SORBONNE UNIVERSITÉ

École doctorale de sciences mathématiques de Paris centre

Thèse de doctorat

Discipline : Mathématiques

par

Félix Foutel-Rodier

Scaling limits of branching and coalescing models arising in population biology

dirigée par
Amaury Lambert et Emmanuel Schertzer

Après avis des rapporteurs :

Bénédicte Haas	Sorbonne Paris Nord
Steven Evans	University of Berkeley

Présentée et soutenue publiquement le 15 janvier 2021 devant un jury composé de :

Bénédicte Haas	Sorbonne Paris Nord	Rapportrice
Alison Etheridge	University of Oxford	Examinatrice
Zhan Shi	Sorbonne Université	Examinateur
Julien Berestycki	University of Oxford	Examinateur
Amaury Lambert	Sorbonne Université	Directeur
Emmanuel Schertzer	Sorbonne Université	Directeur

Remerciements

Je tiens tout d'abord à remercier grandement ma rapportrice et mon rapporteur, Bénédicte Haas et Steve Evans, pour avoir accepté de lire ce manuscrit en détail, en un délai aussi court. Je veux aussi remercier chaleureusement Alison Etheridge, Zhan Shi et Julien Berestycki pour avoir accepté de faire partie de mon jury. Je suis honoré par le temps et l'attention que vous avez toutes et tous portés à mon travail.

Il y a deux personnes qui me viennent tout de suite à l'esprit et que je ne pourrai jamais assez remercier. Amaury, merci pour ta générosité, pour ta bienveillance, pour ton engagement sans faille, pour continuer de guider des biologistes vers les mathématiques et de leur faire confiance pour y arriver. Emmanuel, merci pour ton enthousiasme permanent, pour être aussi attentionné, pour tout le temps que tu as passé avec moi, et merci de n'avoir jamais oublié le deuxième tiret de mon nom. Vous avez su me guider au travers de cette thèse à la perfection, et vous resterez pour moi deux modèles, tant mathématiques qu'humains. Cela m'attriste de savoir que pour recevoir cet encadrement exceptionnel, il faut maintenant choisir entre Vienne et Paris, mais je suis sûr que votre duo aura raison de la distance.

Avec deux encadrants aussi formidables vient une équipe tout aussi incroyable. Merci à toutes les smileuseuses et les smileurs que j'ai pu croiser durant ma thèse. Je sais que je ne pourrai jamais retrouver une équipe qui soit si régulière dans les pauses thé, qui aime tant marcher, que ce soit en manifestation ou dans la nature, qui soit si près du SH et avec autant de talent pour les activités manuelles, le kendama ou la musique stochastique. Merci donc à celles et ceux qui sont arrivés il y a un certain temps, Vero, Marc, Jean-Jil, François, Julie, Thomas, Pascal, Florence, Guillaume, François, ou plus récemment, Jasmine, Guillaume, Thomas, Maxime, et sans oublier les deux avec qui j'aurai fait presque toute l'aventure : Pete et Élise, même si je veux bien te concéder pour une fois que tu es arrivée la première. Grâce à vous cela a toujours été un plaisir de venir à SMILE, ce qui est finalement la chose la plus importante de cette thèse.

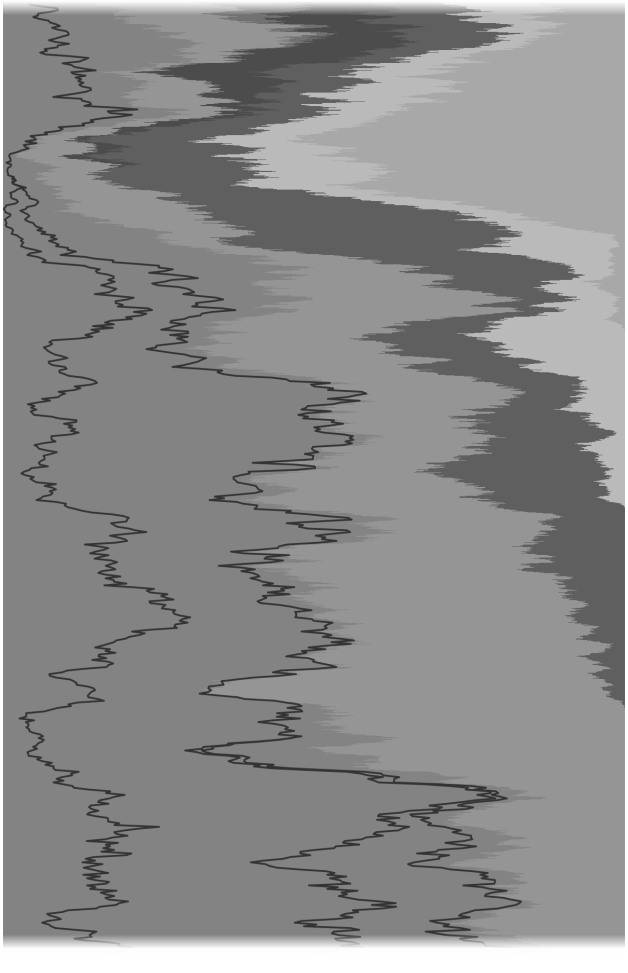
Je voudrais aussi remercier sincèrement de nombreuses personnes qui ont été à mes côtés pendant ces années, et qui ont contribué à les rendre riches et agréables. Merci à celles et ceux avec qui j'ai pu faire toutes ces randonnées, ces courses ou les deux. Mais ce serait vous rendre trop sportifs que de ne pas mentionner toutes ces soirées passées dans un appartement bien connu de la rive droite, ces goûters gourmands et ces dîners dont les qualités de cuisinier du chef n'égalent que sa gentillesse et son attention envers les autres. Merci à tous les membres de Brassens pour ces beaux moments de musique, mais aussi pour ces passages un peu plus « cuivrés », pour ces fous rire en répétition ainsi que pour avoir autant financé la rue Mouffetard depuis 5 ans. Je voudrais remercier tout particulièrement son chef aux multiples talents, sans qui nous serions encore en train de jouer James Bond et Florentiner March devant des octogénaires se bouchant les oreilles. Merci à ceux que j'ai appris à connaître au cours d'une folle année associative et que je retrouve maintenant autour d'une coinche, d'une pétanque ou d'un babyfoot. Je sais que je pourrai toujours compter sur vous. Merci aux bio14, que je retrouve toujours avec plaisir, même si ce n'est plus aussi fréquent qu'avant. Merci à ceux que j'ai rencontré de manière fortuite un matin dans un amphi à Jussieu, avec qui tout a été si naturel et qui m'ont accompagné lors de mes premiers pas dans le monde mathématique. Merci aux membres de Funko, en attendant de pouvoir rejouer ensemble un jour. Merci aux personnes avec qui j'ai passé mes deux premières années parisiennes, dont le rendez-vous dans le quartier latin qui approche à grands pas sera, je l'espère, honoré. Enfin, merci à toutes celles et ceux avec qui j'ai été proche un jour, que j'ai perdu de vue et qui, à mon grand regret, ne liront très certainement jamais ces lignes.

À mes parents. Chaque année qui passe je mesure un peu mieux votre engagement envers mon frère et moi. Merci pour votre soutien de tout instant et pour m'avoir permis de faire ces études dont la touche finale est ce manuscrit. Merci aussi à celui qui m'a toujours précédé et avec qui je suis heureux de partager encore autant de centres d'intérêt.

Il ne me reste plus qu'une seule personne à remercier, sans qui je ne serais pas aussi heureux ni épanoui aujourd'hui.

Contents

Chapter 1



1

Introduction

Illustration. Simulation of various scaling limits of the Wright-Fisher model. Time is going downwards and each color represents an initial type in the population. The thickness of a color region gives the fraction of individuals with the corresponding type, and its dynamics is a Wright-Fisher diffusion. The measure-valued process that records simultaneously the fraction of individuals with each type is a Fleming-Viot process. Finally the black lines correspond to the ancestral lineages of three individuals sampled uniformly at the last time point of the simulation. Their genealogy is a realization of Kingman's coalescent.

1.1 Stochastic models in population biology

Populations play a central role in many areas of biology, including ecology, population genetics, demography, epidemiology, and evolutionary biology, that could be gathered under the name of *population biology.* Yet, the notion of *biological population* is hard to define. A population should be made of *individuals.* Those individuals can represent physical individuals, as plants, humans or viruses, but can also be abstract entities such as genes, populations, and even species. In order to form a population, a collection of individuals should share some common characteristic and interact together. This characteristic can be, for instance, spatial proximity, belonging to the same species, or having the same genotype. Finally, a population is endowed with a notion of ancestry: current individuals are descendants from past individuals in the population.

The notion of population is thus loosely defined. Population biology is a collection of distinct fields that are interested in populations with different biological characteristics, acting at different scales, that study different aspects of them, and do not have access to the same observables of these populations. For instance, demography typically studies the variations in size of animal populations, and relies on time-series of individual counts, whereas diversification is concerned with the number of species and can only directly access the number of past species through fossil records. However, despite the diversity of biological populations that can be studied, an objective shared by of all the fields of population biology is to understand how the interactions between individuals influence the dynamics on a larger scale, that of the population.

In this thesis, I will present and study several probabilistic models of populations, applied mainly to population genetics and epidemiology. In this introductory chapter, I will first give a brief and personal account on the biological questions that have motivated my research during my PhD. Then, I will discuss the general methodology that I have used to study them, and describe some of the tools grounded in probability theory to understand the relationship between microscopic characteristics and macroscopic behavior. Finally, in each subsequent chapter, I will introduce a population model and the biological context that has motivated it, and carry out an analysis of some features of the model using the tools presented in this introduction.

1.1.1 From population genetics . . .

Population genetics studies the variations in frequencies of gene variants, called *alleles*, in a population. It aims at understanding the patterns of genetic diversity that we observe in extant populations, and the evolutionary forces that shaped this diversity. From its very foundation in the 1920s, population genetics has relied extensively on mathematical models to make quantitative predictions that could be tested on observed data. It is not surprising that its three founding fathers, Sewall Wright, J.B.S. Haldane and Ronald Fisher, were accomplished mathematicians, and that the latter of the three also made fundamental contributions to the theory of statistics. Population genetics is a mature field, which has led to rich developments in both probability theory and evolutionary biology. My original academic background being biology, but having a strong taste for mathematics, my initial interest in population genetics lied in this interplay between elegant mathematics and fascinating evolutionary questions. During my PhD, I have mainly focused on two aspects of population genetics: recombination and range expansion. I have also carried out more theoretical works on the representation of random genealogies, which are presented in Chapter 2. Finally, with two other students who share my interest in probability theory and population genetics, I studied a random forest model that encodes the parental relationship of extant individuals in a Moran model, and is not contained is this manuscript [23].

Recombination. In a broad sense, recombination is the formation of new combinations of alleles out of old ones. In a sexually reproducing population, a subset of the chromosomes of the offspring is inherited from one parent, and the rest of the chromosomes from the other parent. If the alleles are not linked, that is, are not on the same strand of DNA, recombination can occur through the random sampling of parental chromosomes during reproduction as in the left panel of Figure 1.1. When alleles are linked, the only way to modify an association of alleles is to break the DNA molecule, and to replace a chunk of DNA by a new one, as in the right panel of Figure 1.1. The most well-known example of such a mechanism

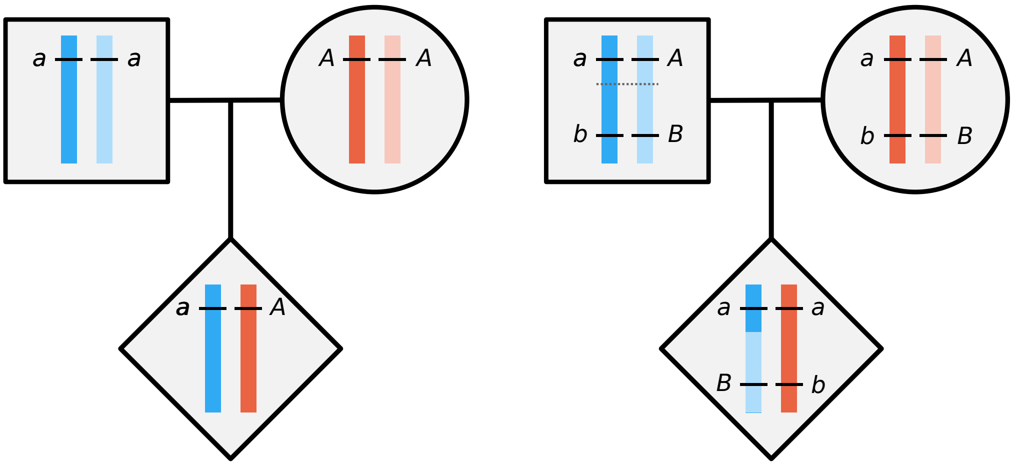


Figure 1.1: Two ways of creating new combinations of alleles. Left: A new combination of unlinked alleles a/A is formed through the sexual reproduction of two individuals with distinct alleles. Right: The new collection of linked alleles (a, B) is obtained after a crossing-over, whose location is indicated by the dashed line.

is the crossing-over that occurs during meiosis in eukaryotes. We will be mostly interested into the latter situation, and, in the remainder on the manuscript, the use of the word recombination will always refer to that situation.

At the population level, due to recombination, individuals inherit chromosomes that are mosaics of those of their ancestors. These mosaics are made of unrecombined segments, also called *identical by descent blocks* (IBD blocks), separated by *junctions* that correspond to past crossing-over events [71, 38]. Compared to the theory for one locus which is now well-established [56, 61], understanding the dynamics of these blocks and junctions remains challenging from both a mathematical and a computational points of view [122, 154]. In Chapter 5, we study the long-term genetic contribution of a focal individual in a constant-size, neutral, well-mixed population experiencing recombination. We provide an expression for the size and location on the genome of the blocks of genetic material left by this individual in the limit of large population size and large recombination rate. A better theoretical understanding of the impact of recombination on IBD blocks could help to formulate a null model for the distribution of these blocks and lead to the development of inference tools based on deviations from this model. Inference methods that leverage the length of IBD blocks have already been proposed to detect past demography [146, 180, 124] and selection [187].

At the species level, recombination can lead to the transfer of genetic material from one species to another through the formation of hybrids and subsequent backcrosses with the recipient species, a phenomenon known as *introgression* [104]. Introgression is now recognized to have had a large impact on the genome of many species [151, 175], including our own species [190, 191]. In Chapter 3, we propose a model that follows the ancestral lineages of a subset of loci sampled in a focal present-day species in backward time. Ancestral lineages are separated into distinct species due to introgression events, and brought back together at speciation events.

We encode the set of species to which the ancestral lineages of the loci belong as a partition, where two loci are placed into the same block at time t if they belong to the same species t units of time ago. Under the modeling assumption that we make, the dynamics of this partition is given by an interesting mathematical object known as Kingman's coalescent with erosion [80].

Additionally, at the population level, the model considered in Chapter 5 provides the dynamics of a single hybrid individual into a large resident population, under the assumption that the hybrid is viable, and as fit as individuals from the recipient species. We can also interpret the results we obtain as providing an expression for the amount of introgressed genetic material and its location on the genome.

Range expansion. The range of a species is the geographical region over which this species is found. The phenomenon by which a species can colonize new habitable areas is known as range expansion. Two well-documented examples of range expansions are the current invasion of cane toads in Australia [212, 208], and the past out-of-Africa expansion of early human populations [108]. Chapter 4 is devoted to a model of expanding population. Our motivation for this study comes from two quite recent observations on the genetics of such populations.

First, it was reported that a neutral or a deleterious allele can rapidly reach a large frequency over a vast region of space during a range expansion, a phenomenon dubbed *gene surfing* [54, 129, 211]. It was further predicted that, due to successive surfing of deleterious mutations, the fitness of the population should decrease along the expansion axis. The resulting fitness loss is called the *expansion load* [172, 29].

Second, there is growing evidence that the rate of loss of genetic diversity during a range expansion is reduced by the presence of an Allee effect [184, 99, 27], that is, when the per-capita growth rate of the population is not maximal at low population size [132]. The intuition behind this effect is quite simple. The range of a population can be divided into a region where it is well-established, the *bulk*, and a boundary region where population densities are lower, the *front*. In the absence of an Allee effect, individuals at the front have the highest growth rate, and are the main genetic ancestors of newly colonized regions. The wave is *pulled* by the few individuals at the front. In the presence of an Allee effect, the highest growth rate is now achieved in a region intermediate between the front and the bulk. There are more individuals that contribute genetically to the newly colonized habitats, and the genetic diversity in that region is higher. The wave is said *pushed*.

The model studied in Chapter 4 is designed to investigate the impact of the pulled or pushed nature of an expansion on the formation of an expansion load during a range expansion.

1.1.2 . . . to epidemiology

In December 2019, a new coronavirus was discovered, SARS-CoV-2, that is the causative agent of a severe respiratory disease, COVID-19. Since then, the world is facing a major sanitary crisis, leading to the shutdown of entire economies and causing fatalities now estimated to more than a million. The research community rapidly took action and started to explore many aspects of the disease. Willing to get involved in this international research effort and with the help of my mixed background in biology and mathematics, I have been the major driver of a modelling project on the COVID-19 epidemic, initiated within the SMILE group to which I belong. The scientific approach of this research project was very different from that of other works presented in this thesis. After having elaborated an epidemic model and studied its large population size limit, I collected epidemiological data from different governmental sources. Then, a large part of my work was to construct an inference framework based on the results that we had obtained and to apply it to these data, which led to Section 6.4 of the present manuscript. Even if the mathematical results of this project do not have the same level of originality as other results that I will present, it was the opportunity for me to bridge the gap between mathematics and data, and to develop a greater expertise in epidemiology.

Epidemiology is the branch of biology that studies the spread of transmissible agents, such as viruses, bacteria, or other types of pathogens. It is a second example of field where mathematical models play an important role. However, in contrast with population genetics, epidemics are more often modeled using deterministic tools, such as dynamical systems or partial differential equations, than with probabilistic methods. In many situations, the deterministic equations that are proposed for the dynamics of the epidemic correspond to the large population limit of some stochastic population model, but this convergence step is not often carried out rigorously, see for instance the discussion in [166], and the macroscopic equations often rely on unrealistic assumptions (exponentially distributed waiting times, constant infectiosity profile). In Chapter 6 and Chapter 7 we propose and study two variants of the same stochastic epidemic model. Our approach is to start by proving the convergence of the model to a deterministic set of equations, and then to use inference methods for deterministic epidemic models to carry out an estimation of some macroscopic epidemiological parameters, as the basic reproduction number or the total number of infected individuals, from measurable individual characteristics, such as the generation time. This probabilistic approach has many benefits, which are discussed in the forthcoming Section 1.1.3.

The epidemiological model considered in this thesis was intended to represent the dynamics of the COVID-19 pandemic. It was designed to take into account two important aspects of this epidemic. First, the COVID-19 is a complex disease. Upon infection, some individuals develop very mild forms of the disease, with few or no symptoms, but remain infectious [6, 164]. Even in the presence of symptoms, it has been estimated that a significant fraction of transmissions could occur before symptom onset [106, 209]. In some cases, infected individuals develop severe respiratory symptoms that require an admission to intensive care unit (ICU), and can eventually lead to death [189]. Moreover, there is a strong impact of cofactors on the severeness of the disease. For instance, age is a major factor of risk [189]. An accurate description of the dynamics of the epidemic should thus be able to take into account this high degree of heterogeneity in the courses of infections among different individuals.

Second, the reaction of many countries to the COVID-19 epidemic has been the enforcement of control measures such as school closures, nation-wide lockdowns, or mandatory face masks [72]. These control measures have large impacts on the contact rate in the population and can drastically reduce the number of transmissions. Moreover, they are typically triggered after close monitoring of some epidemiological quantity of interest, such as the case incidence, the number of occupied hospital or ICU beds, or the number of deaths. Therefore, an epidemiological model for the COVID-19 epidemic should be able to account for temporal changes in the transmission rate, and should as much as possible provide a prediction for the complex set of observables of the epidemic that are mentioned above.

The model proposed in Chapter 6 meets these two expectations. It is a rather general epidemiological model, that could also be used to study the spread of many diseases other than COVID-19. Note, however, that it is "mean field" in the sense that it does not take into account any kind of spatial or social heterogeneity in the contacts made in the population.

1.1.3 The scaling limit approach

The approach used in this thesis to study the problems exposed in the previous section relies on the analysis of stochastic population models. First, these models are often constructed by specifying a set of rules that describe the life, reproduction, and death of individuals in the population. Most of the modeling work is carried out at this step, and the specified rules should reflect the biological phenomenon under consideration. Such models are often referred to as *individual-based models*.

Then, these individual-based models are studied in two steps. First, we prove that, after an appropriate renormalization, some features of the population (its size, its genealogy, etc.) converge under a large population size limit to a continuous object called *scaling limit*. Then we study the scaling limit as an approximation of the discrete individual-based model. Each chapter in this thesis is either devoted to carrying out the first convergence step, and/or to studying the scaling limit of a population model. More precisely, in Chapter 2 we provide new representation results on exchangeable coalescents, which are known to be the scaling limits of genealogies of samples from a large population. In Chapter 3 and Chapter 4 new individual-based models are introduced and we study some properties of their scaling limits, after having identified them using heuristic arguments. Finally, in Chapter 5, Chapter 6 and Chapter 7, we also introduce individual-based models, but the bulk of the work in these chapters is to prove the convergence of these models toward their scaling limits.

In the next section we discuss the interest of this approach. In order to keep the discussion as simple as possible, we will use as an example the celebrated central limit theorem, and its functional version, Donsker's invariance principle.

1.1.4 Invariance principle, universality

It is a general observation in probability theory that the large scale dynamics of many stochastic systems is independent of the fine scale description of that system. When such a situation occurs, the system is said to display an *invariance principle* and the large scale limit is called *universal*. The prime example of this phenomenon is the central limit theorem, and its functional version which provides the convergence of random walks to Brownian motion. We rapidly recall this well-known result in order to illustrate the discussion of this section. Let $(X_i;\, i \geq 1)$ be i.i.d. real random variables, satisfying

$$\mathbb{E}[X_1] = 0, \quad \mathbb{E}[X_1^2] \; = \nu^2,$$

and define the random walk $(S(t);\, t \geq 1)$ as

$$S(0) = 0, \quad \forall t \geq 1,\; S(t) = \sum_{i=1}^{t} X_i.$$

Theorem 1.1 (Donsker's invariance principle)**.** *Let $(B_t;\, t \leq 1)$ be a standard Brownian motion. The following convergence holds in distribution for the uniform topology,*

$$\Big(\frac{1}{\sqrt{N}} S\big(\lfloor Nt \rfloor\big);\, t \in [0,1]\Big) \underset{N\to\infty}{\longrightarrow} \big(\nu B_t;\, t \in [0,1]\big).$$

A proof of this result can be found in [120], Theorem 14.9. We will see other examples of invariance principles for population models in the forthcoming Section 1.2.

Universality is a very desirable property from a modeling perspective. First, a universal limit is robust. The fine scale properties of a system will typically depend on the details of the model under consideration. Studying a universal limit ensures that the conclusions that are derived do not depend too heavily on the modeling assumptions that are made. In a sense, it sorts out the properties that are artifacts of the modeling procedure from the properties that are "intrinsic" to the system.

Second, the invariance principle justifies the use of "toy models". For instance, the random walk verifying $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = 1/2$, which is called the symmetric simple random walk, is highly tractable and enjoys nice combinatorial properties that make many computations feasible. The results obtained for this special case can be used to derive properties of the universal limit, the Brownian motion, which in turn provides a limiting expression for that property for all random walks with finite variance. A well-known example is the use of the reflection principle to derive the distribution of the maximum of a Brownian trajectory (see for instance Section 8 of [25]). Thus, many properties of toy population models, such as the Wright-Fisher model that we will introduce, hold for a wider class of models through the existence of a common scaling limit.

Third, an invariance principle allows us to identify the characteristics of the microscopic system that influence its macroscopic behavior. Typically, the universal limit is an aggregated version of the discrete objects and depends on less parameters. The limiting Brownian motion only depends on the first two moments of the increments of the walk, and not on their entire distribution. From an inference point of view, this indicates the relevant parameters that should be estimated in order to understand and predict the dynamics of the system. Moreover, the universal limit is often easier to simulate, and requires less computational effort to study than the original system. Compare the simulation of one Gaussian random variable to that of the sum of a large number of independent variables. These points will be well illustrated in Chapter 6, where we conduct an estimation of the parameters of the scaling limit of our epidemiological model that fit best the COVID-19 epidemic in France.

Studying population models through their scaling limits often comes at the cost of higher abstraction. Populations are discrete structures that can be described intuitively as a finite collection of individuals that reproduce and die. Taking a scaling limit involves going from this discrete description to a "continuous" version of the population. It requires to identify what are the marginals of interest, as well as suitable state spaces in which the limits live. The success of this approach is conditioned on the prior theoretical development of an adequate framework, with a convenient notion of convergence and enough analytical tools to study the limit. Therefore, studying scaling limits is a challenging task, where deep mathematics can prove useful to the understanding of natural phenomena.

1.2 Scaling limits of classical population models

In this section, we provide examples of classical scaling limits of stochastic population models. These models are simpler than those studied in this thesis. Nevertheless these remarkable examples will be the building blocks of the more complicated scaling limits that we will consider. It is also an opportunity to introduce the formalism needed to study the objects that will be considered in this manuscript.

1.2.1 General neutral population models

All the scaling limits described in this chapter will be derived from the following population model. The model is built out of an array $(\xi_i^n;\, 1 \leq i \leq n < \infty)$ of random variables valued in $\mathbb{Z}_+$, the set of non-negative integers. Generations are discrete and non-overlapping: at each generation, all individuals die and are replaced by a random number of new individuals. For each fixed n, the row $(\xi_1^n, \dots, \xi_n^n)$ gives the offspring sizes when the population is of size n: ξ_i^n is the number of children of the i-th individual. We assume that the vector of offspring sizes is *exchangeable*, in the sense that for any permutation σ of $[n] \coloneqq \{1, \dots, n\}$ we have

$$(\xi_1^n, \dots, \xi_n^n) \overset{\text{(d)}}{=} (\xi_{\sigma(1)}^n, \dots, \xi_{\sigma(n)}^n).$$

Exchangeability amounts to saying that the contribution of each individual to the following generation is identical in distribution. In particular, the number of children is not influenced by any inheritable trait carried by some individual. In evolutionary biology, if an allele does not influence the reproductive success of its carrier, it is called *neutral*. Thus, in our context, exchangeability reflects the fact that all alleles carried by the individuals in the population are neutral. The population model is now constructed as follows.

Definition 1.2. Suppose that the population starts from a random number of individuals denoted by $Z(0)$. Let $Z(t)$ be the population size at some generation $t \geq 0$. Conditional on $Z(t) = n$, label the individuals by $\{1, \dots, n\}$ and let $(\xi_1(t), \dots, \xi_n(t))$ be a copy of $(\xi_1^n, \dots, \xi_n^n)$ independent of the previous generations. Set $\xi_i(t)$ to be the number of offspring of i in the next generation. In particular,

$$Z(t+1) = \sum_{i=1}^{Z(t)} \xi_i(t). \qquad \circ$$

Note that, by exchangeability, the previous construction does not depend on the labeling of the individuals at some generation. This model, as defined above, is too general to be studied. It is simply a convenient way to place into the same framework two very influential population models: the *Galton-Watson process* and the *Cannings model*.

Definition 1.3 (Galton-Watson process)**.** Suppose that the array of offspring sizes $(\xi_i^n;\ 1 \leq i \leq n < \infty)$ is made of i.i.d. variables with common distribution μ. Then the corresponding population model is called a Galton-Watson process. At each generation, each individual gives birth to a number of children distributed as μ, independently of other individuals in the population. ○

Galton-Watson processes are maybe the simplest stochastic population models that one can conceive. Individuals in the population do not interact as they reproduce independently from each other: they only give birth and die. Galton-Watson processes enjoy the *branching property*. Suppose that $(Z(t);\ t \geq 0)$ and $(Z'(t);\ t \geq 0)$ are two independent Galton-Watson processes, started from $Z(0)$ and $Z'(0)$ respectively. Then

$$(Z(t) + Z'(t);\ t \geq 0) \overset{\text{(d)}}{=} (\widetilde{Z}(t);\ t \geq 0) \tag{1.1}$$

where $(\widetilde{Z}(t);\ t \geq 0)$ is a Galton-Watson process started from $Z(0)+Z'(0)$. From the branching property, it is possible to derive very precise results about $(Z(t);\ t \geq 0)$, including the probability of extinction and the long-time asymptotic behavior of the population size. We refer to [3] for a complete introduction to Galton-Watson processes. The branching property (1.1) is the starting point for many extensions of Galton-Watson processes, which have led to the rich theory of branching processes, see for instance [147] for a treatment of a very general class of branching processes.

Galton-Watson processes either die out, or grow to infinity. This is in contradiction with the observation that the size of many natural populations seems to remain close to an equilibrium value, named the *carrying capacity*, see for instance

Chapter 4 in [105]. Therefore, even if Galton-Watson processes have led to important developments in probability theory, they are poor population models. One way to circumvent the previous issue is to reduce the growth rate of the population at large population size [134]. The models of Chapter 4 and Chapter 7 follow this approach. Another way to prevent infinite growth is to assume that the population size is constant. Chapter 2 and Chapter 3 are based on constant-size models. Finally, Galton-Watson processes remain good approximations of the early growth phase of populations. Both Chapter 5 and Chapter 6 study branching approximations of more complex population models. We now introduce Cannings models [36], which are the canonical examples of fixed size population models.

Definition 1.4 (Cannings model)**.** Suppose that the array of offspring sizes fulfills

$$\forall n \geq 1, \quad \sum_{i=1}^{n} \xi_i^n = n.$$

Then, starting from $Z_0 = n$, the population remains of constant size n. The corresponding population model is called a Cannings model. ○

The most celebrated of all Cannings models is the Wright-Fisher model. It corresponds to the Cannings model where

$$(\xi_1^n, \dots, \xi_n^n) \overset{(\mathrm{d})}{=} \mathrm{Multinomial}\Big(n; \tfrac{1}{n}, \dots, \tfrac{1}{n}\Big).$$

The Wright-Fisher model is often described by saying that, at each generation, individuals sample their parent uniformly from the previous generation.

Note that, as the population size is constant, it is possible to define Cannings models for all generations $t \in \mathbb{Z}$ by considering an i.i.d. sequence of vectors $(\xi_1(t), \dots, \xi_n(t);\, t \in \mathbb{Z})$. Actually, a similar extension to all $t \in \mathbb{Z}$ could be made for the more general model of Definition 1.2, provided that the population size $(Z(t);\, t \geq 0)$ admits a stationary distribution.

The description of the population through Definition 1.2 contains much information: from the array $(\xi_i(t);\, t \geq 0,\, i \leq Z(t))$ one can obtain the genealogical relationship between any pair of individuals at any generation. In general we do not need such a detailed description, but are rather interested in more aggregated features of the population. From a mathematical standpoint, this amounts to projecting the distribution of the population on some smaller space, that is, to study *marginals* of the population process. Let us provide some examples of interesting marginals.

The simplest of all marginals is certainly the total population size. In our setting, it is a Markov process $(Z(t);\, t \geq 0)$ valued in $\mathbb{Z}_+$. We could also be interested in writing $Z(t)$ as the sum of the contributions to generation t of each of the initial individuals in the population. This can be encoded as a point measure on $\{1, \dots, Z(0)\}$ defined as

$$Y(t) = \sum_{i=1}^{Z(0)} Z^{(i)}(t)\delta_i$$

where $Z^{(i)}(t)$ is the number of descendants at generation t of the initial individual with label i. Note that the total mass of $Y(t)$ is $Z(t)$.

Another very important marginal is the genealogy of population. For some fixed generation T, the genealogy of the population can be encoded as a process $(\Pi(t);\, t \leq T)$ valued in the partitions of $\{1, \dots, Z(T)\}$ called a *coalescent*, and defined as

$$i \sim_{\Pi(t)} j \iff i \text{ and } j \text{ have a common ancestor at generation } T - t.$$

An alternative way of encoding this genealogy is to record, for each pair of individuals, the time to their *most-recent common ancestor* (MRCA). If we define

$$d(i, j) = \inf\{t \geq 0 : i \text{ and } j \text{ have an ancestor at time } T - t\}$$

then d is called an *ultrametric*, and contains the same information as $(\Pi(t);\, t \leq T)$.

The marginal obtained by further varying the time of observation T corresponds to the dynamical genealogy of the population. Finally, the most informative of all marginals is the vector of offspring sizes $(\xi_i(t);\, t \geq 0,\, i \leq Z(t))$, from which all the features of the populations can be recovered.

Remark 1.5. There are continuous-time analogs of the discrete-time models presented here that have played an important role in population biology. Let us mention the Moran model [158], which is a constant-size continuous-time model very popular in population genetics, and the continuous-time branching processes (see for instance Chapter III of [3]). ○

1.2.2 Feller and Wright-Fisher diffusions

One of the most obvious features of a population is its size. It is thus not surprising that the first stochastic scaling limits that have been derived were those for the dynamics of the population size. The natural framework for this limit is that of diffusions, that is, solutions to stochastic differential equations (SDE) driven by a Brownian motion. Diffusion theory in one dimension has received a lot of attention, and comes with many tools to study the limiting objects, such as the speed measure, the scale function, random time-changes, or the Itô formula. They allow us to obtain expressions for important quantities, such as the distribution of the extinction time, or the probability of fixation of an allele. We refer to [145] for a general introduction to diffusion theory, and to [56, 135] for other accounts directed towards population processes. Let us now introduce the limiting diffusions for the population size of a Galton-Watson process and of a Wright-Fisher model: the *Feller diffusion* and the *Wright-Fisher diffusion*. Several realizations of these processes are displayed in Figure 1.2.

Feller diffusion. We are interested in describing the dynamics of a Galton-Watson process started from a large number N of individuals, where N is a scaling

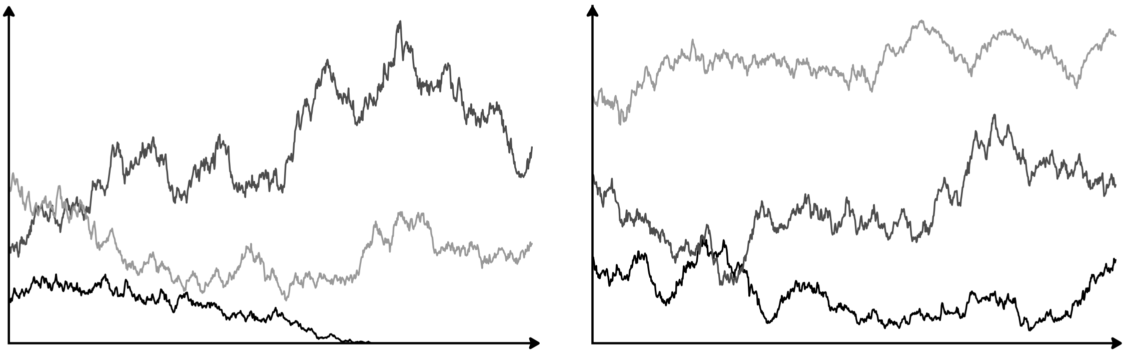


Figure 1.2: Simulation of diffusion limits of population processes. Left: Three independent realizations of the Feller diffusion. Right: Three independent realizations of the Wright-Fisher diffusion.

parameter. Recall the notation ξ_1^N for the number of children of each individual, and $Z^N(t)$ for the total population size at generation t.

We will consider Galton-Watson processes that are "nearly critical", that is, such that

$$\mathbb{E}\Big[\xi_1^N\Big] \ = 1 + \frac{\gamma_N}{N}$$

where γ_N is of order 1. This assumption ensures that the process does not die out or grow to infinity too fast, and that we obtain an interesting scaling limit. Moreover, we will need the following technical assumption

$$\forall a > 0, \quad \limsup_{N\to\infty} \ \mathbb{E}\Big[(\xi_1^N)^2;\, \xi_1^N \geq a\sqrt{N}\Big] = 0. \tag{1.2}$$

The proof of the following result can be found for instance in [169], Proposition 4. The limiting diffusion in the next result is known as the Feller diffusion.

Theorem 1.6. *Suppose that*

$$\frac{Z^N(0)}{N} \longrightarrow x, \quad \gamma_N \longrightarrow \gamma, \quad \mathrm{Var}\Big(\xi_1^N\Big) \longrightarrow \sigma^2.$$

Then, under assumption (1.2)*, the following limit holds in distribution for the Skorohod topology,*

$$\Big(\frac{1}{N}\, Z^N(\lfloor Nt \rfloor);\, t \geq 0\Big) \longrightarrow \Big(Z_t;\, t \geq 0\Big), \tag{1.3}$$

where $(Z_t;\, t \geq 0)$ *is the unique solution to*

$$Z_0 = x, \quad \mathrm{d}Z_t = \gamma Z_t + \sigma\sqrt{Z_t}\,\mathrm{d}B_t,$$

where $(B_t;\, t \geq 0)$ *is a Brownian motion.*

Note that the Feller diffusion enjoys the branching property. If $(Z_t;\, t \geq 0)$ and $(Z'_t;\, t \geq 0)$ denote independent Feller diffusions started from x and y respectively, then a direct application of Itô's formula proves that

$$\Big(Z_t + Z'_t;\, t \geq 0\Big) \overset{(\mathrm{d})}{=} \Big(\widetilde{Z}_t;\, t \geq 0\Big) \tag{1.4}$$

where $(\widetilde{Z}_t;\, t \geq 0)$ is a Feller diffusion started from $x + y$. (The branching property for the Feller diffusion actually follows from that of the Galton-Watson process and the limit (1.3).) Real-valued strong Markov processes that have the branching property (1.4) are called continuous-state branching processes (CSBP), and there exist convergence results similar to (1.3) that prove the convergence of more general Galton-Watson processes to CSBP, see [51].

Wright-Fisher diffusion. We now derive a similar limit for the Wright-Fisher model. We will not be interested into the total population size, as it is fixed to some constant N, but into the frequency of a given allele. Suppose that at some locus there are two alleles a and A in the population. Let us denote by $Y^N(0)$ the number of individuals carrying the allele A at $t = 0$, and assume that each individual inherits the allele of its parent. Let $Y^N(t)$ be the number of individuals with allele A at generation t.

The following result provides the diffusion approximation of the frequency of alleles A in the population. A proof can be found in Chapter 10 of [58].

Theorem 1.7. *Let $(Y^N(t);\, t \geq 0)$ be the number of carriers of allele A in a Wright-Fisher model. Then, if $Y^N(0)/N \to x$, the following convergence holds in distribution for the Skorohod topology,*

$$\Big(\frac{1}{N}\, Y^N(\lfloor Nt \rfloor);\, t \geq 0\Big) \longrightarrow \Big(Y_t;\, t \geq 0\Big),$$

where $(Y_t;\, t \geq 0)$ is the unique solution to

$$Y_0 = x, \quad \mathrm{d}Y_t = \sqrt{Y_t(1 - Y_t)}\, \mathrm{d}B_t,$$

where $(B_t;\, t \geq 0)$ is a Brownian motion.

The limiting diffusion in the previous result is known as the Wright-Fisher diffusion. Similar diffusion limits have been derived for extensions of the Wright-Fisher model that account for mutation or selection [56]. Note that the previous result does not explicitly show that the Wright-Fisher diffusion is a universal limit, as we have only proved the convergence in a very particular case. Nevertheless, a similar limit should hold for the allele frequency of a larger class of Cannings models, under assumptions similar to those of Theorem 1.6.

There is a remarkable connection between the Feller diffusion and the Wright-Fisher diffusion. Consider two independent Feller diffusions denoted by $(Z_t;\, t \geq 0)$

and $(Z'_t;\, t \geq 0)$, started from x and x' respectively, with $\gamma = 0$ and $\sigma = 1$. Let us define

$$\forall t \geq 0, \quad Y_t = \frac{Z_t}{Z_t + Z'_t}, \quad N_t = Z_t + Z'_t.$$

Then an application of Itô's formula shows that $(Y_t;\, t \geq 0)$ solves

$$Y_0 = \frac{x}{x+x'}, \quad \mathrm{d}Y_t = \sqrt{\frac{Y_t(1-Y_t)}{N_t}}\, \mathrm{d}B_t.$$

The previous SDE is similar to that solved by the Wright-Fisher diffusion, except that there is an additional term accounting for the varying total population size: when population size is low, the dynamics of the allele frequencies is faster. This reflects the well-known fact that *genetic drift*, that is, the fluctuations in allele frequencies due to random births and deaths, is stronger for lower population sizes. Moreover, this simple calculation strongly suggests that the "genetic structure" of the Feller diffusion is similar to that of a Wright-Fisher model. A formalization of this idea in a much more general framework can be found in [26]. Even if the population models introduced in Section 1.2.1 seemed rather different at first sight, their large population size scaling limits allow us to draw new connections between them, and to get deeper insight into the fundamental characteristics that differentiate them.

1.2.3 Kingman's coalescent

Recall the definition of a Cannings models of size N constructed from the vector $(\xi_1^N, \dots, \xi_N^N)$. Recall also that the genealogy of the population at generation T can be encoded as a process $(\Pi^N(t);\, t \geq 0)$ valued in the partitions of $[N]$ and defined as

$$i \sim_{\Pi^N(t)} j \iff i \text{ and } j \text{ have a common ancestor at time } T - t.$$

We will not be interested into the genealogy of the whole population, but only into the genealogy of a sample of fixed size n. By exchangeability of the population, this amounts to considering the genealogy of the individuals labeled $\{1, \dots, n\}$, that is to consider the restriction of Π^N to $[n]$, that we denote by Π_n^N.

As we will see, one universal limit of the genealogies of Cannings models is Kingman's coalescent, which was introduced in [128]. It is defined as follows, see Figure 1.3 for a simulation.

Definition 1.8 (Kingman's coalescent)**.** The Kingman coalescent $(\Pi^K(t);\, t \geq 0)$ is a process valued in the partitions of $\mathbb{N}$. It is started from the partition into singletons, and for any n, its restriction $(\Pi_n^K(t);\, t \geq 0)$ to $[n]$ is a Markov process such that each pair of blocks merges at rate one. ○

Before stating the result, we need some additional notation. Let

$$c_N = \frac{\mathbb{E}\big[\xi_1^N(\xi_1^N - 1)\big]}{N-1}, \quad d_N = \frac{\mathbb{E}\big[\xi_1^N(\xi_1^N - 1)(\xi_1^N - 2)\big]}{(N-1)(N-2)}.$$

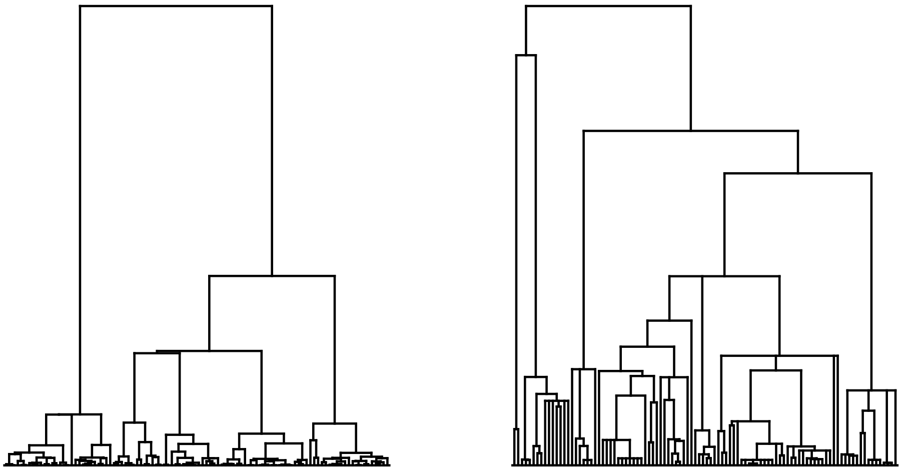


Figure 1.3: Simulations of coalescents. Left: Kingman's coalescent with a sample size of $n = 100$. Right: Beta$(2-\alpha, \alpha)$ coalescent, with a sample size of $n = 100$, and $\alpha = 3/2$.

By exchangeability, c_N is the probability that two individuals find their common ancestor at the previous generation, and d_N is the probability that three individuals all have the same ancestor at the previous generation. The following result, known as Möhle's lemma [155, 156], shows that under some mild assumptions on c_N and d_N the genealogy of Cannings models converge to Kingman's coalescent.

Theorem 1.9 (Möhle's Lemma)**.** *Suppose that*

$$\lim_{N\to\infty} \frac{d_N}{c_N} = 0.$$

Then, for any $n \geq 1$,

$$\Big(\Pi_n^N\big(\lfloor c_N t\rfloor\big);\, t \geq 0\Big) \longrightarrow \Big(\Pi_n^K(t);\, t \geq 0\Big),$$

in distribution for the Skorohod topology.

All of the points of the discussion in Section 1.1.4 are well illustrated by Kingman's coalescent. The limiting object is simpler than the genealogy at fixed N. Kingman's coalescent is a binary tree, so that no more than two lineages coalesce at a time, whereas it is possible to see multiple coalescences at fixed N. Moreover, the limit only depends on the rate of pairwise mergers c_N. Finally, the universality of Kingman's coalescent has made it popular in population genetics, as it is a robust model for the genealogy of neutrally evolving populations. See for instance [216] for an account more directed towards applications in biology.

Remark 1.10. Hidden in the hypothesis of Theorem 1.9 is the assumption that no individual in the population leaves a very large offspring and contributes to a large fraction of the population in the next generation. When such a situation occurs, the limiting genealogy is no longer given by Kingman's coalescent, but by multiple mergers coalescents [195]. A very important subclass of these coalescents

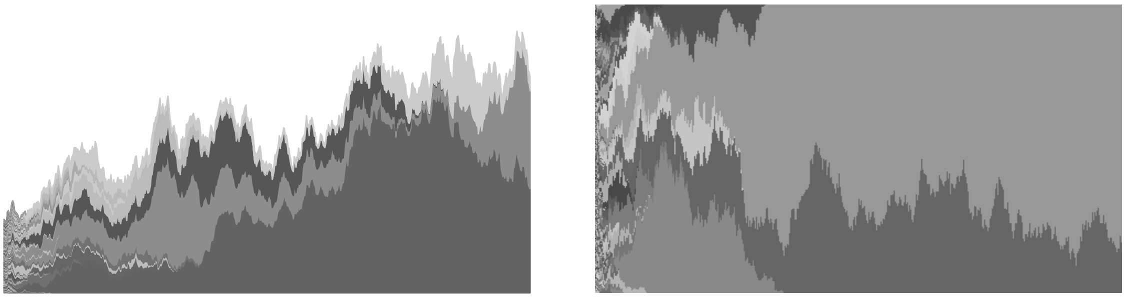


Figure 1.4: Simulation of a Feller branching diffusion (left), and of a Fleming-Viot process (right). Each initial individual is given a different color, and the thickness of each region indicates the size of the progeny of the individual with the corresponding color.

are the Λ-coalescents introduced in [178, 188], see Figure 1.3 for an illustration. Convergence of the genealogies of Cannings models to such coalescents has been for instance conducted in [196] and [157]. ○

Even if Kingman's coalescent is a very popular genealogical model in population genetics, the shape of species trees are better described by birth-death processes [109, 159], which are continuous-time versions of Galton-Watson processes. Interestingly, the approach of sampling individuals in the population and encoding their genealogy as a partition-valued process has not received much attention for Galton-Watson processes, but see [138, 119, 102] for results in this direction, and [19] for a link between Kingman's coalescent and the Feller diffusion. An alternative approach to study the tree shape of birth-death processes that has been very fruitful is to endow the population with a "planar order", and record the coalescence times between consecutive individuals in this order, see Section 1.2.4 for a formal definition. The corresponding genealogy, called a *coalescent point process* (CPP), has been studied for finite-size branching processes [136, 141], and its scaling limit has been given in [179].

1.2.4 Other scaling limits

We end this section by presenting informally the scaling limits of more complex marginals of the population, which lead to more involved objects in the limit.

Superprocesses. In addition to understanding the dynamics of the total population size, it is interesting to separate this total size into the contribution of each initial individual at $t = 0$. Recall that, if the population starts from N_0 individuals, this information can be encoded as a point measure

$$Y^N(t) = \sum_{i=1}^{N_0} Z^{(i)}(t)\delta_i,$$

where $Z^{(i)}(t)$ is the number of descendants at generation t of the individual with label i at $t = 0$. Now, let us rescale both the population size and the label space by N, and define the rescaled process as

$$\widetilde{Y}^N(t) = \frac{1}{N}\sum_{i=1}^{N_0} Z^{(i)}(t)\delta_{i/N}.$$

Suppose that $N_0/N \to x$, then the initial measure $\widetilde{Y}^N(0)$ converges weakly to the Lebesgue measure on $[0, x]$. For the Galton-Watson process and the Wright-Fisher model, it can be shown that the entire rescaled process $(\widetilde{Y}^N(t);\, t \geq 0)$ converges to a measure-valued process. In the former case, the limit is known as a *Feller branching diffusion* and in the latter case it is the *Fleming-Viot process*. Both processes are illustrated in Figure 1.4. More generally, measure-valued population processes are known as *superprocesses*, we refer to [55] for a nice introduction to these notions. Note that, in the original formulation of superprocesses, each individual moves into space, so that the measure does not only encode the offspring sizes in the population, but also the locations of the individuals. Here we have considered the special case where there is no spatial motion.

Dynamical genealogies. Kingman's coalescent corresponds to the genealogy of a sample from the population at one point in time. A naive way to obtain an evolving genealogy is to sample n individuals independently at each generation T, and to consider the collection of coalescents associated to each sample. However, this coalescent-valued process does not enjoy any continuity property, as the sample between two consecutive generations are independent, and it is hard to obtain a scaling limit for this object. This issue can be overcome in several ways.

First, it is possible to sample the individuals in a "smart way", so that the individuals sampled at generation $T+1$ are related to those sampled at generation T and that the coalescent at $T+1$ is similar to that at T. This idea leads to a construction of the evolving Kingman coalescent from the *lookdown process* of [49], that was first proposed in [176, 177]. The offspring of all individuals but one in the population will eventually die out. Thus, individuals can be ranked according to the time at which their offspring goes extinct, in decreasing order. This sequence of initial individuals, ordered by time of extinction of their offspring, is called the *Eves* of the population. The evolving Kingman coalescent constructed from the lookdown process corresponds to the dynamical genealogy of the Eves of the population [133].

Second, the genealogy of a population can be encoded as a metric space. The distance between two individuals is the time to their MRCA. Using the framework of random metric measure spaces [94, 91], it is possible to define a tree-valued process that corresponds to the scaling limits of the evolving genealogy of a Cannings model [91] or of a Galton-Watson process [43].

Finally, the point of view advocated in Chapter 2 is to make use of a *planar* representation of the genealogies. Suppose that, at $t = 0$, the initial individuals are given an order. Then we can define inductively for any T a total order $\preceq_T$ on the individuals alive at generation T as follows. For each set of siblings, choose an arbitrary order. If i and j are not siblings, their order is that of their parents. If

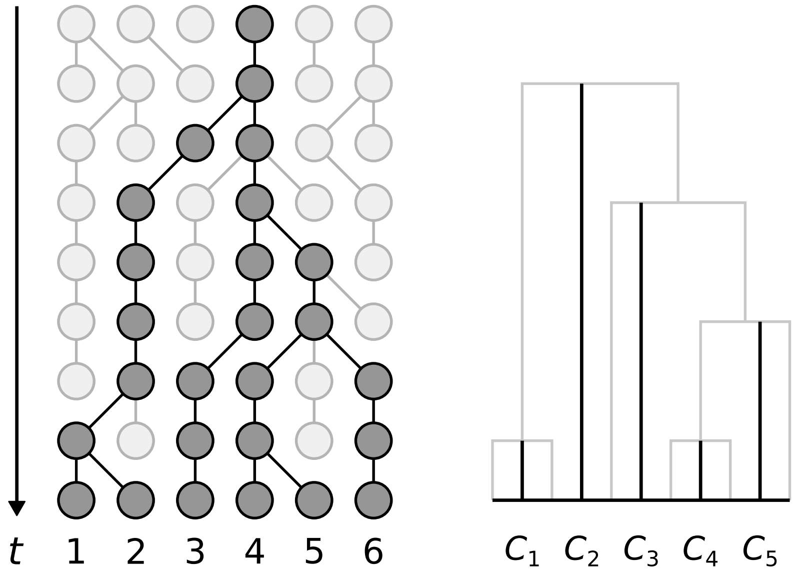


Figure 1.5: Illustration of the planar representation of genealogies.

the individuals at generation T are labeled by $[Z(T)]$ in such a way that $i \preceq_T j$ iff $i \leq j$, their genealogy can be encoded as a vector $(C_1, \dots, C_{Z(T)-1})$, where C_i is the time to the MRCA of i and $i+1$, see Figure 1.5. The order of this vector is crucial, as changing this order leads to distinct genealogies.

In the limit, this vector is conveniently encoded as a point process, and the corresponding genealogy is known as a *comb*. If this point process is a Poisson point process, then the comb is called a CPP, and was already discussed in Section 1.2.3. The CPP associated to the Feller diffusion can be obtained as the depths of the excursions of a Brownian motion below some level [179]. There is also a simple construction of the comb associated to Kingman's coalescent see for instance [140]. In Chapter 2 we will derive a similar representation result for a much broader class of coalescents, and study the associated evolving genealogy.

Stochastic flows. All objects that we have introduced so far are scaling limits of particular characteristics of the population. It is also possible to define a scaling limit for the population as a whole. The idea is to record for each pair of times $s \leq t$ a function $F_{s,t}$ that records the ancestors of the individuals in generation t that lived at generation s. (Formally, this is encoded as a measure in much the same way as for superprocesses, and $F_{s,t}$ is the distribution function of this measure.) Then the collection $(F_{s,t};\, s \leq t)$ is called a *stochastic flow*. The stochastic flow associated to Galton-Watson processes is called the *flow of subordinators* and was introduced in [21], whereas that corresponding to Cannings models is known as the *flow of bridges*, see [21].

All the scaling limits that have been exposed in this section are marginals of a stochastic flow. The Feller branching diffusion and Fleming-Viot process correspond to the measure-valued process $(F_{0,t};\, t \geq 0)$, where $(F_{s,t})$ is a flow of subordinator and a flow of bridges respectively. The genealogy of the population at

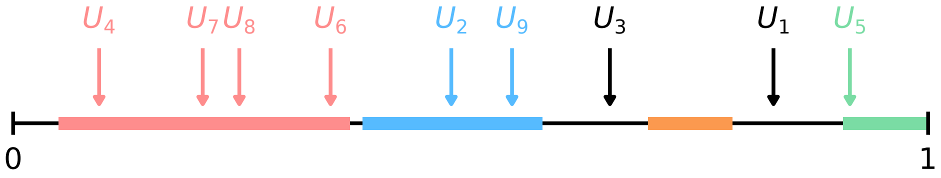


Figure 1.6: Illustration of the paintbox procedure. The interval I is the union of the colored subintervals, and each U_i is represented by an arrow. The partition Π is the partition in colors of the $(U_i;\, i \geq 1)$, that is, $\Pi = \{\{1\}, \{2, 9\}, \{3\}, \{4, 6, 7, 8\}, \{5\}\}$. Note that each variable that does not belong to I forms a singleton block.

time T is encoded by the process $(F_{T-t,T};\, t \geq 0)$, and Kingman's coalescent can be recovered through a sampling procedure involving this process. Finally, the process obtained by varying the observation time T is an evolving genealogy. Stochastic flows of bridges will play an important role in Chapter 2 and Chapter 3, where there will be introduced formally.

1.3 Outline of the thesis

The remainder of the manuscript is divided into six chapters. Each chapter is self-contained, with its own introduction, notation and references. As discussed in the previous section, there are many features of a population that can be studied, and the mathematical formalism that is used depends on the particular marginal under consideration. The chapters have been gathered into three different parts, each corresponding to a different type of formalism being used. The remainder of this section contains some basic results and definitions about the objects that underpin each part, as well as an outline of each chapter.

1.3.1 Exchangeable partition-valued processes

The first part of this manuscript contains two chapters that make use of the framework of exchangeable partition-valued processes. Let us briefly recall some basic facts on exchangeable partitions in order to motivate them. For any permutation σ and partition π of $\mathbb{N}$, we can define a partition $\sigma(\pi)$ whose blocks are given by the following equivalence relation

$$i \sim_{\sigma(\pi)} j \iff \sigma(i) \sim_\pi \sigma(j).$$

A random partition Π of $\mathbb{N}$ is called exchangeable if, for any permutation σ, we have

$$\sigma(\Pi) \overset{(\mathrm{d})}{=} \Pi.$$

A fundamental result due to Kingman shows that any exchangeable partition can be obtained through a procedure called a *paintbox* that we now describe.

Let I be some random open subset of $(0, 1)$, and $(U_i;\, i \geq 1)$ be a sequence of i.i.d. uniform variables on $(0, 1)$. The open set I can be written as a countable union of open intervals. We can define a partition Π of $\mathbb{N}$ by prescribing that

$$i \sim_\Pi j \iff U_i \text{ and } U_j \text{ belong to the same subinterval of } I.$$

See Figure 1.6 for an illustration of the paintbox procedure. It is clear that the partition Π is exchangeable. Kingman's representation theorem states the converse, that is that any exchangeable partition can be represented as a paintbox on some random interval I, see [127] for the original proof, or Theorem 2.1 in [20] for a modern proof. It is not hard to see that the distribution of Π only depends on the sequence of lengths of the subintervals of I, and not on their location. This sequence of lengths is called the *asymptotic frequencies* of Π, and entirely characterizes its law. We refer to [20] for additional results on exchangeable partitions.

