Λ-Fleming-Viot Processes and their Spatial Extensions

Habib Saadi

Department of Statistics

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy (DPhil)

October 2011

Declaration

All material presented in this thesis is my own work, except where otherwise stated.

Signed: *Date:*
Habib Saadi

....

To my family.

Acknowledgements

First of all, I could never thank enough Alison Etheridge for her supervision and mentoring. Not only did she provide me with the perfect conditions for my DPhil, I also learned from her a whole new way of thinking, developing my intuition and appreciating mathematical beauty. I learned as much during our meetings as during the coffee breaks, when she was subtly teaching in her clear and elegant style. These conversations remind me that Mathematics are and should always be a pleasure.
The results in Chapter 2 are the fruit of our collaboration. In particular, the idea of the proof of Theorem 2.9 in the case $B = 1$ is due to Alison.

I am also immensely indebted to Geoff Nicholls, who has acted as my second supervisor. The MCMC moves in Chapter 3 were conceived with his help. I wholly owe him my new birth as a Bayesian, and I must acknowledge his almost-surely infinite patience, as it took some time for my conversion to sink in... Thanks to his unique ability to give the perfect example to illustrate a concept, he illuminated my understanding of most things I thought I knew about Statistics.

My college advisor Bob Griffiths, who has done a lot more than his duty, must know how much I owe him - although he would categorically deny it. He always took time he did not have to answer my questions on Population Genetics, Probability, Computing, and Yoga. It is a unique chance to meet someone who so kindly shares his considerable experience and knowledge!

I have been incredibly lucky to be part of such a friendly and vibrant community in the Statistics Department in Oxford. I would like to thank everybody for contributing to make this place such a positive environment. I

am grateful to Bjarki Eldon, Christina Goldschmidt, Chris Yau, Amandine Véber, Daniel Straulino, Charlotte Dean, Robin Ryder, Luke Miller, Simon Myers, Xavier Didelot, Louigi Addario-Berry, David Steinsaltz, Dario Spano and many others for countless enjoyable and stimulating conversations. I would like to thank more particularly Nick Freeman, Mladen Savov, James Martin, Martin Kolb and Emily Tann for useful feedback on Chapters II and V of this thesis.

We are all indebted to Jan Boylan, Beverley Lane, Jennie McKenzie, the exceptionally competent and dedicated ithelp@stats team, and all the Department staff I have not mentioned.

I would like to thank my office colleagues Sebastian Kelm, Waqar Ali, Philip Schmidt, Jamie Hill, and our regular guests Yoonjoo Choi and Mireille Gomez from the OPIG team, for making these four years truly memorable. We tried our very best to have only meaningless conversations and banter, and it is still a mystery to me how we failed and learned so much from each other... I certainly did learn a lot, mostly about Biology, Bioinformatics and International Relations.

This page is also the perfect occasion to thank all the people who helped me during my interdisciplinary journey. I am very grateful to Hervé Drevon, Guillaume Haberer, Hermun Durand, Dominique Guégan, Jean-Christophe Tavanti, Philippe Fontaine, Thierry Chauveau, Philippe Bougerol, Francis Comets, Sylvie Méléard, Philippe Robert, Philippe Chassaing, Jean Bertoin, Jean Jacod and Matthias Winkel for their support and precious advice.

I would like to thank those who supported me through the busy times of a DPhil: My parents, my brother, my friends Jeremy, Mathew, Sylvester, Cecily and all the others who are far too many to be all mentioned here. I would like to acknowledge the bands AC/DC and Rammstein, who provided the electrifying energy I needed to write. I do apologise if these names horrify you, but you would not be reading this thesis if it were not for their

music!

Finally, I would like to thank my examiners Ellen Baake and Christina Goldschmidt for their helpful comments, and for making the examination a very positive experience.

Abstract

The subject of this thesis is the study of certain stochastic models arising in Population Genetics. The study of biological evolution naturally motivates the construction and use of sometimes sophisticated mathematical models.

We contribute to the study of the so-called Λ models. Our work is divided into two parts. In Part I, we study non-spatial models, introduced in 1999. Although there is a very rich literature concerning the description of genetic diversity thanks to the genealogies arising in these models, we obtain new results by considering the dynamics of the full population. We also contribute by presenting the first Bayesian method that allows us to reconstruct the genealogies generated by these models from data.

In Part II, we study a recent extension of these models to the spatial setting. In particular, we prove a non trivial result concerning the geographical dispersal of a new mutant under this model.

Contents

Preface

The subject of this thesis is the study of certain stochastic models arising in Population Genetics. The study of biological evolution naturally motivates the construction and use of sometimes sophisticated mathematical models. Since the work of Fisher, Wright and Haldane, there has been a tradition of fertile exchange of ideas between mathematics and biology in the study of population genetics. Development of biological knowledge and techniques provides more and more questions to solve and data to analyse, whereas progress made in mathematics and statistics provides models that help in answering these questions. Certain mathematical models actually contributed to shape the way we think about evolution. This is for example the case with the Kingman coalescent (Kingman, 1982).

It is now common knowledge that the information that encodes how every organism functions, develops and repairs itself is contained in the *genotype*. For most species, the genotype presents itself in the form of chromosomes, which are long double strands of DNA. During reproduction, the genetic material of the parents is combined to create the childrens' genetic material. In the case of asexual reproduction, the childrens' genotype is actually a copy of the parental genotype.

A fundamental property of the genotype is the fact that its transmission from parents to children is not perfect. Duplication errors, the mutations, cause variation in the genotype. This mechanism is the primary cause of biological evolution.

If we look at the individual level, a very important step is to understand the consequences of these mutations on the organism function (the *phenotype*). The link between genotype and phenotype is incredibly complex. Although many things are known in biology, integrating them in a quantitative framework is still the object of very active research. However, it is crucial to take into account this aspect, because for a given

environment, certain mutations have a deleterious effect on the organism, while others have a beneficial effect.

Population genetics aims to explain biological evolution by studying the effects of these genetic phenomena at the population level. A population is a community of individuals, usually from the same species. Each of these individuals reproduces and dies. In particular, new genotypes are created as a consequence of mutations. When we look at the genetic composition of a population, we can classify all the different genotypes into a certain number of genetic types, called *alleles*. For example, we may want to know how many people carry a gene known to cause some serious condition. Having defined these genetic types, we can count their frequencies of occurrence within the population.

A priori, it is not obvious why evolution has to be studied from a population perspective. It becomes clearer if we remember that one of the objectives of evolution theory is to explain the existence of species, and the great variety of forms that life can take. As we said earlier, genetics teaches us that every new allele arises through mutation. So any innovation is initially carried by a single individual. The challenge is to understand how this innovation is transferred to the whole population. Stated otherwise, it is important to understand how genetic type frequencies change through time.

A fundamental force that explains this dynamic is called *genetic drift*. We illustrate it in a very simple setting. If an individual is the single carrier of an allele A and dies without reproducing, then the type has become extinct. On the other hand, it may successfully reproduce until all the population at a later time is descended from it and carries allele A. This randomness in allele frequencies due to pure chance in the number of offspring is the genetic drift.

Other phenomena arise when we consider the population. For example, the effect of a mutation on the organism's survival may depend on the genetic types of the other individuals. This is why we measure the effect of a mutation on survival in relative terms. For example, if we have two types a and A, we say that A has a fitness s relatively to a if the probability of survival of A is $1+s$ times larger than the probability of survival of a when both types of individuals are in presence.

The fundamental idea in evolution theory that made Darwin famous is the concept of natural selection. It simply says that if $s > 0$, A will survive better than a, and therefore the frequency of A in the population will increase until it may eventually reach 100%. On the other hand, if $s < 0$, the frequency of A will decrease and A may ultimately disappear from the population.

When one wants to formalise a theory of evolution, probability theory and stochastic processes become fundamental tools. The reason is that mutations arise as random events. At the same time, the lifetimes of individuals, the number of their offspring and their interactions also appear to be inherently random phenomena. The fundamental laws of genetics are probabilistic in nature. A classical application in medicine is the evaluation of the risk that a future child carries a gene causing a disease given the fact that certain members of her family have the disease.

As such, much care has been devoted to constructing stochastic models that help us to understand evolution. A population genetics model needs to:

- Describe the reproduction mechanism
- Model the causes of death (accidents, competition, lack of natural resources,...)
- Represent the mutation process
- Model the genotype-phenotype mapping, although in a simplistic way
- Describe the environment (geography, natural resources).

As we can see, a realistic model would become very quickly intractable. Although people have in mind that in reality all these aspects are important to understand evolution and to analyse biological data, the models actually used in Population Genetics are much simpler. The main objective is to capture a universal behaviour, for example how genetic drift and natural selection act on genetic diversity. To obtain such an understanding, one does not really need all the details of the genotype-phenotype mapping, but only the relative fitness of the few genes we are considering. The models are written in a simple and generic way to reach general conclusions, and the parameters of these models summarise a very large quantity of information.

One of the objectives in Population Genetics is to assess the relative strength of drift, natural selection, mutation and population structure (spatial structure, cultural structure, etc...). In particular, understanding each of these factors separately from the others is important. A model that assumes that there is no natural selection at the genes studied is called a *neutral model*.

This thesis is a contribution in the study of some neutral models. The emphasis is clearly on the mathematical and statistical aspects, rather than on the biological situations that motivate these models. It is divided into two parts. In Part I, which includes Chapters 1 to 3, we study non-spatial models, and in Part II, made of Chapters 4 and 5, we focus on the influence of space.

In Chapter 1, we present the fundamental neutral models and some of their applications. We start with the Wright-Fisher model, its diffusion approximation, and the coalescent approach. We present a significant result, the Ewens sampling formula, and its extension to the full allele distribution thanks to the Poisson-Dirichlet distribution. Then we introduce a more recent class of models, the Λ-coalescent trees and their forward in time equivalents, the Λ-Fleming-Viot processes. We conclude with the lookdown construction, which is a very powerful tool in the study of these models.

In Chapter 2, we present some original results about Λ-Fleming-Viot processes (ΛFV). The objective is to obtain similar results to those related to the Poisson-Dirichlet distribution in the case of the classical Fleming-Viot process. We prove that for a certain class of ΛFV, if we use a new mutation mechanism, we can explicitly identify the stationary distribution of the allele frequencies. The very surprising result is that we still obtain the Poisson-Dirichlet distribution, but with different parameters.

In Chapter 3, we contribute to the statistical analysis of Λ-coalescent trees. There is currently no Bayesian method that allows us to reconstruct Λ-coalescent trees and to infer their parameters. We propose such a method, based on a Reversible-Jump MCMC algorithm. We present the model and the MCMC algorithm we developed, and illustrate it with simulated data in the case of the Eldon-Wakeley coalescent.

In Chapter 4, we present models that take space into account. We start with two classical approaches, the Stepping-Stone model due to Kimura (1953) and Malécot's Formula (Malécot, 1948). In particular we see that discrete and continuous space settings give the same results concerning the large scale behaviour of the probability of

identity. We then introduce a recent model due to Barton et al. (2010a), the Spatial Λ-Fleming-Viot model (SΛFV), that extends the ΛFV processes by incorporating a continuous spatial structure. The main point is to show how the SΛFV overcomes the mathematical difficulties brought by working in a spatial continuum.

In Chapter 5, we present an original result in the study of the SΛFV process. We analyse the geographical range of a newly created mutation. The central question is how far a new mutant type spreads. The result we obtain is particularly suprising, as it states that with probability one, the mutant will stop spreading in finite time, occupying only a bounded set. Because the structure of the SΛFV model is radically different from the ones existing in the literature, we develop an original method for proving this result. The main component is a technique that handles the complex geometry of the problem.

Finally, we conclude by presenting an agenda for future work.

Part I

Λ-Fleming-Viot Processes

Chapter 1

Background

In this Chapter, we present some of the classical models used in population genetics, as well as some recent models developed in the last decade. The concepts, objects and results presented here are used in the original research we present in Chapters 3 and 4. But before talking about the models, let us just remind ourselves a few useful probability distributions.

1.1 Some probability distributions

1.1.1 The Dirichlet distribution

We begin by introducing some notation for the set of frequencies.

Notation 1.1. *We denote by Δ_K the unit K-dimensional simplex*

$$\Delta_K := \{(x_1, \ldots, x_k) : x_j \geq 0, x_1 + \cdots + x_K \leq 1\}$$

Definition 1.2. *The Dirichlet distribution $Dir_K(\alpha_1, \ldots, \alpha_K)$ of order $K \geq 2$ with parameters $\alpha_1, \ldots, \alpha_K > 0$, is a probability distribution on Δ_{K-1} with density*

$$f(x_1, \ldots, x_{K-1}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_K)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_K)} x_1^{\alpha_1 - 1} \ldots x_K^{\alpha_K - 1}, \tag{1.1}$$

where $x_1, \ldots, x_{K-1} \in \Delta_{K-1}$ *and* $x_K := 1 - \sum_{i=1}^{K-1} x_i$.

The special case $K = 2$ *is the Beta distribution with parameters* (α_1, α_2). *When* $\alpha_1 = \cdots = \alpha_K = \alpha$, *the distribution is denoted by* $Dir_K(\alpha)$.

The Dirichlet distribution is used to define random frequencies $X_1, \ldots, X_K > 0$ summing up to 1. We say that K random variables $X_1, \ldots, X_K$ follow a $\mathrm{Dir}_K(\alpha_1, \ldots, \alpha_K)$ distribution if the random vector $(X_1, \ldots, X_{K-1})$ admits a density given by (1.1), and $X_1 + \cdots + X_K = 1$.

A convenient way of constructing a $\mathrm{Dir}_K(\alpha)$ random vector is through the use of Gamma random variables.

Definition 1.3. *The* $Gamma(\varepsilon, \lambda)$ *distribution is a probability distribution on* $\mathbb{R}$ *with density*

$$f(x) = \frac{\lambda^\varepsilon}{\Gamma(\varepsilon)} x^{\varepsilon - 1} e^{-\lambda x}, \quad x > 0,\ \lambda, \varepsilon > 0.$$

Proposition 1.4. *Let* $Y_1, \ldots, Y_K$ *be independent* $Gamma(\varepsilon, 1)$ *random variables, and define* $S = Y_1 + \cdots + Y_K$ *and* $p_i = Y_i/S$, $i = 1, \ldots, K$. *Then* S *is independent from* $(p_1, \ldots, p_K)$, *the distribution of* S *is* $Gamma(K\varepsilon, 1)$, *and* $(p_1, \ldots, p_K)$ *follows a* $Dir_K(\varepsilon)$ *distribution.*

The result follows from a straightforward change of variable calculation to calculate the density of the vector $(p_1, \ldots, p_{K-1}, S)$.

1.1.2 The Poisson-Dirichlet distribution

We are now considering infinite vectors of frequencies summing up to 1. This is why we introduce the following sets.

Notation 1.5. *We denote by* Δ *be the set of* frequencies*:*

$$\Delta := \left\{x = (x_1, x_2, \ldots) : x_1, x_2, \cdots \geq 0, \sum_{i=1}^{\infty} x_i \leq 1\right\}, \tag{1.2}$$

and by Ξ *the set of* ranked frequencies

$$\Xi := \left\{x = (x_1, x_2, \ldots) : x_1 \geq x_2 \geq \cdots \geq 0, \sum_{i=1}^{\infty} x_i \leq 1\right\}. \tag{1.3}$$

Finally, we defined the operator $\mathcal{O}rd$ *the following way:*

$$\begin{aligned} \mathcal{O}rd : \quad \Delta \quad &\longrightarrow \quad \Xi \\ (x_1, x_2, \ldots) &\longmapsto (x_{(1)}, x_{(2)}, \ldots), \end{aligned}$$

where $x_{(1)} \geq x_{(2)} \geq \ldots$ *are the frequencies ranked in decreasing order.*

We allow for the sum of the frequencies to be less than 1, which will happen in certain circumstances.

In the next sections, we will be interested in approximating the $\mathrm{Dir}_K(\varepsilon)$ distribution as K tends to infinity. However, because of the symmetry of the distribution, this is not directly possible. For example, one can see that a vector $(p_1, \ldots, p_K)$ following a $\mathrm{Dir}_K(\varepsilon)$ distribution must satisfy

$$\begin{cases} \mathbb{E}(p_1) = \cdots = \mathbb{E}(p_K) = 1/K \\ \mathbb{E}(p_1) + \cdots + \mathbb{E}(p_K) = 1, \end{cases}$$

and these two relations cannot hold for $K = \infty$.

The solution, proposed in Kingman (1975), is to rank the frequencies in decreasing order. More precisely, consider the probability generating functional G_K of the points $Y_1, \ldots, Y_K$, which is defined for functions $g : [0, \infty) \to \mathbb{R}$ by

$$G_K(g) = \mathbb{E}\big[g(Y_1) \ldots g(Y_K)\big].$$

We are going to rescale ε such that

$$K\varepsilon \longrightarrow \theta \quad \text{as} \quad K \longrightarrow \infty.$$

Under this rescaling, we can show that

$$G_K(g) \xrightarrow[K\to\infty]{} \exp\Big(- \theta \int_0^1 \big(1 - g(u)\big) \frac{e^{-u}}{u} du \Big).$$

The limit is the quantity

$$\mathbb{E}\big[\prod_{i=1}^{\infty} g(X_i)\big],$$

where the X_i's are the points of a Poisson Point Process Π with intensity $\theta e^{-u}/u$. This shows that the set of points $\{Y_1, \ldots, Y_K\}$ converges weakly to Π. We consider the ranked points $X_{(1)} \geq X_{(2)} \geq \ldots$, and we can prove that the sum

$$S_\infty := \sum_{i=1}^{\infty} X_{(i)}$$

is almost surely finite and follows a Gamma distribution.

Definition 1.6. *With the previous definition, the distribution of the infinite vector*

$$\frac{1}{S_\infty}\big(X_{(1)}, X_{(2)}, \dots\big) \in \Xi$$

is called the Poisson-Dirichlet distribution with parameter θ, and is denoted by $\mathcal{PD}(\theta)$.

It turns out that the Poisson-Dirichlet distribution is not very amenable to calculations. However, if we reorder the frequencies in a specific way, we obtain a distribution which is much nicer to work with: the GEM distribution.

1.1.3 The GEM distribution

To define the GEM distribution, we need first to define a size-biased reordering of a vector of frequencies.

Definition 1.7. *Consider a vector of frequencies $x = (x_1, x_2, \dots) \in \Delta$, and define $\overline{x}$ by*

$$\overline{x} = \sum_{j=1}^{\infty} x_j.$$

The size-biased reordering *of x is the random vector $Z = (Z_1, Z_2, \dots) \in \Delta$ defined by*

$$Z_n = x_{\pi(n)},$$

where π is a random permutation of $(1, 2, \dots)$ with distribution

$$\mathbb{P}\big[\pi(1) = i_1, \dots, \pi(n) = i_n\big] = \frac{x_{i_1}}{\overline{x}} \, \frac{x_{i_2}}{\overline{x} - x_{i_1}} \, \cdots \frac{x_{i_n}}{\overline{x} - x_{i_1} - \cdots - x_{i_{n-1}}}.$$

The interpretation is the following. Consider an urn containing balls with colours $(1, 2, \dots)$ such that the colour j has frequency $x_j/\overline{x}$. We first draw a ball at random, and the probability that it is of colour i_1 is $x_{i_1}/\overline{x}$. We remove all the balls of colour i_1 from the urn, and we draw at random a second ball, say of colour i_2. The probability

is now $x_{i_2}/(\overline{x} - x_{i_1})$. If we repeat this process until we empty the urn, the permutation π simply records the order of appearance of the colours.

Definition 1.8. *Consider a vector X following a $\mathcal{PD}(\theta)$ distribution for some $\theta > 0$. By definition, the distribution of the size-biased reordering of X is called the GEM distribution with parameter θ and is denoted by GEM(θ).*

The following property is a direct consequence of the definition.

Proposition 1.9. *If Z is a GEM(θ) random vector, then Z', the size-biased reordering of Z, is also a GEM(θ) random vector.*

The power of the GEM distribution lies in its *stick-breaking* representation, which makes it analytically tractable.

Proposition 1.10. *Consider $(Z_1, Z_2, \dots)$ i.i.d. Beta$(1, \theta)$ random variables. We construct a sequence $(X_1, X_2, \dots)$ of random variables using the following* stick-breaking *scheme:*

$$\begin{cases} X_1 = Z_1 \\ X_n = (1 - Z_1)(1 - Z_2) \dots (1 - Z_{n-1}) Z_n, \ n \geq 2. \end{cases}$$

The distribution of $(X_1, X_2, \dots)$ is GEM(θ).

1.2 Classical models

1.2.1 Wright-Fisher model

For simplicity we assume that our individuals are haploid, which means that they have only one copy of each chromosome. It means the reproduction is asexual, as each individual is able to reproduce alone. This is the case for example for bacteria, which reproduce by cellular division.

1.2.1.1 Reproduction mechanism and genetic drift

Definition 1.11 (K-allele Wright-Fisher model)**.** *The dynamics takes place in discrete time. Consider a population of constant size N. Each individual has one of K allelic types $A_1, \dots, A_K$. Each individual from generation $n+1$ chooses independently his unique parent uniformly at random among the individuals from generation n, and the child has the same type as his parent.*

The above definition simply describes the *reproduction mechanism*. Because population genetics is about the change of allele frequencies, we are going to see how allele frequencies change in this model.

We denote by $Y_k(n)$ the number of individuals of type A_k at time n. The process $\big((Y_1(n), \dots, Y_K(n)),\, n \geq 0\big)$ is a Markov chain with state space the K-dimensional simplex of level N:

$$D_N^K := \{(x_1, \dots, x_K) \in \mathbb{N}^K : x_1 + \cdots + x_K = N\}.$$

Conditionally on $(Y_1(n), \dots, Y_K(n)) = (x_1, \dots, x_K)$, $(Y_1(n+1), \dots, Y_K(n+1))$ has multinomial distribution with parameters $(x_1/N, \dots, x_K/N)$. Therefore we have

$$\begin{aligned} & \mathbb{P}\Big(Y_1(n+1) = y_1, \dots, Y_K(n+1) = y_K \,|\, Y_1(n) = x_1, \dots, Y_K(n) = x_K\Big) \\ = \;& \frac{N!}{y_1! \dots y_K!} \Big(\frac{x_1}{N}\Big)^{y_1} \cdots \Big(\frac{x_K}{N}\Big)^{y_K}. \end{aligned}$$

In particular, the probability that $Y_k(n+1) = N$ is $(x_k/N)^N$. This implies that at every stage, there is a positive probability that the Markov chain hits a state $(0, \dots, 0, N, 0, \dots, 0)$ where it will remain forever. This is called *fixation*. Because the state space D_N^K is finite, fixation will happen with probability one.

Fixation is an illustration of what biologists call *drift*, that is the fact that genetic diversity decreases due to the randomness in the reproduction mechanism. Even if we

started with K alleles, we ultimately lost all but one of them because of chance. Drift is one of the most important evolutionary forces and we are going to devote a lot of effort to studying it.

1.2.1.2 Mutations

We need to introduce now mutations that are going to modify the genetic material of an individual. We are in the situation where there are in total K alleles, so any mutation acting on an individual of type A_i is going to change its type to another one, say A_j. We include this mechanism in the definition of the Wright-Fisher model

Definition 1.12 (Mutations in the Wright-Fisher model)**.** *Consider the reproduction mechanism in the Wright-Fisher model, and denote by A_i the allelic type of the parent. Independently of the other children, the child type changes to A_j, $i \neq j$ with probability μ_{ij}, and remains A_i with probability $1 - \sum_{j \neq i} \mu_{ij}$.*

This formulation is very general, the only limitation being that the number of alleles is finite.

The distribution of $(Y_1(n+1), \ldots, Y_K(n+1))$ conditionally on $(Y_1(n), \ldots, Y_K(n)) = (x_1, \ldots, x_K)$ is also multinomial, but with different parameters. The contribution to allele A_i will come from individuals of type A_j, $j \neq i$ whose offspring do mutate to A_i, and this happens with probability μ_{ji}, as well as from individuals of type A_i whose offspring do not mutate, and this happens with probability $1 - \sum_{j \neq i} \mu_{ij}$. Therefore, if we define the conditional sampling probabilities ψ_i by

$$\psi_i := x_i(1 - \sum_{j \neq i} \mu_{ij}) + \sum_{i \neq j} \mu_{ji} x_j,$$

we obtain the transition probabilities for the allele counts:

$$\begin{aligned} & \mathbb{P}\Big(Y_1(n+1) = y_1, \dots, Y_K(n+1) = y_K \,|\, Y_1(n) = x_1, \dots, Y_K(n) = x_K\Big) \\ = \quad & \frac{N!}{y_1! \dots y_K!} \psi_1^{y_1} \dots \psi_k^{y_K}. \end{aligned}$$

Several results can be derived from these D_N^K-valued Markov chains. For example, they can be found in Ewens (2004, Chap. 3). Because exact calculations in the Wright-Fisher model are rarely possible, a natural way of analysing it is to make approximations as N tends to infinity. On the other hand, the mutation probabilities are in practice small, so we can consider the situation where μ_{ij} tends to zero. There is a formal way of using these approximations in a very powerful way. This is setting of the diffusion approximation.

1.2.2 Wright-Fisher diffusion

1.2.2.1 Diffusion processes

Let $\big(X(t)\big)_{t\geq 0}$ be a time homogeneous Markov process taking values in a metric space E. The *infinitesimal generator* of X is defined by

$$\mathcal{L}f(x) := \lim_{dt \to 0} \frac{\mathbb{E}\big[f\big(X(t+dt)\big) - f(x) \,|\, X(t) = x\,\big]}{dt},$$

where f is a function such that the limit exists. The limit does not depend on t because the process is time homogeneous. A one dimensional *diffusion* is a continuous path real valued process such that its generator $\mathcal{L}$ is a second order differential operator of the form

$$\mathcal{L}f(x) = a(x)f'(x) + \frac{1}{2}b^2(x)f''(x),$$

where $a(x)$ and $b(x)$ satisfy some regularity conditions. To define a diffusion, one only needs to identify $a(x)$ and $b(x)$. If we apply the generator $\mathcal{L}$ to the functions $f(x) = x$ and $f(x) = x^2$, we can see that

$$\lim_{dt\to 0} \frac{\mathbb{E}\big[X(t+dt) - x \,|\, X(t) = x\,\big]}{dt} = a(x) \tag{1.4}$$

$$\lim_{dt\to 0} \frac{\mathbb{E}\big[\big(X(t+dt) - x\big)^2 \,|\, X(t) = x\,\big]}{dt} = b^2(x), \tag{1.5}$$

which is why $a(x)$ and $b^2(x)$ are respectively called the *infinitesimal drift* and the *infinitesimal variance*.

In higher dimensions, the generator of an $\mathbb{R}^n$-valued diffusion is of the form

$$\mathcal{L}f(x_1,\ldots,x_n) = \sum_{i=1}^{n} a_i(x_1,\ldots,x_n)\,\frac{\partial f}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{n} (b\,b^T)_{ij}(x_1,\ldots,x_n)\,\frac{\partial^2 f}{\partial x_i \partial x_j},$$

where the a_i are real-valued functions, b is a $n \times n$ matrix valued function, and b^T is the transpose of b.

1.2.2.2 Convergence to a diffusion

Our objective is to approximate the discrete time, integer-valued Markov chain from the Wright-Fisher model by a continuous time, continuous path process. To understand how this approximation arises, we are going to focus on the case $K = 2$. Because $Y_1 + Y_2 = N$, the process is actually one dimensional, so we can restrict our attention to $Y = Y_1$.

We are going to perform three operations. We first embed $Y(n)$ in continuous time by defining $Y(t) = Y(n)$ for $t = n$, and $Y(t) = Y(n)$ for $n \le t < n+1$. We rescale time by an arbitrary factor, M say, that will tend to infinity. to obtain a continuous path,

at the same time we must rescale the size of the jumps so that they also tend to zero, but at a rate that needs to be determined.

The second operation is that we might need to modify the parameters of our model (we already know we are going to let the population size tend to infinity). This is why we first define a sequence Y_M of discrete time Markov chains for which the transition functions will depend on M. To take into account the time change, we introduce a sequence of continuous time processes $\widehat{Y}_M(t)$ defined by

$$\widehat{Y}_M(t) = Y_M(M\,t).$$

At the same time, we want the jumps size to tend to zero so that the limiting path becomes continuous, so our third operation is to rescale $\widehat{Y}_M$ by an arbitrary factor C_M depending on M. We define

$$X_M(t) = \frac{\widehat{Y}_M(t)}{C_M}.$$

It is this process X_M that we want to converge to a diffusion. To decide how to choose M and C_M, we are going to use equations (1.4) and (1.5). They imply that any diffusion $X(t)$ satisfies

$$\mathbb{E}\big[X(t+dt) - x \,|\, X(t) = x\,\big] = a(x)\,dt + o(dt)$$

$$\mathbb{E}\big[\big(X(t+dt) - x\big)^2 \,|\, X(t) = x\,\big] = b^2(x)\,dt + o(dt)$$

as dt tends to zero. The equivalent quantities for X_M (which is not a homogeneous Markov process) are obtained by formally taking $dt = 1/M$, which gives the following

set of conditions

$$\mathbb{E}\big[X_M(1/M) - x \,|\, X_M(0) = x\,\big] = \frac{a(x)}{M} + o(\frac{1}{M})$$
$$\mathbb{E}\big[\big(X_M(1/M) - x\big)^2 \,|\, X(0) = x\,\big] = \frac{b^2(x)}{M} + o(\frac{1}{M}).$$

Therefore the increment

$$\delta_M := \frac{Y_M(1) - Y_M(0)}{C_M}$$

needs to satisfy

$$M\,\mathbb{E}\big[\delta_M \,|\, Y_M(0) = C_M\,x\,\big] \xrightarrow[M\to\infty]{} a(x) \tag{1.6}$$

$$M\,\mathbb{E}\big[\delta_M^2 \,|\, Y_M(0) = C_M\,x\,\big] \xrightarrow[M\to\infty]{} b^2(x). \tag{1.7}$$

This informal approach can actually be made rigorous, although there is an extra technical condition to be met:

$$M\,\mathbb{E}\big[\delta_M^k \,|\, Y_M(0) = C_M\,x\,\big] \xrightarrow[M\to\infty]{} 0, \quad k \geq 3. \tag{1.8}$$

The following result can be found in Ethier and Kurtz (1986, Chapter 10, Theorem 1.1).

Theorem 1.13 (Diffusion approximation)**.** *If conditions (1.6), (1.7) and (1.8) are met, then the rescaled process* $(X_M)_{t\geq 0}$ *converges to the diffusion with generator*

$$\mathcal{L}f(x) = a(x)f'(x) + \frac{1}{2}b^2(x)f''(x).$$

1.2.2.3 Wright-Fisher diffusion

We return now to the one dimensional Wright-Fisher diffusion. We have introduced some arbitrary time scale parameter M, as well as the corresponding reweighting C_M. In our problem, we want to consider the allele frequency Y_M/N, this is why we parameterise the rescaling by taking $C_M = N$. We just have to identify the time scale parameter M that guarantees (1.6).

We know that $Y_M(1)$ is a binomial random variable with parameters (N, ψ), where ψ is given by

$$\psi = x(1 - \mu_{12}) + \mu_{21}(1 - x).$$

This means that

$$M\,\mathbb{E}\big[\delta_M \,|\, Y_M(0)/N = x\big] = M\Big((1-x)\mu_{21} - x\mu_{12}\Big),$$

and in order to satisfy condition (1.6), it suffices to take $M = N$ and

$$N\mu_{ij} \longrightarrow \beta_{ij} \quad \text{as } N \longrightarrow \infty. \tag{1.9}$$

We can then verify that with this choice conditions (1.7) and (1.8) are met. Therefore the rescaled Markov chain converges weakly to the diffusion with generator

$$\mathcal{L}f(x) = \big((1-x)\beta_{21} - x\beta_{12})\big)f'(x) + \frac{1}{2}x(1-x)f''(x).$$

For the K-dimensional case, the same rescaling holds, namely we measure time in units of N generations, and we rescale the mutation parameters according to (1.9). Recall that we denoted by Δ_K the unit K-dimensional simplex $\Delta_K := \{(x_1, \ldots, x_k) : x_j \geq 0,\, x_1 + \cdots + x_K \leq 1\}$.

Theorem 1.14. *The Wright-Fisher model with K alleles and mutation probabilities verifying*

$$N\mu_{ij} \xrightarrow[N\to\infty]{} \beta_{ij}, \quad 1 \le i, j \le K,$$

can be approximated by a diffusion taking values in Δ_{K-1}, with generator given by

$$\mathcal{L}f(x_1, \ldots, x_{K-1}) = \sum_{i=1}^{K-1} \Big(- x_i \sum_j \beta_{ij} + \sum_j x_j \beta_{ji} \Big) \frac{\partial f}{\partial x_i} + \frac{1}{2} \sum_{i=1}^{K-1} x_i(1 - x_i) \frac{\partial^2 f}{\partial x_i^2} - \sum_{1 \le i < j \le K-1} x_i x_j \frac{\partial^2 f}{\partial x_i \partial x_j},$$

where $x_K = 1 - x_1 - \cdots - x_{K-1}$.

The result is proved in Ethier and Kurtz (1986, Chapter 10, Theorem 1.1). A derivation of the coefficients based on Taylor's theorem is given in Etheridge (2011, Chapter 4, Section 1).

Although in general one cannot find an explicit expression for the stationary distribution of the Wright-Fisher diffusion for $K \ge 3$ alleles, in the special case of *parent independent* mutation, that is $\mu_{ij} = \mu_j$ for all i, one is actually able to calculate its stationary distribution. Any stationary distribution $\widetilde{\pi}$ must satisfy

$$\int \mathcal{L}f(x) \widetilde{\pi}(dx) = 0 \tag{1.10}$$

for every bounded and twice differentiable function f. We assume that $\widetilde{\pi}$ has density $\pi(x)$. An integration by parts of equation (1.10) provides a partial differential equation for π that one can solve. The details of the following result can be found in Durrett(2004), Theorem 8.5.

Theorem 1.15. *When the mutation probability is parent independent, the stationary distribution for the Wright-Fisher diffusion is the Dirichlet distribution with parameters* $(2\beta_1, \ldots, 2\beta_K)$*, with density*

$$C\, x_1^{2\beta_1 - 1} \ldots x_K^{2\beta_K - 1},$$

where C *is the normalising constant.*

An interesting special case is when the mutant is chosen uniformly among the other $K - 1$ types. In this situation, the mutation rates are all equal:

$$\beta_1 = \cdots = \beta_K = \frac{\varepsilon}{2}.$$

The density becomes

$$\frac{\Gamma(K\varepsilon)}{\Gamma(\varepsilon)^K} \left(x_1 \ldots x_K\right)^{\varepsilon - 1}.$$

If we let the number of alleles K tend to infinity and rescale the mutation rate ε such that $K\varepsilon \longrightarrow \theta$ we are now in the situation that allowed us to define the Poisson-Dirichlet distribution.

