools for seed bank detection**. arXiv preprint, arXiv:1907.13549. Joint work with: J. Blath, J. Koskela and M. Wilke Berenguer. Chapter 1.2.2, 1.3, 6.2-6.8.

Contents

Glossary

t	Continuous time unit	*p*.10
$(X(t),Y(t))_{t\geq 0}$	Our main diffusion process	*p*.10
u_1,u_2	Mutation rates in island 1	*p*.10
u_1',u_2'	Mutation rates in island 2	*p*.9
$(W(t),W'(t))_{t\geq 0}$	Standard 2-dimensional Wiener process	*p*.10
(x,y)	Starting points	*p*.10
(N,M)	Total n. of active and dormant individuals (when finite)	*p*.10
K	Relative seed bank size	*p*.10
c	Migration rate	*p*.10
(α,α')	Coalescent-relative population sizes	*p*.11
c'	Usually, short for K · c	*p*.11
$\mathcal{A}^{(1)}$	Infinitesimal generator of (X,Y)	*p*.11
β	Key parameter in the weak seed bank model	*p*.14
$(n^{(1)},n^{(2)})$	Size of a sample ($n^{(1)}$ active, $n^{(2)}$ dormant individuals)	*p*.14
n	Size of a sample; when applicable, $n=\sum_i n^{(i)}$	*p*.14
$(N(t),M(t))_{t\geq 0}$	Moment dual of the seed bank diffusion	*p*.154
$\bar{N}_{ij}$	Transition rates for the dual process	*p*.15
u	Mutation parameter in infinitely many alleles model	*p*.17
$\mathbb{P}_x$	Probability measure	*p*.19
$S(\cdot,\cdot)$	Duality function	*p*.19
$(\sigma(t),b(t))_{t\geq 0}$	Diffusion and drift terms of a diffusion process	*p*.21
$(a(t))_{t\geq 0}$	$(\sigma(t)\sigma(t)^T)_{t\geq 0}$	*p*.29
$\mathcal{P}$	A set of polynomials	*p*.29
r	Reward function	*p*.31
(∂,∂)	Death (=absorbing) state	*p*.32
$\mathcal{A}^{(2)}$	Infinitesimal generator of $(N(t),M(t))_{t\geq 0}$	*p*.32
$M_{\bar{n}}$	$\bar{n}$-th mixed moment of (X,Y), i.e. $\mathbb{E}[X^{n_1}Y^{n_2}]$	*p*.36
a_j,a_j',D_{ij}	Constants from the recursive formula	*p*.36
μ	Stationary distribution of (X,Y)	*p*.37
P,G	Projection- and Q-matrix for a nonstandard semigroup	*p*.59
$(\tilde{N}(t),\tilde{M}(t))_{t\geq 0}$	Backwards-in-time ancient ancestral lines process	*p*.60
$(\tilde{X}(t),\tilde{Y}(t))_{t\geq 0}$	Forwards-in-time ancient ancestral material process	*p*.61
H	Sample heterozygosity	*p*.85
F_{ST}	Wright's measure of population structure	*p*.89
T_i	Waiting time	*p*.92
$F_{ST}^{(0)}$	Slatkin's approximation for the F_{ST}	*p*.95
$(\zeta_1^n,\dots,\zeta_{n-1}^n)$	Site frequency spectrum	*p*.98
T_{MRCA}	Time to most recent common ancestor	*p*.99

1 Introduction and basic models

This chapter is based on [K1] (Subchapter 2.1), [K2] (Subchapter 1) and [K3] (Subchapters 2.2 and 3).

1.1 Motivation

For centuries, mathematics in general and the theory of probability in particular have been used to model and understand real-world events which involve some form of randomness. One of the areas where probability, in the form of stochastic models, was introduced to this aim is population genetics, where the mathematical approach allowed us to gain deep insights into evolutionary mechanisms. In this context, the *Wright-Fisher process* is a widespread probabilistic model in mathematical population genetics. It is defined as a Markov process which describes the relative frequency of two competing alleles in a well-mixed, fixed-size, panmictic population developing in discrete generations. The Wright-Fisher model played a core role in the history of mathematical population genetics: since regarded as a sort of "toy model", it has been taken as the basis for more complex population models. The Wright-Fisher process and generalizations have been thoroughly investigated, starting with the pioneering work of Wright ([Wri31]) and Fisher ([Fis99]). A further important discovery was that of the *Wright-Fisher diffusion* (see [Fel51, Kim55, Kim64]), a continuous-time, continuous-state-space Markov process approximating the neutral allele frequencies in a large haploid population on a macroscopic timescale.

Another milestone was the introduction of the *Kingman coalescent* ([Kin82a, Kin82b]), defined as the partition-valued ancestral process describing the genealogy of a (present-day) sample from the Wright-Fisher diffusion. This retrospective viewpoint, which allows us to see the entire topic with a backwards-in-time approach ("Where did the population come from?"), turned out to be very fruitful.

The introduction of population structure was a further important extension of the modeling frame. Most real-life populations are not panmictic, but geographically structured, as for example North Sea cod species (see [EW08]). This led to the definition of the so-called structured coalescent ([Her94, Not90]) as a way of modeling genetic frequency processes in a variety of cases, all involving some form of population structure.

In fact, in the presence of population structure, e.g. in the guise of the two-island model ([Wri31, Mor59]), many new effects appear. In particular, the genealogy of a sample taken from the subdivided population may be described by a *structured coalescent* instead of the classical Kingman coalescent, in which two lines may merge only at times when both are in the same island. Yet, other qualitative features remain unchanged, including the fact that the structured coalescent with two islands still "comes down from infinity", and that the Wright-Fisher diffusion with two islands (without mutation) will eventually fixate. In this model, there seems to be no explicit characterization of the stationary distribution, though recursion formulas may still be found (see e.g. [NG93, FGH03, KZH08] for results in this direction).

Another interesting feature that was implemented into the Wright-Fisher model was the effect of latency or dormancy ([KKL01]). This means that we allow for the possibility that some individuals of the population are in a latent or dormant state. Seed banks, that is, reservoirs of dormant individuals that can potentially be resuscitated in the future, are common in many communities of macroscopic (e.g. plants) and microscopic (e.g. bacteria) organisms. They increase the persistence of genotypes and are important for the diversity and development of populations. In particular, microbial dormancy is common in a range of ecosystems, and there is evidence that the ecology and evolution of microbial communities are strongly influenced by seed bank dynamics. It has been observed that more that 90% of microbial biomass in soils is metabolically inactive; see [LJ11] and [SL18] for recent overviews on this subject.

In our models, this phenomenon is reflected in the possibility that the direct ancestor of a current-time individual does not come from the previous generation, but can instead be found several generations in the past. The distance is modeled by a random variable with values in the natural numbers, the main cases being a random variable with finite support (weak seed bank, see e.g. [KKL01]) or geometrically distributed of order N (strong seed bank). Such scenarios with seed bank are less well analyzed, and in fact only recently, in [BGKWB16], the *Wright-Fisher diffusion with strong seed bank* and its dual, the *seed bank coalescent*, have been introduced as mathematical objects (see also [LM15], in which the same dual has been obtained as scaling limit of the genealogy in a metapopulation model with peripatric speciation). While at first glance similar to the two-island model and the structured coalescent, the seed bank diffusion and its dual exhibit some remarkable qualitative differences. For example, the seed bank coalescent *does not* come down from infinity, and its expected time to the most recent common ancestor is unbounded as the sample size increases (see [BGKWB16] for details). One of the chief goals of this thesis is to extend the comparison between these two models by looking at them through several lenses, with a specific focus on boundary behavior; another one is to prove other key features of the relatively new seed bank model, in particular regarding scaling limits.

In parallel to the introduction of new and increasingly complex extensions of the Wright-Fisher model, several statistical tools have been introduced, especially markers allowing biologists to see quickly whether it makes sense to use one of our models to predict the behavior of a real-world population whose data have been collected. One of the simplest markers is *sample heterozygosity*, or SH for short, which gives us the probability for two randomly sampled individuals to have different genes. Further key measures of population structure are Wright's F_{ST}, which gives us an index of subpopulation differentiation, and the expected site frequency spectrum (SFS).

While some basic mathematical models have been derived and predict unique patterns of genetic variability in idealized scenarios ([KKL01, LJ11, ZT12, BGE$^+$15, BGKWB16, dHP17]), statistical tools to infer the presence of 'weak' or 'strong' seed banks are still largely missing. A basic statistical theory for seed banks, which is able to analyze patterns of population structure and genetic variability, is still in its infancy and further development is urgently needed. The third and last chief goal of this work is to contribute to this development, computing F_{ST}, SH and the expected SFS in some important cases.

It is worth pointing out that several tools from stochastics are needed. The most used among them are: theory of stochastic differential equations (SDEs), in particular diffusion equations; moment duality; diffusion limits; theory of continuous-time Markov chains.

1.2 The basic models

As we already hinted at, there are two types of model: forwards-in-time and backwards-in-time. Those two classes of models are strongly linked via moment duality (see Chapter 2). Forwards-in-time models are usually described through their generator or as the diffusion process solving a system of SDEs, while backwards-in-time models are described as continuous-time Markov chains with a finite or countable state space.

1.2.1 Forwards-in-time models

The seed bank diffusion (SBD). The *Wright-Fisher diffusion with seed bank* was recently introduced in [BGKWB16] as the forwards-in-time scaling limit of a bi-allelic Wright-Fisher model (with type space $\{a, A\}$) that describes a population where individuals may stay inactive in a *dormant form* such as seeds or spores (in the *seed bank*), essentially "jumping" a significant (geometrically distributed) number of generations, before rejoining the *active* population. For an active population of size N and a seed bank size $M = \lfloor N/K \rfloor$, $K > 0$, under the classical scaling of speeding up time by a factor N the a-allele frequency process $(X^N(t))_{t\geq 0}$ in the active and $(Y^N(t))_{t\geq 0}$ in the dormant population converge to the (unique strong) solution $(X(t), Y(t))_{t\geq 0}$ of a two-dimensional SDE. In [BGE$^+$15] the model was extended to include mutation in both the active and the dormant population in which case the limiting process is the solution to the SDE given in Definition 1.1 below. Since the population model and limiting result are completely analogous to the case without mutation we refrain from details and instead refer to [BGKWB16], Section 2.

Definition 1.1 (Seed bank diffusion with mutation)**.** Let $(W(t))_{t\geq 0}$ be a standard Brownian motion, u_1, u_2, u_1', u_2' be finite, non-negative constants and c, K finite, positive constants.

The *Wright-Fisher diffusion with seed bank* with parameters $u_1, u_2, u_1', u_2', c, K$, starting in $(x, y) \in [0,1]^2$, is given by the $[0,1]^2$-valued continuous strong Markov process $(X(t), Y(t))_{t\geq 0}$ that is the unique strong solution of the initial value problem

$$\begin{aligned}
\mathrm{d}X(t) &= \big[-u_1 X(t) + u_2(1 - X(t)) + c(Y(t) - X(t))\big]\mathrm{d}t + \sqrt{X(t)(1 - X(t))}\mathrm{d}W(t),\\
\mathrm{d}Y(t) &= \big[-u_1' Y(t) + u_2'(1 - Y(t)) + Kc(X(t) - Y(t))\big]\mathrm{d}t, \qquad (1)
\end{aligned}$$

with $(X(0), Y(0)) = (x, y) \in [0,1]^2$.

The fact that the initial value problem (1) admits a unique strong solution which is a two dimensional continuous strong Markov diffusion is a standard application of the theorem by Yamada and Watanabe ([YW71], Theorem 1) for 2-dimensional diffusions.

The first coordinate process $(X(t))_{t\geq 0}$ can be interpreted as describing the fraction of a-alleles in the limiting *active population*, while $(Y(t))_{t\geq 0}$ gives the fraction of a-alleles in the limiting *dormant population*, i.e. in the seed bank. The parameters

u_1, u_2 describe the mutation rates from a to A, respectively from A to a, in the active population, and u_1', u_2' the corresponding values in the seed bank. Note that the mutation rates may differ for active and dormant individuals. K fixes the so-called *relative seed bank size* ($M = \lfloor N/K \rfloor$) and c is the rate of migration between the active population and seed bank, i.e. initiation of dormancy and resuscitation. For more details on the biological background see [BGKWB16] and [BGE+15].

The two-island diffusion (TID).

A natural extension of this model can be obtained by (potentially) adding noise in the second coordinate. For parameters $u_1, u_2, u_1', u_2', \alpha, \alpha' \geq 0$, $c, c' > 0$ and independent standard Brownian motions $(W(t))_{t\geq 0}, (W'(t))_{t\geq 0}$ consider the initial value problem

$$\begin{aligned} \mathrm{d}X(t) &= \big[- u_1 X(t) + u_2(1- X(t)) + c(Y(t)- X(t))\big]\mathrm{d}t+ \alpha\sqrt{X(t)(1-X(t))}\mathrm{d}W(t), \\ \mathrm{d}Y(t) &= \big[- u_1' Y(t) + u_2'(1- Y(t)) + c'(X(t)- Y(t))\big]\mathrm{d}t+ \alpha'\sqrt{Y(t)(1-Y(t))}\mathrm{d}W'(t), \end{aligned} \tag{2}$$

with $(X(0), Y(0)) = (x, y) \in [0,1]^2$. Existence and uniqueness of the solution are again standard with Theorem 3.2 in [SS80]. For $\alpha = 1$, $\alpha' = 0$ and $c' = cK$ ($K > 0$; this condition will be assumed in all models, unless stated otherwise) this is the seed bank diffusion. For $\alpha' > 0$ we obtain the diffusion of Wright's *two-island model* initially introduced in [Wri31] and considered in this form for example in [KZH08].

Remark 1.2. $(\alpha)^2$ and $(\alpha')^2$ are also called the relative coalescent rates in the 1st and 2nd subpopulation, respectively. Therefore, it makes sense from the biological point of view to see them as functions of the relative seed bank size K. In particular, $\alpha' \sim \sqrt{K}$, while assuming α as constant, is consistent with previous literature (see e.g. [Her94], Section 3.3.1). For the rest of the thesis, we will keep this relationship between α' and K in the background since we usually assume K as fixed anyway. However, at all points in which we directly investigate dependence of any quantity in function of K, we will assume α' to be proportional to $\sqrt{K}$.

Remark 1.3. The infinitesimal generator of this process, which we will denote from now on with $\mathcal{A}^{(1)}$, was calculated in [BGKWB16]:

Lemma 1.4. *The domain $\mathcal{D}(\mathcal{A}^{(1)})$ contains $C^2([0,1]^2)$, the space of twice continuously differentiable (and thus bounded) functions on the domain. Moreover, $\mathcal{A}^{(1)}$ is given by*

$$\begin{aligned} \mathcal{A}^{(1)}(f)(x,y) =& \big[- u_1 x + u_2(1-x) + c(y-x)\big] \frac{df(x,y)}{dx} \\ &+ \big[- u_1' y + u_2'(1-y) + Kc(x-y)\big] \frac{df(x,y)}{dy} \\ &+ \frac{\alpha^2}{2} x(1-x) \frac{d^2 f(x,y)}{dx^2} + \frac{(\alpha')^2}{2} y(1-y) \frac{d^2 f(x,y)}{dy^2}. \end{aligned} \tag{3}$$

Remark 1.5 (Extension to multiple seed banks)**.** It is straightforward to extend the system (1) to several (e.g. geographically) subdivided seed banks. This means that, instead of having one single seed bank of relative sizes $\frac{1}{K}$, we model k seed banks of

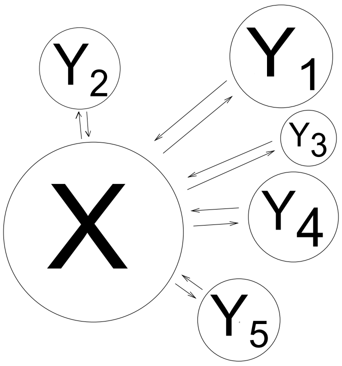


Figure 1: Extension of the model to multiple seed banks ($k = 5$)

relative size $\frac{1}{K_i} \in (0, \infty)$, $i = 1, \ldots, k$, and an active population of relative size 1. Moreover, we suppose that different seed banks may have different mutation and migration rates. We denote the frequency process for the active population by $(X(t))_{t \geq 0}$ and for the i-th seed bank by $(Y_i(t))_{t \geq 0}$, $i = 1, \ldots, k$. Similarly, we consider mutation rates u_1, u_2 in the active population, u_1^i, u_2^i in the i-th seed bank and migration rate $c_1, \ldots, c_k$, where c_i is the rate for migration from the active population to the i-th seed bank (see also figure 1). Then the *seed bank diffusion with k seed banks* is given by the $k + 1$ interacting SDEs

$$\begin{aligned} \mathrm{d}X(t) &= \big[-u_1 X(t) + u_2(1 - X(t)) + \sum_{i=1}^{k} c_i (Y_i(t) - X(t)) \big] \mathrm{d}t \\ &\qquad\qquad\qquad\qquad + \sqrt{X(t)(1 - X(t))} \mathrm{d}W(t), \\ \mathrm{d}Y_i(t) &= \big[-u_1^i Y_i(t) + u_2^i (1 - Y_i(t)) + K_i c_i (X(t) - Y_i(t)) \big] \mathrm{d}t \end{aligned} \tag{4}$$

with initial value $(X(0), Y_1(0), \ldots, Y_k(0)) = (x, y_1, \ldots, y_k) \in [0, 1]^{k+1}$. Note that the only source of randomness is in the active population $(X(t))$ (see also Chapter 3.3).

One can also define an even wider class of models as a joint generalization of both the island model and the seed bank model:
Imagine a population which is subdivided in l subpopulations. Every subpopulation has mutation rates u_1^i and u_2^i, $i = 1, \ldots, l$, a coalescent-relative population size α^i, $i = 1, \ldots, n$ (which can be equal to 0 for some subpopulations, but not for all of them) and, for every two subpopulations i and j, there is migration from subpopulation i to subpopulation j with a relative migration rate c^{ij}. This results in the following system of SDEs:

$$\begin{aligned} \mathrm{d}X_i(t) = \big[-u_1^i X_i(t) + u_2^i (1 - X_i(t)) + &\sum_{j=1, j \neq i}^{l} c^{ij} (X_j(t) - X_i(t)) \big] \mathrm{d}t \\ &+ \alpha^i \sqrt{X_i(t)(1 - X_i(t))} \mathrm{d}W_i(t), \end{aligned} \tag{5}$$

with initial value $(X_1(0), \ldots, X_l(0)) = (x_1, \ldots, x_l) \in [0, 1]^l$. Here, the $W_i, i = 1, .., l$ are independent Brownian motions. Of course, if $l = 2$ and $\alpha^1 = 1$, $\alpha^2 = 0$, we get

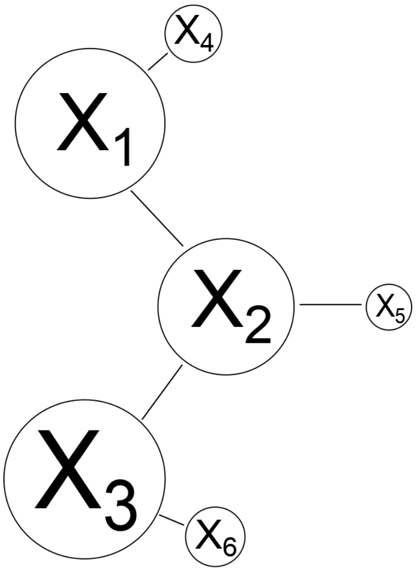


Figure 2: Joint generalization of the two models, as in Formula (5). Imagine X_1-X_3 as “active subpopulations" and X_4-X_6 as “dormant subpopulations". Then, α^1, α^2 and α^3 are strictly positive, while α^4, α^5 and α^6 are zero. Moreover, we can see in the figure that some migration rates, like e.g. $c^{4,6}$, are zero as well.

the seed bank diffusion.
Notice that den Hollander and Pederzani ([dHP17]) investigated these seed bank models on an infinite torus.

1.2.2 Backwards-in-time models

Kingman's coalescent (K). The standard model of genetic ancestry in the absence of a seed bank is the *coalescent* (or *Kingman's coalescent*) [Kin82a], which describes ancestries of samples of size $n \in \mathbb{N}$ from a large selectively neutral, panmictic population of size $N \gg n$ following e.g. a Wright-Fisher model. Measuring time in units of N and tracing the ancestry of a sample of size $n \ll N$ backwards in time results in a coalescent process Π^n as $N \to \infty$. Formally, Π^n is defined as a partition-valued continuous-time Markov process starting at the trivial partition of $\{1, \ldots, n\}$, i.e. $(\{\{1\}, \{2\}, \ldots, \{n\}\})$ where any two blocks merge at rate 1. For many purposes, the block-counting process, describing only the number of blocks independently of their size, is sufficient. It is defined as the continuous-time Markov process on $\{1, \ldots, n\}$, starting in n, jumping from k to $k-1$ at rate $\binom{k}{2}$ and getting absorbed in 1.

A rooted ancestral tree is formed once the most recent common ancestor of the whole sample is reached. We denote this scenario by K. This model is currently the standard null model in population genetics (see e.g. [Wak09] for an introduction) and arises from a large class of population models.

'Weak' seed banks and delayed coalescents (`W`). The basic coalescent model was extended in [KKL01] to incorporate a *'weak' seed bank effect.* In this model, an individual does not always inherit its genetic material from a parent in the previous generation, but rather from a parent that was alive a random number of generations ago. The random separation is assumed to be modeled by a bounded random variable with mean β^{-1} for some $\beta \in (0,1]$. Again, after measuring time in units of N and tracing the ancestry of a sample of size $n \ll N$ as above, it can be shown that the genealogy is still given by a coalescent, but now each pair of lineages merges to a common ancestor independently with rate β^2, as opposed to 1. Thus, the effect of the seed bank is to stretch the branches of the Kingman coalescent by a constant factor [KKL01, BGKS13]. We call the corresponding coalescent a 'delayed coalescent' and denote this 'weak' seed bank scenario by `W`. It should be noted that the overall topological tree structure is identical to that under Kingman's coalescent. Thus, for example, the normalized frequency spectrum of a sample in the infinitely many sites model remains unchanged [BGE$^+$15], and in fact the delayed coalescent with mean delay β^{-1} and population-rescaled mutation rate $u/2 > 0$ is statistically identical to Kingman's coalescent with population-rescaled mutation rate $u/(2\beta^2)$. Nevertheless, the seed bank does have potentially important consequences e.g. for the estimation of effective population size and mutation rates in the presence of prior information, or some other means of resolving the lack of identifiability.

'Strong' seed banks and the 'seed bank coalescent' (`S`). As in [BGKWB16], we want to extend the Wright-Fisher framework to a model with a classical 'active' population of size N and a separate 'seed bank' of comparable size $M := \lfloor N/K \rfloor$, for some $K > 0$, allowing for 'migration' (of a fraction of c/N individuals) between the two subpopulations. Considering samples of size $n^{(1)} \ll N$ and $n^{(2)} \ll N$ from the active and dormant population, respectively, and again measuring time in units of N as before, the genealogy is now described for $N \to \infty$ by the so-called *seed bank coalescent (without mutation)* [BGKWB16], in which active lineages fall dormant at rate c, and dormant lines resuscitate at rate cK. The seed bank coalescent is defined through the following steps:

• Define the space of marked partitions $\mathcal{P}_n^{\{p,s\}}$ as follows: $\mathcal{P}_n$ being the set of partitions of $[n] := \{1, 2, \dots, n\}$, $n = n^{(1)} + n^{(2)}$, let $|\pi|$ be the number of blocks of the partition π for any $\pi \in \mathcal{P}_n$. Then,

$$\mathcal{P}_n^{\{p,s\}} := \big\{(\zeta, u) : \zeta \in \mathcal{P}_n, u \in \{p, s\}^{|\zeta|}\big\}.$$

That is, every block of the partition is endowed either with a p- or with an s-label.

• Considering two marked partitions $\pi, \pi' \in \mathcal{P}_n$, we say that $\pi \prec_p \pi'$ if π' can be constructed by merging exactly two blocks of π carrying a p-label, and the resulting block in π' obtained from the merger again carries a p-label. Similarly, define $\pi \prec_s \pi'$. We use the notation $\pi \bowtie \pi'$ if π' can be constructed by changing the label of precisely one block of π.

• Define the seed bank n-coalescent $(Z(t))_{t \geq 0}$ as a continuous-time Markov chain on $\mathcal{P}_n^{\{p,s\}}$ with the Q-matrix given by

$$Q_{\pi,\pi'} = \begin{cases} 1 & \text{if } \pi \prec_p \pi', \\ c & \text{if } \pi \bowtie \pi', \text{ and a } p \text{ has been replaced by an } s, \\ Kc & \text{if } \pi \bowtie \pi', \text{ and an } s \text{ has been replaced by a } p, \\ 0 & \text{else with } \pi \neq \pi', \\ -\sum_{\pi' \neq \pi} Q_{\pi,\pi'} & \text{if } \pi = \pi' \end{cases}$$

and starting at $\{\{1\}^p, \ldots, \{n^{(1)}\}^p, \{n^{(1)}+1\}^s, \ldots, \{n^{(1)}+n^{(2)}\}^s\}$. Notice that in the seed bank n-coalescent, dormant lineages cannot merge.

- The seed-bank coalescent is then defined as the unique Markov process which, for all $n \in \mathbb{N}$, if started in $(n^{(1)}, n^{(2)})$ with $n = n^{(1)} + n^{(2)}$, is equal in finite-dimensional distributions to the seed bank n-coalescent[1].

Moreover, the block-counting process of the seed bank n-coalescent is given as follows:

Definition 1.6 (Block-counting process of the seed bank coalescent)**.** Consider the space $E := \mathbb{N}_0 \times \mathbb{N}_0$ equipped with the discrete topology. Let $c, K > 0$. The *block-counting process of the seed bank coalescent* $(N(t), M(t))_{t \geq 0}$ is defined as the continuous time Markov chain with values in E, started in $(n^{(1)}, n^{(2)})$, with jump rates given by:

$$\bar{N}_{(n,m),(\bar{n},\bar{m})} = \begin{cases} \binom{n}{2} & \text{if } n \geq 1, \ (\bar{n}, \bar{m}) = (n-1, m), \\ cn & \text{if } n \geq 1, \ (\bar{n}, \bar{m}) = (n-1, m+1), \\ cKm & \text{if } m \geq 1, \ (\bar{n}, \bar{m}) = (n+1, m-1), \end{cases}$$

for every $(n, m) \in \mathbb{N}_0 \times \mathbb{N}_0$ (with the convention $\binom{1}{2} = \binom{0}{2} = 0$) and zero otherwise off the diagonal.

Thus, the properties of the ancestral process are drastically changed, and we speak of a *strong seed bank*, denoting this scenario by `S`. The *seed bank coalescent* has a very different site frequency spectrum compared to the classical (`K`) and weak seed bank (`W`) scenarios [BGE+15].

See [BGE+15] for a full description of the modeling assumptions and the derivation of the seed bank coalescent parameters c, K.

The coalescent process can be interpreted as follows. Lineages are labeled as *active* (first component) or *dormant* (second component). Each pair of active lineages merges to a common ancestor independently with rate 1, as before in the classical Kingman coalescent, while dormant lineages are not allowed to merge. Further, each active lineage becomes dormant independently with rate $c > 0$, and each dormant lineage becomes active with rate $Kc > 0$. The parameter c describes the migration rate between active and dormant populations, and K denotes the relative seed bank size.

Difficulties arise when adding *mutation* to the model. We will discuss the details in Chapter 3.

[1] The seed bank coalescent can be obtained via the projective limit of n-coalescents for $n \to \infty$.

The two island model and the structured coalescent (TI). Having modeled a strong seed bank as a separate population linked to the active one via migration, it is natural to investigate its relation to Wright's two island model [Her94, Wak09]. In the simplest case (which we assume throughout) there are two populations (1 and 2) of respective sizes N and $M = \lfloor N/K \rfloor$, with a fixed fraction of $\lfloor c/N \rfloor$ individuals migrating both from 1 to 2 and from 2 to 1 in every generation. With time measured in units of $N \to \infty$ generations, considering sample sizes $n^{(1)} \ll N$ from island 1 and $n^{(2)} \ll M$ from island 2, this model gives rise to a similar ancestral process as the strong seed bank coalescent except that all pairs of lineages in subpopulation 2 may also merge to a common ancestor independently with rate K, leading to the following coalescent process, which we call the *structured coalescent* [Her94], describing the ancestry of a geographically structured population with migration. This process is also a continuous time Markov chain with values in $\mathcal{P}_n$ which is defined exactly like the seed bank coalescent except for the following two differences:
• $Q_{\pi,\pi'} = \alpha^2$ if $\pi \prec_p \pi'$;
• $Q_{\pi,\pi'} = (\alpha')^2$ if $\pi \prec_s \pi'$, that is, if π' can be constructed by merging exactly two blocks of π carrying an s-label, and the resulting block in π' obtained from the merger again carries an s-label.
Again, we can define a block-counting process with values in $E := \mathbb{N}_0 \times \mathbb{N}_0$ and jump rates equal to

$$\bar{N}_{(n,m),(\bar{n},\bar{m})} = \begin{cases} \alpha^2\binom{n}{2} & \text{if } (\bar{n},\bar{m}) = (n-1,m), \\ (\alpha')^2\binom{m}{2} & \text{if } (\bar{n},\bar{m}) = (n,m-1), \\ cn & \text{if } (\bar{n},\bar{m}) = (n-1,m+1), \\ cKm & \text{if } (\bar{n},\bar{m}) = (n+1,m-1), \end{cases}$$

for every $(n,m) \in \mathbb{N}_0 \times \mathbb{N}_0$ (with the convention $\binom{1}{2} = \binom{0}{2} = 0$) and zero otherwise off the diagonal.
We denote this scenario by TI.

1.3 Models of mutation

We consider three popular models of genetic diversity and mutation: the infinite alleles model (IAM), the finite alleles model (FAM) (which we will usually take to be the two alleles model for brevity, but our results generalize to any finite number of alleles), and the infinite sites model (ISM), each of which we outline below.

The infinite alleles model (IAM). Given a coalescent tree distributed according to either K, W, S, or TI, a sample of genetic data from the infinite alleles model is generated by assigning an arbitrary allele to the most recent common ancestor, and simulating mutations along the branches of the coalescent tree with population-rescaled mutation rate $u > 0$. Each mutation results in a new (parent-independent) allele that has never existed in the population previously, and alleles are inherited along lineages in the absence of mutation. In the cases K and W, a sample of size $n \in \mathbb{N}$ is then described by a tuple of length n, $k := (k_1, \dots, k_n)$, in which k_i is the number of lineages carrying allele i (in some fixed but arbitrary ordering of observed alleles). Note that $k_1 + \ldots + k_n = n$, where the vector is padded with zero entries if fewer than n distinct alleles are observed for notational convenience. In the cases `S` and `TI`, we need to distinguish lineages from the two sub-populations (active and dormant resp. two islands). We consider a sample of size $n := n^{(1)} + n^{(2)}$ (where $n^{(i)}$ is the sample size on subpopulation/island i) by the pair $(k^{(1)}, k^{(2)})$, where both tuples are of length n and $k_j^{(i)}$ counts the number of j-alleles on island i. The (somewhat outdated) infinite alleles model is appropriate when the data is uninformative enough that it is possible to discern when two alleles are different, but no further information is available, such as is the case for data obtained by electrophoresis [HL66].

The finite alleles model (FAM). Here, we consider a finite set of possible alleles, which we identify with $\{1, \ldots, d\}$. The type of the most recent common ancestor is sampled from some probability mass function $\pi = (\pi_1, \ldots, \pi_d)$, which is usually the stationary distribution, and mutations occur along the branches of the coalescent tree at rate u as before. At a mutation event, a new allele is sampled from a $d \times d$ stochastic matrix P, and alleles are inherited along branches in the absence of mutation as before. Using similar notation as for the IAM, under `S` and `TI`, a sample of size $n := n^{(1)} + n^{(2)}$ is described by a pair $(\mathbf{n}^{(1)}, \mathbf{n}^{(2)})$ of vectors of allele frequencies, now each of length d. In this thesis, we usually take $d = 2$, and set $u_2 := uP_{12}$ as well as $u_1 := uP_{21}$ for notational brevity. We also fix $(\pi_1, \pi_2) = (u_2/(u_1 + u_2), u_1/(u_1 + u_2))$, which corresponds to a population evolving at stationarity.

The finite alleles model is much richer than the infinite alleles model, but it is also less tractable. The main difficulty is the possibility of back-mutations, which are lineages that mutate and later revert back to their original allele via a reverse mutation. A compromise between these two extremes is the infinite sites model, often suitable to treat real DNA sequence data.

The infinite sites model (ISM). In this model we identify the ancestral locus with the unit interval $[0, 1]$. Mutations, which continue to occur on the branches of the coalescent tree with rate u, always occur at distinct locations, and are inherited along the branches of the tree so that the allele of an individual is the list of all mutations along its ancestral line. Thus, the whole history of mutations up to the root is retained. Under `S` and `TI`, a sample of size $n := n^{(1)} + n^{(2)}$ is specified by the triple $(\mathbf{t}, \mathbf{n}^{(1)}, \mathbf{n}^{(2)})$, where $\mathbf{t} := (t_1, \ldots, t_k)$ is the list of all observed alleles, and $n_j^{(i)}$ is the observed frequency of allele t_j on island i. In the simpler cases `K` and `W`, it suffices to consider $(\mathbf{t}, \mathbf{n})$. For details on this parametrization of the infinite sites model and its relation to coalescent models see e.g. [BB08].

Remark 1.7. In scenarios `S` and `TI`, the mutation rate is allowed to differ between active and dormant lineages, and we denote the respective rates by u and u' when necessary. It is an open question whether mutations take place on dormant lineages in nature, perhaps at a reduced rate [SL18].

It is well known that all four coalescent models are dual to their respective Wright-Fisher diffusions, the exact form of which depends on the accompanying mutation model. Using the notation from the previous subchapter, duals to scenarios `K`, `W`, and

$\mathtt{S}$ can be recovered as special cases: for $\mathtt{K}$ we set $\alpha = 1$ and $c = 0$, for $\mathtt{W}$ we take $\alpha = \beta$ and $c = 0$, and for $\mathtt{S}$ we take $\alpha = 1$ and $\alpha' = 0$. For scenarios $\mathtt{K}$ and $\mathtt{W}$ we only consider the $X(t)$-coordinate, while in scenario $\mathtt{S}$, the $X(t)$-coordinate corresponds to the active population, while $Y(t)$ is the seed bank.

1.4 Outline of the thesis

The rest of the thesis is organized as follows.
Chapter 2 presents some preliminaries. We summarize without proof the main mathematical tools we need, with a focus on moment duality, generators, polynomial diffusions and phase-type distribution theory.
In Chapter 3, we have compiled some basic facts concerning the forwards-in-time diffusion processes. In particular, we make sure that these admit a unique stationary distribution characterized by its mixed moments, thus allowing us to directly refer to those in the rest of the thesis.
Chapter 4 is devoted to the study of the topic of boundary behavior. Here, it is of interest to know whether the diffusion process can reach the boundaries of the state space in finite time with positive probability. Our main result, which is stated and proved in subchapter 4.3, shows that this depends on the mutation rates and that the critical value is $1/2$ if $\alpha = 1$.
In Chapter 5 we will be concerned with scaling limits. That is, we wish to investigate what happens if the migration rate or the relative size of the seed bank goes to 0 or infinity. We will first provide a detailed exposition and formalization of this problem. In particular, we will see how a time-rescaling is always necessary. As a result, we will introduce a new process called the *ancient ancestral lines process* as the “fast” limit for the migration rate going to 0.
In Chapter 6 we proceed with the study of some measures of population structure. In particular, we will introduce the sample heterozygosity H, Wright’s F_{ST} and the expected site frequency spectrum (SFS). The tool we will mostly use for this aim is phase-type distribution theory. We can show that both the F_{ST} and the expected, non-normalized SFS can give us useful information for distinguishing between the models $\mathtt{S}$ and $\mathtt{TI}$.
As for prerequisites, the reader is expected to be familiar with basic stochastic calculus, in particular with the concept of Markov process and stochastic differential equations (SDEs). The standard stochastic analysis notation we will use is taken from an excellent book on these topics, namely the one by Ethier and Kurtz ([EK86]).

2 Methods

One of the peculiarities of diffusion processes is that they can be applied to a plethora of real-life situations, from physics to biology and finance. Indeed, in this chapter we present some mathematical methods that stem originally from other branches of mathematics, but that can be applied to the analysis of the models presented in Chapter 1 as well. The emerging results will then be used in the rest of this thesis.
We will look at (moment) duality, three arguments related to boundary behavior (speed and scale and the McKean and Lyapunov arguments), polynomial diffusion theory and phase-type distribution theory.
For every topic, we will see an application in an easy case. We will pick either the one-dimensional Wright-Fisher diffusion (with mutation), defined as the $[0,1]$-valued Markov process $(X(t))_{t\geq 0}$ which is the unique strong solution of the SDE

$$\mathrm{d}X(t) = (-u_1 X(t) + u_2(1-X(t)))\mathrm{d}t + \sqrt{X(t)(1-X(t))}\mathrm{d}W(t),\ u_1, u_2 \geq 0.$$

Its generator, acting on continuous and twice differentiable functions on $[0,1]$, is

$$\mathcal{A}(f)(x) = (-u_1 x + u_2(1-x))\frac{\partial f}{\partial x} + \frac{x(1-x)}{2}\frac{\partial^2 f}{\partial x^2}.$$

Alternatively, we will choose the standard Kingman coalescent defined in Chapter 1 (or its block-counting process).
A comprehensive outline of these topics, however, falls outside of the aim of this work; therefore, we will confine ourselves to the results we use in the rest of the thesis. For more details, we will provide key references at the appropriate places.

2.1 Duality

A convenient way to study the behavior of diffusions in population genetics has proven to be the usage of duality for Markov processes. The concept of duality is defined as follows (see e.g. [JK14]):

Definition 2.1. Let

$$X = (\Omega_1, \mathcal{F}_1, (X(t))_{t\geq 0}, (\mathbb{P}_x)_{x\in E}) \text{ and } Y = (\Omega_2, \mathcal{F}_2, (Y(t))_{t\geq 0}, (\mathbb{P}_y)_{y\in F})$$

be two Markov processes, taking values in two Polish state spaces E and F endowed with Borel σ-algebras. Then, X and Y are dual to each other with respect to a bounded, measurable function $S : E \times F \to \mathbb{R}$ if for all $x \in E$, $y \in F$ and $t \geq 0$,

$$\mathbb{E}_y[S(x, Y(t))] = \mathbb{E}_x[S(X(t), y)].$$

The art lies in finding the duality function that suits us best.

Some of the most used duality functions in the case $E = F = \mathbb{R}$ are:

- $S(x,y) = \mathbb{1}_{x\leq y}$ (Siegmund duality),
- $S(x,y) = \mathbb{1}_{x\wedge y=0}$ (coalescing dual for interacting particle systems),
- $S(x,y) = x^y$ (moment duality).

In this thesis, we will use moment duality only, a key tool in diffusion theory which is also used in the contexts of random walks ([HA07], [JK14]), queuing theory ([Asm08]), and interacting particle systems ([Lig12]).

Example: 1-dimensional Wright-Fisher diffusion We can show that the block-counting process of the Kingman coalescent is the (moment) dual of the Wright-Fisher diffusion. The result is obtained by proving the assumptions of Proposition 1.2 in [JK14], from which we also take the notation, which we will state here:

Theorem 2.2. *Let $(Z_1(t))_{t\geq 0}, (Z_2(t))_{t\geq 0}$ be Markov processes taking values in two Polish state spaces E and F endowed with Borel σ-algebras, with generators $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}$, and let S, the real-valued duality function, be bounded and continuous. Define P and $\bar{P}$ as the semi-groups corresponding to $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$, respectively. Then, if $S(x,\cdot)$, $P_tS(x,\cdot) \in \mathcal{D}(\mathcal{G}^{(2)})$ for all x and $t \geq 0$ and $S(\cdot,y)$, $\bar{P}_tS(\cdot,y) \in \mathcal{D}(\mathcal{G}^{(1)})$ for all y and $t \geq 0$, and if*

$$\mathcal{G}^{(2)}S(x,\cdot)(y) = \mathcal{G}^{(1)}S(\cdot,y)(x)$$

for all x, y, then $(Z_1(t))$ and $(Z_2(t))$ are dual with respect to S.

In the case in which $(Z_1(t))_{t\geq 0} = (X(t))_{t\geq 0}$ is the Wright-Fisher diffusion with mutation and $(X_2(t))_{t\geq 0} = (N(t))_{t\geq 0}$ the Kingman block-counting process started in $k \in \mathbb{N}$, we have:

$$\begin{aligned}
\mathcal{G}_1(f)(x) &= (-u_1x + u_2(1-x))\frac{\partial V}{\partial x} + \frac{x(1-x)}{2}\frac{\partial^2 f}{\partial x^2},\ f \in C^2([0,1]);\\
\mathcal{G}_2(f)(n) &= \binom{n}{2}[f(n-1) - f(n)]\mathbb{1}_{n\geq 2},\ f : \{1,\dots,k\} \to \mathbb{R};\\
S(x,\cdot) &: n \to x^n \in \mathcal{D}(\mathcal{G}^{(2)});\ S(\cdot,n) : x \to x^n \in \mathcal{D}(\mathcal{G}^{(1)});\\
P_tS(x,\cdot) &: n \to \mathbb{E}_x[X(t)^n] \in \mathcal{D}(\mathcal{G}^{(2)});\ \bar{P}_tS(\cdot,n) : x \to \mathbb{E}_n[x^{N(t)}] \in \mathcal{D}(\mathcal{G}^{(1)}),
\end{aligned}$$

where in the first two lines we used the formulae for the infinitesimal generator of a diffusion process and a pure jump process, respectively. Moreover,

$$\begin{aligned}
\mathcal{G}_2(S)(x,\cdot)(n) &= \binom{n}{2}(x^{n-1} - x^n)\\
&= \frac{n(n-1)}{2}x^{n-1}(1-x)\\
&= \frac{x(1-x)}{2}n(n-1)x^{n-2}\\
&= \frac{x(1-x)}{2}\frac{\partial^2 x^n}{\partial x^2}\\
&= \mathcal{G}_1(S)(\cdot,n)(x),
\end{aligned}$$

(with the convention $\binom{1}{2} = \binom{0}{2} = 0$) which closes the example.

Similarly, we can prove that the block-counting process of the seed bank coalescent is the (moment) dual of the seed bank diffusion (see [BGKWB16], Theorem 2.8); we will see that moment duality fits the bill in our case as well, linking together the forwards-in-time and the backwards-in-time two-island processes described in Chapter 1.

2.2 Boundary behavior arguments

In this subchapter, we ask ourselves the question whether it is possible to find a necessary condition for the diffusion process not to reach a certain boundary in finite time almost surely. This is formalized as follows: for any boundary[2] B and stochastic process X, define the first hitting time[3]

$$\tau_B^X := \inf\{t \geq 0 \mid X(t) \in B\}.$$

We say X *will never hit* B *started from the interior*, if

$$\mathbb{P}^{\mu_0}\left\{\tau_B^X < \infty\right\} = 0$$

for any initial distribution μ_0 such that $\mu_0(B) = 0$. For the sake of simplicity, we will call a boundary B *accessible* if for the relevant stochastic process, the statement that it will never hit B started from the interior does *not* hold.