Coalescents, ultrametric spaces, combs. In Chapter 2, we study general exchangeable coalescents, that is, partition-valued processes that are exchangeable, non-decreasing, and *a priori* non-Markovian. As explained in Section 1.2.3, these processes encode the genealogy of a sample from a population, and the canonical example of exchangeable coalescents is Kingman's coalescent. The main result of this chapter proves that all exchangeable coalescents admit a planar representation in the sense of Section 1.2.4. This requires to derive an extension of Kingman's paintbox construction for exchangeable coalescents, which involves sampling from an interval-valued process called a *nested interval-partition.*

After having derived this result, we investigate two of its consequences. First, we show that any exchangeable coalescent can be represented as a sample from an ultrametric measure space. This extends a well-known connection between separable ultrametric measure spaces and exchangeable coalescents without dust. This result requires a non-trivial extension of the framework of ultrametric measure spaces to incorporate non-separable spaces, which is another contribution of our work. Second, we provide a new representation of the evolving Kingman coalescent discussed in Section 1.2.4 using nested interval-partitions. This representation can be easily adapted to more general dynamical genealogies, such as dynamical genealogies whose one-dimensional marginal is a Λ-coalescent.

This chapter is joint work with Amaury Lambert and Emmanuel Schertzer. It has been accepted for publication in the *Annals of Applied Probability* [79].

Kingman's coalescent with erosion. In Chapter 3, we study a fragmentation-coalescence process known as Kingman's coalescent with erosion. In this process, any pair of blocks merges at rate one, and any integer is *eroded*, that is, is removed from its block and placed into a singleton block, at rate d. Our initial interest in this process was to describe the backward in time dynamics of a diversification model incorporating introgression and speciation, which has been discussed briefly in Section 1.1.1.

An interesting feature of fragmentation-coalescence processes is that they display stationary distributions [18]. We show two results on this stationary distribution. First, we give an expression for the asymptotic frequencies of its blocks in terms of a Fleming-Viot process discussed in Section 1.2.4. This expression can also be formulated in terms of an infinite sequence of independent Wright-Fisher diffusions, conditioned on non-extinction. Second, we investigate the asymptotic properties of the restriction of the stationary distribution to $[n]$, for large n. We show that there are asymptotically $\sqrt{2dn}$ blocks, and provide the limit of the empirical distribution of the blocks sizes. This results is proved by coupling Kingman's coalescent with erosion to another process that we have introduced and that can be of independent interest, called Kingman's coalescent with immigration.

From a modelling perspective, our results predict that there are asymptotically $\sqrt{2dn}$ species that are ancestral to a set of n loci in a present-day focal species. This prediction is unrealistically high. An explanation for this discrepancy is that we have not taken into account the fact that hybridization is less likely to occur between more distantly related species [41]. Backwards in time, the ancestral lineages of the focal species will have a tendency to "cluster" together into the same species. See [152] for a model where this effect is taken into account.

This chapter is joint work with Amaury Lambert and Emmanuel Schertzer. It has been published in *Electronic Journal of Probability* [80].

1.3.2 Branching processes in population genetics

The two objects studied in the second part are related to the theory of measure-valued branching processes. In both cases, each individual in the population has a characteristic which can be seen as a point in some space E. In Chapter 4, this characteristic is the spatial location and the number of deleterious mutations carried by an individual, so that $E = \mathbb{R} \times \mathbb{Z}_+$. In Chapter 5, it is the block of ancestral genetic material inherited by an individual, so that $E = \mathcal{I}(\mathbb{R}_+)$, the set of intervals of $\mathbb{R}_+$. The population can then be conveniently encoded as a random point measure on E, in a very similar way to the encoding of the Fleming-Viot process and of the Feller branching diffusion discussed in Section 1.2.4, but where the measure also encodes the location in E of the individuals. In both Chapter 4 and Chapter 5 we are interested in describing the scaling limit of the empirical measure of the location of the individuals for a large population size. We refer to [55] for an introduction to the theory of super-processes which correspond to the latter scaling limit, to [199] for an introduction to branching random walks, which are the discrete analogous of super-processes, and to [146] for a very general account on measure-valued branching processes.

The spatial Muller's ratchet. In Chapter 4, we consider a population expanding on a linear space represented by the real line. The genetic structure of the population is so that individuals can only accumulate deleterious mutations through time, leading to what is known as a Muller's ratchet. We observe that, as the population expands, "spatial clicks" of the ratchet occur, in the sense that the number of deleterious mutations of the fittest individuals at the front decreases. These repeated spatial clicks of the ratchet lead to a decrease in fitness at the front during the expansion, and thus to the formation of an expansion load, as described in Section 1.1.1. Using numerical simulations, we are able to study the impact of the pushed or pulled nature of the expansion on the rate at which spatial clicks occur, see Section 1.1.1 for a brief definition of pushed and pulled expansions. Moreover, we derive heuristically the scaling limit of the location and number of mutations of individuals in the population. It is a system of reaction-diffusion partial differential equations. By reducing this system of equations to a well-known one-dimensional PDE, we can prove that it admits a collection of travelling wave solutions, and provide an expression for the wave speed of these solutions.

Chapter 4 is joint work with Alison Etheridge and has been published under a slightly different form in *Theoretical Population Biology* [78].

A branching process with recombination. In Chapter 5 we study the branching approximation of a Wright-Fisher model with recombination. In the latter model, the population has fixed size N, and each individual carries a single haploid continuous chromosome represented by the interval $[0, R]$. At each generation, each individual samples two parents uniformly from the previous generation and inherits either a chromosome from one of its parents, or, with a small probability, a recombined chromosome which is a mixture of that of its two parents, and where the location of the crossing-over is chosen uniformly on the chromosome, see Figure 1.1. We consider a branching approximation of this model, and follow forward in time the genetic contribution of one focal individual at $t = 0$. We provide the large population size, large chromosome size limit of the empirical distribution of the lengths of blocks of genome that descend from this focal individual. We also provide an expression for the location of these blocks on the chromosome in terms of a Brownian CPP discussed in Section 1.2.4, under the same limiting regime.

This chapter is work in progress with Amaury Lambert and Emmanuel Schertzer.

1.3.3 Branching processes in epidemiology

The last part of the manuscript consists of two chapters that study related epidemiological models. The two models are based on the notion of general branching processes, also called Crump-Mode-Jagers processes (CMJ processes). General branching processes are population models where the ages at which individuals give birth can have a very general distribution. The life-history of each individual i is given by a random variable X_i living in some general spaces Ω. This variable X_i encodes the ages at which i dies and gives birth, and any desired characteristic of the life of i. The only requirement is that the variables (X_i) are i.i.d. for different individuals. We refer to [205] for an introduction to CMJ processes, and see Figure 1.7 for an illustration.

From individual-based models to McKendrick-von Foerster PDEs (I). In Chapter 6, we model the spread of COVID-19 with a CMJ process. Each birth in the population now represents a new infection, and the age of an individual is the time elapsed since her infection. The variable X_i is defined as a stochastic process

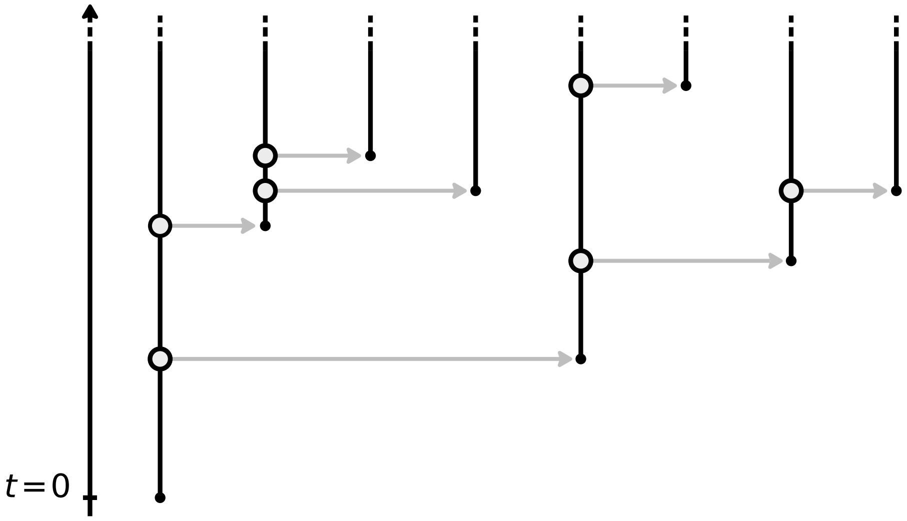


Figure 1.7: Illustration of a Crump-Mode-Jagers process. Each black vertical line represents an individual, and the grey dots correspond to the times at which this individual reproduces. At each such time, a new, independent individual is placed in the population.

$(X_i(a);\, a \geq 0)$ such that $X_i(a)$ is the *compartment* to which i belongs at age a. A compartment usually corresponds to a stage of the disease, but can also represent the health condition, the job category or the real age of an individual. Classical examples of compartments are the exposed compartment, when the individual is not yet infectious, the infectious compartment, or the removed compartment when the individual has recovered from the disease or is dead. Modelling the COVID-19 epidemic requires to add many other compartments, as discussed in Section 1.1.2.

Our main result proves the convergence of the empirical measure of ages and compartments in the population towards a deterministic scaling limit. In the large population size limit, the age structure converges to the solution of a PDE of the McKendrick-von Foerster type, and the number of individuals in each compartment is recovered by integrating the one-dimensional marginals of $(X(a);\, a \geq 0)$ against this age structure. Then, we use this expression to estimate some key parameters of the COVID-19 epidemic in France, such as the number of infected individuals and the basic reproduction number R_0, between March 2020 and May 2020, which corresponds to the lockdown period.

This chapter has been submitted to *Theoretical Population Biology* [77]. Due to my main contribution to this project, I have been listed as the first author.

From individual-based models to McKendrick-von Foerster PDEs (II). Chapter 7 is an extension of the previous model that accounts for the saturating number of individuals that are susceptible to the disease. The epidemiological model is the same as in Chapter 6, except that each infection is targeted towards a uniformly chosen individual in a population of size N. If this individual has already been infected, this infection is discarded, otherwise this individual becomes infected. As in Chapter 6, we provide the scaling limit of the empirical measure of ages and compartments in the population. We prove that, in the large population size limit, the age distribution of the population converges to a non-linear version of the McKendrick-von Foerster PDE derived in Chapter 6. If it is assumed that the process $(X(a);\, a \geq 0)$ is a Markov process, the latter PDE reduces to a system of ODE of the SIR type, which are popular models for the spread of diseases.

The previous chapter is work in progress with Jean-Jil Duchamps and Emmanuel Schertzer.

References for Chapter 1

[3] K. B. Athreya and P. E. Ney. *Branching Processes.* Grundlehren der mathematischen Wissenschaften. Springer-Verlag Berlin Heidelberg, 1972.

[6] Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, and M. Wang. Presumed asymptomatic carrier transmission of COVID-19. *JAMA* **323** (2020), 1406–1407.

[18] J. Berestycki. Exchangeable fragmentation-coalescence processes and their equilibrium measures. *Electronic Journal of Probability* **9** (2004), 770–824.

[19] J. Berestycki and N. Berestycki. Kingman's coalescent and Brownian motion. *ALEA, Latin American Journal of Probability and Mathematical Statistics* **6** (2009), 239–259.

[20] J. Bertoin. *Random Fragmentation and Coagulation Processes.* Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2006.

[21] J. Bertoin and J.-F. Le Gall. Stochastic flows associated to coalescent processes. *Probability Theory and Related Fields* **126** (2003), 261–288.

[23] F. Bienvenu, J.-J. Duchamps, and F. Foutel-Rodier. The Moran forest (2020). arXiv: 1906.08806.

[25] P. Billingsley. *Convergence of Probability Measures.* Second edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1999.

[26] M. Birkner, J. Blath, M. Capaldo, A. M. Etheridge, M. Möhle, J. Schweinsberg, and A. Wakolbinger. Alpha-stable branching and Beta-coalescents. *Electronic Journal of Probability* **10** (2005), 303–325.

[27] G. Birzu, O. Hallatschek, and K. S. Korolev. Fluctuations uncover a distinct class of traveling waves. *Proceedings of the National Academy of Sciences* **115** (2018), E3645–E3654.

[29] L. Bosshard, I. Dupanloup, O. Tenaillon, R. Bruggmann, M. Ackermann, S. Peischl, and L. Excoffier. Accumulation of deleterious mutations during bacterial range expansions. *Genetics* **207** (2017), 669–684.

[36] C. Cannings. The Latent Roots of Certain Markov Chains Arising in Genetics: A New Approach, I. Haploid Models. *Advances in Applied Probability* **6** (1974), 260–290.

[38] N. H. Chapman and E. A. Thompson. The effect of population history on the lengths of ancestral chromosome segments. *Genetics* **162** (2002), 449–458.

[41] J. A. Coyne and H. A. Orr. Patterns of speciation in Drosophila. *Evolution* **43** (1989), 362–381.

[43] A. Depperschmidt and A. Greven. Tree-valued Feller diffusion (2019). arXiv: `1904.02044`.

[49] P. Donnelly and T. G. Kurtz. Particle representations for measure-valued population models. *Annals of Probability* **27** (1999), 166–205.

[51] T. Duquesne and J.-F. Le Gall. *Random Trees, Lévy Processes and Spatial Branching Processes.* Astérisque, 2002.

[54] C. A. Edmonds, A. S. Lillie, and L. L. Cavalli-Sforza. Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences* **101** (2004), 975–979.

[55] A. M. Etheridge. *An Introduction to Superprocesses.* Vol. 20. University Lecture Series. American Mathematical Society, 2000.

[56] A. M. Etheridge. *Some Mathematical Models from Population Genetics. École d'Été de Probabilités de Saint-Flour XXXIX-2009.* Vol. 2012. Lecture Notes in Mathematics. Springer Science & Business Media, 2011.

[58] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence.* Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1986.

[61] W. J. Ewens. *Mathematical Population Genetics. I. Theoretical Introduction.* Second edition. Interdisciplinary Applied Mathematics. Springer, New York, NY, 2004.

[71] R. A. Fisher. A fuller theory of "junctions" in inbreeding. *Heredity* **8** (1954), 187–197.

[72] S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J. W. Eaton, M. Monod, P. N. Perez-Guzman, N. Schmit, L. Cilloni, K. E. C. Ainslie, M. Baguelin, A. Boonyasiri, O. Boyd, L. Cattarino, L. V. Cooper, Z. Cucunubá, G. Cuomo-Dannenburg, A. Dighe, B. Djaafara, I. Dorigatti, S. L. van Elsland, R. G. FitzJohn, K. A. M. Gaythorpe, L. Geidelberg, N. C. Grassly, W. D. Green, T. Hallett, A. Hamlet, W. Hinsley, B. Jeffrey, E. Knock, D. J. Laydon, G. Nedjati-Gilani, P. Nouvellet, K. V. Parag, I. Siveroni, H. A. Thompson, R. Verity, E. Volz, C. E. Walters, H. Wang, Y. Wang, O. J. Wat-

son, P. Winskill, X. Xi, P. G. T. Walker, A. C. Ghani, C. A. Donnelly, S. Riley, M. A. C. Vollmer, N. M. Ferguson, L. C. Okell, S. Bhatt, and Imperial College COVID-19 Response Team. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584** (2020), 257–261.

[77] F. Foutel-Rodier, F. Blanquart, P. Courau, P. Czuppon, J.-J. Duchamps, J. Gamblin, É. Kerdoncuff, R. Kulathinal, L. Régnier, L. Vuduc, A. Lambert, and E. Schertzer. From individual-based epidemic models to McKendrick-von Foerster PDEs: A guide to modeling and inferring COVID-19 dynamics (2020). arXiv: 2007.09622.

[78] F. Foutel-Rodier and A. M. Etheridge. The spatial Muller's ratchet: Surfing of deleterious mutations during range expansion. *Theoretical Population Biology* **135** (2020), 19–31.

[79] F. Foutel-Rodier, A. Lambert, and E. Schertzer. Exchangeable coalescents, ultrametric spaces, nested interval-partitions: A unifying approach (2019). arXiv: 1807.05165.

[80] F. Foutel-Rodier, A. Lambert, and E. Schertzer. Kingman's coalescent with erosion. *Electronic Journal of Probability* **25** (2020), 33 pp.

[91] A. Greven, P. Pfaffelhuber, and A. Winter. Convergence in distribution of random metric measure spaces (Λ-coalescent measure trees). *Probability Theory and Related Fields* **145** (2009), 285–322.

[94] M. Gromov. *Metric structures for Riemannian and non-Riemannian spaces.* Vol. 152. Progress in Mathematics. Birkhäuser Boston, 1999.

[99] O. Hallatschek and D. R. Nelson. Gene surfing in expanding populations. *Theoretical Population Biology* **73** (2008), 158–170.

[102] S. C. Harris, S. G. G. Johnston, and M. I. Roberts. The coalescent structure of continuous-time Galton–Watson trees. *Annals of Applied Probability* **30** (2020), 1368–1414.

[104] R. G. Harrison and E. L. Larson. Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity* **105** (2014), 795–809.

[105] A. Hastings. *Population Biology. Concepts and Models.* Springer-Verlag New York, 1997.

[106] X. He, E. H. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y. C. Lau, J. Y. Wong, Y. Guan, X. Tan, X. Mo, Y. Chen, B. Liao, W. Chen, F. Hu, Q. Zhang, M. Zhong, Y. Wu, L. Zhao, F. Zhang, B. J. Cowling, F. Li, and G. M. Leung. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine* **26** (2020), 672–675.

[108] B. M. Henn, L. L. Cavalli-Sforza, and M. W. Feldman. The great human expansion. *Proceedings of the National Academy of Sciences* **109** (2012), 17758–17764.

[109] J. Hey. Using phylogenetic trees to study speciation and extinction. *Evolution* **46** (1992), 627–640.

[119] S. G. G. Johnston. The genealogy of Galton-Watson trees. *Electronic Journal of Probability* **24** (2019), 35 pp.

[120] O. Kallenberg. *Foundations of Modern Probability*. Second edition. Probability and its Applications. Springer-Verlag New York, 2002.

[122] J. Kelleher, A. M. Etheridge, and G. A. T. McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Computational Biology* **12** (2016), 1–22.

[124] É. Kerdoncuff, A. Lambert, and G. Achaz. Testing for population decline using maximal linkage disequilibrium blocks. *Theoretical Population Biology* **134** (2020), 171–181.

[127] J. F. C. Kingman. The representation of partition structures. *Journal of the London Mathematical Society* **18** (1978), 374–380.

[128] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications* **13** (1982), 235–248.

[129] S. Klopfstein, M. Currat, and L. Excoffier. The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution* **23** (2006), 482–490.

[132] A. M. Kramer, B. Dennis, A. M. Liebhold, and J. M. Drake. The evidence for Allee effects. *Population Ecology* **51** (2009), 341–354.

[133] C. Labbé. From flows of Λ-Fleming-Viot processes to lookdown processes via flows of partitions. *Electronic Journal of Probability* **19** (2014), 49 pp.

[134] A. Lambert. The branching process with logistic growth. *Annals of Applied Probability* **15** (2005), 1506–1535.

[135] A. Lambert. Population Dynamics and Random Genealogies. *Stochastic Models* **24** (2008), 45–163.

[136] A. Lambert. The contour of splitting trees is a Lévy process. *Annals of Probability* **38** (2010), 348–395.

[138] A. Lambert. The coalescent of a sample from a binary branching process. *Theoretical Population Biology* **122** (2018), 30–35.

[140] A. Lambert and E. Schertzer. Recovering the Brownian coalescent point process from the Kingman coalescent by conditional sampling. *Bernoulli* **25** (2019), 148–173.

[141] A. Lambert and T. Stadler. Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. *Theoretical Population Biology* **90** (2013), 113–128.

[145] J.-F. Le Gall. *Brownian Motion, Martingales, and Stochastic Calculus*. Graduate Texts in Mathematics. Springer International Publishing, 2016.

[146] H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature* **475** (2011), 493–496.

[147] Z. Li. *Measure-Valued Branching Markov Processes*. Probability and Its Applications. Springer-Verlag Berlin Heidelberg, 2011.

[151] J. Mallet, N. Besansky, and M. W. Hahn. How reticulated are species? *BioEssays* **38** (2016), 140–149.

[152] J. Marin, G. Achaz, A. Crombach, and A. Lambert. The genomic view of diversification. *Journal of Evolutionary Biology* **33** (2020), 1387–1404.

[154] G. A. T. McVean and N. J. Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360** (2005), 1387–1393.

[155] M. Möhle. Robustness results for the coalescent. *Journal of Applied Probability* **35** (1998), 438–447.

[156] M. Möhle. Weak convergence to the coalescent in neutral population models. *Journal of Applied Probability* **36** (1999), 446–460.

[157] M. Möhle and S. Sagitov. A classification of coalescent processes for haploid exchangeable population models. *Annals of Probability* **29** (2001), 1547–1562.

[158] P. A. P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society* **54** (1958), 60–71.

[159] H. Morlon, M. D. Potts, and J. B. Plotkin. Inferring the dynamics of diversification: A coalescent approach. *PLOS Biology* **8** (2010), 1–13.

[164] H. Nishiura, T. Kobayashi, T. Miyama, A. Suzuki, S.-m. Jung, K. Hayashi, R. Kinoshita, Y. Yang, B. Yuan, A. R. Akhmetzhanov, and N. M. Linton. Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). *International Journal of Infectious Diseases* **94** (2020), 154–155.

[166] G. Pang and É. Pardoux. Functional limit theorems for non-Markovian epidemic models (2020). arXiv: `2003.03249`.

[169] É. Pardoux. *Probabilistic Models of Population Evolution. Scaling Limits, Genealogies and Interactions*. Stochastics in Biological Systems. Springer International Publishing, 2016.

[172] S. Peischl, I. Dupanloup, M. Kirkpatrick, and L. Excoffier. On the accumulation of deleterious mutations during range expansions. *Molecular Ecology* **22** (2013), 5972–5982.

[175] E. Pennisi. Shaking up the Tree of Life. *Science* **354** (2016), 817–821.

[176] P. Pfaffelhuber and A. Wakolbinger. The process of most recent common ancestors in an evolving coalescent. *Stochastic Processes and their Applications* **116** (2006), 1836–1859.

[177] P. Pfaffelhuber, A. Wakolbinger, and H. Weisshaupt. The tree length of an evolving coalescent. *Probability Theory and Related Fields* **151** (2011), 529–557.

[178] J. Pitman. Coalescents with multiple collisions. *The Annals of Probability* **27** (1999), 1870–1902.

[179] L. Popovic. Asymptotic genealogy of a critical branching process. *Annals of Applied Probability* **14** (2004), 2120–2148.

[180] P. Ralph and G. Coop. The geography of recent genetic ancestry across Europe. *PLOS Biology* **11** (2013), 1–20.

[184] L. Roques, J. Garnier, F. Hamel, and É. K. Klein. Allee effect promotes diversity in traveling waves of colonization. *Proceedings of the National Academy of Sciences* **109** (2012), 8828–8833.

[187] P. C. Sabeti, D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419** (2002), 832–837.

[188] S. Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability* **36** (1999), 1116–1125.

[189] H. Salje, C. Tran Kiem, N. Lefrancq, N. Courtejoie, P. Bosetti, J. Paireau, A. Andronico, N. Hozé, J. Richet, C.-L. Dubost, Y. Le Strat, J. Lessler, D. Levy-Bruhl, A. Fontanet, L. Opatowski, P.-Y. Boelle, and S. Cauchemez. Estimating the burden of SARS-CoV-2 in France. *Science* **369** (2020), 208–211.

[190] S. Sankararaman, S. Mallick, M. Dannemann, K. Prüfer, J. Kelso, S. Pääbo, N. Patterson, and D. Reich. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507** (2014), 354–357.

[191] S. Sankararaman, S. Mallick, N. Patterson, and D. Reich. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Current Biology* **26** (2016), 1241–1247.

[195] J. Schweinsberg. Coalescents with Simultaneous Multiple Collisions. *Electronic Journal of Probability* **5** (2000), 50 pp.

[196] J. Schweinsberg. Coalescent processes obtained from supercritical Galton–Watson processes. *Stochastic Processes and their Applications* **106** (2003), 107–139.

[199] Z. Shi. *Branching Random walks. École d'Été de Probabilités de Saint-Flour XLII-2012.* Vol. 2151. Lecture Notes in Mathematics. Springer, Cham, 2015.

[205] Z. Taïb. *Branching Processes and Neutral Evolution.* Vol. 93. Lecture Notes in Biomathematics. Springer Berlin Heidelberg, 1992.

[208] R. Tingley, M. Vallinoto, F. Sequeira, and M. R. Kearney. Realized niche shift during a global biological invasion. *Proceedings of the National Academy of Sciences* **111** (2014), 10233–10238.

[209] Z.-D. Tong, A. Tang, K.-F. Li, P. Li, H.-L. Wang, J.-P. Yi, Y.-L. Zhang, and J.-B. Yan. Potential presymptomatic transmission of SARS-CoV-2, Zhejiang province, China, 2020. *Emerging Infectious Disease journal* **26** (2020), 1052.

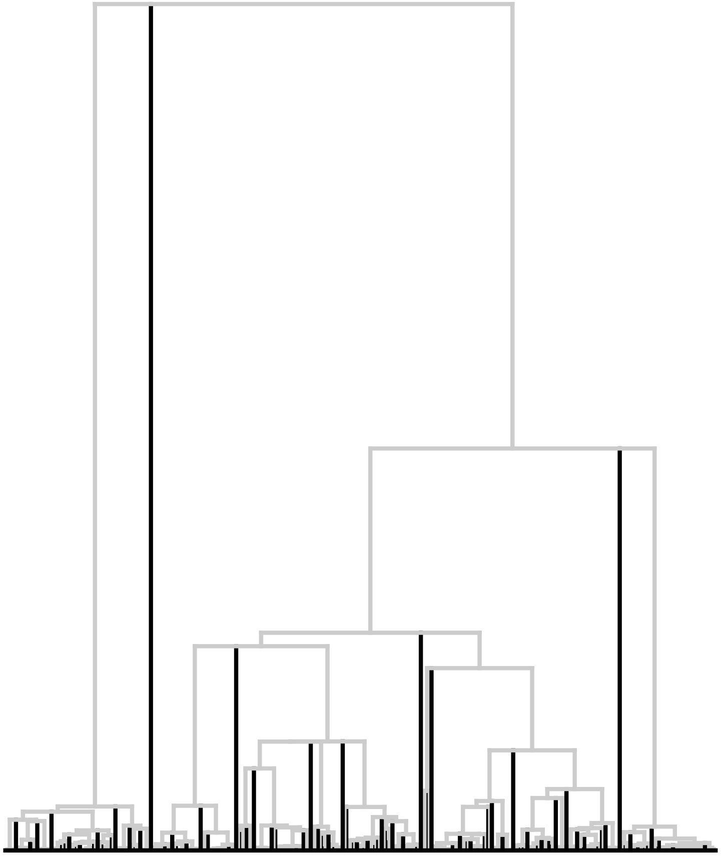
[211] J. M. J. Travis, T. Münkemüller, O. J. Burton, A. Best, C. Dytham, and K. Johst. Deleterious mutations can surf to high densities on the wave front of an expanding population. *Molecular Biology and Evolution* **24** (2007), 2334–2343.

[212] M. C. Urban, B. L. Phillips, D. K. Skelly, and R. Shine. The cane toad's (*Chaunus* [*Bufo*] *marinus*) increasing ability to invade Australia is revealed by a dynamically updated range model. *Proceedings of the Royal Society B: Biological Sciences* **274** (2007), 1413–1419.

[216] J. Wakeley. *Coalescent theory. An introduction.* Roberts & Company Publishers, 2008.

Exchangeable partition processes

Chapter 2



2

Exchangeable coalescents, ultrametric spaces, nested interval-partitions: A unifying approach

This chapter is joint work with Amaury Lambert and Emmanuel Schertzer. It has been accepted for publication in the *Annals of Applied Probability* [79].

Illustration. Simulation of a Kingman comb, whose transition rates are given in Remark 2.32. The black vertical lines correspond to the teeth of the comb, and the corresponding ultrametric tree is pictured in grey.

2.1 Introduction

2.1.1 Ultrametric spaces and exchangeable coalescents

In this paper we extend earlier work from [142] on the comb representation of ultrametric spaces. An ultrametric space is a metric space (U, d) such that the metric d fulfills the additional assumption

$$\forall x, y, z \in U, \quad d(x, y) \leq \max(d(x, z), d(z, y)).$$

In applications, ultrametric spaces are used to model the genealogy of entities co-existing at the same time. The distance between two points x and y of an ultrametric space is interpreted as the time to the most recent common ancestor (MRCA) of x and y. For instance, in population genetics ultrametric spaces model the genealogy of homologous genes in a population. Another example can be found in phylogenetics where ultrametric spaces are used to model the evolutionary relationships between species.

In population genetics and more generally in biology we do not have access to the entire population (that is to the entire ultrametric space) but only to a sample from the population. To model the procedure of sampling we equip the ultrametric space with a probability measure μ (also referred to as the sampling measure), yielding the notion of ultrametric measure spaces.

Definition 2.1. A quadruple $(U, d, \mathscr{U}, \mu)$ is called an ultrametric measure space (UMS) if the following holds.

(i) The distance d is an ultrametric on U which is $\mathscr{U} \otimes \mathscr{U}$ measurable.

(ii) The measure μ is a probability measure defined on $\mathscr{U}$.

(iii) The family $\mathscr{U}$ is a σ-field such that

$$\forall x \in U, \forall t > 0, \quad \{y \in U : d(x,y) < t\} \in \mathscr{U}$$

and $\mathscr{U} \subseteq \mathscr{B}(U)$, where $\mathscr{B}(U)$ is the usual Borel σ-field of (U, d).

If $\mathscr{U} = \mathscr{B}(U)$, we say that $(U, d, \mathscr{U}, \mu)$ is a Borel UMS. ◦

Remark 2.2. This definition might be surprising as we would naively expect a UMS to be any ultrametric space with a probability measure on its Borel σ-field. However the previous naive definition is not satisfying for several reasons, that are exposed in Section 2.4.1. Notice that if (U, d) is separable, then $\mathscr{U} = \mathscr{B}(U)$ and point (i) always holds. We thus recover the usual definition of an ultrametric measure space. ◦

A sample from a UMS is an i.i.d. sequence $(X_i)_{i \geq 1}$ distributed according to μ. The genealogy of the sample is usually encoded as a partition-valued process, $(\Pi_t)_{t \geq 0}$ called a *coalescent*. For any time $t \geq 0$, the blocks of the partition Π_t are given by the following relation

$$i \sim_{\Pi_t} j \iff d(X_i, X_j) \leq t. \tag{2.1}$$

The process $(\Pi_t)_{t \geq 0}$ has two major features. First a well-known characteristic of ultrametric spaces is that for a given t the balls of radius t form a partition of the space that gets coarser as t increases. This implies that given $s \leq t$, the partition Π_t is coarser than Π_s. Second, if σ denotes a finite permutation of $\mathbb{N}$ and $\sigma(\Pi_t)$ is the partition of $\mathbb{N}$ whose blocks are the images by σ of the blocks of Π_t, we have

$$(\Pi_t)_{t \geq 0} \overset{(\mathrm{d})}{=} (\sigma(\Pi_t))_{t \geq 0}.$$

We call any càdlàg partition valued process that fulfills these two conditions an *exchangeable coalescent* (note that the process $(\Pi_t)_{t \geq 0}$ is not necessarily Markovian).

2.1.2 Combs in the compact case

Combs and ultrametric spaces. In this section, we address similar questions in the much simpler framework of comb metric spaces which have been introduced recently by [142] to represent *compact* ultrametric spaces. A comb is a function

$$f \colon [0, 1] \to \mathbb{R}_+$$

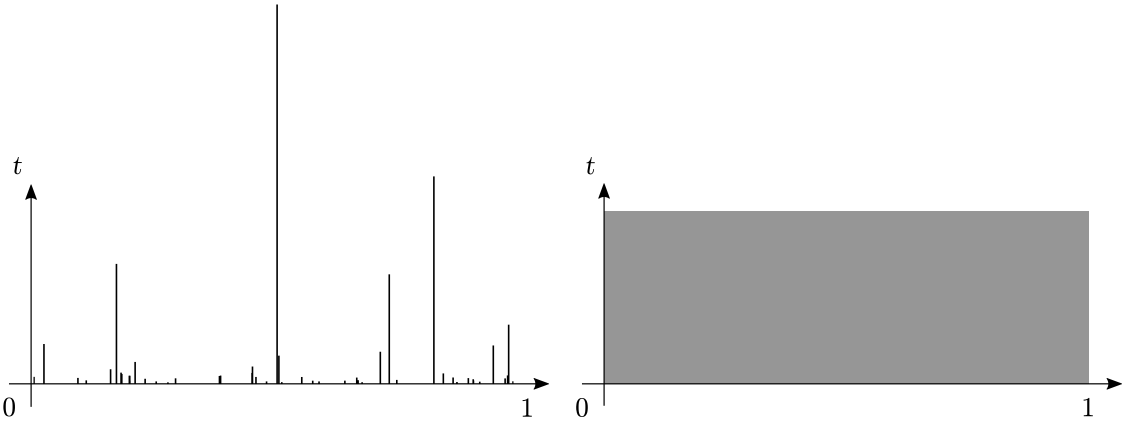


Figure 2.1: Representation of two nested interval-partitions. A point (x, t) is plotted in dark if $x \notin I_t$. Left panel: A realization of the Kingman comb, a tooth of size y at location x represents that $f(x) = y$. Right panel: The star-tree comb, an example of a nested interval-partition that cannot be represented as an original comb.

such that for any $\varepsilon > 0$ the set $\{f \geq \varepsilon\}$ is finite (see Figure 2.1 left panel). To any comb is associated a comb metric d_f on $[0, 1]$ defined as

$$\forall x, y \in [0,1], \quad d_f(x,y) = \mathbb{1}_{\{x \neq y\}} \sup_{[x \wedge y, x \vee y]} f.$$

In general d_f is only a pseudo-metric on $[0, 1]$ and it is easy to verify that it is actually ultrametric. One of the main results in [142] shows that any compact ultrametric space is isometric to a properly completed and quotiented comb metric space (see Theorem 3.1 in [142]).

Exchangeable coalescents. We also will be interested in the relation between combs and exchangeable coalescents. Any comb metric space $([0, 1], d_f)$ can be naturally endowed with the Lebesgue measure on $[0, 1]$. Sampling from a comb can be seen as a direct extension of Kingman's paintbox procedure. More precisely, given a comb f, we can generate an exchangeable coalescent $(\Pi_t)_{t \geq 0}$ by throwing i.i.d. uniform random variables $(X_i)_{i \geq 1}$ on $[0, 1]$ and declaring that

$$i \sim_{\Pi_t} j \iff \sup_{[X_i \wedge X_j, X_i \vee X_j]} f \leq t.$$

For the sake of illustration, we recall the comb representation of the Kingman coalescent stated in [128]. The Kingman comb is constructed out of an i.i.d. sequence $(e_i)_{i \geq 1}$ of exponential variables with parameter 1, and of an independent i.i.d. sequence $(U_i)_{i \geq 1}$ of uniform variables on $[0, 1]$. We define the sequence $(T_i)_{i \geq 2}$ as

$$T_i = \sum_{j \geq i} \frac{2}{j(j-1)} e_j.$$

The Kingman comb f_K is defined as

$$f_K = \sum_{i \geq 2} T_i \mathbb{1}_{U_i}.$$

See Figure 2.1 left panel for an illustration of a realization of the Kingman comb. The paintbox based on f_K is a version of the Kingman coalescent (see Section 4.1.3 of [20]).

More generally, the assumption that $\{f \geq \varepsilon\}$ is finite implies that the coalescent $(\Pi_t)_{t\geq 0}$ obtained from a paintbox based on f has only finitely many blocks for any $t > 0$. This property is usually referred to as "coming down from infinity". It has been shown in [137] that any coalescent which comes down from infinity can be represented as a paintbox based on a comb, see Proposition 3.2.

2.1.3 General combs

One of the objectives of this work is to extend Theorem 3.1 of [142] and Proposition 3.2 of [137] to any ultrametric space (not only compact) and to any exchangeable coalescent (i.e., beyond the "coming down from infinity" property). From a technical point of view, we note that this extension is conceptually harder, and requires the technology of exchangeable nested compositions which were absent in [142]. This point will be discussed further in Section 2.2.1.

In order to deal with non-compact metric spaces, we need to generalize the definition of a comb by relaxing the condition on the finiteness of $\{f \geq \varepsilon\}$. We will encode combs as functions taking values in the open subsets of $(0,1)$. Any open subset I of $(0,1)$ can be decomposed into an at-most countable union of disjoint intervals denoted by $(I_i)_{i\geq 1}$. For this reason we will call an open subset of $(0,1)$ an *interval-partition* and each of the intervals I_i is an *interval component* of I. The space of interval-partitions is conveniently topologized with the Hausdorff distance on the complement, d_H, defined as

$$d_H(I, \tilde{I}) = \sup\Big\{d(x, [0,1] \setminus \tilde{I}), x \notin I\Big\} \vee \sup\Big\{d(x, [0,1] \setminus I), x \notin \tilde{I}\Big\}.$$

We propose to generalize the notion of comb to the notion of *nested interval-partition.*

Definition 2.3. A nested interval-partition is a càdlàg function $(I_t)_{t\geq 0}$ taking values in the open subsets of $(0,1)$ verifying

$$\forall s \leq t, \quad I_s \subseteq I_t.$$

Sometimes nested interval-partitions will be called generalized combs or even simply combs. ○

Let us briefly see how this definition extends the initial comb of [142]. Starting from a comb function f, we can build a nested interval-partition $(I_t)_{t\geq 0}$ as follows

$$\forall t > 0, \quad I_t = \{f < t\} \setminus \{0, 1\}$$

and

$$I_0 = \mathrm{int}(\{f = 0\})$$

where $\mathrm{int}(A)$ denotes the interior of the set A.

Conversely if $(I_t)_{t\geq 0}$ is a nested interval-partition we can define a comb function $f_I\colon [0,1] \to \mathbb{R}_+$ as

$$f_I(x) = \inf\{t \geq 0 : x \in I_t\}.$$

In general f_I does not fulfill that $\{f_I \geq t\}$ is finite. A necessary and sufficient condition for this to hold is that for any $t > 0$, I_t has finitely many interval components, and the summation of their lengths is 1. If the latter condition is fulfilled, we say that I_t is proper or equivalently that it has no dust.

A nested interval-partition naturally encodes a (pseudo-)ultrametric d_I on $[0,1]$ defined as

$$\begin{aligned} d_I(x,y) &= \inf\{t \geq 0 : x \text{ and } y \text{ belong to the same interval of } I_t\} \\ &= \sup_{[x,y]} f_I \end{aligned}$$

for $x < y$. We call the ultrametric space $([0,1], d_I)$ the *comb metric space* associated to $(I_t)_{t\geq 0}$. In order to turn $([0,1], d_I)$ into a UMS, we need to define an appropriate σ-field and a sampling measure. The interval $[0,1]$ is naturally endowed with the usual Borel σ-field $\mathscr{B}([0,1])$ and the Lebesgue measure. However, the usual Borel σ-field does not fulfill the requirements of Definition 2.1 in general because two points that belong to the same interval component of I_0 are indistinguishable in the metric d_I. This can be addressed by considering a slightly smaller σ-field as follows.

Let $(I_i^0)_{i\geq 1}$ be the interval components of I_0. We define a σ-field $\mathscr{I}$ on $[0,1]$ as

$$\mathscr{I} = \Big\{ A \cup \bigcup_{i\in M} I_i^0 : A \in \mathscr{B}([0,1] \setminus I_0) \text{ and } M \subseteq \mathbb{N} \Big\}$$

where $\mathscr{B}([0,1] \setminus I_0)$ denotes the usual Borel σ-field on $[0,1] \setminus I_0$. It is clear that $\mathscr{I} \subseteq \mathscr{B}([0,1])$. We call a *comb metric measure space* associated to $(I_t)_{t\geq 0}$ the quadruple $([0,1], d_I, \mathscr{I}, \mathrm{Leb})$, where Leb is the restriction of the Lebesgue measure to $\mathscr{I}$. The following lemma shows that the Lebesgue measure on $\mathscr{I}$ satisfies the requirements of Definition 2.1, and that a comb metric measure space is a UMS.

Lemma 2.4. *Any comb metric measure space* $([0,1], d_I, \mathscr{I}, \mathrm{Leb})$ *is a UMS.*

Proof. Let us first prove that (iii) holds. For $x \in [0,1]$ and $t \geq 0$, let $I_t(x)$ denote the interval component of I_t to which x belongs if $x \in I_t$, or let $I_t(x) = \{x\}$ else. Then for $t > 0$ we have

$$\{y \in [0,1] : d_I(x,y) < t\} = \bigcup_{s<t} I_s(x) \in \mathscr{I}.$$

It remains to show that $\mathscr{I} \subseteq \mathscr{B}_I([0,1])$, where $\mathscr{B}_I([0,1])$ denotes the σ-field induced by d_I. It is sufficient to prove that for all $x, y \notin I_0$, we have $(x,y) \in \mathscr{B}_I([0,1])$. Let $z \in (x,y)$ and suppose that $z \in I_t$ for all $t > 0$. Then $I_{t_z}(z) \subseteq (x,y)$ for a small enough t_z, and thus

$$\{z' \in [0,1] : d_I(z,z') < t_z\} \subseteq (x,y).$$

Otherwise if $z \notin I_{t_z}(z)$ for some t_z, then $\{z' \in [0,1] : d_I(z,z') < t_z\} = \{z\}$. We can now write

$$(x,y) = \bigcup_{z \in (x,y)} \{z' \in [0,1] : d_I(z,z') < t_z\} \in \mathscr{B}_I([0,1])$$

which proves that point (iii) of the definition is fulfilled.

Let $I_{t-} = \bigcup_{s<t} I_s$, then

$$\Big\{(x,y) \in [0,1]^2 : d(x,y) < t\Big\} = \Delta_0 \cup \bigcup_{x \in I_{t-}} I_t(x) \times I_t(x),$$

where $\Delta_0 = \{(x,y) \in ([0,1] \setminus I_0)^2 : x = y\}$. As there are only countably many interval components of I_t, the union on the right-hand side is countable, and this set belongs to the product $\mathscr{I} \otimes \mathscr{I}$. This proves that point (i) holds and that the comb metric measure space is a UMS. □

For later purpose, let us denote by U_I the completion of the quotient space of $\{f_I = 0\}$ by the relation $x \sim y$ iff $d_I(x,y) = 0$. (This completion can be realized explicitly by adding countably many "left" and "right" faces to the comb, see Section 2.4.5.)

Finally, as in the compact case, an exchangeable coalescent $(\Pi_t)_{t\geq 0}$ can be obtained from a nested interval-partition $(I_t)_{t\geq 0}$ out of an i.i.d. uniform sequence $(X_i)_{i\geq 1}$ by defining

$$i \sim_{\Pi_t} j \iff X_i \text{ and } X_j \text{ belong to the same interval component of } I_t. \qquad (2.2)$$

Notice that this definition is a multidimensional extension of the original Kingman paintbox procedure, see e.g. the beginning of Section 2.3.2 of [20].

Remark 2.5. The coalescent obtained through this sampling procedure is not càdlàg in general. As a coalescent is a non-decreasing process, we can (and will) always suppose that we work with a càdlàg modification of the coalescent. ○

Remark 2.6. We have defined two natural ways of sampling a coalescent from a nested interval-partition. First, one can realize the extended paintbox procedure described in equation (2.2). Second, one can consider the comb metric measure space associated to the nested interval-partition and sample the coalescent according to equation (2.1). It is not hard to see that the coalescent obtained through (2.1) is the càdlàg version of the one obtained through (2.2). ○

We will now demonstrate that nested interval-partitions form a large enough framework to answer our two initial problems: representing any exchangeable coalescent as a paintbox on a comb and representing general ultrametric measure spaces.

2.1.4 Comb representation of exchangeable coalescents

General comb representation. We start by showing that one can always find a comb representation of any coalescent. First notice that this representation cannot be unique. For example taking the reflection of a comb about the vertical line in the middle of the segment $[0,1]$ yields a new comb but does not change the associated coalescent. In many applications we will not be interested in this order but only in the genealogical structure of the comb. For this reason we introduce the following relation.

Definition 2.7. Two generalized combs are paintbox-equivalent if their associated coalescents are identical in law. Being paintbox-equivalent is an equivalence relation, we denote by $\mathfrak{I}$ the quotient space. ○

Given $I \in \mathfrak{I}$ we denote by ρ_I the distribution on the space of coalescents of the paintbox based on any representative of $\mathfrak{I}$. We provide the following version of Kingman's representation theorem (e.g. see [20] Theorem 2.1) for exchangeable coalescents.

Theorem 2.8. *Let $(\Pi_t)_{t\geq 0}$ be an exchangeable coalescent. There exists a unique distribution ν on $\mathfrak{I}$ such that*

$$\mathbb{P}\Big((\Pi_t)_{t\geq 0} \in \cdot\Big) = \int_{\mathfrak{I}} \rho_I(\cdot)\nu(\mathrm{d}I).$$

Remark 2.9. It is interesting to relate this result to the original theorem from Kingman. A *mass-partition* is a sequence $\beta = (\beta_i)_{i\geq 1}$ such that

$$\beta_1 \geq \beta_2 \geq \cdots \geq 0, \quad \sum_{i\geq 1} \beta_i \leq 1.$$

Kingman's representation theorem states that any exchangeable partition can be obtained through a paintbox based on a random mass-partition, and that this correspondence is bijective. A mass-partition can be seen as the ranked sequence of the lengths of the interval components of an interval-partition. Now notice that two interval-partitions are paintbox-equivalent, i.e. induce the same exchangeable partition, iff they have the same associated mass-partition. In this one-dimensional setting, any paintbox-equivalence class of interval-partitions can be identified with a random mass-partition. In a similar way, it would be natural to try to identify the elements of $\mathfrak{I}$ with mass-partition valued processes, also called *mass-coalescents.* However, one can easily find two different equivalence classes of $\mathfrak{I}$ that have the same associated mass-coalescent, see Figure 2.2. ○

Remark 2.10. A result very similar to Theorem 2.8 has been obtained in [76], Theorem 4, in the context of hierarchies. Roughly speaking, an exchangeable hierarchy is obtained from an exchangeable coalescent by "forgetting about time". In this sense, an exchangeable coalescent carries more information, and this part of our work can be seen as an extension of [76]. However, the forthcoming Section 2.3 and Section 2.4 heavily rely on the knowledge of the coalescence times, and

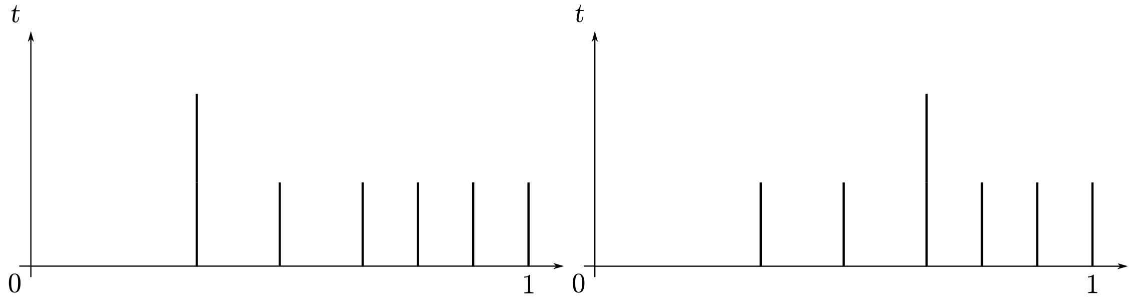


Figure 2.2: An example of two nested interval-partitions that have the same mass-coalescent but different coalescents. For both processes, the initial mass-partition is $(\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, 0, \dots)$, then $(\frac{2}{3}, \frac{1}{3}, 0, \dots)$ and finally $(1, 0, \dots)$. However, for the process on the left-hand side the first blocks to merge are those of mass $1/6$ and $1/9$, whereas for the right-hand process, the blocks of mass $1/6$ first merge with the block of size $1/3$.

could not have been achieved in the framework of hierarchies. We have dedicated Section 2.A to the explanation of the links between the present work and [76]. ∘

Λ-coalescents. Most of the efforts made in the study of exchangeable coalescents have been devoted to the special case of Λ-coalescents [178, 188]. These coalescents are parametrized by a finite measure Λ on $[0,1]$, and their restriction to $[n] := \{1, \dots, n\}$ is a Markov chain whose transitions are the following. The process undergoes a transition from a partition π with b blocks to a partition obtained by merging k blocks of π at rate $\lambda_{b,k}$ given by

$$\lambda_{b,k} = \int_{[0,1]} x^{k-2}(1-x)^{b-k}\Lambda(\mathrm{d}x).$$

The next proposition states that we can always find a Markovian comb representation of a Λ-coalescent. Moreover in Section 2.3 we provide an explicit description of its transition.

Proposition 2.11. *Let $(\Pi_t)_{t\geq 0}$ be a Λ-coalescent. There exists $(I_t)_{t\geq 0}$ a Markov nested interval-partition such that the coalescent obtained from the paintbox based on $(I_t)_{t\geq 0}$ is distributed as $(\Pi_t)_{t\geq 0}$.*

Remark 2.12 (Combs and the flow of bridges)**.** The flow of bridges introduced by [21] represents the dynamics of a population whose genealogy is given by a Λ-coalescent. We will show that we can build a nested interval-partition from the flow of bridges and that it has the same distribution as the Markov nested interval-partition of Proposition 2.11, see Section 2.3. ∘

Remark 2.13. There exists a natural extension of the Λ-coalescents called the coalescents with simultaneous multiple collisions or Ξ-coalescents [195]. All our results carry over to Ξ-coalescents, however for the sake of clarity we will focus on the case of Λ-coalescents. ∘

A coalescent process models the genealogy of a population living at a fixed observation time. Many works have been concerned with the dynamical genealogy obtain by varying the observation time of the population. For example, in [176, 177] the authors study some statistics of the dynamical genealogy, namely the time to the MRCA and the total length of the genealogy. In [92] the genealogy is encoded as a metric space (a real tree, see [59]) and the authors introduce the tree-valued Fleming-Viot process, a process bearing the entire information on the dynamical genealogy. This encoding requires to work with metric space-valued stochastic processes, and with the rather technical Gromov-weak topology for metric spaces.

We address such questions in the framework of combs in Section 2.3.3. We show that we can naturally encode a dynamical genealogy as a comb-valued process called the *evolving comb*. This process is a Markov process, whose semi-group can be explicitly described. In the particular case of coalescents that come down from infinity, the semi-group of the evolving comb takes a particularly simple form in terms of sampling from an independent comb.

2.1.5 Comb representation of ultrametric spaces

The second main aim of this paper is to provide a comb representation of ultrametric measure spaces in the same vein as Theorem 3.1 of [142]. We will only state our results informally and refer to Section 2.4 for the precise statements.

We first introduce the *Gromov-weak topology* on the space of UMS and show that any UMS is indistinguishable from a comb metric space in this topology. To do so, we realize a straightforward extension of the work developed in [91, 94] which is focused on separable metric measure spaces. In short, starting from a UMS we can obtain a coalescent by sampling from it as described in Section 2.1.1. We say that a sequence of UMS converges to a limiting UMS in the Gromov-weak sense if the corresponding coalescents converge weakly as partition-valued stochastic processes (see Section 2.4.2 for a more precise definition). We are now ready to state our representation result, which is a direct application of Theorem 2.8.

Theorem 2.14. *For any UMS $(U, d, \mathscr{U}, \mu)$ there exists a comb metric measure space that is indistinguishable in the Gromov-weak topology from $(U, d, \mathscr{U}, \mu)$.*

Proof. As we have identified any UMS with the distribution of its coalescent, two UMS are indistinguishable iff their coalescents have the same distribution. Theorem 2.8 shows that we can always find a nested interval-partition $(I_t)_{t\geq 0}$ such that the coalescent obtained from a paintbox based on $(I_t)_{t\geq 0}$ is distributed as the coalescent obtained by sampling from $(U, d, \mathscr{U}, \mu)$. As noticed in Remark 2.6, the coalescent obtained by sampling in the comb metric measure space $([0,1], d_I, \mathscr{I}, \mathrm{Leb})$ has the same distribution as the coalescent obtained from the paintbox based on $(I_t)_{t\geq 0}$, and thus this comb metric measure space is indistinguishable from $(U, d, \mathscr{U}, \mu)$. □

The comb representation given by Theorem 2.14 is rather weak, since it only ensures that we can find a comb that has the same sampling structure as a given UMS. We would like to be more precise and obtain an isometry result as in the compact case. This is not possible in general, and we have to consider separately the separable case and the non-separable case.

The separable case. In the separable case, the coalescent contains all the information about the UMS. More precisely, the Gromov reconstruction theorem ensures that two complete separable UMS that are indistinguishable in the Gromov-weak topology have the supports of their measures in isometry, see e.g. [94], Section $3.\frac{1}{2}.5$ or [91], Proposition 2.6. The following refinement of Theorem 2.14 in the separable case is a direct consequence of the Gromov reconstruction theorem and of Theorem 2.14, see Section 2.4.6 for a proof.

Corollary 2.15. *Let $(U, d, \mathscr{U}, \mu)$ be a complete separable UMS. There exists a comb metric measure space $(U_I, d_I, \mathscr{I}, \mathrm{Leb})$ such that the support of μ is isometric to (U_I, d_I), and such that the isometry maps μ to* Leb*.*

Additionally, any separable ultrametric space (U, d) can be endowed with a probability measure whose support is the whole space U, see Lemma 2.53. This result combined with Corollary 2.15 yields the following representation result for complete separable ultrametric spaces, which is the direct extension of Theorem 3.1 of [142] to the separable case.

Proposition 2.16. *Let (U, d) be a complete separable ultrametric space. We can find a nested interval-partition such that (U_I, d_I) is isometric to (U, d).*

A proof of this proposition is provided in Section 2.4.6. Notice that the proof of the previous proposition is very different from the original proof of [142] which is no longer valid for non-compact UMS.

The general case. In general, two UMS that are associated to the same coalescent are not isometric. This essentially comes from the fact that a coalescent only bears the information about a sequence of "typical" points of the UMS, and that a non-separable UMS may contain more information than the topology generated by these "typical" points. The main idea of our approach relies on a new decomposition that we now expose.

A UMS $(U, d, \mathscr{U}, \mu)$ can be seen as the leaves of a tree. We show that we can decompose this tree into two parts. The first part is a separable tree that we call the backbone. Secondly, one can then recover the tree from the backbone by grafting some "simple" subtrees on the backbone. By "simple", we mean that each of those subtrees has the sampling properties of a star-tree, in the sense that all points sampled in the same subtree are at the same distance to each other. See Figure 2.3 for an illustration of this decomposition, and Definition 2.43 for a precise definition of the backbone. An object very similar to the backbone is studied in [95] but the construction of the backbone from a general UMS is not considered there.

Our result states that if two UMS have complete backbones and are associated to the same coalescent, then the backbones are in isometry in a way that preserves the star-trees attached to it. We say that the two UMS are in *weak isometry*, see

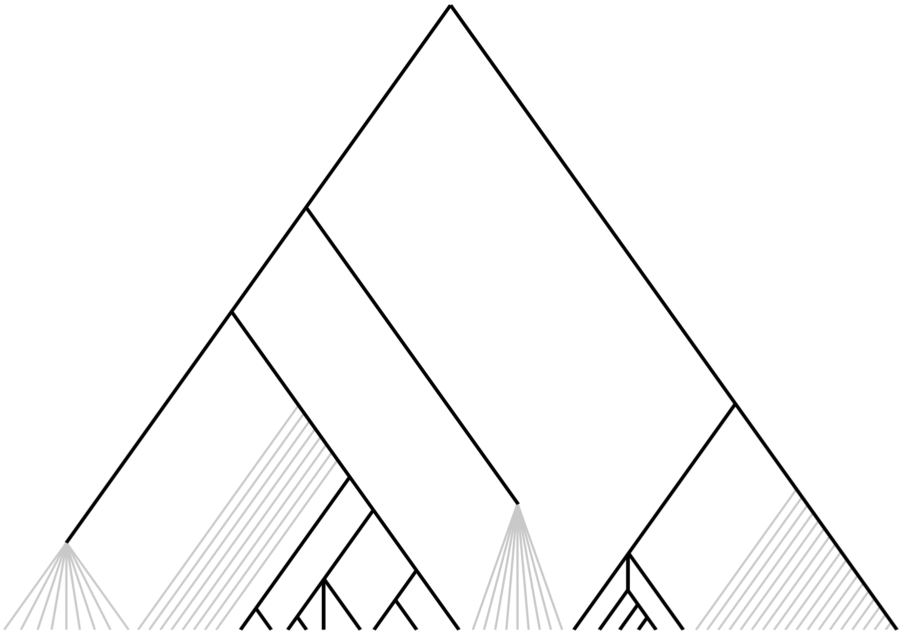


Figure 2.3: Illustration of the backbone decomposition. The dark thick lines represent the backbone. An element of the tree is represented in grey if its descendance has zero mass.

Definition 2.46. We provide the following version of the Gromov reconstruction theorem in the case of general UMS.

Proposition 2.17. *Let $(U, d, \mathscr{U}, \mu)$ and $(U', d', \mathscr{U}', \mu')$ be two UMS with complete backbones. These UMS are indistinguishable in the Gromov-weak topology iff $(U, d, \mathscr{U}, \mu)$ and $(U', d', \mathscr{U}', \mu')$ are in weak isometry.*

An equivalent reformulation of the previous proposition is stated in Section 2.4.4, see Proposition 2.47, and proved at the end of Section 2.4.4. As a consequence of Proposition 2.17 and Theorem 2.14, we have the following version of Theorem 3.1 of [142] in the general case. See Section 2.4.5 for a proof.