Theorem 1.16. *If* $K\varepsilon \longrightarrow \theta$ *as* $K \longrightarrow \infty$*, the ranked frequencies for the Wright-Fisher diffusion follow asymptotically the* $\mathcal{PD}(\theta)$ *distribution.*

This theorem is a weak convergence result. The Wright-Fisher diffusion is an object that allows us to manipulate an infinite population, but the number K of alleles remains finite. To actually work with infinitely many alleles we would need to construct an infinite dimensional diffusion. This is essentially what the Fleming-Viot process does (Fleming and Viot (1979)). We discuss the Fleming-Viot process in §1.2.4.

If we are interested in the distribution of types in a sample from our population,

there exists a much simpler approach, the coalescent method. Among other things, it allows us to work easily with infinitely many alleles.

1.2.3 The Kingman coalescent

1.2.3.1 Genealogy in the neutral case

We return to the discrete time Wright-Fisher model with population size N. So far we have been looking at the dynamics of the allele frequencies. Another approach consists in sampling $n \leq N$ individuals from the population at time 0 and tracing their genealogy backwards in time. We first take $n = 2$. Because of the uniform and independent sampling, the probability that two individuals have the same parent at generation -1 is $1/N$. Conditionally on the event where they have distinct parents, the probability that they have the same grand-parent at generation -2 is also $1/N$. Calling τ the time at which the two individuals have their *most recent common ancestor*, then from what preceded τ is a geometric random variable with success probability $1/N$. In particular, $\mathbb{E}(\tau) = N$, which means we have to go back N generations on average to find the most recent common ancestor for a sample of size two. This is the exact time scale we found in the forward in time diffusion approximation.

More generally, if we look at a sample of size n, we say that k lineages *coalesce at time* t when they all simultaneously merge into a common ancestor at the same time t. The probability that $n \geq 3$ individuals have the same parent in the previous generation is $1/N^{n-1} << 1/N$. The probability that two parents have each two children in the sample is of order $1/N^2$. More generally, the only non trivial event with probability of order $1/N$ is two lineages coalescing, all the other events have probabilities of order $o(1/N)$. These calculations allow us to describe the asymptotic genealogy of a sample of size n when the population size N tends to infinity. While there are n lineages, the probability of two lineages coalescing in the previous generation is $\binom{n}{2}/N$, so the time τ_n^N to the first coalescence event is approximately geometric with parameter $\binom{n}{2}/N$.

After τ_n^N, we have $n-1$ lineages, and because of independence between generations, the same argument applies. Therefore we have to wait an additional geometric time with parameter $\binom{n-1}{2}/N$ until a pair of lineages coalesce.

To obtain a continuous-time approximation of the genealogy, we are going to measure time in N generations, so we consider the random variable τ_k^N/N, and we have

$$\mathbb{P}(\tau_k^N \geq Nx) = \left(1 - \frac{k(k-1)}{2N}\right)^{Nx-1} \xrightarrow[N\to\infty]{} e^{-\binom{k}{2}x}.$$

Therefore τ_k^N/N converges in distribution to an exponential distribution with parameter $\binom{k}{2}$.

In order to properly define a genealogy, we first introduce some notation.

Notation 1.17. *We denote by $[n]$ the set of integers $\{1,\dots,n\}$, and by $\mathcal{R}_n$ the set of partitions of $[n]$. For $\eta \in \mathcal{R}_n$, let $|\eta|$ be the number of blocks of η. For $\eta, \xi \in \mathcal{R}_n$, we say that ξ is a* refinement of η*, and write $\xi \subset \eta$, if every block of η is union of blocks of ξ.*

We label by $\{1,\dots,n\}$ the individuals from our sample. We can represent a genealogy by an $\mathcal{R}_n$-valued Markov process $(\pi_t)_{t\geq 0}$ by saying that two integers i and j are in the same block of the partition π_t if and only if the corresponding individuals have a common ancestor at time t in the past. We can now define the Kingman n-coalescent.

Definition 1.18. *The $\mathcal{R}_n$-valued continuous time Markov chain with transition rates satisfying*

$$\forall \eta \neq \xi, \qquad Q_{\xi\eta} = \begin{cases} \binom{|\xi|}{2} & \text{if } \xi \subset \eta \text{ and } |\eta| = |\xi| - 1, \\ 0 & \text{otherwise,} \end{cases}$$

is called the Kingman n-coalescent.

We start with n lineages, and while there are k lineages, we wait for an exponential time T_k with rate $\binom{k}{2}$. At time T_k we sample uniformly a pair of lineages and merge them, and the process restarts afresh with now $k-1$ lineages.

Our previous calculations are the central argument of the following result due to Kingman (1982):

Theorem 1.19. *If we measure time in units of N generations, the genealogy of the Wright-Fisher model converges weakly to the n-coalescent.*

The theorem Kingman proved is actually much stronger. It shows that the coalescent approximation is valid for a large number of models. The idea is to generalise the multinomial sampling in the Wright-Fisher to an *exchangeable* family of random variables $(\nu_1,\ldots,\nu_N)$ satisfying $\nu_1 + \cdots + \nu_N = N$.

Definition 1.20 (Exchangeable random variables)**.** *A family of random variables $(\nu_1, \ldots, \nu_N)$ is said to be* exchangeable *if for every permutation σ of $\{1,\ldots,N\}$,*

$$(\nu_{\sigma(1)},\ldots,\nu_{\sigma(N)}) \overset{\mathcal{D}}{=} (\nu_1,\ldots,\nu_N),$$

where the equality is in distribution.

Definition 1.21 (Cannings model)**.** *Consider a population of fixed size N, and for every $n \geq 0$ a family of N exchangeable random variables $(\nu_1(n), \ldots, \nu_N(n))$ satisfying*

$$\nu_1(n) + \cdots + \nu_N(n) = N.$$

Furthermore, take the random vectors $(\nu_1(n),\ldots,\nu_N(n))$, $n \geq 0$ to be independent. If we label by $\{1,\ldots,N\}$ the individuals of generation n, we can construct generation $n+1$ by defining $\nu_j(n)$ to be the number of children of individual j.

We notice that this definition does not record the exact families, but only their sizes. This is because exchangeability implies that conditionally on $(\nu_1, \dots, \nu_N)$, all the collections of families with sizes $(\nu_1, \dots, \nu_N)$ are equally likely.

Take $n \leq N$ and $\xi, \eta \in \mathcal{R}_n$ such that ξ is a refinement of η. Let $a = |\eta|$ and $b = |\xi|$, and denote by b_j the number of blocks from ξ merged to constitute the j^{th} block of η. For a sample of size n, the transition probabilites $P_{\xi\eta}$ of the Cannings model genealogy are given by

$$P_{\xi\eta} = \frac{(N)_a}{(N)_b} \mathbb{E}\big[(\nu_1)_{b_1} \dots (\nu_a)_{b_a}\big], \tag{1.11}$$

where $(N)_b := N(N-1)\dots(N-b+1)$. This equation is at the origin of the following result.

Theorem 1.22 ((Kingman, 1982)). *Suppose the following conditions are satisfied:*

$$\begin{cases} Var(\nu_1) \xrightarrow[N\to\infty]{} \sigma^2 \\ \\ \sup\limits_{N\geq 1} \mathbb{E}[\nu_1^k] < \infty, \quad k \geq 2. \end{cases}$$

Then the transition probabilites $P_{\xi\eta}$ satisfy the asymptotic relation

$$P_{\xi\eta} = \delta_{\xi\eta} + \frac{\sigma^2}{N} Q_{\xi\eta} + o(\frac{1}{N}), \tag{1.12}$$

where Q is the transition matrix for the Kingman n-coalescent. Moreover, if time is measured in units of N/σ^2 generations, the genealogical process for the Canning's model converges weakly to the n-coalescent.

1.2.3.2 Introducing mutations

The power of the coalescent method is that it allows one to model the genetic composition of a finite sample without worrying about the whole population.

We can now introduce mutations between K alleles on the n-coalescent tree. Because we are measuring time in units of N generations, we can use the same rescaling as for the Wright-Fisher diffusion. Therefore, on the coalescent time scale, mutations happen along each lineage according to a continuous time Markov chain with transitions β_{ij}. Starting from the root of the tree, we run the mutation process forward in time along the branches. In the special case of parent-independent mutation, we can equivalently run the process backward in time, which is much more efficient in statistical applications.

We can model explicitly an infinite number of alleles through what is called the *infinitely many alleles model*.

Definition 1.23. *Mutations follow the infinitely many alleles model if every new mutation has never been observed before. Furthermore, every mutation happens at the same rate $\theta/2$.*

Given a sample of size n, a question of interest is to describe the sample allele frequencies. We denote by α_j the number of alleles that are carried by j individuals in the sample. The following result is fundamental in population genetics:

Theorem 1.24 (Ewens sampling formula)**.** *Under the infinitely many alleles model,*

$$\mathbb{P}\big(\alpha_1 = a_1, \dots, \alpha_n = a_n\big) = \frac{\theta^k}{\theta_{\uparrow n}} \frac{n!}{a_1! \dots a_n!\, 1^{a_1} \dots n^{a_n}}, \tag{1.13}$$

where $k = a_1 + \cdots + a_n$ is the number of alleles in the sample, and $\theta_{\uparrow n} = \theta(\theta+1)\dots(\theta+n-1)$.

To prove this, we first notice that the genetic type of an individual is given by the first mutation we encounter when we follow its lineage backward in time. This is why we can reconstruct the genetic composition of our sample by running a killed Kingman coalescent. When there are k lineages, the coalescence rate is $\binom{k}{2}$, and the

mutation rate is $\theta/2$ times the number of lineages, that is $k\theta/2$. When a coalescence happens, we merge the corresponding lineages, and when a mutation happens, we kill the corresponding lineage. Therefore, while there are $k+1$ lineages, the probability of a coalescence is $k/(\theta+k)$ and the probability of a killing is $\theta/(\theta+k)$. We recognise here the backward in time transition probabilities for a Hoppe urn model.

Definition 1.25 (Hoppe's urn)**.** *Consider an urn containing coloured balls, each of mass* 1*, and a black ball with mass* θ*. At each step, select a ball at random proportionally to its mass. If the ball is couloured, replace this ball in the urn with an additional ball of the same colour. If the ball is black, replace the black ball in the urn and add a ball with a new colour.*

Therefore, we need to prove that the distribution of the colours in Hoppe's urn satisfies (1.24). This is done by induction on n (See Durrett 2004 for more details).

There is a strong connection between Ewens' sampling formula and the Poisson-Dirichlet distribution. The following theorem can be found in Kingman (1977).

Theorem 1.26. *Given* θ*, an allele distribution for a sample of size* n *satisfies the Ewens sampling formula for all* $n \geq 0$ *if and only if it converges weakly to the Poisson-Dirichlet distribution as* n *tends to infinity.*

1.2.4 The Fleming-Viot process

We notice that either rescaling backward in time or forward in time, the infinitely many alleles model is strongly connected to the Poisson-Dirichlet distribution. However, we have obtained the $\mathcal{PD}(\theta)$ distribution only as a weak limit. We would like to be able to model an infinite population with infinitely many alleles by a process that is stationary under the $\mathcal{PD}(\theta)$ distribution. Such a process was initially constructed by Fleming and Viot, and has since been extended to the larger class of the Λ-*Fleming-Viot processes*.

Before presenting the construction of Fleming and Viot, we introduce a much simpler object, the Moran model.

Definition 1.27 (Neutral Moran model)**.** *Consider a population of N individuals. It is said to evolve according to the Moran model if at rate $\binom{N}{2}$, a pair of individuals is sampled, one of them dies, the other survives and produces an identical copy.*

The Moran model is different from the Cannings models because it is directly constructed in continuous time. This model is important because it has two fundamental properties:

- If we introduce K alleles, then the process of the frequencies converges to the Wright-Fisher diffusion as N tends to infinity. In particular, there is no need to rescale the time.
- If we sample n individuals from the population of size N, then their genealogy is exactly a Kingman n-coalescent.

Thanks to these two properties, the Moran model establishes a strong connection between the Wright-Fisher diffusion and the Kingman coalescent. The second property is a *strong duality* between these the Moran model and the Kingman coalescent, namely that we are able to construct both processes simultaneously on the same probability space.

The idea to obtain a model of an infinite population with infinitely many alleles is to extend the Moran model. The problem is that as N tends to infinity, the Moran model does not converge to a stochastic process. One way to overcome this is the lookdown construction, due to Donnelly and Kurtz (1996), and we will discuss it at the end of this section.

The method of Fleming and Viot is to construct an infinite population model via its generator in such a way that the restriction of this generator to a population of size N

coincides with the generator of the Moran model. To allow for infinitely many alleles, the type space is taken to be $[0,1]$. The state space is going to be the set of probability measures on $[0,1]$, so we take a probability measure ρ on $[0,1]$, and we consider test functions G of the form

$$G(\rho) = \int_{[0,1]^p} f(a_1, \ldots, a_p)\, \rho(da_1) \cdots \rho(da_p),$$

where $p \in \mathbb{N}$, and f is a bounded function from $[0,1]^p$ to $\mathbb{R}$.

The generator of the Fleming-Viot process $\mathcal{L}$ is given by

$$(\mathcal{L}G)(\rho) = \sum_{i,j=1}^{p} \frac{1}{2} \int (f_{ij}(a_1, \ldots, a_p) - f(a_1, \ldots, a_p))\, \rho(da_1) \cdots \rho(da_p), \tag{1.14}$$

where $f_{i,j}(a_1, \ldots, a_p) = f(a_1, \ldots, a_i, \ldots, a_i \ldots, a_p)$, that is the value a_i is copied to the j^{th} position. If we take $p = N$, we recover exactly the generator for the Moran model. To introduce mutations in a way consistent with the infinitely many alleles, we proceed exactly as we did for the Kingman coalescent, that is

$$(\mathcal{L}G)(\rho) = \sum_{i,j=1}^{p} \frac{1}{2} \int (f_{ij}(a_1, \ldots, a_p) - f(a_1, \ldots, a_p))\, \rho(da_1) \cdots \rho(da_p) \tag{1.15}$$

$$+ \frac{\theta}{2} \sum_{i=1}^{p} \int (f_i(y; a_1, \ldots, a_p) - f(a_1, \ldots, a_p))\, dy, \tag{1.16}$$

where $f_i(y; a_1, \ldots, a_p) = f(a_1, \ldots, y, \ldots, a_p)$, that is the value y is copied in the i^{th} position.

The proof that the Poisson-Dirichlet distribution is stationary for the Fleming-Viot process relies on calculations with the generator. Details can be found in Ethier and Kurtz (1994).

A powerful tool for manipulating this object is the lookdown construction, due to Donnelly and Kurtz (1996). The principal idea is to relabel the individuals in the Moran model in such a way that models corresponding to different values of N can be constructed on the same probability space. We can then pass to the infinite population limit. To do this we modify the dynamics of the Moran model in such a way that the trajectories of the process with N individuals are exactly the restrictions of the trajectories of the process with $N + 1$ individuals, and such that the dynamics of the allele numbers is the same as in the original Moran model. This is achieved by relabelling the individuals in a convenient way. We give a full description in the next section, where we present this lookdown construction in a more general case.

1.3 Recent Models

The models we presented in the previous section are centered around the Kingman coalescent and the Wright-Fisher diffusion (or the Fleming-Viot process when the number of alleles is infinite). We saw that the Kingman coalescent is the limiting object for a large number of models. This is also true for the Wright-Fisher diffusion. Since 1999, several authors have generalised the methods used so far and obtained other limiting models describing different biological situations. We still have both backward and forward in time approaches, but the coalescent method is the more intuitive one with which to start.

1.3.1 Λ-coalescent processes

Λ-coalescent processes were introduced independently in Donnelly and Kurtz (1999), Pitman (1999) and Sagitov (1999). We are going to follow the presentation of Sagitov (1999), which extends the method used in Kingman (1982) for the convergence to the n-coalescent.

The starting point is the analysis of the transition matrix $P_{\xi\eta}$ for the genealogy in the Canning's model. The objective is to obtain a generalisation of equation (1.12) of the form

$$P_{\xi\eta} = \delta_{\xi\eta} + \frac{1}{T_N} Q_{\xi\eta} + o(\frac{1}{T_N}), \tag{1.17}$$

where Q and T_N are to be determined. The restriction on $Q_{\xi\eta}$ is that it should not allow more than one block from η to be the result of merging at least two blocks from ξ. Using expression (1.11), a first necessary condition for (1.17) to hold is that

$$\frac{\mathrm{Var}(\nu_1)}{N} \xrightarrow[N\to\infty]{} 0, \tag{1.18}$$

which is more general than the one obtained in the Kingman coalescent. In particular, it allows $\mathrm{Var}(\nu_1)$ to tend to infinity. The corresponding time scale is then

$$T_N = \frac{N}{\mathrm{Var}(\nu_1)}.$$

To identify Q, arguments from the theory of infinitely divisible distributions lead to the second necessary condition

$$NT_N\mathbb{P}\big[\nu_1 > Nx\big] \xrightarrow[N\to\infty]{} \int_x^1 \frac{1}{y^2}\,\Lambda(dy), \tag{1.19}$$

where $x \in (0,1)$, and $\Lambda(dy)$ is a probability measure on $[0,1]$.

Finally, the exclusion of simultaneous mergers leads to the third necessary condition:

$$N^{-a}\,T_N\mathbb{E}\big[(\nu_1-1)^2\ldots(\nu_a-1)^2\big] \xrightarrow[N\to\infty]{} 0 \qquad a \geq 2. \tag{1.20}$$

Before stating the result, we need to define a k-merger.

Definition 1.28. *For $\xi, \eta \in \mathcal{R}_n$, we say that (ξ,η) is k-merger if $\xi \subset \eta$, $|\eta| = |\xi| - k + 1$*

and k blocks from $|\xi|$ are merged into one of η .

In particular, simultaneous mergers are ruled out by this definition.

Theorem 1.29. *If conditions (1.18), (1.19) and (1.20) are met, then (1.17) is satisfied, and the transition matrix Q given by*

$$\forall \eta \neq \xi, \qquad Q_{\xi\eta} = \begin{cases} \displaystyle\int_0^1 x^{k-2}(1-x)^{|\xi|-k}\Lambda(dx) & \text{if } (\xi,\eta) \text{ is a } k\text{-merger}, \\ 0 & \text{otherwise.} \end{cases}$$

The main difference between the Λ-coalescent and the Kingman colescent is the possibility for k-mergers with $k \geq 2$. When there are b lineages, the rate at which k specific lineages merge into one is given by

$$\lambda_{b,k} := \int_{[0,1]} x^{k-2}(1-x)^{b-k}\Lambda(dx). \tag{1.21}$$

Sagitov (1999) gives a construction that gives some intuition about the form of the rates $\lambda_{b,k}$. We first recall de Finetti's theorem. An infinite sequence of random variables $(X_n)_{n\geq 1}$ is said to be *exchangeable* if, for every $n \geq 1$ and for every $i_1, \ldots, i_n \in \mathbb{N}$, the random variables $(X_{i_1}, \ldots, X_{i_n})$ are exchangeable.

Theorem 1.30. *(De Finetti's Theorem)*
Any exchangeable sequence of random variables $(I_n)_{n\geq 1}$ taking values 0 or 1 have a joint probability mass function given by

$$\mathbb{P}(I_1 = x_1, \ldots, I_n = x_n) = \int_{[0,1]} u^s(1-u)^{n-s}\Lambda(dx),$$

where $s := x_1 + \cdots + x_n$, and Λ is a probability distribution on $[0,1]$.

We consider now a population of size N. We suppose there are b lineages at time t, and that we have b exchangeable random variables $(I_1, \ldots, I_b)$. The lineage i is marked

with I_i, and at time $t+1$ we merge all the lineages marked with a 1 at time t. We denote by Z_t the number of lineages at time t. With this mechanism $(Z_t)_{t\geq 0}$ is a Markov chain. Its transition probability $p(b,j) := \mathbb{P}(Z_{t+1} = j | Z_t = b)$ is given by

$$\begin{aligned} p_N(b, b-k+1) &= \sum_{x_1,\ldots,x_b} \mathbb{P}(I_1 = x_1, \ldots, I_b = x_b) \mathbb{1}_{x_1+\cdots+x_b=k} \\ &= \binom{b}{k} \mathbb{P}(I_1 = 1, \ldots, I_k = 1, I_{k+1} = 0 \ldots, I_b = 0) \\ &= \binom{b}{k} \int_{[0,1]} x^k (1-x)^{b-k} \Lambda_N(dx), \end{aligned} \tag{1.22}$$

where $k \geq 2$. The second equality comes from the exchangeability, and the last one is the application of de Finetti's theorem. We make explicit the dependence of the distribution Λ_N on the population size N.

To see the Λ-coalescent, we need to make $N \to \infty$, and rescale the time accordingly. The technical condition for the result to hold is that there exists a sequence $B_N \to \infty$ such that

$$B_N \; x^2 \; \Lambda_N(dx) \xrightarrow[N\to\infty]{} \Lambda(dx),$$

where $\Lambda(dx)$ is a probability measure. The weak convergence implies that the probability (1.22) that k lineages merge satisfies

$$B_N p_N(b, b-k+1) \xrightarrow[N\to\infty]{} \binom{b}{k} \int_{[0,1]} x^{k-2} (1-x)^{b-k} \Lambda(dx).$$

We recognize in the right-hand side the definition (1.21) of the transition rate in the Λ-coalescent. The term B_N gives the time scale of the coalescent approximation.

The Kingman coalescent is recovered as a special case of the Λ-coalescent processes by taking $\Lambda = \delta_0$.

1.3.2 Λ-Fleming-Viot process

In §1.2.4, we observed a duality between the Kingman coalescent and the Wright-Fisher diffusion. We have the same situation for the Λ-coalescent models.

A Λ-Fleming-Viot processes (ΛFV) is a Markov process taking values in the set of probability measure on $[0,1]$. Such processes were introduced independently in Donnelly and Kurtz (1999) and Bertoin and Le Gall (2003). We first give the definition of a ΛFVP by providing its generator. Let Λ be a finite measure on $[0,1]$, and define

$$\lambda_{p,j} := \int_{[0,1]} u^j (1-u)^{p-j} \frac{\Lambda(du)}{u^2}. \tag{1.23}$$

Take a probability measure ρ on $[0,1]$, $p \in \mathbb{N}$, and f a bounded function from $[0,1]^p$ to $\mathbb{R}$. We define the function G by

$$G(\rho) = \int_{[0,1]^p} f(a_1, \ldots, a_p)\, \rho(da_1) \cdots \rho(da_p). \tag{1.24}$$

Consider now $J \subset \{1, \ldots, p\}$, and denote by $|J|$ the cardinality of J. For every $(a_1, \ldots, a_p) \in [0,1]^p$, define $(a_1^J, \ldots, a_p^J)$ to be

$$\begin{cases} a_i^J = a_{\min(J)} & \text{if } i \in J, \\ a_i^J = a_i & \text{if } i \notin J. \end{cases} \tag{1.25}$$

The generator $\mathcal{L}$ of a ΛFV process, acting on a function G of the form (1.24), satisfies

$$(\mathcal{L}G)(\rho) = \sum_{J \subset \{1,\ldots,p\}, |j| \geq 2} \lambda_{p,|J|} \int \left(f(a_1^J, \ldots, a_p^J) - f(a_1, \ldots, a_p) \right) \rho(da_1) \cdots \rho(da_p). \tag{1.26}$$

The interpretation of the generator follows the construction from Sagitov (1999) that we described. At rate $\lambda_{p,|J|}$, the set of individuals J is sampled to take part in the

reproduction event. The parent has index $\min(J)$, so it replaces all the other individuals in J. When $\Lambda(\{0\}) = 0$, the generator can be written as

$$(\mathcal{L}G)(\rho) = \int_{(0,1]} \int_{[0,1]^p} \Big[G\Big((1-u)\rho + u\delta_a\Big) - G\big(\rho\big) \Big] \, \rho(da) \frac{\Lambda(du)}{u^2}, \tag{1.27}$$

where δ_a is the unit point mass in a. For this point see Birkner et al. (2005).

From now on, we suppose that the assumption $\Lambda(\{0\}) = 0$ is fulfilled. The case $\Lambda(\{0\}) = 1$ corresponds to the Fleming-Viot process. Allowing for $\Lambda(\{0\}) > 0$ simply introduces an additional component, and we prefer to ignore it to ease the notation.

Thanks to (1.27), the behaviour of the process is now clear. At rate $\rho(da)\Lambda(du)/u^2$, a point a is sampled from $[0,1]$ according to the distribution ρ, and ρ becomes $(1-u)\rho + u\delta_a$. Biologically, it means that an individual of type a reproduces, and by doing so, replaces a proportion u of the whole population by its own offspring. The total rate at which the process leaves state ρ is given by

$$\overline{\Lambda} := \int_{(0,1]} \frac{\Lambda(du)}{u^2}. \tag{1.28}$$

When $\overline{\Lambda}$ is finite, the process is a continuous time Markov chain, with holding times $\mathrm{Exp}(\overline{\Lambda})$. Otherwise, during every finite time interval, an infinite number of jumps take place.

1.3.3 Corresponding lookdown construction

A fundamental tool in studying measure-valued processes is their representation thanks to a system of discrete interacting particles. Donnelly and Kurtz (1996, 1999) provide the method for such constructions in great generality. This method has been applied explicitly to ΛFV processes in Birkner et al. (2005) and Birkner et al. (2009). We briefly explain what this construction is.

Let Π be a Poisson point process on $\mathbb{R}_+ \times [0,1]$, with intensity measure

$$dt \otimes \frac{\Lambda(du)}{u^2}. \tag{1.29}$$

Consider a population of N individuals evolving in continuous time. Each individual i at time t has a type $\widehat{Y}_i^N(t) \in [0,1]$. For every point (t,u) of Π, mark with a cross at time t each individual independently with probability u. Let J_t be the set of all marked individuals. Then, if $|J_t| \geq 2$,

$$\forall i \in J_t, \widehat{Y}_i^N(t) = \widehat{Y}_{min(J_t)}^N(t-). \tag{1.30}$$

This means that all marked individuals *look down* to the individual with the lowest ranking, and copy its type.

To give a full description of the dynamics, we assume that the type of an individual remains constant between two reproduction events, that is

$$\widehat{Y}_i^N(t) \neq \widehat{Y}_i^N(t-) \text{ iff } i = \min(J_t).$$

The fundamental property of this construction is that all processes

$$\left(\widehat{Y}_1^N(t), \ldots, \widehat{Y}_N^N(t)\right)_{(t\geq 0)}$$

for $N = 1, 2, \ldots$ are constructed on the same probability space, using the same realisation of the Poisson process Π. What we obtained is an explicit construction of an $[0,1]^\infty$-valued Markov Process. The connection with the measure-valued process is made when we consider the random measure

$$\widehat{Z}(t) := \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \delta_{\widehat{Y}_i(t)}. \tag{1.31}$$

The limit of the empirical measures relies on the exchangeability of the variables $(\widehat{Y}_1^N(t), \cdots, \widehat{Y}_N^N(t))$. For the details, see Birkner et al. (2009, section 2.2) and Donnelly and Kurtz (1996). The fundamental result is that $(\widehat{Z}(t))_{t\geq 0}$ is a Markov process with generator given by (1.27).

Remark 1.31. *Birkner et al. (2009, Section 2.2) actually treats a more general setting, the Ξ-coalescents, which is a larger family that includes Λ-coalescents. Furthermore, they use the* modified *lookdown construction from Donnelly and Kurtz (1999). However, as explained in Birkner et al. (2009, Remark 2.7), it is perfectly valid to use the classical (and simpler) lookdown construction as we did here.*

In the next Chapter, we develop some original work in which we prove a stationarity result in the infinitely many sites model for certain ΛFV models similar to the one connecting the Poisson-Dirichlet to the Fleming-Viot process.

Chapter 2

Results on some Λ-Fleming-Viot processes

2.1 Introduction

We are interested in this chapter in describing the stationary distribution of the allele frequencies for a class of Λ-Fleming-Viot processes with mutations. Thanks to the duality between a Λ-Fleming-Viot process and a Λ-coalescent, the current literature usually treats this question from the point of view of the Λ-coalescent. For example, Berestycki et al. (2007), Berestycki et al. (2008) and Berestycki et al. (2011a) obtain partial results for stationary distribution of allele frequencies corresponding to the Beta-coalescent processes, for which the Λ measure is taken to be the Beta$(2-\alpha,\alpha)$ probability distribution with $1 < \alpha < 2$.

We are taking a different approach by focussing on the forward-in-time Λ-Fleming-Viot process. We recall that when $\Lambda(\{0\}) = 0$, the total rate at which jumps happen in a ΛFV is given by

$$\overline{\Lambda} = \int_{(0,1]} \frac{\Lambda(du)}{u^2}.$$

We consider the special case where the reproduction events happen at finite rate, although the population is infinite. This is the case if and only if the measure $u^{-2}\Lambda(du)$ is a finite measure on $(0,1]$. In this setting, we study first the dynamics of allele frequencies in the infinitely many alleles model, when mutations occur at a constant rate θ per lineage. We show that there is a *dust formation*, that is the probability measure $Z(t)$ representing the allele frequencies is not a purely atomic measure as was the case in the Kingman coalescent. The reason why alleles have zero frequency is because mutations happen much faster than reproduction events, therefore new mutations have less time to see their frequency increase thanks to reproduction. This dust is an increase in genetic diversity because drift is not strong enough.

We then introduce another mutation mechanism, the *Bernoulli* mutation mechanism, where a mutation arises only during a reproduction event. When an individual reproduces, with a probability p all of her offspring are mutants with the same type, and with a probability $1-p$ they have the same type as their parent. With this modified mechanism, we obtain the following result.

Theorem 2.1. *Suppose $u^{-2}\Lambda(du)$ is a Beta(1,B) distribution for some $B > 0$. Under the Bernoulli mutation mechanism with mutation probability p, the $\mathcal{PD}(pB)$ distribution is stationary for the ranked allele frequencies in the corresponding ΛFV process.*

The interest of this result is that we are able to describe the full distribution of alleles, not only its sampling properties, or the asymptotic behaviour for large samples. Such a result is rare enough to be emphasized. The surprising phenomenon is that the $\mathcal{PD}(pB)$ appears in a model that is completely different from the usual Kingman coalescent.

The rest of the chapter is organised as follows. In §2.2, we study the dynamics of the allele frequencies under the classical Poisson mutation mechanism and demonstrate the creation of dust. In §2.3, we introduce the Bernoulli mutation mechanism and prove

the stationarity of the Poisson-Dirichlet distribution by using the GEM distribution. Finally, we discuss the possible generalisations of this result.

2.2 Classical Poisson mutation mechanism

2.2.1 Infinitely many alleles model

In the previous chapter, we provided the lookdown construction for ΛFV processes without the mutation mechanism. We simultaneously constructed a $[0,1]^\infty$-valued Markov process $(\widehat{Y}_1(t), \widehat{Y}_2(t), \dots)_{t\geq 0}$ and a measure valued process $(\widehat{Z}(t))_{t\geq 0}$. We are going to complete the lookdown construction to introduce mutations according to the standard Poisson mechanism with the infinitely-many alleles model.

We construct now a new system of particles $\big(Y_1(t), Y_2(t), \dots\big)_{t\geq 0}$, and we keep the same reproduction structure as before. Let Π be Poisson point process on $\mathbb{R}_+ \times [0,1]$, with intensity measure

$$dt \otimes \frac{\Lambda(du)}{u^2}. \tag{2.1}$$

Each individual i at time t has a type $Y_i(t) \in [0,1]$. For every point (t,u) of Π, mark with a cross at time t each individual independently with probability u. Let J_t be the set of all marked individuals. Then, if $|J_t| \geq 2$,

$$\forall i \in J_t, Y_i(t) = Y_{min(J_t)}(t-). \tag{2.2}$$

For the mutation mechanism, we introduce a family $(\mathcal{M}_i)_{i\geq 1}$, of independent Poisson point processes on $\mathbb{R}_+ \times [0,1]$ with intensity

$$\theta dt \otimes da, \tag{2.3}$$

where $\theta > 0$ is the mutation rate. We take these Poisson point processes also independent from Π. The mutations are recorded the following way. For every $i = 1, 2, \ldots$, for every $(t, a) \in \mathcal{M}_i$

$$Y_i(t) = a. \tag{2.4}$$

To complete the definition of the process, we need to make precise that

$$\Big(Y_i(t) \neq Y_i(t-)\Big) \text{ iff } \Big(i = \min(J_t) \text{ or } (t, Y_i(t)) \in \mathcal{M}_i\Big).$$

This simply means that the type of an individual remains constant unless modified by a reproduction event or a mutation event.

The name *Poisson mutation mechanism* comes from the Poisson point process $(\mathcal{M}_i)_{i\geq 1}$ used. The infinitely many alleles are introduced via the Lebesgue measure da on the type space. Thanks to the properties of the Poisson point processes, the fact that da has no atoms ensures that all the allelic types are distinct.

As before, by taking the limit for the empirical measure, we obtain a ΛFV process with the infinitely many alleles mutation mechanism:

$$Z(t) := \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \delta_{Y_i(t)}. \tag{2.5}$$

2.2.2 Process of ranked frequencies

We are interested in understanding the distribution of allele frequencies at stationarity. This is why we do not need to keep track of the specific labels of the alleles, but simply the list of frequencies. This means that we do not need to work with the full measure $Z(t)$, but simply the frequencies of its atoms. This is why we introduce the process of the ranked frequencies taking values in the set Ξ of *ranked frequencies* that

we introduced in Chapter 2:

$$\Xi = \big\{x = (x_1, x_2, \ldots) : x_1 \geq x_2 \geq \cdots > 0, \sum_{i=1}^{\infty} x_i \leq 1\big\}. \tag{2.6}$$

Notation 2.2. *Consider a measure ν on $[0,1]$ with decomposition between atomic and diffuse components given by*

$$\nu(d\omega) = \sum_{i=1}^{K} x_i \delta_{\omega_i}(d\omega) + f(\omega)d\omega, \tag{2.7}$$

where $d\omega$ is the Lebesgue measure, f the density of the diffuse part, the ω_i's are the atoms of ν, and the x_i's are the corresponding frequencies. We define the ranking operator $\mathcal{R}$ *acting on probability measures on $[0,1]$ by*

$$\mathcal{R}(\nu) := \big(x_{(1)}, x_{(2)}, \ldots\big) \in \Xi, \tag{2.8}$$

where the $x_{(i)}$'s are the ranked frequencies from (2.7).

Definition 2.3. *The process of ranked frequencies with mutations is defined by*

$$\forall t,\ X(t) := \mathcal{R}(Z(t)). \tag{2.9}$$

2.2.3 Formation of dust

In the absence of mutations, we have seen at the end of §1.3.2 that when the total rate $\overline{\Lambda}$ is finite, $(\widehat{Z}(t))_{t\geq 0}$ is actually a continuous-time Markov chain (CTMC), with jumps

$$\rho \longmapsto (1-u)\rho + u\delta_a$$

happening at rate $\rho(a)\, da\, \Lambda(du)/u^2$.