2.2.1 Speed and scale

For one-dimensional diffusions, the simplest method used in order to study their boundary behavior (as well as many other quantities) is by using *speed and scale* (see e.g. [Eth11]). This method relies on the fact that every one-dimensional diffusion process can be transformed to a standard Brownian motion via a transformation of the state space first and a time-change afterwards. This is synthesized in the following

Theorem 2.3. *(taken from [Eth11], p.45-49, [EK86], Ch.8, Problem 4.1) Take an interval* $[l, r]$*, a drift function* $b \in C^0([l, r])$*, (that is, continuous) and a diffusion function* $\sigma^2 \in C^0([l, r])$ *with* σ^2 *bounded from below on any compact interval* $I \subsetneq (l, r)$ *by an* $\epsilon_I > 0$*. Moreover, assume the SDE*

$$\mathrm{d}X(t) = \frac{\sigma^2(X(t))}{2}\mathrm{d}W_t + b(X(t))\mathrm{d}t$$

admits a unique strong solution $X = (X(t))_{t\geq 0}$ *which is a Markov process on* $[l, r]$*. Then, denoting the scale function by*

$$S(x) := \int_{x_0}^{x} \exp\Big(-\int_{\eta}^{y} \frac{2b(z)}{\sigma^2(z)}\mathrm{d}z\Big)\mathrm{d}y$$

and the speed measure by

$$M(x) := \int_{x_0}^{x} \frac{1}{\sigma^2(z)S'(z)}\mathrm{d}z$$

(x_0*,* η *being any points in* (l, r)*) and defining*

$$u(x) = \int_{x_0}^{x} M\mathrm{d}S \text{ and } v(x) = \int_{x_0}^{x} S\mathrm{d}M,$$

X *will hit the boundary* $b \in \{l, r\}$ *started from the interior if and only if* $|u(b)| < \infty$*.*

[2] In this thesis, this term has no topological meaning; a boundary is simply a measurable set.

[3] This is also a stopping time provided the stochastic process has a.s. continuous paths and the boundary set is either open or closed. Fortunately, all of the choices for X and B in this thesis satisfy these basic properties.

Example: 1-dimensional Wright-Fisher diffusion In our toy model, we can prove that the boundary $\{0\}$ is accessible for the 1-dimensional Wright-Fisher diffusion process if and only if $u_2 < \frac{1}{2}$: In the 1-dimensional Wright-Fisher diffusion process,

$$[l, r] = [0, 1],\ b(x) = -u_1 x + u_2(1-x) \text{ and } \sigma^2(x) = x(1-x),$$

both terms respecting the hypotheses. Assuming that $u_1,\ u_2 > 0$, we can calculate

$$\begin{aligned} s(y) &= \exp\big(-\int_\eta^y \frac{2b(z)}{\sigma^2(z)}\mathrm{d}z\big) = \exp\big(-2\int_\eta^y \frac{-u_1 z + u_2(1-z)}{z(1-z)}\mathrm{d}z\big) \\ &= (1-y)^{-2u_1} y^{-2u_2} (1-\eta)^{2u_1} \eta^{2u_2} = C_1 (1-y)^{-2u_1} y^{-2u_2}, \end{aligned}$$

$C_1 > 0$ being a constant in y. Moreover,

$$m(z) := \frac{1}{s(z)\sigma^2(z)} = \frac{(1-z)^{2u_1-1} z^{2u_2-1}}{C_1}.$$

Therefore, for any $x_0, y_0 \in (0,1)$,

$$\begin{aligned} \int_{x_0}^x M(y)\mathrm{d}S(y) &= \int_{x_0}^x \int_{y_0}^y m(z)\mathrm{d}z\ s(y)\mathrm{d}y \\ &= \int_{x_0}^x \int_{y_0}^y (1-z)^{2u_1-1} z^{2u_2-1}\mathrm{d}z\ (1-y)^{-2u_1} y^{-2u_2}\mathrm{d}y. \end{aligned}$$

Denote

$$w(y) := \int_{y_0}^y (1-z)^{2u_1-1} z^{2u_2-1}\mathrm{d}z.$$

Since

$$\int_0^1 (1-z)^{2u_1-1} z^{2u_2-1}\mathrm{d}z = \frac{\Gamma(2u_1)\Gamma(2u_2)}{\Gamma(2(u_1+u_2))},$$

w is a bounded function; moreover, it does not converge to 0 for $y \to 0$. Therefore, we get that

$$\Big|\int_{x_0}^0 M(y)\mathrm{d}S(y)\Big| < \infty \text{ if and only if } u_2 < 1/2.$$

However, this approach cannot be used for the two-dimensional structured models from Chapter 1, since there is in general no way to turn a two-dimensional diffusion process into a Brownian motion by a change of time and space. Therefore, we have to solve our problem by other means. Two possible approaches are the usage of the *McKean argument*, which has as basic idea that a continuous martingale cannot converge to infinity with positive probability because it has to oscillate constantly, and using the so-called *Lyapunov argument*, which works with the infinitesimal generator.

2.2.2 The McKean argument

The McKean argument, introduced in 1969 ([McK69], p.47, Problem 7), is a tool which is used in order to prove accessibility or inaccessibility of certain boundaries. It is based on using continuous local martingales on random intervals and has been used for multiple purposes ([Bru91], [MPS11], [FL16]). Let us show it applied to a well-known example:

Proposition 2.4. *In the standard one-dimensional Wright-Fisher diffusion with mutation and rescaled random genetic drift (i.e. the one defined as the unique strong solution of the SDE*

$$\mathrm{d}X(t) = (-u_1 X(t) + u_2(1 - X(t)))\mathrm{d}t + \alpha\sqrt{X(t)(1 - X(t))}\mathrm{d}W(t))$$

for an $\alpha > 0$, the diffusion will never hit the boundary $\{0\}$ started from the interior if $\alpha^2 \leq 2u_2$.

Of course, this result can be proved using the method of scale function and speed measure as before. However, we can also use a martingale argument (see, for example, [Alf15], Exercise 1.2.18 or Exercise 6.1.3), which we will call the *McKean argument.*

Proof. We start by calculating the stochastic integral $\int_0^t \frac{1}{X(s)}\mathrm{d}X(s)$ for any $t \in (0, \tau_0)$:

$$\begin{aligned}
\int_0^t \frac{1}{X(s)}\mathrm{d}X(s) &= \int_0^t \frac{u_2}{X(s)}\mathrm{d}s - (u_1 + u_2)\int_0^t \mathrm{d}s + \int_0^t \alpha\sqrt{\frac{1 - X(s)}{X(s)}}\mathrm{d}W(s) \\
&= \int_0^t \frac{u_2}{X(s)}\mathrm{d}s - (u_1 + u_2)t + \int_0^t \alpha\sqrt{\frac{1 - X(s)}{X(s)}}\mathrm{d}W(s).
\end{aligned}$$

Define $\tau_0 := \inf\{t \geq 0 : X(t) = 0\}$. Now, we use the previous result to calculate $\ln \frac{X(t)}{X(0)}$, for any $t < \tau_0$, via the Ito formula[4]

$$\begin{aligned}
\ln \frac{X(t)}{X(0)} &= \int_0^t \frac{1}{X(s)}\mathrm{d}X(s) - \int_0^t \frac{1}{2X(s)^2}\mathrm{d}[X](s) \\
&= \int_0^t \frac{u_2}{X(s)}\mathrm{d}s - (u_1 + u_2)t + \int_0^t \alpha\sqrt{\frac{1 - X(s)}{X(s)}}\mathrm{d}W(s) \\
&\qquad - \int_0^t \frac{\alpha^2 X(s)(1 - X(s))}{2X(s)^2}\mathrm{d}s \\
&= \int_0^t \frac{u_2 - \frac{\alpha^2}{2}}{X(s)}\mathrm{d}s + t\Big(\frac{\alpha^2}{2} - u_1 - u_2\Big) + \int_0^t \alpha\sqrt{\frac{1 - X(s)}{X(s)}}\mathrm{d}W(s).
\end{aligned}$$

Now, we assume that $\alpha^2 \leq 2u_2$ and remind that $\mathbb{P}\{X(0) = 0\} = 0$. In this case, the first term is non-negative, and by exponentiating we get the approximation

$$X(t) \geq X(0) \exp\Big(\big(\frac{\alpha^2}{2} - u_1 - u_2\big)t + M(t)\Big),$$

with

$$M(t) := \int_0^t \alpha\sqrt{\frac{1 - X(s)}{X(s)}}\mathrm{d}W(s).$$

[4] The Ito formula can be used in this case even if the function $x \to \log x$ is not from $\mathbb{R}$ to $\mathbb{R}$. In the step of the proof of the Ito formula where we approximate the integrand with polynomials on a compact interval, we just use $[1/M, M]$ instead of $[-M, M]$ as usual. Doing the limit for $M \to \infty$ gives us the validity of the formula on $(0, +\infty)$, which is enough because of the hypothesis $t < \tau_0$.

We will now proceed by contradiction; thus, assume $\mathbb{P}\{\tau_0 < \infty\} > 0$. $(M(t))$ is a local martingale on $[0, \tau_0)$, since it is an integral of an adapted locally bounded process against Brownian motion.

Thus, we have the equation

$$\mathbb{1}_{\{\tau_0<\infty\}} X(t \wedge \tau_0) \geq X(0) \exp\Big(\frac{1}{2}\big(\alpha^2 - 2(u_1 + u_2)\big)(t \wedge \tau_0) + M(t \wedge \tau_0)\Big) \mathbb{1}_{\{\tau_0<\infty\}}.$$

For $t \to \infty$, the left-hand side will converge almost surely to 0 by definition of τ_0 and continuity of our process. Thus,

$$\frac{1}{2}\big(\alpha^2 - 2(u_1 + u_2)\big)(t \wedge \tau_0) + M(t \wedge \tau_0) \to -\infty$$

for almost every path for which $\tau_0 < \infty$ as $t \to \infty$. The only possibility for this is that $M(t \wedge \tau_0) \to -\infty$.

Because of this, we have that

$$\mathbb{1}_{\{\tau_0<\infty\}} M(t \wedge \tau_0) \to -\infty \mathbb{1}_{\{\tau_0<\infty\}}$$

almost surely for $t \to \infty$. But since $(M(t \wedge \tau_0))_{t \geq 0}$ is a continuous local martingale as well, by a corollary of Dambis-Dubins-Schwarz ([RY99], Prop. 1.8) it must either have an almost sure finite limit or its limsup must be $+\infty$ almost surely. But neither is the case (with positive probability, it "stabilizes" at $-\infty$ and its paths are thus bounded from above), which is impossible. Thus, $\tau_0 = \infty$ a.s. and by symmetry (notice that $(1-X(t))_{t\geq 0}$ is a classical Wright-Fisher process as well, which represents the frequency process of A-alleles), if $\alpha^2 \leq 2u_1$ and we start in (0,1), $\tau_1 = \infty$ a.s. as well.

□

2.2.3 The Lyapunov argument

Another method for analyzing the behavior of a diffusion is the so-called Lyapunov argument. While the McKean argument is based on a martingale obtained via the usage of the Ito formula, the Lyapunov argument is based on the Dynkin formula and on Foster-Lyapunov inequalities for the infinitesimal generator ([Kus67], [MT92], [MT93]). The Lyapunov argument is usually used in order to see whether a stochastic process will explode (that is, reach a point at infinity) in finite time, but can be applied to our case without too many problems. More specifically, we can use two results, called (CD0) and (CD1) in [MT93], where they are presented in the context of respectively Theorem 2.1 and Theorem 3.1. Both of them can be applied to our models using the following result.

Proposition 2.5. *Let $X = (X(t))_{t\geq 0}$ be a Borel right*[5] *Markov process with values in $E \subseteq \mathbb{R}^d$ endowed with a boundary $B \subseteq E \cup \{\infty\}$. Define $\{O_n : n \in \mathbb{Z}_+\}$ as a fixed*

[5] For the (quite technical) definition, see e.g. [Sha88]. For our aims, it is enough to know that all strong Markov processes on a Borel space with a.s. cádlág paths and whose semigroup maps bounded functions into bounded functions are Borel right processes; see [Mey66].

family of open precompact sets with $O_n \uparrow E \setminus B$. Moreover, the following hypotheses hold:
a) $X(0) = x \in E \setminus B$ a.s.;
b) there exists a continuous function $V : E \backslash B \to [0, +\infty)$ for which, for every sequence of points in $E \setminus B$ (x_n) with $x_n \to x \in B$, $V(x_n) \to \infty$;
c) for every $n \in \mathbb{N}$, the Markov process restricted on O_n has a generator $\mathcal{A}^n$ for which $V \in \mathcal{D}(\mathcal{A}^n)$, $\mathcal{A}^n(V)$ is a continuous function and

$$\mathcal{A}^n(V)(x) \leq CV(x) + D$$

for every $x \in E \setminus B$ and some constants (independent from n) $C, D \geq 0$.
Then, X will never hit the boundary B started from the interior, i.e. $\tau_B^X = \infty$ a.s.

Remark 2.6. • In ([MT93]), a V as in the definition is called a *norm-like function.*
• It is usually not hard to choose a fitting sequence of O_ns. For example, if $B \subsetneq E$, we can choose

$$O_n = \{x \in E : d(x, B) > \frac{1}{n}\},$$

with d the Euclidean distance; for a boundary at ∞,

$$O_n = \{x \in E : ||x|| < n\}.$$

Proof. We prove the thesis similarly to the proof of result (CD0). For this aim, we need to go through the following steps:

1) We remove B from our state space E. We then define a Markov process $(\check{X}(t))_{t \geq 0}$ with state space $E \setminus B$ which has the same behavior as X when X is in $E \setminus B$, e.g. by defining

$$\check{X}(t) := X(t - \gamma_t);\ \gamma_t := \sup\{u \in [0, t] : X(u) \in B\},$$

with the convention $\sup \emptyset = 0$.
2) We introduce the stopping time τ defined as the time when the process first enters the boundary, that is,

$$\begin{aligned} \tau := \inf\{t \in [0, \infty) : X(t) \in B\} &= \lim_{n \to \infty} \inf\{t \in [0, \infty) : X(t) \notin O_n\} \\ &= \lim_{n \to \infty} \inf\{t \in [0, \infty) : \check{X}(t) \notin O_n\} \end{aligned}$$

by definition of limit and construction of $\check{X}$.

3) We define a sequence of stopping times

$$\tau_m := \inf\{t \in [0, \infty) : \check{X}(t) \notin O_m\}.$$

Define $X^m(t) := X(t \wedge \tau_m)$, and notice that for the resulting stochastic process, $(X^m(t)) = (\check{X}^m(t))$ almost surely by construction (since the paths are almost surely continuous, we will exit O_m - and thus stop the process - before we enter B).

4) Notice that, without loss of generality, $D = 0$; else define $\hat{V}(x) = V(x) + D/C$ (notice that $\hat{V}$ is still norm-like) and proceed accordingly.

5) Define the *weak*[6] *extended generator*[7] $\mathcal{G}$ of a Markov process X as the operator acting on all measurable functions $h : E \setminus B \times (0, \infty) \to \mathbb{R}$ for which, for all $(x, t) \in E \setminus B \times (0, \infty)$, the limit

$$\mathcal{G}(X)(h)(x,t) := \lim_{\epsilon \to 0} \frac{\mathbb{E}_x[h(X(\epsilon), t+\epsilon)] - h(x,t)}{\epsilon}$$

exists pointwise, $|\mathcal{G}(X)(h)|$ is bounded on all compact sets and

$$\lim_{\epsilon \to 0} \mathbb{E}_x[\mathcal{G}(X)(h)(X(\epsilon), t+\epsilon)] = \mathcal{G}(X)(h)(x,t).$$

6) Denote with $\mathcal{A}^m$ the (standard) infinitesimal generator of X^m, and with $\mathcal{G}^m$ the weak extended generator. Notice that $\mathcal{A}^m(f)$ and $\mathcal{G}^m(h)$ are just restrictions of $\mathcal{A}(f)$ and $\mathcal{G}(h)$, for any function $f \in D(\mathcal{A})$ and $h \in D(\mathcal{G})$.

7) Notice that $V : E \setminus B \to \mathbb{R}_+$ has a restriction in the domain of $\mathcal{A}^m$ for every m, and that $\overline{\{V \leq x\}}$ is a compact subset of $E \setminus B$ for every $x \in (0, \infty)$.

8) Apply $\mathcal{G}^m$ to

$$h(x,t) := V(x)e^{-Ct}.$$

Then, we get

$$\begin{aligned}
&\lim_{\epsilon \to 0} \frac{\mathbb{E}_x[V(X(\epsilon))e^{-C(t+\epsilon)}] - V(x)e^{-Ct}}{\epsilon} = e^{-Ct} \lim_{\epsilon \to 0} \Big(\frac{\mathbb{E}_x[V(X(\epsilon))e^{-C\epsilon}] - V(x)}{\epsilon} \Big) \\
=&e^{-Ct} \lim_{\epsilon \to 0} \Big(\frac{\mathbb{E}_x[V(X(\epsilon))] + e^{-C\epsilon}V(x) - e^{-C\epsilon}V(x) + V(x)}{\epsilon} \Big) \\
=&e^{-Ct} \Big(\lim_{\epsilon \to 0} e^{-C\epsilon} \frac{\mathbb{E}_x[V(X(\epsilon))] - V(x)}{\epsilon} + V(x) \lim_{\epsilon \to 0} \frac{e^{-C\epsilon} - 1}{\epsilon} \Big) \\
=&e^{-Ct} \Big(\mathcal{A}^m(V)(x) - CV(x) \Big)
\end{aligned}$$

Moreover, using hypothesis c) and Step 4), we get that

$$\mathcal{G}^m((h)(x,t)) = e^{-Ct}(\mathcal{A}^m(V)(x) - CV(x)) \leq 0.$$

9) Now, let $t_m := \tau_m \wedge t$ be a bounded stopping time for any $t > 0$. Then, we can use formula (8) in [MT93] ("Dynkin's formula"), which states that

$$\mathbb{E}^x[\int_0^{T_m} \mathcal{G}(X)(f)(X(s), s) \text{ d}s] = \mathbb{E}^x[f(X(T_m), T_m)] - f(X(0), 0),$$

[6] In the sense of pointwise limit. See also [MT93], Formula 5.

[7] Usually the extended generator is described as the operator which, applied on a function h, yields a function U with $\mathbb{E}_x[h(X(t), t)] = h(x, 0) + \mathbb{E}_x[\int_0^t U(X(s), s) \text{ d}s]$. However, we wish to give a more intuitive, even if slightly stronger, definition since it is enough in our case.

where $\mathcal{G}$ is the extended generator of X, f is in the domain of $\mathcal{G}(X)$, $T_m := \min\{\tau_m, m, \zeta\}$ and ζ is a stopping time for X. This formula stems directly from applying Doob's optional stopping theorem to the martingale representation of the extended generator (Formula 3 in [MT93])

$$\int_0^t \mathcal{G}(X)(f)(X(s), s) \, \mathrm{d}s = f(X(t), t) - f(X(0), 0) - M_f(t),$$

with $(M_f(t))_{t\geq 0}$ a martingale. Applying Dynkin's formula to our process with $\zeta = \tau$ and $h = f$ yields

$$\mathbb{E}[V(X^m(t_m))e^{-Ct_m}] = V(X(0)) + \mathbb{E}\big[\int_0^{t_m} \mathcal{G}_m(h)(X(s), s)\mathrm{d}s\big] \leq V(X(0))$$

(which is finite by hypothesis a)).

10) Define

$$M(t) := e^{-Ct}V(X^m(t))\mathbb{1}_{\tau_m \geq t}.$$

We want to show that $(M(t))_{t\geq 0}$ is a $(\mathcal{F}^X(t))_{t\geq 0}$-supermartingale. The process is bounded by

$$\sup_{x:d(x,B)\geq 1/m} V(x) < \infty$$

(in the $B \subseteq E$ case, and analogously else); moreover, the supermartingale property is given by inspecting both terms in the expression (holding almost surely)

$$\mathbb{E}[M(t)|\mathcal{F}^X(s)] = \mathbb{E}[M(t)|\mathcal{F}^X(s)]\mathbb{1}_{s>\tau_m} + \mathbb{E}[M(t)|\mathcal{F}^X(s)]\mathbb{1}_{s\leq\tau_m} :$$

the first one is zero since $M(t) = 0$ for every $t \geq \tau_m$, and thus, for the same reason, equal to $M(s)$; for the second one, using the strong Markov property,

$$\begin{aligned}
\mathbb{E}[M(t)|\mathcal{F}^X(s)]\mathbb{1}_{s\leq\tau_m} &= e^{-Ct}\mathbb{E}_{X^m(s)}[V(X^m(t-s))\mathbb{1}_{\tau_m\geq(t-s)}]\mathbb{1}_{s\leq\tau_m} \\
&= e^{-Cs}\mathbb{E}_{X^m(s)}\big[e^{-C(t-s)}V(X^m(t-s))\mathbb{1}_{\tau_m\geq(t-s)}\big]\mathbb{1}_{s\leq\tau_m} \\
&= e^{-Cs}\mathbb{E}_{X^m(s)}\big[e^{-C((t-s)\wedge\tau_m)}V(X^m((t-s)\wedge\tau_m))\mathbb{1}_{\tau_m\geq(t-s)}\big]\mathbb{1}_{s\leq\tau_m} \\
&\leq e^{-Cs}\mathbb{E}_{X^m(s)}\big[e^{-C((t-s)\wedge\tau_m)}V(X^m((t-s)\wedge\tau_m))\big]\mathbb{1}_{s\leq\tau_m} \\
&\leq e^{-Cs}\mathbb{E}_{X^m(s)}[V(X^m((t-s)\wedge\tau_m)e^{-C((t-s)\wedge\tau_m)}]\mathbb{1}_{s\leq\tau_m} \\
&\leq e^{-Cs}V(X_m(s))\mathbb{1}_{s\leq\tau_m} = M(s)\mathbb{1}_{s\leq\tau_m} \text{ a. s.}
\end{aligned}$$

having used step 9) in the second to last step.

11) Finally, using Doob's maximal inequality and the monotone convergence theorem, for any $a > 0$:

$$\mathbb{P}\{\sup M(t) \geq a\} = \mathbb{P}\Big\{\sup_{t<\tau_m} V(X(t))e^{-Ct} \geq a\Big\} \leq \frac{V(X(0))}{a}$$

and in the limit,

$$\mathbb{P}\{\sup_{t<\tau} V(X(t))e^{-Ct} \geq a\} \leq \frac{V(X(0))}{a}.$$

Thus, on the one hand,

$$\frac{V(X(0))}{a} \geq \mathbb{P}\Big\{\sup_{t<\tau} V(X(t))e^{-Ct} \geq a\Big\} \geq \mathbb{P}\Big\{\sup_{t<\tau} V(X(t))e^{-Ct} \geq a, \tau < \infty\Big\}.$$

But on the other hand,

$$\Big\{\sup_{t<\tau} V(X(t))e^{-Ct} \geq a\Big\} \supseteq \{\tau < \infty\}$$

since

$$\lim_{t\to\tau} V(X(t))e^{-Ct} = \infty \text{ on } \{\tau < \infty\}.$$

Therefore,

$$V(X(0)) \geq a\ \mathbb{P}\{\tau < \infty\} \text{ for all } a > 0.$$

This means that $\mathbb{P}\{\tau < \infty\} = 0$, i.e. $(V(X(t)))_{t\geq 0}$ cannot reach the boundary in finite time with positive probability, which concludes our proof. □

Notice that this algorithm works only in one direction; in the eventuality that our process `does` reach the boundary (with positive probability), this has to be proved by other means: see the following subchapter.

Example: 1-dimensional Wright-Fisher diffusion We can apply the Lyapunov argument to our "toy model", namely the one-dimensional Wright-Fisher diffusion. This stochastic process has as state space $[0,1]$. For simplicity, we just look at boundary behavior at $\{0\}$. The generator of the diffusion, acting on twice differentiable functions, is

$$\mathcal{A}(f)(x) = (-u_1x + u_2(1-x))\frac{\partial f}{\partial x} + \frac{x(1-x)}{2}\frac{\partial^2 f}{\partial x^2}.$$

Here, if we use the above algorithm, set $O_n := (\frac{1}{n}, 1)$ and $\mathcal{D}(\mathcal{A}^n)$ is the set of twice differentiable functions from $(\frac{1}{n}, 1)$ to $\mathbb{R}$, which means that a function with a singularity at 0 can still be in $\mathcal{D}(\mathcal{A}_n)$. We use this to define $V(x) := -\log x$. For this choice of V, the sufficient condition for non-explosivity becomes

$$(u_1 + u_2 - \frac{u_2}{x}) + \frac{1-x}{2x} \leq -C\log x + D,$$

or, putting all constants together,

$$\frac{\frac{1}{2} - u_2}{x} + C\log x \leq D'.$$

This inequality holds for all $x \in (0,1]$ if and only if $u_2 \geq 1/2$, which matches with the result obtained via scale function and speed measure.

2.3 Polynomial Diffusion Theory

In order to treat the other direction of the boundary behavior problem, *polynomial diffusions* come to our help. Polynomial diffusion theory was introduced by Wong in 1964 ([Won64]) and developed in the context of financial mathematics in the last decade. In particular, it can be used in several topics in finance including market models for interest rates ([CKRT12], [LR14], [FL16], [LP17]). However, we will see that the formulas fit our diffusion processes as well.

We define a polynomial diffusion (as in [FL16], see also the Remark after Definition 2.1 therein) as a diffusion process, taking values on a subset of $\mathbb{R}^n$, of the form

$$\mathrm{d}Z(t) = b(Z(t))\mathrm{d}t + \sigma(Z(t))\mathrm{d}W(t),$$

where b consists of polynomials of degree at most 1, $a := \sigma\sigma^T$ of polynomials of degree at most 2 and W is an n-dimensional standard Brownian motion.

It is now easy to see that the diffusion process on $[0,1]^2$ given by (2) is a polynomial diffusion: it fits the bill with $Z := (X, Y)$,

$$\begin{aligned} b(x,y) &:= \begin{pmatrix} -u_1 x + u_2(1-x) + c(y-x) \\ -u_1' y + u_2'(1-y) + c'(x-y) \end{pmatrix} \\ \text{and } \sigma(x,y) &:= \begin{pmatrix} \alpha\sqrt{x(1-x)} & 0 \\ 0 & \alpha'\sqrt{y(1-y)} \end{pmatrix}. \end{aligned}$$

Now, define $\mathcal{P} = \{x, 1-x, y, 1-y\}$, where we abuse notation using x for the map $(x,y) \mapsto x$, and similarly for the other polynomials. Then, the state space of our diffusion can be defined by

$$[0,1]^2 = \{x \in \mathbb{R}^2 : \forall p \in \mathcal{P}\, p(x) \geq 0\}.$$

Hence we can make use of Theorem 5.7 in [FL16].

Theorem 2.7. *Let X be a weak solution of the (multi-dimensional) stochastic differential equation*

$$\mathrm{d}X(t) = \sigma(X(t))\mathrm{d}W(t) + b(X(t))\mathrm{d}t.$$

Denote as $\mathcal{A}$ the generator of the related stochastic process and E its state space. Suppose that

$$E = \{x \in \mathbb{R}^d |\ p(x) \geq 0\ \forall p \in \mathcal{P}\}$$

for a set of polynomials $\mathcal{P}$. Moreover, suppose that

$$\mathbb{P}\{p(X(0), Y(0)) = 0\} = 0 \textit{ for all } p \in \mathcal{P}$$

and that

$$\int_0^t \mathbb{1}_{p(X(s))=0}\ \mathrm{d}s = 0,\ t \geq 0,\ p \in \mathcal{P},\ a.s.$$

For every $p \in \mathcal{P}$, consider a vector of polynomials h for which $a\nabla p = hp$ with $a := \sigma\sigma^T$. Then:

(i) If there exists a neighborhood U of $E \cap \{p = 0\}$ such that

$$2\mathcal{A}(p) - h^T\nabla p \geq 0 \textit{ on } E \cap U,$$

then $p(X(t)) > 0$ for all $t > 0$, a.s.

(ii) If

$$2\mathcal{A}(p) - h^T\nabla p = 0 \text{ on } \mathbb{R}^d \cap \{p = 0\},$$

and additionally p changes sign on $\mathbb{R}^d$, then $p(X(t)) > 0$ for all $t > 0$, a.s.

(iii) For any $\bar{x} \in E \cap \{p = 0\}$, if

$$\mathcal{A}(p)(\bar{x}) \geq 0 \text{ and } 2\mathcal{A}(p)(\bar{x}) - h^T(\bar{x})\nabla p(\bar{x}) < 0,$$

then for any $T > 0$ there exists $\epsilon > 0$ such that if $||X(0) - \bar{x}|| < \epsilon$ almost surely, then $p(X(t))$ has zeroes in $(0, T]$ with positive probability.

Example: 1-dimensional Wright-Fisher diffusion The boundary behavior problem for the classical Wright-Fisher diffusion can be fully solved using this theorem as well. In that case,

$$b(x) = -u_1x + u_2(1 - x) \text{ and } a(x) = x(1 - x).$$

Moreover,

$$[0, 1] = \{x \in \mathbb{R} : \forall p \in \mathcal{P},\ p(x) \geq 0\} \text{ for } \mathcal{P} = \{x, 1 - x\}.$$

We investigate the boundary behavior at 0 by picking $p(x) = x$. Then, the condition 2.7 is satisfied because of the absence of atoms at the boundaries for the stationary distribution, which follows directly from ([Eth11], Example 3.25). And last, let us note that the assumption that $\{t \geq 0 \mid p(X(t), Y(t)) = 0\}$ be a Lebesgue null set is indeed not required when $\mathbb{P}\{p(X(0), Y(0)) = 0\} = 0$, as can be easily seen from the proof.
Moreover, $h(x) = 1 - x$ and thus,

$$2\mathcal{A}(p) - h^T\nabla p = -2u_1x + (1 - x)(2u_2 - 1).$$

This yields that:
i) $2\mathcal{A}(p) - h^T\nabla p$ is non-negative in a neighborhood of 0 if and only if $u_2 > 1/2$;
ii) $2\mathcal{A}(p) - h^T\nabla p = 0$ on $\{p = 0\}$, and p changes sign on $[0, 1]$;
iii) $\mathcal{A}(p)(0) = u_2$ and $2\mathcal{A}(p) - h^T\nabla p(0) = 2u_2 - 1$.

Therefore, the diffusion process will not reach 0 in finite time a.s. if $u_2 \geq 1/2$; however, if $u_2 < 1/2$, the process will reach 0 in finite time with positive probability if started close enough to 0.

2.4 Phase-type Distribution Theory

Phase-type distributions are a class of probability distributions introduced by Neuts in 1975 ([Neu75b])[8] and used both in the context of finance and of population genetics by Bladt et al. ([Bla05], [HSB19]).
Using the notation from [HSB19], we have a Markov jump process $(M(t))_{t\geq 0}$ with a finite state space $\{x_1, x_2, \ldots, x_{p+1}\}$, where exactly one state x_{p+1} is absorbing, while

[8] This source is hard to find. See also ([Neu75a])

all the others are transient. Therefore, the Q-matrix of the Markov chain must have the form

$$Q = \begin{bmatrix} \mathbf{S} & \mathbf{s} \\ \mathbf{0} & 0 \end{bmatrix},$$

with $\mathbf{S}$ a $p \times p$ matrix and the vector $\mathbf{s} \neq 0$.

Let π be an initial distribution, and assume that $\pi(\{x_{p+1}\}) = 0$. Then, if we define τ as the (a.s. finite) absorption time in x_{p+1} of $(M(t))$ started in π, we say that τ is *phase-type distributed.*

Moreover, the following results (Theorem 2.5 in the aforementioned paper and Theorem 8.1.2 in [BN17] respectively) apply:

Let $r : \{x_1, \dots, x_p\} \to \mathbb{R}_+$ be a so-called reward function. Then, for

$$Y := \int_0^\tau r(M(t))\mathrm{d}t,$$

and any $n \in \mathbb{N}$,

$$\mathbb{E}[Y^n] = \pi((-\mathbf{S})^{-1}\Delta(r)))^n\mathbf{e}, \tag{6}$$

where $\mathbf{e}$ is a vector of ones and $\Delta(r)$ is the diagonal matrix with diagonal entries $r := (r(x_1), \dots, r(x_p))$. Furthermore, the Laplace transform of Y is given by

$$L_Y(u) := \mathbb{E}[e^{-uY}] = \pi(\Delta(ur) - \mathbf{S})^{-1}\mathbf{s}. \tag{7}$$

Example: 1-dimensional Kingman coalescent

If we take a sample of $n = 3$, the block-counting process of the Kingman coalescent is defined as the continuous-time Markov chain M with state space $\{1, 2, 3\}$, $M(0) = 3$, jumping from 3 to 2 at rate 3, from 2 to 1 at rate 1, and getting absorbed in 1.

From the definition, we see that the time to the most recent common ancestor τ is phase-type distributed. Formally, 3, 2 and 1 take the roles respectively of x_1, x_2 and x_3, and in addition:

$$Q = \begin{bmatrix} -3 & 3 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \; p = 2, \; \mathbf{S} = \begin{bmatrix} -3 & 3 \\ 0 & -1 \end{bmatrix}, \; \mathbf{s} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \; \pi = (1, 0), \; r = (1, 1).$$

Thus, equations (6) and (7) can be applied here: the moments of τ are equal to[9]

$$\mathbb{E}[\tau^n] = (1, 0)\left(\begin{bmatrix} \frac{1}{3} & 1 \\ 0 & 1 \end{bmatrix} \right)^n \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

And finally, its Laplace transform is equal to

$$L_\tau(u) := \mathbb{E}[e^{-u\tau}] = \pi \frac{1}{(u+1)(u+3)} \begin{bmatrix} u+1 & 3 \\ 0 & u+3 \end{bmatrix} \mathbf{s} = \frac{3}{(u+1)(u+3)}.$$

[9] Of course, the specific formulas for pure-death processes in [HSB19] could have been used here as well, but we have refrained from it for illustrative purposes.

3 Basic results for the Wright-Fisher diffusion with two subpopulations

The aim of this chapter is to provide a solid foundation for all the results that will come afterwards. First, we want to show that the stationary distribution of the Wright-Fisher diffusion with two subpopulations is unique and can be characterized by the means of mixed moments. Moreover, in this Chapter we will also give two basic results. First, we will see a possible transformation of the system of two SDE's characterizing the seed bank diffusion into a system composed of a stochastic delay differential equation for X and an equation which gives Y in function of X. Then, we will see that none of the diffusion processes presented in the thesis is reversible.
This chapter is based on [K1].

3.1 Moment dual of the seed bank diffusion with mutation

The *moment dual of the seed bank diffusion (without mutation)* $(N(t), M(t))_{t\geq 0}$ has already been introduced in Chapter 1. If we want to introduce *mutation* to the model, difficulties might arise. However, there is more than one way of incorporating this mechanism into a dual. We comment on this as well as on the motivation behind our strategy – adding a *death state* (∂, ∂) to the state-space – below, but let us formally introduce our dual first.

Definition 3.1 (Moment dual of the diffusion (2))**.** Consider the space $E := \mathbb{N}_0 \times \mathbb{N}_0 \cup \{(\partial, \partial)\}$ equipped with the discrete topology. Let $u_1, u_2, u_1', u_2', \alpha, \alpha' \geq 0$, $c, c' > 0$. Define $(N(t), M(t))_{t\geq 0}$ to be the continuous time Markov chain with values in E with conservative generator $\mathcal{A}^{(2)}$ given by:

$$\mathcal{A}^{(2)}_{(n,m),(\bar{n},\bar{m})} = \begin{cases} \alpha^2\binom{n}{2} + nu_2 & \text{if } (\bar{n},\bar{m}) = (n-1, m), \\ (\alpha')^2\binom{m}{2} + mu_2^{'} & \text{if } (\bar{n},\bar{m}) = (n, m-1), \\ nu_1 + mu_1^{'} & \text{if } (\bar{n},\bar{m}) = (\partial, \partial), \\ cn & \text{if } (\bar{n},\bar{m}) = (n-1, m+1), \\ c'm & \text{if } (\bar{n},\bar{m}) = (n+1, m-1), \end{cases}$$

for every $(n, m) \in \mathbb{N}_0 \times \mathbb{N}_0$ (with the convention $\binom{1}{m} = \binom{0}{m} = 0$ for all m) and zero otherwise off the diagonal, started in $(n^{(1)}, n^{(2)}) \in E$.
We will call this process the *moment dual of the diffusion* (2).

The name of the process will be justified in Lemma 3.2 below. This dual arises in the context of *sampling duality.* See [GS18] for a thorough introduction to the concept. It is based on the idea that the question "What is the probability of sampling n individuals of type a at time t, if the frequency of type a is x at time 0?" can be answered in two ways: One, looking forward in time at the diffusion which will give precisely the frequency of type a individuals at time t, but also two, tracing back the genealogy to the number of ancestors of the sample present at time 0 and using the frequency x. It is precisely in this latter question that one realizes the need of an artificial *death state.* In order for all n individuals in the sample to be of type a at

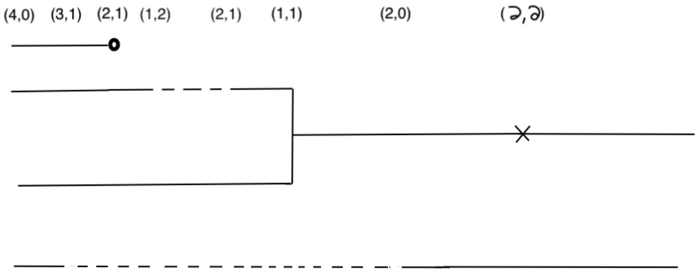


Figure 3: The coalescent corresponding to the process defined in Definition 3.1. The dashed and black lines correspond to the two islands respectively. When a forward-mutation of type $A \mapsto a$ occurs, the line is ended, since it ensures all its leaves to be of type a. A forward-mutation of type $a \mapsto A$ renders it impossible to have all leaves of type a and the process jumps to the death state.

time t it is imperative that we do not encounter a mutation from type a to A in the forward sense, i.e. a mutation from A to a in the coalescent time, on their ancestral lines. Hence, the process $(N(t), M(t))_{t\geq 0}$ is killed off as soon as this happens, since the probability for the sample to be of type a only is now 0. At the same time, if we encounter a mutation of type A to a in the forward sense, i.e. a mutation from type a to A tracing backwards, we are assured all descendants of that line are of type a with probability 1 and we can stop tracing it, whence the process is reduced by one line. See Figure 3 for an illustration.

It is trivial to extend the dual process in Definition 3.1 to a general structured coalescent. A structured-mutation moment dual is new in the literature, as far as we know. These moment duals differ from the *weighted* moment dual for the Wright-Fisher diffusion with mutation introduced in [EG09] and studied in [GJL16] and [EGT10]. The small difference between our construction for mutation and the construction in [EG09], namely the addition of the extra state (∂, ∂), makes our dual compatible with the presence of selection as in [KN97].

The following are straightforward, but important observations on the duals: Note that in the case of $u_1 + u_2 + u_1' + u_2' > 0$, the moment dual of the general diffusion will reach either $\{(0,0)\}$ or $\{(\partial, \partial)\}$ in finite time a.s. (for any starting point $(n^{(1)}, n^{(2)}) \in E$), whereas for $u_1 + u_2 + u_1' + u_2' = 0$ it will reach the set $\{(1,0),(0,1)\}$ in finite time ($\mathbb{P}$-a.s.) and then alternate between these two states. Furthermore observe that, whenever the dual of the general diffusion is started in some $(n^{(1)}, n^{(2)}) \in E$, it will stay in $\{0, \ldots, n^{(1)} + n^{(2)}\} \times \{0, \ldots, n^{(1)} + n^{(2)}\} \cup \{(\partial, \partial)\}$, hence the state space in this case is, indeed, finite.

Lemma 3.2. *Let $S : [0,1]^2 \times E \to [0,1]$ be defined as*

$$S((x,y),(n,m)) := x^n y^m \mathbb{1}_{\mathbb{N}_0 \times \mathbb{N}_0}((n,m))$$

for any $(x,y) \in [0,1]^2$ and $(n,m) \in E$ and let $u_1, u_2, u_1', u_2', \alpha, \alpha' \geq 0$, $c, c' > 0$. Then for every $(x,y) \in [0,1]^2$, $(n,m) \in E$ and for any $t \geq 0$

$$\mathbb{E}^{x,y}[S\big((X(t),Y(t)),(n,m)\big)] = \mathbb{E}_{n,m}[S\big((x,y),(N(t),M(t))\big)],$$

where $(N(t),M(t))_{t\geq 0}$ is defined in Definition 3.1 and $(X(t),Y(t))_{t\geq 0}$ is the solution to the SDE in equation (2)*.*

Proof. Since $S:[0,1]^2\times E\to[0,1]$ is continuous (in the product topology of $[0,1]^2\times E$), the result follows by proving the assumptions of Proposition 1.2 in [JK14] (see Chapter 2.1): Recall the generator $\mathcal{A}^{(1)}$ of $(X(t),Y(t))_{t\geq 0}$ from (3) and observe that for any bounded function $h:E\to\mathbb{R}$, the generator of $(N(t),M(t))_{t\geq 0}$ is given by $\mathcal{A}^{(2)}h((\partial,\partial))=0$ and

$$\begin{aligned}\mathcal{A}^{(2)}h(n,m)=&\left[\alpha^2\binom{n}{2}+nu_2\right][h(n-1,m)-h(n,m)]\mathbb{1}_{\mathbb{N}}(n)\\&+\left[(\alpha')^2\binom{m}{2}+mu_2'\right][h(n,m-1)-h(n,m)]\mathbb{1}_{\mathbb{N}}(m)\\&+c[h(n-1,m+1)-h(n,m)]\mathbb{1}_{\mathbb{N}}(n)\\&+c'[h(n+1,m-1)-h(n,m)]\mathbb{1}_{\mathbb{N}}(m)\\&+[nu_1+mu_1'][h(\partial,\partial)-h(n,m)],\end{aligned}$$

for any $(n,m)\in\mathbb{N}_0\times\mathbb{N}_0$ with the convention that $\binom{1}{2}=0$. Let P and $\bar{P}$ be the semigroups corresponding to $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(2)}$ respectively. Since $(N(t)+M(t))_{t\geq 0}$ is monotonically non-increasing, the assumptions that $S\big((x,y),(n,m)\big),P_tS\big((x,y),(n,m)\big)$ are in the domain of $\mathcal{A}^{(2)}$ and $S\big((x,y),(n,m)\big),\bar{P}_tS\big((x,y),(n,m)\big)$ are in the domain of $\mathcal{A}^{(1)}$ are readily verified.