Corollary 2.18. *Let $(U, d, \mathscr{U}, \mu)$ be a UMS with a complete backbone. There exists a nested interval-partition $(I_t)_{t\geq 0}$ such that, up to the addition of a countable number of points, the comb metric measure space $([0,1], d_I, \mathscr{I}, \mathrm{Leb})$ is weakly isometric to $(U, d, \mathscr{U}, \mu)$.*

2.1.6 Outline

The rest of the paper is divided into three parts. In Section 2.2 we introduce the notion of composition and nested composition which will be our main tool to study combs. Section 2.2.1 introduces the existing material on random compositions. In Section 2.2.2 we define exchangeable nested compositions and prove the representation theorem linking combs and nested compositions. The proof of Theorem 2.8 is given in Section 2.2.3. In Section 2.3 we restrict our attention to the case of Λ-coalescents. We define there the notion of a Λ-comb and study a family of nested compositions emerging from the Λ-coalescents. The proof of Proposition 2.11 is given in Section 2.3.2. The evolving comb is introduced and studied in Section 2.3.3. Finally in Section 2.4 we envision combs as ultrametric spaces. A precise outline of this section is given at the beginning of Section 2.4.

2.2 Combs and nested compositions

The objective of this section is to prove Theorem 2.8 on the comb representation of exchangeable coalescents. As was already mentioned in introduction, the correspondence between combs and exchangeable coalescents cannot be bijective. Roughly speaking, this comes from the fact that a nested interval-partition inherits an order from $[0,1]$, and that changing this order does not modify the associated coalescent. However, we will show in Section 2.2.2 that there is a bijective correspondence between nested interval-partitions and exchangeable nested compositions, the ordered version of exchangeable coalescents. Exchangeable nested compositions will be our main tool to study combs.

We start this section by recalling existing results and material on exchangeable compositions developed in [86, 48] and then show how to extend them to nested compositions.

2.2.1 Exchangeable compositions

In combinatorics, a composition of $[n]$ (resp. $\mathbb{N}$) is a partition of $[n]$ (resp. $\mathbb{N}$) with a total order on the blocks. We write $\mathcal{C} = (\pi, \leq)$ for a composition of $\mathbb{N}$ where π is the partition and $\leq$ the order on the blocks. The blocks of the partition π can always be labeled in increasing order of their least element, i.e. the blocks of π are denoted by $(A_1, A_2, \dots)$ and are such that for any $i, j \geq 1$,

$$i \leq j \iff \min(A_i) \leq \min(A_j).$$

Let σ be a finite permutation of $\mathbb{N}$, we denote by $\sigma(\mathcal{C})$ the composition whose blocks are $(\sigma(A_1), \sigma(A_2), \dots)$ and such that the order of the blocks is

$$\sigma(A_i) \leq \sigma(A_j) \iff A_i \leq A_j.$$

For example, for $n = 5$, consider $\mathcal{C}^n$ the composition

$$\mathcal{C}^n = \{2,3\} \leq \{5\} \leq \{1,4\}.$$

With our labeling convention, we have $A_1 = \{1,4\}$, $A_2 = \{2,3\}$ and $A_3 = \{5\}$ (A_1 needs not be the first block of $\mathcal{C}$ for the order $\leq$). If $\sigma = (2,1,3,5,4)$, the composition $\sigma(\mathcal{C}^n)$ is given by

$$\sigma(\mathcal{C}^n) = \{1,3\} \leq \{4\} \leq \{2,5\}.$$

A random composition $\mathcal{C}$ of $\mathbb{N}$ is called *exchangeable* if for any finite permutation σ,

$$\mathcal{C} \overset{(\mathrm{d})}{=} \sigma(\mathcal{C}).$$

The author of [86] provides a procedure to build an exchangeable composition of $\mathbb{N}$ from any interval-partition I called the *ordered paintbox*. Let $(V_i)_{i\geq 1}$ be an i.i.d. sequence of uniform $[0,1]$ variables. Let $\mathcal{C}$ be the composition of $\mathbb{N}$ whose blocks are given by the relation

$$i \sim j \iff V_i \text{ and } V_j \text{ belong to the same interval component of } I$$

and the order of the blocks is

$$A \leq A' \iff V_i \leq V_j, \quad \forall i \in A,\ \forall j \in A'.$$

The main result of [86] shows that any exchangeable composition of $\mathbb{N}$ can be obtained as an ordered paintbox based on a random interval-partition (see Theorem 11 in [86]). We now give a proof of this result that differs from the original proof of [86]. We make use of de Finetti's theorem in a similar way as Aldous' proof of Kingman's theorem, see e.g. the proof of Theorem 2.1 in [20]. The original proof of [86] relies on a reversed martingale argument combined with the method of moments.

Theorem 2.19 ([86])**.** *Let $\mathcal{C}$ be an exchangeable composition of $\mathbb{N}$. There exists on the same probability space a random interval-partition I and an independent i.i.d. sequence $(V_i)_{i\geq 1}$ of uniform $[0,1]$ variables such that the ordered paintbox based on I by the sequence $(V_i)_{i\geq 1}$ is a.s. $\mathcal{C}$.*

Before showing the theorem we need a technical lemma. Any composition $\mathcal{C} = (\pi, \leq)$ can be encoded as a total preorder $\preceq$ on $\mathbb{N}$ defined as

$$i \preceq j \iff B_i \leq B_j$$

where B_i (resp. B_j) is the block containing i (resp. j). The blocks of π can be recovered from $\preceq$ by the following relation

$$i \sim j \iff i \preceq j \text{ and } j \preceq i$$

and the order $\leq$ by

$$B \leq B' \iff i \preceq j, \quad \forall i \in B, \forall j \in B'.$$

Lemma 2.20. *Let $\mathcal{C}$ be an exchangeable composition of $\mathbb{N}$. We can find an exchangeable sequence of $[0,1]$-valued random variables $(\xi_i)_{i\geq 1}$ such that*

$$i \preceq j \iff \xi_i \leq \xi_j.$$

Proof. Let D_i be the set of integers lower than i

$$D_i = \{k : k \preceq i\}.$$

It is immediate that the partition $(D_i \setminus \{i\}, \mathbb{N} \setminus \{i\} \setminus D_i)$ is an exchangeable partition of $\mathbb{N} \setminus \{i\}$. Thus Kingman's representation theorem (see e.g. Theorem 2.1 in [20]) ensures that the limit

$$\xi_i = \lim_{n\to\infty} \frac{1}{n} \operatorname{Card}(D_i \cap [n])$$

exists a.s. Fix a finite permutation σ whose support lies in $[n]$, i.e. such that $\sigma(i) = i$ for $i \geq n$. For $m \geq n$, the distribution of $(\operatorname{Card}(D_i \cap [m]))_{i\geq 1}$ is invariant by the action of σ. Taking the limit, the distribution of the sequence $(\xi_i)_{i\geq 1}$ is also invariant by the action of σ, and thus it is an exchangeable sequence.

We need to show that

$$i \preceq j \iff \xi_i \leq \xi_j.$$

The only difficulty here is to show that $\xi_i \leq \xi_j$ implies $i \preceq j$. Suppose that $i \not\preceq j$, we need to show that

$$\xi_i - \xi_j = \lim_{n\to\infty} \frac{1}{n} \operatorname{Card}\Big((D_i \setminus D_j) \cap [n]\Big) > 0.$$

The partition $(D_j \setminus \{i,j\},\ D_i \setminus \{i,j\} \setminus D_j,\ \mathbb{N} \setminus \{i,j\} \setminus D_i)$ is an exchangeable partition of $\mathbb{N} \setminus \{i,j\}$. Another interesting consequence of Kingman's theorem is that in any exchangeable partition, the blocks are either singletons or have positive asymptotic frequencies. According to this, it is sufficient to show that a.s. $D_i \setminus D_j$ has at least two elements that are not i. Consider B_i (resp. B_j) the block to which i (resp. j) belongs. The set $D_i \setminus D_j$ is the reunion of all the blocks B such that $B_j < B \leq B_i$. Thus $D_i \setminus D_j$ is a singleton iff $B_i = \{i\}$ and there exists at most one singleton block B such that $B_j < B < B_i$. Let $n \geq 1$ and consider the block sizes and order of $\mathcal{C}^n$ as fixed. Exchangeability shows that the labels inside the blocks are chosen uniformly among all the possibilities. In particular this shows that the probability that $(D_i \setminus D_j) \cap [n]$ is a singleton goes to 0 as n goes to infinity. □

Now Theorem 2.19 is essentially a corollary of the previous lemma and of de Finetti's theorem.

Proof of Theorem 2.19. Let $(\xi_i)_{i\geq 1}$ be as above. Applying de Finetti's theorem we know that there exists a random measure μ such that conditionally on it the sequence $(\xi_i)_{i\geq 1}$ is i.i.d. distributed as μ. Consider the distribution function F_μ of μ, and its generalized inverse

$$F_\mu^{-1}(x) = \inf\{r : F_\mu(r) > x\}.$$

The interval-partition associated with μ, I_μ, is defined as the set of flats of F_μ^{-1}:

$$I_\mu = \Big\{x \in [0,1] : \exists y < x < z,\ F_\mu^{-1}(y) = F_\mu^{-1}(z)\Big\}.$$

The measure μ has the property that if X is distributed as μ, then μ-a.s. $F_\mu(X) = X$. Conditioning on μ, this can be seen from the definition of the sequence $(\xi_i)_{i\geq 1}$ and the law of large numbers:

$$F_\mu(\xi_1) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{\xi_j \leq \xi_1\}} = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{j \preceq 1\}} = \xi_1 \quad \mu\text{-a.s.}$$

In the terminology of [86] this shows that the measure μ is *uniformized.* A uniformized measure has an atomic and a diffuse part. The support of the diffuse part is $[0,1]\setminus I_\mu$ and coincides with the Lebesgue measure. The atomic part is supported by the right endpoints of the interval components of I_μ. If $J = (\ell, r)$ is an interval component of I_μ, the measure μ has an atom of mass $r - \ell$ located at r.

Let $(J_k)_{k\geq 1}$ be the interval decomposition of I_μ, and write $J_k = (\ell_k, r_k)$. Let $(X_i)_{i\geq 1}$ be an independent i.i.d. sequence of uniform variables, we define

$$V_i = \begin{cases} \xi_i & \text{if } \xi_i \notin I_\mu \\ (r_k - \ell_k)X_i + \ell_k & \text{if } \xi_i = r_k. \end{cases}$$

In words, the variables from the sequence $(\xi_i)_{i\geq 1}$ which are equal to the atom r_k are uniformly dispersed over the interval J_k. The previous remarks on the structure of uniformized measures show that conditionally on μ, the sequence $(V_i)_{i\geq 1}$ is i.i.d. uniform on $[0,1]$. The conditional distribution does not depend on μ, thus the sequence $(V_i)_{i\geq 1}$ is independent of μ and of I_μ.

We only need to show that the ordered paintbox based on I_μ using the sequence $(V_i)_{i\geq 1}$ is $\mathcal{C}$ a.s. This is plain from the design of the sequence. □

We end this section with a technical result already present in [86] (see Proposition 9) which we will require. Let $\mathcal{C}$ be an exchangeable composition of $\mathbb{N}$ and $\mathcal{C}^n$ its restriction to $[n]$. Let us denote by n_i the size of the i-th block of $\mathcal{C}^n$. The *empirical interval-partition* associated to $\mathcal{C}_n$ is given by

$$I^n = \Big(0, \frac{n_1}{n}\Big) \cup \Big(\frac{n_1}{n}, \frac{n_1+n_2}{n}\Big) \cup \cdots \cup \Big(\frac{n_1+\cdots+n_{k-1}}{n}, 1\Big).$$

Here is a more pictorial way of constructing I^n. Divide $[0,1]$ in intervals of size $1/n$ and label them from 1 to n in such a way that $i \preceq j$ iff the block with label i is before the block with label j. Then I^n is obtained by merging the intervals whose labels are in the same block of the composition. The next result states that the interval-partition representing $\mathcal{C}$ in Theorem 2.19 can be obtained as the limit of the empirical interval-partitions.

Proposition 2.21. *If $\mathcal{C}$ is an exchangeable composition of $\mathbb{N}$, I the interval-partition obtained from Theorem 2.19 and $(I^n)_{n\geq 1}$ the sequence of empirical interval-partitions associated to $\mathcal{C}$, we have*

$$\lim_{n\to\infty} d_H(I^n, I\setminus\{0,1\}) = 0 \quad a.s.$$

Proof. Let μ, $(\xi_i)_{i\geq 1}$ and I_μ be as in the proof of Theorem 2.19. De Finetti's theorem ensures that

$$\lim_{n\to\infty} \mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i} = \mu \quad \text{a.s.}$$

in the sense of weak convergence of probability measures. The interval-partition I_{μ_n} coincides with the empirical interval-partition I^n and as was already noticed in [86], the weak convergence of μ_n to μ implies the convergence of I_{μ_n} to I in the Hausdorff topology. □

Remark 2.22. This also shows that the representation obtained through Theorem 2.19 is unique in distribution. The interval-partition I is a.s. recovered from I^n whose distribution is fully determined by $\mathcal{C}$. ○

2.2.2 Exchangeable nested compositions

Gnedin's theorem sets up a correspondence between random interval-partitions and exchangeable compositions. We want to find a similar correspondence between nested interval-partitions and exchangeable nested compositions, the ordered version of exchangeable coalescents. A nested composition of $[n]$ (resp. $\mathbb{N}$) is a càdlàg process $(\mathcal{C}_t)_{t\geq 0}$ taking values in the compositions of $[n]$ (resp. $\mathbb{N}$) such that, as t increases, only adjacent blocks of the composition merge. More precisely, if $(\mathcal{C}_t)_{t\geq 0}$ is a nested composition, for any $s \leq t$, the blocks of $\mathcal{C}_t$ are obtained by merging blocks of $\mathcal{C}_s$, and if $A \leq B$ are two blocks of $\mathcal{C}_s$ that merge, they also merge with any block C such that $A \leq C \leq B$.

Naturally we say that $(\mathcal{C}_t)_{t\geq 0}$ is an exchangeable nested composition of $\mathbb{N}$ if for any finite permutation σ we have

$$(\mathcal{C}_t)_{t\geq 0} \overset{\text{(d)}}{=} (\sigma(\mathcal{C}_t))_{t\geq 0}.$$

We can extend the ordered paintbox construction to nested compositions. Let $(I_t)_{t\geq 0}$ be a nested interval-partition, and $(V_i)_{i\geq 1}$ an independent i.i.d. uniform sequence. Let $\mathcal{C}_t$ be the composition obtained from the ordered paintbox based on I_t by $(V_i)_{i\geq 1}$. Then it is immediate that $(\mathcal{C}_t)_{t\geq 0}$ is an exchangeable nested composition. Notice that this is only true because we have used the same sequence $(V_i)_{i\geq 1}$ for all times t.

Remark 2.23. Similarly to Remark 2.5, the nested composition obtained from an ordered paintbox is not càdlàg in general. Again it admits a unique càdlàg modification and we shall always consider this modification. ○

We have the following direct reformulation of Theorem 2.19 in the framework of nested compositions.

Theorem 2.24. *Let $(\mathcal{C}_t)_{t\geq 0}$ be an exchangeable nested composition of $\mathbb{N}$. We can find on the same probability space a nested interval-partition $(I_t)_{t\geq 0}$ and an independent i.i.d. sequence $(V_i)_{i\geq 1}$ of uniform variables such that a.s. the ordered paintbox* *based on* $(I_t)_{t\geq 0}$ *with* $(V_i)_{i\geq 1}$ *is* $(\mathcal{C}_t)_{t\geq 0}$*. This nested interval-partition is unique in distribution.*

Proof. Existence. For any $t \geq 0$, $\mathcal{C}_t$ is an exchangeable composition of $\mathbb{N}$. We can apply Theorem 2.19 distinctly for $t \in \mathbb{Q}_+$ to find on the same probability space a collection of interval-partitions $(I_t)_{t\in\mathbb{Q}_+}$ such that for any $t \in \mathbb{Q}_+$ the ordered paintbox based on I_t is $\mathcal{C}_t$. Let I_t^n be the empirical interval-partition associated to $\mathcal{C}_t \cap [n]$. The fact that $(\mathcal{C}_t)_{t\geq 0}$ is a nested composition ensures that $(I_t^n)_{t\in\mathbb{Q}_+}$ is a nested interval-partition. Taking the limit as n goes to infinity shows that $(I_t)_{t\in\mathbb{Q}_+}$ is also a nested interval-partition. It admits a unique càdlàg extension given by

$$I_s = \operatorname{int}\Big(\bigcap_{\substack{t\geq s\\ t\in\mathbb{Q}_+}} I_t\Big).$$

Let $(V_i)_{i\geq 1}$ be the i.i.d. uniform sequence given by Theorem 2.19 applied at time $t = 0$. To see that $(V_i)_{i\geq 1}$ is independent of $(I_t)_{t\geq 0}$, one can do the exact same steps as in the proof of Theorem 2.19 but using a vectorial version of de Finetti's theorem (see Section 2.B).

We now show that for any $t \in \mathbb{Q}_+$, a.s.

$$i \sim_t j \iff V_i \text{ and } V_j \text{ are in the same interval of } I_t \tag{2.3}$$

where $\sim_t$ is the relation given by the blocks of $\mathcal{C}_t$.

Let $n \geq 1$ and divide the interval $[0, 1]$ in n intervals of size $1/n$. We label the intervals from 1 to n in the same order as the variables $V_1, \dots, V_n$. Let $t \in \mathbb{Q}_+$, the first step is to notice that the empirical interval-partition I_t^n can be recovered by merging the blocks of size $1/n$ whose labels belong to the same block of $\mathcal{C}_t$. Now, let $V_i^{(n)}$ (resp. $V_j^{(n)}$) be the right-hand extremity of the interval with label i (resp. j). Using twice the law of large numbers shows that $V_i^{(n)}$ and $V_j^{(n)}$ converge to V_i and V_j respectively. Moreover, we know that I_t^n converges a.s. to I_t. If we suppose that $V_i < V_j$ and $i \sim_t j$, then for any $n \geq 1$, $(V_i^{(n)}, V_j^{(n)}) \subset I_t^n$, and taking the limit shows that $(V_i, V_j) \subset I_t$. Conversely if $(V_i, V_j) \subset I_t$, using the convergence, for n large enough we have $(V_i^{(n)}, V_j^{(n)}) \subset I_t^n$ and thus i and j are in the same block of $\mathcal{C}_t$.

That relation (2.3) holds a.s. for any $t \geq 0$ will follow by right-continuity. However we have to be careful, in general the nested composition obtained from an ordered paintbox is not càdlàg. By continuity, the relation (2.3) only holds a.s. for all times t when $(\mathcal{C}_t)_{t\geq 0}$ is continuous. The original nested composition $(\mathcal{C}_t)_{t\geq 0}$ is recovered by considering a càdlàg modification of the nested composition obtained though an ordered paintbox based on $(I_t)_{t\geq 0}$.

Uniqueness. The uniqueness will come from the following convergence result

$$\lim_{n\to\infty} \sup_{t\geq 0} d_H(I_t^n, I_t) = 0 \quad \text{a.s.}$$

We start by showing the convergence. Let $\varepsilon > 0$, we can split $[0, 1]$ into a finite number of pairwise disjoint intervals of length smaller than ε denoted by $J_1, \dots, J_p$.

Given a combination of such intervals, $J = J_{i_1} \cup \cdots \cup J_{i_k}$, let f_J^n denote the fraction of variables $V_1, \ldots, V_n$ which belong to J. Then for any $\eta > 0$ using the law of large numbers we can a.s. find a large enough N_J such that

$$\forall n \geq N_J, \quad |\mathrm{Leb}(J) - f_J^n| < \eta.$$

Let N be large enough such that this condition is fulfilled for all possible combinations of intervals.

We now show that a.s.

$$\forall t \geq 0, \forall n \geq N, \quad d_H(I_t^n, I_t) \leq \eta + \varepsilon.$$

Let $x \notin I_t$, and $J_x = (\ell_x, r_x)$ be the interval such that $x \in J$ (in case x is the boundary of two intervals, we choose the left interval). First suppose that $\ell_x = 0$ or $r_x = 1$. By construction $0, 1 \notin I_t^n$, thus $d(x, 0) < \varepsilon$ or $d(x, 1) < \varepsilon$. In the other case, the variables $(V_i)_{i \geq 1}$ which are in $[0, \ell_x]$ and those in $[r_x, 1]$ are not in the same interval component of I_t, and by construction of the paintbox, their labels are not in the same block of $\mathcal{C}_t$. For $n \geq 1$, let f_1^n (resp. f_2^n) denote the frequency of the variables $(V_i)_{i \leq n}$ belonging to $[0, \ell_x]$ (resp. $[0, r_x]$). The previous remark shows that there is a point $y \in [f_1^n, f_2^n]$ which does not belong to I_t^n. For $n \geq N$ we know that $y \in [\ell_x - \eta, r_x + \eta]$ and thus $d(x, y) \leq \eta + \varepsilon$. This shows

$$\forall t \geq 0, \forall n \geq N, \quad \sup_{x \notin I_t} d(x, [0,1] \setminus I_t^n) \leq \eta + \varepsilon.$$

Similarly consider $x_n \notin I_t^n$. If $x_n \in \{0, 1\}$, clearly $d(x_n, [0,1] \setminus I_t) = 0$. In the other case the point x_n is the separation between two intervals of I_t^n. These two intervals can be seen as an agglomeration of blocks of size $1/n$ whose labels belong to the same block of I_t. Let i (resp. j) be the label of the right-most (resp. left-most) block of size $1/n$ of the left interval (resp. right interval) separated by x_n. The rules of the paintbox construction imply that V_i and V_j are not in the same interval of I_t, thus there exists $V_i \leq y_n \leq V_j$ such that $y_n \notin I_t$. The value of x_n is exactly the frequency of variables $V_1, \ldots V_n$ which belong to $[0, y_n]$. Let $J_{y_n} = (\ell_{y_n}, r_{y_n})$ be the interval to which y_n belongs, and f_1^n, f_2^n be as above the frequency of the n first variables in $[0, \ell_{y_n}]$ and $[0, r_{y_n}]$. As $\ell_{y_n} \leq y_n$, we know that $f_1^n \leq x_n$, and similarly $x_n \leq f_2^n$. Thus for $n \geq N$, $x_n \in [\ell_{y_n} - \eta, r_{y_n} + \eta]$ and $d(x_n, y_n) \leq \eta + \varepsilon$. This shows

$$\forall t \geq 0, \forall n \geq N, \quad \sup_{x \notin I_t^n} d(x, [0,1] \setminus I_t) \leq \eta + \varepsilon.$$

Thus, a.s. $(I_t^n)_{t \geq 0}$ converges uniformly to $(I_t)_{t \geq 0}$.

To get uniqueness, it is sufficient to notice that the distribution of the sequence $((I_t^n)_{t \geq 0};\, n \geq 1)$ is determined uniquely by that of $(\mathcal{C}_t)_{t \geq 0}$. As we can recover a.s. $(I_t)_{t \geq 0}$ from $((I_t^n)_{t \geq 0};\, n \geq 1)$, the distribution of $(I_t)_{t \geq 0}$ is also determined by that of $(\mathcal{C}_t)_{t \geq 0}$. □

Remark 2.25. This also proves Proposition 2.21 in a more detailed way. ○

2.2.3 Uniform nested compositions, proof of Theorem 2.8

We recall that $\mathfrak{I}$ stands for the quotient space of combs for the paintbox-equivalence relation. To be entirely rigorous we need to define a suitable σ-field on $\mathfrak{I}$. By definition of $\mathfrak{I}$ a paintbox based on any of the representatives of a class yields the same distribution on the space of coalescents. We can identify each class with this distribution and endow $\mathfrak{I}$ with the weak convergence topology of probability measures on the space of coalescents. We consider the associated Borel σ-field. This approach bears similarity with the Gromov-weak topology introduced in [91], more on this can be found in Section 2.4.

The first step to find a comb representation of a given exchangeable coalescent $(\Pi_t)_{t\geq 0}$ is to order the blocks of $(\Pi_t)_{t\geq 0}$ to obtain a nested composition. We will do that using the notion of uniform nested composition that we now introduce.

Definition 2.26. Let $(\mathcal{C}_t)_{t\geq 0}$ be an exchangeable nested composition of $\mathbb{N}$ and $(\Pi_t)_{t\geq 0}$ be the associated coalescent. We say that $(\mathcal{C}_t)_{t\geq 0}$ is uniform if for any $n \geq 1$, conditionally on $(\Pi^n_t)_{t\geq 0}$, the order of the blocks of $(\mathcal{C}^n_t)_{t\geq 0}$ is uniform among all the possible orderings, i.e. all the orderings such that $(\mathcal{C}^n_t)_{t\geq 0}$ is a nested composition. ○

The following lemma shows that any exchangeable coalescent can be turned into a uniform exchangeable nested composition.

Lemma 2.27. *Let $(\Pi_t)_{t\geq 0}$ be an exchangeable coalescent. There exists a uniform exchangeable nested composition $(\mathcal{C}_t)_{t\geq 0}$ whose associated coalescent is $(\Pi_t)_{t\geq 0}$.*

Proof. We proceed by induction. For $n = 1$ there is a unique trivial possible order on the blocks. Suppose that we have built for n an order on the blocks of $(\Pi^n_t)_{t\geq 0}$ such that only adjacent blocks can merge, we call such an order an order *consistent with the genealogy*. Then there are finitely many orders on the blocks of $(\Pi^{n+1}_t)_{t\geq 0}$ that extend the previous order and are consistent with the genealogy. More precisely, if $n+1$ is in a block of Π^{n+1}_0 the extension is unique. If $n+1$ is a singleton of Π^{n+1}_0, suppose that $\{n+1\}$ coalesce at some point and that k blocks are involved in this coalescence event. Then there are k consistent extensions: $\{n+1\}$ can be placed between any of the $k-1$ other blocks, or at the left-most (resp. right-most) position. If $\{n+1\}$ does not coalesce, the singleton can be placed at any position between blocks that do not coalesce. We pick one of these orders independently and uniformly.

By induction, we have built on the same probability space as $(\Pi_t)_{t\geq 0}$ a nested composition of $\mathbb{N}$ whose blocks merge according to $(\Pi_t)_{t\geq 0}$. It is easily checked from the construction that $(\mathcal{C}_t)_{t\geq 0}$ is a uniform nested composition. It remains to show that it is exchangeable. Fix $0 \leq t_1 < \cdots < t_p$, and let $c_1, \dots, c_p$ be compositions of $[n]$, whose block partitions are $\pi_1, \dots, \pi_n$ respectively. Fix some trajectory $\Pi^n := (\Pi^n_t)_{t\geq 0}$ of the coalescent. Let us denote by $O(\Pi^n)$ the number of orderings of the blocks of Π^n_0 yielding a nested composition, and let $O(c_1, \dots, c_p; \Pi^n)$ be the number of such orderings verifying that $\mathcal{C}^n_{t_i} = c_i$, for $i \in \{1, \dots, p\}$. Then for any permutation σ of $[n]$, the following direct calculation

$$\begin{aligned}\mathbb{P}\Big(\mathcal{C}^n_{t_1} = c_1, \dots, \mathcal{C}^n_{t_p} = c_p\Big) &= \mathbb{E}\Big[\frac{O(c_1, \dots, c_p;\, \Pi^n)}{O(\Pi^n)} \mathbb{1}_{\{\Pi^n_{t_1} = \pi_1, \dots, \Pi^n_{t_p} = \pi_p\}}\Big] \\ &= \mathbb{E}\Big[\frac{O(c_1, \dots, c_p;\, \sigma(\Pi^n))}{O(\sigma(\Pi^n))} \mathbb{1}_{\{\sigma(\Pi^n_{t_1}) = \pi_1, \dots, \sigma(\Pi^n_{t_p}) = \pi_p\}}\Big] \\ &= \mathbb{E}\Big[\frac{O(\sigma^{-1}(c_1), \dots, \sigma^{-1}(c_p);\, \Pi^n)}{O(\Pi^n)} \mathbb{1}_{\{\Pi^n_{t_1} = \sigma^{-1}(\pi_1), \dots, \Pi^n_{t_p} = \sigma^{-1}(\pi_p)\}}\Big] \\ &= \mathbb{P}\Big(\mathcal{C}^n_{t_1} = \sigma^{-1}(c_1), \dots, \mathcal{C}^n_{t_p} = \sigma^{-1}(c_p)\Big)\end{aligned}$$

proves that the nested composition is exchangeable. □

Proof of Theorem 2.8. Let $(\Pi_t)_{t \geq 0}$ be an exchangeable coalescent. Let $(\mathcal{C}_t)_{t \geq 0}$ be the uniform nested compositions obtained through Lemma 2.27. Invoking Theorem 2.24 shows that there exists a comb representation $(I_t)_{t \geq 0}$ of $(\Pi_t)_{t \geq 0}$. The uniqueness is immediate from the definition of the quotient. □

2.3 Comb representation of Λ-coalescents

In this section, we restrict our attention to the well-studied case of Λ-coalescents. A process $(\Pi_t)_{t \geq 0}$ is a Λ-coalescent if for any $n \geq 1$, its restriction $(\Pi^n_t)_{t \geq 0}$ to $[n]$ is a Markov process such that starting from a partition with b blocks, any k blocks coalesce at rate

$$\lambda_{b,k} = \int_{[0,1]} x^{k-2}(1-x)^{b-k} \Lambda(\mathrm{d}x)$$

for a finite measure Λ on $[0,1]$.

The broad aim of this section is to find a Markovian comb representation of a given Λ-coalescent, and to provide its transitions. Recall from the last section the path followed to obtain a comb associated to an exchangeable coalescent. The first step is to order the blocks of the coalescent to get a nested composition, and then to use Theorem 2.24 to define a comb. Here we will follow this path in the special case of Λ-coalescents where we can have an explicit description of both the nested composition and the comb.

Let us first define the nested composition associated to a Λ-coalescent. Consider the modified transition rates

$$\tilde{\lambda}_{b,k} = \frac{1}{b-k+1} \binom{b}{k} \lambda_{b,k}.$$

Let $n \geq 1$, we define a Markov chain $(\mathcal{C}^n_t)_{t \geq 0}$ taking values in the space of composition of $[n]$ as follows. Starting from c, a composition of $[n]$ with b blocks, any k adjacent blocks merge at rate $\tilde{\lambda}_{b,k}$. These transition rates have a natural combinatorial interpretation. Consider $(\Pi^n_t)_{t \geq 0}$ the restriction to $[n]$ of a Λ-coalescent. Starting from a partition with b blocks, there are $\binom{b}{k}$ ways of merging k distinct blocks. Thus the total transition rate from b to $b-k+1$ blocks is $\binom{b}{k}\lambda_{b,k}$. Given that k blocks merge, the blocks that merge are chosen uniformly among the $\binom{b}{k}$ possible choices. Starting from a composition with b blocks, there are only $b-k+1$ ways to merge k adjacent blocks. Thus, the total transition rate of $(\mathcal{C}^n_t)_{t\geq 0}$ from b to $b-k+1$ blocks is the same as $(\Pi^n_t)_{t\geq 0}$, but instead of choosing uniformly k blocks among the $\binom{b}{k}$ possibilities, we choose k *adjacent* blocks among the $b-k+1$ possibilities.

We now extend this sequence of nested compositions to a nested composition of $\mathbb{N}$. To fully determine the distribution of $(\mathcal{C}^n_t)_{t\geq 0}$ we have to specify an initial distribution. We will always assume in this section that the process $(\mathcal{C}^n_t)_{t\geq 0}$ starts from the composition of $[n]$ composed of only singletons ordered uniformly. Using the Markov projection theorem (see e.g. [123], Section 6.3), it is not hard to see that the sequence of processes $((\mathcal{C}^n_t)_{t\geq 0};\, n \geq 1)$ is sampling consistent, i.e. that the restriction of $(\mathcal{C}^{n+1}_t)_{t\geq 0}$ to $[n]$ is distributed as $(\mathcal{C}^n_t)_{t\geq 0}$. Using the Kolmogorov extension theorem we can find $(\mathcal{C}_t)_{t\geq 0}$ an exchangeable nested composition of $\mathbb{N}$ whose projections to $[n]$ is distributed as $(\mathcal{C}^n_t)_{t\geq 0}$ for all $n \geq 1$. The process $(\mathcal{C}_t)_{t\geq 0}$ is a nested composition whose blocks merge according to a Λ-coalescent.

Lemma 2.28. *Let $(\Pi_t)_{t\geq 0}$ be the coalescent associated to $(\mathcal{C}_t)_{t\geq 0}$. Then $(\Pi_t)_{t\geq 0}$ is a Λ-coalescent. Moreover for any $t \geq 0$, conditionally on Π^n_t, the composition $\mathcal{C}^n_t$ is obtained by ordering uniformly the blocks of Π^n_t.*

Proof. Let $(\mathcal{C}^n_t)_{t\geq 0}$ and $(\Pi^n_t)_{t\geq 0}$ be the restriction to $[n]$ of $(\mathcal{C}_t)_{t\geq 0}$ and $(\Pi_t)_{t\geq 0}$ respectively. Let $\hat{Q}_n$ be the generator of $(\mathcal{C}^n_t)_{t\geq 0}$ and Q_n be the generator of a Λ-coalescent on $[n]$. The result will follow by using a Markov projection theorem from [183], see their Theorem 2. To apply this result, we need to find a probability kernel L_n from the space of partitions of $[n]$ to the space of compositions of $[n]$ such that for any function f from the space of compositions of $[n]$ to $\mathbb{R}$,

$$\forall \pi, \quad \hat{Q}_n L_n f(\pi) = L_n Q_n f(\pi)$$

and such that the initial distribution of $(\mathcal{C}^n_t)_{t\geq 0}$ is the push-forward by L_n of the initial distribution of $(\Pi^n_t)_{t\geq 0}$.

Let f be such a function. For π a partition of $[n]$, let $\mathcal{C}_\pi$ be the random composition of $[n]$ obtained by ordering the blocks of π uniformly. We set

$$\forall \pi, \quad L_n f(\pi) = \mathbb{E}[f(\mathcal{C}_\pi)].$$

Our choice of initial distribution for $(\mathcal{C}_t)_{t\geq 0}$ ensures that the second condition holds. A straightforward generator calculation shows that the above equality is fulfilled and that the desired result holds. See Section 2.C for the details of the calculation. □

Using Theorem 2.24, the nested composition $(\mathcal{C}_t)_{t\geq 0}$ defines a unique nested interval-partition $(I_t)_{t\geq 0}$ that we call the Λ-*comb*. In the remainder of the section we want to show that the Λ-comb is a Markov process and give its transitions. We will express the transitions in terms of composition of bridges that we now introduce.

We say that a function $B\colon [0,1] \to [0,1]$ is a bridge if it is of the form

$$B(x) = x(1 - \sum_{i\geq 1} \beta_i) + \sum_{i\geq 1} \beta_i \mathbb{1}_{\{x \leq V_i\}}$$

for a random mass-partition β and an independent i.i.d. sequence $(V_i)_{i\geq 1}$ of uniform $[0,1]$-valued variables. To any bridge we associate an interval-partition defined as

$$I(B) = \mathrm{int}\Big([0,1] \setminus B([0,1])\Big)$$

where $B([0,1])$ is the range of B. We can ask if the converse holds. The correct notion to answer this question is that of uniform order.

Definition 2.29. Let I be a random interval-partition and $\mathcal{C}$ be the composition of $\mathbb{N}$ obtained through an ordered paintbox based on I. We say that I has a uniform order if for any $n \geq 1$, the order of the blocks of $\mathcal{C} \cap [n]$ is uniform. ○

The following lemma shows that having a uniform order is a necessary and sufficient condition for an interval-partition to be represented by a bridge. See Section 2.3.1 for a proof.

Lemma 2.30. *Let I be a random interval-partition. There exists a bridge B such that $I(B) = I$ iff I has a uniform order. If I has a uniform order, the bridge B such that $I(B) = I$ is unique in distribution.*

Notice that for any $t \geq 0$, the Λ-comb I_t at time t has a uniform order. We will denote by B^{I_t} the bridge associated to I_t through Lemma 2.30. We are now in position to provide the transitions of the Λ-comb.

Proposition 2.31. *Let $(I_t)_{t\geq 0}$ be the Λ-comb. The process $(I_t)_{t\geq 0}$ is Markovian, and for any $s, t \geq 0$, conditionally on I_t,*

$$I_{t+s} \overset{(\mathrm{d})}{=} I(B^{I_t} \circ B'_s) \tag{2.4}$$

where B'_s is an independent bridge distributed as B^{I_s}.

Remark 2.32. In the coming down from infinity case we have a simpler description of the semi-group of the Λ-comb. Suppose that $(I_t)_{t\geq 0}$ starts from an interval-partition I_0 with b blocks and no dust. Then any k adjacent blocks of I_0 merge at rate $\tilde{\lambda}_{b,k}$. ○

The above proposition shows that the Λ-comb can be represented in terms of composition of independent bridges. As a direct corollary, we provide an alternative construction of the Λ-comb based on the flow of bridges of [21]. A flow of bridges is a collection $(B_{s,t})_{s\leq t}$ of bridges which fulfills the following three conditions:

(i) For any $s < r < t$, $B_{s,t} = B_{s,r} \circ B_{r,t}$ (cocycle property).

(ii) For any $t_1 < \cdots < t_p$, the bridges $(B_{t_1,t_2}, \ldots, B_{t_{p-1},t_p})$ are independent, and B_{t_1,t_2} is distributed as B_{0,t_2-t_1} (stationarity and independence of the increments).

(iii) The bridge $B_{0,t}$ converges to the identity map Id as $t \downarrow 0$ in probability in Skorohod topology.

It can be seen from the cocycle property that the interval-partition-valued process $(I(B_{0,t}))_{t\geq 0}$ is a nested interval-partition. A sampling procedure has been defined in [21] to obtain a coalescent from a flow of bridges. In our context, sampling from the flow of bridges according to this procedure is the same as doing a paintbox based on $(I(B_{0,t}))_{t\geq 0}$. An important result from [21] states that given a Λ-coalescent $(\Pi_t)_{t\geq 0}$, there exists a unique flow of bridges whose associated coalescent is distributed as $(\Pi_t)_{t\geq 0}$ (see Theorem 1 in [21]). We call it the Λ-flow of bridges. As a corollary of this correspondence and of Proposition 2.31, we are able to show that the comb associated to the Λ-flow of bridges is the Λ-comb introduced above from the transition rates.

Corollary 2.33. *Let Λ be a finite measure on $[0,1]$, and let $(I_t)_{t\geq 0}$ be the Λ-comb and $(B_{s,t})_{s\leq t}$ be the Λ-flow of bridges. Then*

$$(I_t)_{t\geq 0} \overset{\text{(d)}}{=} (I(B_{0,t}))_{t\geq 0}.$$

Proof. Let $p \geq 1$ and $0 \leq t_1 < \cdots < t_p$. Using the Markov property of $(I_t)_{t\geq 0}$ and the expression of the transitions (2.4) we know that

$$\Big(I_{t_1}, \ldots, I_{t_p}\Big) \overset{\text{(d)}}{=} \Big(I_{t_1}, I(B^{I_{t_1}} \circ B'_1), \ldots, I(B^{I_{t_1}} \circ B'_1 \circ \cdots \circ B'_{p-1})\Big),$$

where $(B'_1, \ldots, B'_{p-1})$ are independent bridges and for $1 \leq k \leq p-1$, B'_k is distributed as $B^{I_{t_{k+1}-t_k}}$.

Let $(B_{s,t})_{s\leq t}$ be the Λ-flow of bridges. Then from the cocycle property

$$\Big(I(B_{0,t_1}), \ldots, I(B_{0,t_p})\Big) = \Big(I(B_{0,t_1}), \ldots, I(B_{0,t_1} \circ B_{t_1,t_2} \circ \cdots \circ B_{t_{p-1},t_p})\Big).$$

Moreover as the flow of bridges has independent and stationary increments, the bridge $(B_{t_1,t_2}, \ldots, B_{t_{p-1},t_p})$ are and have the same distribution as above. □

2.3.1 Proof of Lemma 2.30

We will need the following continuity result.

Lemma 2.34. *The map $I\colon B \mapsto I(B)$ that maps a bridge to its associated interval-partition is continuous when the space of interval-partitions is endowed with the Hausdorff topology and the space of bridges with the Skorohod topology.*

Proof. Let B^n be a sequence of bridges that converge to B in the Skorohod topology. We know that we can find a sequence of continuous bijections λ_n from $[0,1]$ to $[0,1]$ such that

$$\lim_{n\to\infty} \|\lambda_n - \mathrm{Id}\|_\infty = 0$$

and

$$\lim_{n\to\infty} \|B - B^n \circ \lambda_n\|_\infty = 0.$$

Let $I = I(B)$ and $I_n = I(B^n)$. As the interval-partitions are obtained from bridges, we can re-write the Hausdorff distance as

$$d_H(I, I_n) = \sup_{x\in[0,1]} \inf_{y\in[0,1]} |B^n(x) - B(y)| \vee \sup_{x\in[0,1]} \inf_{y\in[0,1]} |B^n(y) - B(x)|.$$

We have

$$\sup_{x\in[0,1]} \inf_{y\in[0,1]} |B(x) - B^n(y)| \le \sup_{x\in[0,1]} |B(x) - B^n(\lambda_n(x))|$$

and

$$\sup_{x\in[0,1]} \inf_{y\in[0,1]} |B(y) - B^n(x)| \le \sup_{x\in[0,1]} \left| B(\lambda_n^{-1}(x)) - B^n(x) \right|$$

and thus

$$\lim_{n\to\infty} d_H(I, I^n) = 0,$$

which ends the proof. □

Proof of Lemma 2.30. First suppose that I is of the form $I(B)$ for some bridge B. Consider B^{-1} the generalized inverse of B. Let $(V_i)_{i\ge1}$ be i.i.d. uniform variables and $\mathcal{C}$ be the composition obtained through an ordered paintbox using these variables. By construction of the ordered paintbox and as B^{-1} is non-decreasing, the order of the blocks of $\mathcal{C}$ is given by the order of the variables $(B^{-1}(V_i))_{i\ge1}$. Conditionally on the bridge these variables are i.i.d. and thus their order is uniform.

Now let I be an interval-partition with a uniform order and $\mathcal{C}$ be the composition obtained by an ordered paintbox. We will first consider the case where I has finitely many interval components and no dust. The fact that the order of the blocks of the composition $\mathcal{C}$ is uniform shows that the order of the interval components of I is uniform (each block of $\mathcal{C}$ corresponds to an interval of I). Let K be the number of blocks of I, and let $V_1^* < \dots < V_K^*$ be the order statistics of independent uniform variables. Suppose that β_1 is the length of the left-most interval of I, β_2 that of the second left-most, etc. then

$$\forall u \in [0,1], \quad B(u) = \sum_{i=1}^K \beta_i \mathbb{1}_{\left\{V_i^* \le u\right\}}$$

is a bridge such that $I(B) = I$. Indeed, since the order of the intervals is uniform, there is a uniform permutation σ of $[K]$ independent of $V_1^*, \dots, V_K^*$, such that $(\beta_{\sigma(i)})$ is ranked in nonincreasing order. This shows that

$$B(u) = \sum_{i=1}^K \beta_{\sigma(i)} \mathbb{1}_{\left\{V_{\sigma(i)}^* \le u\right\}}$$

indeed defines a bridge. This also shows the uniqueness in distribution of B.

Let us turn to the general case. Let $n \geq 1$ and consider I^n the empirical interval-partition associated to $\mathcal{C} \cap [n]$. By assumption the interval-partition I^n has a uniform order, thus using the above argument we can find a unique bridge B^n such that $I(B^n) = I^n$. We know that I^n converges a.s. to I. Let β_n (resp. β) be the mass-partition associated to I^n (resp. I). As the function that maps an interval-partition to its mass-partition is continuous, we have that β_n converges a.s. to β (see e.g. [20] Proposition 2.2). We can now make use of another continuity result, namely Lemma 1 from [21], to show that the sequence of bridges $(B^n)_{n\geq 1}$ converges in distribution to a bridge B obtained from the mass-partition β. Using Lemma 2.34, we know that $I(B^n)$ converges in distribution to $I(B)$. By uniqueness of the limit, we get that

$$I \overset{\text{(d)}}{=} I(B),$$

and that B is unique. □

2.3.2 Proof of Proposition 2.31

We will first prove Proposition 2.31 for empirical interval-partitions and then take the limit. We start by proving the following lemma, which is the direct reformulation of Proposition 2.31 for empirical interval-partitions.

Lemma 2.35. *Let $\mathcal{C}_0^n$ be an exchangeable composition of $[n]$ with a uniform order on its blocks, and let $(\mathcal{C}_t^n)_{t\geq 0}$ be the Markov process started from $\mathcal{C}_0^n$ with transitions $(\tilde{\lambda}_{b,k};\ 2 \leq k \leq b < \infty)$. If $(I_t^n)_{t\geq 0}$ denotes the empirical nested interval-partition associated to $(\mathcal{C}_t^n)_{t\geq 0}$, then conditionally on $\mathcal{C}_0^n$,*

$$I_t^n \overset{\text{(d)}}{=} I(B_0^n \circ B_t),$$

where B_0^n and B_t are independent bridges such that $I(B_0^n) = I_0^n$ and $I(B_t) = I_t$, the Λ-comb at time t.

Proof. Let us denote by $(A_1, \dots, A_K)$ the blocks of $\mathcal{C}_0^n$ in order of their least element. As $\mathcal{C}_0^n$ has a uniform order on its blocks, according to Lemma 2.30 we can find $(U_1, \dots, U_K)$ such that conditionally on K these are i.i.d. uniform variables on $[0, 1]$ and

$$\forall r \in [0,1], \quad B_0^n(r) = \frac{1}{n}\sum_{i=1}^{K} \mathrm{Card}(A_i)\mathbb{1}_{\{U_i \leq r\}}$$

defines a bridges satisfying $I(B_0^n) = I_0^n$. Let B_t be independent and such that $I(B_t) = I_t$. To each interval component of I_0^n corresponds a unique block A_i of $\mathcal{C}_0^n$, and thus a unique jump time U_i of B_0^n. We claim that $I(B_0^n \circ B_t)$ is obtained by merging the intervals of I_0^n whose jump times belong to the same interval component of I_t. To see this, notice that by definition $I(B_0^n \circ B_t)$ is the set of flats of $(B_0^n \circ B_t)^{-1} = B_t^{-1} \circ (B_0^n)^{-1}$. Thus x and y belong to the same flat of $(B_0^n \circ B_t)^{-1}$ iff $(B_0^n)^{-1}(x)$ and $(B_0^n)^{-1}(y)$ belong to the same flat of B_t^{-1}, that is to the same interval component of I_t. The claim is proved by further noting that $(B_0^n)^{-1}(x)$ is the jump time of the interval component of I_0^n to which x belongs.

The previous procedure can be rephrased in terms of an ordered paintbox. The interval-partition $I(B_0^n \circ B_t)$ is obtained by labeling uniformly the K blocks of I_0^n, sampling a composition $\mathcal{C}'_t$ of $[K]$ according to an ordered paintbox based on I_t and merging the intervals of I_0^n whose labels belong to the same block of $\mathcal{C}'_t$. As I_t is the Λ-comb at time t, the composition $\mathcal{C}'_t$ is distributed as $\mathcal{C}_t^K$, the nested composition at time t obtained by merging K initial singleton blocks ordered uniformly according to the rates $(\tilde{\lambda}_{b,k};\, 2 \leq k \leq b < \infty)$. Thus $I(B_0^n \circ B_t)$ can be obtained by letting its intervals merge at rate $(\tilde{\lambda}_{b,k};\, 2 \leq k \leq b < \infty)$, and is distributed as I_t^n. □

Proof of Proposition 2.31. Let $(I_t)_{t\geq 0}$ be the Λ-comb, and $(V_i)_{i\geq 1}$ be an independent sequence of i.i.d. uniform variables on $[0,1]$. Denote by $(\mathcal{C}_t^n)_{t\geq 0}$ the nested composition of $[n]$ obtained by an ordered paintbox based on $(I_t)_{t\geq 0}$ using the sampling variables $(V_i)_{i\geq 1}$, and let $(I_t^n)_{t\geq 0}$ be the corresponding empirical nested interval-partition. According to Lemma 2.28 the interval-partition I_t has a uniform order, and thus there exists a bridge B_t such that $I(B_t) = I_t$. Conditionally on B_t, the sequence

$$\forall i \geq 1, \quad \xi_i = B_t^{-1}(V_i)$$

is i.i.d. We denote by μ_t the (random) law of ξ_1 conditionally on B_t, and by μ_t^n its empirical distribution defined as

$$\mu_t^n = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}.$$

Note that B_t is the distribution function of μ_t. If B_t^n denotes the distribution function of μ_t^n, then B_t^n is a bridge such that $I(B_t^n) = I_t^n$. It follows from Lemma 2.35 that

$$(I_t^n, I_{t+s}^n) \overset{\text{(d)}}{=} (I_t^n, I(B_t^n \circ B'_s)) \tag{2.5}$$

where B'_s is an independent bridge distributed as B^{I_s}. The result will follow by taking the limit in (2.5).

According to the Glivenko-Cantelli theorem (see for instance Proposition 4.24 in [120]), the sequence of bridges $(B_t^n)_{n\geq 1}$ converges almost surely to B_t in the uniform topology. Thus $B_t^n \circ B'_s$ converges a.s. in the uniform topology to $B_t \circ B'_s$, and by Lemma 2.34 and Proposition 2.21 the right-hand side of (2.5) converges a.s. to $(I_t, I(B_t \circ B'_s))$. Moreover, according to Proposition 2.21, the left-hand side converges a.s. to (I_t, I_{t+s}) and we have proved that (2.4) holds.

It remains to show that $(I_t)_{t\geq 0}$ is Markovian. As $(\mathcal{C}_t)_{t\geq 0}$ is obtained from $(I_t)_{t\geq 0}$ through the ordered paintbox procedure, it is sufficient to prove that $(\mathcal{C}_t)_{t\geq 0}$ is Markovian. This follows from standard arguments from measure theory by noting that the filtration of $(\mathcal{C}_t)_{t\geq 0}$ is induced by that of its restrictions to $[n]$, and that all of these restrictions are Markov. □

2.3.3 Dynamical combs

As mentioned in the introduction, an exchangeable coalescent models the genealogy of a population observed at a given time. By varying the observation time we obtain a dynamical genealogy that has been named the *evolving coalescent*. There has been much interest into studying evolving coalescents. For example, if the coalescent at a fixed time is the Kingman coalescent, the authors of [176, 177] have studied statistics of the evolving coalescent using a look-down representation, the authors of [92] studied the dynamics of the entire tree structure using the framework of the Gromov-weak topology. Evolving coalescents such that the coalescent at a fixed time is a more general Λ-coalescent have also been considered, see e.g. [126] for the case of Beta-coalescents and [197] for the Bolthausen-Sznitman coalescent.

In this section we show that the previous results on the Markov property of the Λ-comb allow us to define a comb-valued process, the *evolving comb*, such that sampling from the evolving comb at a fixed time yields a Λ-coalescent. The evolving comb contains all the information about the dynamical genealogy but does not require the cumbersome framework of random metric spaces endowed with the Gromov-Hausdorff topology as in [92]. For the sake of clarity we will only consider the evolving Kingman comb where we have an explicit construction of the genealogy at a fixed time.

We will build the evolving Kingman comb by defining its semi-group. Recall that when the coalescent associated to a nested interval-partition comes down from infinity, the comb can be represented using a comb function, see Section 2.1.2. Let f be a deterministic comb function and $s > 0$, we want to describe the genealogy of the population at time s given that its genealogy at time 0 is encoded by f. The procedure we follow is illustrated in Figure 2.4. Recall the Kingman comb construction discussed in introduction. Let $(e_i)_{i\geq 1}$ be a sequence of i.i.d. exponential variables, and $(U_i)_{i\geq 1}$ a sequence of i.i.d. uniform $[0,1]$ variables. For $i \geq 1$, we set

$$T_i = \sum_{k\geq i+1} \frac{2}{k(k-1)} e_k.$$

The Kingman comb is given by

$$f_K = \sum_{i\geq 1} T_i \mathbb{1}_{U_i}.$$

It is known from [140], Proposition 3.1, that the above construction generates the comb associated to the flow of bridges, i.e. the Λ-comb associated to the Kingman coalescent. There are only finitely many teeth of f_K that are larger than s, i.e. such that $T_i \geq s$, say N_s. Let σ be their order, e.g. $\sigma(1)$ is the label of the left-most tooth. Consider $V_1^* < \cdots < V_{N_s+1}^*$ the order statistics of $N_s + 1$ independent i.i.d. uniform variables. For $1 \leq k \leq N_s$ let M_k be the greatest tooth of f in the interval (V_k^*, V_{k+1}^*), i.e.

$$M_k = \sup_{(V_k^*, V_{k+1}^*)} f.$$

We define new variables $(\hat{T}_i)_{i\geq 1}$ as follows

$$\forall i > N_s, \quad \hat{T}_i = T_i,$$

and

$$\forall i \leq N_s, \quad \hat{T}_{\sigma(i)} = M_i + s.$$

We define

$$\hat{f}_K = \sum_{i\geq 1} \hat{T}_i \mathbb{1}_{U_i}.$$

Geometrically, the comb $\hat{f}_K$ is obtained through a cutting and pasting procedure illustrated in Figure 2.4.

The above construction defines an operator given by

$$P_t F(f) = \mathbb{E}[F(\hat{f}_K)],$$

for all continuous bounded functions F. We will show below that the family of operators $(P_t)_{t\geq 0}$ is a semi-group. Thus we can define a comb-valued Markov process $(\mathcal{I}^r)_{r\geq 0}$ whose transitions are given by the above construction. We call the process $(\mathcal{I}^r)_{r\geq 0}$ the *evolving Kingman comb.*

Lemma 2.36. *The family of operators $(P_t)_{t\geq 0}$ is a semi-group. Moreover the Kingman comb is a stationary distribution of the evolving Kingman comb.*

Proof. Let $s, t \geq 0$, let f be a deterministic comb. We call f_t the comb obtained through the above procedure at level t starting from f, and f_{t+s} the one obtained according to the above procedure at level s, but using f_t as starting comb. We need to show that f_{t+s} is distributed as f'_{t+s}, the comb obtained at level $t+s$ starting from f.

It is sufficient to show that the portion of the comb f_{t+s} lying between level 0 and $t+s$ is distributed as a Kingman comb truncated at height $t+s$. To show that, it is more convenient to see combs as nested interval-partitions. The procedure described above can be rephrased in terms of composition. Suppose that f_{t+s} has K truncated teeth at time s, this defines $K+1$ intervals of $[0,1]$. For each of these intervals of f_{t+s}, we throw a uniform variable. Two intervals merge at the first moment when their corresponding variables belong to the same subinterval of f_t. This is exactly the description of the ordered paintbox procedure. Thus, using the Markov property of the Kingman comb we know that f_{t+s}, between level 0 and $t+s$, is distributed as the truncation of a Kingman comb. This argument also shows that the Kingman comb is a stationary distribution. □

This construction can be easily extended to the case of Λ-coalescents that come down from infinity, even though we do not have an explicit construction of the comb in this case. In short, to obtain the evolving comb at time s, one needs to sample independently a new comb, erase the portion lying above height s and replace it by teeth sampled from the original comb. In the general case, we have to define the transition of the evolving comb using composition of bridges.

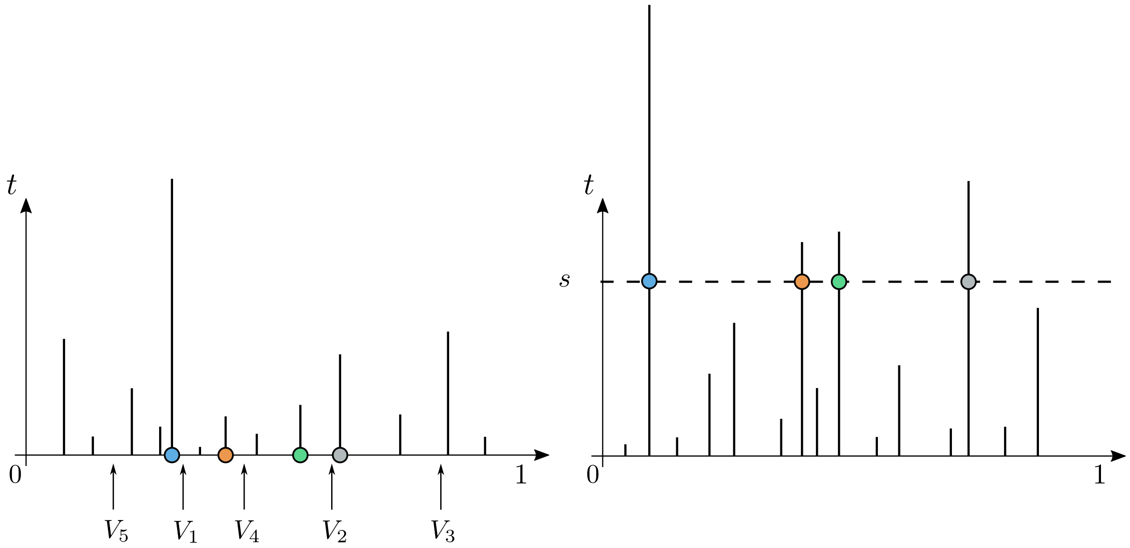


Figure 2.4: Transition of the evolving Kingman comb. The comb at time s, $\hat{f}_K$, is represented on the right, and the initial comb f is on the left. To obtain $\hat{f}_K$, one has first to erase the part of the right comb lying above level s. Here we have erased $N_s = 4$ teeth. Then throw $N_s + 1$ uniform variables $V_1, \dots, V_{N_s+1}$, this defines N_s intervals between these variables, here $(V_5, V_1), (V_1, V_4)$, (V_4, V_2) and (V_2, V_3). Finally take the largest tooth of f in each of these intervals, represented with a coloured root, and paste it in place of the erased tooth.

Again, the evolving comb can be built from the flow of bridges. Let $(B_{s,t})_{t\geq 0}$ be a Λ-flow of bridges, for any time r we can build a nested interval-partition by setting

$$(I^r_t)_{t\geq 0} = (I(B_{r,r+t}))_{t\geq 0}.$$

Then, using a similar argument as in the proof of Corollary 2.33 we could show that the comb-valued process $(\mathcal{I}^r)_{r\geq 0} = ((I^{-r}_t)_{t\geq 0})_{r\geq 0}$ is distributed as the evolving comb introduced above. As a remark this provides a càdlàg modification of the evolving comb, and the Feller property of the flow of bridges ensures that the evolving comb is a Feller process.