However, once we introduce mutations as a Poisson process along the lineages, the

total rate at which mutations happen is infinite, and $Z(t)$ and $X(t)$ are not necessarily continuous time Markov chains. We investigate in this section the behaviour of $X(t)$.

To exploit the jump chain structure of the reproduction mechanism, we denote by $T_1, T_2, \dots$ the times of the points of the Poisson point process Π. The random variables $T_1, T_2, \dots$ are i.i.d. exponential random variables with parameter $\overline{\Lambda}$.

The following result is fundamental.

Proposition 2.4. *On the event* $\{T_1 > t\}$,

$$X(t) = e^{-\theta t}\, X(0). \tag{2.10}$$

As a consequence,

$$\sum_{i=1}^{\infty} X_i(t) = e^{-\theta t} \sum_{i=1}^{\infty} X_i(0) < 1 \quad \textit{for } 0 < t < T_1.$$

We call this phenomenon dust formation.

Proof. We use the lookdown construction. Introduce $\mu_t := Z(t)$ and $\mu_t^R := \mathcal{R}(\mu_t)$. For all $a \in [0,1]$, using (2.5), we have the following almost sure convergence:

$$\mu_t(\{a\}) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{Y_i(t) = a\}. \tag{2.11}$$

We want to determine μ_t^R, that is the ordering of the frequencies of the atoms of μ_t. Therefore, we need to determine the set of alleles a such that $\mu_t(\{a\} > 0)$.

In the infinitely many alleles mechanism, every mutation creates a new allele. Consider such an allele a created by mutation between time 0 and t. We are looking at the event $T_1 > t$, which means that the allele a is carried by a single individual in the population. Therefore, (2.11) implies that $\mu_t(\{a\}) = 0$, and the allele a does not

contribute to μ_t^R. All mutations created before T_1 will not contribute to μ_t^R, even if the set of all new alleles constitutes a positive fraction of the population.

For an allele a, define L_t^N to be the set of all individuals with level smaller than N carrying that allele at time t:

$$L_t^N := \{1 \leq i \leq N : Y_i(t) = a\}, \tag{2.12}$$

and rephrasing (2.11), we have the almost sure convergence

$$\mu_t(\{a\}) = \lim_{N\to\infty} \frac{|L_t^N|}{N}. \tag{2.13}$$

Consider now an allele a present in the population at time 0. Because every allele created by mutation is different, every time a mutation hits an individual carrying the allele a, $|L_t^N|$ is decreased by one. Therefore, almost surely, the only individuals carrying allele a at time t are those carrying a at time 0 and not hit by a mutation event, that is

$$|L_t^N| = \sum_{i\in L_0^N} \mathbb{1}\{\mathcal{M}_i\big((0,t]\times[0,1]\big) = 0\}, \tag{2.14}$$

where $\mathcal{M}_i$ is the Poisson point process in (2.3). Define

$$\epsilon_i(t) := \mathbb{1}\{\mathcal{M}_i\big((0,t]\times[0,1]\big) = 0\}. \tag{2.15}$$

By construction of the Poisson point processes, for every $t > 0$, the $\epsilon_i(t)$ for $i = 1, 2, \ldots$ are independent Bernoulli random variables with parameter $e^{-\theta t}$, and they are independent from L_0^N. By application of the strong law of large numbers and using

(2.13) , we have:

$$\frac{|L_t^N|}{N} = \frac{|L_0^N|}{N} \frac{1}{|L_0^N|} \sum_{i \in L_0^N} \epsilon_i(t)$$

$$\xrightarrow[N\to\infty]{a.s.} e^{-\theta t}\, \mu_0(\{a\}). \tag{2.16}$$

The left-hand side limit provides

$$\mu_t(\{a\}) = e^{-\theta t}\, \mu_0(\{a\}),$$

and we conclude the proof by reordering the frequencies of the atoms of μ_t.

□

The interpretation of this result is that mutations erode deterministically at the same constant rate θ all the allele frequencies, without changing the relative frequencies of the atoms of the distribution. A consequence of the presence of mutations is the creation of *dust*, that is new alleles which each have a frequency 0, but a total weight given by $1 - e^{-\theta t}$. This corresponds to the diffuse part of the measure $Z(t)$, which is not present in the ranked frequencies $X(t)$. This is why the sum of the weights is not 1.

2.2.4 Direct construction of the ranked frequencies process

The previous proposition motivates the definition of the following discrete time Markov chain.

Definition 2.5. *Consider the following independent sequences of random variables:*

$$(T_1, T_2, \dots) \text{ i.i.d } Exp(\overline{\Lambda}),$$
$$(U_1, U_2, \dots) \text{ i.i.d } u^{-2}\Lambda(du)/\overline{\Lambda},$$
$$(V_1, V_2, \dots) \text{ i.i.d } U([0,1]).$$

Given $\widehat{M}_0 \in \Xi$, construct a Markov chain $(\widehat{M}_n)_{n\geq 0}$ using the following recurrence. Writing $\widehat{M}_{n-1} = (x_1, x_2, \dots)$, then:

- *if* $e^{-\theta T_n}(x_1 + \cdots + x_{j-1}) < V_n \leq (x_1 + \cdots + x_j)e^{-\theta T_n}$,

$$\widehat{M}_n = \mathcal{O}rd\Big((1-U_n)e^{-\theta T_n}x_j + U_n, \\ (1-U_n)e^{-\theta T_n}x_1, \dots, (1-U_n)e^{-\theta T_n}x_{j-1}, (1-U_n)e^{-\theta T_n}x_{j+1}, \dots\Big),$$

- *and if* $e^{-\theta T_n}\sum_{i=1}^{\infty} x_i < V_n$,

$$\widehat{M}_n = \mathcal{O}rd\big(U_n, (1-U_n)e^{-\theta T_n}x_1, (1-U_n)e^{-\theta T_n}x_2, \dots\big).$$

The random variable V_n allows us to sample either a non-mutant alleleic type if V_n is smaller than $\sum x_i \exp(-\theta T_n)$, the total mass of non-mutants at time T_n, or one of the mutant alleleic types if V_n is larger than $\sum x_i \exp(-\theta T_n)$.

Definition 2.6. *Consider the random variables $(T_1, T_2, \dots)$ from Definition (2.5), the corresponding realisation of the Markov chain $\widehat{M}_n$, and define the times $(S_n)_{n\geq 0}$ by $S_0 := 0$ and for $n \geq 1$, $S_n := \sum_{i=1}^{n} T_i$. We define the process $(H(t))_{t\geq 0}$ in the following way:*

$$\forall n \geq 0,\ \forall t \in [\ S_{n-1}, S_n), \quad H(T_n) = e^{-\theta(t-S_{n-1})}\widehat{M}_{n-1}.$$

When $t \in [\ S_{n-1}, S_n)$, there is no reproduction event, and Proposition 2.4 estab-

lished the deterministic erosion of all frequencies at rate θ. This accounts for the term $e^{-\theta(t-S_{n-1})}$. At time S_n, an individual is sampled from the population to be the parent. With probability $1 - e^{-\theta S_n} \sum_{i=1}^{\infty} x_i$, the parent is a mutant, and its frequency is increased to U_n. Otherwise, the parent is sampled from the existing frequencies. And this is the description of the Markov chain $\widehat{M}$.

Using Proposition 2.4, we have just proved the following.

Theorem 2.7. *We have the equality in distribution*

$$\big(X(t)\big)_{t\geq 0} \stackrel{\mathcal{D}}{=} \big(H(t)\big)_{t\geq 0}. \tag{2.17}$$

A consequence of this representation is that if $\mathbb{P}(U_1 = 1) = 0$, then for every $t > 0$, $\sum_{i=1}^{\infty} X_i(t) < 1$, which means that the dust is permanent.

This representation also shows that the study of the stationarity is not easy in this situation. A particular difficulty is the fact that the probability that the individual sampled to be the parent is a mutant depends on the amount of time between two successive reproduction events.

This provides a mathematical motivation for taking a different mutation mechanism that avoids this dust creation. The idea is to replace the random mutation probability $1 - e^{-\theta S_n} \sum_{i=1}^{\infty} x_i$ by a constant probability p.

2.3 Bernoulli mutation mechanism

2.3.1 The model

We recall that we introduced in Chapter 1 the set of unordered frequencies Δ defined by

$$\Delta = \big\{x = (x_1, x_2, \ldots) : x_1, x_2, \cdots \geq 0, \sum_{i=1}^{\infty} x_i \leq 1\big\}. \tag{2.18}$$

Definition 2.8. *Let $0 \leq p \leq 1$, and Φ a probability measure on $[0,1]$. We define a Δ-valued Markov chain $(M_n)_{n\geq 0}$ the following way. Sample U independently of $\sigma(M_0, \ldots, M_n)$ according to Φ, and, conditionally on $M_n = (x_1, x_2, \ldots)$:*

with probability $(1-p)\, x_j$,

$$M_{n+1} : = \Big((1-U)x_j + U, (1-U)x_1, \ldots, (1-U)x_{j-1}, (1-U)x_{j+1}, \ldots\Big) \tag{2.19}$$

and with probability $p + (1-p)\Big(1 - \sum_{j=1}^{\infty} x_j\Big)$,

$$M_{n+1} : = \Big(U, (1-U)x_1, (1-U)x_2, \ldots\Big) \tag{2.20}$$

The interpretation is the following. Conditionally on $M_n = (x_1, x_2, \ldots)$, we first consider a fraction U of the population that is going to be modified. U is independent from M_n and is distributed according to the probability measure $\Phi(du)$. With probability $1-p$ there is no mutation and we sample a parent from the population. If the allele of the sampled individual has frequency x_j, it is sampled with probability x_j, and its frequency is increased to $(1-U)\, x_j + U$, whereas all the other alleles have their frequency reduced to $(1-U)x_i$, $i \neq j$. We then reorder our frequencies by moving the j^{th} frequency in first position. In the case where $\sum_{j=1}^{\infty} x_j < 1$, the allele of the sampled individual may have frequency 0, and this happens with probability $1 - \sum_{j=1}^{\infty} x_j$. The allele frequency of this new allele becomes U, and is inserted in the first position of the vector M_{n+1}.

With probability p, there is a mutation, and a new allele is introduced with frequency U. We bring the newly created frequency in the first position, without changing the ordering of the remaining frequencies.

In the previous section we worked with ranked frequencies, but now we decided to

not rank the frequencies. First we note that it does not change anything: what counts is the set of frequencies, not the order we use to label them.

The reason why we do not wish to rank the frequencies is because of the way we are going to prove our stationarity result. It appears that in our case, the easiest way to manipulate the Poisson-Dirichlet distribution is through its size-biased reordering, the GEM distribution, which is a measure on Δ.

2.3.2 Main Result

Theorem 2.1 was expressed in terms of a ΛFV process, which is a continuous time measure-valued process. However, because the jump rates of this process are all equal to $\overline{\Lambda}$, and because $M_{n+1} \neq M_n$ for all $n \geq 0$ almost surely, Theorem 2.1 can be expressed equivalently in terms of the embedded Markov chain $(M_n)_{n\geq 0}$.

Theorem 2.9. *If $\Phi(du)$ is a Beta$(1, B)$ distribution for $B > 0$, then the GEM(pB) distribution is stationary for the Markov Chain $(M_n)_{n\geq 0}$.*

Considering the dynamics of the model, it is not immediate what one should expect for the stationary distribution. The first approach to study of this model consisted in looking at the ranked frequencies. The first step was to calculate, for the case $B = 1$, the frequency spectrum, which is the density of frequencies if they are represented as a point process on $[0, 1]$. The discrete time structure allowed us to establish a recurrence equation for the frequency spectrum, whose solution corresponded with the Poisson Dirichlet distribution. It was conjectured from this that at least when $B = 1$, the stationary distribution should be Poisson-Dirichlet.

The major difficulty in proving this conjecture is handling the order of the frequencies. The method developed in Bertoin (2008) seems very promising because it does not depend on any ranking of the frequencies. It relies on the description of the Poisson-Dirichlet distribution thanks to the Ewens sampling formula. This method

works nicely for the problem studied in Bertoin (2008) because the number of transitions in the corresponding Markov chain is very small. However, in our case, seems to be analytically intractable.

An alternative approach is through the GEM representation of the Poisson-Dirichlet. The difficulty is to find a suitable ordering under which the frequencies follow the GEM distribution. It turns out that right ordering is the very natural one in which the moist recently modified frequency is brought to the first position while keeping the order of the remaining frequencies unchanged. As we will see in the next sections, this fact is fundamental.

Our objective is to understand how the GEM appears naturally in the Markov chain of the ΛFV we are considering. To achieve this we need to solve three main problems. The first is to handle the ordering in an easy way. This is achieved by finding another representation of M_1 thanks to an operator $\mathcal{T}_p$ introduced in Definition 2.14. The second problem is to connect the GEM distribution with the operator $\mathcal{T}_1$, that is in absence of mutations. We achieve this by reinterpreting the dynamics of $\mathcal{T}_1$ in terms of the stick-breaking construction of the GEM. Finally, we need to incorporate mutations, and for this we use a new representation of the GEM in terms of a Poisson process.

The rest of this section is organised as follows. In §2.3.3, we provide some results about the GEM distribution. In §2.3.4, we prove the representation of M_1 in terms of $\mathcal{T}_p(M_0)$. Finally, we show in §2.3.5 that the GEM(pB) is stationary for $\mathcal{T}_p$.

2.3.3 Poisson construction of the GEM distribution

The following result allows us to manipulate easily the GEM distribution. It can be found in Arratia et al. (2003), and is originally due to Ignatov (1982).

Proposition 2.10. *Let $0 \leq E_1 \leq E_2 \leq \dots$ be the points of a Poisson Point Process with constant intensity $\theta > 0$. Define the points $0 \leq Y_1 \leq Y_2 \leq \dots$ by the transforma-*

tion

$$Y_n = 1 - e^{-E_n}. \tag{2.21}$$

Then the infinite vector $(Y_1, Y_2 - Y_1, Y_3 - Y_2, \dots)$ *has distribution GEM(*θ*).*

Proof. By definition of the GEM(θ) distribution, we need to prove that the random variables $Z_1 := Y_1$ and $Z_n := (Y_n - Y_{n-1})/(1 - Y_{n-1})$, $n \geq 2$ are i.i.d. $Beta(1, \theta)$.

$$\begin{aligned}
&\mathbb{P}(Z_1 < x_1, Z_2 < x_2, \dots, Z_n < x_n) \\
&= \mathbb{P}\Big(Y_1 < x_1, \frac{Y_2 - Y_1}{1 - Y_1} < x_2, \dots, \frac{Y_n - Y_{n-1}}{1 - Y_{n-1}} < x_n\Big) \\
&= \mathbb{E}\Big[\mathbb{1}_{Y_1 < x_1} \mathbb{1}_{\frac{Y_2 - Y_1}{1 - Y_1} < x_2} \dots \mathbb{1}_{\frac{Y_{n-1} - Y_{n-2}}{1 - Y_{n-2}} < x_{n-1}} \mathbb{P}\Big(\frac{Y_n - Y_{n-1}}{1 - Y_{n-1}} < x_n \,|\, Y_1, \dots Y_{n-1}\Big)\Big].
\end{aligned} \tag{2.22}$$

We define the function Ψ by $\Psi(x) := -\log(1 - x)$. With this notation, the relation (2.21) can be written as $E_n = \Psi(Y_n)$. We can now calculate explicitly the last term in (2.22):

$$\begin{aligned}
&\mathbb{P}\Big(Y_n < Y_{n-1} + (1 - Y_{n-1})x_n \,|\, Y_1, \dots, Y_{n-1}\Big) \\
&= \mathbb{P}\Big(Y_n < Y_{n-1} + (1 - Y_{n-1})x_n \,|\, Y_{n-1}\Big) \\
&= \mathbb{P}\Big(\Psi(Y_n) - \Psi(Y_{n-1}) < \Psi\Big(Y_{n-1} + (1 - Y_{n-1})x_n\Big) - \Psi(Y_{n-1}) \,|\, Y_{n-1}\Big) \\
&= \mathbb{P}\Big(E_n - E_{n-1} < \Psi\Big(Y_{n-1} + (1 - Y_{n-1})x_n\Big) - \Psi(Y_{n-1}) \,|\, Y_{n-1}, E_{n-1}\Big).
\end{aligned}$$

Thanks to the properties of the Poisson Point Process with constant intensity θ, $E_n - E_{n-1}$ is an exponential random variable with parameter θ, independent from $E_{n-1} =$

$\Psi(Y_{n-1})$, therefore

$$\begin{aligned}
&\mathbb{P}\Big(Y_n < Y_{n-1} + (1 - Y_{n-1})x_n \,|\, Y_1, \dots, Y_{n-1}\Big) \\
&= 1 - \exp\bigg(- \theta \Big(\Psi\Big(Y_{n-1} + (1 - Y_{n-1})x_n\Big) - \Psi(Y_{n-1})\Big)\bigg) \\
&= 1 - \exp\bigg(\theta \log\Big(\frac{1 - Y_{n-1} - (1 - Y_{n-1})x_n}{1 - Y_{n-1}}\Big)\bigg) \\
&= 1 - (1 - x_n)^{\theta}. \qquad (2.23)
\end{aligned}$$

When we use (2.23) into (2.22), we obtain the following recurrence relation:

$$\begin{aligned}
&\mathbb{P}(Z_1 < x_1, Z_2 < x_2, \dots, Z_n < x_n) \\
&= \Big(1 - (1 - x_n)^{\theta}\Big)\, \mathbb{P}(Z_1 < x_1, Z_2 < x_2, \dots, Z_{n-1} < x_{n-1}).
\end{aligned}$$

This recurrence is solved immediately, and we obtain

$$\begin{aligned}
&\mathbb{P}(Z_1 < x_1, Z_2 < x_2, \dots, Z_n < x_n) \\
&= \prod_{j=1}^{n} \Big(1 - (1 - x_j)^{\theta}\Big),
\end{aligned}$$

which shows that the Z_i, $i \geq 1$ are indeed i.i.d $Beta(1, \theta)$ random variables. □

Definition 2.11. *Consider a random vector $X = (X_1, X_2, \dots)$, and take $0 < p \leq 1$. We say that a random vector $Z = (Z_1, Z_2, \dots)$ is a* p-grouping *of X if there exist $I_1, I_2, \dots$ i.i.d. Bernoulli random variables with parameter p, independent of X, such*

that

$$\begin{cases} Z_1 = \sum_{j=1}^{S_1} X_j \\ Z_n = \sum_{j=S_{n-1}+1}^{S_n} X_j,\ n \geq 2. \end{cases}$$

where the sequence $(S_n)_{n\geq 1}$ *is defined by:*

$$\begin{cases} S_1 = \inf\{j \geq 1 : I_j = 1\} \\ S_n = \inf\{j > S_{n-1} : I_j = 1\},\ n \geq 2. \end{cases} \tag{2.24}$$

Every frequency of X is marked independently with a 1 with probability p, and we merge contiguous frequencies by adding them.

Corollary 2.12. *Let* $X = (X_1, X_2, \dots)$ *be a GEM(B) distribution, and take* $0 < p \leq 1$. *Then every p-grouping* Z *of* X *follows a GEM(pB) distribution.*

Proof. We first introduce $Y_n := X_1 + \cdots + X_n$, and we consider the sequence of points $(E_n)_{n\geq 1}$ defined by $E_n = -\log(1 - Y_n)$. Thanks to Proposition (2.10), the points $(E_n)_{n\geq 1}$ are the points of a Poisson process with rate B. Consider the sequence $(S_n)_{n\geq 1}$ appearing in construction (2.24) of the p-grouping Z. Define the points $(E'_n)_{n\geq 1}$ by setting $E'_n = E_{S_n}$. By construction, this new set of points $(E'_n)_{n\geq 1}$ is a thinning of the Poisson point process $(E_n)_{n\geq 1}$ with thinning probability p. This implies that $(E'_n)_{n\geq 1}$ are the points of a Poisson point process with intensity pB. By construction, we have $Z_1 = 1 - \exp(-E'_1)$, and for $n \geq 2$, $Z_n = (1 - \exp(-E'_n)) - (1 - \exp(-E'_{n-1}))$. By applying Proposition 2.10 to the points $(E'_n)_{n\geq 1}$, we conclude that Z follows a GEM(pB) distribution. □

Corollary 2.13. *For every GEM(B) random vector* $Z = (Z_1, Z_2, \dots)$, *and for every*

$0 < p \leq 1$, *there exists a GEM(B/p) random vector* X *such that* Z *is a p-grouping of* X.

Proof. Introduce $Y_n := Z_1 + \cdots + Z_n$, and consider the Poisson point process with rate B defined by $E'_n = -\log(1-Y_n)$. Take an independent Poisson point process with rate $(1-p)B/p$, with points $E''_1, E''_2, \ldots$. Using the properties of Poisson point processes, the set of points $\{E'_1, E'_2, \ldots\} \cup \{E''_1, E''_2, \ldots\}$ forms a Poisson point process with points $E_1 < E_2 < \ldots$ with rate B/p, and the process $(E'_n)_{n\geq 1}$ is a thinning of $(E_n)_{n\geq 1}$ with thinning probability p. Proposition 2.10 allows one to construct the GEM(B/p) vector X defined by $X_1 = 1 - \exp(-E_1)$ and for $n \geq 2$, $X_n = (1 - \exp(-E_n)) - (1 - \exp(-E_{n-1}))$, and the conclusion follows. □

2.3.4 Another representation of the sampling mechanism

A key step in understanding how the GEM distribution appears in the Markov chain M_n is to handle the ordering. We prove in Proposition 2.16 that if M_0 is sampled from a GEM distribution, then M_1 is equal in distribution to the simpler object $\mathcal{T}_p(M_0)$, where the operator $\mathcal{T}_p$ is defined below. This representation of M_1 is exploited in the next section to prove Theorem 2.9.

Definition 2.14. *Take* $0 < p \leq 1$ *and* Φ *a probability distribution on* $[0,1]$. *Define a random operator* $\mathcal{T}_p$ *that transforms a vector* $(x_1, x_2, \ldots)$ *in the following way. First we sample* U *independently of* $\sigma(M_0, \ldots, M_n)$ *according to* Φ, *then:*

with probability $(1-p)$,

$$\mathcal{T}_p(x_1, x_2, \ldots) = \big(U + (1-U)x_1, (1-U)x_2, (1-U)x_3, \ldots\big)$$

and with probability p,

$$\mathcal{T}_p(x_1, x_2, \ldots) = \big(U, (1-U)x_1, (1-U)x_2, \ldots\big)$$

The difference between this operator and the transitions for our Markov chain $(M_n)_{n\geq 0}$ introduced in Definition 2.8 is in the sampling mechanism. For the operator $\mathcal{T}_p$, we always sample the frequency x_1, whereas for $(M_n)_{n\geq 0}$ we perform a size-biased sampling, in which each frequency x_j has a probability x_j of being sampled. Before stating and proving Proposition 2.16, we need a key fact about size-biased reordering.

Lemma 2.15. *Take $x = (x_1, x_2, \dots) \in \Delta$, with $\overline{x} = \sum_{j=1}^{\infty} x_j$ and let $\pi(x)$ be a size-biased reordering of x given by*

$$\pi(x) := (x_{\pi(1)}, x_{\pi(2)}, \dots),$$

where

$$\mathbb{P}\big[\pi(1) = i_1, \dots, \pi(n) = i_n\big] = \frac{x_{i_1}}{\overline{x}} \frac{x_{i_2}}{\overline{x} - x_{i_1}} \cdots \frac{x_{i_n}}{\overline{x} - x_{i_1} - \cdots - x_{i_{n-1}}}. \tag{2.25}$$

Let J be a random variable with conditional distribution with respect to π given by

$$\mathbb{P}\big[J = j \,|\, \pi\big] = \frac{x_{\pi(j)}}{\overline{x}}.$$

The vector $\nu(x)$ defined by

$$\nu(x) := \begin{cases} (x_{\pi(1)}, x_{\pi(2)}, \dots) & \text{if } J = 1, \\ (x_{\pi(J)}, x_{\pi(1)}, \dots, x_{\pi(J-1)}, x_{\pi(J+1)}, \dots) & \text{if } J \geq 2, \end{cases} \tag{2.26}$$

is also a size-biased reordering of x.

The interpretation of the result is the following. Given frequencies $x_1, x_2, \dots$, we first run a size-biased reordering, that is we sample each frequency once, and we rank the frequencies according to their order of appearance. This is the resulting permutation π. We then size-biased sample a single frequency, and its position along π is recorded

by J. We reorder π by bringing the selected frequency $x_{\pi(J)}$ in the first position.

Proof. We introduce the notation $\nu(x) = (x_{\nu(1)}, x_{\nu(2)}, \dots)$. We need to show that

$$\mathbb{P}\big[\nu(1) = i_1, \dots, \nu(n) = i_n\big] = \frac{x_{i_1}}{\overline{x}} \frac{x_{i_2}}{\overline{x} - x_{i_1}} \cdots \frac{x_{i_n}}{\overline{x} - x_{i_1} - \cdots - x_{i_{n-1}}}.$$

In the case $n = 1$, we have

$$\begin{aligned}
\mathbb{P}\big(\nu(1) = i_1\big) &= \mathbb{E}\Big[\mathbb{P}\Big(\nu(1) = i_1 \,|\, \pi\Big)\Big] \\
&= \mathbb{E}\Big[\sum_{j=1}^{\infty} \mathbb{P}\Big(\pi(J) = i_1 \,|\, \pi, J = j\Big)\, \mathbb{P}\Big(J = j \,|\, \pi\Big)\Big] \\
&= \mathbb{E}\Big[\sum_{j=1}^{\infty} \mathbb{1}_{\pi(j)=i_1} \frac{x_{\pi(j)}}{\overline{x}}\Big] \\
&= \frac{x_{i_1}}{\overline{x}} \sum_{j=1}^{\infty} \mathbb{P}\Big(\pi(j) = i_1\Big).
\end{aligned}$$

If we denote by π^{-1} the inverse of the permutation π, then $\pi^{-1}(i_1)$ denotes the time at which the frequency x_{i_1} is sampled during the construction of the size-biased reordering. The last sum of probabilities can be written as

$$\sum_{j=1}^{\infty} \mathbb{P}\Big(\pi(j) = i_1\Big) = \mathbb{P}\Big(\bigcup_{j=1}^{\infty} \Big\{\pi^{-1}(i_1) = j\Big\}\Big),$$

and the event $\bigcup_{j=1}^{\infty} \{\pi^{-1}(i_1) = j\}$ has probability one, because it is simply the disjoint union of all the values taken by the random variable $\pi^{-1}(i_1)$. This proves the formula for case $n = 1$.

In the general case, we proceed similarly by conditioning on π and then conditioning

on the values taken by J.

$$\begin{aligned}
&\mathbb{P}\Big(\nu(1)=i_1,\ldots,\nu(n)=i_n \,|\, \pi\Big) \\
&= \mathbb{P}\Big(\pi(1)=i_1,\ldots,\pi(n)=i_n \,|\, \pi\Big)\, \mathbb{P}\Big(J=1 \,|\, \pi\Big) \\
&+ \sum_{j=2}^{n-1} \mathbb{P}\Big(\pi(j)=i_1,\pi(1)=i_2,\ldots,\pi(j-1)=i_j, \\
&\qquad\qquad \pi(j+1)=i_{j+1},\ldots,\pi(n)=i_n \,|\, \pi\Big)\, \mathbb{P}\Big(J=j \,|\, \pi\Big) \\
&+ \sum_{j=n}^{\infty} \mathbb{P}\Big(\pi(j)=i_1,\pi(1)=i_2,\ldots,\pi(n-1)=i_n \,|\, \pi\Big)\, \mathbb{P}\Big(J=j \,|\, \pi\Big).
\end{aligned}$$

After taking the expectation with respect to π, we have

$$\begin{aligned}
&\mathbb{P}\big[\nu(1)=i_1,\ldots,\nu(n)=i_n\big] \\
&= \frac{x_{i_1}}{\overline{x}}\, \mathbb{P}\Big(\pi(1)=i_1,\ldots,\pi(n)=i_n\Big) \\
&+ \frac{x_{i_1}}{\overline{x}} \sum_{j=2}^{n-1} \mathbb{P}\Big(\pi(j)=i_1,\pi(1)=i_2,\ldots,\pi(j-1)=i_j, \\
&\qquad\qquad \pi(j+1)=i_{j+1},\ldots,\pi(n)=i_n\Big) \\
&+ \frac{x_{i_1}}{\overline{x}} \sum_{j=n}^{\infty} \mathbb{P}\Big(\pi(j)=i_1,\pi(1)=i_2,\ldots,\pi(n-1)=i_n\Big) \\
&= \frac{x_{i_1}}{\overline{x}}\, \mathbb{P}\Big(1 \le \pi^{-1}(i_2) \le \cdots \le \pi^{-1}(i_n) \le n\Big).
\end{aligned}$$

The last equality comes from the fact that the sums of probabilities correspond to a partition of the event

$$A_1 := \left\{1 \le \pi^{-1}(i_2) \le \cdots \le \pi^{-1}(i_n) \le n\right\},$$

which is the event that frequencies $i_2, \dots, i_n$ come in this specific order, with possibly the frequency i_1 inserted before, in, or after the block $i_2, \dots, i_n$. To calculate the probability of A_1, we construct another vector of frequencies y obtained by removing the frequency i_1

$$y = (x_1, \dots, x_{i_1-1}, 0, x_{i_1+1}, \dots),$$

and we consider a size-biased permutation π' of y defined by

$$\mathbb{P}\big[\pi'(1) = j_1, \dots, \pi'(n) = j_n\big] = \frac{y_{j_1}}{\overline{y}} \frac{y_{j_2}}{\overline{y} - y_{j_1}} \cdots \frac{y_{j_n}}{\overline{y} - y_{j_1} - \cdots - y_{j_{n-1}}},$$

where $\overline{y} = \overline{x} - x_{i_1}$ is the sum of the new vector y. Because x and y differ at only one frequency, we can easily construct π and π' on the same probability space such that we have the following equality between events:

$$\Big\{1 \le \pi^{-1}(i_2) \le \cdots \le \pi^{-1}(i_n) \le n\Big\} = \Big\{\pi'(1) = i_2, \dots, \pi'(n-1) = i_n\Big\}.$$

Therefore, using the definition of π', we have

$$\begin{aligned} &\mathbb{P}\Big(1 \le \pi^{-1}(i_2) \le \cdots \le \pi^{-1}(i_n) \le n\Big) \\ &= \frac{x_{i_2}}{\overline{y}} \frac{x_{i_3}}{\overline{y} - x_{i_2}} \cdots \frac{x_{i_n}}{\overline{y} - x_{i_2} - \cdots - x_{i_{n-1}}}, \end{aligned}$$

and because $\overline{y} = \overline{x} - x_{i_1}$, we obtain

$$\mathbb{P}\big[\nu(1) = i_1, \dots, \nu(n) = i_n\big] = \frac{x_{i_1}}{\overline{x}} \frac{x_{i_2}}{\overline{x} - x_{i_1}} \cdots \frac{x_{i_n}}{\overline{x} - x_{i_1} - \cdots - x_{i_{n-1}}},$$

which concludes this proof. □

Here is the main result for this section.

Proposition 2.16. *Let $\Phi(du)$ be a probability distribution on $[0,1]$, and consider the*

corresponding Markov Chain $(M_n)_{n\geq 0}$ *and operator* $\mathcal{T}_p$*. We have*

$$\big(M_0 \sim GEM(\theta)\big) \text{ implies } \big(M_1 \overset{\mathcal{D}}{=} \mathcal{T}_p(M_0)\big),$$

where $\theta > 0$*, the symbol* $\sim$ *means "is sampled from", and the equality is in distribution.*

Proof. Consider $X = (X_1, X_2, \dots)$ a GEM(θ) random vector and $U > 0$. We are going to construct a coupling of the two random vectors M_1 and $\mathcal{T}_p(X)$ and show that they have the same distribution. By definition of the GEM distribution, there exists a $\mathcal{PD}(\theta)$ random vector Z such that X is a size-biased reordering of Z. Using the same notation as in Lemma 2.15, there exists a permutation π satisfying (2.25) such that

$$X = \pi(Z).$$

Let W be a random variable which is uniform on $[0, 1]$, independent of X and U. Let J a random variable independent from W and U such that

$$\mathbb{P}\big[J = j \,|\, X\big] = X_j.$$

We construct M_1 and $\mathcal{T}_p(X)$ the following way:

$$M_1 = \begin{cases} \Big(U,\, (1-U)X_1,\, (1-U)X_2, \dots\Big) & \text{if } W < p, \\ \Big(U + (1-U)X_J,\, (1-U)X_1, \dots, & \\ \qquad\qquad (1-U)X_{J-1},\, (1-U)X_{J+1}, \dots\Big) & \text{if } W \geq p, \end{cases}$$

$$\mathcal{T}_p(X) = \begin{cases} \Big(U, (1-U)X_1, (1-U)X_2, \dots\Big) & \text{if } W < p, \\ \Big(U + (1-U)X_1, (1-U)X_2, \dots\Big) & \text{if } W \geq p. \end{cases}$$

Consider a measurable subset $A \subset \Delta$. We want to show that

$$\mathbb{P}\big(M_1 \in A\big) = \mathbb{P}\big(\mathcal{T}_p(X) \in A\big).$$

We have

$$\begin{cases} \mathbb{P}\big(M_1 \in A\big) = \mathbb{P}\big(M_1 \in A \,|\, W < p\big)\mathbb{P}\big(W < p\big) + \mathbb{P}\big(M_1 \in A \,|\, W \geq p\big)\mathbb{P}\big(W \geq p\big), \\ \mathbb{P}\big(\mathcal{T}_p(X) \in A\big) = \mathbb{P}\big(\mathcal{T}_p(X) \in A \,|\, W < p\big)\mathbb{P}\big(W < p\big) \\ \qquad\qquad + \mathbb{P}\big(\mathcal{T}_p(X) \in A \,|\, W \geq p\big)\mathbb{P}\big(W \geq p\big). \end{cases}$$

By construction, it is clear that

$$\mathbb{P}\big(M_1 \in A \,|\, W < p\big) = \mathbb{P}\big(\mathcal{T}_p(X) \in A \,|\, W < p\big),$$

so we only have to show that

$$\mathbb{P}\big(M_1 \in A \,|\, W \geq p\big) = \mathbb{P}\big(\mathcal{T}_p(X) \in A \,|\, W \geq p\big).$$

Because W is independent from X, U and J, this is equivalent to saying that the two vectors

$$Y^1 = \Big(U + (1-U)X_J,\, (1-U)X_1, \ldots, (1-U)X_{J-1},\, (1-U)X_{J+1}, \ldots\Big)$$

and

$$Y^2 = \Big(U + (1-U)X_1, (1-U)X_2, \ldots\Big)$$

have the same distribution. We notice that

$$\begin{cases} Y^1 = \mathcal{T}_0\big(\nu(Z)\big) \\ Y^2 = \mathcal{T}_0\big(X\big), \end{cases}$$

where ν is defined as in (2.26). Because X is a size-biased reordering of Z, we can apply Lemma 2.15, and we obtain that $\nu(Z)$ is also a size-biased reordering of Z. Therefore X and $\nu(Z)$ have the same distribution, and we have the equality in distribution

$$\mathcal{T}_0\big(\nu(Z)\big) \overset{\mathcal{D}}{=} \mathcal{T}_0\big(X\big),$$

which concludes the proof. □

2.3.5 Connection between the process and the GEM

The results of the previous section will allow us to work with the simpler ordering provided by the operator $\mathcal{T}_p$. Therefore we are going to prove the stationarity result for $\mathcal{T}_p$. We treat first the case $p = 1$.