As $S((x,y),(\partial,\partial))=0$ for all $(x,y)\in[0,1]^2$, we immediately see

$$\mathcal{A}^{(1)}S\big((x,y),(\partial,\partial)\big)=0=\mathcal{A}^{(2)}S((x,y),(\partial,\partial))$$

for any $(x,y)\in[0,1]^2$. Furthermore, if we fix $(x,y)\in[0,1]^2$ and $(n,m)\in\mathbb{N}_0\times\mathbb{N}_0$,

$$\begin{aligned}\mathcal{A}^{(1)}S\big((x,y),(n,m)\big)&=[-u_1x+u_2(1-x)+c(y-x)]\,nx^{n-1}y^m\\&\quad+\frac{\alpha^2}{2}x(1-x)n(n-1)x^{n-2}y^m\\&\quad+\big[-u_1'y+u_2'(1-y)+c'(x-y)\big]\,mx^ny^{m-1}\\&\quad+\frac{(\alpha')^2}{2}y(1-y)m(m-1)x^ny^{m-2}\\&=\left[\alpha^2\binom{n}{2}+nu_2\right]\left[x^{n-1}y^m-x^ny^m\right]\\&\quad+\left[(\alpha')^2\binom{m}{2}+mu_2'\right]\left[x^ny^{m-1}-x^ny^m\right]\\&\quad+c\Big[x^{n-1}y^{m+1}-x^ny^m\Big]+c'\Big[x^{n+1}y^{m-1}-x^ny^m\Big]\\&\quad+(nu_1+mu_1')[0-x^ny^m]\\&=\mathcal{A}^{(2)}S\Big((x,y),(n,m)\Big).\end{aligned}$$

□

This duality now allows us to use the process $(N(t), M(t))_{t\geq 0}$ to study the mixed moments of $(X(t), Y(t))_{t\geq 0}$ from which we can draw conclusions on the limiting behavior of the diffusions itself. The case with and without mutation differs strongly in this behavior.

Lemma 3.3. *In the absence of mutation, in the general diffusion given in* (2) *with* $\alpha = \alpha' = 1$,

$$\frac{yc + xc'}{c + c'} = \lim_{t\to\infty} \mathbb{E}_{n,m}[x^{N(t)}y^{M(t)}] = \lim_{t\to\infty} \mathbb{E}^{x,y}[X(t)^n Y(t)^m]$$

for all $(n, m) \in \mathbb{N}_0 \times \mathbb{N}_0 \setminus \{(0, 0)\}$ *and all* $(x, y) \in [0, 1]^2$. *Moreover,* $(X(t), Y(t))_{t\geq 0}$ *converges* $\mathbb{P}$*-a.s. to a random variable* (X_∞, Y_∞) *with values in* $[0, 1]^2$ *whose distribution is given by*

$$\frac{yc + xc'}{c + c'}\delta_{(1,1)} + \frac{(1 - y)c + (1 - x)c'}{c + c'}\delta_{(0,0)}$$

Proof. Proof almost identical to those of Proposition 2.9 and Corollary 2.10 in [BGKWB16].

□

Note that this is in line with the results for the one-dimensional Wright-Fisher diffusion. In particular, the two-island diffusion without mutation will fixate in finite time at one of the corner points $(0, 0)$ or $(1, 1)$, which means extinction of either of the two alleles.

Proposition 3.4. *Let* $u_1, u_2, u_1', u_2', \alpha, \alpha' \geq 0$, $c, c' > 0$ *and assume that at least one mutation rate among* u_1, u_2, u_1', u_2' *is non-zero. Then, for every* $(n, m) \in \mathbb{N}_0 \times \mathbb{N}_0$ *and for every* $(x, y) \in [0, 1]^2$

$$\lim_{t\to\infty} \mathbb{E}^{x,y}[X(t)^n Y(t)^m] = \mathbb{P}_{n,m}\left\{\lim_{t\to\infty}(N(t), M(t)) = (0, 0)\right\}.$$

Proof. Fix $(x, y) \in [0, 1]^2$ and $(n, m) \in \mathbb{N}_0 \times \mathbb{N}_0$. Then

$$\begin{aligned}\lim_{t\to\infty} \mathbb{E}^{x,y}[X(t)^n Y(t)^m] &= \lim_{t\to\infty} \mathbb{E}^{x,y}\big[\underbrace{X(t)^n Y(t)^m \mathbb{1}_{\mathbb{N}_0\times\mathbb{N}_0}(n, m)}_{=S((X(t),Y(t)),(n,m))}\big] \\ &= \lim_{t\to\infty} \mathbb{E}_{n,m}\left[x^{N(t)}y^{M(t)}\mathbb{1}_{\mathbb{N}_0\times\mathbb{N}_0}(N(t), M(t))\right] \\ &= \mathbb{E}_{n,m}\left[\lim_{t\to\infty} x^{N(t)}y^{M(t)}\mathbb{1}_{\mathbb{N}_0\times\mathbb{N}_0}(N(t), M(t))\right] \\ &= \mathbb{P}_{n,m}\left\{\lim_{t\to\infty}(N(t), M(t)) = (0, 0)\right\},\end{aligned}$$

where the last three equalities follow from the duality in Lemma 3.2, bounded convergence and the fact that $(N(t), M(t))_{t\geq 0}$ is absorbed in $(0, 0)$ or (∂, ∂) *in finite time* $\mathbb{P}$-a.s., respectively. (We use the convention that $0^0 = 1$.) □

3.2 Recursive Formula for the mixed moments in stationarity

In this subchapter, we want to tackle the problem of characterizing the stationary distribution μ of our diffusion. We will see that it exists and is unique, except in pathological cases. Unfortunately, we cannot calculate the stationary distribution explicitly, but the following results show that we can characterize it by means of mixed moments:

Lemma 3.5. *For $l \in \mathbb{N}$, let $(X(t))_{t\geq 0}$ be the solution of* (5) *for some fixed parameters. Denote*

$$M_{\overline{n}} := \lim_{t\to\infty} \mathbb{E}\Big[\prod_{i=1}^{l}(X_i(t))^{n^i}\Big]$$

for all $\overline{n} = (n^1, \dots, n^l) \in \mathbb{N}^n$. Using an argument similar to the one used in Lemma 3.3 or in Proposition 3.4, one can easily ensure that this limit exists for all choices of parameters and starting distributions. Define $(N(t))_{t\geq 0} = (N_1(t), \dots, N_l(t))_{t\geq 0}$ as the multi-dimensional moment dual with mutation of $(X(t))_{t\geq 0}$, defined analogously as in the previous subchapter. Let $\bar{N}_{\overline{n}}$ be the rate at which the process $(N(t))_{t\geq 0}$ leaves the starting state $\overline{n}$. That is,

$$\bar{N}_{\overline{n}} = \sum_{i=1}^{l} n^i\big[u^i + \sum_{j\neq i} c^{ij}\big] + \sum_{i=1}^{l}\binom{n^i}{2}(\alpha^i)^2$$

with $u^i := u_1^i + u_2^i$ and the convention $\binom{1}{m} = \binom{0}{m} = 0$ for all m. Then the following recursion holds:

$$M_{\overline{n}} = \frac{1}{\bar{N}_{\overline{n}}}\Big[\sum_{i=1}^{l} n^i\big[u_2^i M_{\overline{n}-e_i} + \sum_{j\neq i} c^{ij} M_{\overline{n}-e_i+e_j}\big] + \sum_{i=1}^{l}\binom{n^i}{2}(\alpha^i)^2 M_{\overline{n}-e_i}\Big]. \tag{8}$$

Proof. Consider the process $(N(t))_{t\geq 0}$ with starting condition $N(0) = \overline{n}$ and let $\tau = \inf\{t > 0 : N(t) \neq \overline{n}\}$. For any $\overline{n} \in E$ such that $\overline{n} \notin \{(0,0,...,0), (\partial, \dots, \partial)\}$, τ is an almost surely finite stopping time. Then we can apply Proposition 3.4 and write

$$\begin{aligned}
M_{\overline{n}} =& \mathbb{P}_{\overline{n}}\{\lim_{t\to\infty} N(t) = (0,0,\dots,0)\}\\
=& \sum_{\overline{m}\in E} \mathbb{P}_{\overline{n}}\{\lim_{t\to\infty} N(t) = (0,0,...,0) | N(\tau) = \overline{m}\}\mathbb{P}_{\overline{n}}\{N(\tau) = \overline{m}\}\\
=& \sum_{\overline{m}\in E} M_{\overline{m}}\mathbb{P}_{\overline{n}}\{N(\tau) = \overline{m}\}.
\end{aligned}$$

Writing explicitly the values of $\mathbb{P}_{\overline{n}}\{N(\tau) = \overline{m}\}$ leads to the statement. □

Remark 3.6. In the case of the two-island diffusion (2), we can write the recursive formula in a nicer way:

Lemma 3.7. *Assume $u_1 + u_2 + u_1' + u_2' > 0$. Let $(X_1, X_2) = (X, Y)$, which is defined as the solution to* (2) *from Chapter 2. Then $M_{0,0} = 1$ and the following recursion holds for all $(n, m) \in \mathbb{N}_0 \times \mathbb{N}_0 \setminus \{(0,0)\}$:*

$$M_{n,m} = \frac{1}{D_{n,m}}\left(a_n M_{n-1,m} + a_m' M_{n,m-1} + cnM_{n-1,m+1} + c'mM_{n+1,m-1}\right), \tag{9}$$

where

$$\begin{aligned}
a_n &:= \alpha^2\binom{n}{2} + nu_2, \\
a'_m &:= (\alpha')^2\binom{m}{2} + mu'_2, \text{ and} \\
D_{n,m} &:= \alpha^2\binom{n}{2} + (\alpha')^2\binom{m}{2} + (u_2+u_1)n + (u'_1+u'_2)m + cn + c'm,
\end{aligned}$$

with the convention that $c' = Kc$, $\binom{1}{2} = 0$, *and* $M_{-1,k} = M_{k,-1} = 0$ *for any* $k \in \mathbb{N}$.

An interesting application of our moment duality formula is the characterization of the long term behavior of $(X(t), Y(t))_{t\geq 0}$ in the neutral case:

Proposition 3.8. *Let* $u_1, u_2, u'_1, u'_2, \alpha, \alpha' \geq 0$, $c, c' > 0$. *Assume* $u_1+u_2+u'_1+u'_2 > 0$ *in* (2). *Then the diffusion is ergodic in the sense that*

$$\mathbb{P}^{x,y}\left\{(X(t),Y(t)) \in \cdot\,\right\} \xrightarrow{w} \mu, \quad \text{for } t \to \infty,$$

with μ *the unique invariant distribution and for all starting points* $(x,y) \in [0,1]^2$, *where* $\xrightarrow{w}$ *denotes weak convergence of measures. Furthermore* μ *is characterized by*

$$\forall\, n,m \in \mathbb{N}_0: \qquad \int_{[0,1]^2} x^n y^m \, d\mu(x,y) \;=\; M_{n,m}. \tag{10}$$

Proof. The unique solvability of the moment problem on $[0,1]^2$ yields existence of a unique distribution μ such that (10) which in particular implies

$$1 = M_{0,0} = \int_{[0,1]^2} x^0 y^0 \mathrm{d}\mu(x,y) = \mu([0,1]^2).$$

From the definition of the $M_{n,m}, n, m \in \mathbb{N}_0$, we know that

$$\begin{aligned}
&\lim_{t\to\infty} \int_{[0,1]^2} p(x,y)\mathrm{d}\mathbb{P}^{\bar{x},\bar{y}}\left\{(X(t),Y(t)) \in \,\cdot\,\right\} \\
&\qquad = \lim_{t\to\infty} \mathbb{E}^{\bar{x},\bar{y}}[p(X(t),Y(t))] = \int_{[0,1]^2} p(x,y)\mathrm{d}\mu(x,y)
\end{aligned}$$

for any polynomial p on $[0,1]^2$ (and any $(\bar{x},\bar{y}) \in [0,1]^2$). Since the polynomials are dense in the set of continuous (and bounded) functions on $[0,1]^2$ we can conclude that

$$\mathbb{P}^{x,y}\left\{(X(t),Y(t)) \in \cdot\,\right\} \xrightarrow{w} \mu.$$

□

In the case where we have multiple islands (Equation (5) in Subchapter 1.2.1), we can give an alternative proof of the general result on stationary distributions which is based on matrix theory:

Proposition 3.9. *Suppose that it is possible to go from any island to any other island in finite time with positive probability, i.e. that all islands are connected. Then:*
1) The seed bank diffusion admits a unique stationary distribution μ if at least one mutation rate among the u_1^i's and the u_2^i's is non-zero.
2) In case there is purely directional mutation (that is, either all the u_1^i's or all the u_2^i's are zero), the Dirac delta in the corresponding absorbing point[10] *is the unique stationary distribution.*
3) If all mutation rates are zero, for every $\gamma \in [0,1]$ there is a stationary distribution given by $\gamma\delta_{(1,1,\ldots,1)} + (1-\gamma)\delta_{(0,0,\ldots,0)}$.

Proof. First, we can suppose, without loss of generality, that at least one coalescent-relative population size α^i is non-zero. (Else, no coalescence would be possible.)
The strategy of the proof is as follows: first, we show that all the $M_{\bar{n}}$s are uniquely determined. Given this, the uniqueness of the stationary distribution follows from Hausdorff's moment problem.
First of all, observe that the recursive formula gives $M_{\bar{n}}$ as a function of the $M_{\bar{n}-e_i}$'s and the $M_{\bar{n}+e_i-e_j}$'s only. From now on, let us denote by $E_{\bar{n}}$ the equation which is obtained by plugging $\bar{n}$ into the recursive formula. Our strategy is then to calculate the $M_{\bar{n}}$'s starting from those for which $|\bar{n}| := \sum_i n_i = 1$. This is obtained first by solving the equation system made out of

$$E_{(1,0,\ldots,0)}, E_{(0,1,\ldots,0)}, \ldots, E_{(0,\ldots,0,1)}$$

and then continuing in the same fashion for $|\bar{n}| = 2, 3, \ldots$ until desired. That is, in the k-th step of the algorithm, we set up the equation system P_k made of all $E_{\bar{n}}$ with $|\bar{n}| = k$. Notice that by then, the $M_{\bar{n}-e_i}$'s will be already known (provided the recursion is closed). Hence, the number of unknowns of P_k is equal to its number of equations, namely the cardinality of the set of all $\bar{n}$ for which[11] $|\bar{n}| = k$.
Therefore, we just have to prove that, for every natural number k, P_k admits a unique solution. Since the number of equations and unknowns of P_k is the same, it has a unique solution if and only if the coefficient matrix, which we will denote by W^k, is invertible. The matrix W^k is given by

$$W^k_{\bar{n},\bar{m}} = \begin{cases} \bar{N}_{\bar{n}} & \text{for } \bar{n} = \bar{m}, \\ -c^{ij} & \text{for } \bar{n} - e_i + e_j = \bar{m}, \\ 0 & \text{else} \end{cases} \tag{11}$$

with the convention $\binom{1}{m} = \binom{0}{m} = 0$ for all m. Of course, every row of the matrix 'stands for' exactly one vector of length n adding up to k.
Observe that all rows of W^k are weakly dominant[12], since

$$\bar{N}_{\bar{n}} - \sum_{i,j} c^{ij} = \sum_i n^i u^i + \sum_i \binom{n^i}{2} \alpha^i \geq 0.$$

[10] That is, $(1,1,\ldots 1)$ if all u_1^i's are zero and $(0,0,\ldots 0)$ if all u_2^i's are zero.

[11] Using a "balls into boxes argument", we can find out that this system in made of $\binom{k+n-1}{k}$ equations

[12] We recall that the k-th row of a square matrix M_{ij} is called weakly dominant if $|M_{kk}| \geq \sum_{j \neq k} |M_{kj}|$ and strictly dominant if the corresponding strict inequality holds; the matrix itself is weakly/strictly dominant if all of its rows are weakly/strictly dominant.

From now on, we suppose without loss of generality that $\alpha^1 \neq 0$.
Now we want to know under which conditions W^k has at least one strictly dominant row. This is the case when $k > 1$, because then in the row linked to the equation $E_{ke_1}(= E_{(k,0,\dots,0)})$,

$$\binom{n^1}{2}\alpha^1 = \binom{k}{2}\alpha^1 > 0.$$

Moreover, this is the case if there is mutation: if, for instance, $u^i \neq 0$, then in the row linked to the equation E_{ke_i},

$$\sum_i n^i u^i = ku^i > 0.$$

However, if $k = 1$ and $u^i = 0$ for all i, then n^i is always either 0 or 1 and thus by convention $\binom{n^i}{2} = 0$. This means that all rows of W^1 are zero-sum, since

$$\sum_i n^i u^i + \sum_i \binom{n^i}{2}\alpha^i = 0.$$

These remarks allow us to prove all three parts of the theorem:

1) Let at least one mutation rate u^i_j ($i \in \{1,\dots,n\}$, $j \in \{1,2\}$) be non-zero, which ensures the presence of strictly dominant rows. Then, we can use a corollary of the Levy-Desplanques theorem ([JH85], Theorem 6.2.27) which says that in a weakly dominant irreducible matrix with at least one strictly dominant row the determinant is non-zero. But the matrix is irreducible since we have supposed that all islands are connected, which concludes the first part of our proof.

2) For the one-directional mutation case, the fact that there is a single stationary distribution holds due to part 1 of the proof. Moreover, it is the Dirac delta in one absorbing point, as we can show:
Suppose $u^i_2 = 0$ for all i's. Given the uniqueness of the solution in our case and the fact that the diffusion is concentrated in $[0,1]^n$, it is enough to prove that the recursive formula for $k = 1$ is satisfied if $M_{e_i} = 0$ for all i's. In our case, the recursive formula is:

$$M_{e_i}(u^i + \sum_{j\neq i} c^{ij}) = \sum_{j\neq i} c^{ij} M_{e_j},$$

and $M_{e_i} = 0$ for all i's is a (and therefore the only) solution. If $u^i_1 = 0$ for all i's, then $u^i_2 = u^i$ and the recursive formula says

$$M_{e_i}(u^i + \sum_{j\neq i} c^{ij}) = \sum_{j\neq i} c^{ij} M_{e_j} + u^i,$$

which has as a solution $M_{e_i} = 1$ for all i's.
3) This is proved in much the same way as Lemma 3.3. □

3.3 A stochastic delay differential equation

Note that the only source of randomness in the two-dimensional system (1) and also in its generalization to many seed banks (4) is the one-dimensional Brownian motion $(W(t))_{t\geq 0}$ driving the fluctuations in the active population. This, together with the special form of the seed bank(s), allows us to reformulate this system as an essentially one-dimensional stochastic delay differential equation. Recall the notation from Remark 1.5 and abbreviate, for convenience, $u^i := u_1^i + u_2^i$, $i \in \{1, \dots, k\}$.

Proposition 3.10. *The solution to* (4) *with initial values* $x, y_1, \dots, y_k \in [0,1]$ *is a.s. equal to the solution of the unique strong solution of system of* stochastic delay differential equations

$$\begin{aligned} dX(t) = & \sum_{i=1}^{k} c_i \bigg(y_i e^{-(u^i+K_i c_i)t} + \int_0^t e^{-(u^i+K_i c_i)(t-s)}(u_2^i + K_i c_i X(s))\, ds - X(t) \bigg)\, dt \\ & + \big[- u_1 X(t) + u_2(1 - X(t)) \big]\, dt + \sqrt{X(t)(1 - X(t))}\, dW(t), \\ dY_i(t) = & \bigg(- y_i(u^i + K_i c_i) e^{-(u^i+K_i c_i)t} \\ & - (u^i + K_i c_i) \int_0^t e^{-(u^i+K_i c_i)(t-s)}(u_2^i + K_i c_i X(s))\, ds + u_2^i + K_i c_i X(t) \bigg)\, dt, \end{aligned} \tag{12}$$

for $i \in \{1, \dots, k\}$ *with the same initial condition.*

Proof of Proposition 3.10. Let $(X(t), Y_1(t), \dots, Y_k(t))_{t\geq 0}$ be the unique strong solution of (4). Recall e.g. from [RY99, Proposition 3.1] that for continuous semimartingales Z, W, we have the integration by parts formula,

$$\int_0^t W(s) dZ(s) = W(t)Z(t) - Z(0)W(0) - \int_0^t Z(s) dW(s) - \langle Z, W \rangle(t),$$

where $\langle \cdot, \cdot \rangle$ denotes the covariance process and $t \geq 0$. Note that for every differentiable deterministic function f, since $\langle Z, f \rangle \equiv 0$, this reduces to

$$f(t)Y_i(t) - f(0)Y_i(0) = \int_0^t f(s) \mathrm{d}Y_i(s) + \int_0^t f'(s) Y_i(s) \mathrm{d}s.$$

Substituting the expression for $\mathrm{d}Y_i(t)$ from (4), we obtain that

$$\begin{aligned} f(t)Y_i(t) - f(0)Y_i(0) = & \int_0^t f(s) \big[- u_1^i Y_i(s) + u_2^i(1 - Y_i(s)) + K_i c_i (X(s) - Y_i(s)) \big] \mathrm{d}s \\ & + \int_0^t f'(s) Y_i(s) \mathrm{d}s. \end{aligned} \tag{13}$$

Letting $f(t) := e^{(u^i+K_i c_i)t}, t \geq 0$, equation (13) simplifies to

$$f(t)Y_i(t) - f(0)Y_i(0) = \int_0^t e^{(u^i+K_i c_i)s}(K_i c_i X(s) + u_2^i) \mathrm{d}s.$$

This can be rewritten given the initial value $y_i = Y_i(0)$ as

$$e^{(u^i+K_ic_i)t}Y_i(t) = y_i + \int_0^t e^{(u^i+K_ic_i)s}(K_ic_iX(s) + u_2^i)\mathrm{d}s.$$

By dividing on both sides by $e^{(u^i+K_ic_i)t}$ we finally get

$$Y_i(t) = y_ie^{-(u^i+K_ic_i)t} + \int_0^t e^{(u^i+K_ic_i)(s-t)}(K_ic_iX(s) + u_2^i)\mathrm{d}s. \tag{14}$$

Plugging this into the first line of the system in (4) proves that the unique strong solution of (4) is a strong solution to (12). On the converse, let $(X(t), Y_1(t), \dots, Y_k(t))_{t\geq 0}$ now be a solution to (12). (We already know that there exists at least one.) Using (14) we immediately see that $(X(t))_{t\geq 0}$ solves the first equation in (4). Likewise, using (14) in the right-hand side of the last k equations in (12), we obtain the last k equations of (4). Since (4) has a unique solution, this must then hold for (12), too, and the two solutions coincide $\mathbb{P}$-almost surely.

□

Note that the the first equation in (12) does not depend on $Y_i, i = 1, \dots, k$, and that the latter equations for the Y_i are in turn deterministic functions of X, so that the system of SDDEs is essentially one-dimensional.

Remark 3.11. Let us consider an interesting special case of the above result to reveal its structure: It is an immediate corollary from the above that the seed bank diffusion solving (1) with parameters $c = 1$, $K = 1$, $u_1 = u_2 = u_1' = u_2' = 0$, started in $X(0) = x = y = Y(0) \in [0, 1]$ is a.s. equal to the unique strong solution of the stochastic delay differential equations

$$\begin{aligned}
\mathrm{d}X(t) &= \Big(xe^{-t} + \int_0^t e^{-(t-s)}X(s)\mathrm{d}s - X(t)\Big)\mathrm{d}t + \sqrt{X(t)(1-X(t))}\mathrm{d}W(t),\\
\mathrm{d}Y(t) &= \Big(-ye^{-t} - \int_0^t e^{-(t-s)}X(s)\mathrm{d}s + X(t)\Big)\mathrm{d}t,
\end{aligned} \tag{15}$$

with the same initial condition. This now provides an elegant interpretation of the *delay* in the SDDE induced by the seed bank. Indeed, it shows that the *type* (a or A) of any “infinitesimal” resuscitated individual, is determined by the *active* population present an *exponentially distributed time ago* (with a cutoff at time 0), *which the individual spent dormant in the seed bank.* The net effect is positive if the frequency of a-alleles at that time was higher than the current frequency, and negative if it was lower. This is the forward-in-time equivalent of the model for seed banks or dormancy in the coalescent context as formulated in [KKL01], where the seed bank is modeled by having individuals first choose a generation in the past according to some measure μ and then choosing their ancestor uniformly among the individuals present in that generation. The seed bank model given in [BGKWB16] is obtained when μ is chosen to be geometric, i.e. memoryless, like the exponential distribution. This indicates that a forward-in-time model for more general dormancy models is to be searched among SDDEs rather than among SDEs.

Such a reformulation is of course not feasible for the two island model, which is driven by two independent sources of noise.

3.4 Non-reversibility

Given the existence of an invariant distribution, the question of reversibility arises naturally. The classical Wright-Fisher frequency process is reversible. However, the diffusion process of the two-island model is not, as shown in [KZH08] (Theorem 4). It turns out, that the seed bank diffusion with mutation (1) is not reversible in general, either:

Proposition 3.12. *If $c, u_1, u_2 \neq 0$ and $u_1' = u_1$, $u_2' = u_2$, then the seed bank diffusion process is not reversible.*

Proof. We recall that a process (X, Y) with stationary distribution μ is called reversible if for all $f, g \in \mathcal{D}(\mathcal{A})$,

$$\mathbb{E}^\mu[f((X,Y)(\cdot))\mathcal{A}g((X,Y)(\cdot))] = \mathbb{E}^\mu[g((X,Y)(\cdot))\mathcal{A}f((X,Y)(\cdot))].$$

where $\mathcal{A}$ is defined as the generator of the process. We recall that the generator of the seed bank diffusion is equal to

$$\mathcal{A}^{(1)}(f)(x,y) = [u_2 - (c + u_1 + u_2)x + cy]\frac{\partial}{\partial x}f(x,y)$$
$$+[u_1 - (Kc + u_1 + u_2)y + Kcx]\frac{\partial}{\partial y}f(x,y) + \frac{x(1-x)}{2}\frac{\partial^2}{\partial x^2}f(x,y),$$

and that its domain $\mathcal{D}(\mathcal{A}^{(1)})$ contains $C^2([0,1]^2)$, the space of twice differentiable functions. Assuming, by contradiction, that the process is reversible and plugging in $f(x,y) = x$ and $g(x,y) = y$, we get

$$u_2\mathbb{E}^\mu[X - Y] + c\mathbb{E}^\mu[-(K-1)XY - Y^2 + KX^2] = 0.$$

But using the recursive formula (9) from Chapter 3.2, from which we take the notation, we get[13]:

- by solving the equation system

$$\begin{cases} M_{1,0} &= \frac{1}{D_{1,0}}\big(a_1 + cM_{1,0}\big) \\ M_{0,1} &= \frac{1}{D_{0,1}}\big(a_1' + cKM_{0,1}\big), \end{cases}$$

that the first term is 0;

- by solving the equation system

$$\begin{cases} M_{2,0} &= \frac{1}{D_{2,0}}\big(a_2M_{1,0} + 2cM_{1,1}\big) \\ M_{1,1} &= \frac{1}{D_{1,1}}\big(a_1M_{0,1} + a_1'M_{1,0} + cM_{0,2} + KcM_{2,0}\big) \\ M_{0,2} &= \frac{1}{D_{0,2}}\big(a_2'M_{0,1} + 2KcM_{1,1}\big), \end{cases}$$

that $M_{20} > M_{11} > M_{02}$, which ensures that

$$(K-1)M_{11} + M_{02} < (K-1)M_{11} + M_{11} = KM_{11} < KM_{20}$$

and, as a consequence, that the second term is always positive. This contradicts our hypothesis. □

[13] We omit the details since this method of calculating the mixed moments at stationarity recursively will be thoroughly explained in Chapter 6.

4 Boundary behavior

In this chapter, we investigate the boundary behavior of the seed bank and the two island diffusion.

Just like there seems to be no explicit characterization of the stationary distribution for the diffusion models introduced in Chapter 1, a full boundary classification was still lacking. In this case, the standard Feller approach via speed measure and scale function (see Chapter 2) breaks down, since the two island model leads to a two-dimensional diffusion.

First, we give two straightforward results. The first one concerns the presence of atoms at the boundaries in the stationary distribution, ensuring there are none. The second one gives us a partial boundary classification by means of comparisons with stochastic processes where one coordinate is defined as constant.

Then, we move on towards our main result, which gives a full classification of all boundaries. It can be proved in two possible ways: one, using a technique called McKean's argument, which is based on the martingale convergence theorem on stochastic intervals and is suitable also in multi-dimensional settings. Two, adapting Lyapunov's argument, which uses the infinitesimal generator of a stochastic process to infer whether it can reach the boundary in finite time, to our case, where we have a compact state space. Either of these two arguments gives a necessary condition for any boundary not being reached in finite time almost surely. The corresponding sufficient condition, which completes our analysis of both models, is proved using a result from the theory of polynomial diffusions which we already stated in Chapter 2. This chapter is based on [K1] (Subchapters 1-3).

4.1 Atoms on the boundaries

We begin the observation that in the presence of mutation the marginals of the stationary distribution μ of the general diffusion (2) have no atoms at the boundaries, extending the analogous observations made for the two-island model in [KZH08] with use of the same argument.

Proposition 4.1. *Let $(X(t), Y(t))_{t\geq 0}$ be the solution to (2). Assume*

$$u_1 u_2 u_1' u_2' > 0$$

and recall that μ denotes the unique invariant distribution of $(X(t), Y(t))_{t\geq 0}$. Then, for any $t > 0$, we have

$$\mathbb{P}^{\mu}\left\{X(t) \in \{0,1\}\right\} = \mathbb{P}^{\mu}\left\{Y(t) \in \{0,1\}\right\} = 0.$$

Proof. For convenience, we only prove the statement for the seed bank diffusion with $u_1 = u_1'$ and $u_2 = u_2'$; the generic proof is almost the same. Notice that, by construction of μ (see (10)) and since it is the invariant measure of $(X(t), Y(t))_{t\geq 0}$

$$M_{n,m} = \mathbb{E}^{\mu}[X^n(t) Y^m(t)]$$

for any $t \geq 0$. Recall the recursion from Lemma 3.7 for $m = 0$:

$$M_{n,0} = \frac{1}{D_{n,0}} \left(a_n M_{n-1,0} + cn M_{n-1,1} \right) \leq M_{n-1,0} \frac{a_n + cn}{D_{n,0}},$$

since $M_{n,m+1} \leq M_{n,m}$ for any $n, m \in \mathbb{N}_0$. Iterating this observation yields

$$M_{n,0} \leq \underbrace{M_{0,0}}_{=1} \prod_{k=1}^{n} \frac{a_k + ck}{D_{k,0}} = \prod_{k=1}^{n} \frac{\alpha^2 \binom{k}{2} + k u_2 + ck}{\alpha^2 \binom{k}{2} + (u_2 + u_1)k + ck}.$$

Since we assumed in particular that $u_1 > 0$, one can check that $M_{n,0} \to 0$ for $n \to \infty$ which in turn implies

$$\mathbb{P}^{\mu} \left\{ X(t) = 1 \right\} = 0.$$

The other three cases are analogous. □

4.2 Results via comparison

Differences between the seed bank diffusion and the two island model become immediately visible with respect to accessibility of boundary points and absorption in finite time. This correspondence remains in the following, more detailed description of the boundary behavior of the seed bank diffusion as in (1). With the help of the definitions from Chapter 2.2, we can state the next result, giving an incomplete boundary classification:

Proposition 4.2.
In the seed bank diffusion model (1):
a) If $u_2 > \frac{1}{2}$, then X will never hit 0 started from the interior. If $u_2 + c < \frac{1}{2}$, the boundary 0 is accessible for X.
b) If $u_1 > \frac{1}{2}$, then X will never hit 1 started from the interior. If $u_1 + c < \frac{1}{2}$, the boundary 1 is accessible for X.
c) Y will never hit 1 and 0 from the interior.

The intermediate case remains - for now - open.

Proof. a)-b) Let us consider the system of SDE's

$$\begin{aligned} \mathrm{d}X_1(t) &= \big[- u_1 X_1(t) + u_2(1 - X_1(t)) + c(X_2(t) - X_1(t)) \big] \mathrm{d}t \\ &\qquad\qquad\qquad\qquad + \sqrt{X_1(t)(1 - X_1(t))} \mathrm{d}W(t), \\ X_2(t) &\equiv 0. \end{aligned} \tag{16}$$

with $(X_1(0), X_2(0)) = (X(0), Y(0))$. Then we have (obviously) that $X_2(t) \leq Y(t)$ for all $t \geq 0$ and, applying the Ikeda-Watanabe Theorem (Theorem 43 in [RW00]) on the first SDE and using the fact that, according to the Yamada-Watanabe theorem our SDE admits a unique strong solution (so that we can use the same noise), we get that

$X_1 \leq X$ a.s.
But we can also write the first SDE as

$$\mathrm{d}X_1(t) = \big[-(u_1+c)X_1(t) + u_2(1-X_1(t))\big]\mathrm{d}t + \sqrt{X_1(t)(1-X_1(t))}\mathrm{d}W(t),$$

and this is the SDE linked to the Wright-Fisher diffusion with mutation (and parameters u_1+c and u_2). Therefore, if the boundary $\{0\}$ is not accessible for X_1, the same holds for X. In Chapter 2, we found out that the lower boundary is not accessible if $u_2 > \frac{1}{2}$. Moreover, we can say that if the boundary $\{0\}$ is accessible for X_1, then it is accessible for X as well; this happens when $u_1 + c < \frac{1}{2}$.
To find a similar result for the boundary $\{1\}$, the proof works in the same way; consider the system of SDEs

$$\begin{aligned} \mathrm{d}Y_1(t) &= \big[-u_1Y_1(t) + u_2(1-Y_1(t)) + c(Y_2(t)-Y_1(t))\big]\mathrm{d}t \\ &\qquad\qquad\qquad\qquad + \sqrt{Y_1(t)(1-Y_1(t))}\mathrm{d}W(t), \\ Y_2(t) &\equiv 1. \end{aligned} \tag{17}$$

Then the same machinery works (obviously with inverted inequalities) and the first SDE can be written as

$$\mathrm{d}Y_1(t) = \big[-u_1Y_1(t) + (u_2+c)(1-Y_1(t))\big]\mathrm{d}t + \sqrt{Y_1(t)(1-Y_1(t))}\mathrm{d}W(t),$$

which is the SDE linked to the Wright-Fisher diffusion with mutation (and parameters u_1 and u_2+c). Therefore, if the boundary $\{1\}$ is not accessible for Y_1, the same holds for X; this is the case when $u_1 > \frac{1}{2}$. As in the previous case, from the analysis of this SDE we conclude also that the boundary $\{1\}$ is accessible if $u_2 + c < \frac{1}{2}$.
c) The system of SDEs which has to be considered in this case is

$$\begin{aligned} \mathrm{d}Z_2(t) &= \big[-u_1'Z_2(t) + u_2'(1-Z_2(t)) + Kc(Z_1(t)-Z_2(t))\big]\mathrm{d}t \\ Z_1(t) &\equiv 0. \end{aligned} \tag{18}$$

The first one can be written as

$$\mathrm{d}Z_2(t) = \big[-(u_1'+Kc)Z_2(t) + u_2'(1-Z_2(t))\big]\mathrm{d}t.$$

The stationary distribution of this ODE is simply a Dirac delta in its mean, which is equal to

$$\frac{u_2'}{u_1'+u_2'+Kc}.$$

Therefore, for every (non-zero) choice of parameters, by comparison we find out that the boundary $\{0\}$ is not accessible for Y. If, on the other side, $Z_1(t) \equiv 1$, the first SDE becomes

$$\mathrm{d}Z_2(t) = \big[-u_1'Z_2(t) + (u_2'+Kc)(1-Z_2(t))\big]\mathrm{d}t,$$

and all the previous considerations hold again, the mean being now equal to

$$\frac{u_2'+Kc}{u_1'+u_2'+Kc}.$$

□

4.3 Main Results

The previous subchapter illustrates that actually, each of the parameters u_1, u_2, u_1', u_2' is responsible for the value of exactly one of the probabilities in Proposition 4.1. We will see that this principle still holds true for the main result. We will use the notation introduced in the statement of Theorem 2.7.

Theorem 4.3. *Let* $(X(t), Y(t))_{t\geq 0}$ *be the solution to* (2) *with* $u_1, u_2, u_1', u_2', \alpha, \alpha' \geq 0$, $c, c' > 0$. *Moreover, assume that for the starting distribution* μ_0,

$$\mathbb{P}^{\mu_0}\left\{(X(0), Y(0)) \in (0,1)^2\right\} = 1.$$

Then,

(i) X *will never hit 0 started from the interior if and only if* $2u_2 \geq \alpha^2$.

(ii) X *will never hit 1 started from the interior if and only if* $2u_1 \geq \alpha^2$.

(iii) Y *will never hit 0 started from the interior if and only if* $2u_2' \geq (\alpha')^2$.

(iv) Y *will never hit 1 started from the interior if and only if* $2u_1' \geq (\alpha')^2$.

Remark 4.4 (Strategy of the proof)**.** We will actually prove a slightly stronger statement for the 'only if' direction. We state the true statement for the case of (i):

Let $2u_2 < \alpha^2$. Then, for any $s > 0$ there exists an $\varepsilon > 0$ such that

$$\mathbb{P}\left\{\|(X(0), Y(0)) - (0,0)\| < \varepsilon\right\} = 1 \quad \Rightarrow \quad \mathbb{P}\left\{\tau_0^X \leq s\right\} > 0,$$

where τ_0^X is equal to the first hitting time in $\{0\} \times [0,1]$ and as such a "τ_B^X" as defined at the beginning of Subchapter 2.2.

One direction of this result can be obtained by viewing our diffusion in the context of *polynomial diffusions*, which we introduced in Chapter 2.3, and using the result shown there. The other direction will be proved with the help of the McKean argument, which we have shown in subchapter 2.2.

Proof of Theorem 4.3. and begin with a short observation helpful for both parts of the proof.

Take $p \in \mathcal{P} := \{(x,y) \mapsto x, (x,y) \mapsto 1-x, (x,y) \mapsto y, (x,y) \mapsto 1-y\}$. Define

$$\tau_p := \inf\{t \geq 0 \mid p(X(t), Y(t)) = 0\}.$$

Note that each of the τ_p is a τ_B^X as well, hence, we want to prove that $\mathbb{P}^{\mu_0}\{\tau_p < \infty\}$ is either zero or strictly positive, depending on the parameters.
Let $p_0(x,y) := x \in \mathcal{P}$. For $h_{p_0}(x,y) := (\alpha^2(1-x), 0)^T$ we have

$$\begin{aligned} a\nabla p_0(x,y) &= \begin{pmatrix} \alpha^2 x(1-x) & 0 \\ 0 & (\alpha')^2 y(1-y) \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &= x \begin{pmatrix} \alpha^2(1-x) \\ 0 \end{pmatrix} \\ &= p_0(x,y) h_{p_0}(x,y). \end{aligned}$$

Similarly, let $p_1(x,y) := 1 - x \in \mathcal{P}$. For $h_{p_1}(x,y) := (-\alpha^2 x, 0)^T$ we have

$$\begin{aligned} a\nabla p_1(x,y) &= \begin{pmatrix} \alpha^2 x(1-x) & 0 \\ 0 & (\alpha')^2 y(1-y) \end{pmatrix} \begin{pmatrix} -1 \\ 0 \end{pmatrix} \\ &= (1-x) \begin{pmatrix} -\alpha^2 x \\ 0 \end{pmatrix} \\ &= p_1(x,y) h_{p_1}(x,y). \end{aligned}$$

Part 1: We begin proving the 'only if' statements, as they rely on Theorem 2.7. Let $\bar{z} := (0,0) \in \{p_0 = 0\}$. Then

$$\mathcal{A}^{(1)} p_0(\bar{z}) = u_2 \geq 0$$

and

$$2\mathcal{A}^{(1)} p_0(\bar{z}) - h_{p_0}(\bar{z})^T \nabla p_0(\bar{z}) = 2u_2 - \alpha^2 < 0$$

where the latter holds if and only if $2u_2 < \alpha^2$. Hence the 'only if' in (i) follows by Theorem 2.7.

In the same way, let $\bar{z} := (1,1) \in \{p_1 = 0\}$. Then

$$\mathcal{A}^{(1)} p(\bar{z}) = u_1 \geq 0$$

and

$$2\mathcal{A}^{(1)} p(\bar{z}) - h_p(\bar{z})^T \nabla p(\bar{z}) = 2u_1 - \alpha^2 < 0$$

and again, the latter holds if and only if $2u_1 < \alpha^2$. Therefore, the 'only if' in (ii) follows from Theorem 2.7 as well.

The analogous statements in (iii) and (iv) hold by symmetry.

Part 2: We now turn to the proof of the 'if' statements, which is more involved and uses a type of McKean's argument as in [MPS11], Section 4.1, as it follows closely the proof of Proposition 2.2 in [LP17].