2.4 Combs and ultrametric spaces

In this section we envision combs as random UMS. Random metric measure spaces have already been studied in [91, 94]. A key working hypothesis there is that the metric spaces are *separable*. In terms of combs and coalescents, separability translates into absence of dust (see Section 2.4.6). While separability is a very natural hypothesis when considering metric measure spaces, restricting our attention to combs without dust seems arbitrary, as dust has not raised any difficulty so far. In this section we provide a straightforward extension of the framework of random metric measure spaces to account for non-separable UMS.

Let us recall the heuristic of our approach and give a short outline of this section. After a discussion on the assumptions of Definition 2.1 in Section 2.4.1, we define a topology on the space of UMS in Section 2.4.2 by saying that a sequence of UMS converges if the associated sequence of coalescents converges weakly as probability measures. In the separable case, the Gromov reconstruction theorem (see Section $3.\frac{1}{2}.5$ of [94]) ensures that spaces that are indistinguishable have the support of their measures in isometry. In general this result does not hold, we want to obtain a similar result for general UMS. In order to do that, we introduce in Section 2.4.3 the notion of a backbone of a UMS. An UMS can be seen as the leaves of a tree. This tree can be decomposed into 1) a separable part, that we call the backbone and 2) additional subtrees grafted on this backbone. Even though these subtrees can have a complex geometry, from a sampling standpoint they behave as star-trees (recall Figure 2.3). In Section 2.4.4, we show that if two UMS are indistinguishable in the Gromov-weak topology, then they are weakly isometric, in the sense that we can find an isometry between their backbones and a measure-preserving correspondence between the star-trees attached to them (see Proposition 2.47 for a rigorous statement). Finally Section 2.4.5 is dedicated to showing Corollary 2.18, i.e. that we can always find a comb metric space weakly isometric to a given UMS with complete backbone, and Section 2.4.6 is devoted to showing Corollary 2.15 and Proposition 2.16 which are the analogous results in the complete and separable case.

2.4.1 Discussion of Definition 2.1

Recall Definition 2.1 of a UMS from the introduction. This definition has two differences with the "naive" definition of a UMS (that is, any ultrametric space endowed with a probability measure on its Borel σ-field). First, we impose a measurability condition on the metric d. Second we allow the measure μ to be defined on a σ-field that is smaller than the usual Borel σ-field. In this section, we start with a discussion of the assumptions of Definition 2.1.

Let $\mathcal{P}_{\mathrm{coal}}$ denote the state space of coalescents, endowed with its usual Borel σ-field (see [20] Lemma 2.6), and let Π be the map defined as

$$\Pi \colon \begin{cases} U^{\mathbb{N}} \to \mathcal{P}_{\mathrm{coal}} \\ (x_i)_{i\geq 1} \mapsto (\Pi_t)_{t\geq 0}, \end{cases}$$

where

$$i \sim_{\Pi_t} j \iff d(x_i, x_j) \leq t.$$

The following simple lemma proves that the measurability of d is the minimal requirement so that the coalescent obtained by sampling from U is a measurable process.

Lemma 2.37. *The map Π is measurable when $U^{\mathbb{N}}$ is endowed with the product σ-algebra $\mathscr{U}^{\otimes\mathbb{N}}$ iff the distance d is $\mathscr{U} \otimes \mathscr{U}$ measurable.*

Proof. Notice that by definition of Π we have

$$\{d(x_1, x_2) \leq t\} = \{1 \sim_{\Pi_t} 2\}$$

which yields the "only if" part of the proof.

To prove the converse implication, let π be a partition of $[n]$ and define

$$R_{i,j} = \begin{cases} \{d(x_i, x_j) \leq t\} & \text{if } i \sim_\pi j, \\ \{d(x_i, x_j) > t\} & \text{if } i \not\sim_\pi j. \end{cases}$$

Then

$$\left\{ \Pi_t{}_{|[n]} = \pi \right\} = \bigcap_{i,j \leq n} R_{i,j},$$

which ends the proof. □

We now turn to the second point of the definition. Roughly speaking, the Borel σ-field of a non-separable ultrametric space tends to be large, and fewer measures can be defined on it. It is natural to ask whether all coalescents (especially coalescents with dust) can be represented as samples from ultrametric measure spaces, endowed with their natural Borel σ-field. (We call such ultrametric spaces *Borel UMS.*) It turns out that this question can be linked to a deep measure-theoretic problem known as the Banach-Ulam problem. It can be formulated as follows: can we find a space X and a probability measure μ defined on the power set of X such that $\mu(\{x\}) = 0$ for all $x \in X$? The next proposition connects our question to the Banach-Ulam problem. Note that point (iii) yields a positive answer to the problem.

Proposition 2.38. *The following statements are equivalent.*

(i) *There exists an exchangeable coalescent with dust that can be obtained as a sample from a Borel UMS.*

(ii) *Any exchangeable coalescent can be obtained as a sample from a Borel UMS.*

(iii) *There exists an extension of the Lebesgue measure to all subsets of* $\mathbb{R}$.

This proposition is proved in Section 2.F. Proposition 2.38 shows that answering our initial question, that is, representing coalescents with dust as samples from Borel UMS, amounts to finding an extension of the Lebesgue measure to all subsets of $\mathbb{R}$. A treatment of the latter problem requires advanced tools from set theory. Let us recall some basic facts about it. The interested reader is referred to [82] for a complete account on this question and on the Banach-Ulam problem.

A consequence of the various results stated in [82] is that point (iii) of the previous proposition has a greater *consistency strength* than the usual axioms Zermelo-Fraenkel-Choice (ZFC) of set theory. This means that, if ZFC is consistent, further assuming that there exists *no* extension of the Lebesgue measure does not lead to any contradiction. However, even under the assumption that ZFC is consistent, it *cannot* be shown that there is no contradiction in assuming the existence of an extension of the Lebesgue measure.

In other words, assuming that ZFC is consistent, one can safely work under the hypothesis that no extension of the Lebesgue measure exists, and thus that no coalescent with dust can be obtained by sampling from Borel UMS. On the contrary, even assuming that ZFC is consistent, we cannot be sure that further assuming that the Lebesgue measure can be extended (and thus that coalescents with dust are obtained as samples from Borel UMS) will not lead to a contradiction. However, according to the discussion in Remark 1E(e) in [82], it is extremely unlikely that such a contradiction exists, as many consequences of the existence of an extension of the Lebesgue measure have been explored without leading to any contradiction so far.

Remark 2.39. There is a short direct proof that, if the continuum hypothesis and the axiom of choice both hold, there can be no extension of the Lebesgue measure to all subsets of $\mathbb{R}$, see for instance the end of Section 3 of Chapter 2 of [24]. As it is well-known that the continuum hypothesis is relatively consistent with ZFC, this shows that the converse of point (iii) of Proposition 2.38 is also relatively consistent with ZFC.

The greater consistency strength of (iii) is a consequence of Corollary 2E of [82], which states that (iii) is equiconsistent with the existence of a measurable cardinal. Measurable cardinals are instances of (strongly) inaccessible cardinals, whose existence is well-known to have greater consistency strength than ZFC alone, see for instance Theorem 12.12 in [118]. ○

Obviously, all these considerations go far beyond the scope of the current work. The approach we propose is to let the sampling measure be defined on a σ-field smaller than the usual Borel σ-field, namely $\mathscr{U}$. The previous discussion shows that this is not a necessary assumption to be able to represent all coalescents as samples from UMS, but that without it we would need to assume the existence of an extension of the Lebesgue measure to all subsets of $\mathbb{R}$. However, we hope that this short digression has led the reader to the conclusion that, as allowing the sampling measure to be defined on $\mathscr{U}$ avoids the aforementioned set-theoretic issues, it is a more natural framework in which discussing coalescent theory on non-separable UMS than having to assume that one of the statement of Proposition 2.38 holds.

Let us finally discuss the last point of Definition 2.1. This point can be reformulated in terms of the ball σ-field which is defined as follows.

Definition 2.40. Let (U, d) be an ultrametric space. The ball σ-field denoted by $\mathscr{U}_\mathrm{b}$ is the σ-field induced by the open balls of (U, d), that is,

$$\mathscr{U}_\mathrm{b} = \sigma(\{B(x, t) : x \in U, t > 0\}),$$

where

$$\forall x \in U,\, \forall t > 0, \quad B(x, t) = \{y \in U : d(x, y) < t\}. \qquad ○$$

Example 2.41. Consider any set U endowed with the metric

$$\forall x, y \in U, \quad d(x,y) = \mathbb{1}_{\{x \neq y\}}.$$

In this case $\mathscr{U}_\mathrm{b}$ is the countable-cocountable σ-field. ○

The last point of Definition 2.1 can now be rephrased as $\mathscr{U}_\mathrm{b} \subseteq \mathscr{U} \subseteq \mathscr{B}(U)$. It is important to notice that if $\mathscr{B}(U)$ denotes the Borel σ-field of (U,d), then $\mathscr{U}_\mathrm{b} \subseteq \mathscr{B}(U)$ always holds. In that sense, our definition of a UMS should be seen as a generalization of the naive definition as more measures can be defined on $\mathscr{U}_\mathrm{b}$ than on $\mathscr{B}(U)$. The converse statement, i.e. that $\mathscr{B}(U) \subseteq \mathscr{U}_\mathrm{b}$, does not hold in general, as Example 2.41 shows. Nevertheless, in the important case where (U,d) is separable, we have that $\mathscr{U}_\mathrm{b} = \mathscr{B}(U)$, and the ultrametric d is $\mathscr{B}(U) \otimes \mathscr{B}(U)$-measurable. We thus recover the usual framework of metric measure spaces.

Remark 2.42. The ball σ-field appears in other contexts where the underlying metric space is not separable, for example when considering the space of càdlàg functions with the uniform topology, as in [25], Section 6 and Section 15. ○

2.4.2 The Gromov-weak topology

We now define the Gromov-weak topology on the space of UMS. Let $(U, d, \mathscr{U}, \mu)$ be a UMS, and consider $(X_i)_{i \geq 1}$ an i.i.d. sequence distributed as μ. Recall that we define an exchangeable coalescent through the set of relations

$$i \sim_{\Pi_t} j \iff d(X_i, X_j) \leq t.$$

Alternatively, we can see this coalescent as a random pseudo-ultrametric on $\mathbb{N}$ defined as

$$\forall i, j \geq 1, \quad d_\Pi(i,j) = d(X_i, X_j).$$

Both objects encode the same information, as d_Π can be recovered from $(\Pi_t)_{t \geq 0}$ through the equality

$$\forall i, j \geq 1, \quad d_\Pi(i,j) = \inf\{t \geq 0 : i \sim_{\Pi_t} j\}.$$

The distribution of this pseudo-ultrametric is called the *distance matrix distribution* of the UMS.

We use distance matrix distributions to define a topology on the space of UMS. Consider a sequence $(U_n, d_n, \mathscr{U}_n, \mu_n)_{n \geq 1}$ of UMS, and denote by $(\nu_n)_{n \geq 1}$ the associated sequence of distance matrix distributions. We say that the sequence $(U_n, d_n, \mathscr{U}, \mu_n)_{n \geq 1}$ converges in the Gromov-weak topology to $(U, d, \mathscr{U}, \mu)$ if $(\nu_n)_{n \geq 1}$ converges weakly to ν, the distance matrix distribution of (U, d, μ), in the space of probability measures on $\mathbb{R}_+^{\mathbb{N} \times \mathbb{N}}$.

2.4.3 Backbone

It is well known that any ultrametric space (U, d) can be seen as the leaves of a tree. This is illustrated in Figure 2.3. Formally, we work on the space $U \times \mathbb{R}_+$ and consider the pseudo-metric

$$d_T\big((x,s),(y,t)\big) = \max\Big(d(x,y) - \frac{s+t}{2}, \frac{|t-s|}{2}\Big).$$

Let T be the space $U \times \mathbb{R}_+$ quotiented by the equivalence relation

$$z \sim z' \iff d_T(z,z') = 0.$$

Then the space (T, d_T) is a real tree (see [59], Definition 3.15) whose leaves can be identified with (U, d).

Definition 2.43 (Backbone of T)**.** Define

$$f\colon \begin{cases} U \to \mathbb{R}_+ \\ x \mapsto \inf\{t \geq 0 : \mu(B(x,t)) > 0\}, \end{cases}$$

(note that f is measurable since $\mathscr{U}_\mathrm{b} \subseteq \mathscr{U}$) and let

$$\mathcal{S} := \{(x,t) \in T : t \geq f(x)\}.$$

The space $\mathcal{S}$ will be referred to as the backbone of the tree T, and we denote by $d_\mathcal{S}$ the distance d_T restricted to $\mathcal{S}$. ○

Let us now motivate the next result that will be fundamental to our approach. In words, Proposition 2.44 states that even if the underlying UMS is not separable, the backbone is always a separable tree. Secondly, one can recover the whole tree from the backbone by grafting some "simple" subtrees on the skeleton. By "simple", we mean that each of those subtrees has the sampling properties of a star-tree. Let us be more explicit about this last statement and discuss an example.

Consider the space $[0,1] \times \{0,1\}$ endowed with the ultrametric

$$\forall x,y \in [0,1],\, \forall a,b \in \{0,1\}, \quad d\big((x,a),(y,b)\big) = \begin{cases} 1 & \text{if } x \neq y, \\ 1/2 & \text{if } x = y \text{ and } a \neq b, \\ 0 & \text{if } (x,a) = (y,b). \end{cases}$$

The space $([0,1] \times \{0,1\}, d)$ is a star-tree where each branch splits in two at height $1/2$ (see Figure 2.5 left panel), we call it the bifurcating star-tree. We endow this space with the product measure of the Lebesgue measure on $[0,1]$ and the uniform measure on $\{0,1\}$, defined on the usual product Borel σ-field. Consider two independent random variables (X, A) and (Y, B) distributed according to the above measure. We see that these two variables lie at distance $1/2$ iff $X = Y$ and $A \neq B$, which happens with probability 0. Thus, from a sampling point of view, all

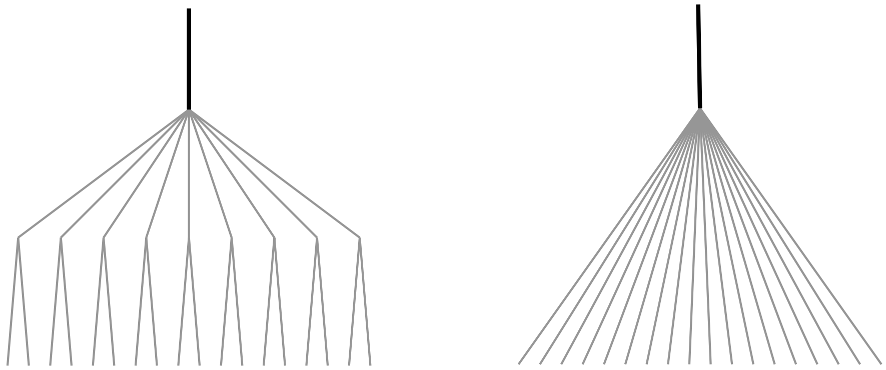


Figure 2.5: Left panel: The bifurcating star-tree. Right panel: The bifurcating star-tree simplified according to the metric $\tilde{d}$. In both cases, the backbone is illustrated with a bold black line and the subtrees attached to it with thin grey lines.

points of the space lie at distance 1 from one another, i.e. the bifurcating star-tree is a star-tree (see Figure 2.5 right panel).

This examples illustrates the more general phenomenon that from the measure point of view, the subtrees attached to the backbone behave like star-trees. More formally, consider a UMS $(U, d, \mathscr{U}, \mu)$. We introduce the distance

$$\forall x, y \in U, \quad \tilde{d}(x,y) = \mathbb{1}_{\{x \neq y\}} \inf\{t \geq 0 : d(x,y) \leq t \text{ and } \mu(B(x,t)) > 0\},$$

which replaces each subtree attached to the backbone by a star-tree. The point (iii) of the following proposition shows that the coalescent obtained by sampling from $(U, d, \mathscr{U}, \mu)$ is the same as the coalescent obtained by sampling from $(U, \tilde{d}, \mathscr{U}, \mu)$.

Proposition 2.44. (i) *The space $(\mathcal{S}, d_{\mathcal{S}})$ is a separable real tree.*

(ii) *The map*

$$\psi\colon \begin{cases} (U, \mathscr{U}) \to (\mathcal{S}, \mathscr{B}(\mathcal{S})) \\ x \mapsto (x, f(x)) \end{cases}$$

is measurable and we define $\mu_{\mathcal{S}} \coloneqq \psi \star \mu$, the pushforward measure (on $(\mathcal{S}, \mathscr{B}(\mathcal{S}))$) of μ by ψ. In particular, the support of $\mu_{\mathcal{S}}$ belongs to the subset of the backbone $\{(x,t) \in \mathcal{S} : t = f(x)\}$.

(iii) *Consider an i.i.d. sequence $(X_i)_{i \geq 1}$ distributed according to μ. Then a.s. for all $i, j \geq 1$, $\tilde{d}(X_i, X_j) = d(X_i, X_j)$.*

Proof. We start by proving (i). The fact that $\mathcal{S}$ is a real tree can be checked directly from the definition. We now show that it is separable. Let $t \in \mathbb{Q}_+$, there are only countably many balls of (U, d) of radius t and positive mass, let us label them $(B_i^t)_{i \geq 1}$. For any $t \in \mathbb{Q}_+$ and $i \geq 1$, let $x_i^t \in B_i^t$. Let us now consider the collection $((x_i^t, t);\, t \in \mathbb{Q}_+, i \geq 1)$. First, since $\mu(B(x_i^t, t)) > 0$, it follows from the definition that $t \geq f(x_i^t)$, and thus $((x_i^t, t);\, t \in \mathbb{Q}_+, i \geq 1)$ is a countable collection of $\mathcal{S}$ and it remains to show that this collection is dense in $\mathcal{S}$.

Let $\varepsilon > 0$ and let $(x,s) \in U \times \mathbb{R}_+$ be in $\mathcal{S}$. We can find $t \in \mathbb{Q}_+$ such that $t > s \geq f(x)$ and $t - s < \varepsilon$. By definition of f, $\mu(B(x,t)) > 0$, and we can find i such that $B(x,t) = B_i^t$. Then $d(x, x_i^t) < t$ and

$$d(x, x_i^t) - \frac{t+s}{2} < d(x, x_i^t) - t + \frac{\varepsilon}{2} < \frac{\varepsilon}{2}$$

and thus $d_T((x,s),(x_i^t,t)) < \varepsilon$. This shows that the collection is dense and that the space is separable.

We now turn to the proof of (ii). Let $(x,t) \in \mathcal{S}$, we denote by

$$C(x,t) = \{(y,s) \in \mathcal{S} : d_T((x,t),(y,t)) = 0\}$$

the *clade* generated by (x,t). In a genealogical interpretation, $C(x,t)$ is the progeny of (x,t) i.e. the subtree that has (x,t) as its MRCA. Notice that this notion can be defined similarly on any rooted tree (here the root is an “infinite point” obtained by letting $t \to \infty$). It is clear that $\psi^{-1}(C(x,t)) = B(x,t)$. Our results is now immediate from the fact that the clades of a rooted separable tree induce the Borel σ-field of the tree. A proof of this fact is given in Section 2.D.

We now prove (iii). It is sufficient to prove that a.s. $d(X,Y) = \tilde{d}(X,Y)$ for X and Y two independent variables distributed as μ. Notice that for any $x, y \in U$, $d(x,y) \leq \tilde{d}(x,y)$. Thus the probability that $d(X,Y) \neq \tilde{d}(X,Y)$ can be written

$$\begin{aligned}\mathbb{P}\Big(d(X,Y) \neq \tilde{d}(X,Y)\Big) &= \iint \mathbb{1}_{\big\{d(x,y) < \tilde{d}(x,y)\big\}} \mu(\mathrm{d}x)\mu(\mathrm{d}y) \\ &= \int \mu(\mathrm{d}x) \int \mu(\mathrm{d}y) \mathbb{1}_{\{d(x,y) < f(x)\}} = 0,\end{aligned}$$

where the last equality can be seen by writing

$$\{x, y \in U : d(x,y) < f(x)\} = \bigcup_{\varepsilon > 0} \{x, y \in U : d(x,y) < f(x) - \varepsilon\}$$

and noticing that each event of the union in the right-hand side has null mass. □

Remark 2.45 (Backbone and marked metric measure space)**.** An object similar to the backbone appears in [95] using the framework of marked metric measure spaces introduced in [44]. We can interpret the backbone as a marked metric measure space where the metric space is U endowed with the backbone metric

$$\bar{d}(x,y) = d_S\Big((x, f(x)), (y, f(y))\Big)$$

and the mark space is $\mathbb{R}_+$. According to this correspondence, backbones are examples of elements of the set $\hat{\mathbb{U}}$ defined in [95]. In [95] the marked metric measure space corresponding to the backbone is either considered as given, or built as the completion of the ultrametric measure space on $\mathbb{N}$ corresponding to the distance matrix distribution. The novelty of the present work is that we start from a general UMS and simplify it to obtain the backbone. This approach requires to identify the measurability assumptions to be made on UMS to avoid the problems that are discussed in Section 2.4.1.

Moreover, the link between backbones and marked metric measure spaces enables us to use the work of [44]. For instance, this provides a metric, the marked Gromov-Prohorov metric, that metrizes the Gromov-weak topology on UMS and ensures that the topology is separable. ○

2.4.4 Isomorphism between backbones

The aim of this section is to introduce the notion of isomorphism between backbones and to prove our reformulation of the Gromov reconstruction theorem.

Definition 2.46. Let $(U, d, \mathscr{U}, \mu)$ and $(U', d', \mathscr{U}', \mu')$ be two UMS with respective backbones $(\mathcal{S}, \mu_{\mathcal{S}})$ and $(\mathcal{S}', \mu'_{\mathcal{S}})$. We say that Φ is an isomorphism from $\mathcal{S}$ to $\mathcal{S}'$ if:

(i) The map Φ is a measure-preserving isometry from $\mathcal{S}$ to $\mathcal{S}'$.

(ii) For every $(x, t) \in \mathcal{S}$, there exists $x' \in U'$ such that $\Phi\big((x, t)\big) = (x', t)$, i.e. Φ preserves the second coordinate.

We say that two UMS are in weak isometry when they have isomorphic backbones. ○

Recall Proposition 2.17 from the introduction. We want to show the following reformulation of Proposition 2.17. In words, this states that having the same distance matrix distribution is equivalent to being weakly isometric.

Proposition 2.47. *Let $(U, d, \mathscr{U}, \mu)$ and $(U', d', \mathscr{U}', \mu')$ be two UMS with respective backbones $(\mathcal{S}, \mu_{\mathcal{S}})$ and $(\mathcal{S}', \mu'_{\mathcal{S}})$. We suppose that the two backbones are complete metric spaces. Then the two spaces $(\mathcal{S}, \mu_{\mathcal{S}})$ and $(\mathcal{S}', \mu'_{\mathcal{S}})$ are isomorphic iff the distance matrix distribution associated $(U, d, \mathscr{U}, \mu)$ and $(U', d', \mathscr{U}', \mu')$ are identical.*

Let us compare this result to the original result from [94]. In the separable case, if two UMS share the same coalescent then the supports of their measures are in isometry. Thus two separable spaces that are indistinguishable in the Gromov-weak topology share the exact same metric structure. The situation is rather different in the general case. Even if two UMS share the same coalescent, they can have rather different metric structures, think of the bifurcating star-tree and the star-tree of Figure 2.5. What Proposition 2.47 states is that in this case there is only a correspondence between coarsenings of the UMS, i.e. the backbones on which all the subtrees are replaced by star-trees. This result is not surprising as the distance matrix distribution only contains the information of a countable number of points, which is not enough to explore the fine metric structure of the UMS.

The "only if" part of Proposition 2.47 is a direct consequence of the following lemma, which shows that the distance matrix distribution of a UMS can be recovered from an i.i.d. sequence of points of the backbone.

Lemma 2.48. *Let $(X_i)_{i\geq 1}$ be an i.i.d. sequence in U sampled according to μ. Then a.s.*

$$\forall i,j \geq 1, \quad d(X_i,X_j) = d_{\mathcal{S}}\Big((X_i,f(X_i)),(X_j,f(X_j))\Big) + \frac{f(X_i)+f(X_j)}{2} \tag{2.6}$$

and

$$\forall i \geq 1, \quad f(X_i) = \inf\{t \geq 0 : \{j : d(X_j,X_i) \leq t\} \text{ is infinite}\}. \tag{2.7}$$

Proof. We know from Proposition 2.44 that for any $i,j \geq 1$, $\tilde{d}(X_i,X_j) = d(X_i,X_j)$ almost surely. Suppose that $(X_i,f(X_i))$ and $(X_j,f(X_j))$ lie at distance 0 in the backbone, then $\tilde{d}(X_i,X_j) = f(X_i) = f(X_j)$ and (2.6) holds. Otherwise notice that $d(X_i,X_j) \geq f(X_i)$ and $d(X_i,X_j) \geq f(X_j)$. Thus

$$d(X_i,X_j) - \frac{f(X_i)+f(X_j)}{2} \geq \frac{|f(X_i)-f(X_j)|}{2}$$

and

$$d_{\mathcal{S}}\Big((X_i,f(X_i)),(X_j,f(X_j))\Big) = d(X_i,X_j) - \frac{f(X_i)+f(X_j)}{2}.$$

The second point of the lemma is a direct consequence of the definition of f and of the observation that if $\mu(B(x,t)) > 0$, then a.s. there are infinitely many $(X_i)_{i\geq 1}$ that belong to this ball. □

It remains to show the converse proposition, i.e. that if two UMS are sampling equivalent then they are in weak isometry. The proof we give is an adaptation of Gromov reconstruction theorem from Section $3.\frac{1}{2}.6$ of [94].

Proof of Proposition 2.47. We say that a sequence $(x_i,t_i)_{i\geq 1}$ in $\mathcal{S}$ is equidistributed if for any $A \in \mathcal{S}$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{(x_i,t_i)\in A\}} = \mu_{\mathcal{S}}(A).$$

A well-known fact is that the empirical measure of an i.i.d. sample converges weakly to the sampling measure. Thus, a.s. an i.i.d. sequence is equidistributed.

Consider the map

$$D \colon \begin{cases} \mathcal{S}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}\times\mathbb{N}} \\ (x_i,t_i)_{i\geq 1} \mapsto \Big(d_{\mathcal{S}}\big((x_i,t_i),(x_j,t_j)\big) + \frac{t_i+t_j}{2}\Big)_{i,j\geq 1}. \end{cases}$$

and let D' be the analogous map for U'. Then Lemma 2.48 shows that the pushforward measure $D\star\mu_{\mathcal{S}}^{\otimes\mathbb{N}}$ is the distance matrix distribution associated to U. Similarly $D' \star {\mu'_{\mathcal{S}}}^{\otimes\mathbb{N}}$ is the distance matrix distribution associated to U'. As we have supposed that the two distance matrix distributions coincide, we can find a sequence $(x_i)_{i\geq 1}$ in U and a corresponding sequence $(x'_i)_{i\geq 1}$ in U' that have the same distance matrix, i.e. such that

$$D\Big((x_i,f(x_i))_{i\geq 1}\Big) = D'\Big((x'_i,f(x'_i)\Big)_{i\geq 1}\Big).$$

We can suppose that these sequences are equidistributed and fulfill equalities (2.6) and (2.7) as all these events have probability 1. Using (2.7) we have

$$\forall i \geq 1, \quad f(x_i) = f(x'_i)$$

and then using (2.6) we obtain

$$\forall i, j \geq 1, \quad d_{\mathcal{S}}\Big((x_i, f(x_i)), (x_j, f(x_j))\Big) = d'_{\mathcal{S}}\Big((x'_i, f(x'_i)), (x'_j, f(x'_j))\Big).$$

We now extend this correspondence to an isomorphism between the backbones. Let $i \geq 1$ and $t \geq f(x_i)$, we set

$$\Phi((x_i, t)) = (x'_i, t).$$

It is clear that Φ is an isomorphism from the set $\{(x_i, t) \in \mathcal{S} : t \geq f(x_i), i \geq 1\}$ to $\{(x'_i, t) \in \mathcal{S}' : t \geq f(x'_i), i \geq 1\}$. It is now sufficient to show that this set is dense to end the proof, by extending Φ to $\mathcal{S}$ by continuity. To see that, let $(x, t) \in \mathcal{S}$. As $t \geq f(x)$, we know that $\mu(\{y \in U : d(x, y) \leq t + \varepsilon\}) > 0$ for any $\varepsilon > 0$. Writing

$$\{y \in U : d(x, y) \leq t + \varepsilon\} = \Big\{y \in U : d_{\mathcal{S}}\Big((x, t+\varepsilon), (y, t+\varepsilon)\Big) = 0\Big\},$$

as $(x_i, f(x_i))_{i \geq 1}$ is equidistributed, we see that we can find $(x_i, f(x_i))$ such that $(x_i, t+\varepsilon) = (x, t+\varepsilon)$. Moreover, it is immediate that $t + \varepsilon \geq f(x_i)$, and we have

$$d_{\mathcal{S}}\Big((x_i, t+\varepsilon), (x, t)\Big) = d_{\mathcal{S}}\Big((x, t+\varepsilon), (x, t)\Big) = \varepsilon.$$

The fact that Φ is measure preserving holds because we have chosen equidistributed sequences. □

Remark 2.49. According to the correspondence between backbones and marked metric measure spaces outlined earlier, Proposition 2.47 is similar to the more general Theorem 1 in [44], which is itself an adaptation of the Gromov reconstruction theorem. However as we only address the case of backbones, we can be more specific. A direct application of Theorem 1 in [44] would only provide an isometry between the supports of the backbones whereas here we obtain a global isometry. ○

Remark 2.50. The results of this section show that the backbone of a UMS contains the same information as the coalescent associated to that UMS. Thus properties of the coalescent can be read off from properties of the backbone. In particular, we can make precise an informal conjecture formulated in the context of exchangeable hierarchies in [76], and addressed in [75], concerning a nice decomposition of the sampling measure μ. Indeed, the sampling measure on the backbone is naturally decomposed into its atoms, its diffuse part on the set $\{(x, t) \in \mathcal{S} : t = 0\}$ of leaves of $\mathcal{S}$ at height 0 and the remaining diffuse part. This decomposition induces three qualitatively different behaviors of the coalescent. In short, points sampled in the atomic part form singletons of the coalescent that all merge at the same time, an event called “broom-like explosion” in [76]. Second, points sampled in $\{(x, t) \in \mathcal{S} : t = 0\}$ always belong to an infinite block of the coalescent for $t > 0$, they form the “iterative branching part”. Finally points sampled in the remaining part of the backbone are singletons of the coalescent that continuously merge with existing blocks. This behavior is referred to as “erosion”. ○

2.4.5 Comb metric measure space, completion of the backbone

An important assumption of Proposition 2.47 is that the backbones of the UMS we consider are complete metric spaces. We will show in this section that the UMS associated to a comb enjoys this property up to the addition of a countable number of points. Let us start with two examples of combs illustrating that the backbone of a comb metric measure space is not in general complete.

First, consider the comb associated to the diadic space. Let $0 < t < 1$ and let k be the only integer such that $t \in \left[2^{-(k+1)}, 2^{-k}\right)$. We set

$$I_t^2 = \bigcup_{0 \leq i \leq 2^{k+1}-1} \left(i2^{-(k+1)}, (i+1)2^{-(k+1)}\right)$$

and for $t \geq 1$ we set

$$I_t^2 = (0,1).$$

The diadic comb is illustrated in Figure 2.6. Now consider the comb metric d_I^2 associated to this comb, and let $x = 2^{-k}$ for some $k \geq 1$. Consider a non-decreasing sequence $(x_n)_{n\geq 1}$ that converges to x. It is not hard to see that $(x_n)_{n\geq 1}$ is Cauchy for d_I^2 but does not admit a limit.

Let us discuss a second example which is not separable. Consider the following comb

$$I_t' = \begin{cases} \emptyset & \text{if } t < 1/2 \\ I_{t-1/2}^2 & \text{otherwise.} \end{cases}$$

This comb is illustrated in Figure 2.6. It is rather clear that the backbone associated to $(I_t')_{t\geq 0}$ is isometric to the backbone obtained from $(I_t^2)_{t\geq 0}$ (notice that here the isometry is not an isomorphism, as the backbone associated to $(I_t')_{t\geq 0}$ is "shifted above by 1/2" from that of $(I_t^2)_{t\geq 0}$). The backbone is not complete for the same reason as above. The following proposition shows that up to the addition of a countable number of points, we can assume that the backbone associated to a comb metric space is complete.

Proposition 2.51. *Consider the comb metric d_I associated to a comb $(I_t)_{t\geq 0}$. We can find a countable set F and an extension $\bar{d}_I$ of d_I to $[0,1] \cup F$ such that $\bar{d}_I$ is ultrametric and the backbone associated to $([0,1] \cup F, \bar{d}_I, \mathscr{I}, \mathrm{Leb})$ is complete.*

Remark 2.52. Here we have implicitly extended the Lebesgue measure to $[0,1] \cup F$ by giving zero mass to F. ∘

A proof of this result is given in Section 2.E. The proof of Corollary 2.18 now directly follows from the various results we have shown.

Proof of Corollary 2.18. Let $(U, d, \mathscr{U}, \mu)$ be a UMS with complete backbone, and let $(\Pi_t)_{t\geq 0}$ be the associated coalescent. Using Theorem 2.8 we can find a nested interval-partition whose associated coalescent is $(\Pi_t)_{t\geq 0}$. We can now use Proposition 2.51 to find a comb metric measure space whose backbone is complete which

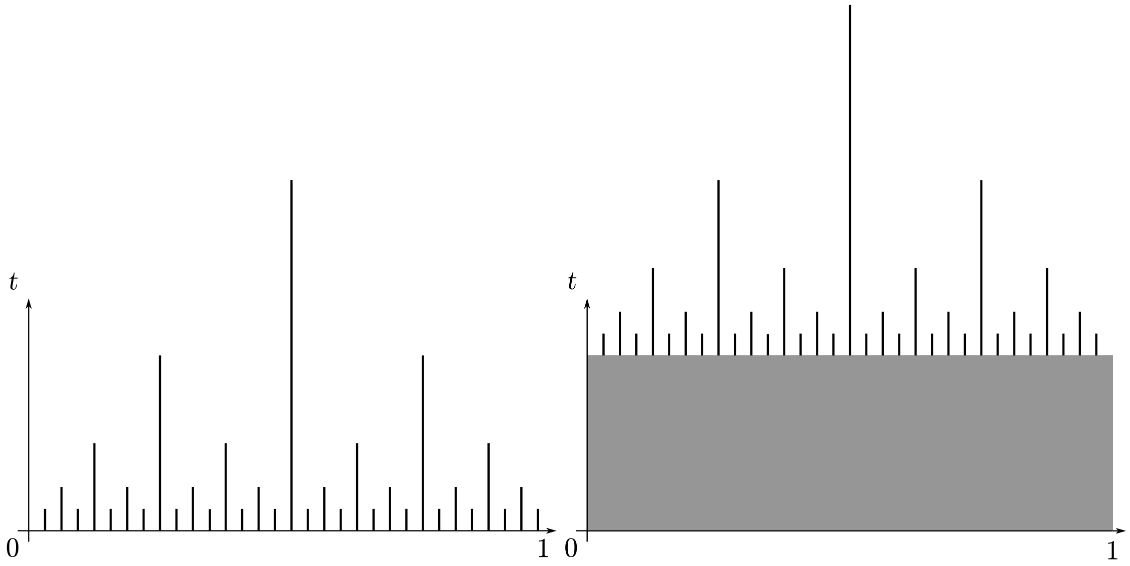


Figure 2.6: Left panel: The diadic comb. Right panel: The comb $(I'_t)_{t\geq 0}$.

has the same distance matrix distribution as $(U, d, \mathscr{U}, \mu)$. Using Proposition 2.47 ends the proof. □

2.4.6 The separable case

In this section we consider the case of separable UMS and prove Corollary 2.15 and Proposition 2.16. The former result states that the weak isometry between backbones can be reinforced to an isometry between the supports of the measures in the case of separable complete UMS. The latter states that any complete separable ultrametric space is isometric to a properly completed comb metric space. Let us start with Corollary 2.15.

Proof of Corollary 2.15. Let $(I_t)_{t\geq 0}$ be a nested interval-partition without dust, and consider the corresponding comb metric measure space $([0,1], d_I, \mathscr{I}, \mathrm{Leb})$. The quotient space of $\{f_I = 0\}$ by the equivalence relation $x \sim y$ iff $d_I(x,y) = 0$ is a separable ultrametric space. Moreover, it is isometric to the subset $\{(x,t) \in \mathcal{S} : t = 0\}$ of the backbone $\mathcal{S}$ of $([0,1], d_I, \mathscr{I}, \mathrm{Leb})$. Thus the quotient space of $(\{f_I = 0\}, d_I)$ can be turned into a complete ultrametric space by adding a countable number of points as in Proposition 2.51, we denote this completion by (U_I, d_I) as in the introduction. As $(I_t)_{t\geq 0}$ has no dust, we have $\mathrm{Leb}(\{f_I = 0\}) = 1$. Thus U_I can be endowed with the pushforward measure of the restriction of Leb to $\{f_I = 0\}$, defined on the Borel σ-field of (U_I, d_I). It is a probability measure, let us denote it by Leb. The space (U_I, d_I, Leb) is a separable complete Borel UMS that has the same distance matrix distribution as the original comb metric measure space $([0,1], d_I, \mathscr{I}, \mathrm{Leb})$.

Let $(U, d, \mathscr{U}, \mu)$ be a complete separable UMS. By restricting our attention to $\mathrm{supp}(\mu)$ we can assume without loss of generality that $\mathrm{supp}(\mu) = U$. Accordinging to Theorem 2.14 we can find a nested interval-partition $(I_t)_{t\geq 0}$ and a corresponding comb metric measure space $([0,1], d_I, \mathscr{I}, \mathrm{Leb})$ whose distance matrix distribution is equal to that of $(U, d, \mathscr{U}, \mu)$. As $\mathrm{supp}(\mu) = U$, for each $t > 0$ we have $\mu(B(x,t)) > 0$. If $(\Pi_t)_{t\geq 0}$ denotes the coalescent obtained by sampling from $(U, d, \mathscr{U}, \mu)$, this shows that for each $t > 0$ all the blocks of Π_t have positive asymptotic frequency. Thus $(I_t)_{t\geq 0}$ has no dust, and we let (U_I, d_I, Leb) be the completion of the comb metric measure space as above. Then (U, d, μ) and (U_I, d_I, Leb) are two complete separable metric measure spaces (in the usual sense) whose distance matrix distributions are equal. Thus, the Gromov reconstruction theorem (see Section $3.\frac{1}{2}.6$ of [94]) proves that we can find a measure-preserving isometry between (U_I, d_I, Leb) and (U, d, μ), which ends the proof. □

We now turn to the proof of Proposition 2.16. We will need the following lemma.

Lemma 2.53. *Any separable ultrametric space* (U, d) *can be endowed with a measure* μ *on its Borel* σ*-field such that* $\mathrm{supp}(\mu) = U$.

Proof. We build the measure by induction. For $n = 1$, as the space is separable there are only countably many balls of radius 1. If there are finitely many such balls, say k balls $B_1, \dots, B_k$, we define

$$\mu(B_i) = \frac{1}{k}.$$

Else we can find an enumeration of the balls, $(B_i)_{i\geq 1}$, and we define

$$\mu(B_i) = \left(\frac{1}{2}\right)^i.$$

Suppose that we have defined $\mu(B)$ for any ball of radius $1/n$. Given a ball B^n of radius $1/n$ there are at most countably many balls $(B_i^{n+1})_{i\geq 1}$ of radius $1/(n+1)$ such that $B_i^{n+1} \subset B^n$. Similarly if there are k balls we define

$$\mu(B_i^{n+1}) = \frac{\mu(B^n)}{k}$$

and if there are countably many balls we define

$$\mu(B_i^{n+1}) = \mu(B^n)\left(\frac{1}{2}\right)^i.$$

A simple application of Caratheodory's extension theorem now provides a probability measure μ defined on the Borel σ-field of (U, d) that extends this measure. It is straightforward from the construction that $\mathrm{supp}(\mu) = U$. □

Remark 2.54. Note that a similar construction was mentioned in [142], where the resulting measure was referred to as the "visibility measure". ○

Proof of Proposition 2.16. Let (U, d) be a separable complete UMS. Lemma 2.53 shows that we can find a measure μ such that $\mathrm{supp}(\mu) = U$. An appeal to Corollary 2.15 now proves the result. □

References for Chapter 2

[20] J. Bertoin. *Random Fragmentation and Coagulation Processes*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2006.

[21] J. Bertoin and J.-F. Le Gall. Stochastic flows associated to coalescent processes. *Probability Theory and Related Fields* **126** (2003), 261–288.

[24] P. Billingsley. *Probability and Measures*. Third edition. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1995.

[25] P. Billingsley. *Convergence of Probability Measures*. Second edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1999.

[44] A. Depperschmidt, A. Greven, and P. Pfaffelhuber. Marked metric measure spaces. *Electronic Communications in Probability* **16** (2011), 174–188.

[48] P. Donnelly and P. Joyce. Consistent ordered sampling distributions: Characterization and convergence. *Advances in Applied Probability* **23** (1991), 229–258.

[59] S. Evans. *Probability and Real Trees. École d'Été de Probabilités de Saint-Flour XXXV-2005*. Vol. 1920. Lecture Notes in Mathematics. Springer, Berlin, Heidelberg, 2007.

[75] N. Forman. Exchangeable hierarchies and mass-structure of weighted real trees. *Electronic Journal of Probability* **25** (2020), 28 pp.

[76] N. Forman, C. Haulk, and J. Pitman. A representation of exchangeable hierarchies by sampling from random real trees. *Probability Theory and Related Fields* **172** (2018), 1–29.

[79] F. Foutel-Rodier, A. Lambert, and E. Schertzer. Exchangeable coalescents, ultrametric spaces, nested interval-partitions: A unifying approach (2019). arXiv: `1807.05165`.

[82] D. H. Fremlin. Real-valued-measurable cardinals. *Set Theory of the Reals, Isreal Mathematical Conference Proceedings*. Vol. 6. 1993, 151–304.

[86] A. Gnedin. The representation of composition structures. *The Annals of Probability* **25** (1997), 1437–1450.

[91] A. Greven, P. Pfaffelhuber, and A. Winter. Convergence in distribution of random metric measure spaces (Λ-coalescent measure trees). *Probability Theory and Related Fields* **145** (2009), 285–322.

[92] A. Greven, P. Pfaffelhuber, and A. Winter. Tree-valued resampling dynamics: Martingale problems and applications. *Probability Theory and Related Fields* **155** (2012), 899–838.

[94] M. Gromov. *Metric structures for Riemannian and non-Riemannian spaces*. Vol. 152. Progress in Mathematics. Birkhäuser Boston, 1999.

[95] S. Gufler. A representation for exchangeable coalescent trees and generalized tree-valued Fleming-Viot processes. *Electronic Journal of Probability* **23** (2018), 42 pp.

[118] T. Jech. *Set Theory*. Third edition. Springer Monographs in Mathematics. Springer-Verlag Berlin Heidelberg, 2003.

[120] O. Kallenberg. *Foundations of Modern Probability*. Second edition. Probability and its Applications. Springer-Verlag New York, 2002.

[123] J. G. Kemeny and J. L. Snell. *Finite Markov Chains*. Undergraduate Texts in Mathematics. Springer-Verlag New York, 1976.

[126] G. Kersting, J. Schweinsberg, and A. Wakolbinger. The evolving beta coalescent. *Electronic Journal of Probability* **19** (2014), 27 pp.

[128] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications* **13** (1982), 235–248.

[137] A. Lambert. Random ultrametric trees and applications. *ESAIM: Procs* **60** (2017), 70–89.

[140] A. Lambert and E. Schertzer. Recovering the Brownian coalescent point process from the Kingman coalescent by conditional sampling. *Bernoulli* **25** (2019), 148–173.

[142] A. Lambert and G. Uribe Bravo. The comb representation of compact ultrametric spaces. *p-Adic Numbers, Ultrametric Analysis and Applications* **9** (2017), 22–38.

[176] P. Pfaffelhuber and A. Wakolbinger. The process of most recent common ancestors in an evolving coalescent. *Stochastic Processes and their Applications* **116** (2006), 1836–1859.

[177] P. Pfaffelhuber, A. Wakolbinger, and H. Weisshaupt. The tree length of an evolving coalescent. *Probability Theory and Related Fields* **151** (2011), 529–557.

[178] J. Pitman. Coalescents with multiple collisions. *The Annals of Probability* **27** (1999), 1870–1902.

[183] C. Rogers and J. Pitman. Markov functions. *The Annals of Probability* **9** (1981), 573–582.

[188] S. Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability* **36** (1999), 1116–1125.

[195] J. Schweinsberg. Coalescents with Simultaneous Multiple Collisions. *Electronic Journal of Probability* **5** (2000), 50 pp.

[197] J. Schweinsberg. Dynamics of the evolving Bolthausen-Sznitman coalecent. *Electronic Journal of Probability* **17** (2012), 50pp.

Appendices for Chapter 2

2.A Exchangeable hierarchies

The aim of this section is to recall some results derived in [76] and discuss the link they have with the current results. Again, we recall that the present work should not be viewed as stemming from the work of [76], but should be viewed as an independent approach bearing similarities that we now expose.

Let X be an infinite space. A hierarchy on X is a collection $\mathcal{H}$ of subsets of X such that

(i) for $x \in X$, $\{x\} \in \mathcal{H}$, $X \in \mathcal{H}$ and $\emptyset \in \mathcal{H}$;

(ii) given $A, B \in \mathcal{H}$, then $A \cap B$ is either A, B or $\emptyset$.

Any ultrametric space encodes a hierarchy that is obtained by "forgetting the time". More precisely, if (U, d) is an ultrametric space, then

$$\mathcal{H} = \{B(x,t),\ x \in X,\ t \geq 0\} \cup \{\{x\},\ x \in X\} \cup \{X, \emptyset\}$$

is a hierarchy. The hierarchy $\mathcal{H}$ encodes the genealogical structure of (U, d), i.e. the order of coalescence of the families, but not the coalescence times.

Remark 2.55. The converse does not hold, there exist hierarchies that cannot be obtained as the collection of balls of an ultrametric space. For example, consider a space X with cardinality greater than the continuum, endowed with a total order $\leq$, and define

$$\mathcal{H} = \{\{y : y \leq x\} : x \in X\} \cup \{\{x\},\ x \in X\} \cup \{X, \emptyset\}. \qquad \circ$$

The main object studied in [76] are exchangeable hierarchies on $\mathbb{N}$. Let σ be a permutation of $\mathbb{N}$, and $\mathcal{H}$ be a hierarchy on $\mathbb{N}$. Then σ naturally acts on $\mathcal{H}$ as

$$\sigma(\mathcal{H}) = \{\sigma(A),\ A \in \mathcal{H}\}.$$

A random hierarchy on $\mathbb{N}$ (see [76] for a definition of the σ-field associated to hierarchies) is called exchangeable if for any permutation σ,

$$\sigma(\mathcal{H}) \overset{(d)}{=} \mathcal{H}.$$

In a similar way that exchangeable coalescents are obtained by sampling in UMS, exchangeable hierarchies are obtained by sampling in hierarchies on measure spaces. Let (X, μ) be a probability space, and consider a hierarchy $\mathcal{H}$ on X. An exchangeable hierarchy $\mathcal{H}'$ can be generated out of an i.i.d. sequence $(X_i)_{i\geq 1}$ by defining

$$\mathcal{H}' = \{\{i \geq 1 : X_i \in A\},\ A \in \mathcal{H}\}.$$

Again, an exchangeable hierarchy can be obtained from an exchangeable coalescent by forgetting the time. Let $(\Pi_t)_{t\geq 0}$ be an exchangeable coalescent. Then

$$\mathcal{H} = \{B,\ B \text{ is a block of } \Pi_t,\ t \geq 0\}$$

is an exchangeable hierarchy.

The main results in [76] show that any exchangeable hierarchy can be obtained by sampling from 1) a random "interval hierarchy" on $[0, 1)$ and 2) a random real-tree. The link with our results now seems straightforward.

An interval hierarchy on $[0, 1)$ is a hierarchy $\mathcal{H}$ on $[0, 1)$ such that all non-singleton elements of $\mathcal{H}$ are intervals. Again, an interval hierarchy can be obtained from a nested interval-partition $(I_t)_{t\geq 0}$ by forgetting the time. The family of sets

$$\begin{aligned}\mathcal{H} = \{I &: I \text{ is an interval component of } I_t, t \geq 0\}\\ &\cup \{\{x\}, x \in [0, 1)\}\\ &\cup \{[0, 1), \emptyset\}\end{aligned}$$

is an interval hierarchy. Theorem 4 in [76] states that any exchangeable hierarchy on $\mathbb{N}$ can be obtained by sampling in a random interval hierarchy. This is the direct equivalent of our Theorem 2.8 that states that any exchangeable coalescent can be obtained by sampling in a random nested interval-partition.

Consider a measure rooted real-tree (T, d, ρ, μ), it can be endowed with a partial order $\preceq$ such that $y \preceq x$ if x is an ancestor of y (see [59]). Then, the fringe subtree of T rooted at $x \in T$ is defined as the set

$$F_T(x) = \{y \in T : y \preceq x\},$$

it is the set of the offspring of x. The natural hierarchy associated to (T, d, ρ) is

$$\mathcal{H} = \{F_T(x),\ x \in T\}.$$

Theorem 5 in [76] states that any exchangeable hierarchy can be obtained by sampling in the hierarchy associated to a random measure rooted real-tree. In our framework, we have seen that a nested interval-partition can be seen as an ultrametric space, and in Section 2.4.5 we have seen how this ultrametric space is embedded in a real-tree. Again we have proved here the reformulation of Theorem 5 from [76].

In a subsequent work, one of the authors has introduced the notion of mass-structural isomorphism [75]. In a nutshell, two trees that are mass-structural isomorphic induce the same exchangeable hierarchy. In our framework, two spaces have the same coalescent iff their backbones are isomorphic. Thus, the mass-structural isomorphism is replaced here by the simpler notion of isomorphism.

Overall, the two works are very similar in the sense that they obtain the same kind of representation results for exchangeable hierarchies and exchangeable coalescents. However the techniques used in the proofs are different, e.g. the work of [76] relies on spinal decomposition whereas the present work relies on nested compositions. Moreover, as an ultrametric space contains "more information" than a hierarchy, our results are not trivially implied by the results in [76], but constitute an extension of their work.

Finally, we wish to stress two things. First, most of the difficulties that Section 2.4 deals with stem from the fact that we consider non-separable metric spaces. These issues and the work that is done here heavily relies on the theory of metric spaces. Seeing genealogies as metric spaces is only possible if we keep the information on the times of coalescence, which is not the case when considering hierarchies.

Second, keeping this information allows us to study genealogies as time-indexed stochastic processes. It is a necessary step to study the Markov property of the combs associated to Λ-coalescents as in Section 2.3. This creates a direct link between the present work and the very rich literature on Λ-coalescents and coalescence theory that is not present in [76]. Moreover, this provides a new approach to the question of dynamical genealogies, with the introduction of the dynamical comb.

2.B Independence of the nested interval-partitions and the sampling variables

Consider an exchangeable nested composition $(\mathcal{C}_t)_{t\geq 0}$, and let $(I_t)_{t\in\mathbb{Q}_+}$ be the nested interval-partition obtained by applying Theorem 2.19 distinctly for any $t \in \mathbb{Q}_+$, and $(V_i)_{i\geq 1}$ be the sequence of i.i.d. uniform variables obtained from Theorem 2.19 applied at time 0. The aim of this section is to show that $(V_i)_{i\geq 1}$ is independent from $(I_t)_{t\in\mathbb{Q}_+}$.

Let $0 = t_0 < t_1 < \cdots < t_p$. We can build a collection of sequences $(\xi_i^{(k)})_{i\geq 1, k=0,\dots,p}$ where for $k = 0,\dots,p$ and $i \geq 1$,

$$\xi_i^{(k)} = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{j \preceq_k i\}},$$

and $\preceq_k$ is the partial order on $\mathbb{N}$ representing $\mathcal{C}_{t_k}$ as in Section 2.2.1. The sequence of vectors $(\xi_i^{(0)},\dots,\xi_i^{(p)})_{i\geq 1}$ is exchangeable. Thus by applying a vectorial version of de Finetti's theorem we know that there exists a measure μ on $[0,1]^{p+1}$ such that conditionally on μ the sequence of vectors is i.i.d. distributed as μ. We can now "spread" the variables $(\xi_i^{(0)})_{i\geq 1}$ using an independent i.i.d. uniform sequence as in the proof of Theorem 2.19 to obtain a sequence $(V_i)_{i\geq 1}$ that is i.i.d. uniform conditionally on μ. Thus the sequence $(V_i)_{i\geq 1}$ is independent of μ. The interval-partitions $(I_{t_0},\dots,I_{t_p})$ can be recovered from the push-forward measures of μ by the coordinate maps on $\mathbb{R}^{p+1}$. Thus $(V_i)_{i\geq 1}$ is independent from $(I_t)_{t\in\mathbb{Q}_+}$.

2.C Generator calculation

Let $n \geq 1$ and let $\hat{Q}_n$ denote the generator of the nested composition $(\mathcal{C}_t^n)_{t\geq 0}$ defined from the transition rates $(\tilde{\lambda}_{b,k};\ 2 \leq k \leq b < \infty)$. Let Q_n be the generator of the restriction to $[n]$ of a Λ-coalescent. Here we show that for any function f, from the space of compositions of $[n]$ to $\mathbb{R}$,

$$\forall \pi, \quad \hat{Q}_n L_n f(\pi) = L_n Q_n f(\pi),$$

where L_n is the operator defined in the proof of Lemma 2.28.

We will need additional notations. The space of partitions and compositions of $[n]$ will be denoted by P_n and S_n respectively. For $\pi, \pi' \in P_n$, we denote by $q_{\pi,\pi'}$ the transition rate from π to π', i.e. $q_{\pi,\pi'} = \lambda_{b,k}$ if π has b blocks and π' is obtained by merging k blocks of π, and $q_{\pi,\pi'} = 0$ otherwise. Similarly for $c, c' \in S_n$ we define $q_{c,c'}$ to be the transition rate from c to c'. Finally, we denote by $O(\pi)$ the set of compositions of $[n]$ whose blocks are given by the partition π, and $\mathrm{Card}(\pi)$ the number of blocks of π. Let $\pi \in P_n$ and denote by b the number of blocks of π, we have

$$\begin{aligned}
\hat{Q}_n L_n f(\pi) &= \sum_{\pi' \in P_n} q_{\pi,\pi'} (L_n f(\pi') - L_n f(\pi)) \\
&= \sum_{\pi' \in P_n} q_{\pi,\pi'} \left(\sum_{c' \in O(\pi')} \frac{1}{\mathrm{Card}(\pi')!} f(c') - \sum_{c \in O(\pi)} \frac{1}{\mathrm{Card}(\pi)!} f(c) \right) \\
&= \sum_{\pi' \in P_n} \sum_{c' \in O(\pi')} q_{\pi,\pi'} \frac{1}{\mathrm{Card}(\pi')!} f(c') - \sum_{c \in O(\pi)} \sum_{k=2}^{b} \frac{1}{\mathrm{Card}(\pi)!} \binom{b}{k} \lambda_{b,k} f(c).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
L_n Q_n f(\pi) &= \sum_{c \in O(\pi)} \frac{1}{\mathrm{Card}(\pi)!} Q_n f(c) \\
&= \sum_{c \in O(\pi)} \frac{1}{\mathrm{Card}(\pi)!} \sum_{c' \in S_n} q_{c,c'} (f(c') - f(c)) \\
&= \sum_{c \in O(\pi)} \frac{1}{\mathrm{Card}(\pi)!} \sum_{c' \in S_n} q_{c,c'} f(c') - \sum_{c \in O(\pi)} \sum_{k=2}^{b} \frac{1}{\mathrm{Card}(\pi)!} \binom{b}{k} \lambda_{b,k} f(c).
\end{aligned}$$

We will end the calculation by showing that for any $c' \in S_n$, the coefficient in front of the term $f(c')$ in the left sum is the same for both expression. Let π' be the partition associated to c'. If π' is not obtained by merging k blocks of π for some k, then the coefficient of the term $f(c')$ in the sum is 0 in both expressions. Now suppose that π' is obtained by merging k blocks of π. In the first expression, we first choose the blocks of π that merge to get π' and then order the resulting partition to get the composition c'. There is only one possible way to do that and obtain a given c'. Thus the coefficient in front of $f(c')$ is $\lambda_{b,k}/(b-k+1)!$. In the second expression, we first choose an order to obtain a composition c, and then merge its blocks to get the composition c'. There are $k!$ possible orderings of π, and then exactly one merger of c that lead to c' (we can take any permutation of the k blocks that merge). Thus the coefficient in front of term $f(c')$ is

$$\frac{k!}{b!}\tilde{\lambda}_{b,k} = \frac{k!}{b!}\frac{1}{b-k+1}\frac{b!}{k!\,(b-k)!}\lambda_{b,k} = \frac{1}{(b-k+1)!}\lambda_{b,k}.$$

2.D Measurability of separable rooted trees

In this section we prove the claim made in the proof of Proposition 2.44 that the Borel σ-field of a separable rooted tree is induced by the clades of the tree. Let us be more specific.