Proposition 2.17. *Suppose $\Phi(du)$ is the Beta$(1, B)$ distribution and $p = 1$. Then*

$$\big(M_0 \sim GEM(B)\big) \Longrightarrow \big(\mathcal{T}_1(M_0) \sim GEM(B)\big).$$

Proof. We are going to use the stick-breaking representation of the GEM distribution given in Proposition 1.10. If we have $M_0 = (X_1, X_2, \dots)$, then there exist i.i.d. Beta$(1, \theta)$ random variables $(Z_1, Z_2, \dots)$ such that:

$$\begin{cases} X_1 = Z_1 \\ X_n = (1 - Z_1)(1 - Z_2) \dots (1 - Z_{n-1})Z_n, \ n \geq 2. \end{cases} \tag{2.27}$$

If we have $\mathcal{T}_1(M_0) = (X'_1, X'_2, \dots)$, then by Definition 2.14 of $\mathcal{T}_1$, there exists a Beta$(1, B)$ random variable U independent of the Z_i's such that

$$\begin{cases} X'_1 = U \\ X'_n = (1-U)X_{n-1}, \ n \geq 2, \end{cases}$$

that is, using (2.27),

$$\begin{cases} X'_1 = Z'_1 \\ X'_n = (1-Z'_1)(1-Z'_2)\dots(1-Z'_{n-1})Z'_n, \ n \geq 2, \end{cases}$$

with $Z'_1 = U$ and $Z'_n = Z_{n-1}$ for $n \geq 2$. The vector $\mathcal{T}_1(M_0) = (X'_1, X'_2, \dots)$ can be written as in (2.27), therefore $\mathcal{T}_1(M_0)$ follows a GEM(B) distribution. □

We are going to extend this result for $0 < p \leq 1$ by using the fact that the p-grouping of a GEM(B) distribution is a GEM(pB) distribution.

Proposition 2.18. *Suppose $\Phi(du)$ is the Beta$(1, B)$ distribution and $0 < p \leq 1$. If M_0 is GEM(pB), then*

$$\big(M_0 \sim GEM(pB)\big) \text{ implies } \big(\mathcal{T}_p(M_0) \sim GEM(pB)\big).$$

Proof. Let $M_0 = (Y_1, Y_2, \dots)$ be a random vector following a GEM(pB) distribution. Using Corollary 2.13, there exists a random vector $(X_1, X_2, \dots)$ following a GEM(B) distribution and $I_1, I_2, \dots$ i.i.d. Bernoulli random variables with parameter p such that

$$\begin{cases} Y_1 = \displaystyle\sum_{j=1}^{S_1} X_j \\ Y_n = \displaystyle\sum_{j=S_{n-1}+1}^{S_n} X_j, \ n \geq 2, \end{cases}$$

where the S_n's are defined by

$$\begin{cases} S_1 = \inf\{j \geq 1 : I_j = 1\} \\ \\ S_n = \inf\{j > S_{n-1} : I_j = 1\},\ n \geq 2. \end{cases}$$

Consider a Beta$(1, B)$ random variable U independent from the X_n's and the I_n's. By construction, the vector

$$\widetilde{X} = \big(U, (1-U)X_1, (1-U)X_2, \dots\big)$$

is a realisation of $\mathcal{T}_1(X_1, X_2, \dots)$. We know from Proposition 2.17 that $\widetilde{X}$ is a GEM(B) random vector.

In the same way, if we introduce a Bernoulli random variable I_0 with parameter p, independent from U, the X_n's and the I_n's, then the vector $(Z_1, Z_2, \dots)$ defined by

$$(Z_1, Z_2, \dots) = \begin{cases} \Big(U + (1-U)Y_1,\ (1-U)Y_2,\ (1-U)Y_3, \dots\Big) & \text{if } I_0 = 0 \\ \Big(U,\ (1-U)Y_1,\ (1-U)Y_2, \dots\Big) & \text{if } I_0 = 1, \end{cases}$$

is a realisation of $\mathcal{T}_p(Y_1, Y_2, \dots)$. We can also see that the vector Z can be obtained as a p-grouping of the vector $\widetilde{X}$, where the sequence of i.i.d. Bernoulli random variables used for the grouping is

$$(I_0, I_1, I_2, \dots).$$

Therefore, by applying Corollary 2.12, we conclude that $Z = \mathcal{T}_p(M_0)$ is a GEM(pB) random vector. □

We now have all the elements for the proof of Theorem 2.9.

Proof of Theorem 2.9. Combining the results of Proposition 2.16 and Proposition 2.18,

we obtain that

$$\big(M_0 \sim \mathrm{GEM}(pB)\big) \text{ implies } \big(M_1 \overset{\mathcal{D}}{=} \mathcal{T}_p(M_0) \sim \mathrm{GEM}(pB)\big),$$

and this completes the proof of Theorem (2.9). □

2.4 Conclusion

We have seen in this Chapter that if we introduce what we called the Bernoulli mutation mechanism, we can identify the stationary distribution for certain Λ-Fleming-Viot processes. The fact that the GEM distribution is the stationary distribution is a hint that looking only at allele frequencies might not be enough to distinguish various Λ-Fleming-Viot processes. We can extend this work in two directions.

The first one is to use this Bernoulli mutation model for other types of coalescent. Our current method does not seem to apply, because it was valid only for $\mathrm{Beta}(1, B)$ probability measures. It would be nice to find some similar results for the case $\mathrm{Beta}(A, B)$, and ideally be able to say something about the β-coalescent.

The second extension of this work is more applied, and consists in developing statistical methods that allow to compare various Λ-coalescent models by reconstructing the genealogies. We make some contribution in this direction in the next Chapter.

Chapter 3

Bayesian estimation of a Λ-coalescent

The work in this Chapter is a contribution to the the statistical inference of Λ-coalescent processes. It is important to try and connect these mathematical objects to the needs of biologists who work on a daily basis with real data. Our achievement is the development of an MCMC algorithm that allows one to perform Bayesian analysis on this type of genealogical model.

3.1 Introduction

As mathematical objects, Λ-coalescent processes have been studied for more than a decade now. There is also a growing literature on detecting evidence of these processes in biological data. For example, see Eldon and Wakeley (2006, 2009). However, the development of powerful statistical methods that allow one to infer genealogical trees on a large scale is still a recent concern. An important contribution is Birkner and Blath (2008) which presents an algorithm that allows one to compute the likelihood of a dataset under the infinitely-many sites model. This method allows for frequentist

inference of mutation rates and parameters on the Λ measures. We provide here a complementary approach, that is an MCMC algorithm that allows for Bayesian inference of the same parameters. Similar work has been done in the case of the Kingman coalescent, see for example Drummond et al. (2002). Basically, we use the same framework, and generalise the method to handle the Λ-coalescent models.

We present the statistical approach in a Bayesian setting, and then describe our innovation concerning the MCMC algorithm. Then we discuss the validity of our implementation in Matlab, before illustrating our algorithm and code with the inference of parameters of interest on simulated data. We conclude by presenting the future work required to convert this work into a powerful statistical tool.

3.2 Statistical model

3.2.1 Form of the data

The type of data we want to analyse is a set of aligned DNA sequences without gaps and of the same lengths. These restrictions come from the mutation model we are considering for the moment. In particular, we do not incorporate insertions and deletions.

We suppose that we have L sequences and S sites. The sequences are stored in an $L \times S$ matrix D with elements chosen from A, C, G and T. For clarity, we use a functional notation to access the matrix, that is $D(2,3)$ is the element at the second line and third column.

The assumption is that there is a genealogical relationship between the sequences from the data, and that this genealogy can be described by a Λ-coalescent tree. The model is structured in two levels:

– The tree $\mathcal{T}$ is a realisation of a Λ-coalescent tree with given Λ measure.

– Conditionally on the genealogical tree $\mathcal{T}$, the data D is generated by a mutation process run along the tree.

3.2.2 Λ-Coalescent process

We need to define formally the genealogical tree $\mathcal{T}$.

Definition 3.1. *Consider two integers L and N such that $L+1 \leq N \leq 2L-1$, an integer Root such that $L+1 \leq Root \leq N$, and a set of nonnegative real numbers $t_1, \ldots, t_N$.*

We can define a genealogical tree $\mathcal{T}$ with L leaves, N nodes, root Root and node ages $t_1, \ldots, t_N$ by specifying a directed tree with nodes $1, \ldots, N$ such that the nodes $1, \ldots, L$ have no children and the node Root has no parent.

For $i \neq Root$, we denote by i_P the parent of i. Formally, we define a cemetary point ∂ such that $Root_P = \partial$. We map each node i, $1 \leq i \leq N$ to its age t_i such that $t_{i_P} > t_i$, that is the parents are older than the children. We fix $t_1 = \cdots = t_L = 0$ and $t_\partial := \infty$.

Notation 3.2. *For a node i, we denote by b_i the number of lineages present just before node i , that is the number of lineages present at all times $t_i - \varepsilon$ for all $\varepsilon > 0$ sufficiently small. We also denote by $|i|$ the number of children of node i.*

Given a probability measure Λ, we remember from Chapter 2 that while there are b lineages, each k-tuple merges at a rate $\lambda_{b,k}$, and that the total transition rate is denoted by λ_b. Given the structure of the Λ-coalescent trees, the probability density of a specific tree $\mathcal{T}$ given a measure Λ is given by

$$L(\mathcal{T}; \Lambda) = \prod_{i=1}^{N} \lambda_{b_i,k_i} \exp\left(- \lambda_{b_i}(t_{i_P} - t_i)\right). \tag{3.1}$$

3.2.3 Mutation process

For the mutation process, we use the finite sites model Felsenstein (1981) with a general time-reversible substitution model. The parameters for this model are:

– a unit rate 4×4 substitution matrix Q,

– a total mutation rate μ.

This means that forward in time, each site mutates independently according to a continuous time Markov chain with intensity matrix μQ. To be able to use the matrix notation, we formally encode the symbols A, C, G and T by taking $A = 1$, $C = 2$, $G = 3$ and $T = 4$. For example, the mutation $A \to G$ happens at rate $\mu Q_{1,3}$. We denote by $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ the unique stationary base frequencies, defined by $\pi Q = 0$.

To generate the data, we sample a sequence of length S according to π, independently for each site, and map it to the node *Root*. Then recursively, we map to each node i an ancestral sequence D_i^A of length S obtained by running the mutation process along the branch connecting i_p to i. The sequences D_i^A at the nodes $i = 1, \ldots, L$ are taken to be the lines of the data matrix D.

Given the parameters Q and μ,and conditionally on the tree $\mathcal{T}$, the probability of the data is given as follows. We first condition on the internal ancestral sequences D_i^A, $i = L+1, ..., N$. Consider a node i and a site c. The probability that $D(i_P, c)$ has mutated to $D(i, c)$ during the time $(t_{i_P} - t_i)$ is given by

$$\left[e^{\mu Q(t_{i_P} - t_i)}\right]\big(D(i_P, c), D(i, c)\big),$$

where the quantity between the square brackets is the matrix exponential of $\mu(t_{i_P} - t_i)Q$, and we take the element at the line $D(i_P, c)$ and column $D(i, c)$ of this matrix. Because the sites evolve independently of each other, and by successive conditioning, we see that the probability of the data D conditionally on the internal ancestral sequences

D_i^A, $i = L+1, ..., N$ is given by

$$\prod_{i=1}^{N} \prod_{c=1}^{S} \left[e^{\mu Q (t_{i_P} - t_i)} \right] \big(D(i_P, c), D(i, c) \big),$$

By taking the expectation with respect to the internal ancestral sequences, we find that the probability of the data D conditionally on the tree $\mathcal{T}$ and the parameters μ, Q and Λ is given by

$$H\Big(D \,;\, \mu, Q, \mathcal{T} \Big) := \mathbb{E} \left[\prod_{i=1}^{N} \prod_{c=1}^{S} \left[e^{\mu Q (t_{i_P} - t_i)} \right] \big(D(i_P, c), D(i, c) \big) \,|\, \mu, Q, \mathcal{T} \right]. \tag{3.2}$$

Finally, the likelihood of the data as a function of the parameters is obtained by taking the expectation with respect to the tree $\mathcal{T}$, that is

$$\int H\Big(D \,;\, \mu, Q, \mathcal{T} \Big) \, L\big(\mathcal{T}; \Lambda \big) \, d\mathcal{T}. \tag{3.3}$$

3.2.4 Parameters and posterior distribution

The parameters in our model are the mutation parameters μ and Q and the probability measure Λ. To remain in the parametric setting, we can further specify Λ. Usually two particular cases are considered, the Eldon-Wakeley coalescent with $\Lambda(du) = \delta_{p_0}(du)$ for some $0 \leq p_0 < 1$, and the Beta coalescent in which $\Lambda(du)$ is a Beta$(2-\alpha,\alpha)$ distribution for $\alpha \in (1, 2)$. We notice that the Eldon-Wakeley model is the Kingman coalescent when we take $p_0 = 0$.

In a general setting, we choose a prior $p(\mu, Q, \Lambda)$. The posterior distribution is then up to a multiplicative constant given by

$$p\big(\mu, Q, \Lambda \,|\, D \big) = \int H\Big(D \,;\, \mu, Q, \mathcal{T} \Big) \, L\big(\mathcal{T}; \Lambda \big) \, d\mathcal{T} \; p\big(\mu, Q, \Lambda \big). \tag{3.4}$$

The main task in this Chapter is to develop an algorithm that allows one to compute this posterior distribution. There are two integrals to evaluate, one with respect to the ancestral sequences in (3.2), and one with respect to the tree $\mathcal{T}$. A first method would be to generate a sequence of random trees with a random DNA sequence at each of the internal nodes, and perform a Monte-Carlo integration for both integrals. We prefer to proceed differently by calculating directly the value of $H(D;\mu,Q,\mathcal{T})$. The reason for this is the existence of the *pruning algorithm*, presented in Felsenstein (1981). It allows one to compute this quantity very effectively thanks to a recursion along the tree $\mathcal{T}$.

Therefore, our strategy for evaluating the distribution (3.4) is to consider the unknown tree $\mathcal{T}$ as a parameter, and consider the distribution

$$p(\mu,Q,\Lambda,\mathcal{T}|\,D) = H\Big(D\,;\,\mu,Q,\mathcal{T}\Big)\,L(\mathcal{T};\Lambda)\;p(\mu,Q,\Lambda). \tag{3.5}$$

To evaluate (3.4), we generate a sequence of variables

$$x = (\mu,Q,\Lambda,\mathcal{T})$$

distributed according to (3.5). The variables μ, Q, and Λ generated in this way are distributed according to (3.4).

We are using a Markov-Chain Monte Carlo approach to simulate x according to (3.5). The novelty is due to the fact that under a Λ coalescent model, the tree $\mathcal{T}$ is a general tree with L leaves and can have between $L+1$ and $2L-1$ nodes, whereas in the Kingman case all the trees are binary and always have $2L-1$ nodes. Therefore, the dimension of x is unknown, which justifies the use of the Reversible Jump MCMC (RJ-MCMC) algorithm.

3.3 MCMC algorithm

3.3.1 Principle of RJ-MCMC

We want to generate a Markov Chain $(X_n)_{n\geq 0}$ with stationary distribution $p(x)$ when the state-space is of varying dimension. The method of RJ-MCMC has been introduced in Green (1995) to solve this problem. The idea is to generalise the classical Metropolis-Hastings algorithm by augmenting the state x with a variable u such that the variable (x, u) has a fixed dimension.

More precisely, suppose the current state of the chain is $X_n = x$, and we want to generate the next state X_{n+1}. The Metropolis-Hastings algorithm is based on an accept-reject procedure. A candidate state x' is generated according to a transition kernel $K(x, x')$. With probability $A(x, x')$ the candidate is accepted and we set $X_{n+1} = x'$, and with probability $1 - A(x, x')$ we reject it and we set $X_{n+1} = x$. The difficulty is in finding the correct acceptance probability $A(x, x')$ when the dimensions of x and x' are different.

An explicit representation of the transition kernel $K(x, x')$ allows us to find the right acceptance probability. Suppose that in order to generate a state x', we first generate a random variable u according to a distribution $g(u)$, and then set $x' = h(x, u)$. Consider the reverse transition, that is a transformation h' and a random variable u' sampled from a distribution g' such that $x = h'(x', u')$. The key idea is to make this transition reversible.

Therefore, we augment the state space and consider the couples (x, u) and (x', u'), a distribution g and a mapping h such that

- u and u' are distributed according to g and g' respectively,
- $(x', u') = h(x, u)$,
- and h is bijective with inverse h'.

We denote by

$$\left|\frac{\partial(x', u')}{\partial(x, u)}\right|$$

the Jacobian of the mapping $(x, u) \mapsto h(x, u)$.

The main result from Green (1995) is that in this setting, taking the acceptance probability to be

$$A(x, x') = \min\left\{1, \frac{p(x')}{p(x)} \frac{g'(u')}{g(u)} \left|\frac{\partial(x', u')}{\partial(x, u)}\right|\right\} \tag{3.6}$$

ensures that the chain X_n is stationary under $p(x)$.

3.3.2 Application to Λ coalescents

To construct an MCMC algorithm, we need to specify the proposal distribution, that is the distribution g and the transformation h, and to calculate analytically the corresponding acceptance probabilities. Furthermore, because we cannot simulate from $p(x)$, we need to make sure that the Markov chain is ergodic, that is the empirical distribution of a trajectory converges to the stationary distribution $p(x)$. It is crucial to verify that the chain converges.

When the state space is large and/or complex, which is the case in our situation, it is useful to partition the transition operator into distinct operators called *moves*. Each move is going to be as simple as possible, but rich enough to enable the chain to converge rapidly.

In our case, the state is of the form $x = (\mu, Q, \Lambda, \mathcal{T})$. There are many moves in the literature that allow one to create efficient MCMC algorithms for the Kingman coalescent. However, there is no method that allows to sample from general trees. Our main contribution in this respect is the *Merge-Split Move*, which allows us to efficiently sample general trees. We present this move in detail, then present the classical moves that we added in our algorithm to ensure a good convergence of the chain. We define

the *proposal ratio* $R(x, x')$ to be the quantity

$$\frac{g'(u')}{g(u)} \left| \frac{\partial(x', u')}{\partial(x, u)} \right|,$$

so that the acceptance probability is given by

$$A(x, x') = \min \left\{ 1, \frac{p(x')}{p(x)} R(x, x') \right\}. \tag{3.7}$$

The proposal ratio is specific to each move.

3.3.3 Merge-Split move

This move changes only the tree $\mathcal{T}$, and allows us to sample accross all the possible tree topologies. Before giving details, we need to introduce some notation.

Notation 3.3. *We denote by i_{PP} the parent of i_P. In the case where $|i_P| = 2$, we denote by $\widetilde{i}$ the only other child of i_P.*

We also need to define a truncated exponential distribution.

Notation 3.4. *Given two parameters $\rho, M > 0$, we denote by $f_{TE}(\rho, M, \cdot)$ the density of an exponential random variable with parameter λ conditioned to be less than M.*

Finally, we need to introduce some quantities that will be used for expressing the proposal ratio.

Definition 3.5. *Given two nodes i and j, we define*

$$M := t_{j_P} - \max(t_i, t_j).$$

If $|i_P| = 2$, we can then introduce

$$t_m := \max(t_i, t_{\widetilde{i}}).$$

Consider the *merging* probability p_M and the *splitting* probability $p_S := 1 - p_M$, and $\rho > 0$. The move is defined as follows.

Sample two nodes i and j without replacement. If one of the following conditions is satisfied, we decide that the move *fails*, that is we directly set $X_{n+1} = x$:

- $i = Root$,
- or $i_P = j$,
- or $i = j_P$,
- or $t_{j_P} \leq t_i$.

With probability p_M, we perform a *Merging* operation. In this case, if $j = Root$ or $i_P = j_P$, the move fails. Otherwise, we "unplug" the subtree subtended by the node i, and we reconnect it under node j_P. More precisely, there are two situations:

(M2) If $|i_P| = 2$, we delete the node i_P from the tree, and the node i becomes a child of the node j_P , without modifying any of the times. The dimension decreases by 1.

(M3) If $|i_P| \geq 3$, the node i becomes a child of the node j_P , without modifying any of the times.. The dimension remains the same.

With probability p_S, we perform a *Splitting* operation. First we generate a random variable δ with density $f_{TE}(\rho, M, \cdot)$. Then, we "unplug" the subtree subtended by the node i, and we reconnect it between the node j_P and the node j. There are also two cases:

(S2) If $|i_P| = 2$, we change the node i_P from being the child of i_{PP} to being the child of j_P, and from being the parent of $\widetilde{i}$ to being the parent of the offspring of j_p.

When $j = Root$, this means that i_P becomes the new root of the tree. The new age of i_P is set to $\max(t_i, t_j) + \delta$. There is no change in the dimension.

(S3) If $|i_P| \geq 3$, the move fails if $i_P = j_P$. Otherwise, we create a new node that is going to be the child of j_P and the parent of i and the offspring of j_P. If $j = Root$, the new node becomes the root of the tree. The age of the new node is set to $\max(t_i, t_j) + \delta$. The dimension is increased by 1.

Each operation has its inverse: $M3$ is paired with $M3$, $S2$ with $S2$ and $M2$ with $S3$. For each of these operations, the Jacobian appearing in (3.6) is equal to one. Therefore, the only terms appearing for the calculation of $R(x, x')$ are the forward and backward densities g and g'. Below, we provide these densities for each of these operations.

3.3.3.1 Case (M2)

$$g = |j_P| \frac{p_M}{N(N-1)},$$
$$g' = f_{TE}(\rho, t_{i_{PP}} - t_m, t_{i_P} - t_m) \frac{p_S}{(N-1)(N-2)}.$$

3.3.3.2 Case (M3)

$$g = |j_P| \frac{p_M}{N(N-1)},$$
$$g' = (|i_P| - 1) \frac{p_M}{N(N-1)}.$$

3.3.3.3 Case (S2)

$$g' = f_{TE}(\rho, M, \delta) \frac{p_S}{N(N-1)},$$
$$g = f_{TE}(\rho, t_{i_{PP}} - t_m, t_{i_P} - t_m) \frac{p_S}{N(N-1)}.$$

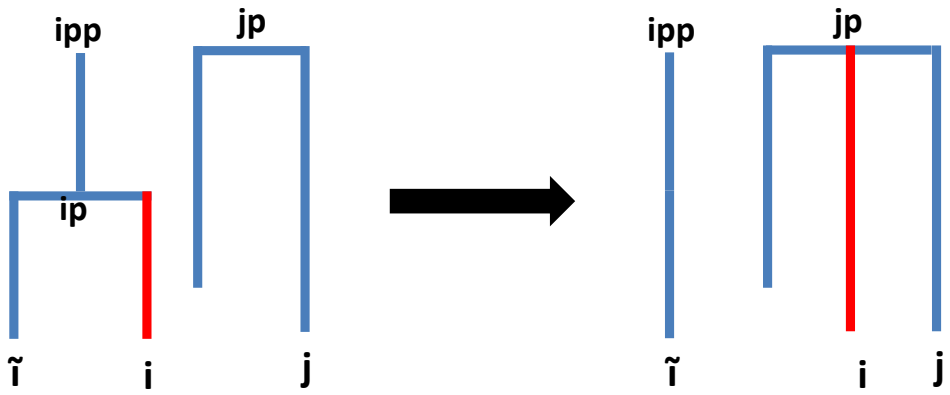

Figure 3.1: Operation M2 - The jump in dimension is -1.

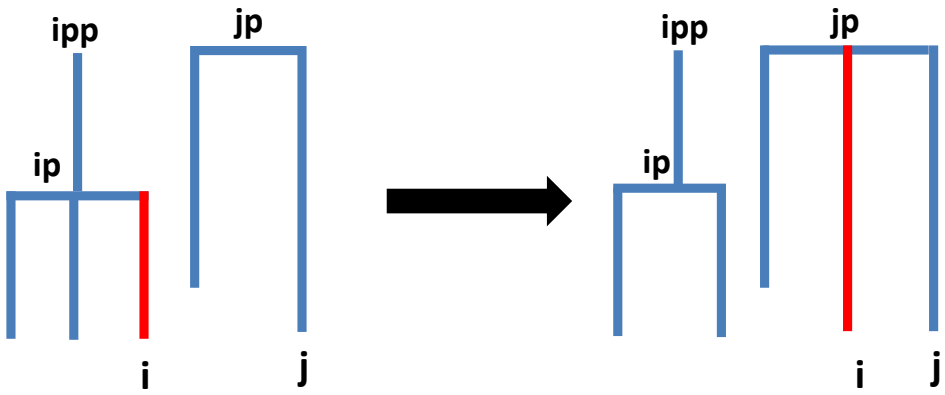

Figure 3.2: Operation M3 - There is no jump in dimension.

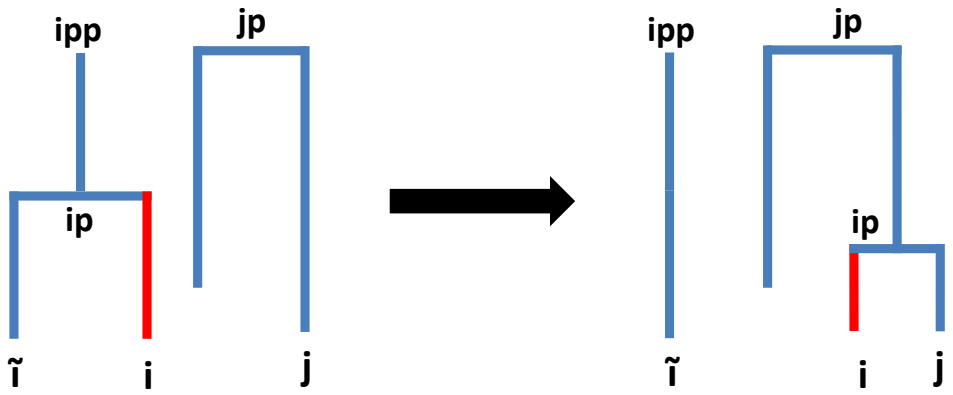

Figure 3.3: Operation S2 - There is no jump in dimension.

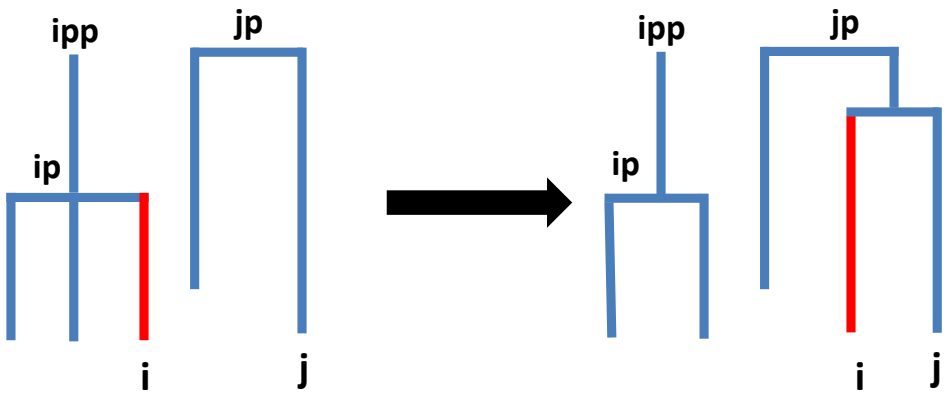

Figure 3.4: Operation S3 - The jump in dimension is +1.

3.3.3.4 Case (S3)

$$g = f_{TE}(\rho, M, \delta)\frac{p_S}{N(N-1)},$$
$$g' = (|i_P| - 1)\,\frac{p_M}{N(N+1)}.$$

For each of these operations, the proposal ratio is given by $R(x, x') = g'/g$.

3.3.4 Other moves

We obviously need to define moves that allow one to modify the coordinates μ, Q and Λ of the state x. We need to parameterise Λ to describe such moves in detail.

To illustrate the method, we just consider the simplest model, and we take $\Lambda(du) = \delta_{p_0}$ for $0 \leq p_0 \leq 1$. We also simplify the situation further by assuming that the parameter Q is known. Therefore we consider it as fixed, and we do not need to estimate it. A random walk for μ and p_0 is enough to sample along these two coordinates.

In practice, it is useful to accelerate the convergence of the chain to add moves that change the age of the nodes. Such moves are made precise in Drummond et al. (2002).

3.4 Implementation

We have developed in Matlab a full RJ-MCMC sampler in the case of the δ_{p_0}-coalescent. It allows to illustrate the method and perform analysis in a reasonable amount of time when the number of sequences L is not too large (less than a hundred). In this section, we provide tests that allow us to verify that the code works correctly so that we can trust its output.

3.4.1 Pruning algorithm

A central piece of the code is a function that computes the conditional distribution $H(D; \mu, Q, \mathcal{T})$ in (3.2) by using the pruning algorithm.

We perform the following test of our implementation of the pruning algorithm. We consider a fixed tree $\mathcal{T}$ with L leaves, and a number of sites S. We generate all the possible datasets D of size $L \times S$, and we partition them into subsets Σ_k, $k = 1, ..., K$. We want to calculate for each k the probability that D belongs to Σ_k. We compute this quantity in two ways:

- First by using the pruning algorithm for each each dataset $D \in \Sigma_k$ and summing the probabilities.
- Second through Monte Carlo simulations with independent realisations of the mutation process along the tree $\mathcal{T}$.

We then compare the two results for each subset Σ_k. We show the plots we obtain with a tree with $L = 8$, $K = 256$ and 10^6 realisations for the Monte Carlo simulation. Figure 3.5 shows a perfect match. We have performed this test with various values of L, K, and for various trees and mutation rates, and the results were all of the same quality. We are therefore confident that our implementation of the pruning algorithm is correct.

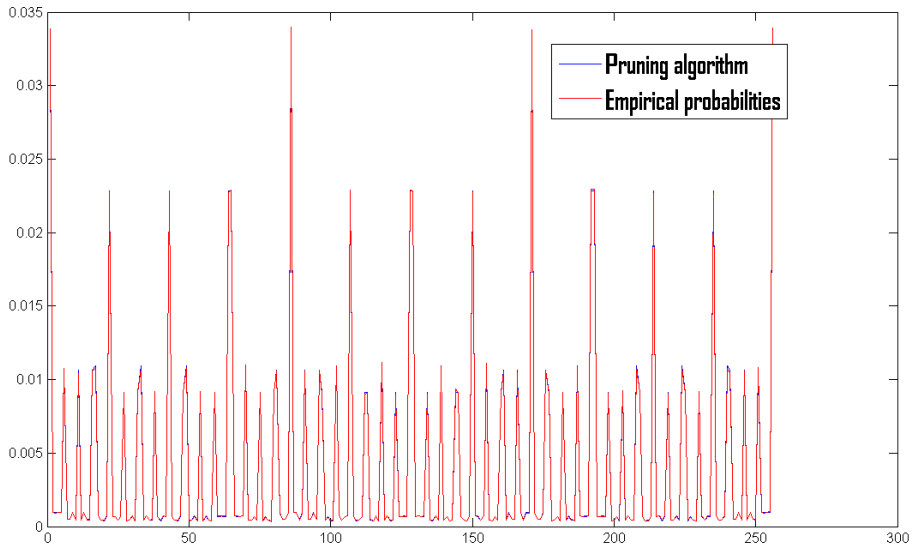

Figure 3.5: Verification of our implementation of the pruning algorithm - We take $L = 8$, and we classify all the possible sequence datasets into $K = 256$ bins. We calculate the probability that a dataset falls in each bin using on one hand the pruning algorithm, and on the other we simulate the process and perform a Monte Carlo evaluation of the probability. The fit between the two distributions is so good that the blue line is completely hidden under the red line.

3.4.2 Merge-Split move

The second test we perform is whether our MCMC sampler for the tree moves is correct. For this we consider a version of the sampler where there is no sequence data D. We want to verify that our MCMC sampler correctly simulates a Λ-coalescent tree. For this, the state space is restricted to the component $\mathcal{T}$, and we take the following two

moves:

– the Merge-Split move

– a local *Node Age Change* that modifies the age of a single node at a time.

The acceptance probability is given by

$$A(\mathcal{T}, \mathcal{T}') = \min\left\{1, \frac{L(\mathcal{T}';\Lambda)}{L(\mathcal{T};\Lambda)} R(x, x')\right\},$$

where the proposal ratio $R(x, x')$ was described in §3.3.3. We generate a sequence of trees with this RJ-MCMC sampler. In parallel, we generate a sequence of independent Λ-coalescent trees with the same number of leaves L and the same measure Λ through to a direct simulation. We compare the empirical measures of the two simulations. For the moment we look at the distribution of the age of the root of the tree (the time to the most recent common ancestor, or TMRCA). We show the results in figure 3.6. We observe a perfect fit of the two histograms. This test is already quite stringent, especially when L is not too large. Although we do not explicitly test for the distribution of the topology of the tree, the moves in the MCMC algorithm involve changes of the topology, so we are confident that not just the distributions of the times are correct, but also the joint distribution of the full genealogical tree $\mathcal{T}$. It would be nice to verify that the distribution of the topologies is correct. However, this is problematic to realise directly because we would need to enumerate all the possible tree topologies, and this is quite difficult when L is large.

However, in the next test we convince ourselves that the distribution of the topology is also correct.