Itō's formula applied to the function $r(x) := \log(p(x))$ gives us for any $t < \tau_p$:

$$\begin{aligned}
&\log p(X(t),Y(t)) = \log p(X(0),Y(0)) \\
&\quad + \int_0^t \partial_1(\log(p(X(s),Y(s)))\mathrm{d}X(s) + \int_0^t \partial_2(\log(p(X(s),Y(s)))\mathrm{d}Y(s) \\
&\quad + \frac{1}{2}\int_0^t \partial_1^2(\log(p(X(s),Y(s)))\mathrm{d}[X](s) + \int_0^t \partial_1\partial_2(\log(p(X(s),Y(s)))\mathrm{d}[X,Y](s) \\
&\quad + \frac{1}{2}\int_0^t \partial_2^2(\log(p(X(s),Y(s)))\mathrm{d}[Y](s) \\
&= \log p(X(0),Y(0)) + \int_0^t \frac{\partial_1 p(X(s),Y(s))}{p(X(s),Y(s))}\mathrm{d}X(s) + \int_0^t \frac{\partial_2 p(X(s),Y(s))}{p(X(s),Y(s))}\mathrm{d}Y(s) \\
&\quad + \frac{1}{2}\int_0^t \frac{\partial_1^2 p(X(s),Y(s)) - (\partial_1 p(X(s),Y(s)))^2}{(p(X(s),Y(s))^2}\mathrm{d}[X](s) \\
&\quad + \frac{1}{2}\int_0^t \frac{\partial_2^2 p(X(s),Y(s)) - (\partial_2 p(X(s),Y(s)))^2}{(p(X(s),Y(s))^2}\mathrm{d}[Y](s) \\
&\quad + \frac{1}{2}\int_0^t \frac{\partial_1\partial_2 p(X(s),Y(s)) \cdot p(X(s),Y(s))}{(p(X(s),Y(s))^2}\mathrm{d}[X,Y](s) \\
&\quad - \frac{1}{2}\int_0^t \frac{\partial_1 p(X(s),Y(s))\partial_2 p(X(s),Y(s))}{(p(X(s),Y(s))^2}\mathrm{d}[X,Y](s).
\end{aligned}$$

Now, we can plug in $\mathrm{d}X(s)$ and $\mathrm{d}Y(s)$ from the SDE's (1). Moreover, for a multidimensional diffusion, the equation $\mathrm{d}[X](t) = \sigma^2(t)\mathrm{d}t$ ([KS98], Proposition 3.2.17) holds, which in our case translates to

$$\begin{pmatrix} \mathrm{d}[X](t) & \mathrm{d}[X,Y](t) \\ \mathrm{d}[X,Y](t) & \mathrm{d}[Y](t) \end{pmatrix} = \begin{pmatrix} \alpha^2 X(t)(1-X(t)) & 0 \\ 0 & \alpha'^2 Y(t)(1-Y(t)) \end{pmatrix}\mathrm{d}t.$$

So, we get

$$\begin{aligned}
\log\, &p(X(t),Y(t)) = \log p(X(0),Y(0)) \\
&+ \int_0^t \frac{\partial_1 p(X(s),Y(s))(-u_1 X(s) + u_2(1-X(s)) + c(Y(s)-X(s)))}{p(X(s),Y(s))}\mathrm{d}s \\
&+ \int_0^t \frac{\partial_1 p(X(s),Y(s))(\alpha^2 X(s)(1-X(s)))}{p(X(s),Y(s))}\mathrm{d}W_1(s) \\
&+ \int_0^t \frac{\partial_2 p(X(s),Y(s))(-u_1' Y(s) + u_2'(1-Y(s)) + Kc(X(s)-Y(s)))}{p(X(s),Y(s))}\mathrm{d}s \\
&+ \int_0^t \frac{\partial_2 p(X(s),Y(s))(\alpha'^2 Y(s)(1-Y(s)))}{p(X(s),Y(s))}\mathrm{d}W_2(s) \\
&+ \frac{1}{2}\int_0^t \frac{\partial_1^2 p(X(s),Y(s))(\alpha^2 X(s)(1-X(s)))}{(p(X(s),Y(s)))^2}\mathrm{d}s
\end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{2}\int_0^t \frac{(\partial_1 p(X(s),Y(s))^2(\alpha^2 X(s)(1-X(s)))}{(p(X(s),Y(s)))^2}\mathrm{d}s\\
&+\frac{1}{2}\int_0^t \frac{\partial_2^2 p(X(s),Y(s))(\alpha'^2 Y(s)(1-Y(s)))}{(p(X(s),Y(s)))^2}\mathrm{d}s\\
&-\frac{1}{2}\int_0^t \frac{(\partial_2 p(X(s),Y(s))^2(\alpha'^2 Y(s)(1-Y(s)))}{(p(X(s),Y(s)))^2}\mathrm{d}s
\end{aligned}
$$

We can see that the sum of the even-numbered lines yields

$$\int_0^t \frac{\mathcal{A}^{(1)}p(X(s),Y(s)}{p(X(s),Y(s))}\mathrm{d}s,^{14}$$

where $\mathcal{A}^{(1)}$ is the generator of the seed bank diffusion, by its definition.
Moreover, we can see that

$$
\begin{aligned}
&(\partial_1 p(X(s),Y(s))^2(\alpha^2 X(s)(1-X(s))) + (\partial_2 p(X(s),Y(s))^2(\alpha'^2 Y(s)(1-Y(s))) =\\
&\big(\partial_1 p(X(s),Y(s)),\quad \partial_2 p(X(s),Y(s))\big)\begin{pmatrix}\alpha^2 X(s)(1-X(s)) & 0\\ 0 & \alpha'^2 Y(s)(1-Y(s))\end{pmatrix}\times\\
&\times\begin{pmatrix}\partial_1 p(X(s),Y(s))\\ \partial_2 p(X(s),Y(s))\end{pmatrix} = \nabla p^T a \nabla p
\end{aligned}
$$

and

$$
\begin{aligned}
&\partial_1 p(X(s),Y(s)\alpha^2 X(s)(1-X(s))\mathrm{d}W(s) + \partial_2 p(X(s),Y(s)\alpha'^2 Y(s)(1-Y(s))\mathrm{d}W'(s) =\\
&\big(\partial_1 p(X(s),Y(s)),\quad \partial_2 p(X(s),Y(s))\big)\begin{pmatrix}\alpha^2 X(s)(1-X(s)) & 0\\ 0 & \alpha'^2 Y(s)(1-Y(s))\end{pmatrix}\times\\
&\times\begin{pmatrix}\mathrm{d}W(s)\\ \mathrm{d}W'(s)\end{pmatrix} = \nabla p^T\sigma^2\mathrm{d}W_2(s),
\end{aligned}
$$

where $W_2(s)$ is a 2-dimensional standard Brownian motion. In the end, this gives us

$$
\begin{aligned}
\log p(X(t),Y(t)) = \log p(X(0),Y(0))&\\
&+\int_0^t \left(\frac{\mathcal{A}^{(1)}p(X(s),Y(s))}{p(X(s),Y(s))} - \frac{1}{2}\frac{\nabla p^T a\nabla p(X(s),Y(s))}{p(X(s),Y(s))^2}\right)\mathrm{d}s\\
&+\int_0^t \frac{\nabla p^T\sigma^2(X(s),Y(s))}{p(X(s),Y(s))}\mathrm{d}W_2(s)
\end{aligned}
$$

and the identity $a\nabla p = h_p p$ yields

$$
\begin{aligned}
&= \log p(X(0),Y(0)) + \int_0^t \frac{2\mathcal{A}^{(1)}p(X(s),Y(s)) - \nabla p^T h_p(X(s),Y(s))}{2p(X(s),Y(s))}\mathrm{d}s\\
&\quad+\int_0^t \frac{\nabla p^T\sigma^2(X(s),Y(s))}{p(X(s),Y(s))}\mathrm{d}W_2(s).
\end{aligned}
$$

[14] This holds only because p is a linear function and thus its second derivative is zero.

Suppose now, we find a constant $\kappa_p > 0$ such that

$$2\mathcal{A}^{(1)}p(x,y) - h_p^T \nabla p(x,y) \geq -2\kappa_p p(x,y) \qquad \text{for all } (x,y) \in [0,1]^2. \tag{19}$$

Then for $M(t) := \log p(X(t), Y(t)) - \log p(X(0), Y(0)) + \kappa_p t$, we have

$$M(t) \geq M(0) + \int_0^t \frac{\nabla p^T \sigma^2(X(s), Y(s))}{p(X(s), Y(s))} \mathrm{d}W_2(s).$$

Since the right-hand side is a stochastic integral with respect to Brownian motion and thus a local martingale, M is a local submartingale on $[0, \tau_p)$; since p is a bounded function on $[0,1]^2$, M is also bounded from above on bounded time intervals. For details concerning stochastic processes on stochastic intervals see for example [Mai77].

This implies, plugging in $p(x,y) = x$, that

$$\log X(t) \geq \log X(0) - \kappa_p t + \int_0^t \alpha^2(1 - X(s))\mathrm{d}W(s).$$

Exponentiating, taking the minimum between t and τ_p and multiplying both sides with $\mathbb{1}_{\{\tau_p < \infty\}}$ yields

$$\mathbb{1}_{\{\tau_p<\infty\}} X(t \wedge \tau_p) \geq \exp(M(t \wedge \tau_p) - \kappa_p(t \wedge \tau_p))\mathbb{1}_{\{\tau_p<\infty\}}.$$

Assume now that $\mathbb{P}\{\tau_p < \infty\} > 0$. For $t \to \infty$, the left-hand side will converge almost surely to 0 by definition of τ_p and continuity of our process. Thus, $M(t\wedge\tau_p) - \kappa_p(t\wedge\tau_p)$ must converge to $-\infty$ for almost every path for which $\tau_p < \infty$ as $t \to \infty$. The only possibility for this is that $M(t \wedge \tau_p) \to -\infty$.

Because of this, we have that

$$\mathbb{1}_{\{\tau_p<\infty\}} M(t \wedge \tau_p) \to -\infty \mathbb{1}_{\{\tau_p<\infty\}}$$

almost surely for $t \to \tau_p$. But since $(M(t \wedge \tau_p))_{t \geq 0}$ is a continuous local submartingale as well, we can conclude, using the same method as in Proposition (2.2.2), that $\tau_p = \infty$ a.s. Therefore, the task left to do is to find suitable constants κ_p.

Recall $p_0(x,y) = x$ and the assumption in (i) that $2u_2 - \alpha^2 \geq 0$. Set

$$\kappa_0 := u_1 + u_2 + c - \frac{\alpha^2}{2} > 0$$

(since $c > 0$) and observe that then

$$\begin{aligned} 2\mathcal{A}^{(1)}p_0(x,y) - h_{p_0}^T \nabla p_0(x,y) &= x(-2u_1 - 2u_2 - 2c + \alpha^2) + y2c + 2u_2 - \alpha^2 \\ &\geq x(-2u_1 - 2u_2 - 2c + \alpha^2) + 2u_2 - \alpha^2 \\ &\geq -2\kappa_0 x = -2\kappa p_0(x,y) \qquad \text{for all } (x,y) \in [0,1]^2. \end{aligned}$$

Hence (19) holds for p_0 and since $\tau_{p_0} = \tau_0^X$, the proof of (i) is completed.

For (ii) we assumed $2u_1 - \alpha^2 \geq 0$ and will use $p_1(x,y) = 1 - x$. Set

$$\kappa_1 := u_2 + c > 0,$$

since then

$$\begin{aligned}2\mathcal{A}^{(1)}p_1(x,y) - h_{p_1}^T \nabla p_1(x,y) &= x(2u_1 + 2u_2 + 2c - \alpha^2) - y2c - 2u_2 \\ &\geq x(2u_1 + 2u_2 + 2c - \alpha^2) - 2c - 2u_2 \\ &\geq -2\kappa_1(1-x) = -2\kappa_1 p_1(x,y) \quad \text{for all } (x,y) \in [0,1]^2.\end{aligned}$$

Again, (19) holds for p_1 and the equality $\tau_{p_1} = \tau_1^X$ completes the proof of (ii).

As before, the remaining statements follow by symmetry. □

Remark 4.5. To prove that the first hitting time at the boundary is a.s. infinite given the condition (19) we could also have used the submartingale convergence theorem in [LR14] (Theorem 4.14), which says:

Proposition 4.6. *Let $\bar{\tau} > 0$ be a stopping time such that a non-decreasing sequence of stopping times (τ_n) with $\tau_n < \bar{\tau}$ for all n, a.s. on $\{\bar{\tau} > 0\}$ and $\tau_n \to \bar{\tau}$ almost surely, exists. Let N be a local supermartingale on $[0, \bar{\tau})$ starting at 0 for which*

$$\sup_n \mathbb{E}[N_{\tau_n}^-] < \infty.$$

Then, $\lim_{t\to\bar{\tau}} N(t)$ exists in $\mathbb{R}$ almost surely.[15]

We can use this theorem with $\tau_n := \inf\{t \geq 0 : p(X(t), Y(t)) \leq \frac{1}{n}\} \wedge s$, $\bar{\tau} = \tau_p \wedge s$ and $N(t) = -M(s \wedge t)$ for any deterministic $s > 0$. Then:

- (τ_n) is a non-decreasing sequence of stopping times, converging almost surely to $\bar{\tau}$ because of the continuity of p, X and Y;
- N is a stopped local supermartingale (starting at 0), and therefore a local supermartingale too;
- N is bounded from below since M is bounded from above on bounded time intervals ($[0, s]$ in this case).

Then, the assumption implies that $\lim_{t\to\tau_p} M(s \wedge t))$ and hence also $\lim_{t\to\tau_p} \log p(X(s \wedge t), Y(s \wedge t))$ exists in $\mathbb{R}$ almost surely. However, if it is possible that $s \geq \tau_p$ with positive probability, this would mean $\lim_{t\to\tau_p} \log p(X(t), Y(t))$ exists and is finite, too, which contradicts the definition of τ_p and M (since $p(X(t), Y(t))$ can't be bounded from below on a neighborhood of τ_p). So we get that $s < \tau_p$ $\mathbb{P}^{\mu_0}$-a.s. for any $s \in \mathbb{R}$ yielding $\tau_p = \infty$ as desired.

Remark 4.7. One could generalize the result to boundaries of more complex-shaped sets. We just have to find a fitting bounded, non-negative function p whose zeros are the set we want to know whether we can reach in finite time.

4.4 Lyapunov argument: applications

Another method for analyzing the boundary behavior of a diffusion is based on Lyapunov inequalities for the infinitesimal generator (see Chapter 2). Now, let us see how this algorithm is concretely implemented, starting with a well-known case.

[15] The main problem with this approach is that the main reference is not peer-reviewed. The proof however seems correct; the main idea is to use supermartingale convergence, with a similar statement being given in [CFR14].

4.4.1 Example I: seed bank diffusion

In the standard seed bank diffusion [1], the (two-dimensional) generator is

$$\begin{aligned}\mathcal{A}^{(1)}(f)(x_1,x_2) &= (-u_1x_1+u_2(1-x_1)+c(x_2-x_1))\frac{\partial f}{\partial x_1}\\ +(-u_1'x_2+u_2'(1-x_2)&+Kc(x_1-x_2))\frac{\partial f}{\partial x_2}+\frac{x_1(1-x_1)}{2}\frac{\partial^2 f}{\partial {x_1}^2}.\end{aligned}$$

We take the boundary $\{0\}\times[0,1]$, which leads to $O_n=(\frac{1}{n},1)\times(0,1)$; moreover, we use logarithmic functions here as well by choosing $V(x_1,x_2):=-\log x_1$. Then, calculating,

$$\mathcal{A}^{(1)}(V)(x_1,x_2)=u_1+u_2+c-\frac{1}{2}-\frac{u_2+cx_2-\frac{1}{2}}{x_1}\leq D+\frac{u_2-\frac{1}{2}}{x_1}$$

for some constant $D\geq 0$, and the non-explosivity condition holds if

$$\frac{u_2-\frac{1}{2}}{x_1}\geq C\log x_1$$

for every $x_1\in(0,1)$ and for some constant $C\geq 0$, which is the case if and only if $u_2\geq 1/2$, the same threshold we obtain by using the McKean argument. The same result holds for the two-island model. Notice, however, that the absence of random genetic drift in the second component reflects itself in the fact that with respect to the boundaries $[0,1]\times\{0\}$ and $[0,1]\times\{1\}$, the seed bank diffusion is always non-explosive, as one can easily calculate in a similar way.

4.4.2 Example II: Wright-Fisher with nonstandard diffusion term

We can also analyze a case which is similar to the one-dimensional Wright-Fisher diffusion, but where the random genetic drift term has an additional parameter p in the exponent in the form

$$\mathrm{d}X(t)=(X(t)(1-X(t)))^{\frac{1}{2}+p}\mathrm{d}W(t)+(-u_1X(t)+u_2(1-X(t)))\mathrm{d}t.$$

We assume that $p\in(0,\frac{1}{2}]$, that is, the diffusion term is smaller than in the classical Wright-Fisher process. This reflects itself in the generator

$$\mathcal{A}(f)(x)=(-u_1x+u_2(1-x))\frac{\partial f}{\partial x}+\frac{(x(1-x))^{1+2p}}{2}\frac{\partial^2 f}{\partial x^2}.$$

We can prove then that the process is non-explosive by choosing the norm-like function (see Subchapter 2.2.3) $V:x\rightarrow\frac{1}{x}$:

$$\mathcal{A}(V)(x)=-(-u_1x+u_2(1-x))\frac{1}{x^2}+(x(1-x))^{2p+1}\frac{1}{x^3},$$

and the non-explosivity condition is satisfied if $(D=0)$

$$-(-u_1x+u_2(1-x))\frac{1}{x}+\frac{(x(1-x))^{2p+1}}{x^2}\leq C\Leftarrow\frac{x^{2p}-u_2}{x}\leq C,$$

which holds for any $p,u_2>0$ and for C big enough.

Remark 4.8. In particular, the above holds for the case $p = 1/2$, which is strongly reminiscent of the Wright-Fisher diffusion with selection in fluctuating environment. For more literature, see ([BEK18]).

4.4.3 Example III: multi-allele diffusion

Another example of interest is the multi-allele diffusion, as discussed by Etheridge in ([Eth11], Lemma 4.1). Imagine we have one gene with N possible alleles in a homogeneous population. Then, the (multi-dimensional) generator of the $(N-1)$-dimensional diffusion process is given by

$$\mathcal{A}(f)(x_1,\dots,x_{N-1}) = \sum_{i=1}^{N-1} \frac{x_i(1-x_i)}{2}\frac{\partial^2 f}{\partial x_i^2} - \sum_{i<j} x_i x_j \frac{\partial^2 f}{\partial x_i \partial x_j}$$
$$+\sum_{i=1}^{N-1}\Big(- x_i \sum_{j=1}^{N} m_{ij} + \sum_{j=1}^{N-1} x_j m_{ji} + (1 - \sum_{k=1}^{N-1} x_k) m_{Ni}\Big)\frac{\partial f}{\partial x_i},$$

with the process living on the simplex $\mathcal{S}_{N-1} := \{(x_1, \dots x_{N-1}) \in [0,1] : \sum x_i \in [0,1]\}$ and m_{ij} denoting the mutation rate from allele i to allele j.
Here, we can consider two types of boundaries. The boundary we will take into consideration first is $\{0\} \times \mathcal{S}_{N-2}$ (i.e., this boundary is reached if the first allele becomes *temporarily extinct*). Analogously to example I, we choose as limiting sets $O_n = \{\mathbf{x} \in \mathcal{S}_{N-1} : x_1 < \frac{1}{n}\}$, and $V(x_1,\dots,x_n) := -\log x_1$ as the norm-like function. Then,

$$\mathcal{A}(V)(x_1,\dots,x_{N-1}) = \frac{1-x_1}{2x_1} - \big(- x_1 \sum m_{1j} + \sum_{j=2}^{N-1} x_j m_{j1} + (1 - \sum_{j=1}^{N-1} x_j) m_{k1}\big)\frac{1}{x_1}.$$

Summing up all constants into an unique constant D, we get

$$\mathcal{A}(V)(x_1,\dots,x_{N-1}) = D + \frac{1}{x_1}\big(\frac{1}{2} - \sum_{j=2}^{N-1} x_j m_{j1} - (1 - \sum_{j=1}^{N-1} x_j) m_{k1}\big)$$

Defining $M := \min_i\{m_{i1}\}$, we get

$$\mathcal{A}(V)(x_1,\dots,x_{N-1}) \leq D + \frac{1}{x_1}\big(\frac{1}{2} - \sum_{j=2}^{N-1} x_j M - (1 - \sum_{j=2}^{N-1} x_j) M\big)$$
$$= D + \frac{1}{x_1}\big(\frac{1}{2} - (1-x_1)M\big) = D' + \frac{1}{x_1}\big(\frac{1}{2} - M\big),$$

having changed the constant accordingly. Thus, the process does not hit the boundary if $M \geq 1/2$. This is both consistent with the two-allele case and reasonable, since the drift due to mutation must be strong enough to tackle the random genetic drift, no matter which point of the boundary the process is pushed towards ("a chain is just as

strong as its weakest link").
The second question we pose ourselves regards the absorption at any vertex of the simplex, e.g. $(0, 0, \dots, 0)$, which is equivalent to *temporary fixation* of an allele. In this case, the norm-like function we choose is $V(\mathbf{x}) = -\log \sum x_i$, so that

$$\begin{aligned}
&\mathcal{A}(V)(x_1, \dots, x_{N-1}) \\
=&\frac{1}{2(\sum x_i)^2} \sum_i x_i(1-x_i) + \frac{1}{(\sum x_i)^2} \sum_{i<j} x_i x_j - \frac{1}{\sum x_i} \sum_{i=1}^{k} (-x_i \sum_j m_{ij} + \sum_j x_j m_{ji}).
\end{aligned}$$

The second and third terms are clearly bounded, so the critical condition is whether

$$\frac{1}{2(\sum x_i)^2} \sum_i x_i(1-x_i) - \frac{1}{\sum x_i} \sum_{i=1}^{k} \sum_j x_j m_{ji}$$

is bounded from above. We can write the term as

$$\frac{1}{2(\sum x_i)^2} \sum_i x_i(1-x_i) - \frac{1}{\sum x_i} \sum_{i=1}^{k} \sum_{j=1}^{k-1} x_j m_{ji} - \frac{1}{\sum x_i} \sum_{i=1}^{k} (1 - \sum_{j=1}^{k-1} x_j) m_{ki},$$

and the second and fourth terms are again bounded, so the critical condition is whether

$$\frac{1}{2(\sum x_i)^2} \sum_i x_i(1-x_i) - \frac{1}{\sum x_i} \sum_{i=1}^{k} m_{ki} \leq \frac{1}{\sum x_i} (\frac{1}{2} - \sum m_{ki})$$

is bounded from above, which holds if $\sum_j m_{kj} \geq 1/2$. This condition makes sense since it is weaker than $m_{kj} \geq 1/2$ for every j - the condition which guarantees that none of the boundaries containing the critical point are reached.

Of course, what is useful to show now is that the bounds we got with the Lyapunov argument are strict. In order to prove this, we proceed in the same way as in the seed bank example, i.e. by using polynomial diffusions. Taking

$$a(x_1, \dots, x_n) = \begin{pmatrix} x_1(1-x_1) & -x_1 x_2 & -x_1 x_3 & \dots & -x_1 x_n \\ -x_1 x_2 & x_2(1-x_2) & -x_2 x_3 & \dots & -x_2 x_n \\ \dots & \dots & \dots & \dots & \dots \\ -x_1 x_n & -x_2 x_n & -x_3 x_n & \dots & x_n(1-x_n) \end{pmatrix},$$

we get that all entries of a are polynomials of degree 2 and that for the generator $\mathcal{A}$

$$\mathcal{A}(f)(x) := \frac{1}{2} Tr(a \nabla^2 f) + b^T \nabla f$$

for some vector of linear terms b, making the multi-allele diffusion process a polynomial diffusion as well. Therefore, we just need to check the two conditions

$$\mathcal{A} p_0(\bar{z}) \geq 0$$

and

$$2\mathcal{A} p_0(\bar{z}) - h_{p_0}(\bar{z})^T \nabla p_0(\bar{z}) < 0,$$

where p_0 is a polynomial that is zero on the critical boundary and $\bar{z}$ is any point on the boundary itself. The first choices that come to mind are, for the first case (temporary extinction),

$$p_0(x_1, \dots x_n) = x_1 \text{ and } \bar{z} \in \{(0,0,\dots,0),(0,1,\dots,0),\dots,(0,0,\dots,0,1)\};$$

for the second case (temporary fixation),

$$p_0(x_1, \dots x_n) = x_1 + x_2 \cdots + x_n \text{ and } \bar{z} = (0,0,\dots,0).$$

Let us start with the first case. Here,

$$\begin{aligned} a\nabla p_0(x_1, \dots x_n) &= (x_1(1-x_1), -x_1x_2, -x_1x_3, \cdots - x_1x_n)^T \\ &= x_1(1-x_1, -x_2, \cdots - x_n)^T = p_0 h_{p_0}. \end{aligned}$$

Moreover,

$$\mathcal{A}(p_0)(\bar{z}) = (-x_1 \sum_{j=1}^{n} m_{1j} + \sum_{j=1}^{n-1} x_j m_{j1} + (1 - \sum_{k=1}^{n-1} x_k) m_{n1}(\bar{z}),$$

which is equal to m_{n1} in the case $\bar{z} = 0$ and m_{k1} in the case $\bar{z} = \delta_k$. Both are non-negative, and in addition

$$hp_0(\bar{z})^T \nabla p_0(\bar{z}) = 1$$

for both choices of $\bar{z}$, which yields as sufficient condition for reaching the boundary with positive probability $m_{k1} < 1/2$ for any $k \in \{2,3,\dots,n\}$. This result is exactly the complement of what we computed using the Lyapunov argument.
In the second case,

$$\begin{aligned} &a\nabla p_0(x_1, \dots x_n) \\ &\quad = \Big(x_1(1-x_1) - x_1 \sum_{j \neq 1} x_j, x_2(1-x_2) - x_2 \sum_{j \neq 2} x_j, \dots x_n(1-x_n) - x_n \sum_{j \neq n} x_j\Big)^T, \end{aligned}$$

and therefore

$$h_{p_0} = \Big(\frac{x_1}{\sum x_i} - x_1, \dots, \frac{x_n}{\sum x_i} - x_n\Big)^T,$$

giving us

$$2\mathcal{A}(p_0)(\bar{z}) - (h_{p_0}^T \nabla p_0)(\bar{z}) = 2 \sum_i m_{ni} - 1.$$

Therefore, the corner point $(0,0,\dots 0)$ will be reached in finite time with positive probability if $\sum_i m_{ni} < 1/2$, which again gives us a complementary result to the one obtained with the Lyapunov argument. Thus, both boundary questions have been answered in all cases.

4.4.4 Example IV: branching processes

The last example concerns branching processes, as defined in [Wat69], Section 3. A two-dimensional branching process with interaction is given by its generator, acting on functions $f \in \mathcal{C}^2([0,\infty)^2)$,

$$\mathcal{A}(f)(x_1,x_2) = \alpha^2 x_1 \frac{\partial^2 f}{\partial x_1^2} + \beta^2 x_2 \frac{\partial^2 f}{\partial x_2^2} + (ax_1 + bx_2)\frac{\partial f}{\partial x_1} + (cx_1 + dx_2)\frac{\partial f}{\partial x_2}.$$

In the case where $b = c = 0$, this is the two-dimensional process introduced by Feller in [Fel51] (Section 6) to describe a population with two competing alleles and no fixed size (unlike e.g. the Wright-Fisher process), where the author remarks that the gene frequency is not even a Markov process.
Here, the question we want to pose ourselves is whether the process will reach ∞ in finite time (thus, a proper "explosion"), not 0. For this aim, we define $O_n = [0,n)^2$ and $V(x_1,x_2) := \log(x_1 + x_2 + 1)$, yielding the result

$$\mathcal{A}(V)(x_1,x_2) = -\frac{\alpha^2 x_1}{(x_1+x_2+1)^2} - \frac{\beta^2 x_2}{(x_1+x_2+1)^2} + \frac{(a+c)x_1 + (b+d)x_2}{x_1+x_2+1}.$$

But since all the terms are clearly bounded for $x_1 + x_2 \to \infty$, the non-explosivity criterion holds.

4.5 Discussion

In this chapter, we have completely solved our questions regarding the boundary behavior problem by finding a necessary and sufficient condition. One inequality could be proved using the polynomial diffusion method from Chapter 2.3, while the other could be shown both with the McKean argument and the Lyapunov argument. Since the first one works with the SDE representation and the second one with the infinitesimal generator, both could be useful throughout future work. Overall, the Lyapunov argument makes for smoother proofs and could therefore be preferred when both methods are applicable.

Moreover, a generalization of the Lyapunov argument can be found in [SV07], where a time component is added to the norm-like function. Using this trick, it is possible to find a condition for the process to hit the boundary [Theorem 10.2.1]. The case where the norm-like function is independent from time simply reduces to the Lyapunov argument. Another important publication, giving a sort of a generalization of the McKean argument, is given in [Ruf15], where a local martingale Z, taking the place of the local martingale M we constructed throughout the proof, is given. The main result is obtained applying Girsanov's Theorem to the Radon-Nikodym derivative of Z. For further information, we refer to the mentioned publication.

5 Scaling Limits

5.1 Introduction

It has been reported that in various bacterial species, single individuals may stay inactive for extremely long periods of time, several orders of magnitude longer than the reproductive time ([LJ11], [ADLM99]). Further, it can be expected that in such a scenario, one observes non-classical behavior of the genealogy over long time-scales (e.g., 'extinct' types may be reintroduced after long time periods, though this should happen rarely). These considerations motivate the investigation of the behavior of the above system when migration between active and dormant states (rate c) and reproduction (rate 1) happen on different time-scales.

A stochastic process with this characteristic is usually found as a so-called *scaling limit*, i.e. by investigating convergence of Markov processes where both time and a certain variable are rescaled fittingly.
In particular, scaling diffusion limits of certain processes, which are obtained by letting the population size go to infinity and speeding up time accordingly, appear frequently in population genetics (see [BGKWB16], [Eth11], [KZH08], [Möh98] for results related to the context). However, in many of those cases the authors are concerned with the limit of a sequence of discrete-time Markov chains, while the aim of this chapter is to find the scaling limits of the continuous-time diffusion processes introduced in Chapter 1 for $c \to 0, c \to \infty, K \to 0, K \to \infty$, and $\alpha' \to 0$. Interestingly, we will see that the limit can be a jump diffusion. In this case, many classical results fail; instead, we refer to a duality argument to establish convergence in finite-dimensional distributions.

To be more specific, our approach in order to obtain the scaling limit of the seed bank diffusion processes (1) will be as follows. First, we take a sequence of dual (block-counting) processes (as in Definition 1.6), which, as we already saw, are continuous-time Markov chains, for $c \to 0$ and time sped up accordingly. Then, we will transform them to discrete-time Markov chains via a time-discretization and use a theorem ([Möh98], Thm 1) giving convergence towards a time-continuous limiting process. Then, we will show that this limiting process is also the limit of the original (continuous-time) sequence, closing the proof of the convergence result for the dual processes. Finally, we will show again, using a duality argument, that the diffusion processes converge towards a two-dimensional Markov process with jumps in one coordinate. Since the limit exhibits jumps (while the original diffusion processes, by definition, have none), we cannot get any limits on the path space in the Skorohod sense (at least respectively to the topologies J_1 and J_2), but only limits in finite-dimensional distributions.
The chapter is organized as follows: after introducing the methods used, we derive some scaling limits of the seed bank diffusion on different time-scales. The central point will be defining a new ancestral process describing the genealogy in a 'rare-resuscitation' regime and proving that this exact process is the scaling limit for $c \to 0$ and time sped up accordingly. This process is the result of a separation of time-scales phenomenon, and we discuss the corresponding convergence result in detail.
This chapter is based on [K2].

5.2 The ancient ancestral lines process as scaling limit

5.2.1 Intuition

When computing the scaling limit of the diffusion 1 for $c \to 0$, it is easy to see that interesting limits can only be expected when switching to a faster time-scale. Indeed, if one just lets $c \to 0$, then one obtains the trivial (and uninteresting) limit where the active population is completely separate from the dormant population and simply follows a classical Wright-Fisher diffusion. Hence we speed up time by a factor $1/c$, as $c \to 0$, and transition to the new time-scale, where now migration between states happens at rate 1 while reproduction happens 'instantaneously'. While such a separation of time-scales can be expected to lead to an interesting process, the naïve scaling limit

$$\begin{aligned} \mathrm{d}X(t) &= (Y(t) - X(t))\mathrm{d}t + \text{“}\infty\text{”}\sqrt{X(t)(1-X(t))}\mathrm{d}W(t), \\ \mathrm{d}Y(t) &= K(X(t) - Y(t))\mathrm{d}t, \end{aligned} \tag{20}$$

of course does not make sense (observe the "∞") in front of the diffusion coefficient due to speeding up time).

Intuitively, fast reproduction should drive the process immediately towards the boundaries, and only rarely should one switch from 0 to 1 or vice versa due to immigration. Yet, it is not completely obvious how to make this idea rigorous. Fortunately, we can use duality here. We have seen in Chapter 1 that the moment dual of the seed bank diffusion without mutation is defined as the continuous-time Markov chain with values in $E = \mathbb{N}_0 \times \mathbb{N}_0$ characterized by the rates $\bar{N}_{(n,m),(\bar{n},\bar{m})}$ given by:

$$\bar{N}_{(n,m),(\bar{n},\bar{m})} = \begin{cases} \binom{n}{2} & \text{if } (\bar{n},\bar{m}) = (n-1,m), \\ cn & \text{if } (\bar{n},\bar{m}) = (n-1,m+1), \\ cKm & \text{if } (\bar{n},\bar{m}) = (n+1,m-1), \end{cases} \tag{21}$$

(with the convention $\binom{1}{m} = \binom{0}{m} = 0$ for all m), when $(\bar{n},\bar{m})$, $(n,m) \in \mathbb{N}_0 \times \mathbb{N}_0$ and zero otherwise off the diagonal.

In Subchapter 2.1, we saw that this continuous Markov chain process satisfies the moment duality

$$\mathbb{E}_{(x,y)}\big[X(t)^n Y(t)^m\big] = \mathbb{E}_{(n,m)}\big[x^{N(t)} y^{M(t)}\big] \tag{22}$$

for every $t > 0$, for every $(x,y) \in [0,1]$ and for every $n,m \in \mathbb{N}_0$.

In other words, the distribution of the seed bank diffusion at any time t is uniquely determined by the moment dual at said time. It is thus a natural idea to investigate the scaling limit of the moment dual under the same scaling assumption, which is potentially easier to get than the one for the original diffusion, and hope to obtain a well-defined limit which still provides information about the scaling limit of the original diffusion. Still, we encounter technical challenges since the limiting objects might not have standard semi-groups. Indeed, when speeding up time in the continuous-time Markov chain, some transition rates diverge to ∞, thus obstructing direct Q-matrix computations and producing states that are vacated immediately. This phenomenon

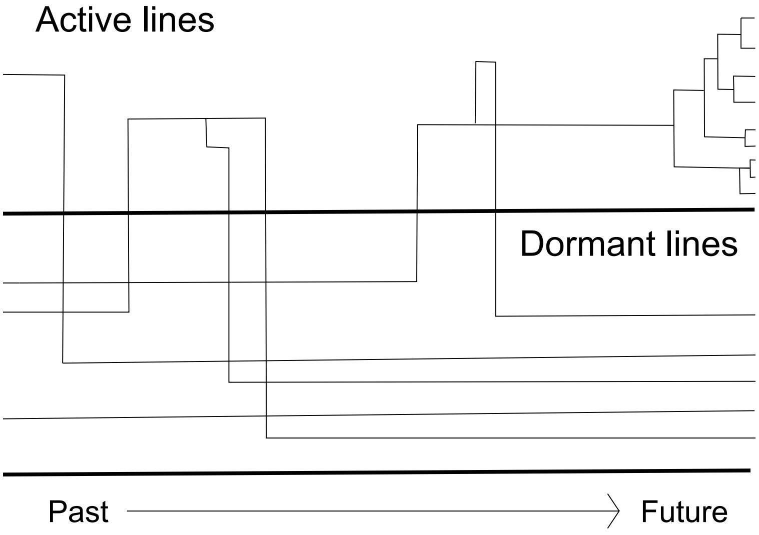


Figure 4: For minuscule migration rate c, coalescent events occur rapidly in comparison to migration events as depicted above. As a result, passing to the limit on a scale that normalizes the migration rates, means coalescence events occur instantaneously and any randomness lies in the migration events.

is frequently observed when dealing with "separation-of-time-scales phenomena"(cf. for example [Wak09, Chapter 6]) and can in the best case scenario still lead to a scaling limit with "degenerate" (non-standard) transition semigroup of the form

$$Pe^{Gt}, \quad t \geq 0,$$

where P is a *projection* to a subspace of the original state space as a result of "immediately vacated states" and G a "classical" conservative Q-matrix. For *discrete-time* Markov chains, this situation was considered e.g. in [Möh98] and also [BBE13]; see further [MN16]. Since this might be of general interest, we give, a detailed "recipe" for such proofs for continuous-time Markov chains in Chapter 5.3.

Subsequently, we apply this strategy to our model in Chapter 5.4 and obtain the following results: Recall that we are interested in the (very) long term effect of a seed bank in which individuals change their state rarely. Heuristically, this means that the switch between active and inactive takes a long time to happen as depicted in Figure 4. In this setting, the time that any pair of active lines needs to find a common ancestor becomes negligible compared with the time that an ancestral line takes to change its state. The consequence of this is that, in the scaling limit, two active ancestral lines coalesce instantaneously while each line changes of state after a random time of order one, as described by the *ancient ancestral lines process*:

5.2.2 Definitions

Definition 5.1. [The ancient ancestral lines process] Let $(n_0, m_0) \in \mathbb{N}_0 \times \mathbb{N}_0$. The *ancient ancestral lines process* is the continuous-time Markov chain $(\tilde{N}(t), \tilde{M}(t))_{t\geq 0}$

with initial value $(\tilde{N}(0), \tilde{M}(0)) = (n_0, m_0)$, taking values in the state space

$$E_{(n_0,m_0)} := \{0, \dots, n_0 + m_0\}^2,$$

with transition matrix

$$\Pi(t) := Pe^{tG}, \qquad t > 0$$

and $\Pi(0)$ is the identity on E. Here, P is a projection given by

$$P_{(n,m),(\bar{n},\bar{m})} := \begin{cases} 1, & \text{if } \bar{n} = 1,\, n \geq 1,\, \bar{m} = m, \\ 1, & \text{if } \bar{n} = n = 0,\, \bar{m} = m, \\ 0, & \text{otherwise,} \end{cases} \tag{23}$$

for all sensible (n, m), $(\bar{n}, \bar{m}) \in E_{(n_0.m_0)}$ and G is a matrix of the form

$$G_{(n,m),(\bar{n},\bar{m})} := \begin{cases} Km, & \text{if } \bar{n} = 1,\, n \geq 0,\, \bar{m} = m - 1, \\ 1, & \text{if } \bar{n} = 0,\, n \geq 1,\, \bar{m} = m + 1, \\ -1 - Km, & \text{if } \bar{n} = 1,\, n \geq 1,\, \bar{m} = m, \\ -Km, & \text{if } \bar{n} = n = 0,\, \bar{m} = m, \\ 0, & \text{otherwise.} \end{cases}$$

The projection acts for any $t > 0$, hence this process ‘immediately’ takes values in the smaller space $\{0, 1\} \times \{0, \dots, m_0 + 1\}$. The first two rates given in the definition of G correspond to the events of resuscitation with immediate coalescence and dormancy. Note that G is, however, not a Q-matrix: for any $\bar{n} \geq 2$ its negative values are off the diagonal.

Using the techniques of Chapter 5.3, we prove that the *ancient ancestral lines process* arises as the scaling limit of the block-counting process of the seed bank coalescent.

5.2.3 Main convergence results

Theorem 5.2. *Denote by $(N^c(t), M^c(t))_{t\geq 0}$ the block counting process of the seed bank coalescent as defined in Definition 1.6 with migration rate $c > 0$ and assume that it starts in some $(n_0, m_0) \in \mathbb{N} \times \mathbb{N}$, $\mathbb{P}$-a.s.*

Furthermore let $(\tilde{N}(t), \tilde{M}(t))_{t\geq 0}$ be the ancient ancestral lines process from Definition 5.1 with the same initial condition. Then, for any sequence of migration rates $(c_\kappa)_{\kappa\in\mathbb{N}}$ with $c_\kappa \to 0$ for $\kappa \to \infty$

$$\left(N^{c_\kappa}\left(\frac{1}{c_\kappa}t\right), M^{c_\kappa}\left(\frac{1}{c_\kappa}t\right)\right)_{t\geq 0} \xrightarrow{f.d.d.} \left(\tilde{N}(t), \tilde{M}(t)\right)_{t\geq 0},$$

as $\kappa \to \infty$.

Recall from (22) that for each fixed $\kappa \in \mathbb{N}$ the process $(N^{c_\kappa}(t), M^{c_\kappa}(t))_{t\geq 0}$ is the moment dual of the seed bank diffusion $(X^{c_\kappa}(t), Y^{c_\kappa}(t))_{t\geq 0}$, where again indicate the value of the migration rate by the superscript c_κ. As we will see in Chapter 5.4 this moment duality is the key ingredient that allows to formalize the proof of convergence of this sequence of diffusions and the existence of the limit as a Markov process, which is not a diffusion and "essentially" has state space $\{0,1\} \times [0,1]$, as described in Figure 5.

Theorem 5.3. *Let $(X^c(t), Y^c(t))_{t\geq 0}$ be the seed bank diffusion given in Definition 1.1 with migration rate $c > 0$. There exists a Markov process $(\tilde{X}(t), \tilde{Y}(t))_{t\geq 0}$ on $[0,1]^2$ such that for any sequence of migrations rates with $c_\kappa \to 0$ for $\kappa \to \infty$*

$$\left(X^{c_\kappa}\left(\frac{1}{c_\kappa}t\right), Y^{c_\kappa}\left(\frac{1}{c_\kappa}t\right)\right)_{t\geq 0} \xrightarrow{f.d.d.} (\tilde{X}(t), \tilde{Y}(t))_{t\geq 0}$$

as $\kappa \to \infty$. Furthermore, $(\tilde{X}(t), \tilde{Y}(t))_{t\geq 0}$ is characterized as the moment dual of the ancient ancestral lines process from Definition 5.1.

Much like its dual, the limit $(\tilde{X}(t), \tilde{Y}(t))_{t\geq 0}$ is "degenerate" in the sense that it does not have a generator because of the discontinuity of its semi-group in $t = 0$. However, as we will prove in Proposition 5.9, if started in $\{0,1\} \times [0,1]$ it coincides in distribution with a jump-diffusion taking values in $\{0,1\} \times [0,1]$ whose generator is given by

$$\begin{aligned}\bar{\mathcal{A}}^{(1)} f(x,y) &= y(f(1,y) - f(0,y))\mathbb{1}_{\{0\}}(x) + (1-y)(f(0,y) - f(1,y))\mathbb{1}_{\{1\}}(x) \\ &\quad + K(x-y)\frac{\partial f}{\partial y}(x,y).\end{aligned}$$

In particular this means that the limit $(\tilde{X}(t), \tilde{Y}(t))_{t\geq 0}$ instantaneously jumps into the smaller state space $\{0,1\} \times [0,1]$.

As in the case of the continuous-time Markov chains, we again state the general result of translating the convergence through duality in Subchapter 5.4.2 and the apply it with the *ancient material scaling* in Chapter 5.4.3.

Remark 5.4. Since we believe the methodology used to prove Theorems 5.2 and 5.3 can be applied in many situations, helping those interested in scaling limits of Markov processes that experience a separation of scales, we have separated the general methodology from the example of the *ancient material scaling.*

For continuous-time Markov chains, this is done in Chapter 5.3 with its key innovation being Lemma 5.5.

If the processes of interest, on the other hand, are the moment duals of a sequence of continuous-time Markov chains, then the convergence in finite dimensional distributions of one sequence of processes can be translated into convergence in finite dimensional distributions for the other, propagating a separation of time-scales where applicable. The strategy of proof is to use the commutative diagram depicted in Figure 6. This is the content of Theorem 5.6 in Chapter 5.4.2 and allows us to prove the convergence of a family of diffusions into a non trivial, non diffusion Markov process.

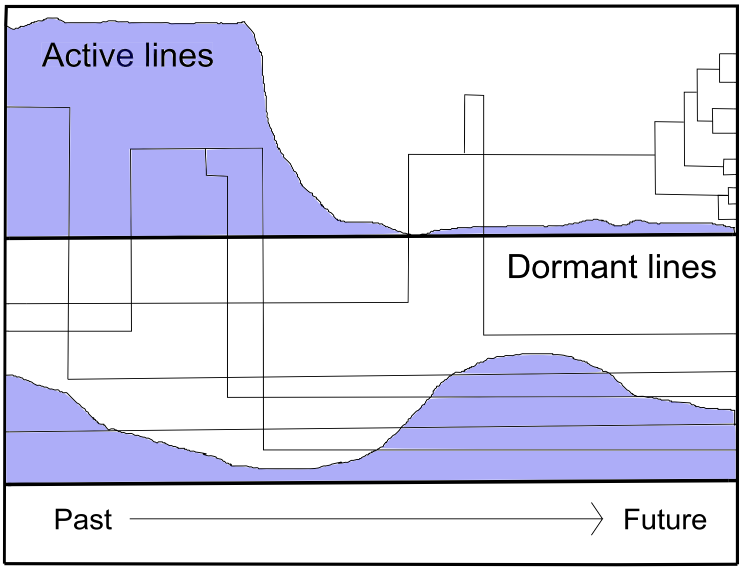


Figure 5: For minuscule migration rate c, most of the time, the active population will be almost homogeneous, i.e. the frequency process of the active population will be very close to one of its boundaries. However, the (rare) migration events of the opposite type from the dormant population will prevent it from staying in that boundary. From time to time, one of these migrations might lead to a change of the predominant type in the active population. This sweep will be extremely fast (of order of the inverse of c), and thus instantaneous in the limit. Due to the homogeneity of the active population, the seed bank will mostly receive individuals of the type dominant in the active population at that time and will thus evolve almost deterministically.

$$\begin{tikzcd}
(N^c(\frac{t}{c}), M^c(\frac{t}{c})) \arrow[r, leftrightarrow, "\text{Duality}"] \arrow[d] & (X^c(\frac{t}{c}), Y^c(\frac{t}{c})) \arrow[d] \\
(\tilde{N}(t), \tilde{M}(t)) \arrow[r, leftrightarrow, "\text{Duality}"] & (\tilde{X}(t), \tilde{Y}(t))
\end{tikzcd}$$

Figure 6: Commutative diagram summarizing the relations between the processes considered. The moment duality of the prelimits and the limits is used to conclude the convergence in f.d.d. on the right from the convergence of the processes on the left.