We consider a separable real-tree (T, d) with a particular point $\rho \in T$ that we call the root. For $x, y \in T$, we denote by $[x, y]$ the unique geodesic with endpoints x and y (see [59]). Recall from Section 2.A the fringe subtree of T rooted at x equivalently defined as the *clade*

$$C(x) = \{y \in T : x \in [\rho, y]\},$$

see Figure 2.7 for an illustration. The claim is that

$$\sigma(\{C(x),\ x \in T\}) = \mathscr{B}(T).$$

Remark 2.56. Our goal in the proof of Proposition 2.44 is to apply the result to the backbone whose root should be such that clades are the balls of U. This can be done by seeing the backbone as having a root “at infinity”. ○

Let $x \in T$ and $\varepsilon > 0$, we assume that $\varepsilon < d(x, \rho)$. We denote by $B(x, \varepsilon)$ the open ball centered in x with radius ε, and $S(x, \varepsilon)$ the sphere of center x and radius ε, i.e.

$$S(x, \varepsilon) = \{y \in T : d(x, y) = \varepsilon\}.$$

There is a unique point in $a \in [\rho, x] \cap S(x, \varepsilon)$. It is clear that

$$B(x, \varepsilon) = C(a) \setminus \bigcup_{y \in S(x,\varepsilon)\setminus\{a\}} C(y).$$

Let $y \in S(x, \varepsilon)$, and $0 < \eta < \varepsilon$, we denote by y_η the only point in $[y, x]$ such that $d(y_\eta, y) = \eta$. We can write

$$\bigcup_{y \in S(x,\varepsilon)\setminus\{a\}} C(y) = \bigcap_{\eta > 0} \bigcup_{y \in S(x,\varepsilon)\setminus\{a\}} C(y_\eta).$$

The claim is proved if we can show that the union on the right-hand side is countable. This holds due to the separability of (T, d). To see that notice that by uniqueness of the geodesic, if y and y' are such that $y_\eta \neq y'_\eta$, then $d(y, y') > \eta$. Thus if the set $\{y_\eta : y \in S(x, \varepsilon) \setminus \{a\}\}$ is not countable, we can find an uncountable subset of $S(x, \varepsilon)$ such that any two points lie at distance at least η. This is not possible due to separability.

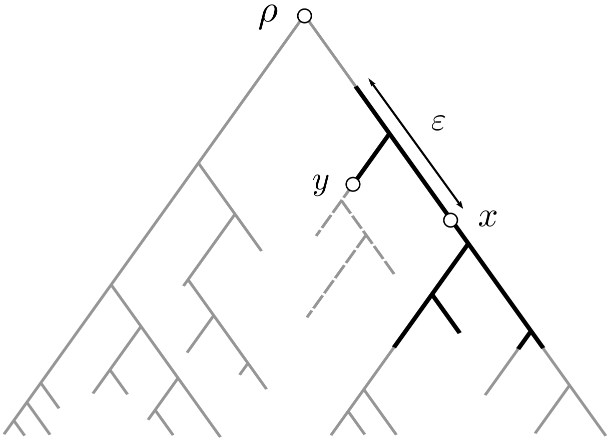


Figure 2.7: A tree rooted at ρ. The ball of radius ε and center x is represented by the black bold lines. An example of $y \in S(x,\varepsilon)$ is given, and its corresponding clade $C(y)$ is represented by grey dashed lines.

2.E Comb completion

In this section we prove Proposition 2.51, i.e. that the backbone of a comb is complete up to the addition of a countable number of points. We start from a nested interval-partition $(I_t)_{t\geq 0}$. We define

$$\mathcal{R} = \{x \in [0,1] : \exists s_x, t_x \text{ s.t. } x \text{ is the right endpoint} \\ \text{of an interval component of } I_u \text{ for } u \in [s_x, t_x]\}$$

and

$$\mathcal{L} = \{x \in [0,1] : \exists s_x, t_x \text{ s.t. } x \text{ is the left endpoint} \\ \text{of an interval component of } I_u \text{ for } u \in [s_x, t_x]\}.$$

We now work with a subset of $[0,1] \times \{0, r, \ell\}$. Let

$$\bar{I} = ([0,1] \times \{0\}) \cup (\mathcal{R} \times \{r\}) \cup (\mathcal{L} \times \{\ell\}).$$

We will simply write x for $(x,0)$, x_r for (x,r) if $x \in \mathcal{R}$ and x_ℓ for (x,ℓ) if $x \in \mathcal{L}$. We extend d_I to $\bar{I}$ in the following way. Let $x < y$, we define

$$\bar{d}_I(x,y) = \bar{d}_I(x,y_\ell) = \bar{d}_I(x_r,y_\ell) = \bar{d}_I(x_r,y) = \sup_{[x,y]} f_I$$
$$\bar{d}_I(x,y_r) = \bar{d}_I(x_r,y_r) = \sup_{[x,y)} f_I$$
$$\bar{d}_I(x_\ell,y) = \bar{d}_I(x_\ell,y_\ell) = \sup_{(x,y]} f_I$$
$$\bar{d}_I(x_\ell,y_r) = \sup_{(x,y)} f_I$$

and $\bar{d}_I(x_r, x_\ell) = f(x)$. We use symmetrized definitions if $x > y$. It is straightforward to check that $\bar{d}_I$ is a pseudo-ultrametric. We will denote by $\mathcal{S}_I$ the backbone associated to this UMS, and $d_{\mathcal{S}_I}$ the restriction of the tree metric to $\mathcal{S}_I$, i.e.

$$\forall (x', t),\, (y', s) \in \mathcal{S}_I, \quad d_{\mathcal{S}_I}\Big((x', t), (y', s)\Big) = \max\left\{\bar{d}_I(x', y') - \frac{t+s}{2}, \frac{|t-s|}{2}\right\}.$$

Lemma 2.57. *The backbone* $(\mathcal{S}_I, d_{\mathcal{S}_I}, \mathrm{Leb})$ *associated to* $(\bar{I}, \bar{d}_I, \mathrm{Leb})$ *is a complete metric space.*

Proof. Consider $(x'_n, t_n)_{n\geq 1}$ a Cauchy sequence in $\bar{I}$ for the metric $d_{\mathcal{S}_I}$. As

$$\frac{|t_n - t_m|}{2} \leq d_{\mathcal{S}_I}\Big((x'_n, t_n), (x'_m, t_m)\Big),$$

the sequence $(t_n)_{n\geq 1}$ is Cauchy and converges to a limit that we denote by t. Each point x'_n can be written as $x'_n = (x_n, a_n)$ with $x_n \in [0,1]$ and $a_n \in \{0, r, \ell\}$. The sequence $(x_n)_{n\geq 1}$ admits a subsequence that converges to a limit x for the usual topology in $[0,1]$. Without loss of generality we can assume that $(x_n)_{n\geq 1}$ is non-decreasing and converges to x.

Using the fact that the sequence is Cauchy, we know that

$$\lim_{n\to\infty} \sup_{m\geq n} \bar{d}_I(x'_n, x'_m) - \frac{t_n + t_m}{2} \leq 0,$$

which directly implies that

$$\lim_{\varepsilon\to 0} \sup_{[x-\varepsilon, x)} f_I \leq t.$$

Suppose that $x \in \mathcal{R}$. By definition of $\bar{d}_I$ and the above remark,

$$\lim_{n\to\infty} \bar{d}_I(x'_n, x_r) - \frac{t_n + t}{2} \leq 0.$$

Thus the sequence $(x'_n, t_n)_{n\geq 1}$ converges to (x_r, t).

Now suppose that $x \notin \mathcal{R}$. We claim that

$$\lim_{\varepsilon\to 0} \sup_{[x-\varepsilon, x)} f_I = f(x).$$

As $x \notin \mathcal{R}$ we directly know that

$$\lim_{\varepsilon\to 0} \sup_{[x-\varepsilon, x)} f_I \geq f(x).$$

Suppose that the above limit is strictly greater than $f(x)$. Then we can find a non-decreasing sequence $(y_n)_{n\geq 1}$ converging to x in the usual topology such that $f(y_n) \downarrow \lambda > f(x)$ as n goes to infinity. Let $\eta < \lambda - f(x)$. Notice that the set $\{y \in [0,1] : f(y) > \lambda - \eta\}$ is closed in the usual topology, as it is the complement of $I_{\lambda-\eta}$. This shows that x belongs to this set, which is a contradiction. Our claim is proved. Similarly to above, it is now immediate that

$$\lim_{n\to\infty} \bar{d}_I(x'_n, x) - \frac{t_n + t}{2} \leq 0.$$

and that $(x'_n, t_n)_{n\geq 1}$ converges to (x, t). □

Remark 2.58. This completion is already present in the compact case in [142]. In this case, we have $\mathcal{R} = \mathcal{L} = \{f_I > 0\}$. ○

2.F The link between dust and the Banach-Ulam problem

In this section we prove Proposition 2.38. We prove this result by constructing a solution to the so-called Banach-Ulam problem. This problem can be formulated as follows: is it possible to find a space X with a probability measure μ on the power-set $\mathcal{P}(X)$ of X such that $\mu(\{x\}) = 0$ for all $x \in X$?

Recall that a UMS $(U, d, \mathscr{U}, \mu)$ is called a Borel UMS if $\mathscr{U}$ is the Borel σ-field of (U, d). The support of the measure μ, $\operatorname{supp}(\mu)$, is defined as the intersection of all balls with positive mass. Equivalently, it can be defined as

$$\operatorname{supp}(U) = \{x \in U : \forall t > 0,\ \mu(B(x,t)) > 0\}.$$

We start with the following lemma, which gives a necessary and sufficient condition for the coalescent sampled from U to have dust in terms of the support of μ.

Lemma 2.59. *Let $(U, d, \mathscr{U}, \mu)$ be a UMS, and let $(\Pi_t)_{t\geq 0}$ be the associated coalescent. Then $(\Pi_t)_{t\geq 0}$ has dust iff $\mu(\operatorname{supp}(\mu)) < 1$.*

Proof. Let $(X_i)_{i\geq 1}$ be an i.i.d. sequence in U distributed as μ and let $(\Pi_t)_{t\geq 0}$ be the coalescent obtained as above. We say that i is in the dust of the coalescent if there exists $t > 0$ such that $\{i\}$ is a singleton block of Π_t. We show that a.s.

$$i \text{ is in the dust } \iff X_i \notin \operatorname{supp}(\mu).$$

Suppose that $X_i \in \operatorname{supp}(\mu)$. Then for any $t > 0$, $\mu(B(X_i, t)) > 0$, thus a.s. there are infinitely many other variables $(X_j)_{j\geq 1}$ in $B(X_i, t)$. Thus X_i is in an infinite block of Π_t. Conversely suppose that i is not in the dust, i.e. that for any $t > 0$, $\{i\}$ is not a singleton block. Using Kingman's representation theorem for exchangeable partitions, we know that the block of i is a.s. infinite and has a positive asymptotic frequency f_i. The law of large numbers shows that $f_i = \mu(B(X_i, t)) > 0$. □

Proof of Proposition 2.38. Let us first show that (i) implies (iii). Let $(U, d, \mathscr{U}, \mu)$ be a Borel UMS with associated coalescent $(\Pi_t)_{t\geq 0}$. Suppose that $(\Pi_t)_{t\geq 0}$ has dust. According to Lemma 2.59, we have $\mu(\operatorname{supp}(\mu)) < 1$. Consider $t > 0$ and let $(B^t_\alpha)_{\alpha\in A_t}$ be the collection of open balls of radius t with zero mass, where A_t is just an index set. We know that

$$\bigcup_{t>0} \bigcup_{\alpha\in A_t} B^t_\alpha = U \setminus \operatorname{supp}(\mu).$$

Using the continuity from below of the measure μ, we can find an $\varepsilon > 0$ such that $\mu(\bigcup_{\alpha\in A_\varepsilon} B^\varepsilon_\alpha) > 0$. We now consider the equivalence relation

$$x \sim y \iff d(x,y) < \varepsilon$$

and denote by X the quotient space of $\bigcup_{\alpha\in A_\varepsilon} B_\alpha^\varepsilon$ for the relation $\sim$. We define the quotient map as

$$\varphi\colon \begin{cases} U \to X \\ x \mapsto \{y \in U : d(x,y) < \varepsilon\}. \end{cases}$$

We claim that φ is continuous when U is equipped with the metric topology induced by d, and X is equipped with the discrete topology $\mathcal{P}(X)$. Let $C \subset X$, then

$$\varphi^{-1}(C) = \bigcup_{x\in\varphi^{-1}(C)} B(x,\varepsilon)$$

which is an open subset of U. We call μ_X the push-forward measure of μ by the map φ. The measure $\mu_X/\mu_X(X)$ is a diffuse probability measure defined on $\mathcal{P}(X)$ as required. Thus, $(X, \mathcal{P}(X), \mu_X)$ is a solution to the Banach-Ulam problem.

Using the terminology from [82], this proves that the cardinality of X is a real-valued cardinal (see Notation 1C in [82]). According to Ulam's theorem (see Theorem 1D in [82]), real-valued cardinals fall into two classes: atomlessly-measurable cardinals and two-valued-measurable cardinals. The cardinal of X is atomlessly-measurable. To see this, one can for example notice that our measurability assumption on d implies that the cardinality of U (and thus that of X) is not larger than the continuum. (If this does not hold, then the diagonal does not belong to the product σ-field $\mathcal{P}(U) \otimes \mathcal{P}(U)$ and the metric d is not measurable.) Finally, using Theorem 1D of [82] proves (iii).

The fact that (ii) implies (i) is obvious, it remains to show that (iii) implies (ii). Suppose that there exists an extension of the Lebesgue measure to all subsets of $\mathbb{R}$, let us denote by $\overline{\mathrm{Leb}}$ its restriction to $[0,1]$. Let $(\Pi_t)_{t\geq 0}$ be any coalescent with dust. By Theorem 2.8 we can find a nested interval-partition $(I_t)_{t\geq 0}$ such that the paintbox based on $(I_t)_{t\geq 0}$ is distributed as $(\Pi_t)_{t\geq 0}$. Let d_I be the corresponding comb metric on $[0,1]$. Then $([0,1], d_I, \mathscr{B}_I([0,1]), \overline{\mathrm{Leb}})$ is a UMS, where $\mathscr{B}_I([0,1])$ refers to the Borel σ-field induced by d_I and $\overline{\mathrm{Leb}}$ is restricted to that σ-field. The coalescent obtained by sampling from this UMS is distributed as $(\Pi_t)_{t\geq 0}$. □

Chapter 3

3

Kingman's coalescent with erosion

This chapter is joint work with Amaury Lambert and Emmanuel Schertzer. It is published in the *Electronic Journal of Probability* [80].

Illustration. Simulation of Kingman's coalescent with erosion. Time is going downwards and each black line separates two blocks, whose sizes are given by the length of the interval between the lines. This simulation is based on the construction of Kingman's coalescent with erosion from a flow of bridges given in Proposition 3.11.

3.1 Introduction

3.1.1 Motivation

In evolutionary biology, speciation refers to the event when two populations from the same species lose the ability to exchange genetic material, e.g. due to the formation of a new geographic barrier or accumulation of genetic incompatibilities. Even if speciation is usually thought of as irreversible, related species can often still exchange genetic material through exceptional hybridization, migration events or sudden collapse of a geographic barrier [186]. This can lead to the transmission of chunks of DNA between different species, a phenomenon known as introgression, which is currently considered as a major evolutionary force shaping the genomes of groups of related species [151]. Our study of Kingman's coalescent with erosion was first motivated by the following model of speciation incorporating rare migration events, depicted in Figure 3.1.

Consider a set of N species, each harboring a genome of n genes indexed by $\{1, \dots, n\}$. We suppose that the species are monomorphic, i.e., that all individuals in the same species carry the same alleles, and that their dynamics is given by a Moran model: at rate one for each pair of species (s_1, s_2), species s_2 dies, s_1 gives birth to a new species, replicates its genome and sends it into the daughter species. Moreover, we assume that the species are closely related and that they retain the ability to exchange genetic material at exceptional migration events. This effect is incorporated into the model by stating that at rate d for each gene $g \in \{1, \dots, n\}$

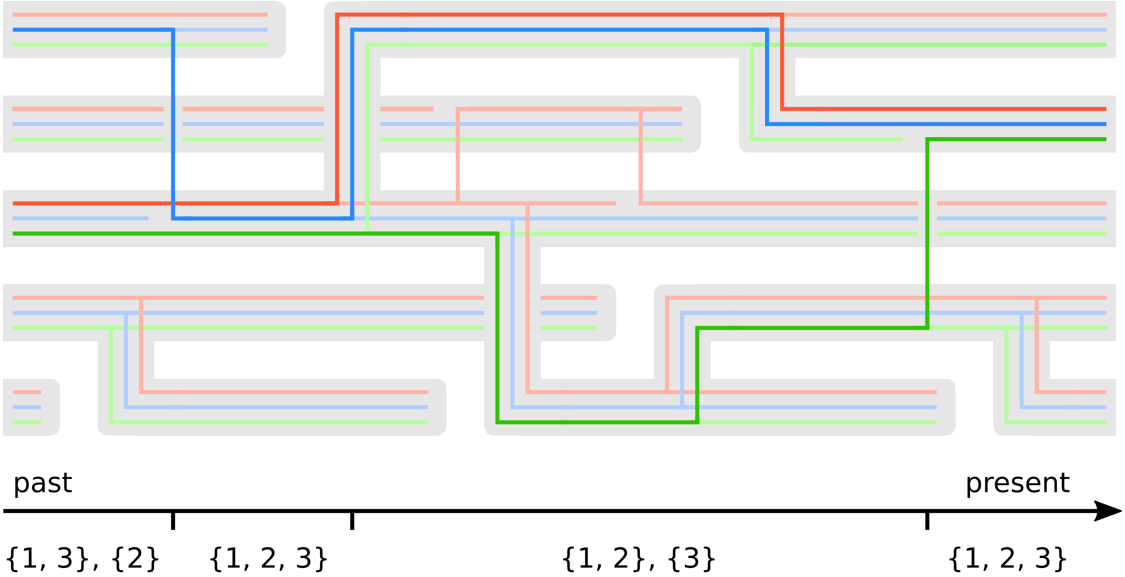


Figure 3.1: Illustration of the model with $N = 5$ species, represented by grey tubes, and $n = 3$ genes, represented by the colored lines inside the tubes. A species can split into two, simultaneously replicating its genome (speciation). A gene can replicate and move from one species to another and then replace its homologous copy in the recipient species (introgression). At present time a randomly chosen species is sampled: the ancestral lineages of its genes are represented with bolder colors. The green lineage is first subject to an introgression event and jumps to a new species. It is then brought back to the same species as the other genes by a coalescence event. The corresponding partition-valued process obtained by assigning the labels 1, 2 and 3 to the red, blue and green gene respectively is given.

and each pair of species s_1 and s_2, the gene g is replicated, the new copy of g is sent from s_1 to s_2 and replaces its homolog in s_2.

The assumption that each migrant transmits at most one gene to the recipient species is strong. A more realistic model should allow any subset of the n genes to be transmitted, at a rate that depends in a complex way on the geometry of the genome due to the biological nature of recombination. However, if recombination is sufficiently strong and if the number of individuals in each species is large, each time a migrant goes from species s_1 to s_2, its genome is rapidly broken into small segments due to frequent back-crosses with the resident. Each of these segments behaves almost independently from the other segments, and has a small probability to reach fixation. Thus, to the first order there should be at most one segment that can reach fixation at a time, as we have assumed.

Now consider a fixed large time T, and sample uniformly one species at that time. We follow backwards in time the ancestral lineages of its genes and the ancestral species to which those genes belong. This induces a process valued in the partitions of $\{1, \dots, n\}$ by declaring that i and j are in the same block at time t if the ancestral lineages of genes i and j sampled at T lie in the same ancestral species at time $T - t$.

At first ($t = 0$), all genes belong to the same ancestral species. Eventually this species receives a successful migrant from another species. Backwards in time, the gene that has been transmitted during this event is removed from its original species and placed in the migrant's original species. Such events occur at rate $(N - 1)d$ for each gene, and the migrant species is then chosen uniformly in the population. Once genes belong to separate species, they can be brought back to the same species by coalescence events, corresponding to genome replication foward in time. Any two species find their common ancestor at rate one, and at such an event the genes from the two merging species are placed back into the same species.

This informal description shows that the partition-valued process has two kinds of transitions: each pair of blocks merges at rate one; each gene is placed in a new uniformly chosen species at rate $(N - 1)d$. Setting the introgression rate to $d_N = d/N$ and letting $N \to \infty$, introgression events occur at rate d for each gene. At each such event the gene is sent to a new species that does not contain any of the other $n-1$ ancestral gene lineages, i.e., it is placed in a singleton block. This is the description of Kingman's coalescent with erosion, that we now more formally introduce.

3.1.2 Kingman's coalescent with erosion

Let $n \geq 1$, we define the n-Kingman coalescent with erosion as a Markov process $(\Pi^n_t)_{t\geq 0}$ taking values in the partitions of $[n] \coloneqq \{1, \dots, n\}$. Its transition rates are the following. Started from a partition π of $[n]$, the process jumps to any partition π' obtained by merging two blocks of π at rate one. Moreover, at rate d for each $i \leq n$, the integer i is "eroded". This means that if C is the block of π containing i, then the process jumps to the partition π' obtained by replacing the block C with the blocks $C \setminus \{i\}$ and $\{i\}$. (Obviously if $C = \{i\}$, i.e., if i is in a singleton block, no such transition can occur.)

Kingman's coalescent with erosion is a special case of the more general class of partition-valued processes called *exchangeable fragmentation-coalescence processes*, introduced and studied in [18]. These processes are a combination of the well-studied fragmentation processes, where blocks can only split, and coalescence processes, where blocks are only allowed to merge. The main new feature of combining fragmentation and coalescence is that they can balance each other so that fragmentation-coalescence processes display non-trivial stationary distributions. In this work we will be interested into describing the stationary distribution associated to Kingman's coalescent with erosion. The following proposition, which is a direct consequence of Theorem 8 of [18], provides the existence and uniqueness of this distribution.

Proposition 3.1 ([18]). *There exists a unique process $(\Pi_t)_{t\geq 0}$ valued in the partitions of $\mathbb{N}$ such that for all $n \geq 1$, the restriction of $(\Pi_t)_{t\geq 0}$ to $[n]$ is distributed as the n-Kingman coalescent with erosion. Moreover, the process $(\Pi_t)_{t\geq 0}$ has a unique stationary distribution Π.*

Kingman's coalescent with erosion is an exchangeable process in the sense that for any finite permutation σ of $\mathbb{N}$,

$$(\sigma(\Pi_t))_{t\geq 0} \overset{(\mathrm{d})}{=} (\Pi_t)_{t\geq 0}.$$

It is then clear that the stationary distribution Π is also an exchangeable partition of $\mathbb{N}$. Exchangeable partitions of $\mathbb{N}$ are often studied through what is known as their asymptotic frequencies. Let $\Pi = (C_1, C_2, \dots)$ be the blocks of the partition Π. Then, Kingman's representation theorem [127] shows that for any i, the following limit exists a.s.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{k\in C_i\}} = f_i.$$

Let $(\beta_i)_{i\geq 1}$ be the non-increasing reordering of the sequence $(f_i)_{i\geq 1}$. We call $(\beta_i)_{i\geq 1}$ the *asymptotic frequencies* of Π. The sequence $(\beta_i)_{i\geq 1}$ is such that

$$\beta_1 \geq \beta_2 \geq \dots \geq 0, \quad \sum_{i\geq 1} \beta_i \leq 1.$$

Such sequences are called *mass-partitions*. Mass-partitions are interesting because exchangeable partitions are entirely characterized by their asymptotic frequencies. The partition Π can be recovered from its asymptotic frequencies $(\beta_i)_{i\geq 1}$ through what is known as a *paintbox procedure*. Conditional on $(\beta_i)_{i\geq 1}$, let $(X_i)_{i\geq 1}$ be an independent sequence such that for $k \geq 1$, $\mathbb{P}(X_i = k) = \beta_k$, and $\mathbb{P}(X_i = -i) = 1 - \sum_{k\geq 1} \beta_k$. Then the partition Π' of $\mathbb{N}$ defined as

$$i \sim_{\Pi'} j \iff X_i = X_j$$

is distributed as Π [127]. We see that i is in a singleton block iff $X_i = -i$. The set of all singleton blocks is referred to as the *dust* of Π, and the partition Π has dust iff $\sum_{i\geq 1} \beta_i < 1$.

The main characteristics of the asymptotic frequencies of the stationary distribution of fragmentation-coalescence processes have already been derived in [18], see Theorem 8. In the case of Kingman's coalescent with erosion, these results specialize to the following theorem.

Theorem 3.2 ([18]). *Let $(\beta_i)_{i\geq 1}$ be the asymptotic frequencies of Π, the stationary distribution of Kingman's coalescent with erosion. Then*

$$\sum_{i\geq 1} \beta_i = 1, \quad \text{and} \quad \forall i \geq 1,\ \beta_i > 0 \quad \text{a.s.}$$

In other words, the partition Π has infinitely many blocks, and no dust.

Before stating our main two results, let us motivate them. Consider a partition $\hat{\Pi}$ obtained from a paintbox procedure on a random mass-partition $(\hat{\beta}_i)_{i\geq 1}$, and denote $\hat{\Pi}^n$ its restriction to $[n]$. There are two sources of randomness in $\hat{\Pi}^n$. One originates from the fact that $(\hat{\beta}_i)_{i\geq 1}$ is random. Moreover, conditional on $(\hat{\beta}_i)_{i\geq 1}$,

$\hat{\Pi}^n$ is obtained by sampling a finite number of variables with distribution $(\hat{\beta}_i)_{i\geq 1}$. Thus, in addition to the randomness of $(\hat{\beta}_i)_{i\geq 1}$, $\hat{\Pi}^n$ is subject to a finite sampling randomness.

Suppose that $\hat{\Pi}$ has finitely many blocks, say N, with asymptotic frequencies $(\hat{\beta}_1, \dots, \hat{\beta}_N)$. When n gets large, the finite sampling effects vanish and the sizes of the blocks of $\hat{\Pi}^n$ resemble $(n\hat{\beta}_1, \dots, n\hat{\beta}_N)$. However, when $\hat{\Pi}$ has infinitely many non-singleton blocks, there always exists a large enough i such that the size of the block with frequency $\hat{\beta}_i$ remains subject to finite sampling effects in $\hat{\Pi}^n$. In this case it is not entirely straightforward to go from the asymptotic frequencies $(\hat{\beta}_i)_{i\geq 1}$ to the size of the blocks of $\hat{\Pi}^n$, as this involves a non-trivial sampling procedure.

In this work our task will be twofold. First, we will investigate the size of the "large blocks" of Π^n by describing the distribution of the asymptotic frequencies $(\beta_i)_{i\geq 1}$. In order to get an insight into the distribution of the "small blocks" of Π^n, we will then also study the empirical distribution of the size of the blocks of Π^n, for large n. Let us now state the corresponding results.

3.1.3 Main results

We show two main results in this work. One is concerned with the size of the large blocks of the stationary distribution of Kingman's coalescent with erosion, and gives a representation of its asymptotic frequencies in terms of an infinite sequence of independent diffusions. The other is concerned with the size of the small blocks and provides the limit of the distribution of the size of a block chosen uniformly from the stationary partition when n is large. Let us start with the former result.

Size of the large blocks. Let $(Y_i)_{i\geq 1}$ be an i.i.d. sequence of diffusions verifying

$$\forall i \geq 1, \quad \mathrm{d}Y_i = (1 - Y_i)\,\mathrm{d}t + \sqrt{Y_i(1 - Y_i)}\,\mathrm{d}W_i,$$

started from 0, and where $(W_i)_{i\geq 1}$ are independent Brownian motions. Each Y_i is distributed as a one-dimensional Wright-Fisher diffusion with mutation, see for example [56], Lemma 4.1. It represents the dynamics of the frequency of a focal allele in a population with constant size, where the mutation rate from any other allele to that focal allele is one, and there are no back mutations, i.e., the mutation rate from the focal allele to any other allele is 0. Moreover, it is known that each Y_i is also distributed as a Wright-Fisher diffusion (without mutation) conditioned on hitting 1 [see for instance 135, Proposition 2.3.4]. Thus we have

$$\forall i \geq 1, \quad \lim_{t\to\infty} Y_i(t) = 1 \quad \text{a.s.}$$

Accordingly, we set $Y_i(\infty) = 1$. We build inductively a sequence of processes $(Z_i)_{i\geq 1}$ and time-changes $(\tau_i)_{i\geq 1}$ as follows. Set

$$\forall t \geq 0, \quad Z_1(t) = Y_1(t), \quad \tau_1(t) = \int_0^t \frac{1}{1 - Z_1(s)}\,\mathrm{d}s.$$

Then, suppose that $(Z_1, \dots, Z_i)$ and $(\tau_1, \dots, \tau_i)$ have been defined, and set

$$\forall t \geq 0, \quad Z_{i+1}(t) = (1 - Z_1(t) - \cdots - Z_i(t))Y_{i+1}(\tau_i(t)),$$
$$\forall t \geq 0, \quad \tau_{i+1}(t) = \int_0^t \frac{1}{1 - Z_1(s) - \cdots - Z_{i+1}(s)} \, \mathrm{d}s.$$

Then we have the following representation of the asymptotic frequencies of the stationary distribution of Kingman's coalescent with erosion.

Theorem 3.3. *Let $(Z_i)_{i\geq 1}$ be the sequence of diffusions defined previously. Then the non-increasing reordering of the sequence $(z_i)_{i\geq 1}$ defined as*

$$\forall i \geq 1, \quad z_i = \int_0^\infty d e^{-dt} Z_i(t) \, \mathrm{d}t,$$

is distributed as the frequencies of the blocks of the stationary distribution of Kingman's coalescent with erosion rate d.

Remark 3.4. Note that the previous result provides a coupling between the stationary distributions of Kingman's coalescent with erosion for various values of the erosion rate d. ○

Let us explain the intuition behind Theorem 3.3. Kingman's coalescent is dual to a measure-valued process called the Fleming-Viot process [55]. The Fleming-Viot process describes the family size distribution of a population with constant size, while Kingman's coalescent gives the genealogy of that population. By a classical duality argument, Kingman's coalescent at time t can be obtained by sampling individuals at time t from a Fleming-Viot process and placing in the same block those that have the same ancestor [21]. The link with Theorem 3.3 is that the diffusions $(Z_i)_{i\geq 1}$ correspond to the family sizes of the initial individuals of a Fleming-Viot process, ordered by extinction time of their descendance, see Section 3.5. The integral transformation is roughly due to the fact that in Kingman's coalescent with erosion, one needs to place in the same block the individuals that have the same ancestor at their last erosion event, which is an exponential variable with parameter d. This heuristic argument is made rigorous in Section 3.5, where Theorem 3.3 is proved.

Size of the small blocks. In order to capture the characteristics of the small blocks of Π^n, we study the empirical measure of the size of the blocks of Π^n. Let M^n be the total number of blocks of Π^n, and let $(|C_1^n|, \dots, |C_{M^n}^n|)$ be their sizes. For each $k \geq 1$, we denote

$$\mu_k^n = \frac{1}{M^n} \operatorname{Card}\Big(\{i : |C_i^n| = k\}\Big)$$

the frequency of blocks of size k. The probability vector $(\mu_k^n)_{k\geq 1}$ is the empirical measure of the size of the blocks of Π^n. We give the following characterization of the asymptotic law of $(\mu_k^n)_{k\geq 1}$ and M^n.

Theorem 3.5. (i) *The following convergence holds in probability*

$$\lim_{n\to\infty} \frac{M^n}{\sqrt{n}} = \sqrt{2d}.$$

(ii) *Moreover, for each $k \geq 1$, the following convergence holds in probability*

$$\lim_{n\to\infty} \mu_k^n = \frac{1}{2^{2k-1}} \frac{1}{k} \binom{2(k-1)}{k-1} = \mathbb{P}(J = k),$$

where J is half the return time to 0 of a simple symmetric random walk.

There is a natural interpretation of the random variable J involved in the previous proposition. Consider a Markov process on $\mathbb{N}$ starting from one that jumps from k to $k+1$ and from k to $k-1$ at rate k. It represents the size of a population where each individual gives birth and dies independently at rate one, and is called a critical binary branching process. Then the total progeny of this process, that is the total number of particles that have lived before the population goes extinct, is distributed as J. Actually, we will show the slightly stronger result that the genealogy of a block sampled uniformly from Kingman's coalescent with erosion is a critical binary branching process, see Remark 3.22.

Remark 3.6. It is interesting to notice that the limiting distribution of the vector $(\mu_k^n)_{k\geq 1}$ is deterministic and does not depend on the erosion coefficient d. ○

Remark 3.7. The convergence of the vector $(\mu_k^n)_{k\geq 1}$ is equivalent to the convergence in probability of the empirical measure of the size of the blocks of Π^n to the distribution of J in the weak topology. ○

Kingman's coalescent with immigration. The proof of Theorem 3.5 is based on the following heuristic. Erosion occurs at a rate proportional to the size of the blocks, i.e., a block of size k is eroded at rate dk, while coalescence events do not take the sizes of the blocks into account. As there are only few blocks with large size in Π^n, and many small blocks, most coalescence events occur between small blocks, while most erosion events occur within these few large blocks. When restricting our attention to small blocks, we can neglect erosion, and consider that pairs of blocks coalesce at rate one, and that new blocks of size one appear at constant rate due to the erosion of the large blocks.

This heuristic led us to consider a process analogous to Kingman's coalescent with erosion, where pairs of blocks coalesce at rate one, but new singleton blocks immigrate according to a Poisson process with rate d. We call this process Kingman's coalescent with immigration, see Section 3.2.1 for a rigorous definition. We will first prove that the genealogy of a block sampled uniformly from Kingman's coalescent with immigration converges, as the immigration rate goes to infinity, to a critical binary birth-death process, see forthcoming Proposition 3.20. Then, we will use this result and a coupling between Kingman's coalescents with immigration and erosion, described in Section 3.2.4, to prove Theorem 3.5.

The main focus of the present work is the stationary distribution of Kingman's coalescent with erosion. We only use Kingman's coalescent with immigration to obtain information about this distribution. However, we believe that Kingman's coalescent with immigration is an interesting object in its own right, which could describe the genealogy of entities sampled at distinct time points. In a population genetics interpretation, Kingman's coalescent models the genealogy of genes that are all sampled at the current time. In this case, a new particle that immigrates corresponds to a gene that has been sampled in the past. Such a multi-temporal sampling could occur for example in two situations: in viral phylodynamics [90, 215] and in macroevolution [201, 107]. Viral phylodynamics is a field of evolutionary biology that studies viral phylogenies and their interaction with various characteristics of the underlying epidemics. Viral sequences are often sampled at several timepoints, corresponding for example to different viral outbreaks. Macroevolution studies the evolutionary history of speciation, extinction. In this context, fossil data correspond to remainders of individuals that have lived and been sampled at some time point in the far past.

Outline. The remainder of the paper is organized as follows. In Section 3.2 we provide two constructions of Kingman's coalescent with immigration, as well as a coupling between Kingman's coalescents with erosion and immigration. Section 3.3 is then devoted to giving the genealogy of the blocks of Kingman's coalescent with immigration. The main result of this section is Proposition 3.15, which is the reformulation of Theorem 3.5 in the immigration case. In Section 3.4, we use Proposition 3.15 and the coupling between Kingman's coalescents with erosion and immigration to prove Theorem 3.5. Finally, we prove Theorem 3.3 in Section 3.5.

Possible extensions. As we have mentioned, Kingman's coalescent is part of the more general class of fragmentation-coalescence processes. We now briefly discuss potential extensions of our results to such processes.

The main ingredient of our study of the size of small blocks is that fragmentation is faster for larger blocks, while coalescence occurs at the same speed regardless of the size of the blocks. This allows us to neglect fragmentation and consider a purely coalescing system where new blocks immigrate due to the fragmentation of the large blocks. This picture remains valid for Λ-coalescents with erosion, but the proofs would be more involved because computations could no longer be made explictly. Morever, we believe that this picture also remains valid for a broad class of binary fragmentation measures. The particles that are removed from the large blocks would no longer be of size one, but should not have time to split on the time-scale when small blocks are formed, yielding a situation similar to the erosion case.

Theorem 3.3 relies on a construction of the stationary distribution of Kingman's coalescent with erosion from a Fleming-Viot process that can be directly generalized to Λ-coalescents with erosion (and even to Ξ-coalescents with erosion) by using the corresponding Λ-Fleming-Viot process. However, the explicit expression of the size of the blocks in terms of independent diffusions cannot be achieved in general. Nevertheless see the end of Section 3.5 for a discussion of a possible extension of Theorem 3.3 to Beta-coalescents with erosion.

Overall, the techniques and ideas we use in this work are not entirely specific to Kingman's coalescent with erosion. Nevertheless, in this case, the proofs are greatly simplified because all calculations can be made explicitly. This reason led us to restrict our attention to Kingman's coalescent with erosion in this work, and to leave possible extensions for future work.

3.2 Kingman's coalescent with immigration

In this section we construct Kingman's coalescent with immigration as a partition-valued process such that pairs of blocks coalesce at rate one and new blocks immigrate at rate d. Then, we give an alternative construction of Kingman's coalescent with erosion from the flow of bridges of [21]. Finally, the coupling between Kingman's coalescents with erosion and with immigration is carried out in Section 3.2.4.

3.2.1 Definition

Consider a Poisson point process on $\mathbb{R}$ with intensity $d\,\mathrm{d}t$, and let $(T_i)_{i\in\mathbb{Z}}$ be its atoms labeled in increasing order such that $T_0 < 0 < T_1$. The sequence $(T_i)_{i\in\mathbb{Z}}$ corresponds to the immigration times of new particles in the system.

Fix $N \in \mathbb{Z}$, we will first define Kingman's coalescent with immigration for the particles that have a label larger that N, and then extend it to all particles by consistency. We do that in the following way. Initially, set

$$\forall t < T_N, \quad \bar{\Pi}^N_t = \emptyset.$$

We now extend $\bar{\Pi}^N_t$ to all real times by induction. Suppose that $\bar{\Pi}^N_t$ has been defined on $(-\infty, T_k)$, for $k \geq N$. We first set

$$\bar{\Pi}^N_{T_k} = \bar{\Pi}^N_{T_k-} \cup \{k\}$$

to represent the immigration of the new particle with label k. We now let each pair of blocks of $\bar{\Pi}^N_t$ coalesce at rate one for $T_k \leq t < T_{k+1}$. One can achieve this by considering, conditional on $\left\{\bar{\Pi}^N_{T_k} = \bar{\pi}_k\right\}$, an independent version $(\Pi^k_t)_{t\geq 0}$ of Kingman's coalescent started from $\bar{\pi}_k$, and setting

$$\forall t < T_{k+1} - T_k, \quad \bar{\Pi}^N_{T_k+t} = \Pi^k_t.$$

We say that the process $(\bar{\Pi}^N_t)_{t\in\mathbb{R}}$ is the *N-Kingman coalescent with immigration rate d*. The following proposition shows that we can extend consistently the N-Kingman's coalescent with immigration to a process taking its values in the partitions of $\mathbb{Z}$.

Proposition 3.8. (i) *There exists a unique process $(\bar{\Pi}_t)_{t\in\mathbb{R}}$, called Kingman's coalescent with immigration rate d, such that for all $N \in \mathbb{Z}$, its restriction to $\{i \in \mathbb{Z} : i \geq N\}$ is distributed as the N-Kingman coalescent with immigration.*

(ii) *With probability one, $\bar{\Pi}_t$ has finitely many blocks for all $t \in \mathbb{R}$.*

Proof. (i) Let $(\bar{\Pi}^N_t)_{t\in\mathbb{R}}$ be a N-Kingman's coalescent with immigration. It is sufficient to show that the restriction $(\bar{\Pi}^{N+1}_t)_{t\in\mathbb{R}}$ of $(\bar{\Pi}^N_t)_{t\in\mathbb{R}}$ to $\{i \in \mathbb{Z} : i \geq N+1\}$ is distributed as a $N+1$-Kingman's coalescent with immigration, and the result will follow from Kolmogorov's extension theorem. Obviously, the immigration times of $(\bar{\Pi}^{N+1}_t)_{t\in\mathbb{R}}$ have the desired distribution. The result is now a simple consequence of the sampling consistency of Kingman's coalescent.

(ii) Let us now prove the second point. Kingman's coalescent has the property of coming down from infinity [128]. This means that even if Kingman's coalescent is started from a partition with an infinite number of blocks, then for all positive times it will have only finitely many blocks. Thus, as the number of immigrated particles is locally finite, Kingman's coalescent with immigration only has a finite number of blocks for all times a.s. □

In the remainder of this work we will make use of the process counting the number of blocks of Kingman's coalescent with immigration. More formally, for $t \in \mathbb{R}$, we define M_t as the (finite) number of blocks of $\bar{\Pi}_t$.

3.2.2 Preliminaries on flows of bridges

The previous construction of the Kingman coalescent with immigration is based on Kolmogorov's extension theorem. The aim of the next two sections is to give an alternative construction of Kingman's coalescent with immigration based on the flow of bridges of [21]. This construction will only be needed in Section 3.4 for the proof of Theorem 3.3. In this section we recall the material on flows of bridges that will be needed.

Bridges. We call a bridge [21] any random function of the form

$$\forall u \in [0,1], \quad B(u) = (1 - \sum_{i\geq 1} \beta_i)u + \sum_{i\geq 1} \beta_i \mathbb{1}_{\{u \geq V_i\}},$$

for some random mass-partition $(\beta_i)_{i\geq 1}$ and an independent i.i.d. sequence of uniform $[0,1]$ variables $(V_i)_{i\geq 1}$. For a bridge B, we define its inverse B^{-1} as

$$\forall u \in [0,1), \quad B^{-1}(u) = \inf\{t \in [0,1] : B(t) > u\}, \quad B^{-1}(1) = 1.$$

Let $(U_i)_{i\geq 1}$ be a sequence of i.i.d. uniform variables. An exchangeable partition $\hat{\Pi}$ of $\mathbb{N}$ can be obtained from B and $(U_i)_{i\geq 1}$ by setting

$$i \sim_{\hat{\Pi}} j \iff B^{-1}(U_i) = B^{-1}(U_j).$$

Let $(C_1, C_2, \dots)$ be the blocks of $\hat{\Pi}$ labeled in decreasing order of their least elements, i.e., such that

$$i \leq j \iff \min(C_i) \leq \min(C_j).$$

To each block C_i is associated a unique random variable V_i' defined as

$$\forall j \in C_i, \quad V_i' = B^{-1}(U_j).$$

If $\hat{\Pi}$ has finitely many blocks, say M, for $i > M$ we set $V_i' = \tilde{V}_i'$ where $(\tilde{V}_i')_{i \geq 1}$ is an independent sequence of i.i.d. uniform random variables. The sequence $(V_i')_{i \geq 1}$ will be referred to as the sequence of ancestors of the blocks of $\hat{\Pi}$. The key results on bridges from [21] is their Lemma 2 that we state here for later use.

Lemma 3.9 ([21])**.** *Consider a bridge B, and let $\hat{\Pi}$ and (V_i') be respectively the partition and sequence of ancestors obtained from B as above. Then $(V_i')_{i \geq 1}$ is independent of $\hat{\Pi}$, and $(V_i')_{i \geq 1}$ is a sequence of i.i.d. uniform variables.*

The standard flow of bridges. A flow of bridges is defined as follows.

Definition 3.10. A flow of bridges is a family of bridges $(B_{s,t})_{s \leq t}$ such that:

(i) For any $s \leq u \leq t$, we have $B_{s,u} \circ B_{u,t} = B_{s,t}$.

(ii) For $t_1 \leq \cdots \leq t_p$, the bridges $B_{t_1,t_2}, \dots, B_{t_{p-1},t_p}$ are independent, and B_{t_1,t_2} is distributed as B_{0,t_2-t_1}.

(iii) The limit $B_{0,t} \to \mathrm{Id}$ as $t \downarrow 0$ holds in probability in the Skorohod space. ○

A flow of bridges encodes the dynamics of a population represented by the interval $[0,1]$. Let $t \in \mathbb{R}$ and $x < y$. If the interval $[x,y]$ is interpreted as a subfamily of the population at time t, then its progeny at time $s \leq t$ is represented by the interval $[B_{s,t}(x-), B_{s,t}(y)]$. (Notice that time is going backward: if t is the present, then $s \leq t$ represents the future of the population.)

By the independence and stationarity of the increments of the flow, the distribution of a flow of bridges is entirely characterized by the distribution of $B_{0,t}$, for $t \geq 0$. We will be particularly interested in the so-called *standard flow of bridges*, that can be described as follows. Let $t \geq 0$ and consider the bridge

$$\forall u \in [0,1], \quad B_{0,t}(u) = \sum_{i=1}^{N_t} \beta_i \mathbb{1}_{\{V_i \leq u\}},$$

where

(i) The process $(N_t)_{t \geq 0}$ is distributed as a pure-death process started at ∞, and going from k to $k-1$ at rate $k(k-1)/2$.

(ii) Conditional on N_t, $(\beta_1, \dots, \beta_{N_t})$ has a Dirichlet distribution with parameter $(1, \dots, 1)$.

(iii) The variables $(V_i)_{i\geq 1}$ form an independent i.i.d. sequence of uniform variables.

Then we know [21] that there exists a flow of bridges $(B_{s,t})_{s\leq t}$ such that $B_{0,t}$ is distributed as above. It is called the standard flow of bridges.

Our interest in the standard flow of bridges is that is represents the dynamics of a population whose genealogy is given by Kingman's coalescent. Let $(U_i)_{i\geq 1}$ be a sequence of i.i.d. uniform variables, and let $\hat{\Pi}_t$ be the partition obtained from the bridge $B_{0,t}$ and the sequence $(U_i)_{i\geq 1}$. We stress that the *same* sequence is used for all t. Then the process $(\hat{\Pi}_t)_{t\geq 0}$ is distributed as Kingman's coalescent started from the partition of $\mathbb{N}$ into singletons [21].

The Fleming-Viot process. One of the main advantages of flows of bridges is that they couple a backward process, giving the genealogy of the population, and a forward process, giving the size of the progeny of the individuals in the population. This forward process is often encoded as a measure-valued process known as a Fleming-Viot process.

Let $(B_{s,t})_{s\leq t}$ be a standard flow of bridges. For each $t \geq 0$, $B_{-t,0}$ is the distribution function of some random measure ρ_t on $[0,1]$. The measure-valued process $(\rho_t)_{t\geq 0}$ is called a Fleming-Viot process [55]. A well-known fact that we will use is that the dynamics of the mass of n fixed disjoint intervals is distributed as the n first coordinates of a $(n+1)$-dimensional Wright-Fisher diffusion. More precisely, let $(I_1, \dots, I_n)$ be n disjoint intervals, and define

$$\forall i \in \{1, \dots, n\}, \forall t \geq 0, \quad X_i(t) = \rho_t(I_i)$$

and

$$\forall t \geq 0, \quad X_{n+1}(t) = 1 - (X_1(t) + \cdots + X_n(t)).$$

Then, if we denote by $(|I_1|, \dots, |I_n|)$ the lengths of the intervals $(I_1, \dots, I_n)$, the process $(X_1, \dots, X_{n+1})$ is distributed as the unique solution to

$$\forall i \in \{1, \dots, n+1\}, \quad \mathrm{d}X'_i = \sum_{\substack{j=1 \\ j\neq i}}^{n+1} \sqrt{X'_i X'_j}\, \mathrm{d}W'_{i,j},$$

started from $(|I_1|, \dots, |I_n|, 1 - |I_1| - \cdots - |I_n|)$, where $(W_{i,j})_{i<j}$ are independent Brownian motions and $W_{i,j} = -W_{j,i}$.

3.2.3 A flow of bridges construction of Kingman's coalescent with immigration

Let $(B_{s,t})_{s\leq t}$ be a standard flow of bridges. We now construct a version of Kingman's coalescent with immigration from $(B_{s,t})_{s\leq t}$. Consider a Poisson point process on $\mathbb{R} \times [0,1]$ with intensity $d\,\mathrm{d}t \otimes \mathrm{d}x$, and let $(T_i, U_i)_{i\in\mathbb{Z}}$ be its atoms, labeled in increasing order of their first coordinate such that $T_0 < 0 < T_1$. Similarly to Section 3.2.1, the times $(T_i)_{i\in\mathbb{Z}}$ correspond to immigration times of new particles. Here the sequence $(U_i)_{i\in\mathbb{Z}}$ represents the location in the population of these immigrated particles.

For each $t \in \mathbb{R}$, we define a partition $\bar{\Pi}_t$ of $\{i \in \mathbb{Z} : T_i \le t\}$ by setting

$$i \sim_{\bar{\Pi}_t} j \iff B^{-1}_{T_i,t}(U_i) = B^{-1}_{T_j,t}(U_j).$$

The following proposition shows that $(\bar{\Pi}_t)_{t\in\mathbb{R}}$ is distributed as Kingman's coalescent with immigration.

Proposition 3.11. *The process $(\bar{\Pi}_t)_{t\in\mathbb{R}}$ defined from the flow of bridges is a version of Kingman's coalescent with immigration rate d.*

Proof. The proof is almost identical to the proof of Corollary 1 of [21]. The main difference is that here the flow of bridges is sampled at various times $(T_i)_{i\in\mathbb{Z}}$ while for the classical Kingman coalescent, the flow of bridges is only sampled at an initial time.

We work conditional on $(T_i)_{i\in\mathbb{Z}}$ and consider these times as fixed. Let $(\bar{\Pi}^N_t)_{t\in\mathbb{R}}$ be the restriction of $(\bar{\Pi}_t)_{t\in\mathbb{R}}$ to $\{i \in \mathbb{Z} : i \ge N\}$. It is sufficient to show that for all $N \in \mathbb{Z}$ the blocks of $(\bar{\Pi}^N_t)_{t\in\mathbb{R}}$ coalesce according to independent versions of Kingman's coalescent between immigration times.

Let $t \in \mathbb{R}$, and let $(C_1, \dots, C_{M_t})$ be the blocks of $\bar{\Pi}^N_t$, where M_t is the number of blocks, and where the blocks are labeled such that

$$i \le j \iff \min(C_i) \le \min(C_j).$$

Similarly to Section 3.2.2, we can define the sequence of ancestors of $\bar{\Pi}^N_t$ by setting

$$\forall j \in C_i, \quad V'_i = B^{-1}_{T_j,t}(U_j),$$

and supplementing it with an independent sequence of i.i.d. uniform variables $(\tilde{V}'_i)_{i\ge1}$, i.e., defining $\forall i > M_t$, $V'_i = \tilde{V}'_i$.

Let us show by induction that for all $k \ge N$,

1. The ancestors $(V^{(k)}_i)_{i\ge1}$ of $\bar{\Pi}^N_{T_k}$ are i.i.d. with uniform distribution.
2. The sequence $(V^{(k)}_i)_{i\ge1}$ is independent of $(\bar{\Pi}^N_t)_{t\le T_k}$.
3. $(\bar{\Pi}^N_t)_{t\le T_k}$ is a version of the N-Kingman coalescent with immigration.

Fix $T_k \le t_1 < \cdots < t_{p+1} \le T_{k+1}$. By induction on p we can suppose that the sequence of ancestors of $\bar{\Pi}^N_{t_p}$, denoted by $(V^{(t_p)}_i)_{i\ge1}$, is independent of $\left((\bar{\Pi}^N_t)_{t\le T_k}, \bar{\Pi}^N_{t_1}, \dots, \bar{\Pi}^N_{t_p}\right)$. Then (i) and (ii) are proved if we can show that the sequence of ancestors of $\bar{\Pi}^N_{t_{p+1}}$ is independent of $\left((\bar{\Pi}^N_t)_{t\le T_k}, \bar{\Pi}^N_{t_1}, \dots, \bar{\Pi}^N_{t_{p+1}}\right)$, and is a sequence of i.i.d. uniform variables.

Let us now call Π^* the partition obtained from the bridge $B_{t_p,t_{p+1}}$ and the sequence $(V^{(t_p)}_i)_{i\ge1}$, i.e.,

$$i \sim_{\Pi^*} j \iff B^{-1}_{t_p,t_{p+1}}(V^{(t_p)}_i) = B^{-1}_{t_p,t_{p+1}}(V^{(t_p)}_j),$$

and let $(V_i^*)_{i\geq 1}$ be the sequence of ancestors of Π^*, i.e.,

$$\forall j \in C_i^*, \quad V_i^* = B^{-1}_{t_p,t_{p+1}}(V_j^{(t_p)}),$$

where $(C_1^*, C_2^*, \dots)$ denote the blocks of Π^* labeled in increasing order of their minimal elements as above. Using the fact that for $u \leq s \leq t$, $B^{-1}_{u,t} = B^{-1}_{s,t} \circ B^{-1}_{u,s}$, we get that for all $N \leq i, j \leq k$,

$$\begin{aligned} i \sim_{\bar{\Pi}_{t_{p+1}}} j &\iff B^{-1}_{t_p,t_{p+1}}(B^{-1}_{T_i,t_p}(U_i)) = B^{-1}_{t_p,t_{p+1}}(B^{-1}_{T_j,t_p}(U_j)) \\ &\iff B^{-1}_{t_p,t_{p+1}}(V^{(t_p)}_{b(i)}) = B^{-1}_{t_p,t_{p+1}}(V^{(t_p)}_{b(j)}) \\ &\iff b(i) \sim_{\Pi^*} b(j) \end{aligned} \tag{3.1}$$

where $b(i)$ denotes the label of the block of $\bar{\Pi}^N_{t_p}$ to which i belongs.

By independence of the increments of the flow of bridges, the bridge $B_{t_p,t_{p+1}}$ is independent of the collection of variables $\left((\bar{\Pi}^N_t)_{t\leq T_k}, \bar{\Pi}^N_{t_1}, \dots, \bar{\Pi}^N_{t_p}, (V_i^{(t_p)})_{i\geq 1}\right)$. Thus, $(B_{t_p,t_{p+1}}, (V_i^{(t_p)})_{i\geq 1})$ are independent of $\left((\bar{\Pi}^N_t)_{t\leq T_k}, \bar{\Pi}^N_{t_1}, \dots, \bar{\Pi}^N_{t_p}\right)$, and hence $(\Pi^*, (V_i^*)_{i\geq 1})$ are independent of $\left((\bar{\Pi}^N_t)_{t\leq T_k}, \bar{\Pi}^N_{t_1}, \dots, \bar{\Pi}^N_{t_p}\right)$. Using Lemma 3.9, we get that Π^* is independent of $(V_i^*)_{i\geq 1}$. This shows that $(V_i^*)_{i\geq 1}$ is independent of $\left((\bar{\Pi}^N_t)_{t\leq T_k}, \bar{\Pi}^N_{t_1}, \dots, \bar{\Pi}^N_{t_p}, \Pi^*\right)$. Using (3.1), we see that $\bar{\Pi}^N_{t_{p+1}}$ can be recovered from $\bar{\Pi}^N_{t_p}$ and Π^*. Thus, the variables $\left((\bar{\Pi}^N_t)_{t\leq T_k}, \bar{\Pi}^N_{t_1}, \dots, \bar{\Pi}^N_{t_{p+1}}\right)$ are independent of $(V_i^*)_{i\geq 1}$.

In order to end the proof of the claim we need to distinguish two cases. First, suppose that $t_{p+1} < T_{k+1}$. Then, due to our labeling convention, we have that $(V_i^*)_{i\geq 1} = (V_i^{(t_{p+1})})_{i\geq 1}$ (up to the auxiliary variables $(\tilde{V}_i)_{i\geq 1}$ that play no role). Conversely, if $t_{p+1} = T_{k+1}$, then one of the variables $(V_i^*)_{i\geq 1}$ has to be replaced by the ancestor U_{k+1} of the block $\{k+1\}$. More precisely, if $\bar{\Pi}^N_{T_{k+1}}$ has M_{k+1} blocks, again by labeling convention, the block $\{k+1\}$ has label M_{k+1}. Thus, $(V_i^{(t_{p+1})})_{i\geq 1}$ is recovered by setting $V_i^{(t_{p+1})} = V_i^*$ for $i \neq M_{k+1}$, and $V_i^{(t_{p+1})} = U_{k+1}$ for $i = M_{k+1}$. It is straightforward to see that, as U_{k+1} is independent of all other variables, $(V_i^{(t_{p+1})})_{i\geq 1}$ remains a sequence of i.i.d. of uniform variables, independent of $\left((\bar{\Pi}^N_t)_{t\leq T_k}, \bar{\Pi}^N_{t_1}, \dots, \bar{\Pi}^N_{t_{p+1}}\right)$ and thus the fact that points (i) and (ii) of the claim hold.

For $k \geq N$ and $t < T_{k+1} - T_k$ consider the partition Π^k_t of $\mathbb{N}$ defined as

$$i \sim_{\Pi^k_t} j \iff B^{-1}_{T_k,T_k+t}(V_i^{(k)}) = B^{-1}_{T_k,T_k+t}(V_j^{(k)})$$

As the sequence $(V_i^{(k)})_{i\geq 1}$ is i.i.d. uniform and independent of $(\bar{\Pi}^N_t)_{t\leq T_k}$, the process $(\Pi^k_t)_{t\geq 0}$ is a version of Kingman's coalescent started from the partition into singletons, independent of $(\Pi^k_t)_{t\geq 0}$. Using equation (3.1), we have that

$$i \sim_{\bar{\Pi}^N_{T_k+t}} j \iff b(i) \sim_{\Pi^k_t} b(j),$$

where $b(i)$ denotes the label of the block of $\bar{\Pi}^N_{T_k}$ to which i belongs. In other words, $(\bar{\Pi}^N_{T_k+t})_{0\leq t<T_{k+1}-T_k}$ is obtained by letting the blocks of $\bar{\Pi}^N_{T_k}$ coalesce according to

an independent version of Kingman's coalescent. This proves that $(\bar{\Pi}_t^N)_{t \le T_{k+1}}$ is distributed as a N-Kingman coalescent with immigration, and ends the proof of the result. □

3.2.4 Coupling erosion and immigration

We now explain the coupling between Kingman's coalescents with erosion and with immigration. Let $n \ge 1$, consider a Poisson point process P^n on $\mathbb{R}$ with intensity $nd\,\mathrm{d}t$ and let $(T_i)_{i \in \mathbb{Z}}$ be its atoms ordered increasingly such that $T_0 < 0 < T_1$. To each atom of the process we attach a uniform mark in $[n]$. We denote by ℓ_i the mark attached to T_i, so that $(\ell_i)_{i \in \mathbb{Z}}$ is a sequence of i.i.d. uniform variables on $[n]$.