3.4.3 Full sampler

Finally, we want to test whether when we consider a dataset D, the sequence generated by the MCMC sampler actually follows the posterior distribution (3.5). For this we

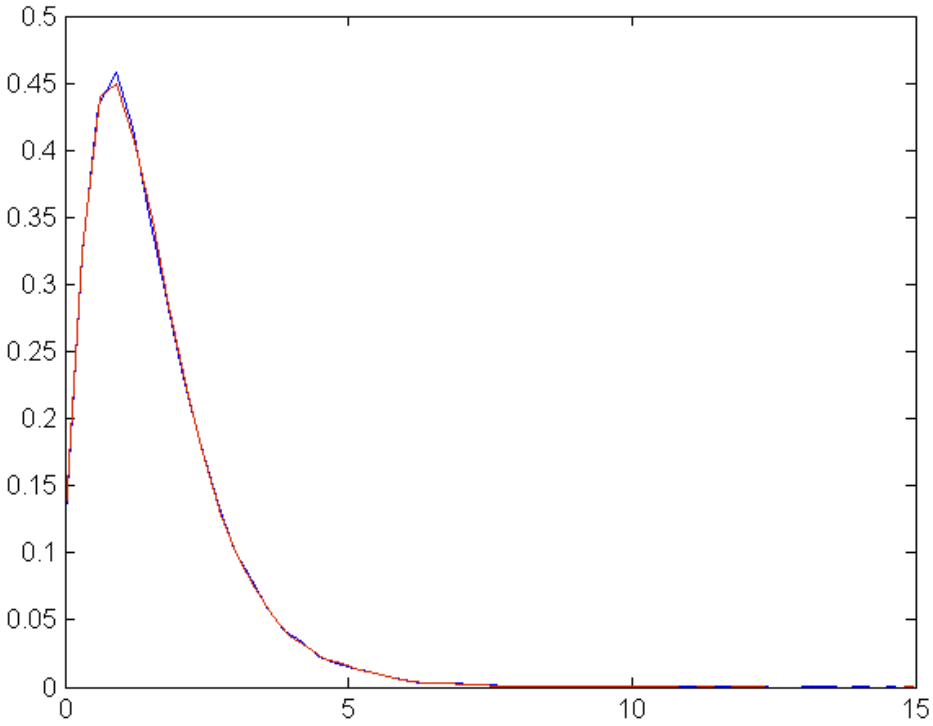

Figure 3.6: Verification of the tree moves - We run the MCMC sampler by considering only the tree moves, and by taking the target distribution $p(\mathcal{T})$ to be the distribution of a δ_{p_0}-coalescent. We record the age of the root, and compare its distribution with an independent direct simulation. We run the MCMC chain for 10^7 iterations and 10^5 independent samples, and we can see that the two distributions fit perfectly.

consider a small problem with $L = 3$ and $S = 4$. We fix a tree $\mathcal{T}$ and parameters μ and p_0, and we generate a dataset D. To be able to compute the posterior distribution (3.5) directly, we need to assume that we know the real values of the parameters μ and p_0. This means we specify the prior distributions to be point masses at the real values.

We are going to compare the posterior distribution of the topology of the tree $\mathcal{T}$ conditionally on the data D using two methods:

– First using the RJ-MCMC sampler

– Second, by a Monte Carlo simulation with independent runs.

In the case of the MCMC method, we just count the empirical frequency for each topology. When using the independent run Monte Carlo approach, we rely on the fact that the posterior distribution of $\Xi(\mathcal{T})$, the topology of $\mathcal{T}$, is given by

$$p(\Xi(\mathcal{T}) = \xi \mid D) = \int_{\mathcal{T}:\, \Xi(\mathcal{T})=\xi} H(D; \mu, Q, \mathcal{T})\, L(\mathcal{T}; \Lambda) d\mathcal{T}.$$

The integral is evaluated by a Monte Carlo integration, and the probability $H(D; \mu, Q, \mathcal{T})$ is calculated by the pruning algorithm, which we have already tested.

For $L = 3$, there are four possible tree topologies, labelled 1, 2, 3 and 4. In figure 3.7 we compare the calculations from the two methods, and once again we have a perfect match.

Verifying that a program is performing as intended is a never ending task. However, the various tests we have are quite strict, and they didn't detect any problem. We can therefore reasonably confidently rely on our implementation to perform some statistical analysis.

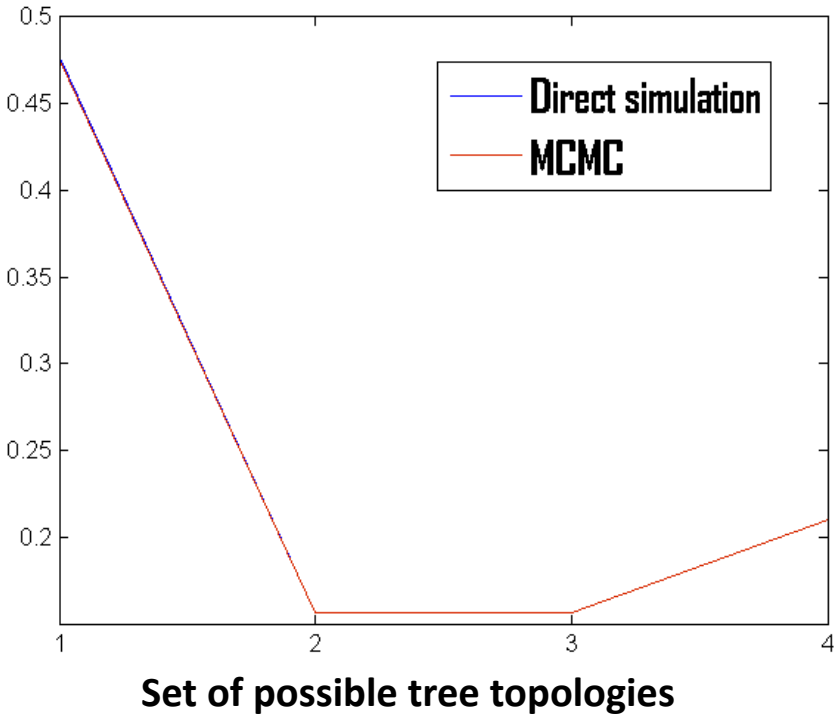

Figure 3.7: Posterior probabilities for all the topologies when $L = 3$ - We generate data with $L = 3$ and $S = 4$, then we compute the posterior probabilities using two methods, the first is our RJ-MCMC sampler, the second is an independent run Monte Carlo using the pruning algorithm. We can see that the fit between the two distributions is perfect.

3.5 Inference on simulated data

We simulate data in the setting we have already introduced, that is given values for the parameters p_0, μ and Q we first sample a δ_{p_0} coalescent tree, then a sequence drawn equilibrium distribution π for the bases, and we run the mutation process along the tree $\mathcal{T}$.

We suppose that the matrix Q is known. We could easily estimate its entries, but this would require running the chain for a much longer time. Later versions of the code will be optimized to run much faster.

We take a uniform prior on $[0, 1]$ for p_0 and a uniform prior on $[0, \mu_{max}]$ for μ, where μ_{max} can be large if we do not want the prior to be too informative.

Therefore, our task is to calculate the posterior distribution of μ and p_0 given the data.

In the example we are discussing, the dataset D was simulated with the parameters $L = 20$, $p_0 = 0.4$, $\mu = 1$ and $S = 80$. The average pairwise difference between two sequences was about 60%. We run our RJ-MCMC with the following moves:

- The Merge-Split move,
- A local *Node Age Change* that modifies the age of a single node at a time,
- Random walks for μ and p_0,
- and a *rescaling move* similar to the one presented in Drummond et al. (2002), in which we sample a random factor ϕ, then multiply each edge length by ϕ and divide the mutation rate by ϕ.

We run the algorithm for 10^6 iterations. In terms of performance, it runs in 5 hours on a desktop computer (Intel 2×2.13GHz). We first verify that the chain has correctly converged, and then we draw a few conclusions about the statistical inference of the parameters.

3.5.1 Convergence of the chain

We plot in figure 3.8 the time series of the parameters μ and p_0, as well as the total length of the tree, which is a proxy for the "size" of the tree. We notice by visual inspection in figure 3.8 that in all these dimensions, the chain mixes very well and converges quickly to a stationary state.

If we look now at the topology of the tree, we simply compare the real tree and a tree sampled at random from the distribution. This is a weak comparison, but the result is still very surprising. The topology of the two trees is identical, as well as the "shape", that is the relative length of the edges. We have verified these features in many simulations. The Merge-Split move and the Node Change Move are very effective for rapidly converging to the right topology and the right relative sizes.

What made the convergence difficult was the identification of the correct "size" of the tree. We understand this if we compute the probability of the data given the right topology and shape of the tree but let the scale l of the tree vary (figure 3.11). There is a region of very high probability around a curve of the form $\mu\, l = \text{constant}$. This is why the rescaling move we added was so important. It allows to sample along this curve. We can see in figure 3.12 that the chain is concentrated on the curve of interest. This is why the chain mixes so quickly.

3.5.2 Discussion

If we look now at the posterior distributions, we see that there is no bias in our estimations for μ and p_0. This is reassuring, because the dataset we simulated was quite informative. In particular, given that the prior on μ was significantly non informative, we have a reasonably accurate estimation of μ.

Concerning p_0, there is a very clear signal around the true value. In particular, this shows that the existence of multiple mergers in the tree is very clearly detected in the data. However, there is some uncertainty about the actual value of p_0. We suspect

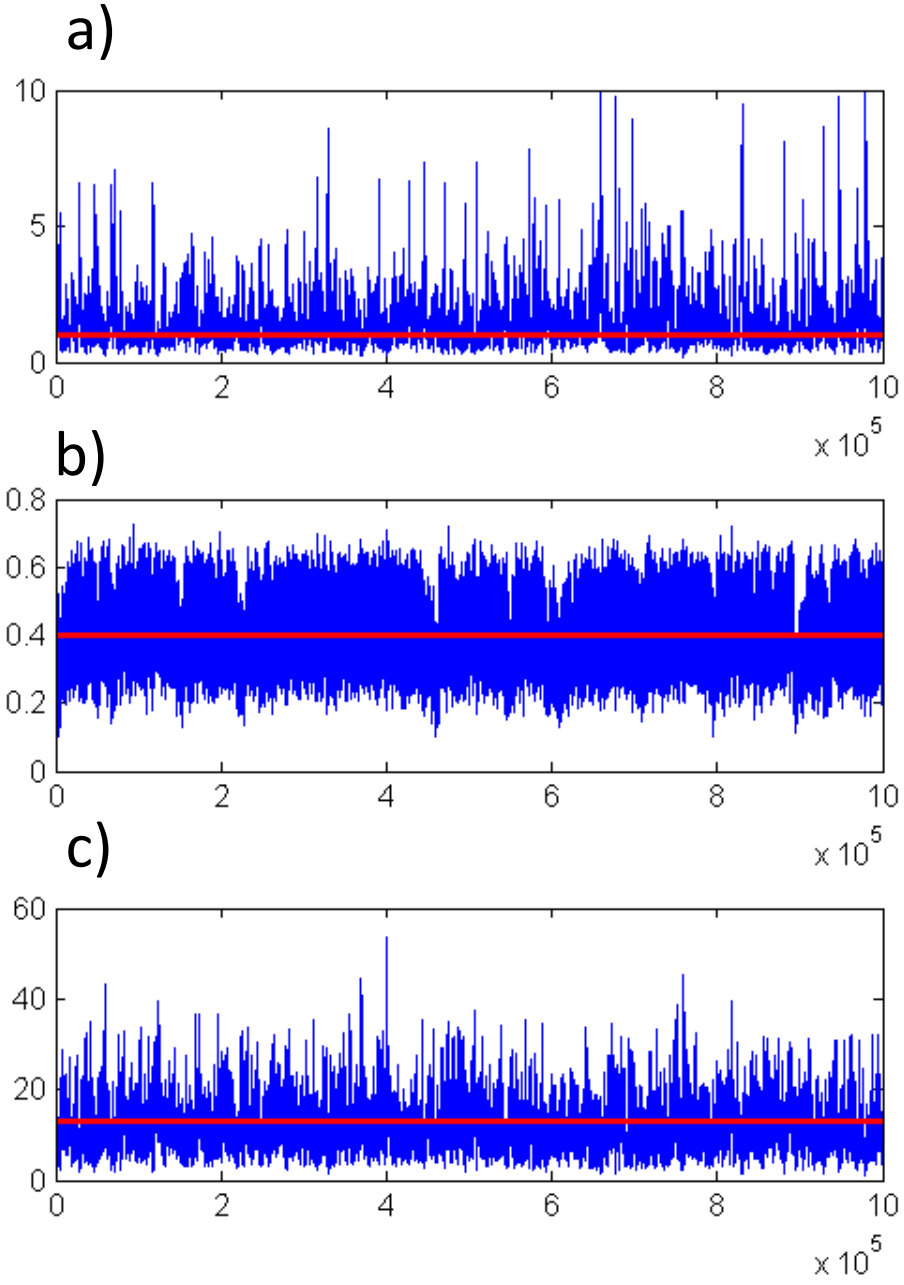

Figure 3.8: Trajectories of the MCMC run for 10^6 iterations - a) Mutation rate μ; b) Fraction p_0; c) Total length of the tree. The horizontal red line is the true value of the parameter in question.

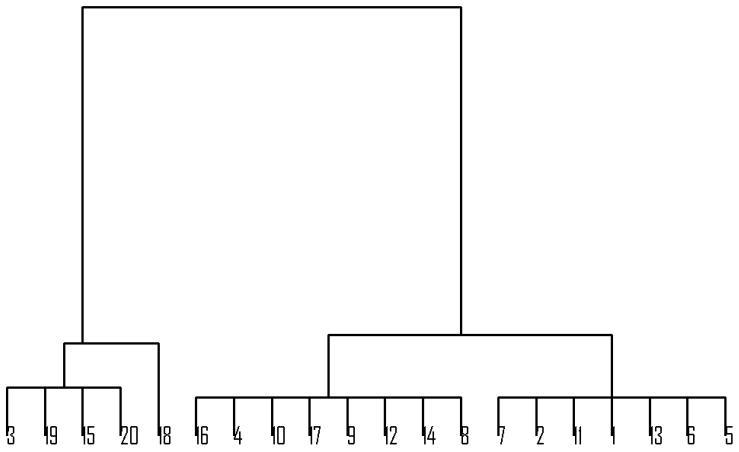

Figure 3.9: Real genealogical tree. - Used to generate the data we analysed

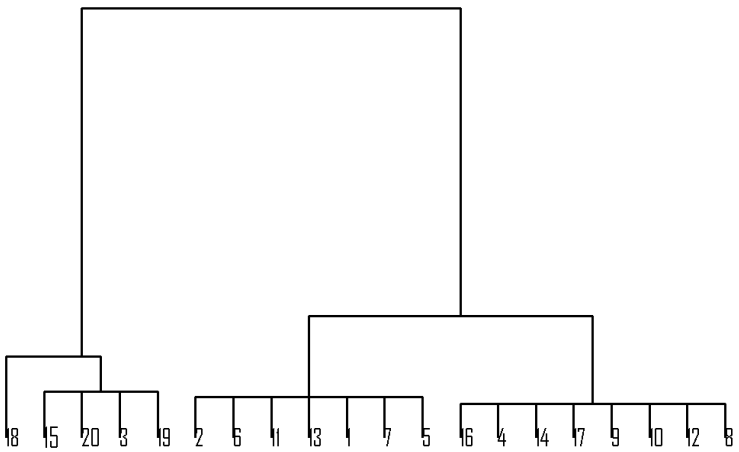

Figure 3.10: Genealogical tree - We sample a tree from the posterior distribution generated by the MCMC algorithm. It is exactly the same topology as the real tree, and the shape of the tree is very close.

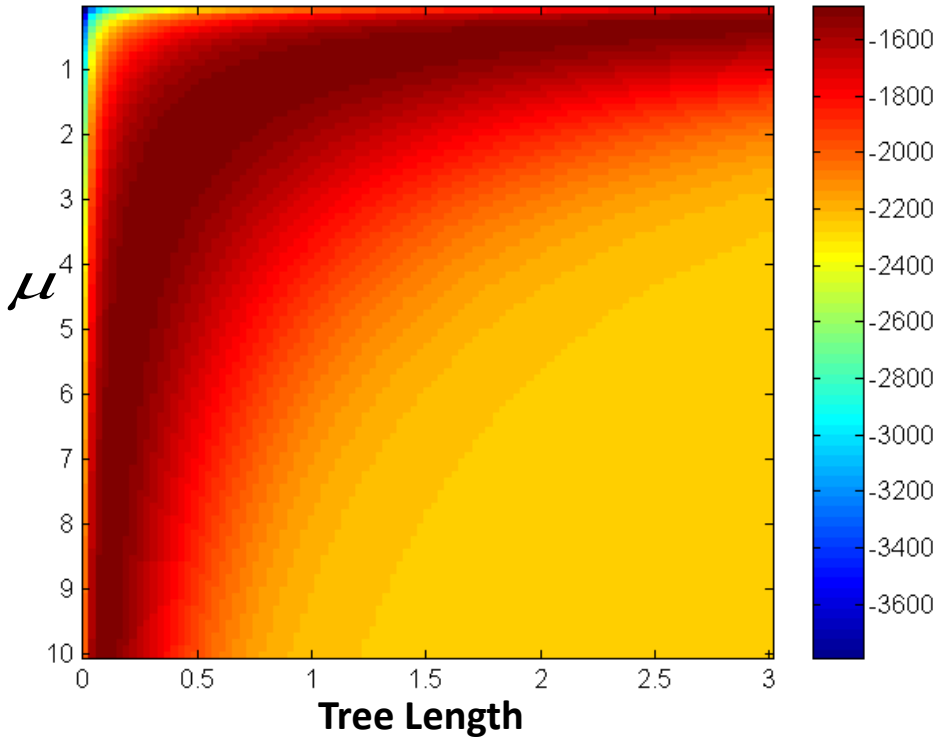

Figure 3.11: Probability of the data as a function of the mutation rate μ and the length of the tree - The red regions correspond to large likelihood values of the parameters. The shape of the curve can make the convergence of the MCMC algorithm quite slow.

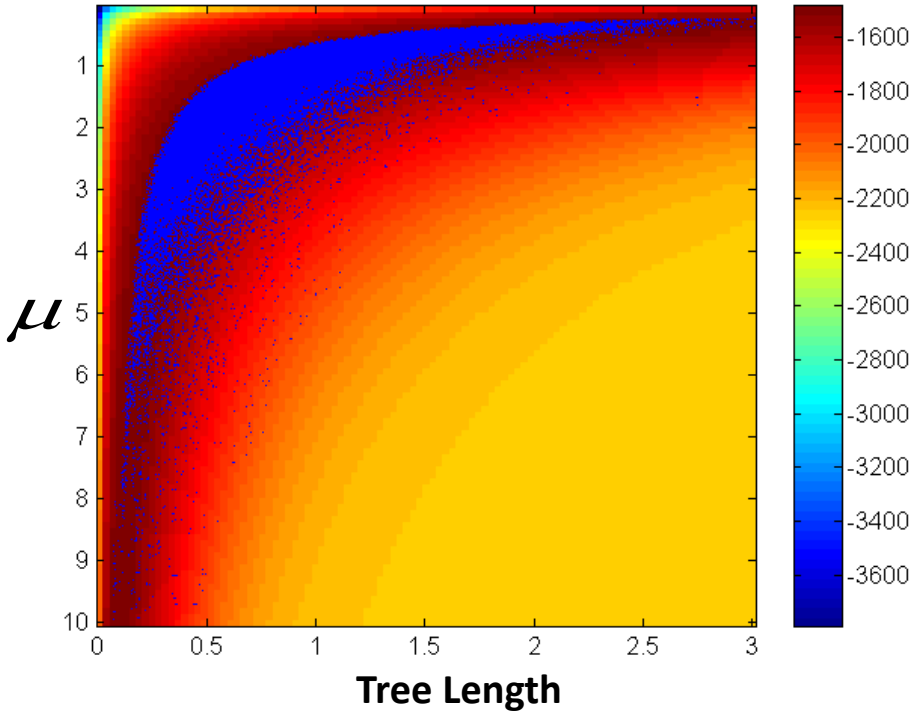

Figure 3.12: Trajectory of the MCMC run on the likelihood curve - The blue points correspond to the positions of the Markov Chain we generate. We see that we sample very efficiently from the regions with a high likelihood. This is because we have a scaling move that allows the chain to move along the hyperbola.

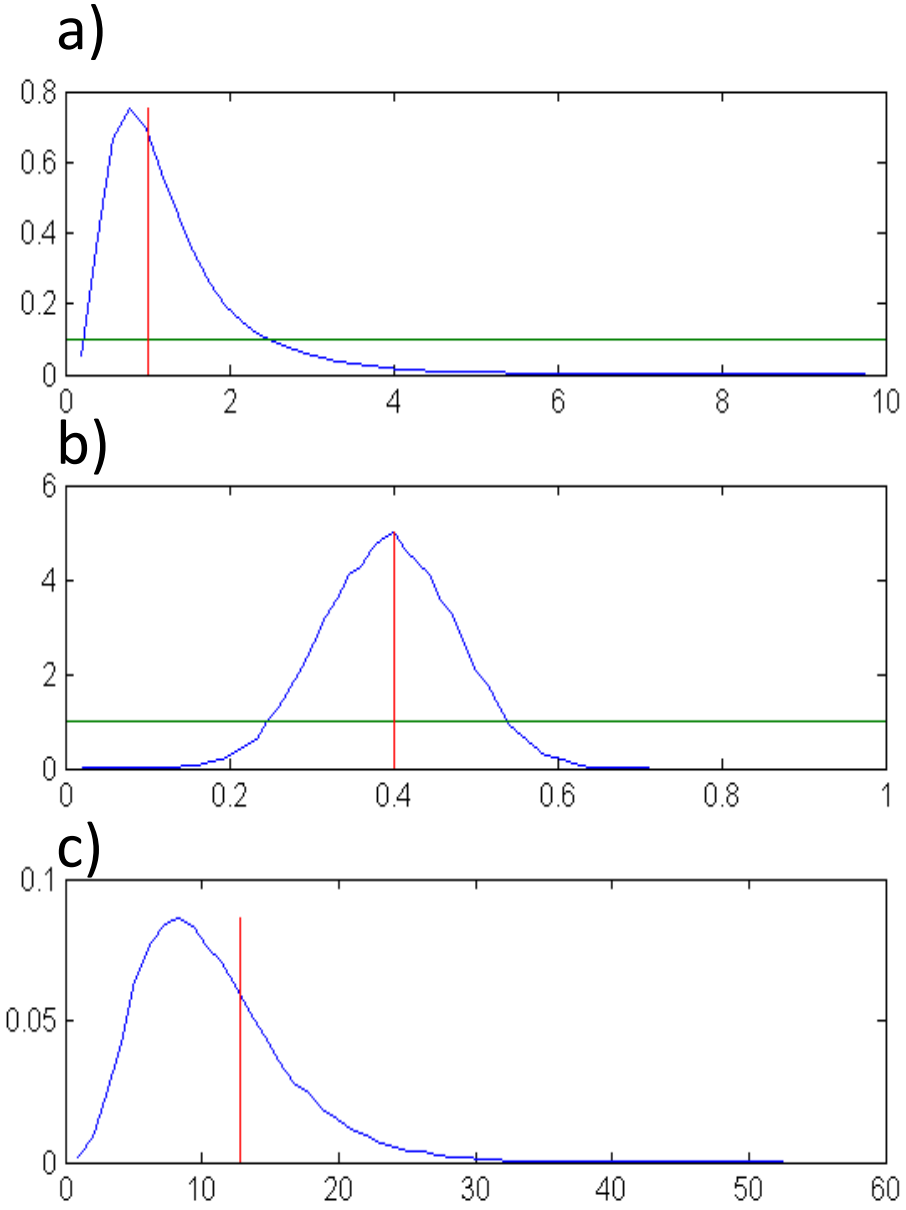

Figure 3.13: Posterior distribution of the parameters given by the MCMC algorithm - a) Mutation rate μ; b) Fraction p_0; c) Total length of the tree. The blue lines are the posterior distributions, the green lines the priors, and the vertical red line the position of the true parameter.

that this is due to the fact that the sample size $L = 20$ is rather small. We would need to optimise our code to be able to handle larger datasets.

3.6 Conclusion

We have explained the details of our main contribution, the Merge-Split move in the framework of the RJ-MCMC algorithm. We also demonstrated, thanks to a prototype implemented in Matlab, that

- Our Merge-Split move is very effective in sampling across the large state space of the Λ-coalescent trees,
- It is possible to incorporate this move into a full MCMC sampler that actually converges (at least in the models we considered),
- We confirmed the results from Birkner and Blath (2008) and Birkner and Blath (2008) that it is possible to detect in DNA sequences the presence of multiple mergers in the genealogy.

This work should be pursued in the following directions:

- Allow for more choices of Λ, in particular the case of the Beta coalescents.
- Develop a piece of software in a powerful language, for example C++, to be able to handle larger datasets in a reasonable amount of time,
- Develop a full Bayesian methodology, in particular by allowing for model comparison between various choices of Λ.

Part II

The Spatial-Λ-Fleming-Viot process

Chapter 4

Spatial models

In this chapter, we focus on models that incorporate spatial structure. The simplest way to analyse the influence of space on genetic diversity is to study the *probability of identity* (POI) as a function of the distance, that is the probability that two individuals separated by a distance x have the same genetic type.

We first present the Stepping Stone Model, due to Kimura (1953). It is a discrete space model which generalises the Wright-Fisher model. For this model, we are able to explicitly calculate the POI, and then derive asymptotic formulae.

We then introduce Malécot's formula, presented in Malécot (1948), which models a population living on a two dimensional continuum. It is also possible to make explicit calculations for the POI, and then derive approximations. However, this model presents certain mathematical flaws, and we discuss them.

Finally, we conclude by presenting the Spatial-Λ-Fleming-Viot process (SΛFV), due to Barton et al. (2010a). It has the major advantage of providing a mathematically coherent continuous-space model. Furthermore, it extends to the spatial setting the Λ-Fleming-Viot processes encountered in Chapter 2. After discussing the difficulties in constructing a continuous-space model, we present the approximate calculations for the POI in the SΛFV process.

4.1 Stepping Stone Model

4.1.1 1D nearest neighbour Stepping Stone Model

4.1.1.1 The model

We consider a population in which the individuals live in separate colonies positioned on an infinite line, and we study the dynamics of a single genetic locus with 2 alleles a and A. Each colony (or "deme") being indexed by an integer i, we define $p_i(t)$ to be the proportion of the population in colony i at generation t which is of type a. We suppose that the only migration allowed is between adjacent colonies, and that at every generation a proportion $m_1/2$ of the individuals living in the colony i migrate to the colony $i-1$, and the same proportion $m_1/2$ migrate to the colony $i+1$. We also suppose that there are mutations : at each generation, a proportion μ_1 of a alleles mutate to A, and a proportion μ_2 of A alleles undergo the reverse mutation. This gives the following dynamics:

$$p_i(t+1) = (1 - m_1 - \mu_1)\, p_i(t) + \mu_2\,(1 - p_i(t)) + \frac{m_1}{2}\,(p_{i-1}(t) + p_{i+1}(t)) + \xi_i(t). \quad (4.1)$$

The last term $\xi_i(t)$ is a random variable which accounts for the random drift due to the reproduction in a finite population. The drift is represented by a Wright-Fisher model applied to each colony with an effective population size N_e. Thus, $\xi_i(t)$ is the change in allele frequencies in a Wright-Fisher model, and conditionally on p_i we have $\mathbb{E}(\xi_i(t)|p_i) = 0$ and $\mathbb{E}(\xi_i(t)^2|p_i) = \frac{p_i(1-p_i)}{2N_e}$.

4.1.1.2 Probability of Identity

We are going to focus on the results presented in two papers by Kimura and Weiss (1964) and Kimura and Weiss (1965). We are interested in the probability of identity for two

individuals sampled from different colonies, or equivalently the correlation between the allele frequencies in these colonies. For this we consider the situation where the population is in equilibrium, and we denote by $\bar{p}_i$ the expectation of the frequency of the allele a in colony i . Taking the expectation in equation (4.1) leads to

$$\bar{p} = \bar{p}_i = \frac{\mu_2}{\mu_1 + \mu_2} \quad \forall i. \tag{4.2}$$

We introduce the total rate of mutation $\mu := \mu_1 + \mu_2$. Thus equation (4.1) becomes:

$$p_i(t+1) = (1 - m_1 - \mu)p_i(t) + \frac{m_1}{2}(p_{i-1}(t) + p_{i+1}(t)) + \mu\bar{p} + \xi_i(t). \tag{4.3}$$

We center the random variables p_i, defining

$$\tilde{p}_i := p_i - \bar{p}_i.$$

If we introduce the notation $\alpha := 1 - m_1 - \mu$, $\beta := m_1/2$, then equation (4.3) reads

$$\tilde{p}_i(t+1) = \alpha\tilde{p}_i(t) + \beta(\tilde{p}_{i-1}(t) + \tilde{p}_{i+1}(t)) + \xi_i(t). \tag{4.4}$$

When the population is in equilibrium, the correlation between frequencies in colonies i and $i + k$ is then simply

$$r_k = \frac{\mathbb{E}[\tilde{p}_i\, \tilde{p}_{i+k}]}{\mathbb{E}[\tilde{p}_i{}^2]},$$

and depends only on the separation k.

As the population is in equilibrium, the calculations to establish a recurrence relation for the r_k are straightforward. They are achieved by multiplying equation (4.4) by the same recurrence relation (4.4) written for $i + k$, and then taking the expectation.

This gives:

$$(\alpha^2 + 2\beta^2 - 1)r_k + 2\alpha\beta(r_{k+1} + r_{k-1}) + \beta^2(r_{k+2} + r_{k-2}) = 0 \tag{4.5}$$

This difference equation is classically solved using the method of the characteristic polynomial. The latter being a fourth degree polynomial, it has four roots $\lambda_1, \ldots, \lambda_4$. The general formula for r_k is then:

$$r_k = \sum_{i=0}^{4} C_i \lambda_i^k, \tag{4.6}$$

where $C_1, \ldots, C_4$ are constants. It is shown in Kimura and Weiss (1964) that we can label the roots $\lambda_1, \ldots, \lambda_4$ such that $\lambda_3 < -1 < \lambda_4 < 0 < \lambda_2 < 1 < \lambda_1$. Since r_k is a correlation, it is bounded between -1 and 1 and $r_0 = 1$, which implies that $C_1 = C_3 = 0$, and

$$C_2 = \frac{R_1}{R_1 + R_2} \quad \text{and} \quad C_4 = \frac{R_2}{R_1 + R_2},$$

where

$$R_1 = \sqrt{(2 - m_1 - \mu)^2 - m_1^2} \quad \text{and} \quad R_2 = \sqrt{(m_1 + \mu)^2 - m_1^2}.$$

The expression for λ_2 and λ_4 is given by

$$\lambda_2 = \frac{1}{m_1}(m_1 + \mu - R_2) \quad \text{and} \quad \lambda_4 = \frac{-1}{m_1}(2 - m_1 - \mu - R_1).$$

We then have

$$r_k = C_2\lambda_2^k + C_4\lambda_4^k.$$

If we consider the case where $\mu << m_1$, that is when the mutation rate is very small compared to the migration rate, then we have the following approximations:

$$R_1 \approx 2\sqrt{1-m_1}, \quad R_2 \approx \sqrt{2m_1\mu}, \quad |\lambda_4| < \lambda_2.$$

From these we deduce that for $\mu/m_1 \to 0$,

$$r_k \approx C_2 \left(1 - \sqrt{\frac{2\mu}{m_1}}\right)^k \quad \text{and} \quad C_2 \to 1.$$

We finally obtain the following behaviour of the correlation r_k for large sampling distance k:

$$r_k \approx \exp\left(-\sqrt{\frac{2\mu}{m_1}}k\right).$$

This exponential decay of the probability of identity with respect to distance is characteristic of dimension 1. As we shall see, this kind of asymptotic result is robust to modifications of the migration scheme.

4.1.2 General 1D Stepping Stone Model

4.1.2.1 Symmetric migration

In this part we take a stepping stone model in one dimension with an arbitrary symmetric migration. At every generation, a proportion $m_j/2$ of the individuals living in the colony i migrate to the colony $i-j$, and the same proportion $m_j/2$ migrate to the colony $i+j$, with $j = 1, 2, \ldots$. The rest of the model is unchanged, and equation (4.4) becomes

$$\tilde{p}_i(t+1) = (1 - \sum_{j=1}^{\infty} m_j - \mu)\tilde{p}_i(t) + \frac{1}{2}\sum_{j=1}^{\infty} m_j(\tilde{p}_{i-j}(t) - \tilde{p}_{i+j}(t)) + \xi_i(t). \tag{4.7}$$

For a more compact notation, we define the shift operator $\mathbf{S}$, which is invertible, by

$$\mathbf{S}\, f(i) = f(i+1) \quad \text{and} \quad \mathbf{S}^{-1} f(i) = f(i-1).$$

We also define the operator $\mathbf{L}$ by

$$\mathbf{L} := \left(1 - \sum_{j=1}^{\infty} m_j - \mu\right) + \frac{1}{2}\sum_{j=1}^{\infty} m_j(\mathbf{S}^j + \mathbf{S}^{-j}).$$

Using vector notation for $\underline{p}(t) := (\dots, p_{-1}(t), p_0(t), p_1(t), \cdots)$ and $\underline{\xi}(t)$, the dynamics of the population (4.7) becomes

$$\underline{\tilde{p}}(t+1) = \mathbf{L}\underline{\tilde{p}}(t) + \underline{\xi}(t) \tag{4.8}$$

4.1.2.2 Covariance calculations

In the previous section, the calculations for the correlation r_k were straightforward thanks to the difference equation (4.5). In the case of the general migration scheme, we need another method. We first need to compute the covariance

$$\rho_k = \mathbb{E}[\tilde{p}_0\, \tilde{p}_k].$$

The population being in equilibrium, we have

$$\mathbb{E}[\tilde{p}_0(t)\, \tilde{p}_k(t)] = \mathbb{E}[\tilde{p}_0(t+1)\, \tilde{p}_k(t+1)],$$

so equation (4.8) gives

$$\mathbb{E}[\tilde{p}_0\, \tilde{p}_k] = \mathbb{E}[(\mathbf{L}\underline{\tilde{p}})_0\, (\mathbf{L}\underline{\tilde{p}})_k].$$

With direct calculations, one can go further and show that

$$\mathbb{E}[\tilde{p}_0\,\tilde{p}_k] = \Big(\mathbf{L}^2\mathbb{E}[\,\tilde{p}_0\underline{\tilde{p}}\,]\Big)_k. \tag{4.9}$$

The set of equations fulfilled by the covariance function ρ_k is then

$$\begin{cases} \big[\,(1-\mathbf{L}^2)\underline{\rho}\,\big]_k = 0, & k \neq 0 \\ \\ \big[\,(1-\mathbf{L}^2)\underline{\rho}\,\big]_0 = \dfrac{-1}{2N_e}\rho_0 + \dfrac{\bar{p}-\bar{p}^2}{2N_e}, \end{cases} \tag{4.10}$$

where the operator $\mathbf{L}$ is applied to the vector $\underline{\rho} = (\dots,\, \rho_{-1},\, \rho_0,\, \rho_1, \cdots)$. The first line is simply obtained by rewriting equation (4.9), whereas the second follows from similar calculations in the case where $k = 0$. Once we know the full covariance vector $\underline{\rho}$, we deduce easily the correlation vector $\underline{\rho}/\underline{\rho}_0$.