Once convergence of the finite dimensional distributions is established, it is natural to wonder if it is possible to prove tightness, in order to obtain weak convergence over the Skorohod space with the J_1-topology. At first glance maybe surprisingly, it is not hard to see that convergence in this sense cannot hold. To see this, observe that there are *jumps* occurring at the events that happen instantaneously in the limit. In these time-points the prelimiting processes visit a state outside the smaller state space of the limit process. In the coalescent set-up, for example, if there is one active and one dormant individual, in order to lose one block the discrete processes go from the state $(1, 1)$ to $(2, 0)$ and then $(1, 0)$ (in quick succession). On the other hand, the limiting process goes directly from $(1, 1)$ to $(1, 0)$. Regardless of the time spent in the state $(2, 0)$ by the prelimiting processes approaching 0, this makes convergence in both the Skorohod J_1- and J_2-topologies impossible. However, the set of such time-points has Lebesgue measure equal to zero, whence convergence in the Meyer-Zheng topology ([Kur91], page 8) should hold.

5.3 Separation of time-scales phenomena for continuous-time Markov chains - a strategy

As thoroughly motivated by the example of the *ancient material scaling* in the introductory subchapter, as a first step, we consider scaling limits of continuous-time Markov chains and extend the results for discrete-time Markov chains from [BBE13, Möh98, MN16].

Given a sequence of continuous-time Markov chains $(\xi^\kappa(t))_{t\geq 0}$, $\kappa \in \mathbb{N}$ with finite state-space E (equipped with a metric d), we want to prove its convergence under some time-rescaling $(c_\kappa)_{\kappa\in\mathbb{N}}$ to a limit $(\xi(t))_{t\geq 0}$ for $\kappa \to \infty$.

The idea behind these proofs has three steps:

i) First, we consider *time discretizations* of the original continuous-time Markov chains by considering the discrete-time Markov chains $\eta^\kappa(i) := \xi^\kappa(i/a_\kappa), i \in \mathbb{N}_0$. The non-negative sequence $(a_\kappa)_{\kappa\in\mathbb{N}}$ with $a_\kappa \to \infty$, will be chosen to ensure the distance between the discretizations and the original processes to be sufficiently small.

ii) Secondly, we employ a generalization of Theorem 1 in [Möh98], namely Lemma 1.7 in [BBE13], to establish *convergence of the discretized processes to the desired continuous-time limit* on the new time scale by speeding up the discretized processes by $b_\kappa = a_\kappa/c_\kappa$, that is, we establish the convergence $(\eta^\kappa(\lfloor b_\kappa t\rfloor))_{t\geq 0} \to (\xi(t))_{t\geq 0}$ in finite dimensional distributions.

iii) Finally, we prove a continuity result to show that the original processes sped up by the factor b_κ/a_κ, i.e. $(\xi^\kappa(b_\kappa t/a_\kappa))_{t\geq 0}$, converges to the same limit $(\xi(t))_{t\geq 0}$ in finite dimensional distributions.

Since the strategy is not at all restricted to the specific examples we consider, but might be of general interest, we will give the details together with the necessary results here.

Step i) Denote by G^κ the Q-matrix of $(\xi^\kappa(t))_{t\geq 0}$ for each $\kappa \in \mathbb{N}$. The rescaling sequence $(a_\kappa)_{\kappa\in\mathbb{N}}$ needs to be chosen such that for $q_\kappa := \max_{e\in E}\left\{-G^\kappa_{e,e}\right\}$

$$a_\kappa \to \infty \qquad \text{and} \qquad \frac{q_\kappa}{a_\kappa} \to 0, \qquad \text{for } \kappa \to \infty. \tag{24}$$

Define the time discretizations

$$\eta^\kappa(i) := \xi^\kappa(i/a_\kappa), i \in \mathbb{N}_0$$

of the original sequence of continuous-time Markov chains $(\xi^\kappa(t))_{t\geq 0}$, $\kappa \in \mathbb{N}$.

As we will see in the proof of Lemma 5.5 in **Step iii)**, (24) will ensure the step-size to be sufficiently fine for the probability of a jump of ξ^κ during one time-step of η^κ to tend to 0.

Step ii) The next step is to apply the known convergence result for discrete-time Markov chains from [BBE13] to the sequence $((\eta^\kappa)(i))_{i\in\mathbb{N}_0}$, the assumptions of which we summarize here for the reader's convenience. Let Π_κ be the transition matrix of $(\eta^\kappa(i))_{i\in\mathbb{N}_0}$.

First, establish a suitable decomposition of Π_κ: For the sequence $(b_\kappa)_{\kappa\in\mathbb{N}}$ with $b_\kappa = a_\kappa/c_\kappa \to \infty$ write

$$\Pi_\kappa = A_\kappa + \frac{B_\kappa}{b_\kappa} \tag{25}$$

where A_κ is a stochastic matrix that contains only entries of order 1 and a_κ^{-1}, and B_κ contains only entries of order 1 and $o(1)$. As we will see below, speeding up time by the factor b_κ leads in the limit to a separation of time-scales, where the entries in A_κ give rise to a projection matrix P acting on the probability distributions on E, effectively restricting the state space of the limiting continuous-time Markov chain to a subspace of E, while the entries of B_κ yield the infinitesimal generator.

In order to prove this, first confirm that

$$\lim_{C\to\infty}\lim_{\kappa\to\infty}\sup_{r\geq Ca_\kappa}\|(A_\kappa)^r - P\| = 0 \tag{26}$$

for some matrix P. Here we equipped the matrices $A = (A(e,\bar{e}))_{e,\bar{e}\in E}$ on E with the matrix norm $\|A\| := \max_{e\in E}\sum_{\bar{e}\in E}|A(e,\bar{e})|$. Note that given (26), the matrix P is necessarily a projection on E, i.e. satisfies $P^2 = P$, as can be checked by a small calculation. To see this, note that $||A_N|| \leq 1$ and consider

$$\begin{aligned}||P^2 - P|| &= ||(P - (A_N)^r + (A_N)^r)^2 - P||\\ &\leq ||(P-(A_N)^r)^2|| + 2||P - (A_N)^r||\,||(A_N)^r|| + ||(A_N)^{2r} - P||\end{aligned}$$

which can be made arbitrarily small by choosing r and N sufficiently large.
Secondly, require that the matrix limit with respect to the matrix norm

$$G := \lim_{\kappa\to\infty} PB_\kappa P \qquad \text{exists.} \tag{27}$$

(of course, such G certainly exists if $B_N \to B$ for some fixed matrix B). Since E is assumed to be finite, convergence in matrix norm is equivalent to point-wise convergence. Then, by [BBE13, Lemma 1.7 and Remark 1.8], we obtain the following convergence (with respect to the matrix norm):

$$\lim_{\kappa\to\infty} \Pi_\kappa^{\lfloor tb_\kappa \rfloor} = \lim_{\kappa\to\infty} \left(A_\kappa + \frac{B_\kappa}{b_\kappa} \right)^{\lfloor tb_\kappa \rfloor} = Pe^{tG} =: \Pi(t) \qquad \text{for all } \; t > 0. \tag{28}$$

Note that since $P = P^2$, we have $PG = GP = G$ and hence $Pe^{tG} = e^{tG}P = P - I + e^{tG}$ for any $t \geq 0$. In particular, $(\Pi(t))_{t\geq 0}$ with $\Pi(0) := \mathrm{Id}_E$ is a (non-standard) semi-group and we denote by $(\xi(t))_{t\geq 0}$ the continuous-time Markov chain it generates.

If, last but not least, $\eta^\kappa(0) \xrightarrow{w} \xi(0)$, equation (28) implies

$$(\eta^\kappa(\lfloor b_\kappa t \rfloor))_{t\geq 0} \xrightarrow{\text{f.d.d.}} (\xi(t))_{t\geq 0}, \quad \text{as } \kappa \to \infty.$$

Here, $\xrightarrow{\text{f.d.d.}}$ denotes convergence of the processes in *finite dimensional distributions.*

Step iii) Lastly, we prove that the conditions given in **Step i)** and **ii)** are sufficient to also the convergence of original continuous-time Markov chains $(\xi^\kappa(t))_{t\geq 0}$ to the same limit $(\xi(t))_{t\geq 0}$ on the faster time-scale b_κ/a_κ as well. This is summarized in the following lemma.

Lemma 5.5. *Let $(\xi^\kappa(t))_{t\geq 0}, \kappa \in \mathbb{N}$ be a sequence of continuous-time Markov chains with finite state space E (equipped with some metric d). Let $(a_\kappa)_{\kappa\in\mathbb{N}}$ and $(b_\kappa)_{\kappa\in\mathbb{N}}$ be non-negative sequences such that $a_\kappa, b_\kappa/a_\kappa \to \infty$.*

Define the sequence of discrete-time Markov chains $(\eta^\kappa(i))_{i\in\mathbb{N}_0}, \kappa \in \mathbb{N}$ by

$$\eta^\kappa(i) := \xi^\kappa\left(\frac{i}{a_\kappa}\right), \quad i \in \mathbb{N}_0.$$

Denote by G^κ the Q-matrix of $(\xi^\kappa(t))_{t\geq 0}$ for each $\kappa \in \mathbb{N}$ and set $q_\kappa := \max_{e\in E}\left\{-G^\kappa_{e,e}\right\}$. If

- *a) $\frac{q_\kappa}{a_\kappa} \to 0$ and*
- *b) $(\eta^\kappa(\lfloor b_\kappa t \rfloor))_{t\geq 0} \xrightarrow{f.d.d.} (\xi(t))_{t\geq 0}$ as $\kappa \to \infty$, and*

then also

$$\left(\xi^\kappa\left(\frac{b_\kappa}{a_\kappa}t\right)\right)_{t\geq 0} \xrightarrow{f.d.d.} (\xi(t))_{t\geq 0} \quad \text{as } \kappa \to \infty.$$

Hence, if the conditions given in **Step i)** and **ii)** hold, then Lemma 5.5 will yield the desired convergence.

Proof. Note that condition a) was chosen precisely such that

$$\mathbb{P}\left\{(\xi^\kappa(t))_{t\geq 0} \text{ has a jump in } \left(0\,,\, \frac{1}{a_\kappa}\right]\right\} \leq 1 - \exp\left(\frac{-q_\kappa}{a_\kappa}\right) \to 0, \quad \kappa \to \infty. \tag{29}$$

Observe that for the distance between $(\xi^\kappa(t))_{t\geq 0}$ and $(\eta^\kappa(t))_{t\geq 0}$ at any time $t \geq 0$ we have

$$d\left(\xi^\kappa\left(\frac{b_\kappa t}{a_\kappa}\right), \eta^\kappa(\lfloor b_\kappa t \rfloor)\right) = d\left(\xi^\kappa\left(\frac{b_\kappa t}{a_\kappa}\right), \xi^\kappa\left(\frac{\lfloor b_\kappa t \rfloor}{a_\kappa}\right)\right) > 0$$

only if the process $(\xi^\kappa(t))_{t\geq 0}$ has a jump in the interval $\left(\frac{\lfloor b_\kappa t\rfloor}{a_\kappa}, \frac{b_\kappa t}{a_\kappa}\right]$. Since its length can be estimated through

$$0 \leq \frac{b_\kappa t}{a_\kappa} - \frac{\lfloor b_\kappa t \rfloor}{a_\kappa} \leq \frac{1}{a_\kappa}$$

we can estimate the probability of this event with (29) and obtain

$$\mathbb{P}\left\{d\left(\xi^\kappa\left(\frac{b_\kappa t}{a_\kappa}\right), \eta^\kappa(\lfloor b_\kappa t \rfloor)\right) > 0\right\} \leq 1 - \exp\left(\frac{-q_\kappa}{a_\kappa}\right) \to 0, \quad \kappa \to \infty. \tag{30}$$

In order to prove the convergence of the finite dimensional distributions, recall that weak convergence of measures is equivalent to convergence in the Prohorov metric (see, e.g. [Whi02], Section 3.2). Hence, assumption b) yields that for all time points $0 \leq t_0, \dots, t_l < \infty$, states $e_0, \dots, e_l \in E$ and any $\varepsilon > 0$ sufficiently small there exists a $\bar{\kappa} \in \mathbb{N}$ such that for all $\kappa \geq \bar{\kappa}$:

$$\mathbb{P}\left\{\eta^\kappa\left(\lfloor b_\kappa t_0 \rfloor\right) = e_0, \dots, \eta^\kappa\left(\lfloor b_\kappa t_l \rfloor\right) = e_l\right\} \geq \mathbb{P}\left\{\xi(t_0) = e_0, \dots, \xi(t_l) = e_l\right\} - \frac{\varepsilon}{2}.$$

Combining this with (30) we see that for all time points $0 \leq t_0 \leq \ldots \leq t_l < \infty$, states $e_0, \dots, e_l \in E$ and any $\varepsilon > 0$ sufficiently small there exists a $\bar{\kappa} \in \mathbb{N}$ such that for all $\kappa \geq \bar{\kappa}$

$$\begin{aligned}
&\mathbb{P}\left\{\xi^\kappa\left(\frac{b_\kappa t_0}{a_\kappa}\right) = e_0, \dots, \xi^\kappa\left(\frac{b_\kappa t_l}{a_\kappa}\right) = e_l\right\} \\
&\quad \geq \mathbb{P}\Bigg\{\eta^\kappa(\lfloor b_\kappa t_0 \rfloor) = e_0, \dots, \eta^\kappa(\lfloor b_\kappa t_l \rfloor) = e_l, \\
&\qquad\qquad d\left(\xi^\kappa\left(\frac{b_\kappa t_0}{a_\kappa}\right), \eta^\kappa(\lfloor b_\kappa t_0 \rfloor)\right) = \dots = d\left(\xi^\kappa\left(\frac{b_\kappa t_l}{a_\kappa}\right), \eta^\kappa(\lfloor b_\kappa t_l \rfloor)\right) = 0\Bigg\} \\
&\quad \geq \mathbb{P}\left\{\eta^\kappa(\lfloor b_\kappa t_0 \rfloor) = e_0, \dots, \eta^\kappa(\lfloor b_\kappa t_l \rfloor) = e_l\right\} - \frac{\varepsilon}{2} \\
&\quad \geq \mathbb{P}\left\{\xi(t_0) = e_0, \dots, \xi(t_l) = e_l\right\} - \varepsilon.
\end{aligned}$$

This implies the convergence of the finite dimensional distributions of $\left(\xi^\kappa\left(\frac{b_\kappa}{a_\kappa} t\right)\right)_{t\geq 0}$ to the finite dimensional distributions of $(\xi(t))_{t\geq 0}$ in the Prohorov metric and hence weakly, which completes the proof. □

5.4 Proof of the main results

5.4.1 Convergence of the coalescent processes

Let us apply these theoretical observations to the *ancestral material scaling limit* to the block-counting process of the seed bank coalescent defined in Definition 1.6

with vanishing migration rate c. If we simply let $c \to 0$, the limiting object will be a (block counting process of the) Kingman coalescent in the active population and a constant population of dormant individuals. However, if we speed up time by a factor $1/c \to \infty$, we obtain a new structure given in Definition 5.1, thus uncovering a separation of time-scales phenomenon. While the exchange of ancestral lineages between active and dormant states here becomes rare in the original timescale, in the new timescale, this migration will still happen at rate 1 while coalescences in the active population now occur almost instantaneously. Hence, in the limit, for each time $t > n$, there will be at most one active line.

Theorem 5.2 establishes the ancient ancestral lines process as scaling limit in finite dimensional distributions of the block-counting process of the seed bank coalescent.

Proof of 5.2. We prove the result using the machinery outlined in the previous chapter with $a_\kappa := c_\kappa^{-2}$ and $b_\kappa := c_\kappa^{-3}$. W.l.o.g. assume $c_\kappa \leq 1$, for all $\kappa \in \mathbb{N}$.

Step i) In analogy to our previous notation abbreviate

$$(\xi^\kappa(t))_{t\geq 0} := (N^{c_\kappa}(t), M^{c_\kappa}(t))_{t\geq 0}$$

and consider a discretized process with time steps of length $a_\kappa^{-1} = c_\kappa^2$ by letting

$$\eta^\kappa(i) := \xi^\kappa(ic_\kappa^2), \qquad i \in \mathbb{N}_0.$$

Recalling the rates of this processes as given in Definition 1.6

$$q_\kappa := \max_{(n,m)\in E_{(n_0,m_0)}} \left\{-\bar{A}^{c_\kappa}_{(n,m),(n,m)}\right\} \leq \binom{n_0+m_0}{2} + c_\kappa(n_0+m_0) + c_\kappa K(n_0+m_0)$$

whence (29) (and therefore (30)) hold as required.

Now, we claim that the transition probabilities of $(\eta^\kappa(i))_{i\in\mathbb{N}_0}$ are

$$\begin{aligned}
&\mathbb{P}\{\eta^\kappa(1) = (\bar{n},\bar{m}) \mid \eta^\kappa(0) = (n,m)\}\\
&\qquad = \mathbb{P}\left\{(N^{c_\kappa}(c_\kappa^2), M^{c_\kappa}(c_\kappa^2)) = (\bar{n},\bar{m}) \mid (N^{c_\kappa}(0), M^{c_\kappa}(0)) = (n,m)\right\}\\
&\qquad = \begin{cases}
\binom{n}{2}c_\kappa^2 + o(c_\kappa^3), & \text{if } \bar{n} = n-1,\ \bar{m} = m,\\
c_\kappa n c_\kappa^2 + o(c_\kappa^3), & \text{if } \bar{n} = n-1,\ \bar{m} = m+1,\\
c_\kappa K m c_\kappa^2 + o(c_\kappa^3), & \text{if } \bar{n} = n+1,\ \bar{m} = m-1,\\
1 - \binom{n}{2}c_\kappa^2 - c_\kappa n c_\kappa^2 - c_\kappa K m c_\kappa^2 + o(c_\kappa^3), & \text{if } \bar{n} = n,\ \bar{m} = m,\\
o(c_\kappa^3), & \text{otherwise.}
\end{cases}
\end{aligned}$$

for any sensible (n,m), $(\bar{n},\bar{m}) \in E_{(n_0,m_0)}$, recalling the convention of $\binom{n}{2} = 0$ for $n \leq 1$. This is proved as follows:

Proof. We first check that this discretisation is fine enough to guarantee that the probability of multiple jumps within the same discretisation period is sufficiently small. Indeed, recall that a jump in the process η^κ occurs only when there is either a coalescence event in the active population, a migration event from active to dormant state ("dormancy"), or a migration event from dormant to active state ("resuscitation"). Suppose η^κ is currently in state (n, m). The three events (coalescence, resuscitation and dormancy) happen respectively with rate $c_\kappa^2\binom{n}{2}, c_\kappa^3 Km$ and $c_\kappa^3 n$. Hence, using the Markov property, the waiting time T_1 for first jump is exponential with total rate $c_\kappa^2\binom{n}{2} + c_\kappa^3 Km + c_\kappa^3 n$. Further, the time T_2 for second event dominates an independent exponential random variable with rate

$$c_\kappa^2\Big(\binom{n+1}{2} + c_\kappa K(m+1) + c_\kappa(n+1)\Big).$$

Hence, we obtain

$$\begin{aligned}\mathbb{P}\Big\{\text{ two jumps within } c_\kappa^2 \text{ time units }\Big\} &\le \mathbb{P}\Big\{T_1 + T_2 < c_\kappa^2\Big\}\\ &\le \mathbb{P}\Big\{T_1 < c_\kappa^2, T_2 < c_\kappa^2\Big\}\\ &\le \Big(1 - \exp\Big[\Big(\binom{n}{2} + c_\kappa Km + c_\kappa n\Big)c_\kappa^2\Big]\Big)\\ &\qquad \times \Big(1 - \exp\Big[\Big(\binom{n+1}{2} + c_\kappa K(m+1) + c_\kappa(n+1)\Big)c_\kappa^2\Big]\Big).\end{aligned}$$

Developing the result in powers of c_κ yields

$$\binom{n}{2}\binom{n+1}{2}c_\kappa^4 + o(c_\kappa^4).$$

Since n, m are trivially bounded by $n_0 + m_0$, we see that the time τ_{c_κ} until at least two jumps fall into a discretisation interval of length c_κ^2 dominates a geometric random variable with success probability $C(n_0, m_0)c_\kappa^4 + o(c_\kappa^4)$ for some suitable constant $C(n_0, m_0)$, which proves the "otherwise" rates.
The other rates are then proved easily; for example,

$$\begin{aligned}&\mathbb{P}\{\eta^\kappa(1) = (n-1, m+1) \mid \eta^\kappa(0) = (n, m)\}\\ &\quad = \mathbb{P}\{\eta^\kappa(T_1) = (n-1, m+1), T_1 < 1, T_2 > 1\} + o(c_\kappa^3)\\ &\quad = \mathbb{P}\{\eta^\kappa(T_1) = (n-1, m+1) | T_1 < 1, T_2 > 1\}\mathbb{P}\{T_1 < 1, T_2 > 1\} + o(c_\kappa^3)\\ &\quad = \mathbb{P}\{\eta^\kappa(T_1) = (n-1, m+1)\}\big(1 - \mathbb{P}\{T_1 > 1\} + o(c_\kappa^3)\big) + o(c_\kappa^3)\\ &\quad = \frac{c_\kappa n}{\binom{n}{2} + c_\kappa Km + c_\kappa n}\Big(1 - \exp\big(c_\kappa^2\big(\binom{n}{2} + c_\kappa Km + c_\kappa n\big)\big)\Big) + o(c_\kappa^3)\\ &\quad = c_\kappa^3 n + o(c_\kappa^3).\end{aligned}$$

□

Step ii) Therefore, if Π_κ is defined as the transition matrix of $(\eta^\kappa(i))_{i\in\mathbb{N}_0}$, we obtain the decomposition

$$\Pi_\kappa = A_\kappa + \frac{B_\kappa}{b_\kappa}$$

with $b_\kappa = c_\kappa^{-3}$ as defined above and

$$(A_\kappa)_{(n,m),(\bar{n},\bar{m})} = \begin{cases} \binom{n}{2}c_\kappa^2, & \text{if } \bar{n} = n-1,\, \bar{m} = m,\\ 1 - \binom{n}{2}c_\kappa^2, & \text{if } \bar{n} = n,\, \bar{m} = m,\\ 0, & \text{otherwise,}\end{cases}$$

and

$$(B_\kappa)_{(n,m),(\bar{n},\bar{m})} = \begin{cases} n + o(1), & \text{if } \bar{n} = n-1,\ \bar{m} = m+1, \\ Km + o(1), & \text{if } \bar{n} = n+1,\ \bar{m} = m-1, \\ -n - Km + o(1), & \text{if } \bar{n} = n,\ \bar{m} = m, \\ o(1), & \text{otherwise.} \end{cases} \tag{31}$$

In order to apply the convergence result, we first need to check condition (26) for our set-up, which reads

$$\lim_{C\to\infty} \lim_{\kappa\to\infty} \sup_{r \geq Cc_\kappa^{-2}} \|(A_\kappa)^r - P\| = 0 \tag{32}$$

for P given in (23). Note that A_κ is a stochastic matrix and denote by $(Z^\kappa(r))_{r\in\mathbb{N}_0}$ the Markov chain associated to it. This is a pure death process in the first component and constant in the second. Then, by the definition of matrix norm, we get

$$\begin{aligned}\|(A_\kappa)^r - P\| &= \max_{(n,m)\in E_{(n_0,m_0)}} \sum_{(\bar{n},\bar{m})\in E_{(n_0,m_0)}} |(A_\kappa)^r_{(n,m),(\bar{n},\bar{m})} - P_{(n,m),(\bar{n},\bar{m})}| \\ &= \max_{n\geq 1, m\geq 0} \Big(|(A_\kappa)^r_{(n,m),(1,m)} - 1| + \sum_{\bar{n}=2}^{n} |(A_\kappa)^r_{(n,m),(\bar{n},m)} - 0| \Big) \\ &= \max_{n\geq 1, m\geq 0} 2\Big(1 - (A_\kappa)^r_{(n,m),(1,m)}\Big) \\ &= 2 \max_{n\geq 1, m\geq 0} \mathbb{P}\big\{Z^\kappa(r) \neq (1,m) \mid Z^\kappa_0 = (n,m)\big\}.\end{aligned}$$

Observe that for all $n \geq 2$ (and all $m \geq 0$)

$$A_{\kappa(n,m),(n-1,m)} = \binom{n}{2} c_\kappa^2 \geq c_\kappa^2.$$

Hence, since we start from n_0 active individuals, the number of time-steps required until full coalescence is dominated by the sum of $n_0 - 1$ independent geometric random variables $\gamma_1^\kappa, \ldots, \gamma_{n_0-1}^\kappa$ with success probability c_κ^2. By Markov's inequality, we get

$$\sup_{\kappa\in\mathbb{N}} \mathbb{P}\left\{\gamma_1^\kappa + \cdots + \gamma_{n_0-1}^\kappa \geq Cc_\kappa^{-2}\right\} \leq \frac{c_\kappa^2}{C}\mathbb{E}\big[\gamma_1^\kappa + \cdots + \gamma_{n_0-1}^\kappa\big] = \frac{(n_0-1)c_\kappa^2}{Cc_\kappa^2} = \frac{(n_0-1)}{C}$$

and with it

$$\lim_{C\to\infty} \lim_{\kappa\to\infty} \sup_{r \geq Cc_\kappa^{-2}} \|(A_\kappa)^r - P\| \leq \lim_{C\to\infty} \lim_{\kappa\to\infty} \sup_{r \geq Cc_\kappa^{-2}} \frac{(n_0-1)}{C} = 0$$

which gives (32). We are now left to establish the matrix-norm limit (27), that is, show that

$$\lim_{N\to\infty} PB_N P \qquad \text{exists} \tag{33}$$

and coincides with the G in Definition 5.1. For this, notice that B_κ converges for $\kappa \to \infty$ uniformly and in the matrix norm (recalling that the state space $E_{(n_0,m_0)}$ is finite), and define

$$B_{(n,m),(\bar{n},\bar{m})} \;:=\; \lim_{\kappa\to\infty}(B_\kappa)_{(n,m),(\bar{n},\bar{m})} \;=\; \begin{cases} n, & \text{if } \bar{n}=n-1,\ \bar{m}=m+1,\\ Km, & \text{if } \bar{n}=n+1,\ \bar{m}=m-1,\\ -n-Km, & \text{if } \bar{n}=n,\ \bar{m}=m,\\ 0, & \text{otherwise.} \end{cases}$$

Through careful entry-by-entry calculations one confirms

$$G = PBP \tag{34}$$

and thus (27). (Of course, this implies, as a consequence, that $PG = GP = G$, due to the fact that $P^2 = P$.) As described in the previous chapter, [BBE13, Lemma 1.7 and Remark 1.8] then yields

$$\lim_{\kappa\to\infty}\Pi_\kappa^{\lfloor tc_\kappa^{-3}\rfloor} = \lim_{\kappa\to\infty}\left(A_\kappa + c_\kappa^3 B_\kappa\right)^{\lfloor tc_\kappa^{-3}\rfloor} = Pe^{tG} =: \Pi(t) \qquad \text{for all } \ t>0,$$

which given $\eta^\kappa(0) = (N^{c_\kappa}(0), M^{c_\kappa}(0)) = (\tilde{N}(0),\tilde{M}(0))$ then implies

$$(\eta^\kappa(\lfloor c_\kappa^{-3}t\rfloor))_{t\ge0} \to (\tilde{N}(t),\tilde{M}(t))_{t\ge0} \quad \text{in finite dimensional distributions, as } \kappa\to\infty,$$

where $(\tilde{N}(t),\tilde{M}(t))_{t\ge0}$ is the ancient ancestral lines process defined in Definition 5.1.

Step iii) Since we have proven the necessary assumptions in Step i) and ii), Lemma 5.5 implies

$$\left(N^{c_\kappa}(c_\kappa^{-1}t), M^{c_\kappa}(c_\kappa^{-1}t)\right)_{t\ge0} = \left(\xi^{c_\kappa}\left(\frac{c_\kappa^{-3}}{c_\kappa^{-2}}t\right)\right)_{t\ge0} \longrightarrow \left(\tilde{N}(t),\tilde{M}(t)\right)_{t\ge0}$$

in finite dimensional distributions for $\kappa\to\infty$ and the proof is complete. □

We would now also like to observe similar scaling limits for the diffusion (1). As we saw in the case of genealogies, rescaling time may lead to a limiting process that is still Markovian, but whose semi-group is not standard. We can, however, use *moment duality* to obtain this limit.

5.4.2 Convergence in finite dimensional distributions from duality

We here present a general result on how to obtain convergence in finite dimensional distributions from moment duality and the analogous convergence of the dual process. This result is independent of whether time is rescaled, too, or not. It is, however, of particular interest in that case, since it might lead to limiting objects, that are rather "ill-behaved" and we will see examples in Chapter 5.4.3 where the limit does not have a generator, hence more standard ways of proving convergence through generator convergence fail.

For any vectors $n := (n_1,\dots,n_d)\in\mathbb{N}_0^d$ and $x := (x_1,\dots,x_d)\in[0,1]^d$, define the *mixed-moment function* $\mathfrak{m}$ as $\mathfrak{m}(x,n) := x_1^{n_1}\cdots x_d^{n_d}$.

Theorem 5.6. *Let $(\phi_\kappa(t))_{t\geq 0}$, $\kappa \in \mathbb{N}_0$, be a sequence of Feller Markov processes taking values in $[0,1]^d$ (for some $d \in \mathbb{N}$) and $(\xi^\kappa(t))_{t\geq 0}$, $\kappa \in \mathbb{N}_0$, a sequence of Markov chains with values in $\mathbb{N}_0^d$ such that they are pairwise moment duals, i.e.*

$$\forall \kappa \in \mathbb{N}_0\ , \forall t \geq 0\ \forall x \in [0,1]^d, n \in \mathbb{N}_0^d :\ \mathbb{E}_n[\mathfrak{m}(x, \xi^\kappa(t))]] = \mathbb{E}^x[\mathfrak{m}(\phi_\kappa(t), n)].$$

As usual, $\mathbb{P}_n$ and $\mathbb{P}^x$ denote the distributions for which ξ, resp. ϕ, start in n, resp. x.

If $(\xi^\kappa)_{\kappa\in\mathbb{N}_0}$ converges to some Markov chain ξ in finite dimensional distributions, *then there exists a Markov process ϕ with values in $[0,1]^d$ such that it is the limit in finite dimensional distributions of $(\phi_\kappa)_{\kappa\in\mathbb{N}_0}$ and the moment dual to ξ, i.e.*

$$\forall t \geq 0\ \forall x \in [0,1]^d, n \in \mathbb{N}_0^d :\ \mathbb{E}_n[\mathfrak{m}(x, \xi(t))]] = \mathbb{E}^x[\mathfrak{m}(\phi(t), n)]. \tag{35}$$

Remark 5.7. At a first glance one might suspect that this result should also hold in a more general set-up as long as the duality function used yields convergence determining families for the respective semi-groups. Indeed, most of the steps of the proof would still go through. However, note that we did not assume existence of a Markovian limit beforehand. For this we use the solvability of Hausdorff's moment problem on $[0,1]^d$ [KS13], which is a match to the moment duality function in our theorem.

Proof. The proof can roughly be split into three steps: We first use duality to prove the convergence of the *one-dimensional* distributions of $(\phi_\kappa)_{\kappa\in\mathbb{N}_0}$. This, together with the Markov property will give us the convergence of the *finite dimensional distributions* of $(\phi_\kappa)_{\kappa\in\mathbb{N}_0}$ to a family of limiting distributions. Then we prove *consistency* of the respective limiting measures and hence by Kolmogorov's Extension Theorem the existence of a *limiting process* ϕ, which must then be Markovian.

Since the mixed-moment function $\mathfrak{m}$ is continuous and bounded as a function on $\mathbb{N}_0^d$, the convergence of the finite dimensional distributions of $(\xi^\kappa)_{\kappa\in\mathbb{N}}$ and the assumed moment duality yield

$$\mathbb{E}^x[\mathfrak{m}(\phi_\kappa(t), n)] = \mathbb{E}_n[\mathfrak{m}(x, \xi_\kappa(t))] \xrightarrow{\kappa\to\infty} \mathbb{E}_n[\mathfrak{m}(x, \xi(t))] := \gamma(n, x, t) \tag{36}$$

for any $t \geq 0$, $x \in [0,1]^d$ and $n \in \mathbb{N}_0^d$. The unique solvability of the Hausdorff moment problem on $[0,1]^d$ [KS13] then gives the existence of a distribution $\mu^{x,t}$ on $([0,1], \mathfrak{B}([0,1]))$ (where $\mathfrak{B}$ is the Borel-σ-algebra) such that

$$\gamma(n, x, t) = \int_{[0,1]^d} \mathfrak{m}(\bar{x}, n) \mathrm{d}\mu^{x,t}(\bar{x}).$$

Since the polynomials are dense in the continuous functions, (36) implies the convergence of the one-dimensional distributions to $(\mu^{x,t})_{t\geq 0}$ (for each starting point $x \in [0,1]^d$).

To check the convergence in finite dimensional distributions, let P_κ be the probability transition function of ϕ_κ and recall that we assumed them to be Feller. For
$0 \leq t_1 < \ldots < t_l < \infty$, $x \in [0,1]^d$ and $n_1, \ldots, n_l \in [0,1]^d$ then observe

$$\begin{aligned}
&\mathbb{E}^x[\mathfrak{m}(\phi_\kappa(t_1), n_1) \cdots \mathfrak{m}(\phi_\kappa(t_l), n_l)] \\
&\quad = \int_{[0,1]} \int_{[0,1]} \cdots \int_{[0,1]} \mathfrak{m}(\bar{x}_1, n_1) \cdots \mathfrak{m}(\bar{x}_l, n_l) P_\kappa(\bar{x}_{l-1}, t_l - t_{l-1}, \mathrm{d}\bar{x}_l) \cdots P_\kappa(x, t_1, \mathrm{d}\bar{x}_1) \\
&\quad \xrightarrow{\kappa\to\infty}: \gamma(n_1, \ldots, n_l, x, t_1, \ldots, t_l).
\end{aligned} \tag{37}$$

The convergence (to some constant $\gamma(n_1, \ldots, n_l, x, t_1, \ldots, t_l)$) follows from the convergence of the one-dimensional distributions if one observes using the Lebesgue convergence theorem that weak convergence also implies the convergence of integrals when the integrand itself converges to a continuous and bounded function. Since our processes are Feller, we can iterate this argument and obtain the convergence above.

Again, by the unique solvability of the Hausdorff moment problem [KS13] we thus obtain the existence of a measure $\mu^{I,x}$ on $([0,1]^I, \mathfrak{B}([0,1])^{\otimes I})$ for any finite set of indices $I = \{t_1, \ldots, t_l\} \subset [0, \infty)$ and starting point $x \in [0,1]^d$ and (37) implies the convergence of the finite-dimensional distributions of $(\phi_\kappa)_{\kappa \in N}$ to a respective $\mu^{I,x}$. Since these $\mu^{I,x}$ are the limits of a consistent family they are themselves consistent and with Kolmogorov's Extension Theorem there exists a unique measure μ^x on the product-space $([0,1]^{[0,\infty)}, \mathfrak{B}([0,1])^{\otimes [0,\infty)})$ which is the distribution of the desired process ϕ. Its Markovianity follows from the respective property of the $(\phi_\kappa)_{\kappa \in \mathbb{N}_0}$

The duality of ϕ and ξ follows from the duality of the prelimiting processes. □

5.4.3 Convergence of the forwards-in-time processes

As an application of Theorem 5.6 we consider the diffusion (1) with the scaling regime of Chapter 5.4, namely, when the migration rate $c \to 0$ while simultaneously speeding up time by a factor $1/c \to \infty$ and obtain Theorem 5.3 stating the convergence of the rescaled diffusions to a Markovian limit $(\tilde{X}(t), \tilde{Y}(t))_{t \geq 0}$.

Proof of Theorem 5.3. Since the moment duality of the block-counting process of the seed bank coalescent and the seed bank diffusion [BGKWB16, Thm. 2.8] holds for every time $t \geq 0$, it is preserved for the time-changed processes $\left(N^{c_\kappa}(t/c_\kappa), M^{c_\kappa}(t/c_\kappa)\right)_{t \geq 0}$ and $\left(X^{c_\kappa}(t/c_\kappa), Y^{c_\kappa}(t/c_\kappa)\right)_{t \geq 0}$. Together with Theorem 5.2, all assumptions of Theorem 5.6 hold and we get the existence of a Markov process $(\tilde{X}(t), \tilde{Y}(t))_{t \geq 0}$ that is the dual of $(\tilde{N}(t), \tilde{M}(t))_{t \geq 0}$. The uniqueness of the moment dual of a Markov process proves that the limit does not depend on the choice of scaling sequence $(c_\kappa)_{\kappa \in \mathbb{N}_0}$. □

Moment duality now allows us to translate our knowledge about the ancient ancestral lines process $(\tilde{N}(t), \tilde{M}(t))_{t \geq 0}$ (Definition 5.1) to $(\tilde{X}(t), \tilde{Y}(t))_{t \geq 0}$. More precisely, since (35) holds in particular for $t > 0$, $m = 0$ and any $n \geq 1$, $x, y \in [0,1]$ we see

$$\begin{aligned}
\mathbb{E}^{x,y}[\tilde{X}(t)^n] &= \mathbb{E}_{n,0}[x^{\tilde{N}(t)} y^{\tilde{M}(t)}] \\
&= x\mathbb{P}_{n,0}\{\tilde{N}(t) = 1, \tilde{M}(t) = 0\} + y\mathbb{P}_{n,0}\{\tilde{N}(t) = 0, \tilde{M}(t) = 1\} \\
&= x(Pe^{tG})_{(n,0),(1,0)} + y(Pe^{tG})_{(n,0),(0,1)} = x(e^{tG})_{(1,0),(1,0)} + y(e^{tG})_{(1,0),(0,1)}.
\end{aligned}$$

We used the fact, that the first component of the ancient ancestral lines process immediately takes values in $\{0,1\}$ in the second equality and the definition of the projection in the last equality. Since the right-hand-side does not depend on $n \geq 1$, we can conclude that

$$\tilde{X}(t) \in \{0,1\} \quad \mathbb{P}^{x,y}\text{-a.s. for any } t > 0 \text{ and any } (x,y) \in [0,1]^2. \tag{38}$$

This small observation has an important consequence: Much like in the case of its dual $(\tilde{N}(t), \tilde{M}(t))_{t\geq 0}$, the transition function of the ancient ancestral lines process $(\tilde{X}(t), \tilde{Y}(t))_{t\geq 0}$ is not right-continuous in 0 and therefore $(\tilde{X}(t), \tilde{Y}(t))_{t\geq 0}$ does not have a classical generator.

Intuitively the reproduction mechanism (in the active population) acts so fast, that fixation (or extinction) *in the active population* happens instantaneously. Whenever there is an invasion from the seed bank of the type extinct in the active population, its chances of fixating are proportional to the fraction of the type in the *dormant* population. The limit is therefore a pure jump process in the active component that moves between the states 0 and 1, while the seed bank component retains its classical behavior. We can formalize this observation if we restrict the process to the smaller state space $\{0, 1\} \times [0, 1]$, see Proposition 5.9 below.

Definition 5.8. Let $(\bar{N}(t), \bar{M}(t))_{t\geq 0}$ be the Markov chain on $\{0, 1\} \times \mathbb{N}_0$ given by the (conservative) Q-matrix

$$\bar{G}_{(n,m),(\bar{n},\bar{m})} = \begin{cases} Km, & \text{if } \bar{n} = 1,\ n \in \{0, 1\},\ \bar{m} = m - 1, \\ 1, & \text{if } \bar{n} = 0,\ n = 1,\ \bar{m} = m + 1, \\ -1 - Km, & \text{if } \bar{n} = n,\ \bar{m} = m, \\ 0, & \text{otherwise.} \end{cases}$$

for any $(n, m), (\bar{n}, \bar{m}) \in \{0, 1\} \times \mathbb{N}_0$.

On the other hand, let $(\bar{X}(t), \bar{Y}(t))_{t\geq 0}$ be the Markov process on $\{0, 1\} \times [0, 1]$ with generator

$$\begin{aligned} \bar{\mathcal{A}}^{(1)} f(x, y) = y(f(1, y) - f(0, y))\mathbb{1}_{\{0\}}(x) + (1 - y)(f(0, y) - f(1, y))\mathbb{1}_{\{1\}}(x) \\ + K(x - y)\frac{\partial f}{\partial y}(x, y). \end{aligned}$$

Note that $\bar{G}$ given above is the restriction of G, the "generator" of $(\tilde{N}(t), \tilde{M}(t))_{t\geq 0}$ from Definition 5.1. Indeed, these processes are essentially the ancestral material processes when started in the smaller state-space:

Proposition 5.9. *The processes $(\bar{N}(t), \bar{M}(t))_{t\geq 0}$ and $(\bar{X}(t), \bar{Y}(t))_{t\geq 0}$ from Definition 5.8 are moment duals, i.e.*

$$\forall t \geq 0\ \forall (x, y) \in [0, 1]^2, (n, m) \in \mathbb{N}_0^2 :\ \mathbb{E}_{n,m}\left[x^{\bar{N}(t)} y^{\bar{M}(t)}\right] = \mathbb{E}^{x,y}\left[\bar{X}(t)^n \bar{Y}(t)^m\right]. \quad (39)$$

$(\bar{N}(t), \bar{M}(t))_{t\geq 0}$ coincides in distribution with $(\tilde{N}(t), \tilde{M}(t))_{t\geq 0}$ if (both are) started in the reduced state-space $\{0, 1\} \times \mathbb{N}_0$.

Likewise, $(\bar{X}(t), \bar{Y}(t))_{t\geq 0}$ coincides in distribution with $(\tilde{X}(t), \tilde{Y}(t))_{t\geq 0}$ if (both are) started in the reduced state-space $\{0, 1\} \times [0, 1]$.

Moment duality of the involved processes will be important for the proof of the last statement.

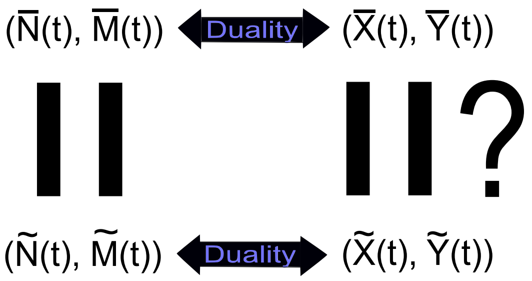


Figure 7: Strategy of the proof of Proposition 5.9. The moment duality of $(\tilde{N}(t), \tilde{M}(t))_{t\geq 0}$ and $(\tilde{X}(t), \tilde{Y}(t))_{t\geq 0}$ is consequence of Theorem 5.3. The semigroups of $(\tilde{N}(t), \tilde{M}(t))_{t\geq 0}$ and $(\bar{N}(t), \bar{M}(t))_{t\geq 0}$ coincide when started in the reduced state-space $\{0,1\} \times \mathbb{N}_0$. We prove the moment duality of $(\bar{N}(t), \bar{M}(t))_{t\geq 0}$ and $(\bar{X}(t), \bar{Y}(t))_{t\geq 0}$, which allows us to conclude, that the semigroups of $(\tilde{X}(t), \tilde{Y}(t))_{t\geq 0}$ and $(\bar{X}(t), \bar{Y}(t))_{t\geq 0}$ also agree when started in $\{0,1\} \times [0,1]$.