Consider $t \in \mathbb{R}$. For each $k \in [n]$, let $\varphi_t(k)$ be the label of the last atom of P^n with mark k before time t, i.e., $\varphi_t(k) \in \mathbb{Z}$ is the unique i such that $\ell_i = k$ and there is no atom T of P^n with $T_i < T \le t$ carrying mark k. Let $(\bar{\Pi}_t)_{t \in \mathbb{R}}$ be Kingman's coalescent with immigration rate nd built from the Poisson process $(T_i)_{i \in \mathbb{Z}}$ as in Section 3.2.1. We define a partition Π_t^n of $[n]$ by setting

$$i \sim_{\Pi_t^n} j \iff \varphi_t(i) \sim_{\bar{\Pi}_t} \varphi_t(j).$$

In words, i and j belong to the same block of Π_t^n iff the most recently immigrated particles of $(\bar{\Pi}_t)_{t \in \mathbb{R}}$ with marks i and j have coalesced before time t. The key point of this construction is that $(\Pi_t^n)_{t \in \mathbb{R}}$ is distributed as Kingman's coalescent with erosion.

Proposition 3.12. *The process $(\Pi_t^n)_{t \in \mathbb{R}}$ is a stationary version of the n-Kingman coalescent with erosion rate d.*

Proof. Let $k \in [n]$. By thinning, the set of atoms of P^n with mark k is a Poisson process on $\mathbb{R}$ with intensity $d\,\mathrm{d}t$, and these processes are independent. Thus new atoms of P^n with mark k arrive at rate d. Let us consider what happens at such an arrival time. Suppose that $\ell_i = k$. Then, by definition, we have $\varphi_{T_i}(k) = i$, as the atom T_i has mark k. Moreover, the particle i is a singleton of the partition $\bar{\Pi}_{T_i}$ (it is the particle that has newly immigrated). Thus at time T_i, the integer k is removed from its block and placed in a singleton block. This is the description of an erosion event, which occur at rate d.

Let us now describe the dynamics between two immigration times, say T_i and T_{i+1}. Conditional on $\bar{\Pi}_{T_i}$, the blocks of $(\bar{\Pi}_t)_{T_i \le t \le T_{i+1}}$ coalesce according to an independent version of Kingman's coalescent started from $\bar{\Pi}_{T_i}$. The labels of the atoms of P^n that are the last atoms with their marks form a subset of $\{j \in \mathbb{Z} : j \le i\}$, say L. By sampling consistency of Kingman's coalescent, the restriction of $(\bar{\Pi}_t)_{T_i \le t \le T_{i+1}}$ to L is also distributed as Kingman's coalescent, starting from the restriction of $\bar{\Pi}_{T_i}$ to L. Thus, as the blocks of $(\Pi_t^n)_{T_i \le t \le T_{i+1}}$ are, up to an independent relabeling, the blocks of the restriction of $(\bar{\Pi}_t)_{T_i \le t \le T_{i+1}}$ to L, any two pairs of blocks of $(\Pi_t)_{T_i \le t \le T_{i+1}}$ coalesce at rate one.

The fact that $(\Pi_t)_{t \in \mathbb{R}}$ is stationary follows from the stationarity of the Poisson point process. □

Combined with the construction of Kingman's coalescent with immigration from the standard flow of bridges, this coupling gives an interesting construction of the stationary distribution of Kingman's coalescent with erosion.

Corollary 3.13. *Let $(B_{s,t})_{s\leq t}$ be a standard flow of bridges, $(T_i)_{i\geq 1}$ be an independent sequence of i.i.d. exponential variables with parameter d, and $(U_i)_{i\geq 1}$ be an independent sequence of i.i.d. uniform variables. Then the partition Π defined by*

$$i \sim_\Pi j \iff B^{-1}_{-T_i,0}(U_i) = B^{-1}_{-T_j,0}(U_j)$$

has the stationary distribution of Kingman's coalescent with erosion rate d.

Proof. Consider a Poisson process P^n on $\mathbb{R} \times [0,1]$ with intensity $nd\,\mathrm{d}t \otimes \mathrm{d}x$, and attach to each atom of P^n a uniform mark on $[n]$. If (T_i, U_i) denotes the last atom of P^n with mark i before $t = 0$, then T_i is exponentially distributed with parameter d, U_i is uniform on $[0,1]$, and all these variables are independent. A combination of Proposition 3.12 and Proposition 3.11 now proves the result. □

Remark 3.14. The construction of Kingman's coalescent with immigration from Section 3.2.1 and the construction with the flow of bridges of Section 3.2.3 only rely on the sampling consistency of Kingman's coalescent. These constructions could be extended directly to a case where the coalescence events occur according to a Λ-coalescent [178, 188]. In particular, the construction of the stationary distribution of Kingman's coalescent with erosion of Corollary 3.13 extends directly to Λ-coalescents with erosion if one replaces the standard flow of bridges by the corresponding Λ-flow of bridges. ∘

3.3 Size of the blocks of Kingman's coalescent with immigration

In this section we study Kingman's coalescent with immigration. The main result we will show is the following.

Proposition 3.15. *Let $n \geq 1$ and consider $(\bar{\Pi}^n_t)_{t\in\mathbb{R}}$ a version of Kingman's coalescent with immigration rate nd. Let $(|\bar{C}^n_1|, \dots, |\bar{C}^n_p|)$ be the size of p blocks chosen uniformly from $\bar{\Pi}^n_0$, then*

$$(|\bar{C}^n_1|, \dots, |\bar{C}^n_p|) \Longrightarrow (J_1, \dots, J_p)$$

where $(J_1, \dots, J_p)$ are i.i.d. variables distributed as the total progeny of a critical binary branching process.

We prove this result by choosing k blocks uniformly from $\bar{\Pi}^n_0$, and counting backwards in time the number of blocks that are ancestors of these blocks, i.e., that will further coalesce to form these blocks. We show that this process converges, under appropriate scaling, to k independent critical binary branching processes, yielding the result. In this section we work in both directions of time. We will index time by t when it is flowing forward, and by s when it is flowing backwards.

We first give a precise definition of the ancestral process counting the number of blocks in Section 3.3.1, along with its basic properties. The convergence is then carried out in Section 3.3.2.

3.3.1 The ancestral process

Let $(\bar{\Pi}_t)_{t\in\mathbb{R}}$ be a version of Kingman's coalescent with immigration rate d. The process $(\bar{\Pi}_t)_{t\in\mathbb{R}}$ is naturally endowed with a notion of ancestry between its blocks. For $t \in \mathbb{R}$, let M_t be the number of blocks of $\bar{\Pi}_t$. Let $(\bar{C}_1, \dots, \bar{C}_{M_t})$ be an enumeration of the blocks of $\bar{\Pi}_t$. We say that this enumeration is exchangeable if conditional on $\{M_t = k\}$, for any permutation σ of $[k]$,

$$(\bar{C}_1, \dots, \bar{C}_k) \overset{\text{(d)}}{=} (\bar{C}_{\sigma(1)}, \dots, \bar{C}_{\sigma(k)}).$$

We can always consider an exchangeable enumeration of the blocks of $\bar{\Pi}_t$ by changing the labels of any enumeration according to an independent uniform permutation.

For $u \leq t$, consider $\bar{\Pi}_t = (\bar{C}_1, \dots, \bar{C}_{M_t})$ and $\bar{\Pi}_u = (\bar{C}'_1, \dots, \bar{C}'_{M_s})$ an enumeration of the blocks of $\bar{\Pi}_t$ and $\bar{\Pi}_u$ respectively. In Kingman's coalescent with immigration, a block present at time u can only coalesce with other blocks. Thus, for any block $\bar{C}'_i$, there is a unique block $\bar{C}_j$ of $\bar{\Pi}_t$ such that $\bar{C}'_i \subseteq \bar{C}_j$. We say that $\bar{C}'_i$ is an ancestor of $\bar{C}_j$. We define the ancestral process of Kingman's coalescent with immigration as the vector counting the number of ancestors of the blocks of $\bar{\Pi}_0$, enumerated in an exchangeable way. This definition is illustrated in Figure 3.2.

Definition 3.16. Let $(\bar{\Pi}_t)_{t\in\mathbb{R}}$ be Kingman's coalescent with immigration, and let $(\bar{C}_1, \dots, \bar{C}_{M_0})$ be the blocks of $\bar{\Pi}_0$ enumerated in an exchangeable order. For $s \geq 0$, let $(\bar{C}'_1, \dots, \bar{C}'_{M_{-s}})$ be the blocks of $\bar{\Pi}_{-s}$. We define the number of ancestors of the i-th block as

$$\mathcal{A}_s(i) = \begin{cases} \mathrm{Card}\Big\{ j \leq M_{-s} : \bar{C}'_j \subseteq \bar{C}_i \Big\} & \text{if } i \in \{1, \dots, M_0\} \\ 0 & \text{if } i > M_0. \end{cases}$$

The process $(\mathcal{A}_s)_{s\geq 0}$ defined as $\mathcal{A}_s \coloneqq (\mathcal{A}_s(1), \mathcal{A}_s(2), \dots)$ is called the ancestral process associated to $(\bar{\Pi}_t)_{t\in\mathbb{R}}$. ○

The process $(\mathcal{A}_s)_{s\geq 0}$ can be seen as a particle system where at time 0, there are M_0 particles with distinct types, and $(\mathcal{A}_s(i))_{s\geq 0}$ records the number of particles with type i. As we have reversed time, each coalescence event now corresponds to the birth of a new particle, and each immigration event to the death of a particle.

Note that relative to the original population model described in the introduction, we have now reversed the time twice. As Kingman's coalescents with erosion and immigration represent genealogies, the future of these processes corresponds

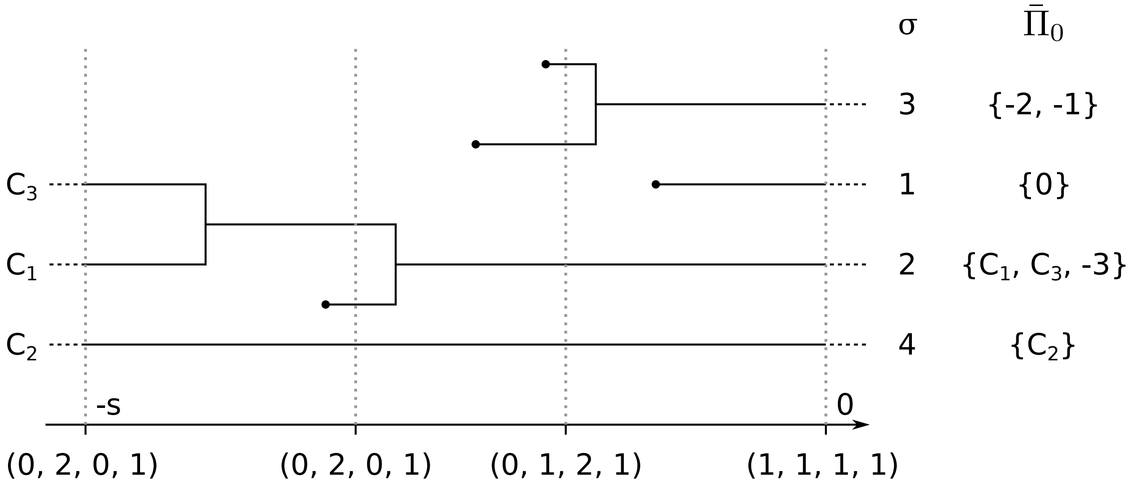


Figure 3.2: In this example, we have $\bar{\Pi}_{-s} = (C_1, C_2, C_3)$. Each black circle represents an immigration event, and the lines merge at the coalescence time of the blocks to which they correspond. At $s = 0$ the blocks of $\bar{\Pi}_0$ are labeled according to the permutation σ, and the value of $(\mathcal{A}_s)_{s\geq 0}$ is given below for some times.

to the past of the population. Therefore, the "ancestors" of the blocks of Kingman's coalescent with immigration actually correspond to the descendants of these individuals in the population point of view.

Recall that $(M_t)_{t\in\mathbb{R}}$ stands for the number of blocks of $(\bar{\Pi}_t)_{t\in\mathbb{R}}$ forward in time. For each $s \in \mathbb{R}$, we define $N_s := M_{-s}$, the number of blocks of $(\bar{\Pi}_t)_{t\in\mathbb{R}}$ backwards in time. The process $(N_s)_{s\geq 0}$ also gives the number of particles of the ancestral process $(\mathcal{A}_s)_{s\geq 0}$, that is we have

$$\forall s \geq 0, \quad N_s = \sum_{i\geq 1} \mathcal{A}_s(i).$$

The following proposition shows that the ancestral process is Markovian. This is a key feature that makes Kingman's coalescent with immigration easier to study than Kingman's coalescent with erosion.

Proposition 3.17. *Let $(\mathcal{A}_s)_{s\geq 0}$ be the ancestral process associated to Kingman's coalescent with immigration rate d, and let $(N_s)_{s\geq 0}$ be the number of particles of $(\mathcal{A}_s)_{s\geq 0}$. Then $(\mathcal{A}_s)_{s\geq 0}$ is a Markov process with initial condition*

$$\forall i \leq N_0,\ \mathcal{A}_0(i) = 1, \quad \forall i > N_0,\ \mathcal{A}_0(i) = 0.$$

Moreover, conditional on $\mathcal{A}_s$:

- *each particle gives birth to a new particle of its type at rate d/N_s.*
- *each particle dies at rate $(N_s - 1)/2$.*

The proof of Proposition 3.17 can be found in Section 3.A, we only sketch it here. We will first show that the process $(M_t)_{t\in\mathbb{R}}$ is a stationary birth-death process, such that conditional on $M_t = k$, a birth occurs at rate d, and a death at rate $k(k-1)/2$. A simple calculation shows that it is actually a reversible process, i.e., with our notation, that $(N_s)_{s\geq 0}$ is distributed as $(M_t)_{t\geq 0}$. When $(N_s)_{s\geq 0}$ jumps from k to $k+1$, a particle has given birth to two particles. By exchangeability of our system, the particle that gives birth is chosen uniformly, i.e., each particle gives birth at the same rate d/k. Similarly, when $(N_s)_{s\geq 0}$ jumps from k to $k-1$ a particle chosen uniformly from the population dies. Thus each particle dies at rate $k(k-1)/(2k) = (k-1)/2$.

Making the above argument rigorous involves counting the number of trajectories of $(\bar{\Pi}_t)_{t\in\mathbb{R}}$ yielding a given trajectory of $(\mathcal{A}_s)_{s\geq 0}$. We postpone it until Section 3.A.

In order to prove Proposition 3.15, we need to keep track of the number of ancestors of k blocks chosen uniformly from $\bar{\Pi}_0$. As we have chosen a uniform labeling of the blocks of $\bar{\Pi}_0$, this amounts to considering the process $(\mathcal{A}_s(1), \dots, \mathcal{A}_s(k);\ s \geq 0)$. Proposition 3.17 directly gives us the distribution of this process.

Corollary 3.18. *The process $(\mathcal{A}_s(1), \dots, \mathcal{A}_s(p), N_s;\ s \geq 0)$ is a Markov process such that conditional on $\{\mathcal{A}_s(1) = a_1, \dots, \mathcal{A}_s(p) = a_p, N_s = k\}$, the process jumps to:*

- $(a_1, \dots, a_i + 1, \dots, a_p, k+1)$ *at rate* $\frac{d}{k} a_i$.
- $(a_1, \dots, a_i - 1, \dots, a_p, k-1)$ *at rate* $\frac{k-1}{2} a_i$.
- $(a_1, \dots, a_p, k+1)$ *at rate* $\frac{d}{k}(k - a_1 - \cdots - a_p)$.
- $(a_1, \dots, a_p, k-1)$ *at rate* $\frac{k-1}{2}(k - a_1 - \cdots - a_p)$.

Proof. We see from the expression of the transition rates of $(\mathcal{A}_s)_{s\geq 0}$ that the rate at which each particle splits or dies only depends on the rest of the population through the total population size N_s. This is enough to prove the result. □

3.3.2 Convergence

We now prove that the process $(\mathcal{A}_s(1), \dots, \mathcal{A}_s(p);\ s \geq 0)$ converges to independent critical binary birth-death processes when time is rescaled by a factor $1/\sqrt{n}$. We start with the following lemma.

Lemma 3.19. *Let M^n have the stationary distribution of $(M^n_t)_{t\geq 0}$, the number of blocks of Kingman's coalescent with immigration rate dn. The sequence $(M^n/\sqrt{n};\ n \geq 1)$ is tight.*

Proof. Let $n \geq 1$ and consider a birth-death process $(X^n_t)_{t\geq 0}$ such that conditional on $\{X^n_t = k\}$, the process jumps to

- $k+1$ at rate dn;
- $k-1$ at rate μ_k,

where the death rate μ_k is defined as

$$\mu_k = \begin{cases} 0 & \text{if } k < \sqrt{2dn}+1, \\ \frac{(\sqrt{2dn}+1)\sqrt{2dn}}{2} & \text{else.} \end{cases}$$

The process $(X_t^n - \lfloor\sqrt{2dn}+1\rfloor;\, t \geq 0)$ is distributed as a simple random walk, reflected at 0. Thus it admits a geometric stationary distribution with parameter γ_n given by

$$\gamma_n = \frac{2dn}{(\sqrt{2dn}+1)\sqrt{2dn}} = \frac{1}{1+\sqrt{\frac{1}{2dn}}}.$$

This shows that the process $(X_t^n)_{t\geq 0}$ also admits a stationary distribution. If X^n has the stationary distribution of $(X_t^n)_{t\geq 0}$, then X^n is distributed as $\lfloor\sqrt{2dn}\rfloor + 1 + Y^n$, where Y^n has a geometric distribution with parameter γ_n.

Hence, for K and n large enough, we have

$$\begin{aligned} \mathbb{P}\Big(X^n \leq K\sqrt{n}\Big) &\leq \mathbb{P}\Big(Y^n \leq K\sqrt{n} - \sqrt{2dn}\Big) \\ &= 1 - \gamma_n^{(K-\sqrt{2d})\sqrt{n}} \\ &= 1 - \exp\Big(-\frac{K-\sqrt{2d}}{\sqrt{2d}}\Big) + o_n(1). \end{aligned}$$

Thus the sequence $(X^n/\sqrt{n};\, n \geq 1)$ is tight.

Recall that $(M_t^n)_{t\geq 0}$ is a birth-death process jumping from k to $k+1$ at rate dn, and from k to $k-1$ at rate $k(k-1)/2 \geq \mu_k$. Its stationary distribution is thus dominated by that of X^n, and this proves the result. □

We now prove our main convergence result. The proof will use a result from Chapter 11 of [58] on the a.s. convergence of rescaled Markov processes. In order to stick to their notation, we introduce

$$\forall s \geq 0, \quad \hat{N}_s^n = N_{s/\sqrt{n}}^n, \quad \hat{\mathcal{A}}_s^n = \mathcal{A}_{s/\sqrt{n}}^n,$$

and

$$\forall x \geq 0, \quad \beta_+(x) = d, \quad \beta_-(x) = \frac{x^2}{2}, \quad F(x) = d - \frac{x^2}{2}.$$

Proposition 3.20. *Let $(\mathcal{A}_s^n)_{s\geq 0}$ be the ancestral process of Kingman's coalescent with immigration rate dn. Then*

$$\Big(\hat{\mathcal{A}}_s^n(1), \dots, \hat{\mathcal{A}}_s^n(p), \frac{\hat{N}_s^n}{\sqrt{n}};\, s \geq 0\Big) \Longrightarrow \Big(X_1(s), \dots, X_p(s), \sqrt{2d};\, s \geq 0\Big),$$

in the sense of convergence in distribution in the Skorohod space, and where the processes $(X_1, \dots, X_p)$ are i.i.d. critical binary birth-death processes, with per-capita birth and death rate $\sqrt{d/2}$.

Proof. We start by showing that the process $(\hat{N}^n_s/\sqrt{n};\ s \geq 0)$ converges to the constant process with value $\sqrt{2d}$. By applying Proposition 3.17 (bearing in mind that in Proposition 3.17 the immigration rate is d, and not dn) the process $(\hat{N}^n_s)_{s\geq 0}$ is a Markov process jumping from

- k to $k+1$ at rate $d\sqrt{n} = \sqrt{n}\beta_+(\frac{k}{\sqrt{n}})$.
- k to $k-1$ at rate $\frac{k(k-1)}{2\sqrt{n}} = \sqrt{n}\beta_-(\frac{k}{\sqrt{n}}) - \frac{1}{2\sqrt{n}}$.

Thus, the process $(\hat{N}^n_s)_{s\geq 0}$ is of the same form as the processes considered in Theorem 2.1 of Chapter 11 of [58], except that the scaling is $\sqrt{n}$ and not n.

Let us consider a stationary version of the process $(\hat{N}^n_s)_{s\geq 0}$. Lemma 3.19 shows that the sequence $(\hat{N}^n_0/\sqrt{n};\ n \geq 1)$ is tight. We can thus find an increasing sequence of indices $(n_k)_{k\geq 1}$ such that the subsequence $(\hat{N}^{n_k}_0/\sqrt{n_k};\ k \geq 1)$ converges in distribution to a limiting variable N. Using Skorohod's representation theorem [see e.g. 25, Theorem 6.7], we can assume that the convergence holds a.s.

Applying Theorem 2.1 of Chapter 11 of [58] shows that the sequence of processes $(\hat{N}^{n_k}_s/\sqrt{n_k};\ s \geq 0,\ k \geq 1)$ converges a.s. uniformly on compact sets to the solution of

$$\dot{x} = F(x) = d - \frac{x^2}{2}, \tag{3.2}$$

started from the random variable N. (The original theorem is given for a different scaling, but the proof is easily adapted to ours.) As each process $(\hat{N}^{n_k}_s)_{s\geq 0}$ is stationary, the limiting process is a stationary solution to (3.2), i.e., it is the constant process with value $\sqrt{2d}$. This shows that each converging subsequence of $(\hat{N}^n_s/\sqrt{n};\ s \geq 0,\ n \geq 1)$ converges to the same constant process, and thus that the entire sequence converges.

Let us now prove the convergence of the ancestral processes. Consider independent Poisson processes $(P^-_i(s))_{s\geq 0}$, $(P^+_i(s))_{s\geq 0}$ for $i \leq p$, and $(P^-_N(s))_{s\geq 0}$, $(P^+_N(s))_{s\geq 0}$. Using e.g. Theorem 4.1 from Chapter 6 of [58], there exists a unique strong solution to the following equation

$$\begin{aligned}
\forall s \geq 0, \forall i \leq p, \quad & X^n_i(s) = P^+_i\Big(\int_0^s \frac{d\sqrt{n}X^n_i(u)}{Y^n(u)}\,\mathrm{d}u\Big) - P^-_i\Big(\int_0^s \frac{X^n_i(u)(Y^n(u)-1)}{2\sqrt{n}}\,\mathrm{d}u\Big),\\
\forall s \geq 0, \forall i \leq p, \quad & Y^n(s) = P^+_N\Big(\int_0^s d\sqrt{n}\Big(1 - \tfrac{\sum_i X^n_i(u)}{Y^n(u)}\Big)\,\mathrm{d}u\Big)\\
& \qquad - P^-_N\Big(\int_0^t \tfrac{Y^n(u)(Y^n(u)-1)}{2\sqrt{n}}\Big(1 - \tfrac{\sum_i X^n_i(u)}{Y^n(u)}\Big)\,\mathrm{d}u\Big) + \sum_{i=1}^p X^n_i(s).
\end{aligned}$$

Moreover, the solution $(X^n_1,\dots,X^n_p,Y^n)$ to the previous equation has the same distribution as $(\hat{\mathcal{A}}^n_s(1),\dots,\hat{\mathcal{A}}^n_s(p),\hat{N}^n_s;\ s \geq 0)$.

As $Y^n/\sqrt{n}$ converges in probability to the constant process with value $\sqrt{2d}$, we can find a subsequence such that

$$\lim_{n\to\infty} \frac{d\sqrt{n}}{Y^n(s)} = \sqrt{\frac{d}{2}}, \quad \lim_{n\to\infty} \frac{(Y^n(s)-1)}{2\sqrt{n}} = \sqrt{\frac{d}{2}} \quad \text{a.s.}$$

holds uniformly in s on compact sets. This is sufficient to show that for each $i \leq p$, the subsequence of processes $(X_i^n(s))_{s\geq 0}$ converges a.s. in the Skorohod space to the solution $(X_i(s))_{s\geq 0}$ of

$$\forall s \geq 0,\, \forall i \leq p, \quad X_i(s) = P_i^+\Big(\int_0^s \sqrt{\frac{d}{2}} X_i(u)\,\mathrm{d}u\Big) - P_i^-\Big(\int_0^s \sqrt{\frac{d}{2}} X_i(u)\mathrm{d}u\Big).$$

This proves that the entire sequence $(X_1^n, \dots, X_p^n)$ converges in probability in the Skorohod topology to the solution of the previous equation. Finally, noting that the solutions of these equations are independent and distributed as critical binary branching processes with branching rate $\sqrt{d/2}$ ends the proof. □

We are now ready to prove Proposition 3.15.

Proof of Proposition 3.15. By construction, the size of p blocks of $\bar{\Pi}^n$ chosen uniformly is given by the total number of particles of $(\hat{\mathcal{A}}_s^n(1), \dots, \hat{\mathcal{A}}_s^n(p);\, t \geq 0)$. Thus, in the limit, the size of these blocks converges to the total size of p independent critical binary branching processes. □

3.4 Proof of Theorem 3.5

In the previous section we have derived the limiting distribution of the sizes of blocks uniformly sampled from Kingman's coalescent with immigration. In this section we make use of the coupling between Kingman's coalescent with immigration and Kingman's coalescent with erosion from Section 3.2.4 to get the analogous result in the erosion case.

We first show the following result.

Corollary 3.21. *Let Π^n have the stationary distribution of the n-Kingman coalescent with erosion. Let $(|C_1^n|, \dots, |C_p^n|)$ be the size of p blocks chosen uniformly from Π^n. Then*

$$(|C_1^n|, \dots, |C_p^n|) \Longrightarrow (J_1, \dots, J_p),$$

where $(J_1, \dots, J_p)$ are i.i.d. variables distributed as the total progeny of a critical binary branching process.

Proof. Recall the coupling between Kingman's coalescent with erosion and Kingman's coalescent with immigration. Let $(T_i)_{i\in\mathbb{Z}}$ be the atoms of a Poisson point process P^n with intensity dn, labeled in increasing order such that $T_0 < 0 < T_1$. Consider an independent i.i.d. sequence of marks $(\ell_i)_{i\in\mathbb{Z}}$ that are uniformly distributed on $[n]$.

Let $\bar{\Pi}_0^n$ be the value at time 0 of the version of Kingman's coalescent with immigration rate nd built from $(T_i)_{i\in\mathbb{Z}}$ as in Section 3.2.1. We know from Proposition 3.12 that we can obtain a version Π^n of the stationary distribution of the n-Kingman coalescent with erosion rate d by placing i and j in the same block of Π^n if the most recent atoms of P^n in $(-\infty, 0]$ with mark i and j both belong to the same block of $\bar{\Pi}_0^n$.

Now let $(\bar{C}_1^n, \dots, \bar{C}_p^n)$ be p blocks chosen uniformly from $\bar{\Pi}_0$, and denote by $(|\bar{C}_1^n|, \dots, |\bar{C}_p^n|)$ their respective sizes. For $k \leq p$, let

$$|C_k^n| = \mathrm{Card}\Big\{i \in \bar{C}_k^n : (T_i, \ell_i) \text{ is the most recent atom in } (-\infty, 0] \text{ with mark } \ell_i\Big\}.$$

Then conditional on $\Big\{|C_1^n| \geq 1, \dots, |C_p^n| \geq 1\Big\}$, $(|C_1^n|, \dots, |C_p^n|)$ are the sizes of p blocks chosen uniformly from Π^n. The result is thus proved if we can show that

$$\lim_{n\to\infty} \mathbb{P}\Big(|C_1^n| = |\bar{C}_1^n|, \dots, |C_p^n| = |\bar{C}_p^n|\Big) = 1.$$

Let us first explain intuitively why the previous claim holds. The ancestors of $\bar{C}_1^n$ have all immigrated on a time-scale of order $1/\sqrt{n}$. On this time-scale, there are of order $\sqrt{n}$ particles that have also immigrated. All these particles receive a uniform label in $[n]$. Thus the probability that an ancestor of $\bar{C}_1^n$ has received the same label as one of the other $\sqrt{n}$ particles, i.e., that it is not the most recent atom with its mark, is of order $1/\sqrt{n}$. Let us make this argument rigorous.

Set

$$\tau_1^n := \min\Big\{T_i : i \in \bar{C}_1^n\Big\}$$

to be the total life-time of the ancestors of the block $\bar{C}_1^n$. (The variable τ_1^n gives the immigration time of the first particle that forms the block $\bar{C}_1^n$.) The total number of particles that have immigrated during the time interval $[\tau_1^n, 0]$ is then $P^n([\tau_1^n, 0])$. Consider the event

$$E_k = \Big\{|\bar{C}_1^n| = k,\ \tau_1^n \in [-\tfrac{s}{\sqrt{n}}, 0],\ P^n([-\tfrac{s}{\sqrt{n}}, 0]) \leq (1+\varepsilon) ds\sqrt{n}\Big\}.$$

On this event, if $|C_1^n| \neq |\bar{C}_1^n|$, then one of the k ancestors of $\bar{C}_1^n$ has received the same label as one of the particles that has immigrated in the time interval $[\tau_1^n, 0]$, that is, the same label as one of the $(1+\varepsilon)dt\sqrt{n}$ most recent atoms of P^n. As the labels are chosen uniformly, the probability that each of the k ancestors has a label distinct from the labels of the other $(1+\varepsilon)ds\sqrt{n} - 1$ most recent particles is

$$\Big(1 - \frac{1}{n}\Big) \dots \Big(1 - \frac{k-1}{n}\Big)\Big(1 - \frac{k}{n}\Big)^{(1+\varepsilon)ds\sqrt{n} - k}$$

which goes to 1 as n goes to infinity for all fixed k. Thus

$$\mathbb{P}\Big(|C_1^n| \neq |\bar{C}_1^n|, E_k\Big) \leq \Big(1 - \frac{1}{n}\Big) \dots \Big(1 - \frac{k-1}{n}\Big)\Big(1 - \frac{k}{n}\Big)^{(1+\varepsilon)ds\sqrt{n} - k},$$

and

$$\begin{aligned}\mathbb{P}\Big(|C_1^n| \neq |\bar{C}_1^n|\Big) \leq \mathbb{P}\Big(\tau_1^n \notin [-\tfrac{s}{\sqrt{n}}, 0]\Big) + \mathbb{P}\Big(|\bar{C}_1^n| \geq K\Big)& \\ + \mathbb{P}\Big(P^n([-\tfrac{s}{\sqrt{n}}, 0]) > (1+\varepsilon)ds\sqrt{n}\Big) + o_n(1).&\end{aligned} \tag{3.3}$$

Now, by Proposition 3.20, the sequence $(-\sqrt{n}\tau_1^n)_{n\geq 1}$ converges in distribution to the total life-time of a binary critical branching process and $(|\bar{C}_1^n|)_{n\geq 1}$ converges to the total progeny of this process. Thus, the first two terms in the above equation can be made as small as desired uniformly in n by taking t and K large enough. For a fixed $\varepsilon > 0$, Chebishev's inequality shows that the last term goes to 0 as n goes to infinity. This proves the result for $p = 1$ and a simple union bound proves the result for any p. □

Remark 3.22. In the previous proof, on the event $\Big\{|\bar{C}_1^n| = |C_1^n|\Big\}$, not only the size of the blocks of Kingman's coalescents with erosion and immigration coincide, but also the genealogy of the blocks. Thus we have shown the slightly stronger result that, in the n-Kingman coalescent with erosion, the genealogy of a block chosen uniformly from the stationary distribution converges to that of a critical binary branching process. ○

We can now prove Theorem 3.5. Recall that μ_k^n denotes the frequency of blocks of size k of Π^n, i.e., if the blocks of Π^n are $(C_1^n, \dots, C_{M^n}^n)$, then

$$\mu_k^n = \frac{1}{M^n} \operatorname{Card}(\{i : |C_i^n| = k\}).$$

Proof of Theorem 3.5. (i) We start by proving that $M^n/\sqrt{n}$ converges to $\sqrt{2d}$ in probability. Let us consider a version $\bar{\Pi}^n$ of the stationary distribution of Kingman's coalescent with immigration rate nd, coupled with a version Π^n of the stationary distribution of Kingman's coalescent with erosion rate d on $[n]$. Let $\bar{M}^n$, resp. M^n, denote the number of blocks of $\bar{\Pi}^n$, resp. Π^n. Recall that the blocks of Π^n are subsets of the blocks of $\bar{\Pi}^n$, where a particle is retained if there are no other particles with the same label that have immigrated after it. Let $|\bar{C}^n|$ be the size of a block of $\bar{\Pi}^n$ chosen uniformly, and let $|C^n|$ be the size of the corresponding block of Π^n. Some blocks of $\bar{\Pi}^n$ are only composed of particles that are not retained to form Π^n. Such blocks have no corresponding blocks in Π^n, and $\bar{M}^n - M^n$ is exactly the number of such blocks. Thus

$$\mathbb{E}\Big[\frac{\bar{M}^n - M^n}{\bar{M}^n}\Big] = \mathbb{P}(|C^n| = 0) \le \mathbb{P}\Big(|C^n| \neq |\bar{C}^n|\Big) \longrightarrow 0,$$

where the convergence holds by (3.3). This shows that $M^n/\bar{M}^n$ goes to 1 in probability. Proposition 3.20 further shows that $\bar{M}^n/\sqrt{n}$ goes to $\sqrt{2d}$ in probability, and thus that $M^n/\sqrt{n}$ also goes to $\sqrt{2d}$ in probability.

(ii) We prove the second point using the method of moments. Let $(|C_1^n|, \dots, |C_p^n|)$ be the sizes of k uniformly sampled blocks of Π^n. Then, as the number of blocks M^n goes to infinity, Corollary 3.21 shows that

$$\lim_{n\to\infty} \mathbb{E}[(\mu_k^n)^p] = \lim_{n\to\infty} \mathbb{P}\Big(|C_1^n| = \dots = |C_p^n| = k\Big) = \mathbb{P}(J = k)^p,$$

where J is the total progeny of a binary critical branching process. The convergence of the moments readily implies convergence in distribution as the limit is a Dirac mass. □

3.5 Asymptotic frequencies of Kingman's coalescent with erosion

In this section we prove Theorem 3.3, which gives a representation of the asymptotic frequencies in terms of independent diffusions. First, we use the flow of bridges construction of Kingman's coalescent with erosion from Corollary 3.13 to give a correspondence between the frequencies of the blocks and the size of the families of a Fleming-Viot process.

3.5.1 Eves of a Fleming-Viot process

Let $(\rho_t)_{t\geq 0}$ be a Fleming-Viot process built from a standard flow of bridges as in Section 3.2.2. For each individual $x \in [0,1]$, denote by

$$\zeta(x) = \inf\{t \geq 0 : \rho_t(\{x\}) = 0\}$$

the extinction time of the offspring of x. It is clear that the set

$$\{x \in [0,1] : \zeta(x) > 0\} = \{x \in [0,1] : \rho_t(\{x\}) > 0 \text{ for some } t \geq 0\}$$

is countable. The elements of this set can actually be enumerated in decreasing order of their extinction time, that is, they can be written $(\mathbf{e}_i)_{i\geq 0}$ with

$$\zeta(\mathbf{e}_1) > \zeta(\mathbf{e}_2) > \dots$$

This fact can be found e.g. in [133], Theorem 1.6. The sequence $(\mathbf{e}_i)_{i\geq 0}$ is called the sequence of *Eves* of $(\rho_t)_{t\geq 0}$, and was introduced in [21] and [133], see also [50] for a similar notion for Continuous-State Branching Processes. The following result shows that the frequencies of the blocks of the stationary distribution of Kingman's coalescent with erosion can be recovered from the size of the offspring of the Eves.

Lemma 3.23. *Let $(\mathbf{e}_i)_{i\geq 1}$ be the Eves of a Fleming-Viot process $(\rho_t)_{t\geq 0}$. Then the non-increasing reordering of the sequence $(z_i)_{i\geq 1}$ defined as*

$$\forall i \geq 1, \quad z_i = \int_0^\infty de^{-dt} \rho_t(\{\mathbf{e}_i\})\, \mathrm{d}t$$

is distributed as the frequencies of the blocks of the stationary distribution of Kingman's coalescent with erosion rate d.

Proof. Consider a flow of bridges $(B_{s,t})_{s\leq t}$, and let $(T_i)_{i\geq 1}$, $(U_i)_{i\geq 1}$ be two independent i.i.d. sequences of exponential variables with parameter d, and uniform variables respectively. Again, as in Corollary 3.13, let Π be the partition of $\mathbb{N}$ defined as

$$i \sim_\Pi j \iff B^{-1}_{-T_i,0}(U_i) = B^{-1}_{-T_j,0}(U_j),$$

which has the stationary distribution of Kingman's coalescent with erosion. We denote by $\Pi = (C_1, C_2, \dots)$ the blocks of Π, ordered in increasing order of their least elements, i.e., such that

$$i \leq j \iff \min(C_i) \leq \min(C_j).$$

Then let us call

$$A_i = B^{-1}_{-T_j,0}(U_j), \ \forall j \in C_i,$$

the ancestor of the block C_i.

As the flow of bridges $(B_{s,t})_{s \leq t}$ is independent of the sequences $(U_i)_{i \geq 1}$ and $(T_i)_{i \geq 1}$, the sequence $(B^{-1}_{-T_i,0}(U_i))_{i \geq 1}$ is exchangeable. Thus, the law of large numbers shows that for any $i \geq 1$,

$$\frac{1}{n} \operatorname{Card}(C_i \cap [n]) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\left\{B^{-1}_{-T_j,0}(U_j) = A_i\right\}} \underset{n \to \infty}{\longrightarrow} \int_0^\infty d e^{-dt} \rho_t(\{A_i\}) \, \mathrm{d}t \quad \text{a.s.}$$

Thus the result is proved if we can show that a.s.

$$\{\mathbf{e}_i : i \geq 1\} = \{A_i : i \geq 1\}.$$

Clearly we have $\zeta(A_i) > 0$, as otherwise the frequency of the block C_i would be zero. Moreover, conditional on the flow of bridges, there exists a.s. some $j \geq 1$ such that

$$(U_j, T_j) \in \Big\{(x,t) : B^{-1}_{-t,0}(x) = \mathbf{e}_i\Big\}$$

as by definition of $\mathbf{e}_i$ this set has positive Lebesgue measure. Thus, a.s. $\mathbf{e}_i$ is the ancestor of some block of Π, and the result is proved. □

In order to prove Theorem 3.3, it remains to show that the sequence of processes $\big(\rho_t(\{\mathbf{e}_1\}), \rho_t(\{\mathbf{e}_2\}), \dots; t \geq 0\big)$ has the same distribution as the sequence of diffusions introduced in Section 3.1.3. In the following section we characterize this distribution, and complete the proof in the last section.

3.5.2 Wright-Fisher diffusion conditioned on its extinction order

Consider a n-dimensional Wright-Fisher diffusion $(X_1, \dots, X_n)$. That is, the collection of processes $(X_1, \dots, X_n)$ is distributed as the unique solution to

$$\forall i \geq 1, \quad \mathrm{d}X_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} \sqrt{X_i X_j} \, \mathrm{d}W_{i,j},$$

where $(W_{i,j})_{i<j}$ are independent Brownian motions, and $W_{j,i} = -W_{i,j}$, and started from an initial condition $(x_1, \dots, x_n) \in (0,1)^n$ verifying $x_1 + \dots + x_n = 1$. The Wright-Fisher diffusion describes the dynamics of a population with constant size, where individuals can be of n different types; X_i denotes the frequency of type i individuals in the population. Each process X_i is eventually absorbed at 0 or 1. We say that the family X_i reaches fixation if it gets absorbed at 1, and that it becomes extinct otherwise. Let

$$\zeta_i = \inf\{t \geq 0 : X_i = 0\}$$

denote its absorption time at 0.

In this section, we study the distribution of $(X_1, \dots, X_n)$ conditional on the event $\{\zeta_n < \dots < \zeta_1\}$. First, notice that as $X_1 + \dots + X_n = 1$, there is exactly one family that reaches fixation. Thus, on the event $\{\zeta_n < \dots < \zeta_1\}$, we have $\zeta_1 = \infty$ and X_1 reaches fixation; X_2 is the last family to go extinct, and X_n is the first family to go extinct. We now express the distribution of the conditioned Wright-Fisher diffusion in terms of the diffusions introduced in Section 3.1.3.

We will work inductively, by first conditioning the process $(X_1, \dots, X_n)$ on ζ_1 being the largest extinction time, then on ζ_2 being the second largest and so on and so forth. The key point is that after conditioning on the fixation of X_1, the remainder of the population, $(X_2, \dots, X_n)$, is distributed as a rescaled, time-changed, unconditioned $(n-1)$-dimensional Wright-Fisher diffusion, independent of X_1.

Let us be more specific and let Y_1 be the solution of

$$\mathrm{d}Y_1 = (1 - Y_1)\,\mathrm{d}t + \sqrt{Y_1(1 - Y_1)}\,\mathrm{d}W_1, \tag{3.4}$$

for some Brownian motion W_1. Notice that Y_1 is distributed as a usual one-dimensional Wright-Fisher diffusion, conditioned on fixation. Consider the fixation time of Y_1 which is defined as

$$S_1 = \inf\{t \geq 0 : Y_1(t) = 1\}.$$

We further define a random time-change τ_1 as

$$\forall t < S_1,\ \tau_1(t) = \int_0^t \frac{1}{1 - Y_1(s)}\,\mathrm{d}s, \quad \forall t \geq S_1,\ \tau_1(t) = \infty.$$

We start by proving the following result.

Lemma 3.24. *Let Y_1 and τ_1 be as above and consider an independent $(n-1)$-dimensional Wright-Fisher diffusion $(X_2, \dots, X_n)$. Then, the process $(Z_1, \dots, Z_n)$ defined as*

$$Z_1 = Y_1, \quad \forall i > 1, \forall t \geq 0,\ Z_i(t) = (1 - Z_1(t))X_i(\tau_1(t)),$$

is distributed as a n-dimensional Wright-Fisher diffusion conditioned on $\{\zeta_1 = \infty\}$.

Remark 3.25. The time $\tau_1(t)$ is infinite with positive probability. However, each of the processes $(X_2, \dots, X_n)$ has an a.s. limit as t goes to infinity. On the event $\{\tau_1(t) = \infty\}$, we take $X_i(\tau_1(t))$ to be this limit, so that the process $(Z_1, \dots, Z_n)$ is now well-defined. ○

Before proving Lemma 3.24, we need the following fact that we prove for the sake of completeness.

Lemma 3.26. *Let $(W_t)_{t \geq 0}$ be a Brownian motion on $\mathbb{R}$ started at 1, and let T_0 be the first time it hits 0. Then for $\alpha \in \mathbb{R}$, a.s.*

$$\int_0^{T_0} W_s^\alpha\,\mathrm{d}s = \begin{cases} \infty & \text{if } \alpha \leq -2 \\ y_\alpha < \infty & \text{if } \alpha > -2. \end{cases}$$

Proof. Let us define

$$\forall t \geq 0, \quad \xi_t = \tilde{W}_t - \frac{t}{2}, \quad \tau(t) = \inf\Big\{ s \geq 0 : \int_0^s \exp(2\xi_u)\, \mathrm{d}u > t \Big\},$$

for a Brownian motion $(\tilde{W}_t)_{t\geq 0}$ with the convention that $\inf \emptyset = \infty$ and $\xi_\infty = -\infty$. The Lamperti representation of positive self-similar processes [143] shows that W_t stopped at T_0 satisfies the equality in distribution

$$(W_{t\wedge T_0})_{t\geq 0} \overset{\text{(d)}}{=} (\exp(\xi_{\tau(t)}))_{t\geq 0}.$$

Thus

$$\int_0^{t\wedge T_0} W_s^\alpha \,\mathrm{d}s \overset{\text{(d)}}{=} \int_0^t \exp(\alpha \xi_{\tau(s)})\,\mathrm{d}s = \int_0^{\tau(t)} \exp((2+\alpha)\xi_s)\,\mathrm{d}s,$$

and

$$\int_0^{T_0} W_s^\alpha \,\mathrm{d}s \overset{\text{(d)}}{=} \int_0^{\infty} \exp((2+\alpha)\xi_s)\,\mathrm{d}s,$$

which yields the result. □

Proof of Lemma 3.24. Consider a n-dimensional Wright-Fisher diffusion $(X_1, \dots, X_n)$. A calculation of Doob's h-transform using the harmonic function

$$h(x_1, \dots, x_n) = \mathbb{P}\Big(\lim_{t\to\infty} X_1(t) = 1 \,\Big|\, X_1(0) = x_1, \dots, X_n(0) = x_n \Big) = x_1$$

shows that the process $(X_1, \dots, X_n)$ conditioned on $\{\lim_{t\to\infty} X_1(t) = 1\} = \{\zeta_1 = \infty\}$ is distributed as the unique solution to the equation

$$\mathrm{d}X_1 = (1 - X_1)\,\mathrm{d}t + \sum_{j=2}^n \sqrt{X_1 X_j}\,\mathrm{d}W_{1,j},$$

$$\forall i \geq 2, \quad \mathrm{d}X_i = -X_i\,\mathrm{d}t + \sum_{\substack{j=1\\ j\neq i}}^n \sqrt{X_i X_j}\,\mathrm{d}W_{i,j},$$

where $(W_{i,j})_{i<j}$ are independent Brownian motions, and $W_{i,j} = -W_{j,i}$. We will prove that the process $(Z_1, \dots, Z_n)$ solves this equation.

Now consider a $(n-1)$-dimensional Wright-Fisher diffusion $(X'_2, \dots, X'_n)$ independent of Y_1 which solves

$$\forall i \geq 2, \quad \mathrm{d}X'_i = \sum_{\substack{j=2\\ j\neq i}}^n \sqrt{X'_i X'_j}\,\mathrm{d}W'_{i,j},$$

where $(W'_{i,j})_{i<j}$ are independent Brownian motions and $W'_{i,j} = -W'_{j,i}$. We start by giving the equation solved by the process $(Y_1, X'_2 \circ \tau_1, \dots, X'_n \circ \tau_1)$. Notice that here, only a subset of the processes are time-changed, and that τ_1 explodes in finite time. For these two reasons, let us realize the time-change carefully.

We transform τ_1 into a family of finite stopping times. Our first task is to prove that τ_1 goes continuously to infinity, we do this using the speed function and scale measures of the diffusion Y_1, see e.g. [56]. If we define $D = 1/Y_1$, then by Itô's formula,

$$\mathrm{d}D = -\sqrt{D-1}D\,\mathrm{d}W_1, \quad \forall t \geq 0,\ [D,D]_t = \int_0^t (D(s)-1)D(s)^2\,\mathrm{d}s.$$

Recall that S_1 stands for the first time when Y_1 hits one. Using Dubins-Schwarz theorem, see for instance Theorem 18.4 of [120], we obtain that

$$\begin{aligned}\int_0^{S_1} \frac{1}{1-Y_1(s)}\,\mathrm{d}s = \int_0^{S_1} \frac{D(s)}{D(s)-1}\,\mathrm{d}s \\ = \int_0^{S_1} \frac{\tilde{W}_1([D,D]_s)}{\tilde{W}_1([D,D]_s)-1}\,\mathrm{d}s = \int_0^{T_1} \frac{1}{(\tilde{W}_1(s)-1)^2\tilde{W}_1(s)}\,\mathrm{d}s\end{aligned}$$

where $\tilde{W}_1$ is a Brownian motion (on a possibly larger probability space) started at $1/Y_1(0)$, and T_1 is the first time when $\tilde{W}_1$ hits 1. We now know from Lemma 3.26 that this integral is a.s. infinite, and thus that τ_1 goes continuously to infinity, and does not "jump to infinity".

Further consider the times

$$\forall i \geq 2,\ S_i = \inf\{t \geq 0 : X'_i(t) = 1\}, \quad S = \min(S_2, \dots, S_n).$$

At time S, one of the families has reached fixation, and thus for $t \geq S$ we have $X'_i(t) = X'_i(S)$. Therefore, for all $t \geq 0$, we have $X'_i(\tau_1(t)) = X'_i(\tau_1(t) \wedge S)$, where the stopping time $\tau_1(t) \wedge S$ is now a.s. finite, and $t \mapsto \tau_1(t) \wedge S$ is continuous. (The continuity requires that τ_1 does not jump to infinity.) Thus, by making a time-change in the following integrals, see e.g. [120], Theorem 17.24, we obtain

$$\begin{aligned}\forall t \geq 0, \quad X'_i(\tau_1(t)) &= X'_i(\tau_1(t) \wedge S) \\ &= \sum_{\substack{j=2 \\ j\neq i}}^n \int_0^{\tau_1(t)\wedge S} \sqrt{X'_i(s)X'_j(s)}\,\mathrm{d}W'_{i,j} \\ &= \sum_{\substack{j=2 \\ j\neq i}}^n \int_0^t \sqrt{X'_i(\tau_1(s)\wedge S)X'_j(\tau_1(s)\wedge S)}\,\mathrm{d}W'_{i,j}(\tau_1(s)\wedge S) \\ &= \sum_{\substack{j=2 \\ j\neq i}}^n \int_0^t \sqrt{\frac{X'_i(\tau_1(s))X'_j(\tau_1(s))}{1-Y_1(s)}}\,\mathrm{d}\tilde{W}_{i,j}\end{aligned}$$

where

$$\forall t \geq 0, \quad \tilde{W}_{i,j}(t) = \int_0^t \sqrt{1-Y_1(s)}\,\mathrm{d}W'_{i,j}(\tau_1(s)\wedge S).$$

A direct computation of the quadratic variations gives

$$\forall i,j,t \geq 0, \quad [\tilde{W}_{i,j}, \tilde{W}_{i,j}]_t = t \wedge S,$$

and the crossed variations are null. Thus a multidimensional version of Dubins-Schwarz theorem, see e.g. Theorem 18.4 in [120], shows that we can find independent Brownian motions $(\hat{W}_{i,j})_{i<j}$ such that $\tilde{W}_{i,j}(t) = \hat{W}_{i,j}(t \wedge S)$. This proves that the time-changed processes solve

$$\forall t \geq 0, \quad X'_i(\tau_1(t)) = \sum_{\substack{j=2 \\ j \neq i}}^{n} \int_0^t \sqrt{\frac{X'_i(\tau_1(s)) X'_j(\tau_1(s))}{1 - Y_1(s)}} \, \mathrm{d}\hat{W}_{i,j}.$$

Finally, setting $\hat{X}_i \coloneqq X'_i \circ \tau_1$ and applying Itô's formula to the process $(Y_1, \hat{X}_2, \dots, \hat{X}_n)$ with the function

$$(x_1, \dots, x_n) \mapsto (x_1, (1-x_1)x_2, \dots, (1-x_1)x_n)$$

we obtain that for all $i \geq 2$,

$$\begin{aligned} \mathrm{d}Z_i &= -\hat{X}_i \, \mathrm{d}Y_1 + (1 - Y_1) \, \mathrm{d}\hat{X}_i \\ &= -\hat{X}_i (1 - Y_1) \, \mathrm{d}t - \hat{X}_i \sqrt{Y_1(1 - Y_1)} \, \mathrm{d}W_1 + \sum_{\substack{j=2 \\ j \neq i}}^{n} \sqrt{(1 - Y_1) \hat{X}_i \hat{X}_j} \, \mathrm{d}\hat{W}_{i,j} \\ &= -Z_i \, \mathrm{d}t - Z_i \sqrt{\frac{Z_1}{1 - Z_1}} \, \mathrm{d}W_1 + \sum_{\substack{j=2 \\ j \neq i}}^{n} \sqrt{\frac{Z_i Z_j}{1 - Z_1}} \, \mathrm{d}\hat{W}_{i,j}, \end{aligned}$$

where $(Z_1, \dots, Z_n)$ is defined as in the statement of the result. A straightforward computation of the quadratic variations shows that $(Z_1, \dots, Z_n)$ is distributed as $(X_1, \dots, X_n)$ conditioned on $\{\zeta_1 = \infty\}$ and proves the result. □

We can now proceed inductively. Let us set up the notation for the proof. Consider i.i.d. processes $(Y_1, \dots, Y_{n-1})$ such that

$$\forall i \geq 1, \quad \mathrm{d}Y_i = (1 - Y_i) \, \mathrm{d}t + \sqrt{Y_i(1 - Y_i)} \, \mathrm{d}W_i$$

where $(W_1, \dots, W_{n-1})$ are independent Brownian motions. We set $\tilde{Z}_1 = Y_1$, and

$$\forall t \geq 0, \quad \tilde{\tau}_1(t) = \int_0^t \frac{1}{1 - \tilde{Z}_1(s)} \, \mathrm{d}s.$$

We then define recursively, for $i < n-1$,

$$\begin{aligned} \forall t \geq 0, \quad \tilde{Z}_{i+1}(t) &= (1 - \tilde{Z}_1(t) - \cdots - \tilde{Z}_i(t)) Y_{i+1}(\tilde{\tau}_i(t)) \\ \forall t \geq 0, \quad \tilde{\tau}_{i+1}(t) &= \int_0^t \frac{1}{1 - \tilde{Z}_1(s) - \cdots - \tilde{Z}_{i+1}(s)} \, \mathrm{d}s. \end{aligned}$$

We finally set $\tilde{Z}_n = 1 - \tilde{Z}_1 - \cdots - \tilde{Z}_{n-1}$.

Proposition 3.27. *The collection of process $(\tilde{Z}_1, \dots, \tilde{Z}_n)$ defined above is distributed as a n-dimensional Wright-Fisher diffusion conditioned on $\{\zeta_n < \cdots < \zeta_1\}$.*

Proof. We prove the result inductively. For $n = 2$, conditioning (X_1, X_2) on its extinction order amounts to conditioning it on the fixation of X_1, and Lemma 3.24 shows that the result holds.

Let $(Y_1, \dots, Y_{n-1})$ be the i.i.d. diffusions defined above. We first define

$$\forall t \geq 0,\ \tilde{Z}'_2(t) = Y_2(t), \quad \forall t \geq 0,\ \tilde{\tau}'_2(t) = \int_0^t \frac{1}{1 - \tilde{Z}'_2(s)}\, \mathrm{d}s$$

and then define inductively, for $i < n-1$,

$$\begin{aligned} \forall t \geq 0, \quad & \tilde{Z}'_{i+1}(t) = (1 - \tilde{Z}'_2(t) - \dots - \tilde{Z}'_i(t)) Y_{i+1}(\tilde{\tau}'_i(t)), \\ \forall t \geq 0, \quad & \tilde{\tau}'_{i+1}(t) = \int_0^t \frac{1}{1 - \tilde{Z}'_2(s) - \dots - \tilde{Z}'_{i+1}(s)}\, \mathrm{d}s, \end{aligned}$$

and $\tilde{Z}'_n = 1 - \tilde{Z}'_2 - \dots - \tilde{Z}'_{n-1}$. By induction, we can suppose that $(\tilde{Z}'_2, \dots, \tilde{Z}'_n)$ is distributed as a $(n-1)$-dimensional Wright-Fisher diffusion conditioned on its extinction order. We first claim that the process defined as

$$\begin{aligned} \forall t \geq 0, \quad & \tilde{Z}_1(t) = Y_1(t), \\ \forall i > 1, \forall t \geq 0, \quad & \tilde{Z}_i(t) = (1 - \tilde{Z}_1(t)) \tilde{Z}'_i(\tilde{\tau}_1(t)) \end{aligned}$$

is distributed as a n-dimensional Wright-Fisher diffusion conditioned on its extinction order.

To see this, let $(X_2, \dots, X_n)$ be a $(n-1)$-dimensional unconditioned Wright-Fisher diffusion, independent of Y_1, and recall the definition of $(Z_1, \dots, Z_n)$ from Lemma 3.24. Consider

$$\zeta'_i = \inf\{t \geq 0 : Z_i(t) = 0\}, \quad \zeta_i = \inf\{t \geq 0 : X_i(t) = 0\}$$

the extinction times of Z_i and X_i. Lemma 3.24 ensures that $(Z_1, \dots, Z_n)$ is distributed as a Wright-Fisher diffusion conditioned on the fixation of Z_1. Thus, the process $(Z_1, \dots, Z_n)$ further conditioned on $\{\zeta'_n < \dots < \zeta'_2\}$ has the distribution of a Wright-Fisher diffusion conditioned on its extinction order. Now notice that

$$\{\zeta'_n < \dots < \zeta'_2\} = \{\zeta_n < \dots < \zeta_2\}.$$

Thus conditioning $(Z_1, \dots, Z_n)$ on $\{\zeta'_n < \dots \zeta'_2\}$ amounts to conditioning $(X_2, \dots, X_n)$ on $\{\zeta_n < \dots < \zeta_2\}$, that is, conditioning it on its fixation order. As $\{\zeta_n < \dots < \zeta_2\}$ is independent of Z_1, conditioning the process $(Z_1, \dots, Z_n)$ on this event is equivalent to replacing $(X_2, \dots, X_n)$ by $(\tilde{Z}'_2, \dots, \tilde{Z}'_n)$ in the construction of $(Z_1, \dots, Z_n)$, and this proves the claim.