We look for a solution for equations (4.10) of the form

$$\rho_k = \frac{1}{2\pi}\int_0^{2\pi} F(\theta)\,\cos k\theta\;d\theta. \tag{4.11}$$

Solving equations (4.10) consists in identifying the function F in the representation (4.11). The first step is to apply the operator $\mathbf{L}$ to the vector $\underline{\phi}$, where $\underline{\phi}_k = \cos(k\theta)$. Direct calculations show that

$$\big[\,(1-\mathbf{L}^2)\underline{\phi}\,\big]_k = \big(1 - H^2(\theta)\big)\,\cos(k\,\theta),$$

where

$$H(\theta) = \sum_{j=0}^{\infty} m_j \cos(j\,\theta), \qquad \text{and} \qquad m_0 = 1 - \sum_{j=1}^{\infty} m_j - \mu.$$

This allows us to express the vector $(1-\mathbf{L}^2)\,\underline{\rho}$ as

$$\left[(1-\mathbf{L}^2)\underline{\rho}\right]_k = \frac{1}{2\pi}\int_0^{2\pi} F(\theta)\,\left(1-H^2(\theta)\right)\,\cos(k\theta)\,d\theta.$$

In particular, the first line of (4.10) becomes

$$\frac{1}{2\pi}\int_0^{2\pi} F(\theta)\,\left(1-H^2(\theta)\right)\,\cos(k\theta)\,d\theta = 0\ , \quad k\neq 0.$$

A natural solution is to take

$$F(\theta) = \frac{C}{1-H^2(\theta)},$$

where C is a constant. This gives the general formula

$$\rho_k = \frac{1}{2\pi}\int_0^{2\pi} \frac{C}{1-H^2(\theta)}\,\cos(k\theta)\,d\theta. \tag{4.12}$$

We can determine the constant C using the second line of (4.10):

$$\frac{1}{2\pi}\int_0^{2\pi} F(\theta)\,\left(1-H^2(\theta)\right)\,d\theta = \frac{-1}{2N_e}\rho_0 + \frac{\bar{p}-\bar{p}^2}{2Ne},$$

which yields

$$C = \frac{\bar{p}(1-\bar{p})-\rho_0}{2N_e}.$$

To identify ρ_0, we plug this expression into (4.12) applied with $k=0$, and this gives the formula for the variance of an allele frequency in one deme:

$$\rho_0 = \frac{\bar{p}(1-\bar{p})}{1+4\pi N_e\left(\displaystyle\int_0^{2\pi}\frac{d\theta}{(1-H^2(\theta))}\right)^{-1}}.$$

Now the covariance vector $\underline{\rho}$ is completely determined. As our primary focus is on the correlation vector $\underline{r}$, using equation (4.12), we simply divide the expression for ρ_k by

ρ_0 to obtain

$$r_k = \frac{\displaystyle\int_0^{2\pi} \frac{\cos(k\theta)}{1-H^2(\theta)}\, d\theta}{\displaystyle\int_0^{2\pi} \frac{1}{1-H^2(\theta)}\, d\theta} \tag{4.13}$$

4.1.2.3 Asymptotics for large sampling distance

We are primarily interested in finding an asymptotic expression for r_k when $k \to \infty$. We first rewrite formula (4.13) in terms of two functions A_1 and A_2 defined by

$$A_1(k) = \int_0^{2\pi} \frac{\cos(k\theta)}{1-H(\theta)}\, d\theta \quad \text{and} \quad A_2(k) = \int_0^{2\pi} \frac{\cos(k\theta)}{1+H(\theta)}\, d\theta.$$

Thus we have:

$$r_k = \frac{A_1(k) + A_2(k)}{A_1(0) + A_2(0)}$$

We want to know the asymptotic behaviour of $A_1(k)$. As

$$A_1(k) = 2\int_0^{\pi} \frac{\cos(k\theta)}{\mu + \sum_{j=1}^{\infty} m_j(1-\cos(j\,\theta))}\, d\theta,$$

we expand $\cos(j\,\theta)$ inside the denominator using

$$1-\cos(j\,\theta) = \sum_{n=1}^{\infty} \frac{(-1)^{2n+1}(j\theta)^{2n}}{(2n)!}.$$

We keep only the dominant term in the expansion, and after introducing the notation $\mathcal{M}_2 := \sum_{j=1}^{\infty} j^2\, m_j$ for the second moment of the migration kernel, we have

$$A_1(k) \sim 2\int_0^{\pi} \frac{\cos(k\theta)}{\mu + \frac{\theta^2}{2}\mathcal{M}_2}\, d\theta.$$

Up to an exponentially small error, we can replace the upper limit of integration by ∞ to obtain

$$A_1(k) \sim 2\int_0^{\infty} \frac{\cos(k\theta)}{\mu + \frac{\mathcal{M}_2}{2}\,\theta^2}\, d\theta.$$

Knowing that the Fourier transform of $e^{-a|t|}$ is $\sqrt{2/\pi}\ a/(a^2+\omega^2)$, we can explicitly calculate the integral, and we obtain

$$A_1(k) \sim \frac{2\pi}{\sqrt{2\mu\,\mathcal{M}_2}} e^{-k\sqrt{2\,\mu/\mathcal{M}_2}}.$$

One can also show that for large k, $A_2(k)$ is negligible compared to $A_1(k)$. The asymptotic behaviour of the correlation when $k \to \infty$ is then given by

$$r_k \sim D_1 \cdot e^{-k\sqrt{2\,\mu/\mathcal{M}_2}}, \tag{4.14}$$

where D_1 is a constant.

This result is a generalisation of the one we found in the previous section for the simple migration scheme. Thus we see that as long as the second moment of the migration kernel is finite, we obtain a very robust result, namely that in one dimension the probability of identity of two individuals decays exponentially for large sampling distances. We want to have a similar result in dimension two, and a natural question is whether the dimension plays a role in the asymptotics.

4.1.3 2D Stepping Stone Model

The general method of solution we just presented applies well in the two dimensional case. We begin by defining rigorously the stepping stone model in two dimensions. Now each colony is indexed by a pair of integers (k_1, k_2), and we allow a general symmetric migration scheme. At every generation, a proportion $m_{ij}/2$ of the individuals living in the colony (k_1, k_2) migrate to the colony $(k_1 - i, k_2 - j)$, and the same proportion migrate to the colonies $(k_1 - i, k_2 + j)$, $(k_1 + i, k_2 + j)$ and $(k_1 + i, k_2 - j)$.

We can use the same formalism as before, namely that the dynamics of the popu-

lation can be expressed by equation (4.8):

$$\underline{\tilde{p}}(t+1) = \mathbf{L}\underline{\tilde{p}}(t) + \underline{\xi}(t).$$

We adapt the notation to the two dimensions. Here $\underline{\tilde{p}}$ is an infinite matrix: $\underline{\tilde{p}} = (p_{i,j})_{i,j\in\mathbb{Z}}$. We need now two shift operators, one for each dimension:

$$\mathbf{S}_1\, f(i,j) = f(i+1,j)$$
$$\mathbf{S}_2\, f(i,j) = f(i,j+1).$$

The operator $\mathbf{L}$ becomes

$$\mathbf{L} := m_0 + \frac{1}{4} \sum_{\substack{i,j=0 \\ (i,j)\neq(0,0)}}^{\infty} m_{ij}(\mathbf{S}_1^i + \mathbf{S_1}^{-i})(\mathbf{S}_2^j + \mathbf{S_2}^{-j}),$$

where

$$m_0 := 1 - \sum_{\substack{i,j=0 \\ (i,j)\neq(0,0)}}^{\infty} m_{ij} - \mu.$$

Kimura and Weiss (1965) show that the general expression for the correlation function in two dimensions is of the form

$$r(k_1,k_2) = \frac{\displaystyle\int_0^{2\pi}\int_0^{2\pi} \frac{\cos(k_1\theta_1)\cos(k_2\theta_2)}{1-H^2(\theta_1,\theta_2)}\, d\theta}{\displaystyle\int_0^{2\pi}\int_0^{2\pi} \frac{1}{1-H^2(\theta_1,\theta_2)}\, d\theta}.$$

It is possible to rewrite $r(k_1,k_2)$ in terms of two functions $A_1(k_1,k_2)$ and $A_2(k_1,k_2)$. This gives

$$r(k_1,k_2) = \frac{A_1(k_1,k_2) + A_2(k_1,k_2)}{A_1(0,0) + A_2(0,0)}.$$

Details are given in Kimura and Weiss (1965) on the derivation of the asymptotic

behaviour of $r(k_1, k_2)$ when the distance sampling is large, or more precisely when $k_1 + k_2 \to \infty$. It can be shown that asymptotically $A_2(k_1, k_2)$ becomes negligible compared to $A_1(k_1, k_2)$, which gives the leading term in the expansion of $r(k_1, k_2)$. We have

$$A_1(k_1, k_2) \sim \frac{1}{\pi\sqrt{2\mathcal{M}_{02}\mathcal{M}_{20}}} K_0(\tilde{R}\sqrt{2\mu}),$$

where

$$\mathcal{M}_{np} := \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} i^n j^p m_{ij}, \qquad \tilde{R} := \sqrt{\frac{k_1^2}{\mathcal{M}_{20}} + \frac{k_2^2}{\mathcal{M}_{02}}},$$

and K_0 is the modified Bessel function of the second kind. $\tilde{R}$ is a distance which is adapted to the geometry of the model. As $K_0(x) \sim \sqrt{\pi/2x}e^{-x}$ when $x \to \infty$, we deduce that

$$r(k_1, k_2) \sim D_2 \cdot \frac{e^{-\tilde{R}\sqrt{2\mu}}}{\sqrt{\tilde{R}}}.$$

In the case where $\mathcal{M}_{20} = \mathcal{M}_{02} = \mathcal{M}_2$, we denote the Euclidean sampling distance by $R := \sqrt{k_1^2 + k_2^2}$. We obtain then

$$r(k_1, k_2) \sim D_2 \cdot \frac{e^{-R\sqrt{2\mu/\mathcal{M}_2}}}{\sqrt{R}}. \tag{4.15}$$

We can see that this formula is different from (4.14), the one in the one-dimensional case. In dimension one the probability of identity decreases exponentially, whereas in dimension two it decreases faster, of order of an exponential divided by the square root of the distance.

4.1.4 POI: Summary

The results found in the stepping stone model for the probability of identity seem to be very robust. The only assumption required is that the migration kernel has finite second moments. Thus equations (4.14) and (4.15) describe a universal behaviour

within the stepping stone model. We will see in the next section that this behaviour is also observed in Malécot's model, which increases the interest of this result.

4.2 Malécot's formula

4.2.1 Model and result

We consider a population of individuals living in a region A of $\mathbb{R}^2$, with at every point $P \in A$ a population density $\delta(P)$. An individual born at a point P has a probability $f(P,Q)dS_Q$ of giving birth to an individual in the area dS_Q centered around the point Q. Calling $g(P,Q)$ the probability density that the parent of an individual born at Q was born at P, by conditioning, we have:

$$g(P,Q) = \frac{\delta(P)f(P,Q)}{\displaystyle\int_A \delta(E)f(E,Q)dS_E}.$$

We study a single genetic locus which has two alleles a and A. We denote by μ the probability of mutations, that is the sum of the probabilities for the mutations $a \to A$ and $A \to a$. We suppose the existence of a stationary state for the population, and we want to calculate at this stationary state the probability $\phi(C,D)$ that two individuals sampled from locations $C, D \in A$ are identical by descent. To calculate this probability, we can establish an equation thanks to a recurrence which is similar in spirit to that which we obtained for the stepping stone model. Starting from two individuals in C and D, we condition on the location of their parents, and we obtain

$$\begin{aligned}\phi(C,D) =&(1-\mu)^2 \int_A\int_A \phi(E,F)g(E,C)g(F,D)dS_E dS_F + \\ &(1-\mu)^2 \int_A \frac{1-\phi(E,E)}{\delta(E)} g(E,C)g(E,D)dS_E.\end{aligned} \tag{4.16}$$

A particular version of this model is when the population density is constant, and

the distribution $g(C, D)$ depends only on the distance CD. More specifically, g is chosen to be Gaussian. Actually these assumptions can't hold together, as we will see in section §4.2.2.

For the following calculations we drop Malécot's notation and rewrite equation (4.16) as

$$\begin{aligned}\phi(y) = (1-\mu)^2 \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} \phi(y+x-z)g(x)g(z)\,dxdz \\ + (1-\mu)^2\,\frac{1-\phi(0)}{\delta}\int_{\mathbb{R}^2} g(x)g(y-x)\,dx,\end{aligned} \tag{4.17}$$

where $g(x) = 1/\sqrt{2\pi\sigma^2}\,\exp(-x^2/2\sigma^2)$. We can solve this equation using the Fourier transform. First we introduce the function f defined by

$$f(y) = \frac{1-\phi(0)}{\delta}\int_{\mathbb{R}^2} g(x)g(y-x)\,dx$$

For a given function ϕ we denote $\hat{\phi}(\omega)$ its Fourier transform. Equation (4.17) becomes

$$\hat{\phi}(\omega) = \ (1-\mu)^2\left(\hat{f}(\omega) + (2\pi)^2\hat{g}(-\omega)\hat{g}(\omega)\hat{\phi}(\omega)\right)$$

which gives immediately

$$\hat{\phi}(\omega) = \frac{(1-\mu)^2\hat{f}(\omega)}{1-(1-\mu)^2(2\pi)^2\hat{g}(\omega)\hat{g}(-\omega)}.$$

Thanks to the Gaussian assumption, explicit calculations are possible, and by inverting the Fourier transform we obtain that

$$\phi(y) = \frac{1-\phi(0)}{\delta 2\pi\sigma^2}K_0\left(\frac{\|y\|}{\sigma}\sqrt{2\log\left(\frac{1}{1-\mu}\right)}\right),$$

where K_0 is the modified Bessel function of the second kind.

As we can see, this formula presents a very serious problem for small values of y. Indeed, $\phi(y) \to \infty$ when $y \to 0$, although it's a probability! This simply reveals very clearly some underlying mathematical problems in the hypotheses made by Malécot.

However, if we look at the behaviour when the sampling distance $\|y\|$ is large and the mutation rate μ is small, we find that

$$\phi(y) \sim \frac{1-\phi(0)}{\delta\pi(2^5\sigma^3\mu)^{1/4}} \cdot \frac{e^{-\|y\|\sqrt{2\mu/\sigma^2}}}{\sqrt{\|y\|}}, \tag{4.18}$$

which is exactly the same asymptotic formula as for the stepping stone model (4.15). The coefficient in the exponential is the same in both cases.

4.2.2 The pain in the torus

If we review the reasoning made to reach Malécot's formula, we identify two implicit hypotheses:

- The individuals reproduce independently according to a critical Poisson branching mechanism with Gaussian spatial dispersion.
- The population has a constant spatial density, and is stationary under this reproduction mechanism.

Felsenstein (1975) is the first paper which describes the incompatibility of these two assumptions. It explains the *clumping* phenomenon, that is the fact that in a spatial branching process, the population concentrates on small and highly populated areas, called clumps. As time goes on, these clumps get more and more populated, and the distance between them tends to infinity. This prevents the population from reaching a spatial equilibrium.

Sawyer (1976) gives a general analysis of the phenomenon. It is shown in this paper that for a spatial branching process with migration, there is clumping if the migration motion is a recurrent diffusion. This is the case for the Brownian motion in two dimensions, which explains why this happens in Malécot's model.

The issues illustrated in Malécot's formula are therefore very general. What is missing in this model is some population regulation that prevents the formation of clumps. In the Stepping Stone model, this regulation was forced because we assumed a constant population size in every deme. What we need in continuous space is to construct a reproduction mechanism that depends on the local population density. Traditionally, doing so leads to complex models for which we are not able to describe the genealogical structure. For many years, a continuous space model with a tractable genealogical structure has been elusive. However, Barton et al. (2010a) provides a breakthrough by proposing a radically different way of incorporating population regulation on a continuum, in the setting of the *Spatial Λ-Fleming-Viot model*.

4.3 The Spatial Λ-Fleming-Viot process

4.3.1 A prelimiting particles model

One simple way to introduce population regulation is to move away from the concept of branching processes. The structure is the following. We consider a discrete population constituted of haploid organisms evolving in $\mathbb{R}^d$. Due to limited resources, the individuals compete for survival. As a consequence, every time there are births at a specific location, some of the neighbours die simultaneously in the process of competition. The key point is the spatial link between births and deaths, which brings the local regulation in this model, and therefore avoids clumping. Each event of simultaneous births and deaths is called a *reproduction event*. This mechanism of reproduction is the central

idea that can be enriched with much more structure. We present one of the simplest versions of this family of models.

We consider a discrete population of individuals living on $\mathbb{R}^d$. We start from an arbitrary spatial distribution for these individuals at time 0. At every reproduction event, one considers the individuals inside a random disc centered around $x \in \mathbb{R}^d$ with radius r. One of these individuals is chosen at random to be the mother for this event. Each of the individuals dies with a probability u (which depends on r). At the same time, the mother gives birth to a Poisson number of children who are scattered uniformly on the disc. Mathematically, this dynamics is described as follows. Let $\mu(dr)$ be a possibly infinite measure on $(0, \infty)$, and for each $r \in (0, \infty)$ let $\nu_r(du)$ be a probability measure on $[0, 1]$. Let Π be a Poisson Point Process on $\mathbb{R}_+ \times \mathbb{R}^d \times \mathbb{R}_+ \times [0, 1]$ with rate $dt \otimes dx \otimes \mu(dr) \otimes \nu_r(du)$. Then, for every point (t, x, r, u) of the point process Π:

- At time t, draw a disc of radius r centered around x.
- If the disc is empty, nothing happens.
- If not, then:
  - i) Pick uniformly an individual in the disc, she will be the mother in this reproduction event.
  - ii) At the same time the mother gives birth to a Poisson number of individuals with mean $\lambda\, u\, V(r)$, uniformly scattered on the disc, with the same type as their mother.

The quantity $V(r)$ is the volume of a ball with radius r. We denote by $(X_t, t \geq 0)$ the process defined by this dynamics, where X_t is the point measure that gives the position of all individuals at time t.

This description of the model is not a complete mathematical construction. One

needs to prove that such a Poisson construction actually leads to a well defined infinite dimensional Markov process. The first step of the mathematical work consists in proving the existence of this process under suitable hypotheses. In particular one has to put constraints on the measure $\mu(dr)$ so that the process does not"explode". This is achieved in Berestycki et al. (2009), where the existence of the process is proved if the following condition is fulfilled:

$$\int_{[0,\infty)}\int_{[0,1]} ur^d(1+r^d)\nu_r(du)\mu(dr) < \infty.$$

We also need to show that this model satisfies our main requirement, namely that there is no clumping. This requirement is met, as an ergodic theorem is proved in Berestycki et al. (2009). It states that for λ sufficiently large, the process converges to a stationary distribution, which rules out the possibility of clumping.

A significant limitation of this model is that the genealogical structure is not easy to describe. One would like the genealogy to be obtained simply by sampling individuals from the population and reversing the history of these lineages. Unfortunately, it is not that simple, because we need to know the past history of individuals outside our sample.

This difficulty is the main motivation for another version of the model, which is obtained as a limit when we let $\lambda \to \infty$. We analyse this limiting model in the next section.

4.3.2 The Spatial ΛFV model

The original idea in extending the previous model consists of identifying the limiting process when we let λ, the density of births at every reproduction event, tend to infinity. The results on this convergence have not been published yet, and are still the object

of ongoing work by Etheridge and Kurtz. However, it is possible to independently construct the limiting process.

4.3.2.1 Dynamics of the limiting model

For such an infinite population, the model describes the genetic composition of the population at every point. For this we suppose that each individual has a type taken from a set K, for example $K = [0, 1]$. At time t the population is represented by a set $\{\rho(t, x, \cdot),\ x \in \mathbb{R}^d\}$ of probability measures on K which give the distribution of the types at every point x. The dynamics of the process is almost the same as in the case of the particles model. We keep the same Poisson Point Process Π on $\mathbb{R}_+ \times \mathbb{R}^d \times \mathbb{R}_+ \times [0, 1]$ with rate $dt \otimes dx \otimes \mu(dr) \otimes \nu_r(du)$. Then, for every point (t, x, r, u) of the point process Π:

i) At time t, draw a disc of radius r centered around x.

ii) Select uniformly at random a point z in the disc. This will be the source of the reproduction event.

iii) Select at random a type k according to the probability distribution $\rho(t-, z, \cdot)$

iv) Update the population after the reproduction event: for every point y in the disc,

$$\rho(t, y, \cdot) = (1 - u)\rho(t-, y, \cdot) + u\delta_k \tag{4.19}$$

At every reproduction event, a proportion u of the individuals at every point inside the disc will be replaced by individuals of type k. This limiting process has the same features as the particles model, but we can easily describe its genealogy.

If we sample n individuals from the population, the genealogy is built thanks to a realisation of the Poisson Point Process Π. Consider a point $(t, x, r, u) \in \Pi$, and a

point z sampled uniformly on the circle with radius r and centre x. Tracing the lineages backward in time, the dynamics is the following. If there are k lineages inside the circle, mark each lineage with a 1 with probability u. All the lineages marked with a 1 are moved to the new location z.

– If only one lineage is marked, it moves to the location z, and this is a pure migration event.

– If several lineages are marked, this a coalescent event with migration. If more than two lineages are marked, it is a multiple merger.

Thus, we have obtained a spatial version of the Λ coalescent processes studied in the previous chapters.

4.3.3 Formal construction

So far, we have given only a description of the process. The formal construction is far from trivial, and is achieved in Barton et al. (2010b). We briefly mention the method. The first step is to define the SΛFV process. The most natural way is to characterise it thanks to its infinitesimal generator.

The random variable X_t is a collection of probability measures on $[0,1]$. Therefore, we consider a collection $(\rho_x, x \in \mathbb{R}^d)$ a probability measures on $[0,1]$, and we take test functions of the form

$$\widetilde{I}_n\big(\rho; \psi, (\chi)_{1\leq i\leq n}\big) = \int_{(\mathbb{R}^d)^n} \psi(x_1, \ldots, x_n) \prod_{i=1}^{n} \langle\, \chi_i\,,\, \rho_{x_i} \rangle \, dx_1 \ldots dx_n, \tag{4.20}$$

where ψ is an integrable continuous function on $(\mathbb{R}^d)^n$, and the χ_i's are continuous functions on $[0,1]$.

Definition 4.1. *The infinitesimal generator $\mathcal{L}$ of the SΛFV process is given by*

$$\begin{aligned}
&\mathcal{L}\widetilde{I}_n(\rho;\psi,(\chi)_{1\le i\le n})\\
=\ &\int_{\mathbb{R}^d}\int_0^\infty\int_0^1\int_{B(y,R)}\int_0^1\int_{(\mathbb{R}^d)^n}\sum_{I\subset\{1,\dots,n\}}\Big(\prod_{j\notin I}\mathbb{1}_{B(y,r)^c}(x_j)\langle\chi_j\,,\,\rho_{x_j}\rangle\Big)\\
&\qquad\times\Big(\prod_{i\in I}\mathbb{1}_{B(y,r)}(x_i)\Big)\times\Big[\prod_{i\in I}\langle\chi_i\,,\,(1-u)\rho_{x_i}+u\delta_k\rangle-\prod_{i\in I}\langle\chi_i\,,\,\rho_{x_i}\rangle\Big]\\
&\qquad\qquad\psi(x_1,\dots,x_n)\,dx_1\dots dx_n\,\rho_z(dk)\frac{dz}{V(r)}\nu_r(du)\,\mu(dr)\,dy.
\end{aligned}\tag{4.21}$$

The main point is to identify necessary conditions on the intensity $\nu_r(du)\mu(dr)$ under which there exists a Markov process $(X_t)_{t\ge0}$ with generator given by (4.21). The approach is to construct the forwards-in-time process using the backwards-in-time genealogy. Because the genealogy involves only a finite number of lineages, it can be directly constructed. A theorem due to Evans (1997) allows to use this finite particles genealogy to construct the full forward-in-time process.

Theorem 4.2 (Barton et al. (2010b))**.** *Under the condition*

$$\int_{[0,\infty)}\int_{[0,1]}ur^d\nu_r(du)\mu(dr)<\infty,\tag{4.22}$$

there exists a Markov process $(X_t)_{t\ge0}$ with generator (4.21).

4.3.4 Probability of identity

4.3.4.1 Probability of identity

To establish the equation for the probability of identity, we sample two individuals separated by $x\in\mathbb{R}^2$, and we consider the time of their most recent common ancestor τ_x. We write $\widetilde{\phi}_\theta(x):=\mathbb{E}[\exp(-\theta\tau_x)]$ the Laplace transform of τ_x. If we consider

the same model with mutations arising independently at rate $\theta/2$, then $\widetilde{\phi}_\theta(x)$ is the probability of identity by descent for two individuals separated by x in the model with mutations.

If we look backward in time at the previous event, there are four possibilities. Either only one of the individuals was hit by a reproduction event (in which case her lineage jumps), or both of them are hit (so the two lineages coalesce), or a mutation appears in one of the lineages, or nothing happens because the reproduction event didn't hit anyone. After reordering the different terms, we obtain an integral equation. The problem with the equation we obtain is that it is not tractable, and even approximate calculations do not seem possible. This is why Barton et al. (2010a) develop another model for which the calculations are easier.

4.3.4.2 Generalization/simplification: Gaussian case

In the previous version, at every point (t, x, r) of the point process Π, the parent was chosen uniformly on a disc $D(x, r)$ centered around x with a radius r, and for each individual in $D(x, r)$ the probability to be affected by the reproduction event was u. This reproduction mechanism is changed in the following way.

Consider a reproduction event centered on x. In the previous situation, the fraction of the population modified at location y was given by

$$u\mathbb{1}_{\|y-x\|<r}.$$

Now, we replace this uniform density on a disc by a Gaussian kernel, given by

$$w(x, y) = u \exp\left(\frac{-\|y-x\|^2}{(2\theta)^2}\right).$$

The second modification concerns the choice of the parent. Before, it was sampled uniformly on the disc with centre x and radius r. This time, we sample the parent

according to another Gaussian kernel, given by

$$v(x,z) = u \exp\left(\frac{-\|y-x\|^2}{(2\alpha\theta)^2}\right),$$

with $\alpha > 0$. With this notation, once the parent location has been sampled according to the kernel $v(x,z)$, we sample the parental type according to $\rho(t-,z,.)$, and we update the population according to an equivalent of equation (4.19) given by

$$\forall y \in \mathbb{R}^d, \quad \rho(t,y,\cdot) = (1-w(x,y))\rho(t-,y,\cdot) + w(x,y)\,\delta_k \tag{4.23}$$

This new model has the same structure as the previous one, but is more amenable to calculations. We denote by $\phi(x)$ the the probability of identity for two individuals sampled at distance x. Similarly to what was done for Malécot's formula, we can obtain an integral equation and approximate its solution for large sampling distances using a Fourier transform. We obtain that

$$\phi(x) \sim \frac{u}{1+\alpha^2} K_0\left(\frac{x}{\sigma/\sqrt{2\mu}}\right)$$

where K_0 is the modified Bessel function of the second kind. We notice that it is the same behaviour as in Malécot's formula and the stepping stone model in two dimension.

4.4 Conclusion

This chapter allowed us to present a few models used in mathematical genetics to model the influence of space. Spatial structure is a very large subject, and we just introduced the minimal notions necessary to present the SΛFV process and understand its importance.

A fundamental feature of this model, that we have not mentioned yet, is its ability to incorporate large scale reproduction events. We refer to Barton et al. (2010b) for a

detailed study of the impact of these large events on the genealogy.

The SΛFV model is relatively new, but there is a growing literature around it. Our main contribution in this respect is the analysis of the geographical dispersal of a new allelic type, and this is the object of the next Chapter.

Chapter 5

Range expansion in the SΛFV process

5.1 Introduction

5.1.1 Motivation

In this chapter we investigate the fate of a new allele arising by mutation in a population described by a Spatial Λ-Fleming Viot process on $\mathbb{R}^d$, $d \geq 1$. More precisely, we want to know how far this newly created allele, say of type *Red*, is going to spread. We focus on the dynamics of the *Red* type, so we can collapse any type space K to the special case $K = \{Red, Blue\}$. If $(\rho_t)_{t\geq 0}$ is the SΛFV we consider, then it is straightforward to see that $(X_t)_{t\geq 0} := (\{\rho_t(x, \{Red\}), x \in \mathbb{R}^d)_{t\geq 0}$ is itself a Markov process. We consider X_t to be function from $\mathbb{R}^d$ to $[0, 1]$, where $X_t(x)$ is the fraction of *Red* population at location $x \in \mathbb{R}^d$. We introduce now some notation we are going to use during the rest of this chapter.

Notation 5.1. • *We denote by* $B(x, r) := \{z \in \mathbb{R}^d, \|z - x\| \leq r\}$ *the closed ball of centre* $x \in \mathbb{R}^d$ *and radius* $r > 0$. *Its volume and surface area as a function of the*

radius r are given respectively by $V(r) := \delta r^d$ and $S(r) := \gamma r^{d-1}$.

- *For every Lebesgue-measurable subset $K \subseteq \mathbb{R}^d$, we denote its volume by $|K|$, and for all $r \in \mathbb{R}$, $K^r := \{x \in \mathbb{R}^d : \inf_{z \in K}(\|z - x\|) \leq r\}$.*
- *For any given function $f : \mathbb{R}^d \to \mathbb{R}$, we denote its support by Supp(f).*
- *We define $\mathcal{S}_c$ to be the set of Borel measurable functions $f : \mathbb{R}^d \to [0, 1]$ with compact support. We endow $\mathcal{S}_c$ with the uniform norm.*
- *For $K \subset \mathbb{R}^d$, δ_K is the function from $\mathbb{R}^d$ to $\{0, 1\}$ such that $\delta_K(x) = 1$ if and only if $x \in K$.*

For the sake of simplicity, we will work with a model slightly different from the one described in Barton et al. (2010b). We change step ii) of the algorithm given in §4.3.2, and instead of sampling the point z uniformly on $B(c, r)$, we simply take $z = c$. Furthermore, we are going to work in the special case where the radius of the event and the proportion affected by the event are fixed, that is $\mu(dr) = \delta_R$ and $\nu_r(du) = \delta_U$ for some constants $R > 0$ and $0 < U < 1$. The condition $U < 1$ is just to simplify the notation. In the next sections we are going to focus on the increase of the range of the population at time t, that is $\bigcup_{0 \leq s \leq t} \text{Supp}(X_s)$. It turns out that when $U < 1$, the range coincides with Supp(X_t). All the results that follow hold in the case $U = 1$ if we substitute *the range* for the *support* of X_t.

The algorithm describing the dynamics of $(X_t)_{t \geq 0}$ is the following. Consider Π the Poisson Point Process on $\mathbb{R}_+ \times \mathbb{R}^d \times [0, 1]$ with rate $dt \otimes dc \otimes dv$. Then, for every point (t, c, v) of Π,

i) Draw a ball $B(c, R)$ of radius R centered around c.

ii) If $X_{t-}(c) \geq v$, then the parent is red, and for every point $y \in B(c, R)$,

$$X_t(y) = (1 - U)X_{t-}(y) + U.$$

iii) If $X_{t-}(c) < v$, then the parent is blue, and for every point $y \in B(c,R)$,

$$X_t(y) = (1-U)X_{t-}(y).$$

We can formally characterise the process $(X_t)_{t\geq 0}$ by providing its generator. First, we introduce new test functions $I_n(\,\cdot\,;\psi)$ of the form

$$I_n\big(f;\psi\big) = \int_{(\mathbb{R}^d)^n} \psi(x_1,\ldots,x_n) \prod_{i=1}^{n} f(x_i)\, dx_1 \ldots dx_n, \tag{5.1}$$

where ψ is a function from $(\mathbb{R}^d)$ to $\mathbb{R}$ such that $\int |\psi(x_1,\ldots,x_n)| dx_1 \ldots dx_n < \infty$, and f is a function from $(\mathbb{R}^d)^n$ to $[0,1]$ corresponding to the value taken by X_t at the point where it is evaluated. The generator $\mathcal{L}$ of the process $(X_t)_{t\geq 0}$ is given by

$$\begin{aligned}\mathcal{L}I_n(f;\psi) = \int_{\mathbb{R}^d} \int_{(\mathbb{R}^d)^n} \sum_{I\subset\{1,\ldots,n\}} &\Big(\prod_{j\notin I} \mathbb{1}_{B(y,R)^c}(x_j) f(x_j)\Big) \times \Big(\prod_{i\in I} \mathbb{1}_{B(y,R)}(x_i)\Big) \\ &\times \Bigg[f(y) \Big(\prod_{i\in I} \big((1-U)f(x_i)+U\big) - \prod_{i\in I} f(x_i)\Big) \\ &\quad + \big(1-f(y)\big) \Big(\prod_{i\in I} (1-U) f(x_i) - \prod_{i\in I} f(x_i)\Big)\Bigg] \\ &\psi(x_1,\ldots,x_n)\, dx_1 \ldots dx_n\, dy.\end{aligned} \tag{5.2}$$

5.1.2 Dynamics of the process

Intuitively, one expects the *Red* population to ultimately become extinct, because it is competing against an infinite *Blue* population. But the interesting question is *how* is it going to disappear. The first scenario one can think of is the following:

Scenario 1

The population becomes extinct in the sense that

$$\forall x \in \mathbb{R}^d,\ X_t(x) \longrightarrow 0 \quad \text{as } t \to \infty,$$

but the population keeps spreading infinitely far, meaning that the surface area of the support of X_t tends to infinity as $t \longrightarrow \infty$. This scenario is for example compatible with the formation of an *invasion front*, that is the presence of high frequencies near the boundary of the support of X_t that allow the *Red* population to be sampled. Another possibility is the dynamics of *thin oil film spreading*, that is the frequencies are always small, apart from certain events where the population reproduces and spreads. The frequencies are large enough so that the *Red* population is sampled infinitely often.

This scenario shows the complexity of the spatial dynamics taking place. To gain some intuition, we consider the situation where there is no space. This corresponds to the ΛFV processes we studied in Chapters 2 and 3, when we take the Λ measure to be $\Lambda(du) = u^2\delta_U(du)$. In this case, the ΛFV process is simply a continuous time Markov chain. The embedded Markov chain $(Z_n)_{n\geq 0}$ for the frequency of the *Red* population in this non spatial model is described by:

$$\begin{cases} Z_{n+1} = (1-U)Z_n + U\,\varepsilon_{n+1}, \\ \varepsilon_{n+1} \,|\, Z_n \ \sim\ \mathrm{B}(Z_n), \end{cases}$$

where $\mathrm{B}(Z_n)$ is a Bernoulli distribution with parameter Z_n. It is straightforward to show that $(Z_n)_{n\geq 0}$ is a nonegative martingale, and therefore converges almost surely. As a consequence, ε_n converges almost surely to 0 or 1.