Proof. The duality of $(\bar{N}(t), \bar{M}(t))_{t\geq 0}$ and $(\bar{X}(t), \bar{Y}(t))_{t\geq 0}$ can be proven through the standard method of generator calculations: Applying $\bar{\mathcal{A}}^{(1)}$ to $S((x,y),(n,m)) := x^n y^m$ for $(n,m) \in \{0,1\} \times \mathbb{N}_0$ and $(x,y) \in \{0,1\} \times [0,1]$, as a function in (x,y) yields

$$\begin{aligned}\bar{\mathcal{A}}^{(1)}(S)((x,y),(n,m)) &= y(y^m - 0^n y^m)\mathbb{1}_{\{0\}}(x) + (1-y)(0^n y^m - y^m)\mathbb{1}_{\{1\}}(x)\\ &\quad + K(x-y)x^n m y^{m-1}\\ &= -(x-y)(y^m - 0^n y^m)\mathbb{1}_{\{0\}}(x) + (x-y)(0^n y^m - y^m)\mathbb{1}_{\{1\}}(x)\\ &\quad + K(x-y)x^n m y^{m-1}\\ &= -n(x-y)y^m + Km(x-y)x^n y^{m-1}\\ &= Kmx^{n+1}y^{m-1} + (-Kmx^n - nx)y^m + ny^{m+1}\end{aligned}$$

where we continue to use $0^0 = 1$, the fact that $n \in \{0,1\}$ and simply sorted the terms by powers of y for easier comparison in the last line.

On the other hand, if we apply $\mathcal{A}^{(1)}$, the generator of the stochastic process having $\bar{G}$ as Q-matrix, to S as a function in $(n,m) \in \{0,1\} \times N_0$, we get

$$\begin{aligned}\mathcal{A}^{(1)}(S)((x,y),(n,m)) &= Km(xy^{m-1} - x^n y^m) + 1(y^{m+1} - xy^m)\mathbb{1}_{\{1\}}(n)\\ &= Km(xy^{m-1} - xy^m)\underbrace{\mathbb{1}_{\{1\}}(n)}_{=n} + Km(xy^{m-1} - y^m)\underbrace{\mathbb{1}_{\{0\}}(n)}_{=1-n}\\ &\quad + 1(y^{m+1} - xy^m)\underbrace{\mathbb{1}_{\{1\}}(n)}_{=n}\\ &= Kmxy^{m-1} + (-Kmnx - Km(1-n) - nx)y^m + ny^{m+1}.\end{aligned}$$

A close look noting that for our choices of variables we have $x^{n+1} = x$ and $nx + (1-n) = x^n$ shows that the two coincide. Since:

- the duality function $S : [0,1]^2 \times \mathbb{N}_0^2 \to \mathbb{R}$, $S(x,y,n,m) = x^n y^m$, is bounded and continuous;
- the functions $S(x,y,\cdot)$ and $P_t S(x,y,\cdot)$, that is, the functions $(n,m) \to x^n y^m$ and $(n,m) \to \mathbb{E}_{(x,y)}[\bar{X}^n(t)\bar{Y}^m(t)]$, are in the domain of $\mathcal{A}^{(1)}$;

- the functions $S(\cdot, n, m)$ and $Q_t S(\cdot, n, m)$, that is, the functions $(x, y) \to x^n y^m$ and $(x, y) \to \mathbb{E}_{(n,m)}[x^{\bar{N}(t)} y^{\bar{M}(t)}]$, are in the domain of $\bar{\mathcal{A}}^{(2)}$;

with $(P_t)_{t \geq 0}$ and $(Q_t)_{t \geq 0}$ transition semigroups of $(\bar{X}, \bar{Y})$ and $(\bar{N}, \bar{M})$ respectively, [JK14, Prop. 1.2] concludes the proof of duality.

Let $\bar{f} : \{0, 1\} \times \mathbb{N}_0 \to \mathbb{R}$ and define f as $f(n, m) := \bar{f}(n, m)$ for $(n, m) \in \{0, 1\} \times \mathbb{N}_0$ and 0 otherwise. Recall that Pe^{tG} is the semi-group of $(\tilde{N}(t), \tilde{M}(t))_{t \geq 0}$ from Definition 5.1. Since $G_{(n,m),(\bar{n},\bar{m})} = \bar{G}_{(n,m),(\bar{n},\bar{m})}$ for any $(n, m) \in \{0, 1\} \times \mathbb{N}_0$, we have

$$Gf(n, m) = \bar{G}\bar{f}(n, m), \qquad (n, m) \in \{0, 1\} \times \mathbb{N}_0.$$

As we also know that $G = PBP$ from (34), it follows that

$$Gf(n, m) = PBPf(n, m) \in \{0, 1\} \times \mathbb{N}_0$$

whence the semi-groups of $(\tilde{N}(t), \tilde{M}(t))_{t \geq 0}$ and $(\bar{N}(t), \bar{M}(t))_{t \geq 0}$ coincide on such f and the two processes are equal in distribution when started in $(n, m) \in \{0, 1\} \times \mathbb{N}_0$, as claimed.

This also implies

$$\mathbb{E}^{x,y}\left[\tilde{X}(t)^n \tilde{Y}(t)^m\right] = \mathbb{E}_{n,m}\left[x^{\tilde{N}(t)} y^{\tilde{M}(t)}\right] = \mathbb{E}_{n,m}\left[x^{\bar{N}(t)} y^{\bar{M}(t)}\right] = \mathbb{E}^{x,y}\left[\bar{X}(t)^n \bar{Y}(t)^m\right] \tag{40}$$

for all $t \geq 0$ and all $(x, y) \in \{0, 1\} \times [0, 1]$ and $(n, m) \in \{0, 1\} \times \mathbb{N}_0$, where we used the dualities from Theorem 5.3 and Proposition 5.9 in the first, respectively last equality.

Recall from (38), that for any $t > 0$ we have $(\tilde{X}(t), \tilde{Y}(t)) \in \{0, 1\} \times [0, 1]$, $\mathbb{P}^{x,y}$-a.s., $(x, y) \in [0, 1]^2$. Since a distribution on $\{0, 1\} \times [0, 1]$ is uniquely determined by its moments of order $(n, m) \in \{0, 1\} \times \mathbb{N}_0$, (40) implies that $(\tilde{X}(t), \tilde{Y}(t)) \sim (\bar{X}(t), \bar{Y}(t))$ for any $t > 0$ (when started in the same $(x, y) \in \{0, 1\} \times [0, 1]$). Since they are both Markovian, this implies that the distributions of $(\bar{X}(t), \bar{Y}(t))_{t \geq 0}$ and $(\tilde{X}(t), \tilde{Y}(t))_{t \geq 0}$ coincide when started in the reduced state-space $\{0, 1\} \times [0, 1]$. □

5.5 Further scaling limits

5.5.1 Imbalanced island size

One can, of course, transfer the previous results from the seed bank directly to the two-island model without mutation. From the similarity of the models it is clear that here, too, scaling the migration rate to $c \to 0$ while speeding up time by $1/c \to \infty$, one will obtain a model with instantaneous coalescences in *both* islands similar to the ancient ancestral lines process from Definition 5.1.

However, one could also consider a two-island model with *different* scalings of the coalescence rates in the islands. We recall that the diffusion process related to the two-island diffusion, describing the frequency process of allele a in the two-allele case, is defined as the solution of the SDEs

$$\begin{aligned} \mathrm{d}X(t) &= c(Y(t) - X(t))\mathrm{d}t + \alpha\sqrt{X(t)(1 - X(t))}\mathrm{d}W(t), \\ \mathrm{d}Y(t) &= cK(X(t) - Y(t))\mathrm{d}t + \alpha'\sqrt{X(t)(1 - X(t))}\mathrm{d}W'(t), \end{aligned} \tag{41}$$

The parameters α resp. α' are associated to the notion of *effective population size* (cf. [Wak09]) so a different scaling corresponds to a significant difference in population size on the two islands. For this we will not only scale the migration rate $c \to 0$ and time as before, but in addition assume that the coalescence rate $\alpha' > 0$ in the second island scales as c, i.e. $\alpha'/c \to 1$. The result is a two-island model with instantaneous coalescences in the first island, but otherwise 'normal' migration and coalescence behavior in the second. For more precision, let $(n_0, m_0) \in \mathbb{N}_0 \times \mathbb{N}_0$ and $(\tilde{N}(t), \tilde{M}(t))_{t\geq 0}$ be the continuous-time Markov chain with initial value $(\tilde{N}(0), \tilde{M}(0)) = (n_0, m_0)$, taking values in the state space $E_{(n_0,m_0)} := \{0, \dots, n_0 + m_0\}^2$, with transition matrix $\Pi(t) := Pe^{tG}$, for $t > 0$ and $\Pi(0)$ equal to the identity on E, where P is given by (23) as before and G is a Q-matrix of the form

$$G_{(n,m),(\bar{n},\bar{m})} := \begin{cases} Km + \binom{m}{2}, & \text{if } \bar{n} = 1,\, n \geq 1,\, \bar{m} = m - 1, \\ Km, & \text{if } \bar{n} = 1,\, n = 0,\, \bar{m} = m - 1, \\ \binom{m}{2}, & \text{if } \bar{n} = 0,\, n = 0,\, \bar{m} = m - 1, \\ 1, & \text{if } \bar{n} = 0,\, n \geq 1,\, \bar{m} = m + 1, \\ -\binom{m}{2} - 1 - Km, & \text{if } \bar{n} = 1,\, n \geq 1,\, \bar{m} = m, \\ -\binom{m}{2} - Km, & \text{if } \bar{n} = n = 0,\, \bar{m} = m, \\ 0, & \text{otherwise.} \end{cases}$$

where the first two rates given in the definition of G correspond to the events of dormancy (potentially with immediate coalescence) and the fourth one corresponds to resuscitation.

The following theorem establishes this Markov chain as scaling limit of a two-island model.

Theorem 5.10. *Denote by $(N^{c,\alpha'}(t), M^{c,\alpha'}(t))_{t\geq 0}$ the block counting process of the structured coalescent (without mutation) as defined in Definition 1.6 with migration rate $c > 0$ and coalescence rate $\alpha' > 0$ in the second island. Assume that it starts in some $(n_0, m_0) \in \mathbb{N} \times \mathbb{N}$, $\mathbb{P}$-a.s..*

Furthermore let $(\tilde{N}(t), \tilde{M}(t)))_{t\geq 0}$ be as defined above with the same initial condition. Then, for any sequence of migration rates $(c_\kappa)_{\kappa\in\mathbb{N}}$ and any sequence of coalescence rates $(\alpha'_\kappa)_{\kappa\in\mathbb{N}}$ with $c_\kappa \to 0$ and $c_\kappa/\alpha'_\kappa \to 1$ for $\kappa \to \infty$

$$\left(N^{c_\kappa,\alpha'_\kappa}\left(\frac{1}{c_\kappa}t\right), M^{c_\kappa,\alpha'_\kappa}\left(\frac{1}{c_\kappa}t\right)\right)_{t\geq 0} \xrightarrow{f.d.d.} \left(\tilde{N}(t), \tilde{M}(t)\right)_{t\geq 0},$$

in finite dimensional distributions as $\kappa \to \infty$.

Proof. The proof is analogous to that of Theorem 5.2, whence we shorten it significantly. Consider, again, the sequences $a_\kappa := c_\kappa^{-2}$ and $b_\kappa := c_\kappa^{-3}$. W.l.o.g. assume $c_\kappa \leq 1$, for all $\kappa \in \mathbb{N}$.

Step i) As before we abbreviate $(\xi^\kappa(t))_{t\geq 0} := (N^{c_\kappa,\alpha'_\kappa}(t), M^{c_\kappa,\alpha'_\kappa}(t))_{t\geq 0}$ and consider a discretized process with time steps of length $a_\kappa^{-1} = c_\kappa^2$ by letting

$$\eta^\kappa(i) := \xi^\kappa(ic_\kappa^2), \qquad i \in \mathbb{N}_0.$$

Recalling the rates of this processes as given in Definition (1.6)

$$\begin{aligned} q_\kappa &:= \max_{(n,m)\in E_{(n_0,m_0)}} \left\{ A^{c_\kappa}_{(n,m),(n,m)} \right\} \\ &\leq (\alpha + \alpha'_\kappa)\binom{n_0+m_0}{2} + c_\kappa(n_0+m_0) + c_\kappa K(n_0+m_0) \end{aligned}$$

whence (29) (and therefore (30)) hold as required.

Step ii) Let Π_κ be the transition matrix of $(\eta^\kappa(i))_{i\in\mathbb{N}_0}$. One can calculate the transition probabilities to obtain the decomposition

$$\Pi_\kappa = A_\kappa + \frac{B_\kappa}{b_\kappa}$$

with $b_\kappa = c_\kappa^{-3}$, A_κ as in (5.4.1) and

$$(B_\kappa)_{(n,m),(\bar{n},\bar{m})} = \begin{cases} \binom{m}{2} + o(1), & \text{if } \bar{n}=n,\ \bar{m}=m-1, \\ n+o(1), & \text{if } \bar{n}=n-1,\ \bar{m}=m+1, \\ Km+o(1), & \text{if } \bar{n}=n+1,\ \bar{m}=m-1, \\ -\binom{m}{2} - n - Km + o(1), & \text{if } \bar{n}=n,\ \bar{m}=m, \\ o(1), & \text{otherwise.} \end{cases} \tag{42}$$

Since A_κ and P are the same as in the proof of Theorem 5.2, we already know that (26) holds. Much like before, the matrix-norm limit (27) exists because B_κ converges for $\kappa\to\infty$ uniformly and in the matrix norm to

$$B := \begin{cases} \binom{m}{2}, & \text{if } \bar{n}=n,\ \bar{m}=m-1, \\ n, & \text{if } \bar{n}=n-1,\ \bar{m}=m+1, \\ Km, & \text{if } \bar{n}=n+1,\ \bar{m}=m-1, \\ -\binom{m}{2} - n - Km, & \text{if } \bar{n}=n,\ \bar{m}=m, \\ 0, & \text{otherwise.} \end{cases}$$

Through careful calculations one confirms $G = PBP$. [BBE13, Lemma 1.7 and Remark 1.8] then yields

$$\lim_{\kappa\to\infty} \Pi_\kappa^{\lfloor tc_\kappa^{-3}\rfloor} = \lim_{\kappa\to\infty} \left(A_\kappa + c_\kappa^3 B_\kappa\right)^{\lfloor tc_\kappa^{-3}\rfloor} = Pe^{tG} =: \Pi(t) \qquad \text{for all } t>0,$$

which given $\eta^\kappa(0) = (N^{c_\kappa}(0), M^{c_\kappa}(0)) = (\tilde{N}(0), \tilde{M}(0))$ then implies

$$(\eta^\kappa(\lfloor c_\kappa^{-3}t\rfloor))_{t\geq 0} \xrightarrow{\text{f.d.d.}} (\tilde{N}(t), \tilde{M}(t))_{t\geq 0} \quad \text{as } \kappa\to\infty.$$

Step iii) Since we have proven the necessary assumptions in Step i) and ii), Lemma 5.5 implies

$$\left(N^{c_\kappa,\alpha'_\kappa}(c_\kappa^{-1}t), M^{c_\kappa,\alpha'_\kappa}(c_\kappa^{-1}t)\right)_{t\geq 0} = \left(\xi^{c_\kappa}\left(\frac{c_\kappa^{-3}}{c_\kappa^{-2}}t\right)\right)_{t\geq 0} \xrightarrow{\text{f.d.d.}} \left(\tilde{N}(t), \tilde{M}(t)\right)_{t\geq 0}$$

for $\kappa\to\infty$ and the proof is complete. □

Moreover, we have a forwards-in-time result as well:

Proposition 5.11. *Denote by* $(X^{c,\alpha'}(t), Y^{c,\alpha'}(t))_{t\geq 0}$ *the* two-island diffusion *with migration rate* $c > 0$ *and island 2 of size* $\alpha' > 0$ *and assume that it starts in some* $(x, y) \in [0,1]^2$, $\mathbb{P}$*-a.s.. Then, the sequence* $(X^{c_\kappa,\alpha'_\kappa}(t), Y^{c_\kappa,\alpha'_\kappa}(t))_{t\geq 0}$ *will converge to a Markovian degenerate limit coinciding in distribution with a Markov process with generator*

$$\begin{aligned}\bar{\mathcal{A}}^{(1)} f(x,y) = y(f(1,y) - f(0,y))\mathbb{1}_{\{0\}}(x) + (1-y)(f(0,y) - f(1,y))\mathbb{1}_{\{1\}}(x) \\ + K(x-y)\frac{\partial f}{\partial y}(x,y) + \frac{1}{2}x(1-x)\frac{\partial^2}{\partial x^2}f(x,y)\end{aligned}$$

whenever started in the smaller state-space $\{0,1\} \times [0,1]$.

The proof of this proposition is analogous to that of Theorem 5.3 as well and will be omitted.

5.5.2 Other

There are several other possibilities to rescale the parameters of the process given in Definition 1.6 in order to obtain sensible limits. In particular, there are many options in which, in contrast to the previous two examples, *time is not scaled.* These scaling regimes can also be treated rigorously with similar methods. However, we refrain from providing full technical details, and conclude this chapter with a heuristic discussion of some of these limits.

A) Instead of having the migration rate $c \to 0$ as in the previous chapter one can also let $c \to \infty$, but without changing time.

This scenario was investigated for the two-island model (with parameters $\alpha' = \alpha = K = 1$, $c = c'$, and no mutation) by Nath and Griffiths in [NG93]. They show that the rates of the block counting process converge to those of a time-changed Kingman coalescent, where the time-change is given by a constant delay by a factor $\kappa = 1/2$. This is rather intuitive: Two lines which switch at high (infinite) rate between the two islands may merge only if both are in the same island simultaneously, which happens roughly 50% of the time. A similar effect in the seed bank scenario ($\alpha' = 0, \alpha = 1$, and no mutation) was shown by Lambert and Ma (Theorem 3.2 in [LM15]), where, for $K = 1$ the constant is $\kappa = 1/4$. Again, this follows the same intuition as in the two-island case, since now coalescence is only possible in one of the subpopulations, so roughly 25% of the time. The following result is just a simple consequence of this Theorem.

Corollary 5.12. *Denote by* $(N^c(t), M^c(t))_{t\geq 0}$ *the block-counting process of the seed bank coalescent (without mutation) with migration rate* $c > 0$ *(and relative seed bank size* $K > 0$*) and by* $(K(t))_{t\geq 0}$ *the block counting process of the standard Kingman coalescent. Then,*

$$\big(N^c(t) + M^c(t)\big)_{t\geq 0} \xrightarrow{w} \big(K(\kappa t)\big)_{t\geq 0}$$

as $c \to \infty$*, where* $\kappa = (\frac{K}{K+1})^2$.

Proof. From [LM15, Thm. 3.2] we know that the processes $\big(N^c(t)+M^c(t)\big)_{t\geq 0}$ converge weakly to a coalescent with the following coalescence rates: If there are l individuals left, the next coalescence happens at rate

$$c_l := \sum_{j=2}^{l} \frac{j(j-1)}{2}\binom{l}{j}\big(\frac{K}{1+K}\big)^j\big(\frac{1}{1+K}\big)^{l-j}.$$

Note that if ξ denotes a binomial random variable with parameters l and $\frac{K}{1+K}$, this can be rewritten as

$$\begin{aligned} c_l = \frac{1}{2}\left(\mathbb{E}[\xi^2]-\mathbb{E}[\xi]\right) &= \frac{1}{2}\left(l\frac{K}{1+K}\frac{1}{1+K}+l^2\left(\frac{K}{1+K}\right)^2 - l\frac{K}{1+K}\right) \\ &= \frac{1}{2}\left((l^2-l)\left(\frac{K}{1+K}\right)^2\right) = \binom{l}{2}\left(\frac{K}{1+K}\right)^2 = \binom{l}{2}\kappa, \end{aligned}$$

which are the transition rates of the time-changed Kingman coalescent. □

Note that in [LM15] the rates lack the factor 1/2. Their *peripatric* coalescent differs from the seed bank coalescent in that the rate for j lineages to coalesce is $j(j-1)$ instead of $j(j-1)/2=\binom{j}{2}$ which originates in their usage of the Moran model in the prelimit and the lack of a factor 2 in the time-rescaling. Hence, their Theorem 3.2 formulated for our seed bank coalescent has the additional factor 1/2.
For the diffusions, we get a Wright-Fisher diffusion with a prefactor in the first component, while the second component satisfies $Y(t)\equiv X(t)$. The limit is in finite-dimensional distributions.

B) For the seed bank coalescent without mutation, the standard Kingman coalescent (with no time-change) arises as scaling limit when $K\to\infty$ (and all other parameters - including time - are fixed). The intuition behind this fact is that as $K\to\infty$, lineages that become inactive resuscitate almost immediately since as the seed bank gets very small, the migration rate from the seed bank cK goes to infinity. So the overall effect of such a vanishing seed bank becomes negligible as lineages are essentially always active and thus coalesce as usually for a Kingman coalescent: The proof, which follows standard techniques, is omitted here.

For the diffusions, we get a Wright-Fisher diffusion in the first component, while the second component satisfies $Y(t)\equiv X(t)$ here as well. The limit is in finite-dimensional distributions.

C) Again, consider the seed bank coalescent without mutation, but this time let $K\to 0$, i.e. the migration rate out of the seed bank $cK\to 0$ (while all other parameters including time are fixed). Here, heuristically, if a lineage becomes inactive, it will not resuscitate for a very long time (large seed bank) and in the limit gets completely stuck. The process thus converges to the *coalescent with freeze* introduced by Dong, Gnedin and Pitman in [DGP07], which precisely describes this situation.

To prove this result it is enough to observe that the moment dual of $(X(t),Y(t))_{t\geq 0}$ is precisely the block counting process of the coalescent with freeze $(N(t),M(t))_{t\geq 0}$, where $M(t)$ is the number of ancestral lines that had been frozen until time t and

$N(t)$ is the number of (not frozen) ancestors at time t. The convergence of $(N^K(t), M^K(t))_{t\geq 0}$ follows a generator calculation, and the result is completed applying Theorem 5.6.

Regarding the limit of the diffusions, the first component is the well-known jump diffusion, while the second component is constant: $Y(t) \equiv Y(0)$ for all $t \geq 0$. The convergence here is weakly on the path space: Corollary 4.8.7 of [EK86] can be applied, so that we get uniform bounds on the generator convergence.

D) Regarding the two-island model with $\alpha = 1$ and no mutation, convergence to the seed bank model (without mutation) holds for $\alpha' \to 0$ (if all other parameters and time are fixed). Intuitively, this follows because the rate of coalescence in one of the two islands becomes negligible. The limit of the corresponding diffusion is obviously the seed bank diffusion and we have weak convergence on the path space, with the [EK86] condition easily checked.

6 Measures of population structure

The aim of this chapter is to provide basic statistical theory to analyze patterns of population structure and genetic variability produced by seed banks. We will also provide tools for parameter estimation and model selection based on genetic data. Notably, we will provide comparisons between patterns of variability under seed banks, and classical models of population structure (as in [Her94]). Both model classes predict similar patterns in diversity, and we will study the extent to which sequence data can differentiate between them. This extends earlier studies ([TLL$^+$11], [BGE$^+$15]), where seed banks were compared to classical, panmictic models. This chapter is organized as follows. We consider several classical statistical quantities: moments at stationarity (which give us a sampling formula), the sample heterozygosity and Wright's F_{ST} (as in [Wri49]), and the site frequency spectrum (normalized or not). These measures are informative about the underlying coalescent scenario, and suited to the different mutation models, to varying degrees. They also differ in the extent to which they are tractable. The sample heterozygosity, Wright's F_{ST} and the (normalized) SFS discard statistical signal, but are readily computed (at least numerically) in many settings. The likelihood function obtained from the sampling distribution fully captures the statistical signal in a data set, but is available for coalescent processes only via computationally intensive Monte Carlo schemes, as can be seen in detail in [K3]. Our results clarify when computationally cheaper summary statistics suffice to distinguish between models, and when the full likelihood is needed. This chapter is based on [K3] (Subchapters 6.2, 6.4, 6.5, 6.7, 6.8).

6.1 Linear Algebra of moments

Since the sample heterozygosities in the two-allele model involve mixed moments in stationarity, we must first derive closed-form formulas for those. The recursive formula (9) introduced in Chapter 3, however, easily allows us to find the ($(n+1)$-dimensional) column vectors $M_n := (M_{0n}, M_{1,n-1}, \dots, M_{n0})^T$:

Theorem 6.1. *Defining the $(n+1) \times (n+1)$ matrix $B_n = (B_n)_{ij}$*[16] *and the $(n+1) \times n$ matrix $A_n = (A_n)_{ij}$ by*

$$(B_n)_{ij} = -ci\delta_{i-1,j} - c'(n-i)\delta_{i+1,j} + D_{j,n-j}\delta_{ij}$$

and

$$(A_n)_{ij} = a_i\delta_{i-1,j} + a'_{i+1}\delta_{ij},$$

recalling from Chapter 3.2 that

$$\begin{aligned}
a_n &:= \alpha^2\binom{n}{2} + nu_2, \\
a'_m &:= (\alpha')^2\binom{m}{2} + mu'_2, \text{ and} \\
D_{n,m} &:= \alpha^2\binom{n}{2} + (\alpha')^2\binom{m}{2} + (u_2+u_1)n + (u'_1+u'_2)m + cn + c'm,
\end{aligned}$$

[16] Here, all indices start at 0.

(with the convention $\binom{1}{m} = \binom{0}{m} = 0$ for all m), we have that

$$M_n = B_n^{-1} A_n B_{n-1}^{-1} A_{n-1} \dots B_1^{-1} A_1.$$

Proof. Because of the definition of B_n and of A_n and of Lemma 3.7, we know that

$$B_n M_n = A_n M_{n-1}.$$

Moreover, using a similar argument as in Theorem 3.9, all rows of B_n are strictly dominant and therefore the matrix is invertible. Therefore, by recursion,

$$M_n = B_n^{-1} A_n B_{n-1}^{-1} A_{n-1} \dots B_1^{-1} A_1.$$

□

From now on, we will assume $c' = Kc$. We now give some examples for the matrices A_i and B_i, in case $\mathtt{S}$[17] with $u_1 = u'_1$ and $u_2 = u'_2$:

$$B_1 = \begin{bmatrix} D_{01} & -cK \\ -c & D_{10} \end{bmatrix} = \begin{bmatrix} u'_2 + u'_1 + Kc & -cK \\ -c & u_2 + u_1 + c \end{bmatrix}, \qquad A_1 = \begin{bmatrix} a'_1 \\ a_1 \end{bmatrix}$$

$$B_2 = \begin{bmatrix} D_{02} & -2Kc & 0 \\ -c & D_{11} & -Kc \\ 0 & -2c & D_{20} \end{bmatrix}, \qquad A_2 = \begin{bmatrix} a'_2 & 0 \\ a_1 & a'_1 \\ 0 & a_2 \end{bmatrix}$$

$$B_3 = \begin{bmatrix} D_{03} & -3Kc & 0 & 0 \\ -c & D_{12} & -2Kc & 0 \\ 0 & -2c & D_{21} & -Kc \\ 0 & 0 & -3c & D_{30} \end{bmatrix}, \qquad A_3 = \begin{bmatrix} a'_3 & 0 & 0 \\ a_1 & a'_2 & 0 \\ 0 & a_2 & a'_1 \\ 0 & 0 & a_3 \end{bmatrix}$$

$$B_4 = \begin{bmatrix} D_{04} & -4Kc & 0 & 0 & 0 \\ -c & D_{13} & -3Kc & 0 & 0 \\ 0 & -2c & D_{22} & -2Kc & 0 \\ 0 & 0 & -3c & D_{31} & -Kc \\ 0 & 0 & 0 & -4c & D_{40} \end{bmatrix}, \qquad A_4 = \begin{bmatrix} a'_4 & 0 & 0 & 0 \\ a_1 & a'_3 & 0 & 0 \\ 0 & a_2 & a'_2 & 0 \\ 0 & 0 & a_3 & a'_1 \\ 0 & 0 & 0 & a_4 \end{bmatrix}$$

This yields, in the case $\mathtt{S}$ with $u_1 = u'_1$ and $u_2 = u'_2$, using the previous calculations, the following M's:

$$M_1^{\mathtt{S}} = \begin{bmatrix} \frac{u_2}{u_1+u_2} & \frac{u_2}{u_1+u_2} \end{bmatrix}',$$

and for the second order moments, assuming that $c = K = 1$,

$$M_{20}^{\mathtt{S}} = \frac{u_2(4u_1^2u_2 + 8u_1u_2^2 + 14u_1u_2 + 4u_2^3 + 14u_2^2 + 12u_2 + 1)}{(u_1+u_2)(4u_1^3 + 12u_1^2u_2 + 14u_1^2 + 12u_1u_2^2 + 28u_1u_2 + 12u_1 + 4u_2^3 + 14u_2^2 + 12u_2 + 1)},$$

$$M_{11}^{\mathtt{S}} = \frac{u_2(4u_1^2u_2 + 8u_1u_2^2 + 14u_1u_2 + u_1 + 4u_2^3 + 14u_2^2 + 12u_2 + 1)}{(u_1+u_2)(4u_1^3 + 12u_1^2u_2 + 14u_1^2 + 12u_1u_2^2 + 28u_1u_2 + 12u_1 + 4u_2^3 + 14u_2^2 + 12u_2 + 1)},$$

$$M_{02}^{\mathtt{S}} = \frac{u_2(4u_1^2u_2 + 2u_1^2 + 8u_1u_2^2 + 16u_1u_2 + 4u_1 + 4u_2^3 + 14u_2^2 + 12u_2 + 1)}{(u_1+u_2)(4u_1^3 + 12u_1^2u_2 + 14u_1^2 + 12u_1u_2^2 + 28u_1u_2 + 12u_1 + 4u_2^3 + 14u_2^2 + 12u_2 + 1)}.$$

[17] That is, in the strong seed bank case; see Subchapter 1.2.2

In particular, M_1 does not depend on c and K. Even more strikingly, if $u_1 = u_1'$ and $u_2 = u_2'$, M_1 does not even depend on α and α' and is thus equal in cases $\mathtt{S}$ and $\mathtt{TI}$. However, this is not true for M_2, as it is easy to see by comparing the above results with those from [KZH08] (Appendix C), which state that in the case with $\alpha = \alpha' = 1$ and $K = c = 1,\ u_1 = u_1',\ u_2 = u_2'$, we get

$$\begin{aligned} M_{20}^{\mathtt{TI}} = M_{02}^{\mathtt{TI}} &= \frac{u_2(u_1 + 5u_2 + 2u_1u_2 + 1 + 2u_2^2)}{2u_1^3 + 6u_1^2u_2 + 5u_1^2 + 6u_1u_2^2 + 10u_1u_2 + u_1 + 2u_2^3 + 5u_2^2 + u_2}, \\ M_{11}^{\mathtt{TI}} &= \frac{u_2(5u_2 + 2u_1u_2 + 1 + 2u_2^2)}{(u_1 + u_2)(2u_1^2 + 4u_2u_2 + 5u_1 + 2u_2^2 + 5u_2 + 1)}. \end{aligned}$$

Remark 6.2. In the 2-allele case, we can apply the sampling formulas from [KZH08] to the Wright-Fisher diffusion with two subpopulations (Formula 2).
Suppose the process $(X(t), Y(t))_{t \geq 0}$ is in its stationary distribution. Take a random sample of size $n^{(1)}$ from the active population and an independent random sample of size $n^{(2)}$ from the dormant population. Denote the number of genes of the allelic type of interest in the active and in the dormant individuals respectively by η and ν. Then, the sampling formulas from [KZH08] (3.1) state that:

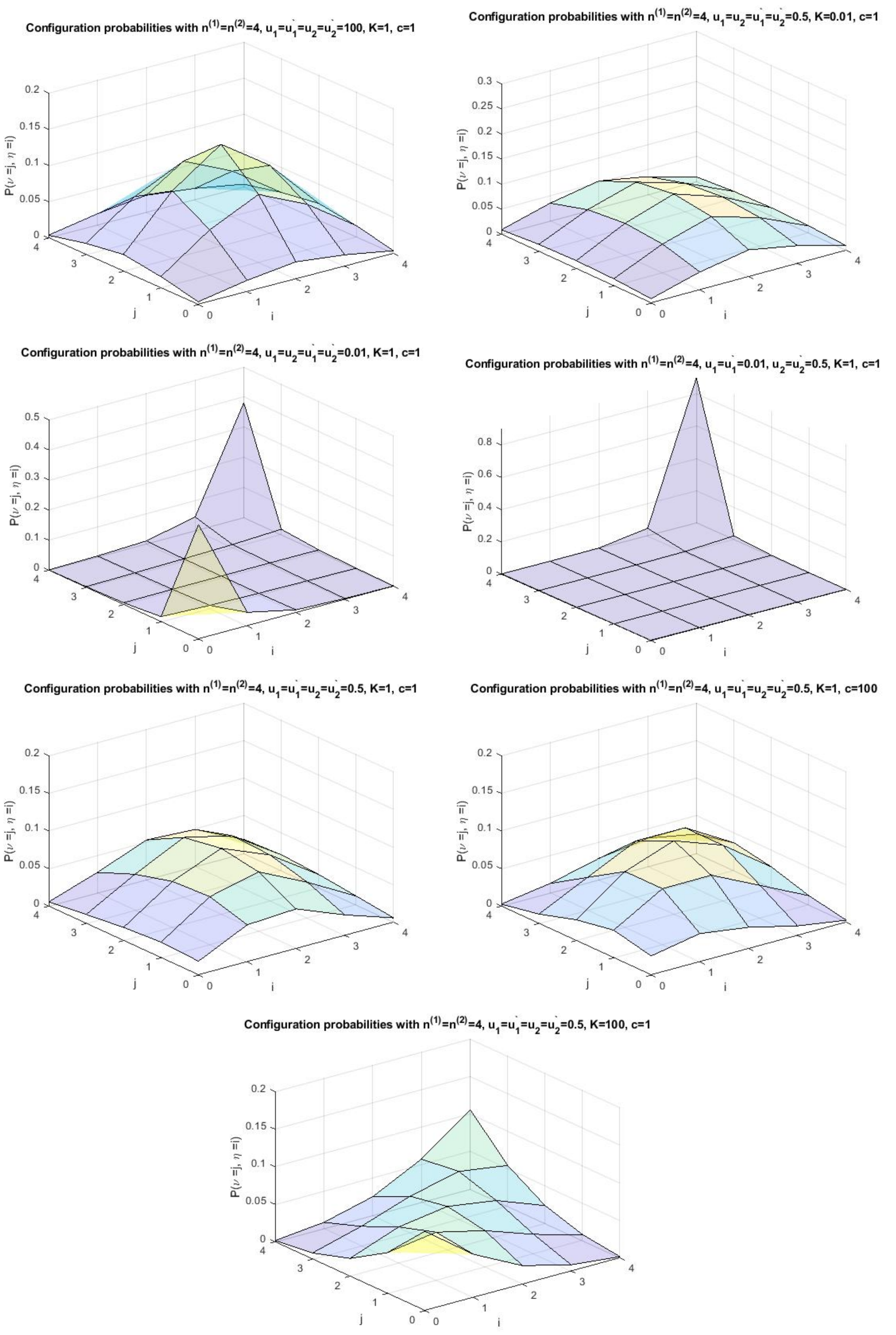
$$\begin{aligned} \mathbb{E}^{\mu}[\nu] &= n^{(2)} M_{01}, \\ \mathbb{E}^{\mu}[\eta] &= n^{(1)} M_{10}, \\ \mathbb{E}^{\mu}[\nu^2] &= n^{(2)} M_{01} + n^{(2)}(n^{(2)} - 1) M_{02}, \\ \mathbb{E}^{\mu}[\eta^2] &= n^{(1)} M_{10} + n^{(1)}(n^{(1)} - 1) M_{20}, \\ \mathbb{E}^{\mu}[\nu\eta] &= n^{(1)} n^{(2)} M_{11}, \end{aligned}$$

and we know all those quantities from Theorem 6.1.

Therefore, we can calculate the means and variances of ν and η, plus their covariance. Moreover, we can calculate the probability of a certain configuration (see also pictures):

$$\begin{aligned} \mathbb{P}^{\mu}\{\nu = j, \eta = i\} &= \binom{n^{(2)}}{i}\binom{n^{(1)}}{j} \mathbb{E}^{\mu}\left[(X(t))^i (1 - X(t))^{n^{(2)} - i} (Y(t))^j (1 - Y(t))^{n^{(1)} - j}\right] \\ &= (-1)^{n^{(1)} + n^{(2)} - i - j} \binom{n^{(2)}}{i}\binom{n^{(1)}}{j} \times \\ &\quad \times \sum_{p=0}^{n^{(1)} - i} \sum_{q=0}^{n^{(2)} - j} \binom{n^{(1)} - i}{p}\binom{n^{(2)} - j}{q} (-1)^{p+q} M_{n^{(1)} - p, n^{(2)} - q}. \end{aligned}$$

This gives us the possibility to try a maximum likelihood approach for parameter inference.



Configuration probabilities with $n^{(1)}=n^{(2)}=4$, $u_1=u_1'=u_2=u_2'=100$, K=1, c=1

Configuration probabilities with $n^{(1)}=n^{(2)}=4$, $u_1=u_2=u_1'=u_2'=0.5$, K=0.01, c=1

Configuration probabilities with $n^{(1)}=n^{(2)}=4$, $u_1=u_2=u_1'=u_2'=0.01$, K=1, c=1

Configuration probabilities with $n^{(1)}=n^{(2)}=4$, $u_1=u_1'=0.01$, $u_2=u_2'=0.5$, K=1, c=1

Configuration probabilities with $n^{(1)}=n^{(2)}=4$, $u_1=u_1'=u_2=u_2'=0.5$, K=1, c=1

Configuration probabilities with $n^{(1)}=n^{(2)}=4$, $u_1=u_1'=u_2=u_2'=0.5$, K=1, c=100

Configuration probabilities with $n^{(1)}=n^{(2)}=4$, $u_1=u_1'=u_2=u_2'=0.5$, K=100, c=1

Figure 8: Configuration probabilities in the seed bank model, for a sample of $n^{(1)} = 4$ active and $n^{(2)} = 4$ dormant and various choices of parameters.

6.2 Sample Heterozygosity in stationarity

The sample heterozygosity H of a population is the probability of finding two different alleles in a sample of size two from the population. In the presence of population substructure, e.g. in the form of a seed bank or separate islands, one often considers both the global resp. local sample heterozygosity, corresponding to samples taken from the total population, resp. from a sub-population. While the sample heterozygosity at stationarity (in the two-allele model with mutation) is well-studied for the Wright-Fisher and the two-island model (see [Her94]), it is less well understood for seed banks.

For $\mathtt{K}$ and $\mathtt{W}$, the sample heterozygosity is defined as

$$\begin{aligned} H^{\mathtt{K}} &:= 2\mathbb{E}^{\mu^{\mathtt{K}}}[X(1-X)], \\ H^{\mathtt{W}} &:= 2\mathbb{E}^{\mu^{\mathtt{W}}}[X(1-X)], \end{aligned}$$

respectively, where $\mathbb{E}^{\mu}$ denotes expectation with respect to the stationary measure μ, and X is the $X(t)$-coordinate from the relative process at stationarity. Analogous definitions hold for local sample heterozygosities under $\mathtt{S}$ and $\mathtt{TI}$:

$$\begin{aligned} H_X^{\mathtt{S}} &:= 2\mathbb{E}^{\mu^{\mathtt{S}}}[X(1-X)], \\ H_Y^{\mathtt{S}} &:= 2\mathbb{E}^{\mu^{\mathtt{S}}}[Y(1-Y)], \\ H_X^{\mathtt{TI}} &:= 2\mathbb{E}^{\mu^{\mathtt{TI}}}[X(1-X)], \\ H_Y^{\mathtt{TI}} &:= 2\mathbb{E}^{\mu^{\mathtt{TI}}}[Y(1-Y)], \end{aligned}$$

and global heterozygosities are defined as weighted averages. The weighting can be done in several ways:
1) We choose the first individual at random, uniformly from the entire population. The second one is chosen uniformly from the same subpopulation the first sampled individual comes from.
2) We flip a fair coin. Depending on the result, we choose both individuals at random from the X- or from the Y-subpopulation.
3) We sample two individuals at random, uniformly from the entire population. If the sample is homogeneous, i.e. if the two individuals come from the same subpopulation, we retain the sample. Else, we discard it and do the sampling again until we get an homogeneous sample.
4) We simply pick two individuals randomly and uniformly from the entire population, regardless of their subpopulations.
Notice that ways 1 to 3 coincide if $K = 1$, that is, if the two subpopulations have the same size, and that, in ways 1 to 3, the sample cannot contain simultaneously elements from the first and from the second subpopulation, while in way 4 it very well can.
In this thesis, we will always choose the first way (when we want a subpopulation-homogeneous sample) or the fourth (when we want to allow for the sample to come from different subpopulations). The corresponding formulas for the global sample

heterozygosity are then, for way 1,

$$\begin{aligned}\frac{KH_X^{\mathtt{S}}+H_Y^{\mathtt{S}}}{K+1}&=\frac{2K}{K+1}\mathbb{E}^{\mu^{\mathtt{S}}}[X(1-X)]+\frac{2}{K+1}\mathbb{E}^{\mu^{\mathtt{S}}}[Y(1-Y)],\\ \frac{KH_X^{\mathtt{TI}}+H_Y^{\mathtt{TI}}}{K+1}&=\frac{2K}{K+1}\mathbb{E}^{\mu^{\mathtt{TI}}}[X(1-X)]+\frac{2}{K+1}\mathbb{E}^{\mu^{\mathtt{TI}}}[Y(1-Y)],\end{aligned}$$

and, for way 4,

$$\begin{aligned}H^{\mathtt{S}}&:=\frac{2K^2}{(K+1)^2}\mathbb{E}^{\mu^{\mathtt{S}}}[X(1-X)]+\frac{2K}{(K+1)^2}\mathbb{E}^{\mu^{\mathtt{S}}}[X(1-Y)+Y(1-X)]\\ &\quad+\frac{2}{(K+1)^2}\mathbb{E}^{\mu^{\mathtt{S}}}[Y(1-Y)],\\ H^{\mathtt{TI}}&:=\frac{2K^2}{(K+1)^2}\mathbb{E}^{\mu^{\mathtt{TI}}}[X(1-X)]+\frac{2K}{(K+1)^2}\mathbb{E}^{\mu^{\mathtt{TI}}}[X(1-Y)+Y(1-X)]\\ &\quad+\frac{2}{(K+1)^2}\mathbb{E}^{\mu^{\mathtt{TI}}}[Y(1-Y)].\end{aligned}$$

We can compute these quantities by using the results from Theorem 6.1; of course, here we are only interested in moments up to order two. Nevertheless, the resulting expressions are complicated if we want to keep all parameters. For example, for S we obtain

$$\begin{aligned}M_{1,0}^{\mathtt{S}}&=\frac{cu_2'+u_1u_2'+u_2u_2'+c'u_2}{cu_1'+cu_2'+u_1u_1'+u_1u_2'+u_2u_1'+u_2u_2'+c'u_1+c'u_2},\\ M_{0,1}^{\mathtt{S}}&=\frac{cu_2'+u_1'u_2+u_2u_2'+c'u_2}{cu_1'+cu_2'+u_1u_1'+u_1u_2'+u_2u_1'+u_2u_2'+c'u_1+c'u_2}.\end{aligned}$$

Expressing the sample heterozygosities in terms of these moments is now immediate:

Proposition 6.3. *With the above notation, we have the representations*

$$\begin{aligned}H_X^{\mathtt{S}}&=2(M_{1,0}^{\mathtt{S}}-M_{2,0}^{\mathtt{S}}),\\ H_Y^{\mathtt{S}}&=2(M_{0,1}^{\mathtt{S}}-M_{0,2}^{\mathtt{S}}),\\ H^{\mathtt{S}}&=\frac{2}{(K+1)^2}\Big((K^2+K)M_{1,0}^{\mathtt{S}}+(K+1)M_{0,1}^{\mathtt{S}}-2KM_{1,1}^{\mathtt{S}}-K^2M_{2,0}^{\mathtt{S}}-M_{0,2}^{\mathtt{S}}\Big),\\ H_X^{\mathtt{TI}}&=2(M_{1,0}^{\mathtt{TI}}-M_{2,0}^{\mathtt{TI}}),\\ H_Y^{\mathtt{TI}}&=2(M_{0,1}^{\mathtt{TI}}-M_{0,2}^{\mathtt{TI}}),\\ H^{\mathtt{TI}}&=\frac{2}{(K+1)^2}\Big((K^2+K)M_{1,0}^{\mathtt{TI}}+(K+1)M_{0,1}^{\mathtt{TI}}-2KM_{1,1}^{\mathtt{TI}}-K^2M_{2,0}^{\mathtt{TI}}-M_{0,2}^{\mathtt{TI}}\Big).\end{aligned}$$

Remark 6.4. The result for the Kingman case is well known, see e.g. [Eth11]:

$$H^{\mathtt{K}}=\frac{4u_1u_2}{(u_1+u_2)(1+2u_1+2u_2)}.$$

Moreover, the weak seed bank diffusion with parameter $\beta>0$ and the one-dimensional diffusion $(X(t))_{t\geq 0}$ given as the unique strong solution of the SDE

$$\mathrm{d}X(t)=\big[-u_1X(t)+u_2(1-X(t))\big]\mathrm{d}t+\beta\sqrt{X(t)(1-X(t))}\mathrm{d}W(t)$$

have the same stationary distribution. Thus, the SH at stationarity can be easily computed in a similar way and

$$H^{\mathtt{W}} = \frac{4u_1u_2}{(u_1+u_2)(\beta^2+2u_1+2u_2)}.$$

Remark 6.5 (Scaling limit as $K \to \infty$)**.** Note that for $K \to \infty$ (that is, the relative seed bank resp. second island size approaches 0), we recover the classical heterozygosities:

$$\begin{aligned} H_X^{\mathtt{S}} &\to H^{\mathtt{K}}, \\ H_X^{\mathtt{TI}} &\to H^{\mathtt{K}}. \end{aligned}$$

This is consistent with the result in Corollary 5.12.