It only remains to show that $\tilde{Z}_{i+1}$ as defined in the proof can be written

$$\forall i > 1, \quad \tilde{Z}_{i+1}(t) = (1 - \tilde{Z}_1(t) - \dots - \tilde{Z}_i(t)) Y_{i+1}(\tilde{\tau}_i(t)).$$

A direct calculation first shows that, for $i > 1$ and $t \geq 0$,

$$\begin{aligned}
\tilde{\tau}_i(t) &= \int_0^t \frac{1}{1-\tilde{Z}_1(s)-\cdots-\tilde{Z}_i(s)}\,\mathrm{d}s \\
&= \int_0^t \frac{1}{1-\tilde{Z}_1(s)-(1-\tilde{Z}_1(s))\tilde{Z}_2'(\tilde{\tau}_1(s))-\cdots-(1-\tilde{Z}_1(s))\tilde{Z}_i'(\tilde{\tau}_1(s))}\,\mathrm{d}s \\
&= \int_0^t \frac{1}{1-\tilde{Z}_2'(\tilde{\tau}_1(s))-\cdots-\tilde{Z}_i'(\tilde{\tau}_1(s))}\;\frac{1}{1-\tilde{Z}_1(s)}\,\mathrm{d}s \\
&= \tilde{\tau}_i'(\tilde{\tau}_1(t)),
\end{aligned}$$

and the result follows. □

We end this section by pointing out the following fact that will be required in the next section. We have only defined the Wright-Fisher diffusion conditioned on its extinction order for an initial condition $(x_1,\dots,x_n)$ such that for all $1 \leq i \leq n$, $x_i > 0$. Nevertheless, the processes Y_i have an entrance boundary at 0. Thus there exists a unique extension of the process $(Y_1,\dots,Y_{n-1})$ started from $(0,\dots,0)$ that remains Feller, see e.g. [120], Theorem 23.3. This shows that a Wright-Fisher diffusion conditioned on its fixation order $(\tilde{Z}_1,\dots,\tilde{Z}_n)$ admits a Feller extension for the initial condition $(0,\dots,0,1)$.

3.5.3 Proof of Theorem 3.3

Let $(\rho_t)_{t\geq 0}$ be a Fleming-Viot process, and let $(\mathbf{e}_i)_{i\geq 1}$ be its Eves. In this section we end the proof of Theorem 3.3 by showing that the distribution of the sequence of processes $(\rho_t(\{\mathbf{e}_1\}), \rho_t(\{\mathbf{e}_2\}),\dots;\, t \geq 0)$ is that of a Wright-Fisher diffusion conditioned on its fixation order.

The result we want to prove is the direct extension of Theorem 4 of [21]. Reformulated in our setting, this theorem proves that $(\rho_t(\{\mathbf{e}_1\});\, t\geq 0)$ is distributed as the solution to (3.4) started from 0. We now give a similar representation for the process $(\rho_t(\{\mathbf{e}_1\}),\dots,\rho_t(\{\mathbf{e}_n\});\, t \geq 0)$ giving the size of the progeny of the first n Eves.

Proposition 3.28. *Let $(\rho_t)_{t\geq 0}$ be a Fleming-Viot process, and $(\mathbf{e}_i)_{i\geq 1}$ be its Eves. Then for any $n \geq 1$, the process $(\rho_t(\{\mathbf{e}_1\}),\dots,\rho_t(\{\mathbf{e}_n\});\, t \geq 0)$ is distributed as $(\tilde{Z}_1,\dots,\tilde{Z}_n)$ where $(\tilde{Z}_1,\dots,\tilde{Z}_{n+1})$ is a $(n+1)$-dimensional Wright-Fisher diffusion conditioned on its extinction order, started from $(0,\dots,0,1)$.*

Proof. We realize a similar computation as in the proof of Theorem 4 of [21]. The proof requires three facts. First notice that

$$\lim_{m\to\infty} \rho_t\Big(\Big(\tfrac{\lfloor m\mathbf{e}_i\rfloor}{m}, \tfrac{\lfloor m\mathbf{e}_i+1\rfloor}{m}\Big]\Big) = \rho_t(\{\mathbf{e}_i\}).$$

Then, if $I_1,\dots,I_n$ are n disjoint intervals of length $1/m$, due to exchangeability of the increments of bridges, the process $(\rho_t(I_1),\dots,\rho_t(I_n);\, t\geq 0)$ is distributed as the process

$$\Big(\rho_t\Big(\Big(0,\tfrac{1}{m}\Big]\Big),\dots,\rho_t\Big(\Big(\tfrac{n-1}{m},\tfrac{n}{m}\Big]\Big);\, t \geq 0\Big)$$

which is distributed as the n first coordinates of a $(n+1)$-dimensional Wright-Fisher diffusion started from $(\frac{1}{m}, \dots, \frac{1}{m}, 1 - \frac{n}{m})$.

Finally, notice that on the event $\{\forall i \neq j \in \{1, \dots, n\},\ \lfloor m\mathbf{e}_i \rfloor \neq \lfloor m\mathbf{e}_j \rfloor\}$, conditioning the process

$$\Big(\rho_t\Big(\big(0, \tfrac{1}{m}\big]\Big), \dots, \rho_t\Big(\big(\tfrac{n-1}{m}, \tfrac{n}{m}\big]\Big);\ t \geq 0\Big)$$

on its extinction order as in Section 3.5.2 is equivalent to conditioning it on the location of the Eves, i.e., on the event $\Big\{\forall k \in \{1, \dots, n\},\ \mathbf{e}_k \in \Big(\frac{k-1}{m}, \frac{k}{m}\Big]\Big\}$.

We can now proceed to the calculation. Let $0 \leq t_1 < \cdots < t_p$ and let $\varphi_1, \dots, \varphi_p$ be continuous bounded functions. Consider $(\tilde{Z}_1, \dots, \tilde{Z}_{n+1})$ a $(n+1)$-dimensional Wright-Fisher diffusion conditioned on its extinction order. Then

$$\begin{aligned}
&\mathbb{E}\Big[\varphi_1\Big(\rho_{t_1}(\{\mathbf{e}_1\}), \dots, \rho_{t_1}(\{\mathbf{e}_n\})\Big) \dots \varphi_p\Big(\rho_{t_p}(\{\mathbf{e}_1\}), \dots, \rho_{t_p}(\{\mathbf{e}_n\})\Big)\Big] \\
&= \lim_{m\to\infty} \sum_{i_1=0}^{m-1} \cdots \sum_{i_n=0}^{m-1} \mathbb{E}\Big[\varphi_1\Big(\rho_{t_1}\Big(\big(\tfrac{i_1}{m}, \tfrac{i_1+1}{m}\big]\Big), \dots, \rho_{t_1}\Big(\big(\tfrac{i_n}{m}, \tfrac{i_n+1}{m}\big]\Big)\Big) \dots \\
&\qquad \varphi_p\Big(\rho_{t_p}\Big(\big(\tfrac{i_1}{m}, \tfrac{i_1+1}{m}\big]\Big), \dots, \rho_{t_p}\Big(\big(\tfrac{i_n}{m}, \tfrac{i_n+1}{m}\big]\Big)\Big) \mathbb{1}_{\big\{\forall k \in \{1,\dots,n\},\ \mathbf{e}_k \in \big(\frac{i_k}{m}, \frac{i_k+1}{m}\big]\big\}}\Big] \\
&= \lim_{m\to\infty} m^n \mathbb{E}\Big[\varphi_1\Big(\rho_{t_1}\Big(\big(0, \tfrac{1}{m}\big]\Big), \dots, \rho_{t_1}\Big(\big(\tfrac{n-1}{m}, \tfrac{n}{m}\big]\Big)\Big) \dots \\
&\qquad \varphi_p\Big(\rho_{t_p}\Big(\big(0, \tfrac{1}{m}\big]\Big), \dots, \rho_{t_p}\Big(\big(\tfrac{n-1}{m}, \tfrac{n}{m}\big]\Big)\Big) \mathbb{1}_{\{\forall k \in \{1,\dots,n\},\ \mathbf{e}_k \in (\frac{k-1}{m}, \frac{k}{m}]\}}\Big] \\
&= \lim_{m\to\infty} \mathbb{E}\Big[\varphi_1\Big(\tilde{Z}_1(t_1), \dots, \tilde{Z}_n(t_1)\Big) \dots \varphi_p\Big(\tilde{Z}_1(t_p), \dots, \tilde{Z}_n(t_p)\Big) \ \Big|\ \tilde{Z}_1(0) = \cdots = \tilde{Z}_n(0) = \tfrac{1}{m}\Big] \\
&= \mathbb{E}\Big[\varphi_1\Big(\tilde{Z}_1(t_1), \dots, \tilde{Z}_n(t_1)\Big) \dots \varphi_p\Big(\tilde{Z}_1(t_p), \dots, \tilde{Z}_n(t_p)\Big) \ \Big|\ \tilde{Z}_1(0) = \cdots = \tilde{Z}_n(0) = 0\Big],
\end{aligned}$$

where, the last line comes from the Feller property of the process $(\tilde{Z}_1, \dots, \tilde{Z}_{n+1})$. □

Our current proof of Theorem 3.3 relies on calculations specific to the Wright-Fisher diffusion. We end this section by discussing a potential alternative proof of this result that would more easily generalize to Beta-coalescents.

The Feller branching diffusion describes the size of a population where different individuals die and reproduce independently. Similarly to the Fleming-Viot process, it is possible to define a measure-valued process, called the Dawson-Watanabe process, that encodes the size of the offspring of each individual in the initial population, see e.g. [55]. (Note that there are no mutations here, i.e., no spatial motion of the particles.) Its total mass is then distributed as a Feller diffusion. Starting from a Dawson-Watanabe process, one can renormalize it by its total mass to obtain a process valued in the space of probability measures. Then the resulting renormalized process is distributed as a time-changed Fleming-Viot process, see [26].

Let us now discuss the results of Section 3.5.2 in the light of this new construction. The key point of Section 3.5.2 is that after removing one family from a

Fleming-Viot process and renormalizing the remainder of the population to have mass one, the resulting process remains distributed as an independent time-changed Fleming-Viot process. Suppose that the Fleming-Viot process has been obtained by renormalizing a Dawson-Watanabe process. Then removing a family from the Fleming-Viot process amounts to removing a family from the original Dawson-Watanabe process. By the branching property, removing this family does not change the distribution of the remainder of the population, which remains distributed as an independent Dawson-Watanabe process. Thus when renormalizing the remainder of the population to have size one, we obtain a new time-changed Fleming-Viot process, independent of the removed family. In other words, the results of Section 3.5.2 essentially originate from the fact that the Fleming-Viot process can be seen as a renormalized branching measure-valued process.

A similar link has been obtained in [26] between the Λ-Fleming-Viot processes associated to Beta-coalescents and a family of α-stable measure-valued branching processes. Thus we believe that one could derive a similar, but less explicit, representation of the asymptotic frequencies of the stationary distribution of the Beta-coalescents with erosion than the one obtained in Theorem 3.3.

References for Chapter 3

[18] J. Berestycki. Exchangeable fragmentation-coalescence processes and their equilibrium measures. *Electronic Journal of Probability* **9** (2004), 770–824.

[21] J. Bertoin and J.-F. Le Gall. Stochastic flows associated to coalescent processes. *Probability Theory and Related Fields* **126** (2003), 261–288.

[25] P. Billingsley. *Convergence of Probability Measures*. Second edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1999.

[26] M. Birkner, J. Blath, M. Capaldo, A. M. Etheridge, M. Möhle, J. Schweinsberg, and A. Wakolbinger. Alpha-stable branching and Beta-coalescents. *Electronic Journal of Probability* **10** (2005), 303–325.

[50] T. Duquesne and C. Labbé. On the Eve property for CSBP. *Electronic Journal of Probability* **19** (2014), 31 pp.

[55] A. M. Etheridge. *An Introduction to Superprocesses*. Vol. 20. University Lecture Series. American Mathematical Society, 2000.

[56] A. M. Etheridge. *Some Mathematical Models from Population Genetics. École d'Été de Probabilités de Saint-Flour XXXIX-2009*. Vol. 2012. Lecture Notes in Mathematics. Springer Science & Business Media, 2011.

[58] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1986.

[80] F. Foutel-Rodier, A. Lambert, and E. Schertzer. Kingman's coalescent with erosion. *Electronic Journal of Probability* **25** (2020), 33 pp.

[90] B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303** (2004), 327–332.

[107] T. A. Heath, J. P. Huelsenbeck, and T. Stadler. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences* **111** (2014), E2957–E2966.

[120] O. Kallenberg. *Foundations of Modern Probability.* Second edition. Probability and its Applications. Springer-Verlag New York, 2002.

[127] J. F. C. Kingman. The representation of partition structures. *Journal of the London Mathematical Society* **18** (1978), 374–380.

[128] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications* **13** (1982), 235–248.

[133] C. Labbé. From flows of Λ-Fleming-Viot processes to lookdown processes via flows of partitions. *Electronic Journal of Probability* **19** (2014), 49 pp.

[135] A. Lambert. Population Dynamics and Random Genealogies. *Stochastic Models* **24** (2008), 45–163.

[143] J. Lamperti. Semi-stable Markov processes. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **22** (1972), 205–225.

[151] J. Mallet, N. Besansky, and M. W. Hahn. How reticulated are species? *BioEssays* **38** (2016), 140–149.

[178] J. Pitman. Coalescents with multiple collisions. *The Annals of Probability* **27** (1999), 1870–1902.

[186] C. Roux, C. Fraïsse, J. Romiguier, Y. Anciaux, N. Galtier, and N. Bierne. Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLOS Biology* **14** (2016), 1–22.

[188] S. Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability* **36** (1999), 1116–1125.

[201] G. J. Slater, L. J. Harmon, and M. E. Alfaro. Integrating fossils with molecular phylogenies improves inference of trait evolution. *Evolution: International Journal of Organic Evolution* **66** (2012), 3931–3944.

[215] E. M. Volz, K. Koelle, and T. Bedford. Viral phylodynamics. *PLoS computational biology* **9** (2013).

Appendices for Chapter 3

3.A Proof of Proposition 3.17

In this section, we prove that the ancestral process of Kingman's coalescent with immigration is Markovian. To do this, consider a version of Kingman's coalescent with immigration $(\bar{\Pi}_t)_{t\in\mathbb{R}}$, and let $(\bar{\Pi}_i)_{i\in\mathbb{Z}}$ be its embedded chain, i.e., the sequence of states visited by $(\bar{\Pi}_t)_{t\in\mathbb{R}}$, where $\bar{\Pi}_0$ is the state at time $t = 0$. We count the number of trajectories of $(\bar{\Pi}_i)_{i\in\mathbb{Z}}$ that produce a given trajectory of $(\mathcal{A}_i)_{i\geq 0}$, the embedded chain of $(\mathcal{A}_t)_{t\geq 0}$.

First, note that given the values of $(\bar{\Pi}_{-n}, \dots, \bar{\Pi}_0)$ and a uniform permutation σ of the blocks of $\bar{\Pi}_0$, one can uniquely reconstruct the values of $(\mathcal{A}_0, \dots, \mathcal{A}_n)$. We now fix a sequence $(a_0, \dots, a_n)$ of possible values of $(\mathcal{A}_0, \dots, \mathcal{A}_n)$, and a partition $\bar{\pi}_{-n}$ with $|a_n|$ blocks, where $|a_n|$ is the total number of particles of a_n. Our first task is to count the number of trajectories of $(\bar{\Pi}_{-n}, \dots, \bar{\Pi}_0)$ starting from $\bar{\pi}_{-n}$, and of labelings σ of the blocks of $\bar{\Pi}_0$ such that $(\mathcal{A}_0, \dots, \mathcal{A}_n) = (a_0, \dots, a_n)$. Before stating the result we need to introduce one notation. The variable $\mathcal{A}_{k+1}$ is obtained from $\mathcal{A}_k$ by splitting or killing one particle. Let us denote ℓ_k the label of this particle. That is, ℓ_k is the unique integer such that

$$|\mathcal{A}_{k+1}(\ell_k) - \mathcal{A}_k(\ell_k)| = 1, \quad \forall i \neq \ell_k,\ |\mathcal{A}_{k+1}(i) - \mathcal{A}_k(i)| = 0.$$

Lemma 3.29. *Fix a sequence of states $(a_0, \dots, a_n)$ of $(\mathcal{A}_0, \dots, \mathcal{A}_n)$, and a partition $\bar{\pi}_{-n}$ of $\{i \in \mathbb{Z} : i \leq -n\}$ with $|a_n|$ blocks. Then the number of trajectories of $(\bar{\Pi}_{-n}, \dots, \bar{\Pi}_0)$ and labelings of the blocks of $\bar{\Pi}_0$ such that $(\mathcal{A}_0, \dots, \mathcal{A}_n) = (a_0, \dots, a_n)$ and $\bar{\Pi}_{-n} = \bar{\pi}_{-n}$ is*

$$\frac{|a_n|!}{2^b} a_0(\ell_0) \dots a_{n-1}(\ell_{n-1}),$$

where b is the number of birth events along the sequence $(a_0, \dots, a_n)$.

Proof. Each trajectory of $(\bar{\Pi}_{-n}, \dots, \bar{\Pi}_0)$ naturally encodes a forest that can be built through the following procedure, which is illustrated in Figure 3.2. Choose any labeling of the blocks of $\bar{\Pi}_{-n}$, and for each block add an initial leaf with the corresponding label. Suppose that the forest corresponding to $(\bar{\Pi}_{-n}, \dots, \bar{\Pi}_{-k})$ has been built. If $\bar{\Pi}_{-k+1}$ is obtained from $\bar{\Pi}_{-k}$ by immigrating a new particle, then add a new isolated vertex. Otherwise, a coalescence event has occurred between two blocks of $\bar{\Pi}_{-k}$. Then add a new internal node and connect it to the nodes corresponding to the two blocks that have coalesced. Once the forest representing

$(\bar{\Pi}_{-n},\dots,\bar{\Pi}_0)$ is built, by construction the nodes corresponding to $\bar{\Pi}_0$ all belong to different trees. We set them to be the roots of their respective trees, and label them according to the partition σ. (Notice that the resulting forest is endowed with some additional structure: the nodes added along the procedure are totally ordered by the induction step at which they have been added.)

Counting trajectories of $(\bar{\Pi}_{-n},\dots,\bar{\Pi}_0)$ now amounts to counting forests. Instead of building the forests by starting from the leaves as above, we build a forest with ancestral sequence $(a_0,\dots,a_n)$ by starting from the roots. Initially, consider a set of $|a_0|$ roots, labeled by $\{1,\dots,|a_0|\}$, that represent the particles of a_0. Nodes can be in two states: active or inactive. Active nodes represent the particles that are still alive in the population while inactive nodes represent the dead particles. Initially all roots are active. We build the forest recursively. Suppose that at step k we have built a forest such that for all i there are $a_k(i)$ nodes that are active in the tree with root i. If a particle with label ℓ_k has died from a_k to a_{k+1}, we inactivate one of the nodes belonging to the tree with root ℓ_k. There are $a_k(\ell_k)$ such nodes. Similarly, if a particle has split from a_k to a_{k+1}, we inactivate one node in the tree ℓ_k, and connect it to two new active nodes. There are again $a_k(\ell_k)$ active nodes in the tree ℓ_k. After step n, we have built a forest with ancestral sequence $(a_0,\dots,a_n)$. We assign the blocks of $\bar{\Pi}_{-n}$ to the remaining active nodes of the forest by choosing one of the $|a_n|!$ permutations of the blocks.

There are

$$|a_n|!\, a_0(\ell_0)\dots a_{n-1}(\ell_{n-1})$$

possible outputs of the previous construction, and all forests with ancestral sequence $(a_0,\dots,a_n)$ can be obtained that way. However, due to symmetries, some forests can be obtained multiple times through this construction. More precisely, at each birth event, the two daughter nodes are indistinguishable. Interchanging the trees corresponding to the offspring of these two nodes yields the same forest. Thus, the actual number of forests with ancestral sequence $(a_0,\dots,a_n)$ is

$$\frac{|a_n|!}{2^b} a_0(\ell_0)\dots a_{n-1}(\ell_{n-1})$$

where b is the number of birth events, and the result is proved. □

Lemma 3.30. *Let $(M_t)_{t\in\mathbb{R}}$ be the process counting the number of blocks of Kingman's coalescent with immigration. Then $(M_t)_{t\in\mathbb{R}}$ is a stationary Markov process such that conditional on $\{M_t = k\}$, it jumps to*

- $k+1$ *at rate* d*;*
- $k-1$ *at rate* $k(k-1)/2$*.*

Moreover $(M_t)_{t\in\mathbb{R}}$ is a reversible process.

Proof. Let us consider a version of Kingman's coalescent with immigration built from a Poisson point process P. Let us first show that $(M_t)_{t\in\mathbb{R}}$ is a Markov process. Conditional on $M_t = k$, each of the $k(k-1)/2$ pairs of blocks coalesce a rate one, and new atoms of P immigrate at rate d. Thus, $(M_t)_{t\in\mathbb{R}}$ goes to $k-1$ at rate $k(k-1)/2$ and to $k+1$ at rate d.

Let us now argue that the family of variables $(M_t)_{t\in\mathbb{R}}$ is tight. Fix $t \in \mathbb{R}$ and let T be the location of the most recent atom of P before time t. Then $t-T$ is exponentially distributed with parameter d, and M_t is distributed as the number of blocks of Π_{t-T}, where $(\Pi_s)_{s\geq 0}$ is a version of Kingman's coalescent started with M_T blocks. Thus M_t is stochastically dominated by the number of blocks Π'_{t-T}, where $(\Pi'_s)_{s\geq 0}$ is a version of Kingman's coalescent started from an infinite number of blocks. As each variable M_t is stochastically dominated by the same variable, the family is tight.

It is not hard to see that a Markov process jumping from k to $k+1$ at rate d, and from k to $k-1$ at rate $k(k-1)/2$ admits a unique stationary distribution. As it is irreducible we have

$$\forall k \geq 1, \quad \mathbb{P}(M_t = i \mid M_s = j) \underset{t\to\infty}{\longrightarrow} \mathbb{P}(M_\infty = i).$$

Thus, using the tightness of $(M_t)_{t\geq 0}$, we have

$$\mathbb{P}(M_t = i) = \sum_{j\geq 1} \mathbb{P}(M_s = j)\mathbb{P}(M_{t-s} = i \mid M_s = j) \underset{s\to-\infty}{\longrightarrow} \mathbb{P}(M_\infty = i).$$

Let us compute the stationary distribution of $(M_t)_{t\in\mathbb{R}}$. As $(M_t)_{t\in\mathbb{R}}$ jumps from k to $k+1$ at rate d and from k to $k-1$ at rate $k(k-1)/2$, a usual calculation shows that its stationary distribution $(\nu_k)_{k\geq 1}$ is

$$\forall k \geq 1, \quad \nu_k \propto \frac{(2d)^k}{k!\,(k-1)!}$$

where the renormalization constant is obtained by summing over all the terms. Thus a direct calculation now proves that $(\nu_k)_{k\geq 1}$ fulfills the detailed balance equation

$$\forall k \geq 1, \quad d\nu_k = \frac{k(k+1)}{2}\nu_{k+1}$$

and thus that $(M_t)_{t\in\mathbb{R}}$ is reversible. □

We are now ready to prove Proposition 3.17

Proof of Proposition 3.17. Recall the notations from Section 3.3.1. As proved in Lemma 3.30, the process $(M_t)_{t\in\mathbb{R}}$ that counts the number of the blocks of Kingman's coalescent with immigration is a reversible Markov process. Thus, the process $(N_t)_{t\geq 0}$ that gives the number of particles of $(\mathcal{A}_t)_{t\geq 0}$ is a stationary process jumping from k to $k+1$ at rate d, and from k to $k-1$ at rate $k(k-1)/2$. Hence, the result is proved if we show that conditional on the sequence of states $(N_0, \dots, N_n)$ visited by $(N_t)_{t\geq 0}$, the type of the particle that dies or splits from $\mathcal{A}_k$ to $\mathcal{A}_{k+1}$ is chosen with a probability proportional to the vector $\mathcal{A}_k$.

Let b denote the number of birth events along the sequence $(a_0, \dots, a_n)$. (Hence, forward in time, there are $n-b$ immigration events.) We have

$$\begin{aligned}&\mathbb{P}(\mathcal{A}_0 = a_0, \dots, \mathcal{A}_n = a_n)\\ &\qquad = \sum_{(\bar\pi_{-n},\dots,\bar\pi_0)} \sum_s \mathbb{P}\Big(\forall i < n,\ \bar\Pi_{-i} = \bar\pi_{-i},\ \sigma = s \,\Big|\, \bar\Pi_{-n} = \bar\pi_{-n}\Big)\mathbb{P}\Big(\bar\Pi_{-n} = \bar\pi_{-n}\Big)\end{aligned}$$

where the sum is taken over all partitions $\bar\pi_{-n}$ of $\{i \in \mathbb{Z} : i \leq -(n-b)\}$ with $|a_n|$ blocks, all trajectories $(\bar\pi_{-n+1}, \dots, \bar\pi_0)$ and labelings s of the blocks of $\bar\pi_0$ such that $(\mathcal{A}_0, \dots, \mathcal{A}_n) = (a_0, \dots, a_n)$. Now notice that the probability of seeing such a trajectory and labeling does only depend on the sequence of number of blocks $(|a_0|, \dots, |a_n|)$. Indeed, conditional on $(|a_0|, \dots, |a_n|)$, two trajectories $(\bar\Pi_{-n}, \dots, \bar\Pi_0)$ are identical up to the choice of the pairs of blocks that merge at each coalescence event, and these pairs are chosen uniformly.

Thus the probability of the event $\{\mathcal{A}_0 = a_0, \dots, \mathcal{A}_n = a_n\}$ is proportional to the number of terms in the sum, and thus to the number of trajectories of $(\bar\Pi_{-n}, \dots, \bar\Pi_0)$ that correspond to this ancestral sequence. Hence, Lemma 3.29 shows that

$$\mathbb{P}(\mathcal{A}_0 = a_0, \dots, \mathcal{A}_n = a_n) = C(|a_0|, \dots, |a_n|)a_0(\ell_0)\dots a_{n-1}(\ell_{n-1}),$$

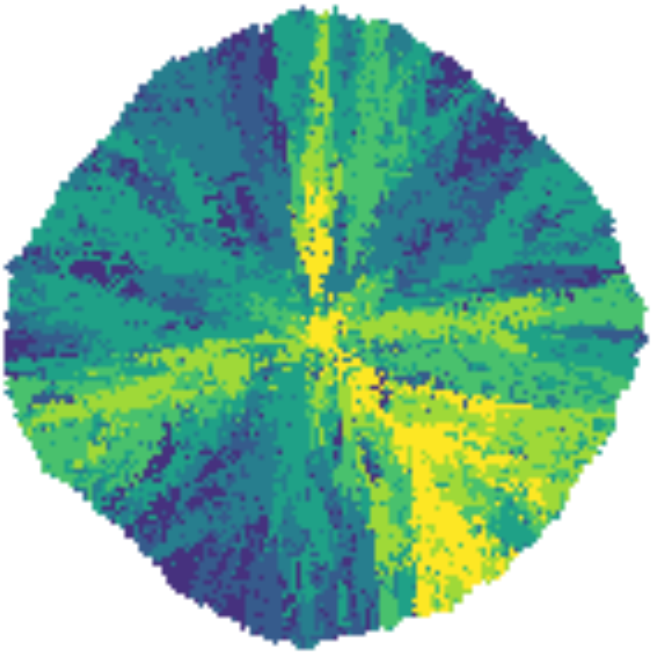
where the coefficient $C(|a_0|, \dots, |a_n|)$ only depends on $(|a_0|, \dots, |a_n|)$. This proves the result. □

Let us end this section by discussing a possible extension to Λ-coalescents. The key point here is that conditional on the block counting process, the particles that die or split are chosen uniformly in the population. This is a consequence of 1) Lemma 3.29 and 2) the fact that all trajectories with a given sequence of number of blocks have the same probability. The second point is a consequence of exchangeability so remains valid for Λ-coalescents. As for Lemma 3.29, the proof could be easily adapted to Λ-coalescents with immigration. (The factor 2^b should be replaced by the product of the number of blocks involved in coalescence events.)

Thus, the only difference between Kingman's coalescent with immigration and more general Λ-coalescents with immigration is that the block counting process is no longer reversible. Hence we cannot obtain a closed form for the transition rates of the corresponding ancestral processes. Nevertheless, we believe that in some cases it should be possible to obtain a result similar to Theorem 3.5 by using the same techniques as in this paper, if one can derive a good enough approximation for the stationary distribution of the number of blocks.

Branching models in population genetics

Chapter 4



4

The spatial Muller's ratchet: Surfing of deleterious mutations during a range expansion

This chapter is joint work with Alison Etheridge. It is published in *Theoretical Population Biology* [78].

Illustration. Simulation of the spatial Muller's ratchet in two dimensions. The interpretation of the colors is described in the caption of Figure 4.6.

4.1 Introduction

Gene surfing and expansion load. The genetics of range expansion is a complex topic that has attracted much attention. In a pioneering work, [54] reported that during a range expansion a neutral mutant appearing in the front of an expansion could rapidly spread over a vast region of space. This phenomenon was further studied in [129] and dubbed gene surfing, see Figure 4.1 for an illustration. Gene surfing originates from two features of range expansions. First, the population density is lower at the range's margin than in its core, where the population has had more time to grow to carrying capacity. Thus a mutant that appears there is already in relatively high frequency among the few individuals in the front. Moreover, individual-level demographic stochasticity, which is the cause of population-level genetic drift, can lead to a further rapid increase of the local frequency of this mutant. Second, population spread can be caricatured by successive founding events, where a few individuals migrate to an empty habitat and grow a new subpopulation. Individuals living at the edge are more likely to be recruited to found these new subpopulations as they are spatially closer to the empty habitats. In other words, individuals that form the subsequent front are sampled from the current front, not from the bulk. Combining these two features, the initial increase in frequency of the mutant at the front (which is the result of the small population size) gets amplified by the successive resampling from the front, and the mutant can reach a high frequency over a large spatial area, see

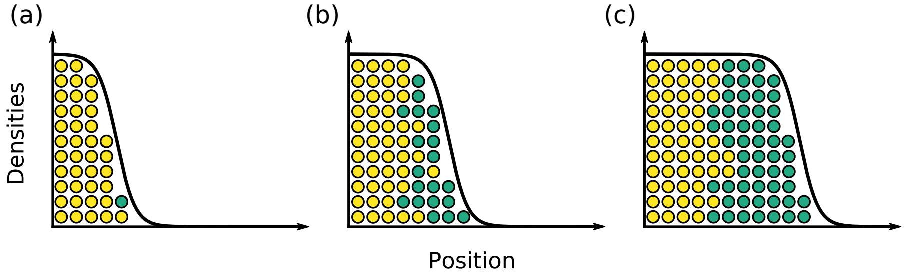


Figure 4.1: Illustration of gene surfing: (a) a mutant (in green) appears in the front of an expanding population; (b) the mutant rapidly reaches fixation in the front; (c) the mutant offspring further expand, leading to a clear allele segregation.

Figure 4.1, panel (c). Gene surfing is now a well-understood phenomenon, that has been assessed both theoretically [54, 129, 211, 99], and empirically using microbial growth experiments [98, 100] or naturally occuring genetic data [88]. See [63, 62] for reviews of this topic.

The very first step of the surfing phenomenon is the local increase in frequency of an allele at the front, resulting from the increased growth rate when population density is low. Increased genetic drift at the front makes reaching fixation easier for both neutral and deleterious mutants. Hence it is not surprising that deleterious mutations can also surf [211]. As deleterious mutations are more frequent than beneficial ones [64], it has been predicted and assessed in [172] that fitness at the front is decreasing during a range expansion, due to successive surfing of deleterious mutations along the expansion axis. This reduction in fitness due to range expansion is known as *expansion load.* Expansion load is thus the additional fitness disadvantage that a population has accumulated during a range expansion, due to a reduced ability of selection to purge deleterious mutations, see [170] for a review. While expansion load has been clearly highlighted using simulations [172, 173, 174, 85], genomic evidence of an expansion load remains a debated topic [47, 200]. The presence of an expansion load has been reported in human populations after the Out of Africa expansion [108, 171], and in plants [87, 218], see [170] for a review.

Impact of an Allee effect. A population exhibits an Allee effect if its maximal per-capita growth rate is achieved at intermediate population density rather than at low population density. The Allee effect is said to be *strong* if the per-capita growth rate is negative at low population density, that is if the population is unable to grow under a certain critical density threshold [30, 206]. Otherwise the Allee effect is *weak.* Allee effects arise in many biological contexts, including for example the presence of cooperation, or the difficulty in finding mates for reproduction [132, 203].

In the context of range expansion, an Allee effect shifts the location of the individuals with the highest per-capita growth rate towards the bulk of the population, see Figure 4.2 panel (a). In the absence of an Allee effect, individuals at the leading edge of the front, where the population density is the lowest, have the highest growth rate. The wave is *pulled* by these individuals in the front. Conversely, if the Allee effect is strong enough, individuals with the highest growth rate are located midway between the front and the bulk of the population. The wave is then *pushed* by the bulk of the population. The distinction between pushed and pulled waves is a well-established paradigm of the reaction-diffusion literature, see [204, 213]. The correspondence between the pulled/pushed nature of the waves and the absence/presence of an Allee effect is not perfect. A weak Allee effect can lead to both pulled and pushed waves, as is for instance the case in the model considered in [27]. Nevertheless, an Allee effect is a necessary condition for the wave to be pushed [204], and a strong Allee effect is a sufficient one.

From a genetic perspective, an Allee effect increases the "effective population size" of the front. Individuals that leave the largest number of offspring are either located in the front in the pulled case, or towards the bulk in the pushed case. Thus, we expect that only the very few individuals far in the front contribute to the genetic pool of the population in a pulled wave, leading to a more drastic loss of diversity than in a pushed wave, see Figure 4.2 for an illustration. Several theoretical studies have assessed the impact of an Allee effect on the genetics of range expansion [184, 83, 99, 27] in a neutral setting, let us briefly review their results.

The authors of [99] considered a stochastic particle system modelling range expansion, and studied the fixation probability of individuals in the front as a function of their locations. Using both simulations and analytical approximations, they provided an expression for the fixation probability, and showed that it reaches a maximum at a location which is shifting towards the bulk of the population as the strength of the Allee effect increases. In [184, 83] the authors considered a deterministic partial differential equation analogous to the celebrated Fisher-KPP equation [70, 130], which has been widely used to model invading populations. They divided the total population into several neutral fractions, and studied the long-term fate of these fractions. They proved that in the pulled case, only the fraction closest to the front is able to follow the expansion, and that all other fractions are left behind. In the long run, the population is only composed of the offspring of the individuals that were initially closest to the front. Conversely in the pushed case all fractions are able to follow the expansion. Asymptotically, all individuals in the population leave progeny that live in the front. The long-term contribution to the front of the various initial fractions can be computed explicitly, with an expression consistent with the approximate fixation probability found in [99]. Finally, [27] used simulations and analytical approximations to study the rate of loss of genetic diversity during a range expansion. They showed that, as the strength of the Allee effect increases, the loss of genetic diversity is slowed down.

All the above studies consistently find that an Allee effect impedes gene surfing, and rescues the genetic diversity in the front of an expanding population. Expansion load originates from successive surfing of deleterious mutants, we thus expect the presence of an Allee effect to reduce expansion load. Nevertheless, the impact of an Allee effect on expansion load has never been explicitly tested. All existing

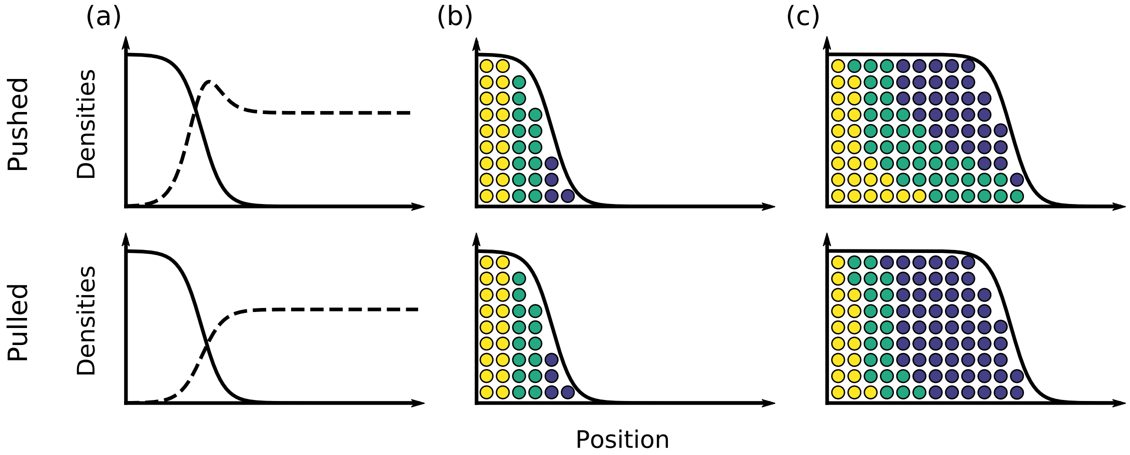


Figure 4.2: Illustration of the impact of an Allee effect. (a) Solid line: population density; dashed line: per-capita growth rate. (b) Particles have been labeled in different colours according to their initial location. (c) Genetic composition of the population after expansion. In the pulled wave, only the few individuals in the front are the founders of the new habitats, while in the pushed wave individuals from the bulk also contribute to the new front.

theoretical studies on expansion load assume logistic growth of the population, in particular no Allee effect [172, 174, 173]. Moreover, the only analytical results available for these models have been derived using a serial founder approximation. In this approximation, the front is supposed to be genetically isolated from the bulk, and at each time-step, a new front is formed by sampling a few individuals from the previous front and letting them grow logistically. Even if this approach yields good approximations of the mean fitness at the front, it misses the continuous gene flow between the bulk and the front that occurs during range expansion, especially in the presence of an Allee effect. The objectives of the present work are thus twofold. First we aim to study the impact of an Allee effect on the expansion load. We will restrict our attention to the case of a weak Allee effect, and will not consider the impact of a strong Allee effect. Second, we aim to build a model that is more amenable to continuous space techniques, in order to take into account the entire dynamics of the expanding population.

A spatial Muller's ratchet. In order to keep the genetic structure of the population as simple as possible, we consider genetic dynamics similar to that of [97], leading to a Muller's ratchet [67, 160]. Muller's ratchet is a mechanism that was first proposed as an explanation for the evolution of recombination, it can be formulated as follows. Consider a population of finite size that can only accumulate deleterious mutations over time. If mutations are irreversible and negatively selected, without drift the population should reach a mutation-selection equilibrium. Nevertheless, due to genetic drift all individuals without mutations are eventually lost. At such a time, in the absence of recombination, the minimal number of mutations in the population is permanently increased by one. We say that the ratchet has clicked. At each click of the ratchet the fitness of the population is decreased and successive clicks of the ratchet should drive it towards extinction. In the presence of recombination, chromosomes without mutations can be recreated by recombining two chromosomes with mutations at different loci, rescuing the population from the ratchet, and thus giving recombination a selective advantage.

In our context, we consider an expanding population where individuals can only accumulate deleterious mutations, without possible reversion or recombination. Due to higher genetic drift at the front, we expect the ratchet to click more often in the front than in the bulk. After a click at the front, we expect the population to be separated into two distinct regions: one towards the front where the ratchet has clicked, and the other towards the bulk where the ratchet has not clicked. Despite their lower fitness, individuals in the front will still be able to colonise new habitat as the low population density guarantees a positive growth rate. Interestingly, individuals without mutations have a positive growth rate when they are placed in a location where the ratchet has clicked, as their fitness is larger. Therefore, the region where the ratchet has not clicked should also be able to expand into the region where it has, at a rate depending on the fitness difference between the two regions. This can be thought of as an inner expansion wave evolving inside the larger expansion wave of the whole population. Each click of the ratchet should create a new inner wave of less unfit individuals. Successive rapid clicks of the ratchet at the front will create an expansion load, but there will be some recovery of fitness at any given location as fitter individuals from the "bulk" invade. This phenomenon will be illustrated numerically in the forthcoming Section 4.3.2. Let us spell out the dynamics of the model more precisely.

4.2 Methods

Description of the model. We consider a population of non-recombining individuals each carrying a single chromosome. The population is subdivided in demes indexed by $\mathbb{Z}$. Each individual is entirely characterized by the number of deleterious mutations it carries and its spatial location. We record as $n_{i,k}(t)$ the number of individuals carrying $k \geq 0$ mutations in deme $i \in \mathbb{Z}$ at time $t \in \mathbb{R}_+$, and let $N_i(t) = \sum_{k\geq 0} n_{i,k}(t)$ be the total population size in deme i. Thus the vector $(n_{i,k}(t);\, i \in \mathbb{Z},\, k \geq 0)$ contains all the information about the population at time t.

Time is continuous and individuals can reproduce, die or migrate according to the following rules. An individual located in deme i and carrying k mutations gives birth to a new individual at rate $\lambda_k(N_i)$, and dies at rate $\delta(N_i)$, where

$$\lambda_k(n) = r(1-s)^k(B\frac{n}{N}+1), \quad \delta(n) = r(B\frac{n}{N}+1)\frac{n}{N}. \tag{4.1}$$

The offspring is located in the same deme as its parent. With probability $1-\mu$, it inherits the same number of mutations as its parent, and with probability μ, it accumulates an additional one. Finally each individual migrates at rate m, and goes to one of the two nearest demes with equal probability.

Let us provide an intuitive description of equation (4.1) and of the parameters of the model. All deleterious mutations have the same fitness effect $s > 0$, and fitness is multiplicative across loci. If $B = 0$, we recover a stochastic version of the usual logistic growth, where r is the Malthusian growth parameter and N is a scaling parameter that can be thought of as the local carrying capacity of the population. Taking $B > 0$ introduces a density-dependence in the growth rate of the population. We think of B as a cooperation parameter, where cooperation acts on the overall growth rate of the population, which will tune the strength of the Allee effect. Notice that the function $n \mapsto \lambda_0(n) - \delta(n)$ is non-negative, and that it reaches its maximum for $n = \max(0, \frac{N(B-1)}{2B})$. Thus, for $B \leq 1$, the per-capita growth rate is maximal for $n = 0$, i.e., there is no Allee effect, while for $B > 1$, we see a weak Allee effect. Increasing B increases the strength of that Allee effect, as it further shifts the location of the maximal per-capita growth rate to higher population density. This parametrization of the Allee effect is similar to the one considered in [27].

Large population scalings. As is usual in population genetics, in order to obtain analytical results about our model, we consider a large population size scaling. We begin with the deterministic infinite population limit. For a fixed value of N, let $(n^N_{i,k}(t);\, t \geq 0,\, i \in \mathbb{Z},\, k \geq 0)$ be a realization of the above model, with population size parameter N and migration rate m_N. Let L be a space renormalization parameter, and for $i \in \mathbb{Z}$ and $x = i/L$ set

$$\forall k \geq 0, \quad u^N_k(x,t) = \frac{n^N_{i,k}(t)}{N},$$

and interpolate the function $u^N_k(\cdot, t)$ linearly between the points $\{i/L : i \in \mathbb{Z}\}$. Then a standard generator calculation (see Section 4.A) suggests that provided the initial condition converges, and $m_N/L^2 \to m$, then, as $N, L \to \infty$, $(u^N_k)_{k\geq 0}$ converges to the solution of

$$\begin{cases} U = \sum_{k\geq 0} u_k, \quad u_{-1} \equiv 0, \\ \forall k \geq 0, \quad \partial_t u_k = m\partial_{xx} u_k + r(BU+1)\Big[u_k\Big((1-\mu)(1-s)^k - U\Big) + \\ \qquad\qquad\qquad\qquad\qquad\qquad \mu\Big((1-s)^{k-1}u_{k-1} - (1-s)^k u_k\Big)\Big], \end{cases}$$

started from the corresponding initial condition. In what follows, we will always assume that selection is weak, and that mutation is low, i.e., that $s, \mu \ll 1$. To the first order in s and μ, the above equation becomes

$$\begin{cases} U = \sum_{k\geq 0} u_k, \quad u_{-1} \equiv 0, \\ \forall k \geq 0, \quad \partial_t u_k = m\partial_{xx} u_k + r(BU+1)\Big(u_k(1-ks-U) + \mu(u_{k-1} - u_k)\Big). \end{cases} \tag{4.2}$$

This limit is deterministic and thus does not take genetic drift into account and we do not observe gene surfing. By retaining terms up to order $1/N$ in a generator calculation, we can derive a diffusion approximation for our model that accounts for finite size fluctuations, see Section 4.A. Under a diffusive scaling, where the population size is further rescaled by a factor L, and time is rescaled by a factor L^2, when N is large, the process $(u_k)_{k\geq 0}$ is approximated by the following system of stochastic partial differential equations

$$\begin{cases} U = \sum_{k\geq 0} u_k, \quad u_{-1} \equiv 0, \\ \forall k \geq 0, \quad \partial_t u_k = m\partial_{xx} u_k + r(BU+1)\Big(u_k(1-ks-U) + \mu(u_{k-1}-u_k)\Big) \\ \qquad\qquad + \sqrt{\frac{r}{N} u_k(BU+1)(1-ks+U)}\dot{W}_k, \end{cases} \tag{4.3}$$

where $(\dot{W}_k)_{k\geq 0}$ are independent space-time white noises.

The above two limits have been obtained by qualitative comparison of the generator of $(u_k^N)_{k\geq 0}$ for large N. We do *not* prove the convergence of the process to any of these limiting objects, which is a highly non-trivial problem. Nevertheless, see [53, 161] for rigorous treatments of similar convergence results.

Simulation setup. When started from a finite number of individuals, our model is a simple continuous-time Markov chain that we simulate using the following classical algorithm: at each iteration of the algorithm, we compute the total transition rate $\bar{w}$ of the population and increment the time t by an exponential variable with mean $1/\bar{w}$. The transition that occurs is then chosen independently with probability proportional to the transition rates. Notice that t will always refer to the "actual time" of the simulation and *not* to the number of iterations. In each simulation, at $t = 0$ only the first 30 demes are occupied, all other demes are empty. The initial number of individuals in the occupied demes, and the distribution of the number of mutations, is chosen according to their deterministic equilibrium value, computed in (4.7). We restricted the spatial domain to 500 demes and used reflecting boundary conditions for the migration, i.e., individuals are not allowed to move outside the domain.

Estimation of the click rate. In order to estimate the click rate, we need to determine from the simulation the moment when the ratchet has clicked. Let us denote by $n_k^{\max}(t)$ the location of the right-most deme containing individuals carrying k mutations at time t, defined as

$$n_k^{\max}(t) = \max\{i \in \{1,\dots,500\} : n_{i,k}(t) > 0\}.$$

We define the approximate first click time T_1 of the ratchet as

$$T_1 = \inf\{t \geq 0 : \exists s \geq t,\, n_1^{\max}(s) - n_0^{\max}(s) > d \text{ and } \forall r \in [t,s],\, n_1^{\max}(r) > n_0^{\max}(r)\}.$$

In words, T_1 is the first moment when individuals with one mutation get ahead of individuals with no mutations, and will get d demes ahead before being caught up.

We set $d = 30$ by observing from the simulations that once $n_1^{\max} - n_0^{\max} > d$, it is very unlikely that the inner wave catches up the front of expansion before the population has colonized the entire habitat.

In order to obtain the mean time to the first click, we averaged T_1 over many simulations for various parameter values. Once the population has expanded, as there are many more individuals in the bulk than in the front, most of the events that occur are reproduction events in the bulk. These events are not relevant for the computation of the click time as individuals far in the bulk will never be able to reach the front. Thus in order to speed the simulations up, a frame of width d, co-moving with the front, was used for the simulations of Figure 4.5. We set the birth, death and migration rates of all individuals in deme i such that $i < n_1^{\max} - d$ to 0, and also prevent individuals in deme $n_1^{\max} - d$ from migrating to the left. We emphasize that the co-moving frame was only used in the simulations of Figure 4.5 and that all other simulations account for the events that occur in the bulk.

Additionally, we define the time between the $k-1$-th and the k-th click as

$$\begin{aligned} T_k = \inf\{t \geq 0 : \exists s \geq t,\ & n_k^{\max}(s) - n_{k-1}^{\max}(s) > d \\ & \text{and } \forall r \in [t,s],\ n_k^{\max}(r) > n_{k-1}^{\max}(r)\} - T_{k-1}. \end{aligned} \tag{4.4}$$

The number of demes that the population has colonized between the $k-1$-th and the k-th click is then given by

$$d_k = n_k^{\max}(T_1 + \cdots + T_k) - n_k^{\max}(T_1 + \cdots + T_{k-1}).$$

Two-dimensional simulations. In the two-dimensional simulations, at $t = 0$ a five by five square of demes is occupied in the centre of the habitat and all other demes are empty. The number of individuals in these demes is chosen as in the one-dimensional case according to the deterministic values computed in (4.7). The simulation is run until $t = 150$. (Recall that $t = 150$ refers to the “actual time” of the simulation, and not to a number of iterations.) We want to record the number of clicks of the ratchet at the colonization time in each deme. One naive way of doing this could be to record for each deme the number of mutations of the first individual that migrates to this deme. However this would produce an extremely noisy picture. Even if the ratchet has not clicked yet, many individuals in the front carry mutations and could by chance migrate first to a new deme. In order to reduce this noise, we have chosen to look at the population at each time unit of the “actual time”, i.e., at $t = 1, 2, \ldots, 150$, and to record the least number of mutations in each newly colonized deme. More precisely, let $n_{i,j;k}(t)$ denote the number of individuals carrying k mutations in deme (i,j), and let $N_{i,j}$ be the total number of individuals in deme (i,j). We define the colonization time $t_{i,j}^{\mathrm{col}}$ and number of clicks at colonization $k_{i,j}^{\mathrm{col}}$ in deme (i,j) to be

$$\begin{aligned} t_{i,j}^{\mathrm{col}} &= \inf\{t \in \{1, \ldots, 150\} : N_{i,j}(t) > 0 \text{ and } N_{i,j}(t-1) = 0\} \\ k_{i,j}^{\mathrm{col}} &= \inf\{k \geq 0 : n_{i,j;k}(t_{i,j}^{\mathrm{col}}) > 0\}. \end{aligned}$$

4.3 Results

4.3.1 Analysis of the deterministic limit

In this section we study the set of reaction-diffusion equations that we obtained by taking the deterministic scaling of the model, namely equation (4.2). Similar reaction-diffusion equations have been widely used to model biological invasions [184, 70]. They are usually studied through their travelling wave solutions. In our context, a travelling wave solution to (4.2) is a solution $(u_k)_{k\geq 0}$ that can be written

$$\forall k \geq 0,\, \forall x \in \mathbb{R},\, \forall t \geq 0, \quad u_k(x,t) = \widehat{u}_k(x-ct),$$

where $c > 0$ is the wave speed and

$$\forall k \geq 0, \quad \widehat{u}_k : \mathbb{R} \to \mathbb{R}_+$$

is the wave shape. A travelling wave solution is thus a constant wave form $(\widehat{u}_k)_{k\geq 0}$ that is shifted at a constant speed c towards the positive reals. Additionally, we impose that the solution connects two stationary points of (4.2), i.e., that

$$\lim_{x\to\infty} \widehat{u}_k(x) = u_k^+, \qquad \lim_{x\to-\infty} \widehat{u}_k(x) = u_k^-,$$

where $(u_k^+)_{k\geq 0}$ and $(u_k^-)_{k\geq 0}$ are two homogeneous solutions to (4.2). Let us first study the non-spatial equivalent of (4.2) to obtain the homogeneous solutions of the system.

Equilibrium of the non-spatial system. Let us consider the following non-spatial version of (4.2),

$$\forall k \geq 0, \quad \frac{\mathrm{d}u_k}{\mathrm{d}t} = r(BU+1)\Big(u_k(1-ks-U) + \mu(u_{k-1}-u_k)\Big). \tag{4.5}$$

Equivalently, this system can be reformulated in terms of the total population size U and of the vector $(p_k)_{k\geq 0} = (u_k/U)_{k\geq 0}$ giving the frequencies of the different types, that we call the *genetic composition* of the population. Equation (4.5) is then equivalent to

$$\begin{cases} \dfrac{\mathrm{d}U}{\mathrm{d}t} = rU(BU+1)(1-s\bar{p}-U), \quad p_{-1} \equiv 0, \\ \forall k \geq 0, \quad \dfrac{\mathrm{d}p_k}{\mathrm{d}t} = r(BU+1)\Big(sp_k(\bar{p}-k) + \mu(p_{k-1}-p_k)\Big) \end{cases} \tag{4.6}$$

where we have set

$$\bar{p} := \sum_{j\geq 0} jp_j$$

to be the mean number of mutations. Up to the non-constant population size, equation (4.6) has already been derived in [57] to describe dynamics of the frequencies of individuals carrying different numbers of mutations in Muller's ratchet.

It is straightforward to see that if $(u_k^*)_{k\geq 0}$ is a stationary point of (4.5), then either it is the trivial null equilibrium, or there exists some $k_0 \geq 0$ such that

$$\begin{cases} \forall k < k_0, \quad u_k^* = 0, \\ \forall k \geq k_0, \quad u_k^* = U^* p_k^* = (1-\mu-k_0 s)e^{-\mu/s}\dfrac{(\mu/s)^{k-k_0}}{(k-k_0)!}. \end{cases} \tag{4.7}$$

Thus, at the equilibrium, the population size is $U^* = 1-\mu-k_0 s$, and the number of mutations has a Poisson distribution with parameter μ/s, shifted by k_0, where k_0 is the number of mutations of the best class.

Recall that a travelling wave solution should connect two stationary points of (4.5). From the above calculation, we conclude that equation (4.2) has at most two different types of travelling waves. Travelling waves that connect the trivial null equilibrium with a non-trivial equilibrium of the form (4.7). Travelling waves connecting two equilibria of the form (4.7) for different values of k_0. The former travelling wave corresponds to the expansion of a population in an empty available habitat. We call it a *population travelling wave*. The latter wave corresponds to the invasion of fitter individuals in a region where the ratchet has clicked. The total population size remains almost constant, but the genetic composition of the population shifts from one Poisson equilibrium to the other. We call it a *genetic travelling wave*. Let us now study these two kinds of waves separately.

Population travelling wave. We first prove the existence of population travelling waves. We can write (4.2) in terms of the total population size and of the genetic composition. Equation (4.2) is then equivalent to

$$\begin{cases} \partial_t U = m\partial_{xx}U + rU(BU+1)(1--s\bar{p}-U), \quad p_{-1}\equiv 0, \\ \forall k\geq 0, \quad \partial_t p_k = m(\partial_{xx}p_k + 2\partial_x \log(U)\partial_x p_k) \\ \qquad\qquad\qquad\qquad\qquad +r(BU+1)\Big(sp_k(\bar{p}-k)+\mu(p_{k-1}-p_k)\Big). \end{cases} \tag{4.8}$$

Suppose that the initial genetic composition is Poisson with parameter μ/s for all $x\in\mathbb{R}$. Then, as the Poisson distribution is a stationary point of (4.5), $(p_k)_{k\geq 0}$ remains Poisson for all x, t, and the equation for U now reads

$$\partial_t U = m\partial_{xx}U + rU(BU+1)(1-\mu-U). \tag{4.9}$$

Up to a scaling in time and space, the above equation has already been considered in [27, 96], and we know from [96] that it admits a travelling wave solution for all speeds $c\geq c_0$, where c_0 is given by

$$c_0 = \begin{cases} 2\sqrt{mr(1-\mu)} & \text{if } B \leq \dfrac{2}{1-\mu} \\ \sqrt{\dfrac{mr}{2B}}(B(1-\mu)+2) & \text{if } B \geq \dfrac{2}{1-\mu}. \end{cases} \tag{4.10}$$

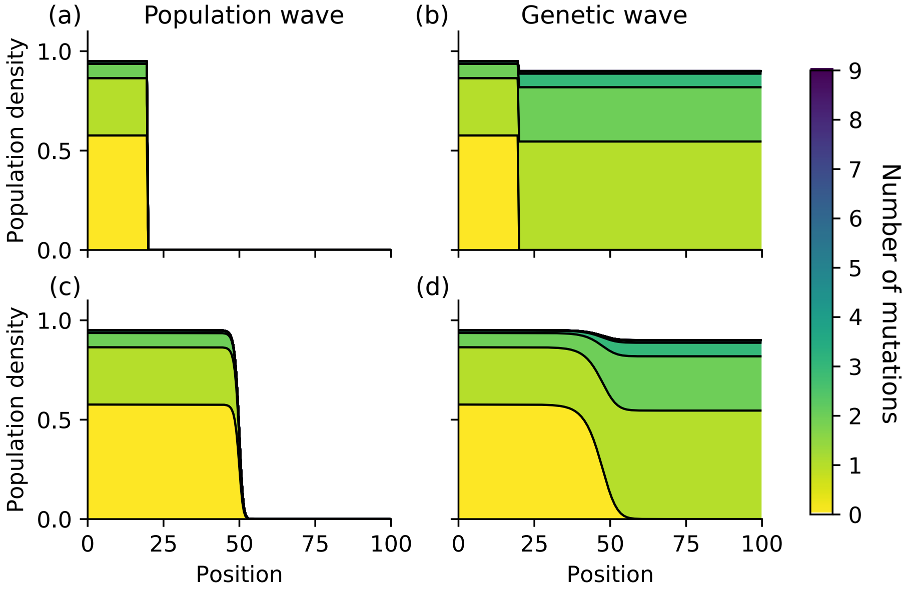


Figure 4.3: Simulation of (4.2) with different initial conditions. (a, c) Population travelling wave at $t = 0$ (a) and $t = 50$ (c). (b, d) Genetic travelling wave at $t = 0$ (b) and $t = 250$ (c); the initial condition is of the form (4.11) for $x_0 = 20$. Parameter values are $r = 1$, $m = 0.1$, $s = 0.05$, $\mu = 0.025$, $B = 0$.

Thus, if $\widehat{U}$ is the wave form of a travelling wave solution with speed c of (4.9), then

$$\forall k \geq 0, \forall x \in \mathbb{R}, \forall t \geq 0, \quad u_k(x,t) = e^{-\mu/s}\frac{(\mu/s)^k}{k!}\widehat{U}(x-ct)$$

is a population travelling wave solution to (4.2). In this case, for $B \leq 2/(1-\mu)$, the population wave is pulled, as it has the same minimal speed as the linearized version of (4.2), while for $B \geq 2/(1-\mu)$ the wave is pushed. In addition to the existence of travelling wave solutions to (4.9), there exist several results concerning the convergence to these travelling waves for various initial conditions, see e.g. Theorem 4.1 and Theorem 4.3 from [2]. These results can be directly adapted to the solutions of (4.2), started from a Poisson genetic composition.