The key fact is that after some finite random time, ε_n will remain constant equal to 0 or 1. It means that almost surely in finite time, either the *Red* or the *Blue* population will be the only one to keep reproducing. The reason behind this phenomenon is

because if the frequency Z_n is not sampled a few times in a row, it is going to decrease geometrically, and become rapidly too small to be sampled again. The same reasoning applies to the *Blue* population.

The situation without space gives the idea of an additional scenario.

Scenario 2

The population becomes extinct in the sense that

$$\sup_{x\in\mathbb{R}^d} X_t(x) \longrightarrow 0 \quad \text{as } t \to \infty,$$

and there exists a random finite set $B \subset \mathbb{R}^d$ and an almost surely finite random time T such that

$$\forall t > T, \text{Supp}(X_t) = B \quad \text{a.s.} \tag{5.3}$$

As we have seen, this scenario can happen if after the time T the *Red* population does not reproduce anymore.

The question is to determine the probability of each of the scenarios, and whether any other scenarios take place with positive probability. Our main result, Theorem 5.1.3, shows that actually scenario 2 happens with probability one. Somehow, in order to prove this result, we need to show that Scenario 1, or any other scenario, happens with probability zero. But given the complexity of the spatial interaction, using upper bounds to prove that these various scenarios have probability zero does not seem to be feasible.

A good idea is to use a martingale argument similar to the one in the non spatial case. In this case, the martingale in question is the total population size, that is the integral of X_t over the whole space. It turns out that the proof is quite straightforward in the case of discrete space. However, with the continuous space, the geometry becomes much more complicated. The proof of scenario 2 in the continuous space setting is the

object of this Chapter.

5.1.3 Main result

Theorem 5.2. *Let $(X_t)_{t\geq 0}$ be a Markov process with generator (5.2). Suppose $X_0 = f$ is deterministic and with bounded support. Then, there exists a random finite set $B \subset \mathbb{R}^d$, and an almost surely finite random time T such that*

$$\forall t > T,\ Supp(X_t) = B \quad a.s. \tag{5.4}$$

Furthermore,

$$\sup_{z\in B} X_t(z) \to 0 \quad a.s.\ \ as\ t \to \infty \tag{5.5}$$

Thanks to the specific choice of $\mu(dr)$ and $\nu_r(du)$ we took, we can study the process $(X_t)_{t\geq 0}$ by relying on the fact that it jumps at finite rate. We can see this by taking $X_0 \in \mathcal{S}_c$ and considering the time of the first change

$$T_1 := \inf\{t > 0 : X_t \neq X_0\}.$$

We can express T_1 as

$$T_1 = \inf\{t > 0 : (t, c, v) \in \Pi, B(c, R) \cap \mathrm{Supp}(X_0) \neq \emptyset\}.$$

Using the definition of Π, the quantity

$$N_t = \#\{(s, c, v) \in \Pi : s \leq t, B(c, R) \cap \mathrm{Supp}(f)\}$$

is a Poisson random variable with parameter

$$\begin{aligned}
&\int_0^t \int_{\mathbb{R}^d} \int_{[0,1]} \mathbb{1}_{\{B(c,R)\cap \mathrm{Supp}(f)\neq\emptyset\}}\, dv\, dc\, ds \\
&= t \int_{\mathbb{R}^d} \mathbb{1}_{\{c\in \mathrm{Supp}(f)^R\}} dc \\
&= t\, |\mathrm{Supp}(f)^R|.
\end{aligned}$$

Therefore N_t is almost surely finite. As a consequence, conditionally on $X_0 = f$, T_1 is exponentially distributed with parameter

$$\lambda(f) := |\mathrm{Supp}(f)^R|. \tag{5.6}$$

We denote by C_1 and V_1 the random variables such that $(T_1, C_1, V_1) \in \Pi$. Using properties of the Poisson point processes, we can argue that

– C_1 is uniform on $\mathrm{Supp}(f)^R$,

– and V_1 is uniform on $[0, 1]$.

Therefore we can construct the process $(X_t)_{t\geq 0}$ as a continuous time Markov chain with total jump rate out of $f \in \mathcal{S}_c$ given by (5.6), and an embedded Markov chain $(Y_n)_{n\geq 0}$ whose transition probabilities are derived from the conditional distribution of C_1 and V_1. We define precisely this Markov chain in §5.2.

If we denote by Δ_n the support of Y_n, our first objective is to prove that there exists some almost surely finite random time κ such that

$$\forall n > \kappa, \quad \Delta_n = B. \tag{5.7}$$

This result is stated in Proposition 5.7. The technical condition $Y_0 = a\delta_{B(C_0,r_0)}$ is easily

removed in Proposition 5.26. Once we have proved (5.7), we can construct explicitly $(X_t)_{t\geq 0}$ thanks to $(Y_n)_{n\geq 0}$ and prove that it does not explode. This is achieved in Proposition 5.28. This allows to transfer the result (5.7) to the process $(X_t)_{t\geq 0}$, which is achieved in §5.5.3, and completes the proof of Theorem 5.2.

Most of the work in this chapter consists in proving (5.7). The major tool is a martingale argument developed in §5.4.1. The total mass of the population, that is the integral of the function Y_n over $\mathbb{R}^d$, is a non negative martingale, and therefore converges almost surely. However, to be able to exploit this convergence, one needs to overcome the difficulty brought by the geometry in continuous space. In particular, one has to handle the dependency between the frequencies at two points x and y in $\mathbb{R}^d$. This crucial step is achieved by considering local averages of the function Y_n, that is the integral of Y_n on a ball with radius R and centre x. In §5.3, we study the properties of the function

$$\Phi_n(x) := \int_{B(x,R)} Y_n(z)\, dz.$$

In §5.4, we combine these geometric properties with the martingale argument to prove (5.7).

5.2 A discrete time Markov chain

5.2.1 Construction

Definition 5.3. *Consider $R > 0$ and $0 < U < 1$. Let y be an $\mathcal{S}_c$-valued random variable. We construct simultaneously random sequences $(C_n)_{n\geq 1}$ and $(V_n)_{n\geq 1}$, a filtration*

$(\mathcal{P}_n)_{n\geq 0}$, and an $\mathcal{S}_c$-valued Markov chain $(Y_n)_{n\geq 0}$, using the following recurrence:

$$\begin{cases} Y_0 = y, \\ \mathcal{P}_0 := \sigma(Y_0), \\ \Delta_0 = \mathit{Supp}(Y_0). \end{cases} \tag{5.8}$$

and for $n \geq 0$,

- $\mathcal{P}_n := \sigma(C_1, \dots, C_n, V_1, \dots, V_n, Y_0, \dots, Y_n)$,
- *conditionally on $\mathcal{P}_n$, C_{n+1} is uniform on Δ_n^R,*
- *V_{n+1} is distributed uniformly on $[0,1]$, independently from $\mathcal{P}_n$ and C_{n+1},*
- *Y_{n+1} is given by the formula*

$$Y_{n+1}(\cdot) = Y_n(\cdot) + U\,\delta_{B(C_{n+1},R)}(\cdot)\left(\mathbb{1}_{\{V_{n+1}\leq Y_n(C_{n+1})\}} - Y_n(\cdot)\right) \tag{5.9}$$

- $\Delta_{n+1} = \mathit{Supp}(Y_{n+1})$.

We introduce the notation $\varepsilon_{n+1} := \mathbb{1}_{\{V_{n+1}\leq Y_n(C_{n+1})\}}$. In particular, the trajectories of $(\Delta_n)_{n\geq 0}$ are given by

$$\Delta_n = \Delta_0 \cup \bigcup_{\substack{1\leq k\leq n,\\ \varepsilon_k = 1}} B(C_k, R). \tag{5.10}$$

Finally, we denote the natural filtration of $(Y_n)_{n\geq 0}$ by $\mathcal{G}_n := \sigma(Y_0, \dots, Y_n)$.

Remark 5.4. *We recall that Δ_n^R is the R-expansion of the set Δ_n.*

For each reproduction event, the random variables C_n and V_n correspond respectively to the centre of the event and the uniform sampling v, whereas R and U are the radius of the event and the proportion of the population that is modified. The random variable ε_{n+1} indicates the types of the parent chosen.

Notation 5.5. *If $\varepsilon_{n+1} = 1$, we say that the event is a* positive sampling event*, because the total red population increases, whereas when $\varepsilon_{n+1} = 0$ we say that it is a* negative sampling event*.*

Equation (5.10) shows that if the cluster Δ_n is to increase, the minimum requirement is that there is a positive sampling.

Remark 5.6. *Expression (5.10) is true because we assumed $U < 1$. In this case, the support and the range of the process coincide. Once a region is occupied by the* Red *population it remains occupied at every finite time. Therefore the cluster Δ_n never shrinks. In the case where $U = 1$ expression (5.10) will remain true if Δ_n is taken to be the range.*

5.2.2 Result of the cluster convergence

The following result is the expression of our main result in the discrete time setting, with the temporary technical condition $Y_0 = a\,\delta_{B(C_0,r_0)}$.

Proposition 5.7. *Suppose $Y_0 = a\,\delta_{B(C_0,r_0)}$, where $a \in [0,1]$, $b, r_0 > 0$ and $C_0 \in \mathbb{R}^d$. Then, there exists an almost surely finite random time κ such that*

$$\forall n > \kappa, \quad \varepsilon_n = 0. \tag{5.11}$$

Therefore, there exists an almost surely bounded random set $B \in \mathbb{R}^d$ such that

$$\forall n > \kappa, \quad \Delta_n = B. \tag{5.12}$$

Most of the remaining of this chapter is devoted to proving this Proposition. We first investigate the geometric properties of the model in the next section.

5.3 Geometry

This section is central, as it allows us to summarise all the geometric properties we need to prove our result, which is the major difficulty.

Remark 5.8. *From now on, unless specified otherwise, we suppose that* $Y_0 = a\,\delta_{B(C_0,r_0)}$.

5.3.1 Y_n is piecewise constant

Definition 5.9. *For notational convenience, we introduce the sequence* $(R_n)_{n\geq 0}$ *such that*

$$\begin{cases} R_0 = r_0, \\ R_n = R, \quad n \geq 1. \end{cases} \tag{5.13}$$

Lemma 5.10. *For every* $n \geq 0$*, for every* $\zeta \subset \{0, \dots, n\}$*, consider the set* $A_{n,\zeta}$ *defined by*

$$A_{n,\zeta} := \big(\bigcap_{m\in\zeta} B(C_m, R_m) \big) \setminus \big(\bigcup_{\substack{j\leq n,\\ j\notin\zeta}} B(C_j, R_j) \big). \tag{5.14}$$

The function Y_n *can be written as*

$$Y_n = \sum_{\zeta\subset\{0,\dots,n\}} \alpha_{n,\zeta}\,\delta_{A_{n,\zeta}}, \tag{5.15}$$

where the sets $A_{n,\zeta}$ *are all disjoint for a given* n*, and* $\alpha_{n,\zeta} \geq 0$.

Proof. We first introduce the shorter notation

$$B_j := B(C_j, R_j),\ j \geq 0.$$

The fact that the sets $A_{n,\zeta}$ are all disjoint for a given n is straightforward, so we just

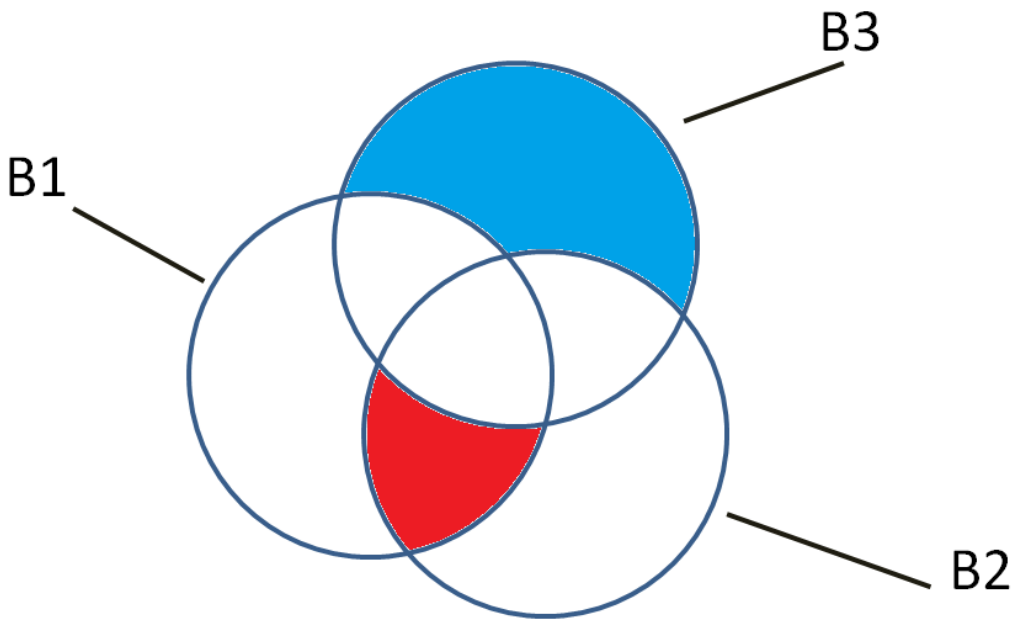

Figure 5.1: Geometrical structure presented in Lemma 5.10 - Consider the three balls B_1, B_2 and B_3. The red region can be expressed as $(B_1 \cap B_2) \setminus B_3$, whereas the blue region can be expressed as $B_3 \setminus (B_1 \cup B_2)$, which illustrates formula (5.14).

need to prove (5.15). But before that, we need to show that

$$\bigcup_{\zeta \subset \{0,\dots,n\}} A_{n,\zeta} = \bigcup_{i=1}^{n} B_i, \tag{5.16}$$

and we proceed by induction on n. The statement is true for $n = 0$. Suppose now that it is true for some given $n \geq 0$. Let $\zeta' \subset \{0, \dots, n+1\}$.

- If $\zeta' = \{n+1\}$, then $A_{n+1,\zeta'} = B_{n+1} \setminus \bigcup_{\zeta \subset \{0,\dots,n\}} A_{n,\zeta}$.
- If $(n+1) \in \zeta'$, and $\zeta' \neq \{n+1\}$ then there exists $\zeta \subset \{0, \dots, n\}$ such that

$$A_{n+1,\zeta'} = B_{n+1} \cap A_{n,\zeta}.$$

– If $(n+1) \notin \zeta'$, then there exists $\zeta \subset \{0,\dots,n\}$ such that

$$A_{n+1,\zeta'} = A_{n,\zeta} \setminus B_{n+1}.$$

Therefore, we see that

$$\begin{aligned}\bigcup_{\zeta' \subset \{0,\dots,n+1\}} A_{n+1,\zeta'} &= \Big(\bigcup_{\zeta \subset \{0,\dots,n\}} B_{n+1} \cap A_{n,\zeta} \Big) \cup \Big(\bigcup_{\zeta \subset \{0,\dots,n\}} A_{n,\zeta} \setminus B_{n+1} \Big) \\ &\qquad \cup \Big(B_{n+1} \setminus \bigcup_{\zeta \subset \{0,\dots,n\}} A_{n,\zeta} \Big) \\ &= \Big(\bigcup_{\zeta \subset \{0,\dots,n\}} A_{n,\zeta} \Big) \cup \Big(B_{n+1} \setminus \bigcup_{\zeta \subset \{0,\dots,n\}} A_{n,\zeta} \Big) \\ &= B_{n+1} \cup \Big(\bigcup_{\zeta \subset \{0,\dots,n\}} A_{n,\zeta} \Big),\end{aligned}$$

and the statement is proven using the inductive hypothesis.

We can return to the proof of expression (5.15). We need to show that for all $n \geq 0$, $\zeta \subset \{0,\dots,n\}$,

$$\Big(x \notin \bigcup_{\zeta \subset \{0,\dots,n\}} A_{n,\zeta} \Big) \text{ implies } \Big(Y_n(x) = 0 \Big), \tag{5.17}$$

and

$$\Big(x, y \in A_{n,\zeta} \Big) \text{ implies } \Big(Y_n(x) = Y_n(y) \Big). \tag{5.18}$$

We prove (5.17) by induction on n, and we use (5.16). Statement (5.17) is satisfied for $n = 0$ because $Y_0 = a\,\delta_{B(C_0,r_0)}$. Suppose now that it is true for some $n \geq 0$, and consider $x \notin \cup_{i=1}^{n+1} B_i$. In particular, $x \notin B_{n+1}$, and using the dynamics equation

(5.9), we find that $Y_{n+1}(x) = Y_n(x)$. Because $x \notin \cup_{i=1}^{n} B_i$, we can use the inductive hypothesis, and we obtain that $Y_n(x) = 0$, which proves (5.17).

To prove (5.18), we also use induction. It is true for $n = 0$. Suppose (5.18) is satisfied for a given $n \geq 0$. Consider $\zeta' \subset \{0, \dots, n+1\}$, and take $x, y \in A_{n+1,\zeta'}$.

– If $\zeta' = \{n+1\}$, then $A_{n+1,\zeta'} = B_{n+1} \setminus \bigcup_{\zeta \subset \{0,\dots,n\}} A_{n,\zeta}$, and in particular

$$x, y \notin \bigcup_{\zeta \subset \{0,\dots,n\}} A_{n,\zeta}.$$

We have $\delta_{B_{n+1}}(x) = \delta_{B_{n+1}}(y) = 1$, and using (5.17) we see that $Y_n(x) = Y_n(y) = 0$.

– If $(n+1) \in \zeta'$ and $\zeta' \neq \{n+1\}$, there exists $\zeta \subset \{0, \dots, n\}$ such that

$$A_{n+1,\zeta'} = B_{n+1} \cap A_{n,\zeta}.$$

Therefore $\delta_{B_{n+1}}(x) = \delta_{B_{n+1}}(y) = 1$, and using the inductive hypothesis, we have $Y_n(x) = Y_n(y)$.

– If $(n+1) \notin \zeta'$, then there exists $\zeta \subset \{0, \dots, n\}$ such that

$$A_{n+1,\zeta'} = A_{n,\zeta} \setminus B_{n+1}.$$

In this case $\delta_{B_{n+1}}(x) = \delta_{B_{n+1}}(y) = 0$, and using the inductive hypothesis, we have $Y_n(x) = Y_n(y)$.

We can express Y_{n+1} using the dynamics equation (5.9), and we obtain:

$$\begin{cases} Y_{n+1}(x) = Y_n(x) + U\,\delta_{B_{n+1}}(x)\,(\varepsilon_{n+1} - Y_n(x)) \\ Y_{n+1}(y) = Y_n(y) + U\,\delta_{B_{n+1}}(y)\,(\varepsilon_{n+1} - Y_n(y)). \end{cases}$$

We have proved that for any choice of ζ', all the terms in the above equations are the same for x and y, therefore $Y_{n+1}(x) = Y_{n+1}(y)$, and we have proved (5.18). □

5.3.2 Variation of the local average

A central tool for the rest of the work is the average of the function Y_n on a ball of radius R and centre $x \in \mathbb{R}^d$. We introduce the following function:

Definition 5.11.

$$\Phi_n(x) := \int_{B(x,R)} Y_n(z)\, dz. \tag{5.19}$$

The main result of this section is the following.

Proposition 5.12. *For every $x, y \in \mathbb{R}^d$,*

$$\Phi_n(y) - \Phi_n(x) \leq \|y - x\| S(R). \tag{5.20}$$

The rest of this section is devoted to proving this inequality. For this we need to introduce some auxiliary functions.

Definition 5.13. *Given $x, y \in \mathbb{R}^d$, for every $n \geq 0$, we define*

$$\begin{aligned} \Lambda_n^{x,y} : [0, \|y - x\|] &\longrightarrow [0, \infty) \\ t \qquad &\longmapsto \Lambda_n(t) := \Phi_n\Big(x + t\, \frac{y - x}{\|y - x\|} \Big). \end{aligned} \tag{5.21}$$

The key property for the proof of Proposition 5.12 is the following.

Proposition 5.14. *$\Lambda_n^{x,y}$ is a continuous, piecewise differentiable function. Moreover, for every point t where $\Lambda_n^{x,y}$ is differentiable, we have:*

$$\frac{d\Lambda_n^{x,y}}{dt}(t) \leq S(R). \tag{5.22}$$

Proof. We first prove that $\Lambda_n^{x,y}$ is continuous and that there is at most a finite number J of points $t_1 < \cdots < t_J$ at which $\Lambda_n^{x,y}$ is not differentiable. Thanks to equation (5.15) from Lemma 5.10, we see that Φ_n is given by

$$\begin{aligned}\Phi_n(x) &= \int_{B(x,R)} \sum_{\zeta \subset \{0,\dots,n\}} \alpha_{n,\zeta} \delta_{A_{n,\zeta}}(z)\, dz \\ &= \sum_{\zeta \subset \{0,\dots,n\}} \alpha_{n,\zeta} |B(x,R) \cap A_{n,\zeta}|,\end{aligned}$$

therefore $\Lambda_n^{x,y}(t)$ is given by

$$\Lambda_n^{x,y}(t) = \sum_{\zeta \subset \{0,\dots,n\}} \alpha_{n,\zeta}\, |B\big(\, x + t\, \frac{y-x}{\|y-x\|}\, , R\big) \cap A_{n,\zeta}|.$$

We simplify the notation by introducing $B_j := B(C_j, R_j),\ j \geq 0$ and

$$x_t := x + t\, \frac{y-x}{\|y-x\|}.$$

The definition (5.14) of the sets $A_{n,\zeta}$ and the inclusion-exclusion formula allow us to prove that for each $\zeta \subset \{0,\dots,n\}$, there exists a $\beta_{n,\zeta} \in \mathbb{R}$ such that

$$\Lambda_n^{x,y}(t) = \sum_{\zeta \subset \{0,\dots,n\}} \beta_{n,\zeta}\, |B\big(\, x_t\, , R\big) \cap \big(\bigcap_{m \in \zeta} B_m\big)|.$$

Given $\zeta \subset \{0,\dots,n\}$, consider the function

$$\begin{aligned} H_\zeta : [0, \|y-x\|] &\longrightarrow \mathbb{R} \\ t \qquad &\longmapsto |\, B(x_t, R) \cap \bigcap_{m \in \zeta} B_m\, |. \end{aligned}$$

The continuity of H_ζ follows from the continuity of the function $t \mapsto x_t$, and this shows that $\Lambda_n^{x,y}$ is continuous.

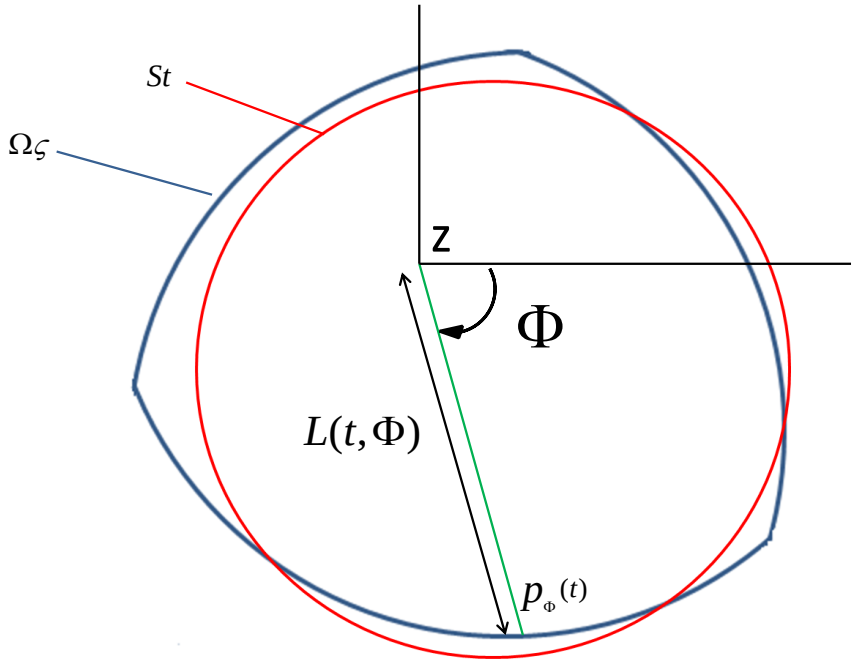

Figure 5.2: Illustration of $p_\phi(t)$ in dimension 2 - The point z is chosen arbitrarily inside the intersection and is the new origin. The angle ϕ is defined in the local polar coordinate system. For a given angle ϕ the point $p_\phi(t)$ is situated on the boundary that is closest to z. For some values of ϕ it belongs to the mobile sphere S_t, and for other values of ϕ it belongs to the static surface Ω_ζ. This is why we partition the set $[0, 2\pi]$.

The set $\bigcap_{m\in\zeta} B_m$ is convex, therefore there exist t_1, t_2 such that

$$H_\zeta(t) > 0 \Leftrightarrow t_1 < t < t_2. \tag{5.23}$$

Consider t such that $t_1 < t < t_2$. Thanks to (5.23), this means we can choose a point z belonging to the interior of $B(x_t, R) \cap \bigcap_{m\in\zeta} B_m$. Because $B(x_t, R) \cap \bigcap_{m\in\zeta} B_m$ is convex, we can express its area in d-hyperspherical coordinates with z as the new origin. Given angular coordinates $\phi := (\phi_1, \ldots, \phi_{d-1})$, we denote by $p_\phi(t)$ the unique point of the surface of $B(x_t, R) \cap \bigcap_{m\in\zeta} B_m$ with angular coordinates ϕ, and $L(t, \phi)$ the distance between z and $p_\phi(t)$. We have:

$$H_\zeta(t) = \int_0^\pi \ldots \int_0^\pi \int_0^{2\pi} \int_0^{L(t,\phi)} r^{d-1} \sin^{d-2}(\phi_1) \sin^{d-3}(\phi_2) \ldots \sin(\phi_{d-2}) dr\, d\phi_{d-1} \ldots d\phi_1.$$

We denote by Ω_ζ the boundary of $\bigcap_{m\in\zeta} B_m$, and by S_t the sphere of centre x_t and radius R. We can find a partition $\Xi_1, \ldots, \Xi_K$ of the space

$$[0, \pi]^{d-2} \times [0, 2\pi]$$

such that

$$\begin{cases} \forall \phi \in \Xi_j,\ p_\phi(t) \in \Omega_\zeta, \\ \text{or} \\ \forall \phi \in \Xi_j,\ p_\phi(t) \in S_t. \end{cases}$$

Therefore we can write $H_\zeta(t)$ as

$$H_\zeta(t) = \sum_{j=1}^{K} \int_{\Xi_j} \int_0^{L(t,\phi)} r^{d-1} \sin^{d-2}(\phi_1) \sin^{d-3}(\phi_2) \ldots \sin(\phi_{d-2}) dr\, d\phi.$$

Thanks to this representation, we see that a sufficient condition for H_ζ to be dif-

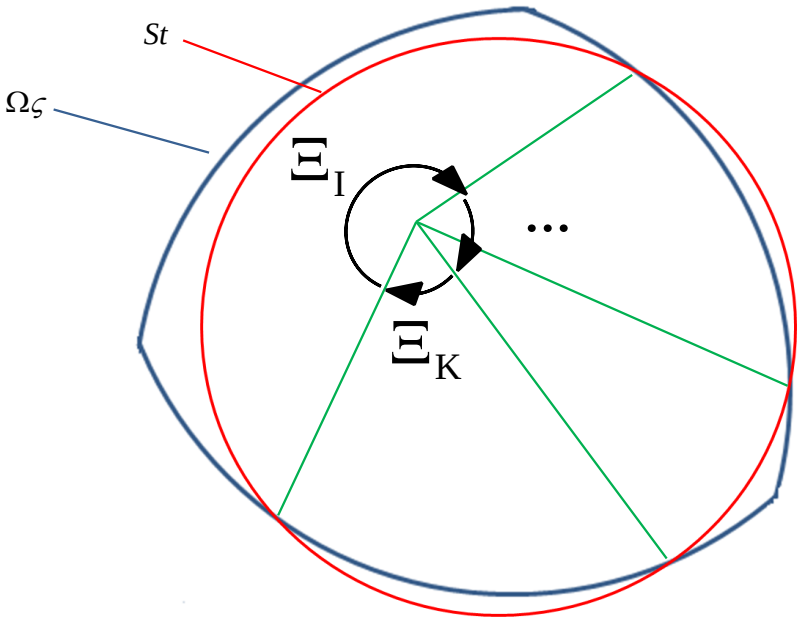

Figure 5.3: Partition Ξ in dimension 2 - The green lines define cones converging at z, such that the intersections of each of these cones with the surface is either entirely included in S_t, or entirely included in Ω_ς. This partition simplifies the study of the dynamics of $L(t, \phi)$ when S_t moves.

ferentiable at t, $t_1 < t < t_2$, is that for every ϕ in the interior of every Ξ_j, the function $t \mapsto L(t,\phi)$ is differentiable at t. In this case, the derivative is given by

$$\frac{dH_\zeta(t)}{dt} = \sum_{j=1}^{K} \int_{\Xi_j} \frac{\partial L}{\partial t}(t,\phi)\, L(t,\phi)^{d-1} \sin^{d-2}(\phi_1) \sin^{d-3}(\phi_2) \ldots \sin(\phi_{d-2})\, d\phi.$$

We focus now on the differentiability of $t \mapsto L(t,\phi)$, where ϕ belongs to the interior of Ξ_j for some j.

Suppose first that $p_\phi(t) \in \Omega_\zeta$. Because the function $t \mapsto x_t$ is continuous, and because z is a fixed point, there exists $h > 0$ such that for every $u \in (t-h, t+h)$, $p_\phi(u) \in \Omega_\zeta$. Therefore, there exists $h > 0$ such that for every $u \in (t-h, t+h)$, $L(u,\phi) = L(t,\phi)$, and $t \mapsto L(t,\phi)$ is differentiable at t.

In the case where $p_\phi(t) \in S_t$, by the same continuity argument, we obtain that for every $u \in (t-h, t+h)$, $p_\phi(u) \in S_u$. Because the distance between z and the projection of z on S_t along the angle ϕ is differentiable, we conclude that $t \mapsto L(t,\phi)$ is also differentiable in this case.

We have proved that for all $t \neq t_1, t_2$, H_ζ is differentiable at t. Given that

$$\Lambda_n^{x,y}(t) = \sum_{\zeta \subset \{0,\ldots,n\}} \beta_{n,\zeta}\, H_\zeta(t),$$

we conclude that there is at most a finite number of points at which $\Lambda_n^{x,y}$ is not differentiable, and this proves the first part of the Proposition.

We need now to show the upper bound (5.22) for the derivative. Suppose $\Lambda_n^{x,y}$ is

differentiable at t. By construction, for all $h \geq 0$, $B(x_{t+h}, R) \subset B(x_t, R+h)$, so

$$\begin{aligned}\Lambda_n^{x,y}(t+h) &= \int_{B(x_{t+h},R)} Y_n(z)dz \\ &\leq \int_{B(x_t,R+h)} Y_n(z)dz \\ &\leq |B(x_t, R+h)|.\end{aligned}$$

Therefore, given that $\Lambda_n^{x,y}(t) = |B(x_t, R)|$, we have

$$\frac{\Lambda_n^{x,y}(t+h) - \Lambda_n^{x,y}(t)}{h} \leq \frac{|B(x_t, R+h)| - |B(x_t, R)|}{h}.$$

Taking the limit as $h \to 0$, we obtain

$$\frac{d\Lambda_n^{x,y}}{dt}(t) \leq \frac{dV(R)}{dR} = S(R), \tag{5.24}$$

where $V(R)$ is the volume of a ball of radius R, and $S(R)$ its surface area. □

Proof of Proposition 5.12. We saw in the previous proof that there is only a finite number of points $t_1, \ldots, t_J$ at which $\Lambda_n^{x,y}$ is not differentiable. By continuity of $\Lambda_n^{x,y}$,

and using Proposition 5.14,

$$\begin{aligned}
&\Phi_n(y) - \Phi_n(x) \\
=&\Lambda_n^{x,y}(\|y-x\|) - \Lambda_n^{x,y}(0) \\
=&\Lambda_n^{x,y}(\|y-x\|) - \Lambda_n^{x,y}(t_J) + \sum_{j=1}^{J-1} \Lambda_n^{x,y}(t_{j+1}) - \Lambda_n^{x,y}(t_j) \\
&\qquad + \Lambda_n^{x,y}(t_1) - \Lambda_n^{x,y}(0) \\
\leq& S(R)(\|y-x\| - t_J) + \sum_{j=1}^{J-1} S(R)(t_{j+1} - t_j) \\
&\qquad + S(R)(t_1 - 0) \\
\leq& \|y-x\|\, S(R).
\end{aligned}$$

□

5.4 Finitely many positive sampling events

5.4.1 A martingale argument

Definition 5.15. *We denote by M_n the total mass of Y_n, that is*

$$M_n = \int_{\mathbb{R}^d} Y_n(z)dz.$$

Lemma 5.16. *The change of the total mass $M_{n+1} - M_n$ is given by*

$$M_{n+1} - M_n = U\left(\varepsilon_{n+1}V(R) - \Phi_n(C_{n+1})\right) \tag{5.25}$$

Proof. Using (5.9),

$$\begin{aligned} M_{n+1} - M_n &= \int_{\mathbb{R}^d} \Big(Y_{n+1}(z) - Y_n(z)\Big) dz \\ &= \int_{\mathbb{R}^d} U\, \delta_{B(C_{n+1},R)}(z)\, \big(\varepsilon_{n+1} - Y_n(z)\big) dz \\ &= U \int_{B(C_{n+1},R)} \Big(\varepsilon_{n+1} - Y_n(z)\Big) dz \\ &= U\, \big(\varepsilon_{n+1} V(R) - \Phi_n(C_{n+1})\big) \end{aligned}$$

□

Proposition 5.17. *$(M_n)_{n\geq 0}$ is a discrete time nonnegative $(\mathcal{G}_n)_{n\geq 0}$-martingale.*

Proof. Thanks to Lemma 5.16 and Definition 5.3, we can calculate explicitly $\mathbb{E}[\, M_{n+1} - M_n \,|\, \mathcal{G}_n\,]$ as follows:

$$\begin{aligned} &\mathbb{E}\big[M_{n+1} - M_n \,|\, \mathcal{G}_n\big] \\ =&\mathbb{E}\left[U\, \mathbb{1}_{\{V_{n+1}\leq Y_n(C_{n+1})\}} \int_{B(C_{n+1},R)} dz - U \int_{B(C_{n+1},R)} Y_n(z) dz \,|\, \mathcal{G}_n\right] \\ =&U \int_{\mathbb{R}^d} \int_{[0,1]} \left[\mathbb{1}_{\{v\leq Y_n(c)\}} \int_{B(c,R)} dz - \int_{B(c,R)} Y_n(z)\, dz\right] dv\, \frac{\mathbb{1}_{c\in\Delta_n^R}}{|\Delta_n^R|}\, dc \\ =&U \int_{\mathbb{R}^d} \left[Y_n(c) \int_{B(c,R)} dz - \int_{B(c,R)} Y_n(z) dz\right] \frac{\mathbb{1}_{c\in\Delta_n^R}}{|\Delta_n^R|} dc \\ =&\frac{U}{|\Delta_n^R|} \left[\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathbb{1}_{z\in B(c,R)}\, \mathbb{1}_{c\in\Delta_n^R}\, Y_n(c) dz\, dc \right. \\ &\qquad \left. - \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathbb{1}_{z\in B(c,R)}\, \mathbb{1}_{c\in\Delta_n^R}\, Y_n(z) dz\, dc\right]. \end{aligned}$$

In particular, for every $f \in \mathcal{S}_c$,

$$\begin{aligned}
&\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathbb{1}_{\{z\in B(c,R)\}}\, \mathbb{1}_{\{c\in \mathrm{Supp}(f)^R\}}\, f(c)dz\, dc \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathbb{1}_{\{c\in B(z,R)\}}\, f(c)dz\, dc \quad \text{since } c \in B(z,R) \Leftrightarrow z \in B(c,R) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathbb{1}_{\{z\in B(c,R)\}}\, f(z)dc\, dz, \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathbb{1}_{\{z\in B(c,R)\}}\, \mathbb{1}_{\{c\in \mathrm{Supp}(f)^R\}}\, f(z)dc\, dz,
\end{aligned}$$

so $\mathbb{E}[\, M_{n+1} - M_n \,|\, \mathcal{G}_n \,] = 0$, which shows that $(M_n)_{n\geq 0}$ is a martingale. □

Definition 5.18. *Let $\alpha > 0$ be a real number such that $0 < \alpha < UV(R)/2$. We then define*

$$\tau_\alpha := \inf\{p \geq 0 : \forall n \geq p, \quad |M_{n+1} - M_n| < \alpha\}. \tag{5.26}$$

In particular, τ_α is not a stopping time, but this is not going to be an issue in what follows.