Example. Consider $K = c = 1, u_1 = u_2 = u_1' = u_2' = 1/2, \alpha = 1$, and $\alpha' = 0$. Then we obtain

$$H^{\mathtt{S}} = \frac{14}{31} \approx 0.4516 \qquad > \qquad H^{\mathtt{K}} = \frac{1}{3}.$$

This indicates that a strong seed bank introduces some amount of population structure, even at stationarity. In the two-island model with $\alpha = 1$ we get, using the same parameter values as before,

$$H^{\mathtt{TI}} = \frac{13}{32} \approx 0.4063,$$

slightly lower, which is consistent with the fact that genetic drift reduces variability.

6.3 Sample Heterozygosity decay

What differentiates K from W is the rate of *decay* of sample heterozygosity in the absence of mutation. Likewise, we can investigate the effect of the presence of a seed bank on the *random genetic drift* in the seed bank diffusion with the help of the sample heterozygosity. In order to do so, let us now compute the decay of the sample heterozygosity in the active population (which can be observed by biologists, unlike the global sample heterozygosity) in the non-mutation case.

In this setting, sample heterozygosity is defined as

$$H^I(t,\bar{x}) := 2\mathbb{E}_{\bar{x}}[X(t)(1-X(t))],$$

for $I \in \{\mathtt{K}, \mathtt{W}, \mathtt{TI}, \mathtt{S}\}$ and a process started in $\bar{x} \in [0,1]$ (cases K and W) or in $\bar{x} \in [0,1]^2$ (cases TI and S). Then, in K (given $u_1 = u_2 = 0$), we obtain (see e.g. ([Dur08]), p.8)

$$H^{\mathtt{K}}(t,x) = 2e^{-t}x(1-x),$$

while in the case W, the mergers happen at rate $\frac{1}{\beta^2}$. This makes the weak seed bank coalescent a time-changed Kingman coalescent, and the sample heterozygosity at time $t > 0$ will thus be

$$H^{\mathtt{W}}(t,x) = 2e^{-\beta^2 t}x(1-x).$$

Thus, the action of genetic drift is delayed by a weak seed bank. For the cases `TI` and `S`, suppose we have initial conditions $(X(0), Y(0)) = (x_0, y_0) \in [0,1]^2$ in

$$\begin{aligned} \mathrm{d}X(t) &= \big[c(Y(t) - X(t))\big]\mathrm{d}t + \alpha\sqrt{X(t)(1-X(t))}\mathrm{d}W(t), \\ \mathrm{d}Y(t) &= \big[c'(X(t) - Y(t))\big]\mathrm{d}t + \alpha'\sqrt{Y(t)(1-Y(t))}\mathrm{d}W'(t). \end{aligned} \tag{43}$$

Then, the sample heterozygosity at time 0 in the active population is

$$H_A(0, (x_0, y_0)) := 2X(0)(1 - X(0)) = 2x_0(1 - x_0). \tag{44}$$

As random genetic drift reduces genetic variability over time, we expect this to converge to 0 just like in the cases `K`, `W`.

Proposition 6.6. *In the absence of mutation, the sample heterozygosity in the active population at time $t > 0$, $H_A(t, (x_0, y_0))$, given that the frequency of an allele at time 0 is equal to $x_0 \in (0,1)$ in the first subpopulation and to $y_0 \in (0,1)$ in the second one, is equal to:*
a) in the seed bank model, defining

$$S = \begin{bmatrix} -2Kc & 2Kc & 0 \\ c & c + Kc & Kc \\ 0 & 2c & -2c-1 \end{bmatrix}$$

and $T(u) := \exp(Su)$ (matrix exponential),

$$H_A^{\mathtt{S}}(t, (x_0, y_0)) = x_0(1 - x_0)T(t)_{3,3} + (x_0 + y_0 - 2x_0y_0)T(t)_{3,2} + y_0(1 - y_0)T(t)_{3,1};$$

b) in the two-island model with $\alpha = 1,\ \alpha' = \sqrt{K}$, same as in (a) but with

$$S = \begin{bmatrix} -2Kc - K & 2Kc & 0 \\ c & c + Kc & Kc \\ 0 & 2c & -2c-1 \end{bmatrix}.$$

Proof. a) It is clear that the sample heterozygosity at time 0 is given by $H_A^{\mathtt{S}}(0, (x_0, y_0)) = x_0(1 - x_0)$. In order to investigate the time evolution of $H_A^{\mathtt{S}}$ for the seed bank diffusion, we apply backwards-in-time arguments related to the seed bank coalescent. Indeed, if we start with two sampled individuals at time $t > 0$, they will have different alleles if and only if their ancestors at time 0 are different (i.e. their lineages have not coalesced) and carry different alleles. In short,

$$\begin{aligned} H_A^{\mathtt{S}}(t, (x_0, y_0)) &:= \mathbb{E}_{(x_0,y_0)}[X(t)(1 - X(t))] \\ &= \mathbb{P}\{A_t\}\bar{H}(0, (x_0, y_0)) \end{aligned}$$

where A_t is the event that two different individuals at time t, both sampled from the active population, have two different ancestors at time 0. $\bar{H}$, meanwhile, is the conditional probability that the two ancestors have different alleles given A_t. However, this probability is tricky to compute since it depends on whether the ancestors at time 0 are active or not. Therefore, we have to decompose A_t as follows:

$$A_t = PP(t)\dot{\cup}PS(t)\dot{\cup}SS(t),$$

where $PP(t)$, $PS(t)$ and $SS(t)$ stand for the events that the ancestors (at time 0) of our sample (picked at time t) are two active individuals, one active and one dormant, and two dormant individuals, respectively (which of course are pairwise disjoint, and all imply that no coalescence has taken place).
And moreover, on each of the three subsets of A_t, the distribution of $\bar{H}$ is easily computable. Denote with $\bar{H}_1$, $\bar{H}_2$ and $\bar{H}_3$ the conditional probabilities that the two ancestors have different alleles given $PP(t)$, $PS(t)$ and $SS(t)$, respectively. Then,

$$\bar{H}_1 = x_0(1-x_0);\ \bar{H}_2 = x_0 + y_0 - 2x_0y_0;\ \bar{H}_3 = y_0(1-y_0).$$

Applying the law of total probability, we get

$$\begin{aligned}\mathbb{P}\{A_t\}\bar{H}(0,(x_0,y_0)) = \mathbb{P}\{PP(t)\}x_0(1-x_0) + \mathbb{P}\{PS(t)\}(x_0 + y_0 - 2x_0y_0)\\ + \mathbb{P}\{SS(t)\}y_0(1-y_0).\end{aligned}$$

Finally, the probabilities of the three subsets of A_t are computable via matrix exponential:

$$\mathbb{P}\{PP(t)\} = \exp(Tt)_{(3,3)};\ \mathbb{P}\{PS(t)\} = \exp(Tt)_{(3,2)};\ \mathbb{P}\{SS(t)\} = \exp(Tt)_{(3,1)},$$

which yields the thesis.
b) All previous considerations can be used for the two-island model with $\alpha = 1$, $\alpha' = \sqrt{K}$, by changing the transition matrix accordingly (see previous considerations). □

However, the matrix exponentials are almost impossible to calculate while leaving t as a parameter. Thus, we only present a plot of the sample heterozygosity in the active population in function of t. (Figure 9) Heuristically, we see that the sample heterozygosity in absence of mutation decays exponentially with parameter approximately 0.2 (seed bank) or 0.45 (two-island).

6.4 Wright's F_{ST}

One of the most prominent statistics for the analysis of population structure is Wright's F_{ST} (see [Wri49]). It is defined as

$$F_{ST} := \frac{p_0 - \bar{p}}{1 - \bar{p}}, \tag{45}$$

where $\bar{p}$ is the probability of identity of two randomly sampled genes from the whole population [Her94, p.73], and p_0 is the probability of identity of two randomly sampled genes from a single subpopulation, itself randomly sampled with probabilities proportional to subpopulation sizes (that is, using method (1) from subchapter 6.2). The specifics of p_0 and $\bar{p}$ depend on the mutation model.

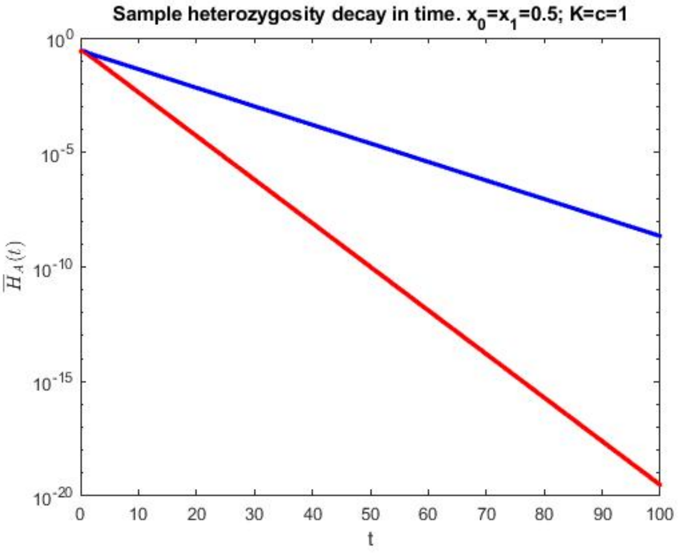


Figure 9: Sample heterozygosity at time t when $x_0 = y_0 = 1/2$, $K = c = 1$, for the strong seed bank model (blue) and the two-island model (red)

Wright's F_{ST} for the two-alleles model. A standard formulation of the F_{ST} for the two-island model is obtained by expressing p_0 and $\bar{p}$ as functions of sample heterozygosities:

$$F_{ST}^{\texttt{TI}} := \frac{(K+1)H^{\texttt{TI}} - KH_X^{\texttt{TI}} - H_Y^{\texttt{TI}}}{(K+1)H^{\texttt{TI}}}.$$

and analogously,

$$F_{ST}^{\texttt{S}} := \frac{(K+1)H^{\texttt{S}} - KH_X^{\texttt{S}} - H_Y^{\texttt{S}}}{(K+1)H^{\texttt{S}}}$$

for the strong seed bank model. For example, for $u_1 = u_2 = 0.5 = u_1' = u_2'$, $c = K = \alpha = 1$, the island model ($\alpha' = 1$) leads to a stronger differentiation than the corresponding seed bank model ($\alpha' = 0$):

$$F_{ST}^{\texttt{S}} = \frac{1}{28} < F_{ST}^{\texttt{TI}} = \frac{1}{13}.$$

This indicates that strong seed banks introduce some population substructure, but that the effect is stronger for the two island model. Intuitively, both subpopulations undergoing genetic drift leads to behavior that is closer to two independent populations than when genetic drift only takes place on one subpopulation.

It is also of interest to see how the F_{ST} depends on the model parameters. In Figure 10, in the first plot we see the F_{ST} as a function of the migration rate c. It approaches 0 as $c \to \infty$ as expected, since this leads to a well-mixed population. A similar result holds if the mutation rates are increased (second plot). Again, this is consistent with the heuristic observation that a strong mutation rate which is equal in both subpopulations further mixes the population. Moreover, from the first two figures we can infer that for equally-sized subpopulations, the presence of islands results in an F_{ST} value which is approximately twice as high as under the seed bank regime.

The third plot shows the dependence of F_{ST} on K. Under both $\mathtt{S}$ and $\mathtt{TI}$, F_{ST} is nearly 0 if the relative population size on either island is very small, resulting in a correspondingly small probability of sampling individuals from the smaller subpopulation both during homogeneous and during generic sampling.

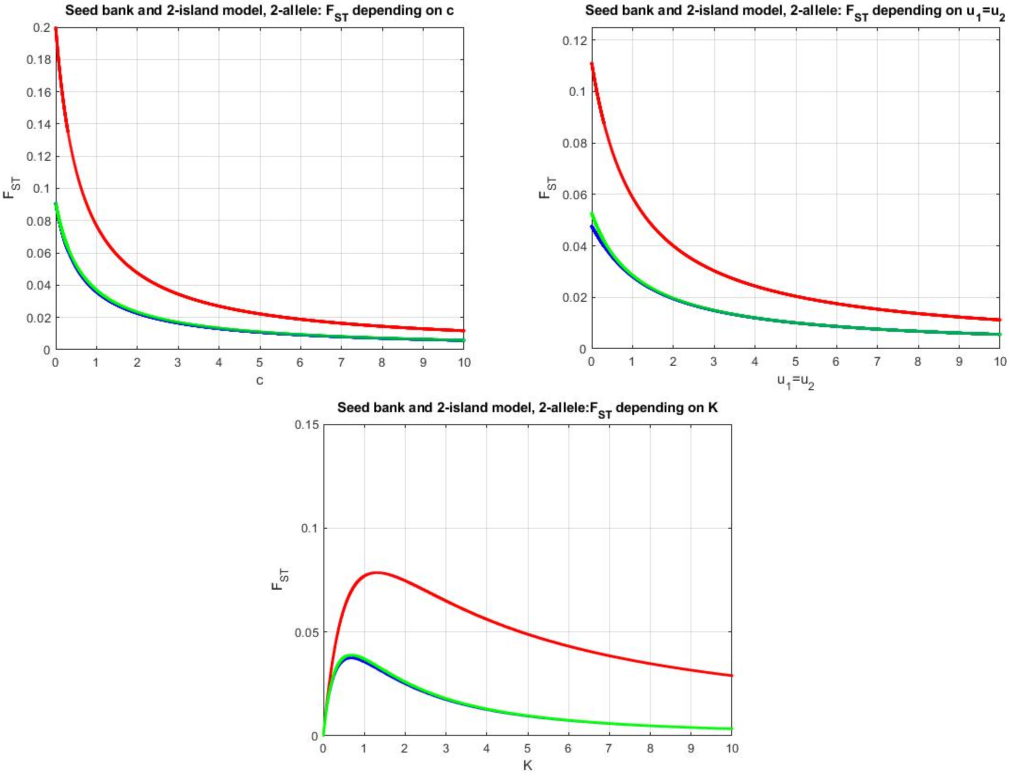


Figure 10: F_{ST} under the two island (red) and strong seed bank models (green for $u_1' = u_2' = 0$, blue for $u_1 = u_2 = u_1' = u_2'$) in function of various parameters. Where not specified, $K = c = 1$, $u_1 = u_2 = u_1' = u_2' = 0.5$.

The plots also show that the case when there is no mutation in the seed bank, i.e. $u_1' = u_2' = 0$, is almost indistinguishable from the case with mutation. For example, we have, provided $u_1 = u_2 = 1/2, K = c = 1$, $F_{ST}^{\mathtt{S}} = \frac{1}{27}$ — a slightly stronger signal than in the case with mutation.

Wright's F_{ST} for the infinite alleles model. Every mutation leads to a new allele under the infinite alleles model. Thus p_0 and $\bar{p}$ from (45) coincide with the probabilities of *identity by descent* (IBD), by which we mean the probability that the two ancestral lineages of the two genes have not been mutated since the time of their most recent common ancestor, which has been expressed in a simple form ([Hud90], Section 4). Let T_0 be the coalescence time of two ancestral lines from the same subpopulation, randomly chosen according to its relative size. Given T_0, the probability of no mutations on each given line is e^{-uT_0}. Since mutations arise conditionally independently (given T_0) we have

$$p_0 = \mathbb{E}[e^{-2uT_0}].$$

A similar quantity has recently been investigated for S in [dHP17] in the case of a finite population on a discrete torus. Similarly, we have

$$\bar{p} = \mathbb{E}[e^{-2u\bar{T}}],$$

where $\bar{T}$ is the coalescence time of two individuals sampled uniformly from the whole population (see e.g. [Her94, p73f]). Hence,

$$F_{ST} = \frac{p_0 - \bar{p}}{1 - \bar{p}} = \frac{\mathbb{E}[e^{-2uT_0}] - \mathbb{E}[e^{-2u\bar{T}}]}{\mathbb{E}[1 - e^{-2u\bar{T}}]}.$$

These quantities can be calculated using the following two results:

Proposition 6.7. *Consider a sample of size $n = 2$ from the seed bank coalescent. Denote by T_1, T_2, and T_3 the coalescence times of two lineages if both lines are initially active (T_1), one is active and the other dormant (T_2), or both are dormant (T_3). Then, for any $u > 0$,*

$$\begin{aligned}
\mathbb{E}\big[e^{-uT_1}\big] &= \frac{1}{1+2c+u} \\
&\quad \Big(1 + \frac{2c^2K(2Kc+u)}{(c+Kc+u)(1+2c+u)(2Kc+u) - 2c^2K(1+2u+2c(1+K))}\Big), \\
\mathbb{E}\big[e^{-uT_2}\big] &= \frac{cK(2Kc+u)}{(c+Kc+u)(1+2c+u)(2Kc+u) - 2c^2K(1+2u+2c(1+K))}, \\
\mathbb{E}\big[e^{-uT_3}\big] &= \frac{2c^2K^2}{(c+Kc+u)(1+2c+u)(2Kc+u) - 2c^2K(1+2u+2c(1+K))}.
\end{aligned}$$

Proof. The proof is based on phase-type distribution theory; we will mainly use formula (7) (see Chapter 2).
This formula fits the bill both for the seed bank and the two-island models. We fix $p = 3$, and associate the states $\{x_1, \ldots, x_4\}$ with "two lineages in the Y-subpopulation", "one lineage in X and one in Y", "two lineages in the X-subpopulation", and "coalescence", respectively. For S we take

$$\mathbf{S} = \begin{bmatrix} -2Kc & 2Kc & 0 \\ c & -c-Kc & Kc \\ 0 & 2c & -2c-1 \end{bmatrix}, \quad \mathbf{s} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

and $\pi = [0,0,1]$ if we start by sampling two active lineages, $\pi = [0,1,0]$ if we sample one active and one dormant lineage, and $\pi = [1,0,0]$ for two dormant lineages.

Since we are looking for the entire time until coalescence, the reward function r

will be constant and equal to 1. Thus, we get

$$\begin{aligned}
\mathbb{E}[e^{-uT_1}] &= \begin{bmatrix}0\\0\\1\end{bmatrix}\left(\begin{bmatrix}u&0&0\\0&u&0\\0&0&u\end{bmatrix}-S\right)^{-1}\begin{bmatrix}0\\0\\-1\end{bmatrix},\\
\mathbb{E}[e^{-uT_2}] &= \begin{bmatrix}0\\1\\0\end{bmatrix}\left(\begin{bmatrix}u&0&0\\0&u&0\\0&0&u\end{bmatrix}-S\right)^{-1}\begin{bmatrix}0\\0\\-1\end{bmatrix},\\
\mathbb{E}[e^{-uT_3}] &= \begin{bmatrix}1\\0\\0\end{bmatrix}\left(\begin{bmatrix}u&0&0\\0&u&0\\0&0&u\end{bmatrix}-S\right)^{-1}\begin{bmatrix}0\\0\\-1\end{bmatrix}.
\end{aligned}$$

Multiplying the matrices gives the claimed expressions.

□

All previous considerations can also be applied to the two-island model TI (see again [Her94], Section 4.3.1). The interpretation of the initial distribution in the two-island model with $\alpha = 1, \alpha' = \sqrt{K}$ is similar, and the full picture differs only slightly; the only difference with respect to the seed bank case is that pairs of individuals may coalesce in both subpopulations. Thus, we set

$$\mathbf{S} = \begin{bmatrix}-2Kc-K & 2Kc & 0\\ c & -c-Kc & Kc\\ 0 & 2c & -2c-1\end{bmatrix}, \quad \mathbf{s} = \begin{bmatrix}K\\0\\1\end{bmatrix}.$$

Now we can prove the following result, giving us the relevant sample heterozygosities in the infinitely many alleles case:

Proposition 6.8 (Sample heterozygosities in IAM)**.** *For the two-island model* TI *with* $\alpha = 1,\ \alpha' = \sqrt{K}$ *and the seed bank model* S*, we have*

$$\begin{aligned}
H_X^{\mathtt{S}} &= 1 - \mathbb{E}[e^{-2uT_1}],\\
H_Y^{\mathtt{S}} &= 1 - \mathbb{E}[e^{-2uT_3}],\\
H^{\mathtt{S}} &= 1 - \frac{K^2}{(1+K)^2}\mathbb{E}[e^{-2uT_1}] - \frac{2K}{(1+K)^2}\mathbb{E}[e^{-2uT_2}] - \frac{1}{(1+K)^2}\mathbb{E}[e^{-2uT_3}],\\
{p_0}^{\mathtt{S}} &= \frac{K}{1+K}\mathbb{E}[e^{-2uT_1}] + \frac{1}{1+K}\mathbb{E}[e^{-2uT_3}],\\
H_X^{\mathtt{TI}} &= 1 - \mathbb{E}[e^{-2uT_1}],\\
H_Y^{\mathtt{TI}} &= 1 - \mathbb{E}[e^{-2uT_3}],\\
H^{\mathtt{TI}} &= 1 - \frac{K^2}{(1+K)^2}\mathbb{E}[e^{-2uT_1}] - \frac{2K}{(1+K)^2}\mathbb{E}[e^{-2uT_2}] - \frac{1}{(1+K)^2}\mathbb{E}[e^{-2uT_3}],\\
{p_0}^{\mathtt{TI}} &= \frac{K}{1+K}\mathbb{E}[e^{-2uT_1}] + \frac{1}{1+K}\mathbb{E}[e^{-2uT_3}],
\end{aligned}$$

where $T_1,\ T_2,\ T_3$ *are the coalescence times of a sample of two introduced in Proposition 6.7. Moreover,* $1 - H^{\mathtt{S}} = \bar{p}^{\mathtt{S}}$ *and* $1 - H^{\mathtt{TI}} = \bar{p}^{\mathtt{TI}}$.

Proof. Both in the seed bank model and in the two-island model, using the previous notation, we can use the formula of total probability to get that

$$\begin{aligned}\mathbb{E}[e^{-uT_0}] &= \frac{K}{1+K}\mathbb{E}[e^{-uT_1}] + \frac{1}{1+K}\mathbb{E}[e^{-uT_3}];\\ \mathbb{E}[e^{-u\bar{T}}] &= \frac{K^2}{(1+K)^2}\mathbb{E}[e^{-uT_1}] + \frac{2K}{(1+K)^2}\mathbb{E}[e^{-uT_2}] + \frac{1}{(1+K)^2}\mathbb{E}[e^{-uT_3}].\end{aligned}$$

The rest of the thesis follows immediately from the definitions. □

Plugging the formulas from Propositions 6.7 and 6.8 into the definition of F_{ST} results in lengthy, but closed-form expressions. In Figure 11, we present simulations comparing models and exploring how F_{ST}s depend on parameters. Notice that all three plots are remarkably similar to those in the 2-allele case.
The same approach can be adapted to the case when mutation rates differ between the two populations — we just need a different reward function.

Given a mutation rate $u \geq 0$ among the active individuals and $u' = \lambda u \geq 0$ among the dormant ones, we now define

$$\begin{aligned} r(x_1) &= \lambda,\\ r(x_2) &= \frac{1+\lambda}{2},\\ r(x_3) &= 1,\end{aligned}$$

reflecting the relative mutation rates. Repeating the computations above with this choice of reward function will again yield closed-form expressions for F_{ST}.

For example, for $u = 1/2$, $c = K = \alpha = 1$, the island model ($\alpha' = 1$) leads to a stronger differentiation than the corresponding seed bank model ($\alpha' = 0$):

$$F_{ST}^{\mathtt{S}} = \frac{5}{113} < F_{ST}^{\mathtt{TI}} = \frac{1}{10}.$$

In the case where there is no mutation among the dormant individuals ($\lambda = 0$), $F_{ST}^{\mathtt{S}} = \frac{1}{21}$. This indicates that also in this case, strong seed banks introduce some amount of population substructure, and the effect is stronger for the two island model. Intuitively, both subpopulations undergoing genetic drift leads to behavior that is closer to two independent populations than when genetic drift only takes place on one subpopulation.

Wright's F_{ST} for the infinite sites model. The central difference between the IAM and the ISM is that all previous mutations on a lineage remain observable in the latter. However, this does not affect the probability of identity by descent of two sampled individuals — they will still carry the same allele if and only if neither ancestral line mutated during the time from their most recent common ancestor to the present. Thus, sample heterozygosity H and F_{ST} under the ISM can be computed in exactly the same way as in the IAM and we refer to the previous section for the explicit formulas.

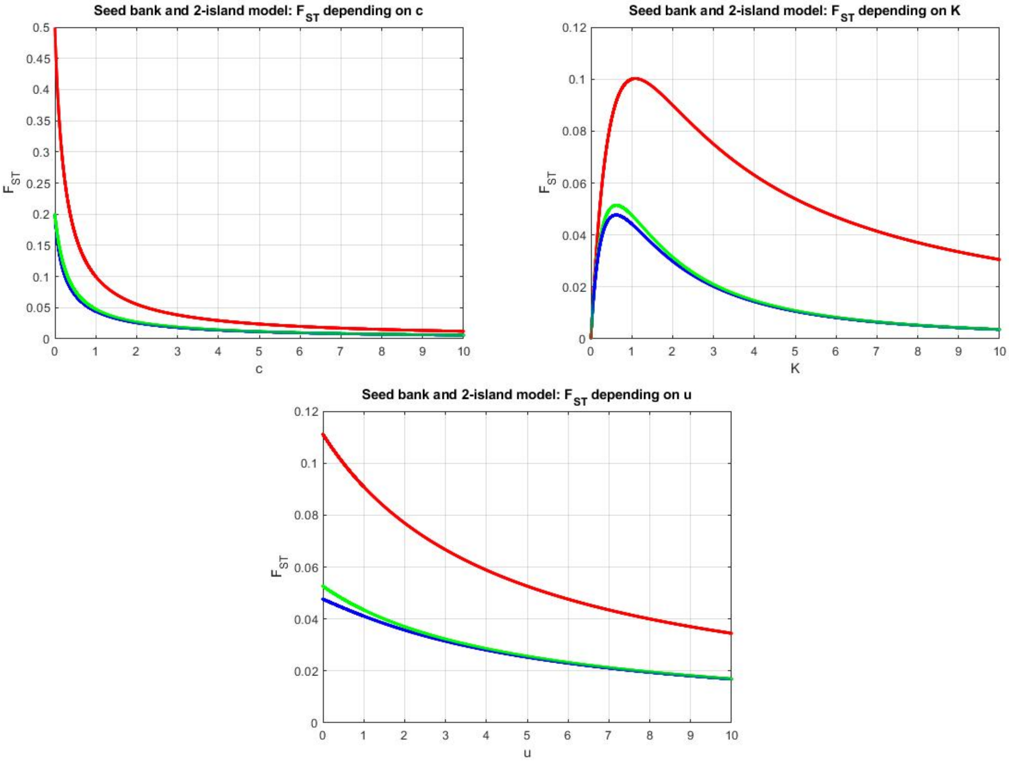


Figure 11: F_{ST} under the two island (red) and strong seed bank models (green for $\lambda = 0$, blue for $\lambda = 1$) in function of various parameters. Where not specified, $K = c = 1, u = 0.5$.

6.5 Slatkin's F^0_{ST}

In [Sla91], Slatkin suggests that F_{ST} can be approximated by its limit as the mutation rate tends to 0; this approximation has the advantage that it can be computed using de l'Hopital's rule:

$$F^0_{ST} := \lim_{u \to 0} F_{ST}(u) = \frac{\mathbb{E}[\bar{T} - T_0]}{\mathbb{E}[\bar{T}]}.$$

These expected values are easily computed and independent of the mutation rate. For example, in the seed bank case with $\lambda = 1$, we just need to apply Formula 6 from Chapter 2.4 with reward vector $(1, 1, 1)$ and initial conditions $(\frac{1}{1+K}, 0, \frac{K}{1+K})$ (used for computing T_0) and $(\frac{1}{(1+K)^2}, \frac{2K}{(1+K)^2}, \frac{K^2}{(1+K)^2})$ (for computing $\bar{T}$) which gives us

$$\mathbb{E}[T_0] = \frac{K}{1+K}\Big(\frac{1}{K^2} + \frac{2}{K} + 1\Big) + \frac{1}{1+K}\Big(\frac{1+2c+K}{2cK^2} + \frac{1+2c}{cK} + 1\Big),$$

and

$$\begin{aligned}\mathbb{E}[\bar{T}] = {} & \frac{K^2}{(1+K)^2}\Big(\frac{1}{K^2} + \frac{2}{K} + 1\Big) + \frac{1}{(1+K)^2}\Big(\frac{1+2c+K}{2cK^2} + \frac{1+2c}{cK} + 1\Big) \\ & + \frac{2K}{(1+K)^2}\Big(\frac{1+2c}{2cK^2} + \frac{1+2c}{cK} + 1\Big).\end{aligned}$$

We obtain the relatively simple explicit expression

$$F^0_{ST} = \frac{K}{2c + 4K + 6cK + 6cK^2 + 2cK^3 + 1},$$

compared with

$$\begin{aligned} F_{ST} = K(c + u + cK)(&2c^2K^4 + 8c^2K^3 + 12c^2K^2 + 8c^2K + 2c^2 \\ &+ 3cK^3u + 9cK^2u + 4cK^2 + 9cKu \\ &+ 5cK + 3cu + c + K^2u^2 + 2Ku^2 + 2Ku + u^2 + u)^{-1} \end{aligned}$$

Remark 6.9. • Note that the F^0_{ST}, like the F_{ST}, tends to zero if $K \to 0$, $K \to \infty$ or $c \to \infty$, but not in the case $c \to 0$.

• Figure 12 shows the relative error when using the F^0_{ST} instead of the F_{ST}.

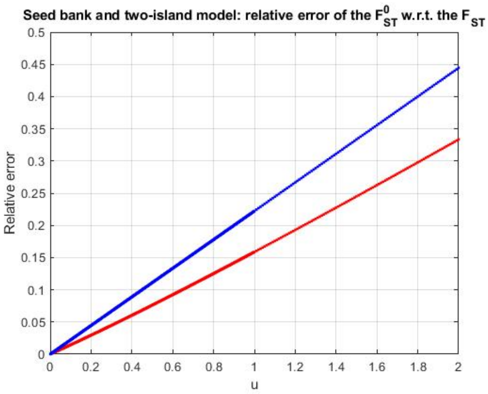


Figure 12: Error using the F^0_{ST} instead of the F_{ST}, under the two island (red) and strong seed bank model with $\lambda = 1$ (blue), given $K = c = 1$.

As we can see, the validity of the F^0_{ST} as an approximation for Slatkin's F_{ST} depends highly on the effective value of u. In the neutral case ($K = c = 1$) the error is 1.5% in the seed bank model and 2.2% in the 2-island model for a (low) value of $u = 0.1$; for a (neutral) value of $u = 0.5$, the errors are respectively 8 and 11%; for a high value of $u = 2$, the errors become 33 and 44% respectively. All in all, the F^0_{ST} seems like a good approximation if u does not get too large.

• The Slatkin approximation is particularly useful in the two-island model, too. In that case, as shown in [Her94](Section 4.2.7), denoting by S_0 the number of nucleotide differences between a pair of genes sampled randomly from the same subpopulation and S the same number, but sampling from the entire population, we get

$$F^{(0)}_{ST} = \frac{\mathbb{E}[S - S_0]}{\mathbb{E}[S]},$$

implying that Slatkin's approximation for F_{ST} can be estimated from DNA sequence data.

6.6 Laplace and Fourier transforms

The introduction of phase-type distribution theory turned out to be revolutionary for solving this kind of F_{ST} problems. The fact that this approach shortens the proofs, sometimes drastically, is highlighted by the messiness of alternative proofs based upon the Laplace and Fourier transforms, as we can see here. For simplicity, we state the proof only in the case $\mathtt{S}$ where $\lambda = 1$.

Proof. Let T_1, T_2, and T_3 be as before. Define W_1, W_2 and W_3 to be the waiting times until the first event in coalescent history of a sample of size two in the above situations (this can be a coalescence or a migration event). By the properties of the seed bank coalescent, the latter random variables are exponentially distributed with respective expected values

$$\mathbb{E}[W_1] = \frac{1}{1+2c}, \quad \mathbb{E}[W_2] = \frac{1}{c(1+K)}, \quad \mathbb{E}[W_3] = \frac{1}{2cK},$$

since in the first scenario, there are two possible transitions: a coalescence happening at rate 1, and a jump from active to dormant state a rate c for each of the two lineages. The other expected values follow from similar considerations. Consequently, the Laplace transforms of these random variables are given by

$$\mathbb{E}\big[e^{-uW_1}\big] = \frac{1+2c}{1+2c+u}; \ \mathbb{E}\big[e^{-uW_2}\big] = \frac{c(1+K)}{c(1+K)+u}; \ \mathbb{E}\big[e^{-uW_3}\big] = \frac{2Kc}{2Kc+u}.$$

We now use these plus memorilessness to compute the Laplace transform of T_1. To this end, let C be the event that the first transition of the sample of size two is a coalescence, and J be the event that the first event is a jump of an active to a dormant lineage. Note that both C and J are independent of W_1. We obtain, due to lack of memory,

$$\begin{aligned}\mathbb{E}\big[e^{-uT_1}\big] &= \mathbb{E}\big[e^{-uT_1}\mathbb{1}_C\big] + \mathbb{E}\big[e^{-uT_1}\mathbb{1}_J\big] = \mathbb{E}\big[e^{-uW_1}\big]\mathbb{P}\{C\} + \mathbb{E}\big[e^{-u(W_1+T_2)}\big]\mathbb{P}\{J\} \\ &= \mathbb{E}\big[e^{-uW_1}\big]\mathbb{E}\Big[\frac{2c}{1+2c}e^{-uT_2} + \frac{1}{1+2c}\Big].\end{aligned}$$

Proceeding in a similar fashion for T_2 and T_3, we obtain the system of equations

$$\left\{\begin{array}{ll} \mathbb{E}\big[e^{-uT_1}\big] & = \mathbb{E}\big[e^{-uW_1}\big]\mathbb{E}\big[\frac{2c}{1+2c}e^{-uT_2} + \frac{1}{1+2c}\big], \\ \mathbb{E}\big[e^{-uT_2}\big] & = \mathbb{E}\big[e^{-uW_2}\big]\mathbb{E}\big[\frac{K}{1+K}e^{-uT_1} + \frac{1}{1+K}e^{-uT_3}\big], \\ \mathbb{E}\big[e^{-uT_3}\big] & = \mathbb{E}\big[e^{-uW_3}\big]\mathbb{E}\big[e^{-uT_2}\big]. \end{array}\right. \tag{46}$$

Inserting the first and the third equation into the second one and simplifying we get:

$$\mathbb{E}\big[e^{-uT_2}\big] = \frac{cK(2Kc+u)}{(c+Kc+u)(1+2c+u)(2Kc+u) - 2c^2K(1+2u+2c(1+K))}.$$

Similarly, we get the other results from Proposition 6.7. □

Remark 6.10. We have computed the Laplace transforms of T_1, T_2, and T_3. Similar arguments also yield their Fourier transform $\mathbb{E}[e^{-2\pi i u T_1}]$. By inverting the Laplace transform, we get that the density of T_1 is equal to

$$\begin{aligned} f_{T_1}(s) = {} & \frac{-2c^2e^{sr_2}K^2r_1 + 2c^2e^{sr_3}K^2r_1 + 2c^2e^{sr_1}K^2r_2 - 2c^2e^{sr_3}K^2r_2 + ce^{sr_1}r_1r_2 - ce^{sr_2}r_1r_2}{(r_1-r_2)(r_1-r_3)(r_2-r_3)} \\ & + \frac{3Kce^{sr_1}r_1r_2 - 3Kce^{sr_2}r_1r_2 + e^{sr_1}r_1^2r_2 - e^{sr_2}r_1r_2^2 - 2c^2K^2e^{sr_1}r_3}{(r_1-r_2)(r_1-r_3)(r_2-r_3)} \\ & + \frac{2c^2K^2e^{sr_2}r_3 - ce^{sr_1}r_1r_3}{(r_1-r_2)(r_1-r_3)(r_2-r_3)} \\ & + \frac{ce^{sr_3}r_1r_3 - 3Kce^{sr_1}r_1r_3 + 3Kcs^{sr_3}r_1r_3 - e^{sr_1}r_1^2r_3 + ce^{sr_2}r_2r_3 - ce^{sr_3}r_2r_3}{(r_1-r_2)(r_1-r_3)(r_2-r_3)} \\ & + \frac{3cKe^{sr_2}r_2r_3 - 3cKe^{sr_3}r_2r_3 + e^{sr_2}r_2^2r_3 + e^{sr_3}r_1r_3^2 - e^{sr_3}r_2r_3^2}{(r_1-r_2)(r_1-r_3)(r_2-r_3)}, \end{aligned}$$

where $r_1 \le r_2 \le r_3$ are the three roots of the polynomial

$$P(x) = x^3 + (3Kc + 3c + 1)x^2 + (2K^2c^2 + 4Kc^2 + 2c^2 + 3Kc + c)x + 2c^2K^2.$$

6.7 The site frequency spectrum (SFS)

Since information about the past mutation history of lineages is retained in the ISM, one may want to consider more informative summary statistics. One of the most frequently used examples is the site frequency spectrum (SFS), which for a sample of size n in the ISM is given by a vector $(\zeta_1^n, \dots, \zeta_{n-1}^n)$, with ζ_i^n denoting the number of sites at which the derived allele is observed i times. In the case where we do not know which allele is ancestral and which is derived, the folded site frequency spectrum $(\eta_1^n, \dots, \eta_{\lfloor n/2 \rfloor}^n)$ can be used instead, where η_i^n is the number of sites where the two variants are observed with multiplicities $i : n-i$ or $n-i : i$. The SFS is well understood for the classical Kingman coalescent `K`, and thus also in the case `W`, since the weak seed bank coalescent is just the Kingman multiplied by a constant ([ZT12], formula 1).

Now, as we will show, it is possible to compute the expected site frequency spectrum in some particular cases in scenarios `K` (for comparison), `S` and `TI`. For this purpose, we first show an efficient way to compute mean times to the most recent common ancestor using phase-type distribution theory. For this aim, for some $n^{(1)}, n^{(2)} \in \mathbb{N}$, define $\mathfrak{C}$ as the totally ordered set of configurations

$$\begin{aligned} \big\{ & (n^{(1)} + n^{(2)}, 0), (n^{(1)} + n^{(2)} - 1, 1), \dots, (0, n^{(1)} + n^{(2)}), (n^{(1)} + n^{(2)} - 1, 0), \dots, \\ & (0, n^{(1)} + n^{(2)} - 1), \dots, (2,0), (1,1), (0,2), (\partial, \partial) \big\} \end{aligned}$$

(where we introduced the elements in increasing order), where (i,j) stands for "i active and j dormant lineages" for all (i,j), except for (∂, ∂), which is a "death state" corresponding to the presence of only one lineage, no matter whether active or dormant. Moreover, define $(Z(t))_{t \ge 0}$ as a continuous-time Markov chain on $\mathfrak{C}$, with the Q-matrix given by

$$Q_{(n,m),(\bar{n},\bar{m})} = \begin{cases} \binom{n}{2} & \text{if } (\bar{n},\bar{m}) = (n-1,m),\ n+m \neq 2, \\ 1 & \text{if } n=2,\ m=0,\ \bar{n}=\bar{m}=\partial, \\ n & \text{if } (\bar{n},\bar{m}) = (n-1,m+1), \\ Km & \text{if } (\bar{n},\bar{m}) = (n+1,m-1), \\ 0 & \text{if } (n,m) = (\partial,\partial), \\ -(\binom{n}{2}+n+Km) & \text{if } (\bar{n},\bar{m}) = (n,m) \neq (\partial,\partial), \\ 0 & \text{else,} \end{cases}$$

for any (n,m), $(\bar{n},\bar{m}) \in \mathfrak{C}$, using the convention $\binom{1}{2} = \binom{0}{2} = 0$, and started in $(n^{(1)}, n^{(2)}) \in \mathfrak{C}$. Then, the following lemma holds:

Lemma 6.11. *Take a sample of $n^{(1)}$ active and $n^{(2)}$ dormant individuals in* `S` *with $n^{(1)} + n^{(2)} \geq 2$. Define $T_{MRCA}^{n^{(1)},n^{(2)}}$ to be the time to the most recent common ancestor of the sample, $\alpha_{n^{(1)},n^{(2)}}$ as a vector of length $|\mathfrak{C}|-1$ being 1 in its $(n^{(2)}+1)$-th entry*[18] *and 0 else, $\bar{Q}$ as the square matrix obtained by removing the last row and column of Q*[19] *and $\boldsymbol{e}$ as the vector of ones of length $|\mathfrak{C}|-1$. Then,*

$$\mathbb{E}[T_{MRCA}^{n^{(1)},n^{(2)}}] = -\alpha_{n^{(1)},n^{(2)}} \bar{Q}^{-1} \boldsymbol{e}.$$

Proof. Proof by direct application of formula (6) from Chapter 2.4 with the initial condition $\pi = \alpha_{n^{(1)},n^{(2)}}$, the reward vector $r = (1,1,\ldots,1)$ and the reduced Q-matrix $\mathbf{S} = \bar{Q}$. □

In `TI` the same result holds with Q defined accordingly. E.g., for $\alpha = 1$, $\alpha' = \sqrt{K}$ we get

$$Q_{(n,m),(\bar{n},\bar{m})} = \begin{cases} \binom{n}{2} & \text{if } (\bar{n},\bar{m}) = (n-1,m),\ n+m \neq 2, \\ K\binom{m}{2} & \text{if } (\bar{n},\bar{m}) = (n,m-1),\ n+m \neq 2, \\ 1 & \text{if } n=2,\ m=0, \bar{n}=\bar{m}=\partial, \\ K & \text{if } m=2,\ n=0, \bar{n}=\bar{m}=\partial, \\ n & \text{if } (\bar{n},\bar{m}) = (n-1,m+1), \\ Km & \text{if } (\bar{n},\bar{m}) = (n+1,m-1), \\ 0 & \text{if } (n,m) = (\partial,\partial), \\ -(K\binom{m}{2}+\binom{n}{2}+n+Km) & \text{if } (\bar{n},\bar{m}) = (n,m) \neq (\partial,\partial), \\ 0 & \text{else.} \end{cases}$$

In `K`, of course, we know from literature that $\mathbb{E}[T_{MRCA}^{n}] = 2(1-1/n)$, from which we can derive that for `W`,

$$\mathbb{E}[T_{MRCA}^{n}] = 2\beta^2(1-1/n).$$

Remark 6.12. Figure 13 gives us the expected times to the most recent common ancestor for a sample made of n active individuals for n from 2 to 15. The values for

[18] Corresponding to the state $(n^{(1)}, n^{(2)}) \in \mathfrak{C}$. That is, the first entry if $n^{(2)} = 0$, the second entry if $n^{(2)} = 1$, and so on.