Proposition 5.19. *The random time τ_α is a.s. finite, and $\forall n > \tau_\alpha$,*

$$\begin{cases} \Phi_n(C_{n+1}) < \dfrac{\alpha}{U} & \text{if } \varepsilon_{n+1} = 0, \\ \Phi_n(C_{n+1}) > V(R) - \dfrac{\alpha}{U} & \text{if } \varepsilon_{n+1} = 1. \end{cases} \tag{5.27}$$

Proof. We know that M_n is a nonnegative martingale, so it converges almost surely when $n \to \infty$. Therefore τ_α is almost surely finite, and by definition, $\forall n > \tau_\alpha, \quad |M_{n+1} - M_n| < \alpha$. Using Lemma 5.16, we observe that

$$|M_{n+1} - M_n| = \begin{cases} U\, \Phi_n(C_{n+1}) & \text{if } \varepsilon_{n+1} = 0, \\ U\, \big(V(R) - \Phi_n(C_{n+1})\big) & \text{if } \varepsilon_{n+1} = 1, \end{cases}$$

which concludes the proof. □

5.4.2 Forbidden region

The concept of a forbidden region will allow us to treat probabilistically the geometric properties established in §5.3.

Lemma 5.20. *For $n \geq 0$,*

$$\begin{aligned} &\{n \geq \tau_\alpha\} \cap \{\varepsilon_k = 1 \text{ infinitely often}\} \\ \subset\ &\{n \geq \tau_\alpha\} \cap \bigcap_{j=n}^{\infty} \{\sup \Phi_j > V(R) - \alpha/U\} \end{aligned} \tag{5.28}$$

Proof. Take $j \geq \tau_\alpha$ such that $\sup \Phi_j \leq V(R) - \alpha/U$. Using Proposition 5.19, this implies that $\varepsilon_{j+1} = 0$. In particular, $\sup \Phi_{j+1} \leq \sup \Phi_j \leq V(R) - \alpha/U$. By induction, we just showed that

$$\begin{cases} j \geq \tau_\alpha \\ \sup \Phi_j \leq V(R) - \alpha/U \end{cases} \implies \begin{cases} j \geq \tau_\alpha \\ \forall k \geq j, \varepsilon_k = 0. \end{cases}$$

The contrapositive of this implication allows to conclude this proof. □

Definition 5.21. *We define the* forbidden region F_n *to be*

$$F_n := \{x \in \mathbb{R}^d : \frac{\alpha}{U} \leq \Phi_n(x) \leq V(R) - \frac{\alpha}{U}\}. \tag{5.29}$$

We also introduce the quantity

$$\psi := V\left(\frac{V(R) - 2\,\alpha/U}{S(R)}\right). \tag{5.30}$$

The reason for the name *forbidden region* is motivated by the following lemma, which tells us that after the time τ_α, if the local averages are always too high, then the points C_{n+1} are forbidden from falling in the region F_n. Furthermore, this lemma

provides a lower bound on the surface area of F_n.

Lemma 5.22.

$$\{j \geq \tau_\alpha\} \cap \{\sup \Phi_j > V(R) - \alpha/U\}$$

$$\subset \{C_{j+1} \notin F_j, |F_j| \geq \psi\} \tag{5.31}$$

Proof. An immediate consequence of Proposition 5.19 is

$$j \geq \tau_\alpha \Rightarrow C_{j+1} \notin F_j.$$

More work is required to obtain the lower bound for the area of the forbidden region. We first define $P_j := \{x \in \mathbb{R}^d : \Phi_j(x) \geq V(R) - \alpha/U\}$. If we assume $\sup \Phi_j > V(R) - \alpha/U$, then P_j is nonempty. We can then take a point $y \in P_j$.

The function Φ_j is continuous, and Δ_j is finite, so the region $N_j := \{x \in \mathbb{R}^d : \Phi_j(x) \leq \alpha/U\}$ is infinite. Indeed, for every x at a distance from Δ_j larger than R, $\Phi_j(x) = 0$. In particular, for a large enough positive number $\bar{R}$, we can consider Γ the sphere of radius $\bar{R}$ and centre y, and the ball $B(y, \bar{R})$ such that

$$\begin{cases} \Delta_j \subset B(y, \bar{R}), \\ \Gamma \subset N_j. \end{cases} \tag{5.32}$$

For $x, y \in \mathbb{R}^d$, we denote by $[x, y]$ the line-segment between x and y. We need the following lemma:

Lemma 5.23. *The point $y \in P_j$ being fixed, for every point $x \in \Gamma$, we can find two*

points x_0, y_0 *such that:*

$$\begin{cases} [x_0, y_0] \subset F_j, \\ [x_0, y_0] \subset [x, y], \\ \|y_0 - x_0\| \geq \dfrac{V(R) - 2\,\alpha/U}{S(R)}. \end{cases} \tag{5.33}$$

By integrating the result of Lemma 5.23 over all the points $x \in \Gamma$, we find that the volume of F_j is larger than the volume of a ball of radius $(V(R) - 2\,\alpha/U)/S(R)$, hence the result. □

Proof of Lemma 5.23. The function $\Lambda_j^{x,y}$ defined in (5.21) is continuous, with $\Lambda_j^{x,y}(0) = 0$ and $\Lambda_j^{x,y}(\|y - x\|) \geq V(R) - \alpha/U$, so there are two points $t_1, t_2 \in [0, \|y - x\|]$ such that $\Lambda_j^{x,y}([t_1, t_2]) = [\alpha/U, V(R) - \alpha/U]$. By application of the continuous function $t \longrightarrow x + t\,(y - x)/(\|y - x\|)$, this means that there are two points $x_0, y_0 \in \mathbb{R}^d$ such that

- $\Phi_j(x_0) = \alpha/U$,
- $\Phi_j(y_0) = V(R) - \alpha/U$,
- $\forall z \in (x_0, y_0), \Phi_j(z) \in (\alpha/U, V(R) - \alpha/U)$.

The last statement is just the fact that $(x_0, y_0) \subset F_j$. By using Corollary 5.12, we find that

$$\begin{aligned} \|y_0 - x_0\|\, S(R) &\geq \Phi_j(y_0) - \Phi_j(x_0) \\ &\geq V(R) - 2\,\alpha/U. \end{aligned}$$

□

We reach now the main point of this section, which is an upper bound for the probability that infinitely many positive sampling events take place.

Proposition 5.24.

$$\begin{aligned}&\mathbb{P}\big(\varepsilon_k = 1 \text{ \textit{infinitely often}}\big)\\ &\leq \sum_{l=0}^{\infty} \mathbb{P}\big(\bigcap_{j=l}^{\infty}\{C_{j+1} \notin F_j, |F_j| \geq \psi\}\big)\end{aligned} \tag{5.34}$$

Proof. As $\tau_\alpha < \infty$ a.s.,

$$\begin{aligned}&\mathbb{P}\big(\varepsilon_k = 1 \text{ i.o.}\big)\\ &= \mathbb{P}\big(\{\tau_\alpha < \infty\} \cap \{\varepsilon_k = 1 \text{ i.o.}\}\big)\\ &= \sum_{n=0}^{\infty} \mathbb{P}\big(\{\tau_\alpha = n\} \cap \{\varepsilon_k = 1 \text{ i.o.}\}\big)\\ &\leq \sum_{n=0}^{\infty} \mathbb{P}\big(\{n \geq \tau_\alpha\} \cap \{\varepsilon_k = 1 \text{ i.o.}\}\big)\end{aligned} \tag{5.35}$$

We can write $\{n \geq \tau_\alpha\} = \bigcap_{j=n}^{\infty}\{j \geq \tau_\alpha\}$ Using this in (5.31), we obtain

$$\begin{aligned}&\{n \geq \tau_\alpha\} \cap \bigcap_{j=n}^{\infty}\{\sup \Phi_j > V(R) - \alpha/U\}\\ \subset &\bigcap_{j=n}^{\infty}\{C_{j+1} \notin F_j, |F_j| \geq \psi\}.\end{aligned}$$

Combining this with (5.28), we have

$$\begin{aligned}&\{n \geq \tau_\alpha\} \cap \{\varepsilon_k = 1 \text{ i.o.}\}\\ \subset &\bigcap_{j=n}^{\infty}\{C_{j+1} \notin F_j, |F_j| \geq \psi\},\end{aligned} \tag{5.36}$$

and the result follows.

□

5.4.3 Finitely many positive sampling events

Proposition 5.25.

$$\mathbb{P}\big(\varepsilon_k = 1 \text{ } \textit{infinitely often}\big) = 0.$$

Proof. Using Proposition 5.24, we simply need to prove that for every $l \geq 0$,

$$\mathbb{P}\big(\bigcap_{j=l}^{\infty}\{C_{j+1} \notin F_j, |F_j| \geq \psi\}\big) = 0.$$

By monotone convergence, we have

$$\mathbb{P}\big(\bigcap_{j=l}^{\infty}\{C_{j+1} \notin F_j, |F_j| \geq \psi\}\big) = \lim_{n\to\infty} \mathbb{P}\big(\bigcap_{j=l}^{n}\{C_{j+1} \notin F_j, |F_j| \geq \psi\}\big).$$

We work on the right-hand side by first conditioning on Δ_0:

$$\mathbb{P}\big(\bigcap_{j=l}^{n}\{C_{j+1} \notin F_j, |F_j| \geq \psi\}\big) = \mathbb{E}\left[\mathbb{P}\big(\bigcap_{j=l}^{n}\{C_{j+1} \notin F_j, |F_j| \geq \psi\} \,|\, \Delta_0\big)\right].$$

We then condition on all but the last reproduction events:

$$\begin{aligned}
&\mathbb{P}\Big(\bigcap_{j=l}^{n}\{C_{j+1} \notin F_j, |F_j| \geq \psi\} \,|\, \Delta_0\Big) \\
=&\,\mathbb{E}\Big[\mathbb{1}\{C_{l+1} \notin F_l, |F_l| \geq \psi\} \cdots \mathbb{1}\{C_n \notin F_{n-1}, |F_{n-1}| \geq \psi\} \\
&\qquad \mathbb{P}\Big(C_{n+1} \notin F_n, |F_n| \geq \psi \,|\, \Delta_0, F_l, C_{l+1}, ..., F_{n-1}, C_n\Big) \,|\, \Delta_0\Big].
\end{aligned} \tag{5.37}$$

We can calculate the last term by conditioning on F_n and Δ_n:

$$\begin{aligned}
&\mathbb{P}\big(C_{n+1} \notin F_n, |F_n| \geq \psi \,|\, \Delta_0, F_l, C_{l+1}, ..., F_{n-1}, C_n, F_n, \Delta_n\big) \\
&= \mathbb{1}_{|F_n|\geq\psi}\mathbb{P}\big(C_{n+1} \notin F_n \,|\, F_n, \Delta_n\big) \\
&= \mathbb{1}_{|F_n|\geq\psi}\Big(1 - \frac{|F_n|}{|\Delta_n^R|}\Big) \\
&\leq 1 - \frac{\psi}{|\Delta_n^R|} \\
&\leq 1 - \frac{\psi}{|\Delta_0^R| + nV(2R)}.
\end{aligned} \tag{5.38}$$

The second and third equalities come from the fact that conditionally on Δ_n, C_{n+1} is sampled uniformly from Δ_n^R, independently of the past. The last inequality comes from (5.10):

$$\Delta_n = \Delta_0 \cup \bigcup_{\substack{1\leq k\leq n,\\ \varepsilon_k=1}} B(C_k, R).$$

In particular, it implies that

$$\begin{aligned}
|\Delta_n^R| \leq& |\Delta_0^R| + |B(C_1, R)^R| + \cdots + |B(C_n, R)^R| \\
\leq& |\Delta_0^R| + n\, V(2R).
\end{aligned}$$

Putting inequality (5.38) into (5.37), we obtain the following upper bound:

$$\begin{aligned}
&\mathbb{P}\Big(\bigcap_{j=l}^{n}\{C_{j+1} \notin F_j, |F_j| \geq \psi\} \,|\, \Delta_0\Big) \\
&\leq \Big(1 - \frac{\psi}{|\Delta_0^R| + nV(2R)}\Big)\mathbb{P}\Big(\bigcap_{j=l}^{n-1}\{C_{j+1} \notin F_j, |F_j| \geq \psi\} \,|\, \Delta_0\Big).
\end{aligned} \tag{5.39}$$

Inequality (5.39) provides a recurrence relation, which we can solve immediately to

obtain

$$\mathbb{P}\big(\bigcap_{j=l}^{n}\{C_{j+1}\notin F_j, |F_j|\geq\psi\}\,|\,\Delta_0\big)\leq\prod_{j=l}^{n}\Big(1-\frac{\psi}{|\Delta_0^R|+jV(2R)}\Big).$$

Taking expectation and then the limit as $n\to\infty$, we obtain

$$\begin{aligned}\mathbb{P}\big(\bigcap_{j=l}^{\infty}\{C_{j+1}\notin F_j, |F_j|\geq\psi\}\big)&\leq\lim_{n\to\infty}\mathbb{E}\,\big[\prod_{j=l}^{n}\big(1-\frac{\psi}{|\Delta_0^R|+jV(2R)}\big)\,\big]\\&\leq\mathbb{E}\,\big[\prod_{j=l}^{\infty}\big(1-\frac{\psi}{|\Delta_0^R|+jV(2R)}\big)\,\big].\end{aligned}$$

We rewrite the infinite random product using logarithms:

$$\begin{aligned}&\prod_{j=l}^{\infty}\big(1-\frac{\psi}{|\Delta_0^R|+jV(2R)}\big)\\&=\exp\Big(\sum_{j=l}^{\infty}\log\big(1-\frac{\psi}{|\Delta_0^R|+jV(2R)}\big)\Big).\end{aligned}$$

After observing that

$$\log\big(1-\frac{\psi}{|\Delta_0^R|+jV(2R)}\big)\underset{j\to\infty}{\overset{a.s.}{\sim}}\frac{-\psi/V(2R)}{j},$$

we conclude that the infinite product is almost surely equal to 0, and

$$\mathbb{P}\big(\bigcap_{j=l}^{\infty}\{C_{j+1}\notin F_j, |F_j|\geq\psi\}\big)=0.$$

□

5.5 Proof of the theorems

5.5.1 Proof of Proposition 5.7

We proved in Proposition 5.25 that with probability one, there are only finitely many sampling events. This means that there exists an almost surely finite random time κ such that

$$\forall n > \kappa, \quad \varepsilon_n = 0. \tag{5.40}$$

We recall the dynamics of the cluster Δ_n described by (5.10):

$$\Delta_n = \Delta_0 \cup \bigcup_{\substack{1 \le k \le n, \\ \varepsilon_k = 1}} B(C_k, R).$$

Therefore, if we define $B := \Delta_\kappa$, we have

$$\forall n > \kappa, \quad \Delta_n = B, \tag{5.41}$$

and the proof of Proposition 5.7 is complete.

We can now generalise Proposition 5.7 by removing the technical condition on the starting point and allowing Y_0 to be any function in $\mathcal{S}_c$.

Proposition 5.26. *Suppose $Y_0 = f \in \mathcal{S}_c$. Then, there exists an almost surely finite random time κ such that*

$$\forall n > \kappa, \quad \varepsilon_n = 0. \tag{5.42}$$

Therefore, there exists an almost surely bounded random set $B \in \mathbb{R}^d$ such that

$$\forall n > \kappa, \quad \Delta_n = B. \tag{5.43}$$

Proof. We proceed by coupling Y with a Markov chain $\widetilde{Y}$ with the same transition probabilities, but started from $\widetilde{Y}_0 = \delta_{B(C_0,r_0)}$ such that $Y_0 \leq \widetilde{Y}_0$. We denote the initial conditions by $\widetilde{Y}_0 = \widetilde{f}$ and $Y_0 = f$. We first build $\widetilde{Y}$ as described in Definition 5.3. We then use the sequences $(\widetilde{C}_n)_{n\geq 1}$ and $(\widetilde{V}_n)_{n\geq 1}$ that we used to construct $\widetilde{Y}$ in the following way. First consider the random sequence Y' defined by $Y'_0 = f$, and for $n \geq 0$,

$$Y'_{n+1} = Y'_n + U\,\delta_{B(\widetilde{C}_{n+1},R)}\left(\mathbb{1}_{\{\widetilde{V}_{n+1}\leq Y'_n(\widetilde{C}_{n+1})\}} - Y'_n\right).$$

We can prove by induction that

$$\forall n \geq 0,\ Y'_n \leq \widetilde{Y}_n. \tag{5.44}$$

It is of course true at $n = 0$, and then we just observe that if $Y'_n \leq \widetilde{Y}_n$, then

$$\begin{aligned}
&\widetilde{Y}_{n+1} - Y'_{n+1} \\
=&\left(1 - U\,\delta_{B(\widetilde{C}_{n+1},R)}\right)\left(\widetilde{Y}_n - Y'_n\right) \\
&\qquad + U\,\delta_{B(\widetilde{C}_{n+1},R)}\left(\mathbb{1}_{\{\widetilde{V}_{n+1}\leq \widetilde{Y}_n(\widetilde{C}_{n+1})\}} - \mathbb{1}_{\{\widetilde{V}_{n+1}\leq Y'_n(\widetilde{C}_{n+1})\}}\right) \\
\geq&0.
\end{aligned}$$

We denote by $\widetilde{\Delta}$ and Δ' the respective sequences of supports, and in particular we have proved that

$$\forall n \geq 0,\ \Delta'_n \subset \widetilde{\Delta}_n. \tag{5.45}$$

We define the sequence of $\left(\sigma(\widetilde{\mathcal{P}}_n, f)\right)_{n\geq 0}$-stopping times $(J_n)_{n\geq 0}$ by setting

$$\begin{cases} J_0 = 0, \\ J_{n+1} = \inf\{k > J_n : \widetilde{C}_k \in \left(\Delta'_{k-1}\right)^R\}. \end{cases} \tag{5.46}$$

We now construct $(Y_n)_{n\geq 0}$, $(C_n)_{n\geq 1}$ and $(V_n)_{n\geq 1}$ by taking

$$\begin{cases} Y_n := Y'_{J_n},\ n \geq 0 \\ C_n := \widetilde{C}_{J_n},\ n \geq 1 \\ V_n := \widetilde{V}_{J_n},\ n \geq 1. \end{cases}$$

We denote by Δ_n the support of Y_n, and we define the filtration $(\mathcal{P}_n)_{n\geq 0}$ to be

$$\begin{cases} \mathcal{P}_0 := \sigma(Y_0), \\ \mathcal{P}_n := \sigma(C_1, \dots, C_n, V_1, \dots, V_n, Y_0, \dots, Y_n), \end{cases}$$

By construction, conditionally on $\mathcal{P}_n$, C_{n+1} is distributed uniformly on Δ_n^R. Because V_{n+1} is independent of $\mathcal{P}_n$, we conclude that the law of Y is the one given in Definition 5.3.

Using (5.44), we see that

$$Y_n = Y'_{J_n} \leq \widetilde{Y}_{J_n}. \tag{5.47}$$

We introduce

$$\begin{cases} \widetilde{\varepsilon}_{n+1} := \mathbb{1}_{\widetilde{Y}_n(\widetilde{C}_{n+1}) \geq \widetilde{V}_{n+1}}, \\ \varepsilon_{n+1} := \mathbb{1}_{Y_n(C_{n+1}) \geq V_{n+1}}. \end{cases}$$

Because $\widetilde{f} = a\,\delta_{B(C_0,r_0)}$, we can use Proposition 5.7, and we obtain that there exists $\widetilde{\kappa}$

almost surely finite such that

$$\forall n > \widetilde{\kappa}, \quad \widetilde{\varepsilon}_n = 0. \tag{5.48}$$

In particular, this implies that there exists κ almost surely finite such that

$$\forall n > \kappa, \quad \widetilde{\varepsilon}_{J_n} = 0. \tag{5.49}$$

Combined with (5.47), this implies that

$$\forall n > \kappa, \quad \varepsilon_n = 0, \tag{5.50}$$

and the conclusion follows. □

5.5.2 The continuous time process is non explosive

We are now going to construct explicitly the process $(X_t)_{t\geq 0}$ with generator (5.2) as a continous time Markov chain, by using $(Y_n)_{n\geq 0}$ as the embedded Markov chain.

Definition 5.27. *Consider an i.i.d sequence $(E_1, E_2, \dots)$ of Exp(1) random variables. We define the jump times $(T_0, T_1, \dots)$ by setting $T_0 = 0$ and*

$$T_n = \frac{E_1}{\lambda(Y_0)} + \cdots + \frac{E_n}{\lambda(Y_{n-1})}, \quad n \geq 1, \tag{5.51}$$

where $\lambda(f) := |(Supp(f))^R|$ for $f \in \mathcal{S}_c$.
We can then define a stochastic process $(X_t)_{t\geq 0}$ by setting

$$\forall n \geq 0, \forall t \in [T_n, T_{n+1}), \quad X_t = Y_n. \tag{5.52}$$

We recall that the set $(\mathrm{Supp}(f))^R$ is the R-dilation of the support of f, that is the

set of points at a distance less than R from the support of f. The quantity $\lambda(f)$ is its surface area, and we have seen in §5.1.3 that this is the correct rate at which the process $(X_t)_{t\geq 0}$ jumps out of the state f. This will be verified in Proposition 5.28 by checking that we have the correct generator.

Proposition 5.28. *The process $(X_t)_{t\geq 0}$ constructed in (5.52) is a non-explosive $\mathcal{S}_c$-valued continuous time Markov chain. Moreover, its generator is given by (5.2).*

Proof. The first thing to verify is that X_t is really defined for all nonnegative t. This is equivalent to saying that

$$\mathbb{P}\big[\, T_n \to \infty \text{ as } n \to \infty \,\big] = 1,$$

that is

$$\mathbb{P}\big[\, \sum_{n=1}^{\infty} \frac{E_n}{\lambda(Y_{n-1})} = \infty \,\big] = 1.$$

We show in Proposition 5.7 that $(\Delta_n)_{n\geq 0}$ converges in a finite number of steps to a bounded set B. This means that almost surely, there is a random time κ such that for all $n \geq \kappa$,

$$\lambda(Y_n) = |\mathrm{B}^R|$$

which implies that

$$\sum_{n=\kappa+1}^{\infty} \frac{E_n}{\lambda(Y_{n-1})} = \frac{\displaystyle\sum_{n=\kappa+1}^{\infty} E_n}{|\mathrm{B}^R|} = \infty \text{ a.s.}$$

Therefore $(X_t)_{t\geq 0}$ is a stochastic process defined for all $t \geq 0$. The Markov property is obvious, and this shows that $(X_t)_{t\geq 0}$ is a non-explosive continuous time Markov chain.

Therefore, we can write the generator of $(X_t)_{t\geq 0}$ for functions $G : \mathcal{S}_c \mapsto \mathbb{R}$ as

$$\mathcal{L}G(f) = \int_{(\mathrm{Supp}(f))^R} \int_0^1 G\big[f + U\delta_{B(c,R)}(\mathbb{1}_{v\leq f(c)} - f)\big] - G(f)\, dv\, dc.$$

If we take $G = I_n(\cdot, \psi)$ as defined in (5.1), the generator of $(X_t)_{t\geq 0}$ takes the form (5.2). □

5.5.3 Proof of Theorem 5.2

We have seen in Proposition 5.28 that the process $(X_t)_{t\geq 0}$ is a non explosive continuous time Markov chain. Therefore, the trajectories of $(X_t)_{t\geq 0}$ are completely described by its embedded Markov chain $(Y_n)_{n\geq 0}$.

In particular, for all $n \geq 0$, for all $t \in [T_n, T_{n+1})$, $\mathrm{Supp}(X_t) = \Delta_n$. Using the result from Proposition 5.26, and the the sequence of times $(T_n)_{n\geq 0}$ defined in (5.51), there exists a finite random set $B \subset \mathbb{R}^d$, and an almost surely finite random time $T := T_\kappa$, such that

$$\forall t > T,\ \mathrm{Supp}(X_t) = B \quad \text{a.s.} \tag{5.53}$$

The second point to prove here is the extinction of the population. From Proposition (5.26),

$$\forall n > \kappa, \quad \varepsilon_n = 0. \tag{5.54}$$

This implies that at every point x, the frequency $(X_t(x))_{t\geq T}$ converges geometrically to zero, which concludes the proof.

5.6 Conclusion

Although the SΛFV is constructed in great generality, our study was restricted to the case $\mu(dr) = \delta_R$ and $\nu_r(du) = \delta_U$. Our result holds because the surface area of Δ_n

increases at most linearly with n.

We could imagine extending the same result using practically the same method to the case where

$$\int_{(0,\infty]} r^d \mu(dr) < \infty,$$

because the process still jumps at finite rate, and the surface area of Δ_n is at most of order $n\mathbb{E}(R)$, where R is a realisation of the random radius. The problem is the fact that the radii are random makes the construction of the Markov chain more complicated. Morally the result remains true in this case, but the proof becomes significantly more complicated.

The situation where

$$\int_{(0,\infty]} r^d \mu(dr) = \infty$$

is radically different, because now the process jumps at an infinite rate. The problem is that we do not have a description of the geometry of the process at time $t > 0$. The behaviour is not obvious, and it cannot be simulated. For us this remains an open question, which would certainly require different techniques.

Chapter 6

Conclusion

Our objective in this thesis was to present in a unified manner the models arising in the "Λ" world, that is the Λ-coalescent processes, the Λ-Fleming-Viot processes and the Spatial Λ-Fleming-Viot processes. We wanted to show that their mathematical structure is quite intuitive, especially when studied with the right objects, and that keeping in mind the forward-backward in time duality helps in solving some difficult problems.

Our first result, the fact that the allele frequencies in certain coalescent processes follow a Poisson-Dirichlet distribution, shows that whereas the traditional method for understanding the distribution of allele frequencies is to use the coalescent approach, actually much can be gained by considering the forward-in-time process. Although this process is a "larger" object, tools like the lookdown construction allow one to analyse it in a very tractable manner.

The second result is that it is not that difficult to modify existing statistical methods to take into account Λ-models. This is particularly true for MCMC methods, where one can innovate significantly by simply creating a move that handles some specificity of the data or the model. We proposed a single move, the Merge-Split move, that allows us to incorporate the Λ coalescents into a rich statistical framework.

Finally, our third result was the proof of some non trivial behaviour brought by incorporating this Λ-behaviour of the reproduction mechanism into a spatial setting. The mathematical result that an allele spreads only finitely far certainly has consequences for the spatial distribution of allele frequencies. It would be very interesting to analyse these aspects. However, our work in this respect showed that this kind of models needs a radically different set of analytic methods, as the existing literature on interacting particle systems treats models with a much simpler structure.

We see future development of the work in this thesis in two aspects. On the mathematical side, many very exciting features of the spatial version for the model need to be investigated. For example, we were looking at the initial condition of the process being a bounded region of $\mathbb{R}^d$. Another situation to investigate is when we split the plane in two halves, the left being red and the right blue. One interesting question is how the interface fluctuates as time tends to infinity. Significant asymptotic results in this direction have been obtained in Berestycki et al. (2011b), but a more detailed analysis of the local structure remains open.

On the applied side, we believe there is a real need for completing the work we have initiated concerning the development of Bayesian methods for Λ-coalescent models. There is growing literature on the theoretical justifications of these models given various biological scenarios. What is also needed is a robust and generic tool that would allow us to perform large scale statistical inference. We have proved in a simple setting that there is no real obstruction for producing such a tool.

Still on the applied side, we didn't have time to develop any analysis of the spatial model. We believe that, because of its universal structure, this model can be a framework that could handle a very large number of biological situations. This is a very desirable feature, because it means we can develop generic methods that could be implemented in many different situations. This is particularly relevant because of

the vast literature concerning spatial genetics, and in particular the large number of techniques used in various fields (genetics, ecology, epidemiology, etc...).

References

Arratia, R., Barbour, A. D., and Tavaré, S. (2003). *Logarithmic Combinatorial Structures*. European Mathematical Society. 45

Barton, N. H., Etheridge, A. M., and Kelleher, J. (2010a). A new model for extinction and recolonization in two dimensions: quantifying phylogeography. *Evolution*, 64:2701–2715. ix, 91, 106, 112

Barton, N. H., Etheridge, A. M., and Véber, A. (2010b). A new model for evolution in a spatial continuum. *Electronic Journal of Probability*, 15:162–216. 110, 111, 113, 116

Berestycki, N., Berestycki, J., and Limic, V. (2011a). Asymptotic sampling formulae and particle system representations for Lambda-coalescents . 33

Berestycki, N., Berestycki, J., and Schweinsberg, J. (2007). Beta-coalescents and continuous stable random trees. *Annals of Probability*, 35:1835–1887. 33

Berestycki, N., Berestycki, J., and Schweinsberg, J. (2008). Small time behavior of beta-coalescents. *Annales de l'Institut Henri Poincaré*, 44:214–238. 33

Berestycki, N., Etheridge, A., and Hutzenthaler, M. (2009). Survival, extinction and ergodicity in a spatially continuous population model. *Markov Processes and Related Fields*, 15:265–288. 108

Berestycki, N., Etheridge, A. M., and Véber, A. (2011b). Large scale behaviour of the spatial Lambda-Fleming-Viot process. *http://arxiv.org/abs/1107.4254*. 154

Bertoin, J. (2008). Two-parameter Poisson-Dirichlet measures and reversible exchangeable fragmentation-coalescence processes. *Combinatorics, Probability and Computing*, 17:329–337. 44, 45

Bertoin, J. and Le Gall, J. F. (2003). Stochastic flows associated to coalescent processes. *Probability Theory and Related Fields*, 126:261–288. 29

Birkner, M. and Blath, J. (2008). Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. *Journal of Mathematical Biology*, 57:435–465. 60, 89

Birkner, M., Blath, J., Capaldo, M., Etheridge, A. M., Möhle, M., Schweinsberg, J., and Wakolbinger, A. (2005). Alpha-stable branching and beta-coalescents. *Electronic*

journal of Probability, 10:303–325. 30

Birkner, M., Blath, J., Möhle, M., Steinrücken, M., and Tams, J. (2009). A modified lookdown construction for the Xi-fleming-viot process withmutation and populations with recurrent bottlenecks. *ALEA Latin American Journal of Probability and Mathematical Statistics*, 6:25–61. 30, 32

Donnelly, P. and Kurtz, T. (1996). A countable representation of the Fleming-Viot measure-valued diffusion. *Annals of Probability*, 24:698–742. 23, 30, 32

Donnelly, P. and Kurtz, T. (1999). Particle representations for measure-valued population models. *Annals of Probability*, 27:166–205. 25, 29, 30, 32

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics.*, 161:13071720. 61, 75, 81

Durrett, R. (2004). *Probability models for DNA sequence evolution.* Springer-Verlag. 22

Eldon, B. and Wakeley, J. (2006). Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172:26211733. 60

Eldon, B. and Wakeley, J. (2009). Coalescence times and FST under a skewed offspring distribution among individuals in a population. *Genetics*, 181:615–629. 60

Etheridge, A. M. (2011). *Some Mathematical Models from Population Genetics.* Springer-Verlag. 15

Ethier, S. N. and Kurtz, T. G. (1986). *Markov processes: Characterisation and convergence.* John Wiley and Sons. 13, 15

Ethier, S. N. and Kurtz, T. G. (1994). Convergence to Fleming-Viot processes in the weak atomic topology. *Stochastic Processes and their Applications*, 54:1–27. 24

Evans, S. N. (1997). Coalescing Markov labelled partitions and a continuous sites genetics model with infinitely many types. *Annales de l'Institut Henri Poincaré*, 33:339178. 111

Ewens, W. J. (2004). *Mathematical Population Genetics.* Springer-Verlag. 10

Felsenstein, J. (1975). A pain in the torus: Some difficulties with the model of isolation by distance. *The American Naturalist*, 109:359–368. 105

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368176. 63, 65

Fleming, W. H. and Viot, M. (1979). Some measure-valued Markov processes in population genetics theory. *Indiana University Mathematics Journal*, 28:817–843. 16

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and bayesian model determination. *Biometrika*, 82:711–732. 66, 67

Ignatov, T. S. (1982). On a constant arising in the asymptotic theory of symmetric

groups, and on Poisson-Dirichlet measures. *Theory of Probability and its applications.*, 27:136–147. 45

Kimura, M. (1953). Stepping stone model of population. *Annual Report National Institute of Genetics, Japan*, 3:62–63. viii, 91

Kimura, M. and Weiss, G. H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49:561–576. 92, 94

Kimura, M. and Weiss, G. H. (1965). A mathematical analysis of the stepping stone model of genetic correlation. *Journal of Applied Probability*, 2:129–149. 92

Kingman, J. F. C. (1977). The population structure associated with the Ewens sampling formula. *Theoretical Population Biology*, 11:274–283. 22

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Applications*, 13:235–248. v, 20, 25

Malécot, G. (1948). *Les mathématiques de l'hérédité*. Masson. viii, 91

Pitman, J. (1999). Coalescents with multiple collisions. *Annals of Probability*, 27:18701702. 25

Sagitov, S. (1999). The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability*, 26:1116–1125. 25, 29

Sawyer, S. (1976). Branching diffusion processes in population genetics. *Advances in Applied Probability*, 8:659–689. 105