[19] Corresponding to the state $(\partial, \partial) \in \mathfrak{C}$.

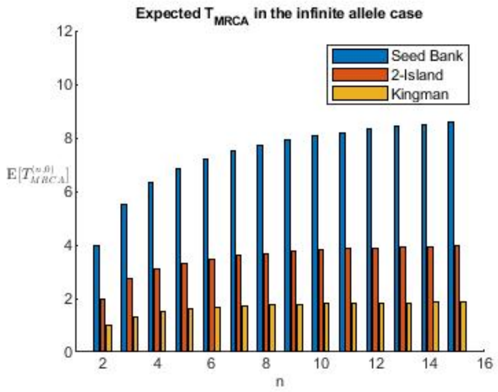


Figure 13: Expected times to the most recent common ancestor. $u = 0.5$; in the seed bank and the 2-island case, $c = K = 1$.

$n = 2$ and $n = 10$ in the seed bank coalescent match with those obtained by Blath, Eldon, González Casanova, Kurt and Wilke Berenguer ([BGE$^+$15], Table 1); however, while in that paper the values are obtained via simulations, our phase-type approach gives us exact results.

Having computed the mean times to the T_{MRCA}, we can now calculate the expected SFS. This can be done in several ways; we will show two of those. The first algorithm we show is quite easy to apply in practice, but has somewhat restrictive hypotheses on the sample:

Proposition 6.13. *Suppose we are either in* K, *in* TI *or in* S, *with generic c and K and all mutation rates equal to $u > 0$. Moreover, suppose we have a sample of total size $n \geq 2$. Additionally, in* S *and* TI, *suppose one of these cases holds:*

- *First case: in* S, *the sample is taken uniformly from active individuals; in* TI, *the sample is taken uniformly from the first island;*
- *Second case: in* S, *the sample is taken uniformly from dormant individuals; in* TI, *the sample is taken uniformly from the second island;*
- *Third case: the number of active individuals (in* S*) or individuals from the first island (in* TI*) is given by a random variable B with $B \sim Bin(n, \frac{K}{1+K})$, resulting in a randomly mixed sampling (the total size of the sample, n, still being fixed a priori).*

Then, the recursive formula

$$\mathbb{E}[\zeta_l^{n-1}] = \frac{n-l}{n}\mathbb{E}[\zeta_l^n] + \frac{l+1}{n}\mathbb{E}[\zeta_{l+1}^n], \quad 1 \leq l \leq n-2,$$

holds (in all three cases and all three models), which, along with the boundary condition

$$\mathbb{E}[\zeta_{n-1}^n] = un\mathbb{E}\Big[T_{MRCA}^{n,0} - T_{MRCA}^{n-1,0}\Big]$$

(S *and* TI, *first case),*

$$\mathbb{E}[\zeta_{n-1}^n] = un\mathbb{E}\Big[T_{MRCA}^{0,n} - T_{MRCA}^{0,n-1}\Big],$$

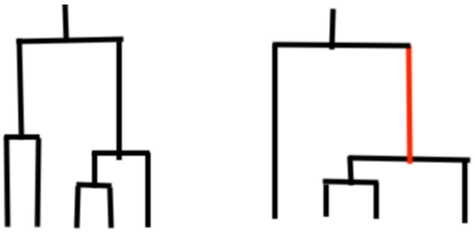


Figure 14: In the ancestral tree on the left, no anti-singletons are possible since the last merger does not involve a block of size 1. In the ancestral tree on the right, however, any mutation on the lineage marked in red will result in an anti-singleton.

(S *and* TI*, second case),*

$$\mathbb{E}[\zeta_{n-1}^{n}] = u\sum_{x=0}^{n}\binom{n}{x}\Big(\frac{K}{K+1}\Big)^{x}\Big(\frac{1}{K+1}\Big)^{n-x}\times$$
$$\times\ \mathbb{E}\Big[nT_{MRCA}^{x,n-x} - xT_{MRCA}^{x-1,n-x} - (n-x)T_{MRCA}^{x,n-x-1}\Big]$$

(S *and* TI*, third case),*

$$\mathbb{E}[\zeta_{n-1}^{n}] = un\mathbb{E}\Big[T_{MRCA}^{n} - T_{MRCA}^{n-1}\Big]$$

(K)*, and the Lemma (which gives us the explicit* $T_{MRCA}^{i,j}$*), gives us* $\mathbb{E}[\zeta_1^n, \ldots, \zeta_{n-1}^n]$ *in closed form for any* n.

Proof. We prove the statement for the cases K and S; the proof in the case TI is almost the same.
First of all, we consider the anti-singletons, i.e. the sites where the original allele can only be observed once in the sample. Consider their mean number τ_n in the case K at first. Imagine the coalescent tree of our sample of size n and compare it to the coalescent tree obtained by removing one of the n individuals. The T_{MRCA} will be different if and only if the last merger involved a block of size 1 containing exactly the individual which we removed, an event which happens with a probability we will denote with p_n.

Secondly, a necessary condition for the presence of anti-singletons is that the last merger (see picture) involves a block of size 1 (else it would be impossible to see a mutation in all individuals save one). If we denote this event with A, then $\mathbb{P}\{A\} = np_n$ by symmetry. Finally, the mean number of anti-singletons given A is, by construction of the infinite sites model, equal to the mutation rate times the difference of T_{MRCA} (that is, the time between the second-to-last and the last coalescences). These three arguments produce the formula

$$\tau_n = u\ \mathbb{E}[T_{MRCA}^{n} - T_{MRCA}^{n-1}]\ n.$$

This algorithm can be extended to the case S by re-adjusting the formula to a sample made of $n^{(1)}$ active and $n^{(2)}$ dormant individuals, with $\tau_{n^{(1)},n^{(2)}}$ the corresponding mean number of anti-singletons, which is then equal to

$$\tau_{n^{(1)},n^{(2)}} = u\mathbb{E}[nT_{MRCA}^{n^{(1)},n^{(2)}} - n^{(1)}T_{MRCA}^{n^{(1)}-1,n^{(2)}} - n^{(2)}T_{MRCA}^{n^{(1)},n^{(2)}-1}].$$

Now, we will use an argument introduced in [SKS16] (part of the proof of Lemma 1), which holds both for `K` and `S`. Take a sample of size $n-1$ and consider the number of sites ζ_l^{n-1} where exactly l individuals exhibit a mutation with $1 \le l \le n-2$. Assume our sample was obtained by taking n individuals first and then removing (randomly and uniformly) one of them we will denote with I. Then, ζ_l^{n-1} is the sum of the number of sites where there are l mutants (in the n-sample) and I is one of them and the number of sites where there are $l+1$ mutants (in the n-sample) and I is not among them. Given the fact that the choice of I is clearly independent of the number of mutants, we get for the corresponding expected values

$$\mathbb{E}[\zeta_l^{n-1}] = \frac{n-l}{n}\mathbb{E}[\zeta_l^n] + \frac{l+1}{n}\mathbb{E}[\zeta_{l+1}^n].$$

However, notice that, by definition of anti-singleton, we have $\mathbb{E}[\zeta_{n-1}^n] = \tau_{n,0}$ (in the first case, and similarly else). Therefore we can calculate $\mathbb{E}[\zeta_j^i]$ recursively using the following algorithm:
1) $n = 2$.
2) Calculate $\mathbb{E}[\zeta_{n-1}^n]$ using the anti-singleton formula.
3) If applicable, calculate $\mathbb{E}[\zeta_{n-2}^n]$ from $\mathbb{E}[\zeta_{n-1}^n]$ and $\mathbb{E}[\zeta_{n-2}^{n-1}]$ using the recursive formula with $l = n-2$.
4) Use the recursive formula to calculate $\mathbb{E}[\zeta_{n-3}^n]$, $\mathbb{E}[\zeta_{n-4}^n]$, ... $\mathbb{E}[\zeta_1^n]$ as well.
5) If $n < i$, increase n by 1 and restart from step 2. Else, we are done.

□

Notice that the recursive formula has the drawback that it is not applicable when the hypothesis $1 \le l \le n-2$ is not met. This is the main reason of the fact that we cannot compute the expected SFS for all samples (in particular, if we have a sample made of $n^{(1)}$ active and $n^{(2)}$ dormant individuals, where $n^{(1)}$ and $n^{(2)}$ are fixed and none is equal to 0) using the method from the proposition.

Moreover, this method does not work in the case where the mutation rates of active and dormant individuals are different. The reason for this is that here the average number of anti-singletons also depends on whether immediately after the second-to-last merger we have an active and a dormant individual or two active ones.
In these cases, we can use a different algorithm, which calculates the expected SFS by considering the entire structured coalescent instead of just its block-counting process.

Remark 6.14. In [K3], a variant of this algorithm, which increases computational efficiency by simplifying the state space via a quotient, is shown. However, in this thesis we will present the algorithm employing the full seed bank coalescent for instructive purposes.

For a sample of $n^{(1)}$ active and $n^{(2)}$ dormant individuals with $n^{(1)}+n^{(2)} = n$, recall the seed bank n-coalescent from Chapter 1 (that is, the coalescing process defined in [BGKWB16]). Moreover, notice that even if $|\mathcal{P}_n^{\{p,s\}}|$, the cardinality of $\mathcal{P}_n^{\{p,s\}}$, can be difficult to calculate, it is clearly finite. Thus, we can define a bijective function (using the notation $[n] := \{1, 2, \dots, n\}$ for $n \in \mathbb{N}$)

$$\Psi : [|\mathcal{P}_n^{\{p,s\}}|] \to \mathcal{P}_n^{\{p,s\}},$$

imposing $\Psi(1) = \{\{1, \dots, n\}^p\}$, $\Psi(2) = \{\{1, \dots, n\}^s\}$, a convention that we will use later. Furthermore, define, for $i, j \in [|\mathcal{P}_n^{\{p,s\}}|]$, $Q_{i,j} := Q_{\Psi^{-1}(i),\Psi^{-1}(j)}$, which induces a matrix as well.

Proposition 6.15. *Suppose we are in* S*, with generic c and K and mutation rates equal to $u > 0$ among the active and $\lambda u \geq 0$ among the dormant individuals.*

Then, if α is defined as the vector of length $|\mathcal{P}_n^{\{p,s\}}| - 2$ with

$$\alpha := \mathbb{1}_{\Psi^{-1}(\{\{1\}^p,\dots,\{n^{(1)}\}^p,\{n^{(1)}+1\}^s,\dots,\{n^{(1)}+n^{(2)}\}^s\})},$$

$\bar{Q}$ as the square matrix of size $|\mathcal{P}_n^{\{p,s\}}| - 2$ obtained by removing the first two rows and columns of Q and r as the function from $\mathcal{P}_n^{\{p,s\}}$ to $[0, \infty)$ defined by

$$r(\pi) = \begin{cases} u & \text{if } \pi \text{ has is a block of size } n-1 \text{ labeled } p, \\ \lambda u & \text{if } \pi \text{ has is a block of size } n-1 \text{ labeled } s, \\ 0 & \text{else,} \end{cases}$$

the following equation holds for the mean number of anti-singletons:

$$\tau_{n^{(1)},n^{(2)}} = -\alpha \bar{Q}^{-1} r.$$

Moreover, if we are in one of the three cases described in Proposition 6.13, we can compute the entire expected SFS and we refer to the statement for the exact formulas.

Proof. Direct consequence of formula (7) from Chapter 2.4, where $\pi = \alpha$, $\mathbf{S} = \bar{Q}$ and r is the reward vector, and Proposition 6.13.

□

In TI the same result holds with Q defined accordingly. E.g., for $\alpha = 1$, $\alpha' = \sqrt{K}$ we get

$$Q_{\pi,\pi'} = \begin{cases} 1 & \text{if } \pi \prec_p \pi', \\ K & \text{if } \pi \prec_s \pi', \\ c & \text{if } \pi \bowtie \pi', \text{ and a p has been replaced by an s,} \\ Kc & \text{if } \pi \bowtie \pi', \text{ and an s has been replaced by a p,} \\ 0 & \text{else with } \pi \neq \pi', \\ -\sum_{\pi' \neq \pi} Q_{\pi,\pi'} & \text{if } \pi = \pi'. \end{cases}$$

Proposition 6.16. *Suppose we are in* S*, with generic c and K and mutation rates equal to $u > 0$ among the active and $\lambda u \geq 0$ among the dormant individuals. For any $\pi \in \mathcal{P}_n^{\{p,s\}}$, define $n_1^j(\pi)$ and $n_2^j(\pi)$ as the number of blocks of π which are of size j and have, respectively, the p-label or the s-label. That is, the number of p- or s- branches with j descendants.*

Moreover, let α and $\bar{Q}$ be defined as before, and, for any $j \in \{1, \ldots, n^{(1)}+n^{(2)}-1\}$, define $r^{(j)}$ as the function from $\mathcal{P}_n^{\{p,s\}}$ to $[0, \infty)$ with

$$r^{(j)}(\pi) := un_1^j(\pi) + \lambda un_2^j(\pi).$$

Then,

$$\mathbb{E}[\zeta_j^{n^{(1)},n^{(2)}}] = -\alpha\bar{Q}^{-1}r^{(j)},$$

giving us the entire expected SFS in the case of a sample of exactly $n^{(1)}$ active and $n^{(2)}$ dormant individuals.

Proof. We take into account the entire coalescent process. By definition, mutations on branches with one descendant give rise to singletons in the SFS, while mutations on branches with two or three descendants give rise to doubletons, tripletons and so on (as seen in [HSB19], Algorithm 3.4; see picture). Therefore, $\zeta_j^{n^{(1)},n^{(2)}}$ can be seen as the number of all mutations on branches with j descendants, and as such, as the sum of $\zeta_j^{n^{(1)},n^{(2)}(p)}$ (number of all mutations on "p" branches with j descendants) and $\zeta_j^{n^{(1)},n^{(2)}(s)}$ (same with "s" branches). If we denote with Z_j the sum of the lengths of all branches with j descendants[20], then $Z_j = Z_j^p + Z_j^s$, where Z_j^p is the sum of the length of all branches with j descendants and a "p" label (where mutations happen at rate u) and Z_j^s the same with "s" (where mutations happen at rate λu). Finally, for the expected values,

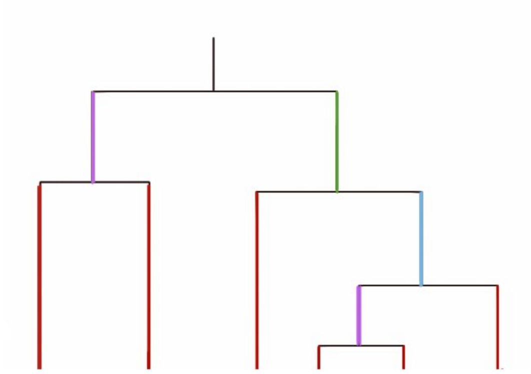


Figure 15: An ancestral tree. Any mutation on a red lineage results in a singleton, i.e. the number of mutations on red lineages is equal to the first entry of the SFS. In the same way, the number of mutations on purple/blue/green... lineages is respectively equal to the second/third/fourth... entry of the SFS.

$$\mathbb{E}[\zeta_j^{n^{(1)},n^{(2)}}] = \mathbb{E}[\zeta_j^{n^{(1)},n^{(2)}(p)} + \zeta_j^{n^{(1)},n^{(2)}(s)}] = \mathbb{E}[uZ_j^p + \lambda uZ_j^s],$$

which, using formula (8) in [HSB19] with $\pi = \alpha$ the initial condition vector, $\mathbf{S} = \bar{Q}$ the reduced Q-matrix and r the reward vector, and taking account of the multiplicity of the branches, yields the thesis.

[20] Notice that some of those branches might coexist at some time point, hence the multiplicity adjustment at the end.

□

Remark 6.17. It is exceedingly complicated to use this theorem in practice, because the state space of our Markov chain (and thus, the dimension of $\bar{Q}$) gets very big even for small n ($\dim(\bar{Q})$=20 for $n = 3$; 92 for $n = 4$). For $n = 2$ and $n = 3$, the results match with those obtained with the recursive formula. This problem can be mostly solved by reducing the state space via a quotient as in [K3], and we refer to the paper for the description of the algorithm in full detail (Section 2.2). We will simply state the equivalency relation $\sim$ inducing the quotient here:
We say that $\pi \sim \pi'$ for $\pi, \pi' \in \mathcal{P}_n^{\{p,s\}}$ if and only if for every $n \in \mathbb{N}$, π and π' have the same amount of p-blocks and of s-blocks of size n. For example,

$$\{\{1\}^p, \{2\}^p, \{3\}^s, \{4\}^s\} \sim \{\{1\}^s, \{2\}^s, \{3\}^p, \{4\}^p\},$$

since both labeled partitions have two p-blocks and two s-blocks of size one each; however,

$$\{\{1\}^p, \{2\}^p, \{3\}^s, \{4\}^s\} \nsim \{\{1, 2\}^p, \{3\}^s, \{4\}^s\},$$

since the former has no p-block of size 2. The elements of $\mathcal{P}_n^{\{p,s\}}/\!\sim$ can then be represented via vectors of length $2n$ (where n is of course the total sample size). In such a vector v, the i-th entry $v(i)$ gives us the number of p-blocks of size i for $i \leq n$ and the number of s-blocks of size $i-n$ for $i > n$. This quotientation reduces the size of the state significantly: e.g. if $n = 3$, $\mathcal{P}_n^{\{p,s\}}/\!\sim$ has just 10 elements, of which two are in the absorbing class, namely, in vector form, $(0, 0, 0, 1, 0, 0, 0, 0)$ and $(0, 0, 0, 0, 0, 0, 0, 1)$; in the phase-type formalism, the matrix $\mathbf{S}$ is therefore only a 8×8 matrix, compared to 20×20 without the quotient. In the case of a sample of active individuals only, this simplification allowed for the algorithm to be implemented; the resulting figures match exactly the ones obtained via Proposition 6.13.

6.8 Expected normalized and normalized expected SFS

Another interesting quantity is the normalized expected site frequency spectrum (NESFS) $(\hat{E\zeta_1^n}, \ldots, \hat{E\zeta_{n-1}^n})$, as introduced in [EBBF15] (p.13). It is defined by

$$\hat{E\zeta_i^n} := \frac{\mathbb{E}[\zeta_i^n]}{\sum_{l=2}^n l\mathbb{E}[T_l]},$$

T_l being the time during which there are l lineages, $l = 2, 3, \ldots n$. In other words, $\sum_{l=2}^n l\mathbb{E}[T_l]$ is the average tree length.[21] The name stems from the fact that the NESFS is a first-order approximation of the expected value of the normalized SFS, which is defined ([EBBF15], p.9) as

$$\hat{\zeta_i^n} := \frac{\zeta_i^n}{\zeta_1^n + \cdots + \zeta_n^n}.$$

The distribution of the resulting random vector is very insensitive to the mutation rate, provided it is not too small, facilitating practical inference when the mutation rate is unknown [EBBF15, Supporting Information, pages SI12 – SI13].

[21] The expected value of T_l can be computed using Formula (6) from Chapter 2.4 with an appropriate reward function.

Figure 16 provide two illustrations on the expected SFS, with and without normalization. Throughout, we set $K = c = u = 1$, and fix a sample size of $n = 15$. Blue bars correspond to `S`, green to `TI` with $\alpha = \alpha' = 1$, and yellow to `K` for comparison.

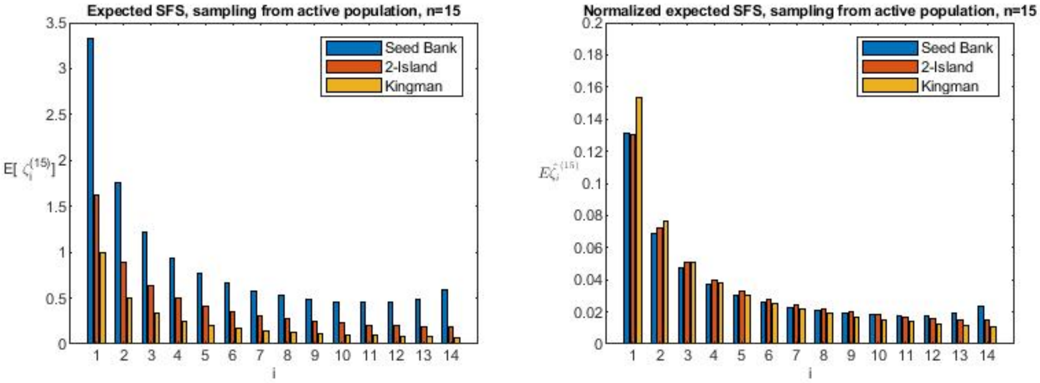


Figure 16: Expected SFS when sampling is purely from the active population.

It is noteworthy that the magnitude of entries in the SFS varies dramatically between the three models, while `S` and `TI` have very similar normalized spectra. The implication is that all three models are straightforward to tell apart if the population-rescaled mutation rate is known, but that a larger sample, or a more informative summary statistic, is needed to distinguish `S` from `TI` when it is unknown.

6.9 Quasi-stationarity and the Yaglom limit

Let us consider a sample of two in the general case. As we have noticed during the calculations of the F_{ST}, we can see the behavior of the two individuals as a Markov chain $(M(t))$, with the state space formed by "two lineages in the Y-subpopulation", "one lineage in X and one in Y", "two lineages in the X-subpopulation", and "coalescence", respectively associated to the abbreviations $\{x_1, \dots, x_4\}$. This Markov chain has a unique absorbing state, namely the coalescence state; from this we can ask ourselves the question: does a limit for the probabilities $\mathbb{P}\{M(t) = i | M(t) \neq 0\}$, i.e. the probability to be in state i at time t conditioned on non-coalescence by that time, exist for $t \to \infty$? The answer, as we will see, is yes, although the closed-form formula for the distribution (which we will call, from now on, the *quasi-stationary distribution*, see [MV12]) is very complicated.
First, let's formalize our thoughts by using the definitions from ([MV12]) (Definitions 1 and 2):

Definition 6.18. If we have a stochastic process $(M(t))_{t\geq 0}$ with state space E and set of transient states E^*, we say $(M(t))$ has a Yaglom limit if there exists a probability measure μ on E^* such that, for any $x \in E^*$ and any measurable set $A \subset E^*$,

$$\lim_{t\to\infty} \mathbb{P}_x\{M(t) \in A | M(t) \in E^*\} = \mu(A).$$

If the previous equation holds for any $t > 0$ (not just in the limit) provided the starting distribution is μ, μ is called a quasi-stationary distribution as well.

Using Theorem 7 in ([MV12]) and imposing $E = \{x_1, x_2, x_3, x_4\}$, we can prove that our Markov chain admits a unique quasi-stationary distribution, which coincides with the Yaglom limit. Moreover, the limit distribution is the unique positive left eigenvector of the Q-matrix of the Markov chain restricted to $E^* = \{x_1, x_2, x_3\}$, the Q-matrix being equal to

$$S = \begin{bmatrix} -2Kc & 2Kc & 0 \\ c & c+Kc & Kc \\ 0 & 2c & -2c-1 \end{bmatrix}$$

in the seed bank model and equal to

$$S = \begin{bmatrix} -2Kc-K & 2Kc & 0 \\ c & c+Kc & Kc \\ 0 & 2c & -2c-1 \end{bmatrix}$$

in the standard two-island model ($\alpha = 1,\ \alpha' = \sqrt{K}$).
However, as in subchapter 6.3, the computation of the matrix exponentials turns out to be messy. Therefore, we will just give a numerical example here. In the case $K = c = 1$, we get approximately

$$(0,289; 0,524; 0,186) \text{ and } (0,219; 0,562; 0,219)$$

as the quasi-stationary distributions on $\{x_1, x_2, x_3\}$. Notice how the distribution in the seed bank model drifts away from the state “two active individuals”, which communicates with the absorbing state, and how the distribution in the two-island model is symmetrical, as we would expect it to be.

6.10 Discussion

While in the two-subpopulation case, the F_{ST} is useless in order to infer on the presence of weak seed banks, this is not the case for both the strong seed bank and the two-island scenarios. In fact, the F_{ST} in the two-island case is approximately twice as big as in the seed bank case, given all other parameters stay the same. However, it is nearly impossible, given the F_{ST}, to infer whether there is mutation among the dormant individuals. Similar results occur in the infinitely many alleles and the infinitely many sites model.

Regarding the SFS, the normalized site frequency spectrum in the ISM is also of little use for distinguishing between the strong seed bank and two-island scenarios. However, the Kingman scenario exhibits a lighter tail and more singletons on average. In contrast, if we have the entire SFS, the three models will differ remarkably.

7 Conclusion

7.1 Conclusions

In this thesis, we addressed the problem of examining many different extensions of the Wright-Fisher diffusion, all exhibiting some form of population structure. Both the case where all subpopulations exhibit the same reproductive mechanisms and the case where some of the subpopulations consist entirely of dormant individuals remaining inactive until resuscitation have been thoroughly examined.

In the first chapter, background regarding the population structures and the mutation models, was presented, while in the second one some tools helping us throughout the thesis were outlined. In the third chapter, the first basic results, ensuring the platform we work on is stable, were provided. In particular, we focused on uniqueness of the stationary distribution and its characterization via mixed moments.

After that, we moved to the main focus of the thesis, which is analyzing similarities and differences between the two-island and the seed bank models. This was mainly done in chapters 4 and 5. The first one tackled the several different ways to handle the boundary behavior problem: if we have two competing alleles, will one of them go temporarily extinct in finite time with positive probability? We saw that, while the speed and scale approach is not applicable here and simple comparisons do not give a sharp bound, both the Lyapunov and the McKean argument combined with a theorem from polynomial diffusions entirely solve our problem; the critical value is $1/2$ (in the case $\alpha = 1$). Moreover, Chapter 5 approached the interesting topic of scaling limits. We saw that the time rescaling is crucial here and that our main allies in this chapter are represented by Möhle's Lemma for the convergence of the coalescent processes and duality for convergence of the diffusion processes. Our contribution was adapting this tool to our purposes in order compute a scaling limit of continuous-time Markov processes. Even more interesting was the fact that in one case, the scaling limit we got was an entirely new, non-trivial diffusion process with jumps.

From a practical point of view, our thesis' main aim was to discuss several popular summary statistics ("measures of population structure") of some population genetic models (mainly, the seed bank and the two-island diffusion), both in the finite-allele and the infinite-allele case. This was done in order to distinguish the patterns of genetic variability produced by those models, and was the main aim of Chapter 6. In particular, the sample heterozygosity, the F_{ST}, which is a measure of how the population differs between subpopulations, and the site frequency spectrum were taken into consideration. Most relevantly, results showed that both the F_{ST} and the expected SFS can be used to distinguish between the seed bank and two-island scenarios. Moreover, the latter, along with the nSFS, can be used in order to differentiate those two from a population where the seed bank is weak or even non-existent (Kingman scenario).

7.2 Future Work

Despite many topics regarding structured extensions of the Wright-Fisher diffusion having been tackled in this thesis, many more have been left for the future.

For example, this thesis mainly focused on the seed bank model and the two island model. Other diffusion models with a geographical structure have only been taken into consideration sparingly. For example, nothing was said about models exhibiting large reproductive events, i.e. individuals giving birth to a positive fraction of the population. This reflects itself in the fact that the diffusion process exhibits jumps here and is the go-to model in the study of genetic patterns of Atlantic cod [EW08].

Regarding the mutation models, an interesting example we did not investigate is the so-called infinitely many genes model [BP14, BHP12]. In this model, reproduction is given by Wright-Fisher diffusion-like dynamics. However, each of the individuals carries a *set* of genes. These sets can be different between the individuals, do not have to be of the same cardinality and only contain genes which don't influence the fitness of the individual (in particular, they can be lost without problems). Every offspring can *lose* each of its parent's genes, or *gain* a new gene for its set, which has never been seen before and which does not override any of the existing ones.

If we only look at topics that were thoroughly handled within the thesis, the easiest possibility for future work is, in our opinion, Chapter 6.3, since there are many ideas that could be exploited to try to compute, or at least approximate, the matrix exponential. We think that it should be doable to prove that the sample heterozygosity decay in absence of mutation is exponential with parameter $1/5$ (in case $\mathtt{S}$), as the plots and simulations strongly suggest. Another topic it would be very interesting to work on is applying the Lyapunov argument to a broader class of Markov processes, since it gave good results (often even reaching the optimal bound) when used hitherto.

References

[ADLM99] S. Ablett, A. H. Darke, P. J. Lillford, and D. R. Martin. Glass formation and dormancy in bacterial spores. *International journal of food science & technology*, 34(1):59–69, 1999.

[Alf15] A. Alfonsi. *Affine diffusions and related processes: simulation, theory and applications*. Springer, 2015.

[Asm08] S. Asmussen. *Applied probability and queues*, volume 51. Springer Science & Business Media, 2008.

[BB08] M. Birkner and J. Blath. Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. *J. Math. Biol.*, 57(3):435–465, 2008.

[BBE13] M. Birkner, J. Blath, and B. Eldon. An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics*, 193(1):255–290, 2013.

[BEK18] N. Biswas, A. Etheridge, and A. Klimek. The spatial Λ-Fleming-Viot process with fluctuating selection, 2018.

[BGE$^+$15] J. Blath, A. González Casanova, B. Eldon, N. Kurt, and M. Wilke-Berenguer. Genetic variability under the seedbank coalescent. *Genetics*, 200(3):921–934, 2015.

[BGKS13] J. Blath, A. González Casanova, N. Kurt, and D. Spanò. The ancestral process of long-range seed bank models. *J. Appl. Probab.*, 50(3):741–759, 2013.

[BGKWB16] J. Blath, A. González Casanova, N. Kurt, and M. Wilke-Berenguer. A new coalescent for seed-bank models. *Ann. Appl. Probab.*, 26(2):857–891, 2016.

[BHP12] F. Baumdicker, W. R. Hess, and P. Pfaffelhuber. The infinitely many genes model for the distributed genome of bacteria. *Genome biology and evolution*, 4(4):443–456, 2012.

[Bla05] M. Bladt. A review on phase-type distributions and their use in risk theory. *ASTIN Bulletin: The Journal of the IAA*, 35(1):145–161, 2005.

[BN17] M. Bladt and B. F. Nielsen. *Matrix-exponential distributions in applied Probability*, volume 81. Springer, 2017.

[BP14] F. Baumdicker and P. Pfaffelhuber. The infinitely many genes model with horizontal gene transfer. *Electronic Journal of Probability*, 19, 2014.

[Bru91] M.-F. Bru. Wishart processes. *Journal of Theoretical Probability*, 4(4):725–751, 1991.

[CFR14] P. Carr, T. Fisher, and J. Ruf. On the hedging of options on exploding exchange rates. *Finance and Stochastics*, 18(1):115–144, 2014.

[CKRT12] C. Cuchiero, M. Keller-Ressel, and J. Teichmann. Polynomial processes and their applications to mathematical finance. *Finance and Stochastics*, 16(4):711–740, 2012.

[DGP07] R. Dong, A. Gnedin, and J. Pitman. Exchangeable partitions derived from Markovian coalescents. *Ann. Appl. Probab.*, 17(4):1172–1201, 2007.

[dHP17] F. den Hollander and G. Pederzani. Multi-colony Wright-Fisher with seed-bank. *Indag. Math. (N.S.)*, 28(3):637–669, 2017.

[Dur08] R. Durrett. *Probability models for DNA sequence evolution.* Springer Science & Business Media, Chicago, 2008.

[EBBF15] B. Eldon, M. Birkner, J. Blath, and F. Freund. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics*, 199(3):841–856, 2015.

[EG09] A.M. Etheridge and R.C. Griffiths. A coalescent dual process in a Moran model with genic selection. *Theoretical Population Biology*, 75(4):320–330, 2009. Sam Karlin: Special Issue.

[EGT10] A.M. Etheridge, R.C. Griffiths, and J.E. Taylor. A coalescent dual process in a Moran model with genic selection, and the λ-coalescent limit. *Theoretical Population Biology*, 78(2):77–92, 2010.

[EK86] S.N. Ethier and T.G. Kurtz. *Markov processes: Characterization and convergence.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986.

[Eth11] A. Etheridge. *Some mathematical models from population genetics*, volume 2012 of *Lecture Notes in Mathematics.* Springer, Heidelberg, 2011. Lectures from the 39th Probability Summer School held in Saint-Flour, 2009, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

[EW08] B. Eldon and J. Wakeley. Linkage disequilibrium under skewed offspring distribution among individuals in a population. *Genetics*, 178(3):1517–1532, 2008.

[Fel51] W. Feller. Diffusion processes in genetics. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability.* The Regents of the University of California, 1951.

[FGH03] R. Fu, A.E. Gelfand, and K.E. Holsinger. Exact moment calculations for genetic models with migration, mutation, and drift. *Theoretical Population Biology*, 63(3):231–243, 2003.

[Fis99] R. A. Fisher. *The genetical theory of natural selection: a complete variorum edition.* Oxford University Press, 1999.

[FL16] D. Filipović and M. Larsson. Polynomial diffusions and applications in finance. *Finance Stoch.*, 20(4):931–972, 2016.

[GJL16] R. C. Griffiths, P. A. Jenkins, and S. Lessard. A coalescent dual process for a Wright–Fisher diffusion with recombination and its application to haplotype partitioning. *Theoretical population biology*, 112:126–138, 2016.

[GS18] A. González Casanova and D. Spanò. Duality and fixation in ξ-Wright–Fisher processes with frequency-dependent selection. *The Annals of Applied Probability*, 28(1):250–284, 2018.

[HA07] M. Hutzenthaler and R. Alkemper. Graphical representation of some duality relations in stochastic population models. *Electronic Communications in Probability*, 12:206–220, 2007.

[Her94] H.M. Herbots. *Stochastic models in population genetics: genealogical and genetic differentiation in structured populations.* PhD thesis, University of London, 1994.

[HL66] J. L. Hubby and R. C. Lewontin. A molecular approach to the study of genic heterozygosity in natural populations. I. the number of alleles at different loci in Drosophila Pseudoobscura. *Genetics*, 54(2):577–594, 1966.

[HSB19] A. Hobolth, A. Siri-Jégousse, and M. Bladt. Phase-type distributions in population genetics. *Theoretical population biology*, 127:16–32, 2019.

[Hud90] R. R. Hudson. Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, 7(1):44, 1990.

[JH85] C. R. Johnson and R. A. Horn. *Matrix analysis.* Cambridge University Press, 1985.

[JK14] S. Jansen and N. Kurt. On the notion(s) of duality for Markov processes. *Probab. Surv.*, 11:59–120, 2014.

[Kim55] M. Kimura. Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences of the United States of America*, 41(3):144, 1955.

[Kim64] M. Kimura. Diffusion models in population genetics. *Journal of Applied Probability*, 1(2):177–232, 1964.

[Kin82a] J. F. C. Kingman. The coalescent. *Stoch Proc Appl*, 13:235–248, 1982.

[Kin82b] J. F. C. Kingman. On the genealogy of large populations. *Journal of applied probability*, 19(A):27–43, 1982.

[KKL01] I. Kaj, S. M. Krone, and M. Lascoux. Coalescent theory for seed bank models. *J. Appl. Probab.*, 38:285–300, 2001.

[KN97] S. M. Krone and C. Neuhauser. Ancestral processes with selection. *Theoretical Population Biology*, 51(3):210–237, 1997.

[KS98] I. Karatzas and S. E. Shreve. Brownian motion. In *Brownian Motion and Stochastic Calculus*, pages 47–127. Springer, 1998.

[KS13] C. Kleiber and J. Stoyanov. Multivariate distributions and the moment problem. *Journal of Multivariate Analysis*, 113:7–18, 2013.

[Kur91] T. G. Kurtz. Random time changes and convergence in distribution under the Meyer-Zheng conditions. *The Annals of probability*, pages 1010–1034, 1991.

[Kus67] H. J. Kushner. Stochastic stability and control. Technical report, Brown U., Providence RI, 1967.

[KZH08] A.R.R. Kermany, X. Zhou, and D.A. Hickey. Joint stationary moments of a two-island diffusion model of population subdivision. *Theoretical Population Biology*, 74(3):226–232, 2008.

[Lig12] T. M. Liggett. *Interacting particle systems*, volume 276. Springer Science & Business Media, 2012.

[LJ11] J. T. Lennon and S. E. Jones. Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nat. Rev. Microbiol.*, 9(2):119–130, 2011.

[LM15] A. Lambert and C. Ma. The coalescent in peripatric metapopulations. *J. Appl. Probab.*, 52(2):538–557, 2015.

[LP17] M. Larsson and S. Pulido. Polynomial diffusions on compact quadric sets. *Stochastic Process. Appl.*, 127(3):901–926, 2017.

[LR14] M. Larsson and J. Ruf. Convergence of local supermartingales and Novikov-Kazamaki type conditions for processes with jumps, 2014.

[Mai77] B. Maisonneuve. Une mise au point sur les martingales locales continues définies sur un intervalle stochastique. *Séminaire de Probabilités (Strasbourg)*, 11:435–445, 1977.

[McK69] H. P. McKean. *Stochastic integrals*, volume 353. American Mathematical Soc., 1969.

[Mey66] P. A. Meyer. *Probability and potentials*, volume 1318. Blaisdell Pub. Co., 1966.

[MN16] M. Möhle and M. Notohara. An extension of a convergence theorem for Markov chains arising in population genetics. *J. Appl. Probab.*, 53(3):953–956, 2016.

[Möh98] M. Möhle. A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. *Adv. in Appl. Probab.*, 30(2):493–512, 1998.

[Mor59] P.A.P. Moran. The theory of some genetical effects of population subdivision. *Austral. J. Bio. Sci.*, 12(2):109–116, 1959.

[MPS11] E. Mayerhofer, O. Pfaffel, and R. Stelzer. On strong solutions for positive definite jump diffusions. *Stochastic processes and their applications*, 121(9):2072–2086, 2011.

[MT92] S. P. Meyn and R. L. Tweedie. Stability of Markovian processes I: Criteria for discrete-time chains. *Advances in Applied Probability*, 24(3):542–574, 1992.

[MT93] S. P. Meyn and R. L. Tweedie. Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, 25(3):518–548, 1993.

[MV12] S. Méléard and D. Villemonais. Quasi-stationary distribution and population processes. *Probab. Surv.*, 9:340–410, 2012.

[Neu75a] M. F. Neuts. Computational uses of the method of phases in the theory of queues. *Computers & Mathematics with Applications*, 1(2):151–166, 1975.

[Neu75b] M. F. Neuts. Probability distributions of phase type. *Liber Amicorum Prof. Emeritus H. Florin*, 1975.

[NG93] H.B. Nath and R.C. Griffiths. The coalescent in two colonies with symmetric migration. *Journal of Mathematical Biology*, 31(8):841–851, 1993.

[Not90] M. Notohara. The coalescent and the genealogical process in geographically structured population. *Journal of mathematical biology*, 29(1):59–75, 1990.

[Ruf15] J. Ruf. The martingale property in the context of stochastic differential equations. *Electronic Communications in Probability*, 20, 2015.

[RW00] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes, and martingales. Vol. 2.* Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000.

[RY99] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edition, 1999.

[Sha88] M. Sharpe. *General theory of Markov processes*, volume 133. Academic press, 1988.

[SKS16] J. P. Spence, J. A. Kamm, and Y. S. Song. The site frequency spectrum for general coalescents. *Genetics*, 202(4):1549–1561, 2016.

[SL18] W. R. Shoemaker and J. T. Lennon. Evolution with a seed bank: the population genetic consequences of microbial dormancy. *Evolutionary applications*, 11(1):60–75, 2018.

[Sla91] M. Slatkin. Inbreeding coefficients and coalescence times. *Genetics Research*, 58(2):167–175, 1991.

[SS80] T. Shiga and A. Shimizu. Infinite-dimensional stochastic differential equations and their applications. *J. Math. Kyoto Univ.*, 20(3):395–416, 1980.

[SV07] D. W. Stroock and S. R. S. Varadhan. *Multidimensional diffusion processes.* Springer, 2007.

[TLL$^+$11] A. Tellier, S. J. Y. Laurent, H. Lainer, P. Pavlidis, and W. Stephan. Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proc. Natl. Acad. Sci. U.S.A.*, 108(41):17052–17057, 2011.

[Wak09] J. Wakeley. *Coalescent Theory: An Introduction.* Roberts & Company Publishers, Greenwood Village, 2009.

[Wat69] S. Watanabe. On two dimensional Markov processes with branching property. *Transactions of the American Mathematical Society*, 136:447–466, 1969.

[Whi02] W. Whitt. *Stochastic-process limits.* Springer Series in Operations Research. Springer-Verlag, New York, 2002.

[Won64] E. Wong. The construction of a class of stationary Markoff processes. *Stochastic processes in mathematical physics and engineering*, 17:264–276, 1964.

[Wri31] S. Wright. Evolution in Mendelian populations. *Genetics*, 16(2):97–159, 1931.

[Wri49] S. Wright. The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354, 1949.

[YW71] T. Yamada and S. Watanabe. On the uniqueness of solutions of stochastic differential equations. *J. Math. Kyoto Univ.*, 11:155–167, 1971.

[ZT12] D. Zivkovic and A. Tellier. Germ banks affect the inference of past demographic events. *Molecular Ecology*, 21(22):5434–5446, 2012.