As a remark, the above population travelling wave connects the null equilibrium with the equilibrium (4.7) for $k_0 = 0$. In a similar way we can find a population travelling wave for all $k_0 \geq 0$, the corresponding wave speed is obtained by replacing the term μ by $\mu + k_0 s$.

Genetic travelling wave. We simulated numerically equation (4.2) with initial condition

$$\forall x \in \mathbb{R}, \forall k \geq 0, \quad u_k(x,0) = \begin{cases} (1-\mu)e^{-\mu/s}\dfrac{(\mu/s)^k}{k!} & \text{if } x \leq x_0 \\ (1-\mu-s)e^{-\mu/s}\dfrac{(\mu/s)^{k-1}}{(k-1)!} & \text{if } x > x_0, \end{cases} \tag{4.11}$$

for some $x_0 \in \mathbb{R}$, see Figure 4.3. The population is initially divided into two regions, one towards the positive reals where the fittest individuals carry one deleterious mutation, the other towards the negative reals where the fittest individuals carry no mutation. The initial condition of the former region approximates the state of the population after one click of the ratchet, while that of the latter region approximates the state of the population when no click has occured. Thus, equation (4.11) corresponds to the situation where fit individuals from the bulk invade a region where the ratchet has clicked once. A travelling wave rapidly forms at the onset of the simulation. Such a wave connects the equilibrium given by (4.7) with $k_0 = 0$, to that with $k_0 = 1$, and corresponds to a genetic travelling wave.

We do not prove the existence of such a wave. Nevertheless, we are able to give an upper bound on the speed at which individuals with no mutations spread. Let us suppose that the total population size U is non-increasing in space, as observed in the simulations. Then, the following bounds would hold,

$$U_* := 1 - \mu - s \leq U \leq U^* := 1 - \mu.$$

Using these bounds in the equation for u_0 we obtain

$$\begin{aligned} \forall x \in \mathbb{R}, \forall t \geq 0, \quad \partial_t u_0 &\leq m \partial_{xx} u_0 + r u_0 (B U^* + 1)(1 - U_* - \mu) \\ &\leq m \partial_{xx} u_0 + s r u_0 (B(1-\mu) + 1). \end{aligned}$$

A classical comparison argument (see Proposition 2.1 in [2]) now shows that u_0 is bounded above by the solution to the linear equation

$$\partial_t v = m \partial_{xx} v + rs(B(1-\mu)+1)v,$$

with initial condition $v(x,0) = \mathbb{1}_{(-\infty, x_0]}$. This linear equation can be solved explicitly, and an argument taken from [184], that we have recalled in Section 4.B, shows that if u_0 is spreading at speed c, then

$$c \leq 2\sqrt{msr(B(1-\mu)+1)}.$$

Comparing this bound with (4.10), we see that for a genetic travelling wave, $c = \mathcal{O}(\sqrt{s})$, while for a population travelling wave, $c = \mathcal{O}(1)$. As we assume weak selection, i.e., $s \ll 1$, genetic travelling waves are much slower than population travelling waves.

4.3.2 Simulations of the model

Spatial clicks at the front. In order to reproduce a range expansion, we considered a population initially at carrying capacity and mutation-selection balance, and then let it expand into an empty region of space. A typical simulation output is shown in Figure 4.4. At the start of the simulation, the population is invading the new habitat at a constant speed, forming a stochastic population travelling wave, see Figure 4.4, panel (a). Within each deme, the total population size and

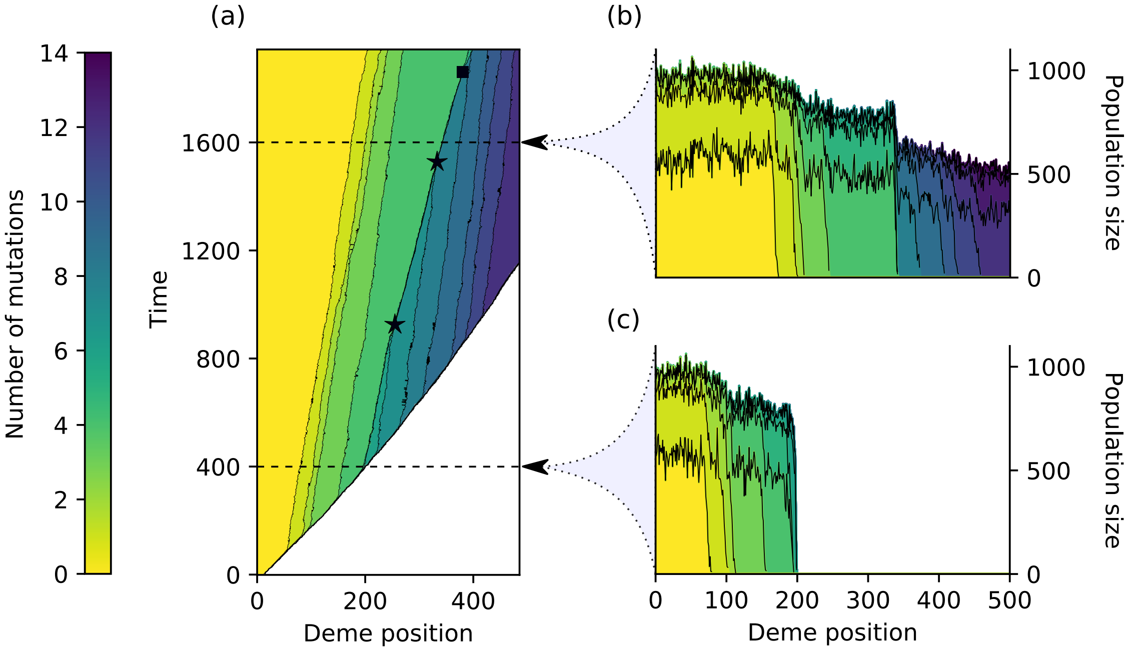


Figure 4.4: Typical simulation of the spatial Muller's ratchet. (a) Time evolution of the population; in each row, the colour gives the number of mutations of the fittest individual in the deme, i.e., the number of clicks of the ratchet in this deme. Black stars indicate genetic waves collisions, and the black square indicates an inner click of the ratchet. (b, c) Genetic composition of the population at time $t = 1600$ and $t = 400$ respectively; the number of individuals carrying a given number of mutations is given by the height of the region with the corresponding colour. Parameter values are $N = 1000$, $r = 1$, $m = 0.1$, $s = 0.05$, $\mu = 0.025$, $B = 0$.

the genetic composition fluctuate around their deterministic values derived in (4.7), see Figure 4.4 panel (c).

Eventually, due to stronger genetic drift at the front, the best type is lost from the front. The population is now divided into two spatial regions: one region towards the front, that has lost the best type; the other region towards the bulk where the best type remains present. By analogy with the non-spatial Muller's ratchet, we call this loss of the best type at the front a spatial click of the ratchet. The region where the ratchet has clicked rapidly approaches a Poisson distribution of mutations, with a slightly decreased total population size as predicted by (4.7).

The situation is now a mixture of the two initial conditions considered in Figure 4.3. The population has not yet colonized all the available demes, and it keeps on spreading according to a population travelling wave (whose speed is decreased due to the spatial click). Nevertheless, the population is now divided into two regions in a similar way to Figure 4.3 panel (b), and we see the formation of an inner genetic wave. The genetic wave is much slower than the population wave. This can be understood from the calculations of Section 4.3.1. We have shown that the speed of genetic waves scales as $\mathcal{O}(\sqrt{s})$, and thus vanishes as the selection coefficient goes to 0. Conversely, for a fixed ratio μ/s, the minimal speed of population waves, provided in equation (4.10), converges to a positive limit as s goes to 0. Thus, as we consider $s \ll 1$, we expect genetic waves to be much slower than population waves, as observed.

After the first spatial click of the ratchet, the population returns to its original selection-mutation balance, except that each individual now bears an additional mutation. Eventually a new spatial click of the ratchet will occur, and subsequent clicks will occur repeatedly during the expansion, see Figure 4.4. We can interpret these results in terms of expansion load. Recall that the expansion load refers to the additional loss in fitness due to range expansion. Initially, the population is at mutation-selection balance, and has a mutational load μ/s. After k clicks of the ratchet, the mean number of mutations at the front is $\mu/s+k$. Thus in our context, the expansion load is given by the number of spatial clicks that the population has experienced. In order to quantify the speed at which expansion load is building, we need to compute the rate at which spatial clicks of the ratchet happen in the population.

Bulk dynamics. After the ratchet has clicked several times, the bulk is divided into regions where the number of mutations of the fittest individuals corresponds to the number of clicks that the region has experienced. Each region has a fitness advantage compared to the adjacent region located towards the front, but has a fitness disadvantage compared to that towards the bulk. Thus each region is able to move forward even as it is being chased, resulting in a sequence of genetic travelling waves, see Figure 4.4 panel (a). We are interested in the dynamics of these genetic waves.

Each wave separates two regions that have accumulated distinct numbers of mutations, and thus have distinct mean fitness. The speed of the wave increases with the fitness difference between these regions. After a single click of the ratchet, this fitness difference is s. All waves that separate regions where the ratchet has clicked once spread at the same average speed, leading to the parallel genetic waves observed in Figure 4.4 panel (a). However, we observe that “double clicks” of the ratchet occur: the best and second best class of individuals can be lost simultaneoulsy from the front, for example this is the case at $t = 400$ in Figure 4.4. In this case, the fitness difference between the two sectors resulting from the click is $2s$, and the corresponding genetic wave spreads faster. It is able to catch up the next genetic waves, leading to wave collisions as indicated by the black stars in Figure 4.4. When two waves collide, the fitness difference between the two regions separated by the resulting wave increases, and thus the wave speeds up.

Interestingly, after several wave collisions, we observe that a genetic wave can split into two waves, as is indicated by the black square in Figure 4.4 (see also Figure 4.8 for an example of simulation where this split occurs earlier). This corresponds to an “inner” click of the ratchet: the best class of the genetic wave is lost from the front (of the genetic wave). The mechanism leading to such an inner click is the same as that leading to spatial clicks at the front of the population wave. Fit individuals at the front of a genetic wave have a high growth rate as they compete with individuals that have accumulated many deleterious mutations. If an individual from the second best class gets ahead, it can rapidly grow a large subpopulation that will further expand, creating a new genetic wave. We expect such an event to occur at a higher rate when the fitness difference between the sectors separated by the genetic wave is large.

The picture that emerges from the analysis of the dynamics of the bulk is the following. We can think of the bulk as a branching-coalescing system of particles: each genetic wave corresponds to a particle located at the front of that wave. Then the system has the following dynamics. Each particle follows a random motion, with an average speed towards the positive reals that increases with the fitness difference of the regions it separates. When two particles meet, i.e., when two genetic waves collide, they merge, and the resulting particle speeds up. Finally, when an inner click of the ratchet occurs, a new particle is created, and the two daughter particles are slower than their mother. Such a branching event occurs at a rate that increases with the fitness difference of the regions separated by the particle.

4.3.3 Impact of the Allee effect on the expansion load

We now aim to study the impact of an Allee effect on the expansion load, i.e., on the click rate in our context. Analysis of the formation of expansion load requires us to take into account genetic drift. Recall that a generator calculation suggests that a good approximation of our model that retains finite size fluctuations is given by equation (4.3). The parameter that controls the strength of the Allee effect is B. However, we see from equation (4.3) that increasing B also increases the noise term, and hence increases genetic drift. Thus the parameter B has two antagonistic effects on the click rate: on the one hand it should reduce the click rate by increasing the strength of the Allee effect, and shifting the nature of the wave from pulled to pushed; on the other hand, it increases the click rate by reinforcing genetic drift, and hence gene surfing. In order to disantangle these two effects, we will study the impact of B on the scaling with N of the click rate. If the pulled/pushed nature of the wave does not impact expansion load, increasing B should only increase the strength of the drift and we expect a similar scaling of the click rate with N for different values of B.

A direct computation of the click rate from (4.3) is not feasible. We thus used simulations to assess the impact of B on the click rate. Starting from an initial condition similar to Figure 4.4, we let the population expand, and record the time T_1 and spatial location $n_1^{\max}(T_1) - n_1^{\max}(0)$ of the first spatial click of the ratchet, see Section 4.2 for a precise definition of these quantities. Both T_1 and $n_1^{\max}(T_1) - n_1^{\max}(0)$ should be inversely related to the click rate. Figure 4.5 panel (a) shows the time of this first click, averaged over 5000 simulation replicates, for various values of B and N. We have performed a similar analysis for the second and third clicks of the ratchet. The results, shown in Figure 4.9, are qualitatively similar.

First, notice that for each fixed value of B, the mean time to the first click increases with N, i.e., the ratchet is slower for large population sizes. This is in agreement with our intuitive understanding of the ratchet, since increasing N reduces the strength of genetic drift and hence reduces gene surfing. Second, for a fixed value of N, increasing B either speeds up the ratchet if N is low, or slows it down if N is large. This observation can be explained intuitively as follows. Recall that B has two antagonistic effects on the click rate: increasing the genetic drift, and increasing the gene flow from the bulk to the front. For low values of N, the population size in the bulk is low, and the gene flow from the bulk restores the genetic diversity at the front less efficiently than for large values of N. Thus increasing the gene flow from the bulk to the front has a larger impact on the click rate for large values of N. For low values of N, the increase of genetic drift with B prevails, while the converse holds for large values of N.

Let us now consider the scaling of the time to the first click, T_1, with N. First, notice that for any value of B, T_1 scales faster than a power law with N. It is clear that N has more impact on T_1 for pushed waves, i.e., for large B, than for pulled waves. In the pulled case, T_1 increases with N very slowly, and the rate of ratchet is only slightly changed by the population size. Conversely, in the pushed case, N has a drastic effect on T_1: the ratchet clicks very fast for low N, but we see almost no click of the ratchet for large N.

We thus conclude that the Allee effect has a large impact on the click rate. For small values of B, the population size has little impact on the building of the expansion load. This reflects the fact that the dynamics is mostly determined by the few individuals in the front, that are almost insensitive to the change in the carrying capacity in the bulk. Conversely, for large values of B, the dynamics of the wave is determined by an intermediate region between the bulk and the front. Increasing N reduces the genetic drift in this region and leads to the large effect of N on T_1 observed in Figure 4.5.

4.4 Discussion

Expansion load originates from the strong genetic drift induced by the low population size at the edge of an expanding population. From a modelling perspective, demography, spatial structure, stochasticity and selection are minimal ingredients to account for expansion load. Each of these features is known to make mathematical treatment harder, and thus building a tractable model for expansion load is challenging. In this work we proposed a model similar in spirit to [174], but with two major differences: we greatly simplified the genetic structure of the population to that of a Muller's ratchet, and we introduced an Allee effect in the population, tuned by the parameter B. This simplification allowed us to prove rigorous results for the deterministic scaling of the model, however an analysis of the stochastic scaling (4.3) where the building of an expansion load occurs remained out of reach.

Among other factors that are known to impact the genetics of range expansion, such as density dependent migration [28] or long distance dispersal [66], we have

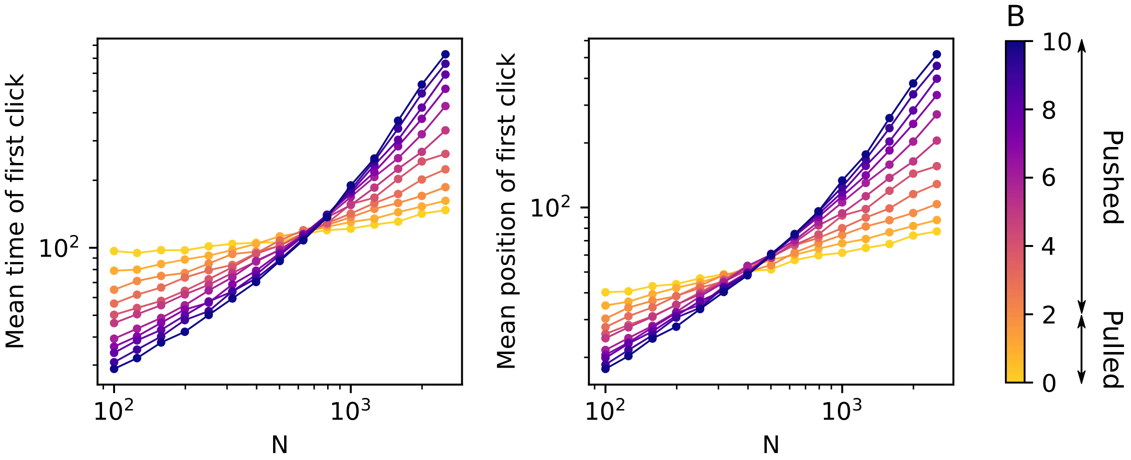


Figure 4.5: Scaling of the click rate and click position with N. The left plot shows the value of T_1, and the right plot the value of $n_1^{\max}(T_1) - n_1^{\max}(0)$, see Section 4.2 for the definitions. Each point is averaged over 5000 simulations. The parameter values are $r = 1$, $\mu = 0.01$, $s = 0.02$, $m = 0.1$.

focused here on the impact of an Allee effect on expansion load. It is already understood from several studies that an Allee effect impedes gene surfing [27, 184, 83]. In agreement with these findings and our intuitive expectations, we have shown that adding an Allee effect to the population slows down the rate at which the ratchet clicks for large population size, and thus reduces the expansion load.

However, [184, 83] predict a sharp qualitative difference between pulled and pushed waves. This disagrees with Figure 4.5 where increasing B continuously changes the scaling of the click rate with N. Nevertheless, note that the results in [184, 83] were obtained in a deterministic setting, and that stochasticity has a tendency to smoothen such transitions. In a stochastic setting, [27] have fitted a power law to the rate of genetic diversity loss in an expanding population, as a function of N. They predicted that the exponent of this power law remains constant outside of the parameter region $B \in [2, 4]$, see their Figure 4 and Figure 5. Again, in our Figure 4.5 we see a change in the scaling on the entire range of B. This discrepancy might be explained by the coupled effect B has on the nature of the wave and the genetic drift.

Moreover, for low N, we observe that increasing B increases the rate of the ratchet. This originates from the complex interaction between Allee effect and genetic drift in our model. More generally, genetic drift depends on the rate at which birth and death events occur in the population. Changing the strength of the Allee effect should modify these rates, and we can expect the Allee effect to interact with genetic drift for a large class of models. The specific form of this interaction should depend on the details of the microscopic model under consideration and the way the Allee effect is implemented. Therefore, we believe that our results cannot be directly transposed to other models of population expansion incorporating an Allee effect. The impact of the Allee effect on the expansion load should depend in a crucial way on its interplay with genetic drift, which is a model dependent feature. In order to illustrate this, we have reproduced the results of Figure 4.5 for an alternative parametrization of our model, see Section 4.C and Figure 4.7. We observe that the results are qualitatively very different. Increasing the strength of the Allee effect reduces the rate of the ratchet for all N in this new parametrization.

Our model also relates to the vast literature on Muller's ratchet, see e.g. [149] and references therein. The effect of spatial structure on the dynamics of Muller's ratchet has already been investigated in [110]. They concluded that, for a fixed total population size, subdividing the population into smaller habitats reinforces Muller's ratchet, since genetic drift is enhanced in each single habitat. The setting we consider is different. The total population size is not fixed as we allow new subpopulations to grow in empty demes. We find that space has two effects on the ratchet. In the bulk of the population, we do not observe clicks of the ratchet. Spatial structure has a stabilizing effect: if the ratchet clicks in one deme, the best type can be reintroduced by migration from the adjacent demes. Conversely, in the front spatial structure causes low population size and thus speeds up the ratchet. Overall, our study shows that spatial structure can interact in a non-trivial way with Muller's ratchet.

We have considered the expansion of a population in a linear one-dimensional habitat. Multiple studies have also been concerned with range expansion in two dimensions, using microbial growth experiments on petri dishes [100, 98, 131, 89] or simulations [62, 172]. The typical set up of these studies is to place a drop containing two labelled strains in the centre of a petri dish, and to let them expand. The colony is rapidly separated into sectors where only one of the two strains is present, and the other is absent, see for example Figure 1 in [89]. These studies have examined the dynamics of these sectors, especially when there exists a fitness difference between the two strains. They have established that the boundary between two sectors should move towards the strain with the lower fitness, and gave an expression for the speed of the boundary in terms of the fitness difference [131, 100].

In the context of the spatial Muller's ratchet, the major expected difference between one and two dimensions is the following. In one dimension, once the ratchet has clicked, best type individuals are trapped in the bulk of the population. The only way to restore the fitness at the front is that the genetic wave of fit individuals catches up with the population wave. We know that this is extremely unlikely, because the genetic wave is much slower than the population wave. In two dimensions, the front is a one-dimensional curve, and a click of the ratchet only removes best type individuals from a small part of it. The remaining best type individuals have a fitness advantage compared to individuals in demes where the ratchet has clicked, and according to the aforementioned studies, they should be able to remove the unfit individuals from the front. Thus, in two dimensions, a click of the ratchet does not irredeemably trap best type individuals in the bulk, and fitness should be restored by migration of fit individuals from parts of the front where the ratchet has not clicked.

In order to assess these predictions and to compare the behavior of our model to previous studies, we have simulated the spatial Muller's ratchet on a two dimen-

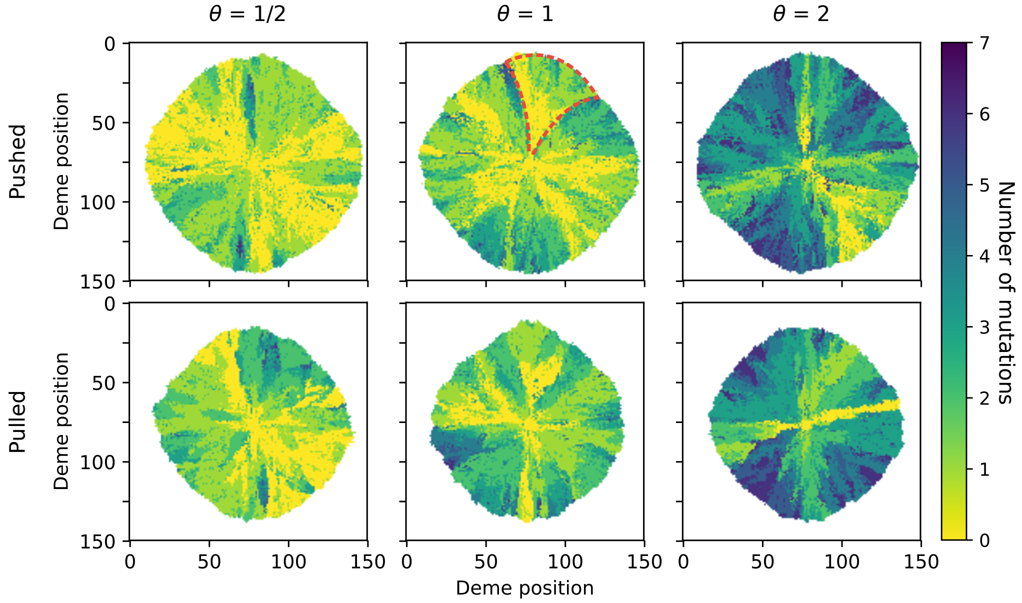


Figure 4.6: Two dimensional simulations, shown at time $t = 150$. All simulations are realized with $N = 300$, $\mu = 0.02$, $r = 1$, $m = 0.1$. We set $B = 0$ in the pulled case, and $B = 3$ in the pushed case. The value of the selection coefficient is chosen in $s \in \{0.01, 0.02, 0.04\}$ to obtain the various values of μ/s. The colour indicates the number of mutations of the best type at the deme colonization, see Section 4.2. The red dashed line indicates a funnel-shaped sector.

sional lattice. Demes are now indexed by $\mathbb{Z}^2$, the reproduction rules within each deme remain the same, but at each migration event individuals now choose one of the four adjacent demes with equal probability. A key difference between our model and the microbial growth experiments is that the cells are non-motile and unable to migrate. In the spatial ratchet, the bulk of the population is dynamical, and fit individuals slowly expand according to genetic waves. In the growth experiments, cells in the bulk remain at their initial location, and the observed patterns correspond to a "frozen record" of the front at the time of colonization. In order to carry out the comparison between our model and the existing studies, we have depicted in Figure 4.6 the number of mutations in each deme at its colonization time, see Section 4.2 for the precise definition of this quantity. We emphasize the fact that this is *not* the state of the entire population at the end of the simulation: many sectors will have been taken over by fit individuals from the bulk and their shapes would not be comparable to that of the sectors obtained in microbial growth experiments. For comparison, we have shown the state of the population in Figure 4.10.

As in the one-dimensional case, we observe in Figure 4.6 clicks of the ratchet leading to the formation of sectors with lower mean fitness. An achievement of the microbial growth experiments on petri dishes is to link the shape of these sectors to their relative fitness. If the strains that are placed on the petri dish have the same fitness, then the sectors should be "cone-shaped": the boundary between two sectors is wandering due to stochastic effects but does not have a preferential direction. Conversely, if one strain is fitter than the other, the sectors of the fitter strain should have a typical "funnel" shape, see for instance Figure 4 in [89] for examples of these two shapes. Most sectors in Figure 4.6 have a cone shape, indicating that the expansion is nearly neutral. Selection is too weak in these simulations to allow fit individuals to efficiently remove unfit individuals from the front. Nevertheless, we have indicated by a red dashed line a sector that has the typical shape of a selectively advantaged strain. Notice that the regions adjacent to this sector have experienced multiple clicks of the ratchet, and thus that the fitness advantage of this sector is large.

Apparently, in the parameter region we have considered, selection is not strong enough to reverse spatial clicks of the ratchet and efficiently restore fitness at the front. The dynamics of the sectors is nearly neutral, and we do not observe any major difference from the one-dimensional case. The pushed/pulled nature of the wave seems to have the same qualitative effect on the clicks of the ratchet as in the one-dimensional case. A better understanding of the two-dimensional case would require a more quantitative and thorough investigation, which goes beyond the scope of the present work.

During a range expansion, the front can accumulate mutations leading to an expansion load, but individuals in the bulk do not bear this additional burden. Thus, at each location in space, fitness should be slowly recovered through migration of fit individuals from the bulk: expansion load is a transient phenomenon [84, 172]. In the spatial Muller's ratchet we have a clear quantification of the rate of this fitness recovery. Fit individuals take over the population through inner genetic waves, with a speed proportional to the square root of their selective advantage. As discussed previously, on the one hand the speed of genetic waves can increase due to wave collisions. On the other hand, the speed of the population wave is decreased by the successive clicks of the ratchet. It is natural to ask whether the population wave is eventually caught up by a genetic wave. More generally, it would be interesting to study the long-term behavior of the spatial Muller's ratchet. A possible starting point is to approximate the dynamics of the bulk of the population by a branching-coalescing particle system as described previously, and to study the asymptotic behavior of this simplified system.

The nature of population travelling waves changes from pulled to pushed at the critical value of $B = 2/(1-\mu)$. It is interesting to ask whether such a transition occurs for genetic waves. From (4.8), the per-capita growth rate of p_0, the frequency of individuals without mutations, is

$$s(BU+1)(\bar{p}-\mu/s).$$

In a genetic travelling wave, the total population size U is almost constant (it ranges from $1-\mu$ to $1-\mu-s$). As we have the constraint $\sum p_i = 1$, we see that, roughly speaking, the maximal per-capita growth rate of p_0 is achieved for lower values p_0. In our intuitive definition of pulled and pushed waves, this corresponds to the pulled case. In our model, genetic waves are always pulled, regardless of the value of B.

This essentially comes from the fact that we have considered a haploid population. The dynamics of the frequency p of a gene with fitness advantage s in a haploid population with local migration is given by the classical Fisher-KPP [70, 130] equation

$$\partial_t p = \partial_{xx} p + sp(1-p)$$

which is the archetypical example of a pulled wave. Considering a diploid population where homozygotes have fitness $1+2\alpha s$ and heterozygotes a fitness $1+(\alpha-1)s$ leads to a special case of the so-called Allen-Cahn equation [13]

$$\partial_t p = \partial_{xx} p + sp(1-p)(2p+\alpha-1)$$

that displays a transition from pulled to pushed waves when varying the parameter α. One could thus obtain pushed genetic waves by considering a diploid population with heterozygote advantage [16].

References for Chapter 4

[2] D. G. Aronson and H. F. Weinberger. Nonlinear diffusion in population genetics, combustion, and nerve pulse propagation. *Partial Differential Equations and Related Topics*. Springer Berlin Heidelberg, 1975, 5–49.

[13] N. H. Barton. The dynamics of hybrid zones. *Heredity* **43** (1979), 341–359.

[14] N. H. Barton, F. Depaulis, and A. M. Etheridge. Neutral evolution in spatially continuous populations. *Theoretical Population Biology* **61** (2002), 31–48.

[16] N. H. Barton and M. Turelli. Spatial waves of advance with bistable dynamics: Cytoplasmic and genetic analogues of allee Effects. *The American Naturalist* **178** (2011), E48–E75.

[27] G. Birzu, O. Hallatschek, and K. S. Korolev. Fluctuations uncover a distinct class of traveling waves. *Proceedings of the National Academy of Sciences* **115** (2018), E3645–E3654.

[28] G. Birzu, S. Matin, O. Hallatschek, and K. S. Korolev. Genetic drift in range expansions is very sensitive to density feedback in dispersal and growth (2019). arXiv: `1903.11627`.

[30] D. S. Boukal and L. Berec. Single-species models of the Allee effect: Extinction boundaries, sex ratios and mate encounters. *Journal of Theoretical Biology* **218** (2002), 375–394.

[47] R. Do, D. Balick, H. Li, I. Adzhubei, S. Sunyaev, and D. Reich. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nature Genetics* **47** (2015), 126–131.

[53] R. Durrett and W.-T. Fan. Genealogies in expanding populations. *The Annals of Applied Probability* **26** (2016), 3456–3490.

[54] C. A. Edmonds, A. S. Lillie, and L. L. Cavalli-Sforza. Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences* **101** (2004), 975–979.

[57] A. M. Etheridge, P. Pfaffelhuber, and A. Wakolbinger. How often does the ratchet click? Facts, heuristics, asymptotics. London Mathematical Society Lecture Note Series. Cambridge University Press, 2009, 365–390.

[62] L. Excoffier, M. Foll, and R. J. Petit. Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics* **40** (2009), 481–501.

[63] L. Excoffier and N. Ray. Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution* **23** (2008), 347–351.

[64] A. Eyre-Walker and P. D. Keightley. The distribution of fitness effects of new mutations. *Nature Reviews Genetics* **8** (2007), 610–618.

[66] J. Fayard, É. K. Klein, and F. Lefèvre. Long distance dispersal and the fate of a gene from the colonization front. *Journal of Evolutionary Biology* **22** (2009), 2171–2182.

[67] J. Felsenstein. The evolutionary advantage of recombination. *Genetics* **78** (1974), 737–756.

[70] R. A. Fisher. The Wave of advance of advantageous genes. *Annals of Eugenics* **7** (1937), 355–369.

[78] F. Foutel-Rodier and A. M. Etheridge. The spatial Muller's ratchet: Surfing of deleterious mutations during range expansion. *Theoretical Population Biology* **135** (2020), 19–31.

[83] J. Garnier, T. Giletti, F. Hamel, and L. Roques. Inside dynamics of pulled and pushed fronts. *Journal de Mathématiques Pures et Appliquées* **98** (2012), 428–449.

[84] K. J. Gilbert, S. Peischl, and L. Excoffier. Mutation load dynamics during environmentally-driven range shifts. *PLOS Genetics* **14** (2018), 1–18.

[85] K. J. Gilbert, N. P. Sharp, A. L. Angert, G. L. Conte, J. A. Draghi, F. Guillaume, A. L. Hargreaves, R. Matthey-Doret, and M. C. Whitlock. Local adaptation interacts with expansion load during range expansion: Maladaptation reduces expansion load. *The American Naturalist* **189** (2017), 368–380.

[87] S. C. González-Martínez, K. Ridout, and J. R. Pannell. Range expansion compromises adaptive evolution in an outcrossing plant. *Current Biology* **27** (2017), 2544–2551.e4.

[88] E. Graciá, F. Botella, J. D. Anadón, P. Edelaar, D. J. Harris, and A. Giménez. Surfing in tortoises? Empirical signs of genetic structuring owing to range expansion. *Biology Letters* **9** (2013), 20121091.

[89] M. Gralka, F. Stiewe, F. Farrell, W. Möbius, B. Waclaw, and O. Hallatschek. Allele surfing promotes microbial adaptation from standing variation. *Ecology Letters* **19** (2016), 889–898.

[96] K.-P. Hadeler and F. Rothe. Travelling fronts in nonlinear diffusion equations. *Journal of Mathematical Biology* **2** (1975), 251–263.

[97] J. Haigh. The accumulation of deleterious genes in a population—Muller's Ratchet. *Theoretical Population Biology* **14** (1978), 251–267.

[98] O. Hallatschek, P. Hersen, S. Ramanathan, and D. R. Nelson. Genetic drift at expanding frontiers promotes gene segregation. *Proceedings of the National Academy of Sciences* **104** (2007), 19926–19930.

[99] O. Hallatschek and D. R. Nelson. Gene surfing in expanding populations. *Theoretical Population Biology* **73** (2008), 158–170.

[100] O. Hallatschek and D. R. Nelson. Life at the front of an expanding population. *Evolution* **64** (2010), 193–206.

[108] B. M. Henn, L. L. Cavalli-Sforza, and M. W. Feldman. The great human expansion. *Proceedings of the National Academy of Sciences* **109** (2012), 17758–17764.

[110] K. Higgins and M. Lynch. Metapopulation extinction caused by mutation accumulation. *Proceedings of the National Academy of Sciences* **98** (2001), 2928–2933.

[129] S. Klopfstein, M. Currat, and L. Excoffier. The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution* **23** (2006), 482–490.

[130] A. Kolmogoroff, I. Petrovsky, and N. Piscounoff. Études de l'équation avec croissance de la quantité de matière et son application à un problème biologique. *Moscow University Bulletin Of Mathematics* **1** (1937), 1–25.

[131] K. S. Korolev, M. J. I. Müller, N. Karahan, A. W. Murray, O. Hallatschek, and D. R. Nelson. Selective sweeps in growing microbial colonies. *Physical Biology* **9** (2012), 026008.

[132] A. M. Kramer, B. Dennis, A. M. Liebhold, and J. M. Drake. The evidence for Allee effects. *Population Ecology* **51** (2009), 341–354.

[149] L. Loewe. Quantifying the genomic decay paradox due to Muller's ratchet in human mitochondrial DNA. *Genetical Research* **87** (2006), 133–159.

[160] H. J. Muller. The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **1** (1964), 2–9.

[161] C. Müller and R. Tribe. Stochastic P.D.E.'s arising from the long range contact and long range voter processes. *Probability Theory and Related Fields* **102** (1995), 519–545.

[170] S. Peischl, I. Dupanloup, L. Bosshard, and L. Excoffier. Genetic surfing in human populations: from genes to genomes. *Current Opinion in Genetics & Development* **41** (2016), 53–61.

[171] S. Peischl, I. Dupanloup, A. Foucal, M. Jomphe, V. Bruat, J.-C. Grenier, A. Gouy, K. J. Gilbert, E. Gbeha, L. Bosshard, E. Hip-Ki, M. Agbessi, A. Hodgkinson, H. Vézina, P. Awadalla, and L. Excoffier. Relaxed selection during a recent human expansion. *Genetics* **208** (2018), 763–777.

[172] S. Peischl, I. Dupanloup, M. Kirkpatrick, and L. Excoffier. On the accumulation of deleterious mutations during range expansions. *Molecular Ecology* **22** (2013), 5972–5982.

[173] S. Peischl and L. Excoffier. Expansion load: recessive mutations and the role of standing genetic variation. *Molecular Ecology* **24** (2015), 2084–2094.

[174] S. Peischl, M. Kirkpatrick, and L. Excoffier. Expansion load and the evolutionary dynamics of a species range. *The American naturalist* **185** (2015), E81–E93.

[184] L. Roques, J. Garnier, F. Hamel, and É. K. Klein. Allee effect promotes diversity in traveling waves of colonization. *Proceedings of the National Academy of Sciences* **109** (2012), 8828–8833.

[200] Y. B. Simons, M. C. Turchin, J. K. Pritchard, and G. Sella. The deleterious mutation load is insensitive to recent population history. *Nature Genetics* **46** (2014), 220–224.

[203] P. A. Stephens, W. J. Sutherland, and R. P. Freckleton. What is the Allee effect? *Oikos* **87** (1999), 185–190.

[204] A. N. Stokes. On two types of moving front in quasilinear diffusion. *Mathematical Biosciences* **31** (1976), 307–315.

[206] C. M. Taylor and A. Hastings. Allee effects in biological invasions. *Ecology Letters* **8** (2005), 895–908.

[211] J. M. J. Travis, T. Münkemüller, O. J. Burton, A. Best, C. Dytham, and K. Johst. Deleterious mutations can surf to high densities on the wave front of an expanding population. *Molecular Biology and Evolution* **24** (2007), 2334–2343.

[213] W. Van Saarloos. Front propagation into unstable states. *Physics Report* **386** (2003), 29–222.

[218] Y. Willi, M. Fracassetti, S. Zoller, and J. Van Buskirk. Accumulation of mutational load at the edges of a species range. *Molecular Biology and Evolution* **35** (2018), 781–791.

Appendices for Chapter 4

4.A Generator computations

Deterministic scaling. Let $(n_{i,k}(t);\, t \in \mathbb{R}_+,\, i \in \mathbb{Z},\, k \geq 0)$ be the process described in Section 4.2, and recall that we have defined the renormalized process as

$$\forall t \geq 0,\, \forall i \in \mathbb{Z},\, \forall k \geq 0, \quad u_k\Big(\frac{i}{L}, t\Big) = \frac{n_{i,k}(t)}{N}.$$

Let $\varphi\colon \mathbb{R} \to \mathbb{R}$ be a twice-differentiable function with compact support, and set

$$\langle u_k, \varphi \rangle = \frac{1}{L} \sum_{i \in \mathbb{Z}} u_k(i/L, t) \varphi(i/L).$$

Recall that U stands for the total renormalized population size

$$U = \sum_{k \geq 0} u_k.$$

We expect a deterministic limit as $N \to \infty$, so in order to identify the limit of the u_k under this scaling, it is enough to consider the generator applied to functions of this form. If G_N is the generator of u_k, then

$$\begin{aligned} G_N \langle \cdot, \varphi \rangle \big(u_k\big) = \frac{1}{NL} \sum_{i \in \mathbb{Z}} \Big(& \frac{m_N}{2} n_{i,k} \Big[\varphi\Big(\frac{i+1}{L}\Big) + \varphi\Big(\frac{i-1}{L}\Big) - 2\varphi\Big(\frac{i}{L}\Big) \Big] \\ & + r(1-s)^k n_{i,k} (BU+1)(1-\mu) \varphi\Big(\frac{i}{L}\Big) \\ & - r n_{i,k} (BU+1) U \varphi\Big(\frac{i}{L}\Big) \\ & + r(1-s)^{k-1} n_{i,k-1} (BU+1) \mu \varphi\Big(\frac{i}{L}\Big) \Big). \end{aligned}$$

Thus we see that, provided u_k is converging and $m_N / L^2 \to m$, the above quantity converges to

$$m \int_{\mathbb{R}} u_k(x) \varphi''(x) \mathrm{d}x + \int_{\mathbb{R}} r\Big((BU+1) u_k ((1-s)^k - U) + \mu ((1-s)^{k-1} u_{k-1} - (1-s)^k u_k) \Big) \varphi(x) \mathrm{d}x,$$

which suggests that in the limit (u_k) solves

$$\partial_t u_k = m \partial_{xx} u_k + r(BU+1)(u_k((1-s)^k - U) + \mu((1-s)^{k-1} u_{k-1} - (1-s)^k u_k)).$$

Stochastic scaling. Consider a function $\varphi\colon \mathbb{R}^{\mathbb{Z}\times\mathbb{N}} \to \mathbb{R}$ that only depends on a finite number of coordinates, and suppose that φ is twice continuously differentiable. Under these assumptions, by making a Taylor expansion of φ at the point $\mathbf{u} = (u_{i,k};\, i \in \mathbb{Z},\, k \geq 0)$ and ignoring terms of order greater than $1/N^2$, we obtain the following expression for the generator $\widetilde{G}_N$ of the process with population size N

$$\begin{aligned}\widetilde{G}_N\varphi(\mathbf{u}) = \sum_{i\in\mathbb{Z}}\sum_{k\geq 0}\Bigg[&\frac{m}{2}Nu_{i,k}\Bigg(\frac{1}{N}\frac{\partial\varphi}{\partial x_{i+1,k}}(\mathbf{u}) - \frac{2}{N}\frac{\partial\varphi}{\partial x_{i,k}}(\mathbf{u}) + \frac{1}{N}\frac{\partial\varphi}{\partial x_{i-1,k}}(\mathbf{u})\\
&+ \frac{1}{2N^2}\frac{\partial^2\varphi}{\partial x_{i+1,k}^2}(\mathbf{u}) + \frac{1}{2N^2}\frac{\partial^2\varphi}{\partial x_{i-1,k}^2}(\mathbf{u}) + \frac{1}{N^2}\frac{\partial^2\varphi}{\partial x_{i,k}^2}(\mathbf{u}) - \frac{1}{N^2}\frac{\partial^2\varphi}{\partial x_{i,k}\partial x_{i-1,k}}(\mathbf{u}) - \frac{1}{N^2}\frac{\partial^2\varphi}{\partial x_{i,k}\partial x_{i+1,k}}(\mathbf{u})\Bigg)\\
&+ r(1-s)^kNu_{i,k}(B\sum_{j\geq 0}u_{i,j}+1)\Bigg((1-\mu)\Bigg(\frac{1}{N}\frac{\partial\varphi}{\partial x_{i,k}}(\mathbf{u}) + \frac{1}{2N^2}\frac{\partial^2\varphi}{\partial x_{i,k}^2}(\mathbf{u})\Bigg)\\
&+ \mu Nu_{i,k}\left(\frac{1}{N}\frac{\partial\varphi}{\partial x_{i,k+1}}(\mathbf{u}) - \frac{1}{N}\frac{\partial\varphi}{\partial x_{i,k}}(\mathbf{u}) + \frac{1}{2N^2}\frac{\partial^2\varphi}{\partial x_{i,k+1}^2}(\mathbf{u}) + \frac{1}{2N^2}\frac{\partial^2\varphi}{\partial x_{i,k}^2}(\mathbf{u}) - \frac{1}{N^2}\frac{\partial^2\varphi}{\partial x_{i,k}\partial x_{i,k+1}}(\mathbf{u})\right)\Bigg)\\
&+ rNu_{i,k}(\sum_{j\geq 0}u_{i,j})(B\sum_{j\geq 0}u_{i,j}+1)\Bigg(-\frac{1}{N}\frac{\partial\varphi}{\partial x_{i,k}}(\mathbf{u}) + \frac{1}{2N^2}\frac{\partial^2\varphi}{\partial x_{i,k}^2}(\mathbf{u})\Bigg)\Bigg].\end{aligned}$$

We suppose $m \ll r$ and that $\mu \ll 1$, so that we can neglect the mixed second order derivatives and discarding terms of $\mathcal{O}(1/N^2)$ we find that the generator of our rescaled process is approximately

$$\begin{aligned}\widetilde{G}_N\varphi(\mathbf{u}) = \sum_{i\in\mathbb{Z}}\sum_{k\geq 0}\frac{\partial\varphi}{\partial x_{i,k}}(\mathbf{u})\Bigg(&\frac{m}{2}(u_{i-1,k}+u_{i+1,k}-2u_{i,k})\\
&+ r(B\sum_{j\geq 0}u_{i,j}+1)(u_{i,k}((1-s)^k - \sum_{j\geq 0}u_{i,j}) + \mu((1-s)^{k-1}u_{i,k-1} - (1-s)^ku_{i,k})\Bigg)\\
&+ \frac{1}{2N}\frac{\partial^2\varphi}{\partial x_{i,k}^2}(\mathbf{u})ru_{i,k}((1-s)^k + \sum_{j\geq 0}u_{i,j})(B\sum_{j\geq 0}u_{i,j}+1).\end{aligned}$$

Thus, for large N, and m fixed such that $s, \mu \ll 1, m \ll r$, our process is well-approximated by the following set of stochastic differential equations,

$$\begin{aligned}\forall i \in \mathbb{Z},\, k \geq 0, \quad \mathrm{d}u_{i,k} = \Bigg(&m\frac{u_{i-1,k}+u_{i+1,k}-2u_{i,k}}{2}\\
&+ r(BU_i+1)(u_{i,k}(1-ks-U_i) + \mu(u_{i,k-1}-u_{i,k}))\Bigg)\mathrm{d}t\\
&+ \sqrt{\frac{1}{N}ru_{i,k}(1-ks+U_i)(BU_i+1)}\mathrm{d}W_{i,k},\end{aligned}$$

where $(W_{i,k};\, i \in \mathbb{Z}, k \geq 0)$ are independent Brownian motions.

Writing

$$\langle u_k, \varphi\rangle = \frac{1}{L}\sum_{i\in\mathbb{Z}}u_k(i/L)\varphi(i/L)$$

as before and speeding up time by a factor of L^2 in order to obtain a diffusive rescaling and scaling $N \mapsto LN$ (corresponding to replacing the population size in a deme by the local population density), we expect that as $L \to \infty$ we should recover the system of stochastic partial differential equations

$$\begin{aligned}\forall k \geq 0, \quad \partial_t u_k = \Big(m\partial_{xx}u_k + r(BU+1)(u_k(1-ks-U) + \mu(u_{k-1}-u_k))\Big) \\ + \sqrt{\frac{1}{N} r u_k(1-ks+U)(BU+1)}\mathrm{d}\dot{W}_k\end{aligned}$$

where $(\dot{W}_k)_{k\geq 0}$ are independent space-time white-noises, see for example [14], Section 2. We emphasize that this derivation is heuristic. Again, see [53, 161] for rigorous treatments of similar convergence results.

4.B Spread of a linear wave

Let us consider the equation

$$\partial_t v = m\partial_{xx}v + \alpha v$$

with bounded initial condition $v(x,0) = v_0(x)$. The solution to this equation is

$$\forall t \geq 0, \forall x \in \mathbb{R}, \quad v(x,t) = \frac{1}{\sqrt{4\pi m t}} \int_{\mathbb{R}} e^{-\frac{(x-y)^2}{4mt}} e^{\alpha t} v_0(y)\mathrm{d}y.$$

Following [184], let $c \geq 0$ be some speed, then

$$\begin{aligned}\forall t \geq 0, \forall x \in \mathbb{R}, \quad v(x+ct,t) &= \frac{1}{\sqrt{4\pi m t}} \int_{\mathbb{R}} e^{-\frac{(x+ct-y)^2}{4mt}} e^{\alpha t} v_0(y)\mathrm{d}y \\ &= \frac{1}{\sqrt{4\pi m t}} \int_{\mathbb{R}} e^{-\frac{(x-y)^2}{4mt}} e^{-\frac{c(x-y)}{2m}} e^{-\frac{c^2 t}{4m}} e^{\alpha t} v_0(y)\mathrm{d}y \\ &\leq \frac{1}{\sqrt{4\pi m t}} e^{\alpha t - \frac{c^2 t}{4m} - \frac{cx}{2m}} \int_{\mathbb{R}} e^{\frac{cy}{2m}} v_0(y)\mathrm{d}y.\end{aligned}$$

Thus, for $c \geq 2\sqrt{m\alpha}$,

$$\forall t \geq 0, \forall x \in \mathbb{R}, \quad v(x+ct,t) \leq \frac{1}{\sqrt{4\pi m t}} e^{-\frac{cx}{2m}} \int_{\mathbb{R}} e^{\frac{cy}{2m}} v_0(y)\mathrm{d}y.$$

Hence, provided the integral is finite and $c \geq 2\sqrt{m\alpha}$ (which is the case when v_0 is Heaviside), $v(x+ct,t)$ goes to 0 uniformly on sets of the form $[A,\infty)$ for $A \in \mathbb{R}$. This shows that the process u of Section 4.3.1 cannot converge to a travelling wave solution with speed larger than $2\sqrt{m\alpha}$, i.e., larger than $2\sqrt{msr(B(1-\mu)+1)}$ in this case.

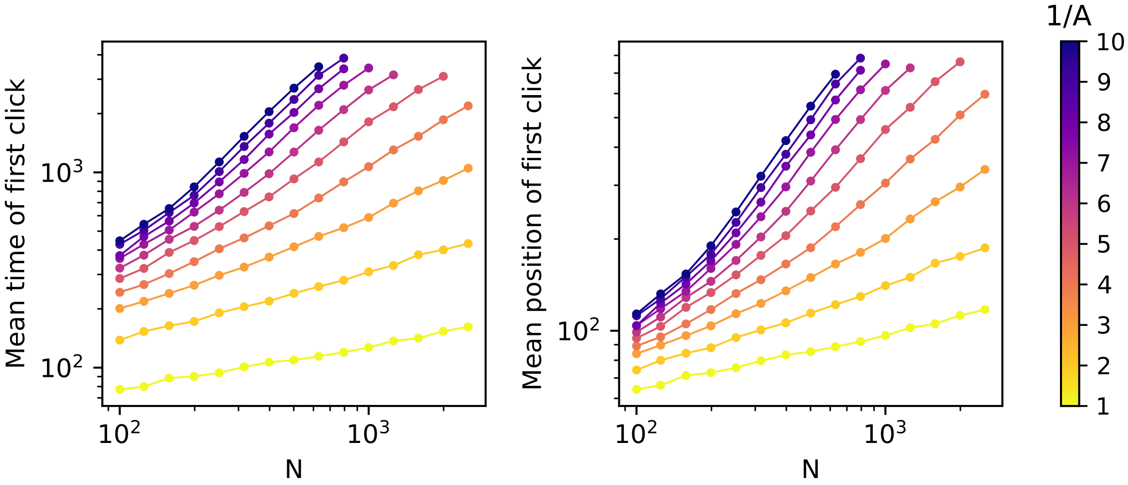


Figure 4.7: Scaling of the click rate and click position with N for the parametrization (4.12). The left plot shows the value of T_1, and the right plot the value of $n_1^{\max}(T_1) - n_1^{\max}(0)$, see Section 4.2 for the definitions. Each point is averaged over 5000 simulations. The parameter values are $\rho = 1$, $\mu = 0.01$, $s = 0.02$, $m = 0.1$.

4.C Alternative parametrization of the Allee effect

In the current work, the birth and death rates have been chosen so that the overall growth rate of the population is a cubic function of the population size n,

$$n(\lambda_0(n) - \delta(n)) = rn(B\frac{n}{N} + 1)(1 - \frac{n}{N}).$$

An alternative parametrization of the same polynomial can be obtained by setting $\rho = rB$ and $A = 1/B$, so that

$$n(\lambda_0(n) - \delta(n)) = \rho n(\frac{n}{N} + A)(1 - \frac{n}{N}). \tag{4.12}$$

In this case, the population exhibits a weak Allee effect for $A \in (0, 1)$ and no Allee effect for $A \geq 1$, so that the strength of the Allee effect is inversely related to the parameter A.

We have reproduced the results of Figure 4.5 for this alternative parametrization. The results are shown in Figure 4.7. In order to make the comparison with Figure 4.5 easier, we have used the values of A corresponding to the values of B used in Figure 4.5. The result is qualitatively very different from Figure 4.5. Increasing the strength of the Allee effect reduces the click rate for all N, whereas this was only the case for large N in Figure 4.5.

4.D Supplementary figures

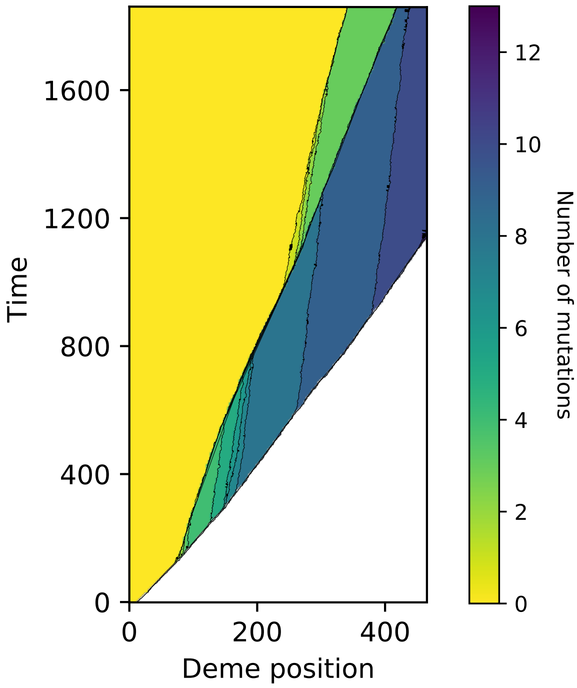


Figure 4.8: Space-time diagram of a simulation of the spatial Muller's ratchet. Each row corresponds to a time value, and the number of mutations of the fittest individual in each deme at that time is represented. Notice that a fast genetic wave rapidly forms, and that it experiences several successive inner clicks of the ratchet. The parameter values are the same as in Figure 4.4.

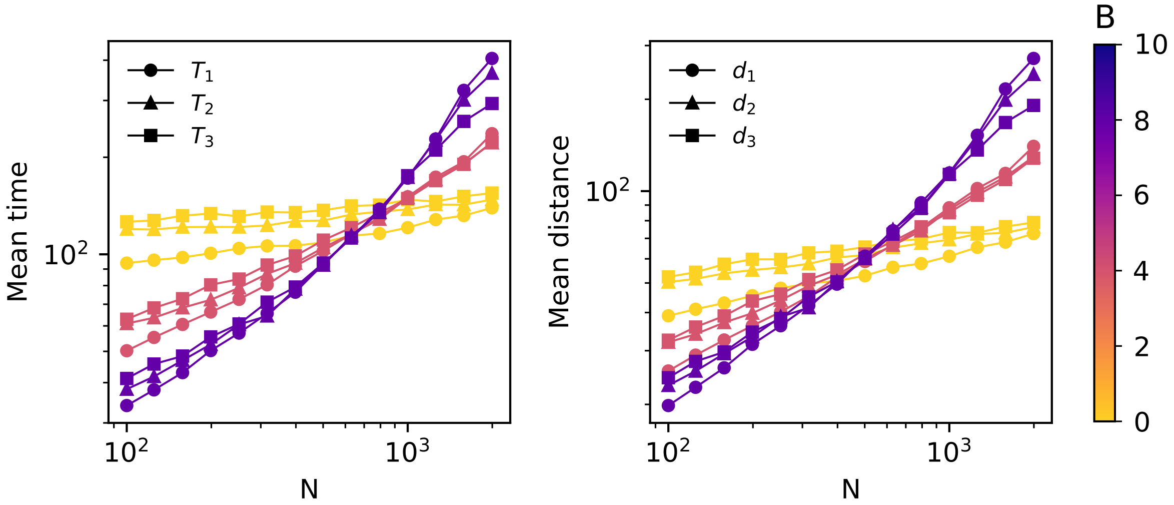


Figure 4.9: Scaling of the first, second and third click rate and click distance with N. Each point is averaged over 5000 simulations. The parameter values are $r = 1$, $\mu = 0.01$, $s = 0.02$, $m = 0.1$.

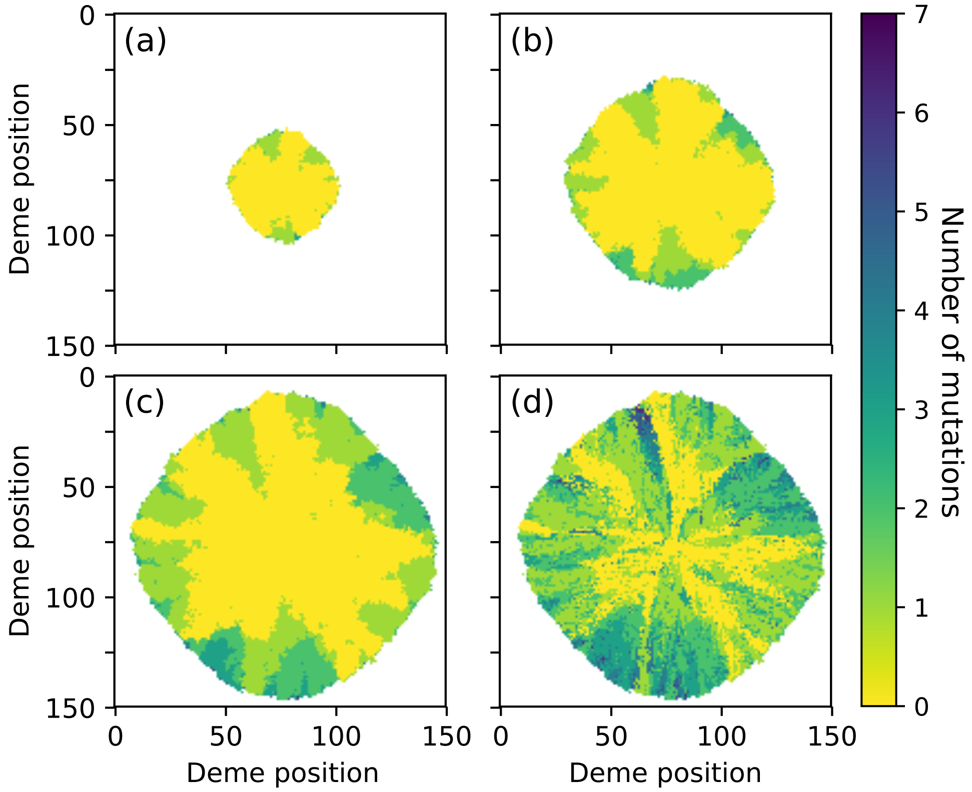
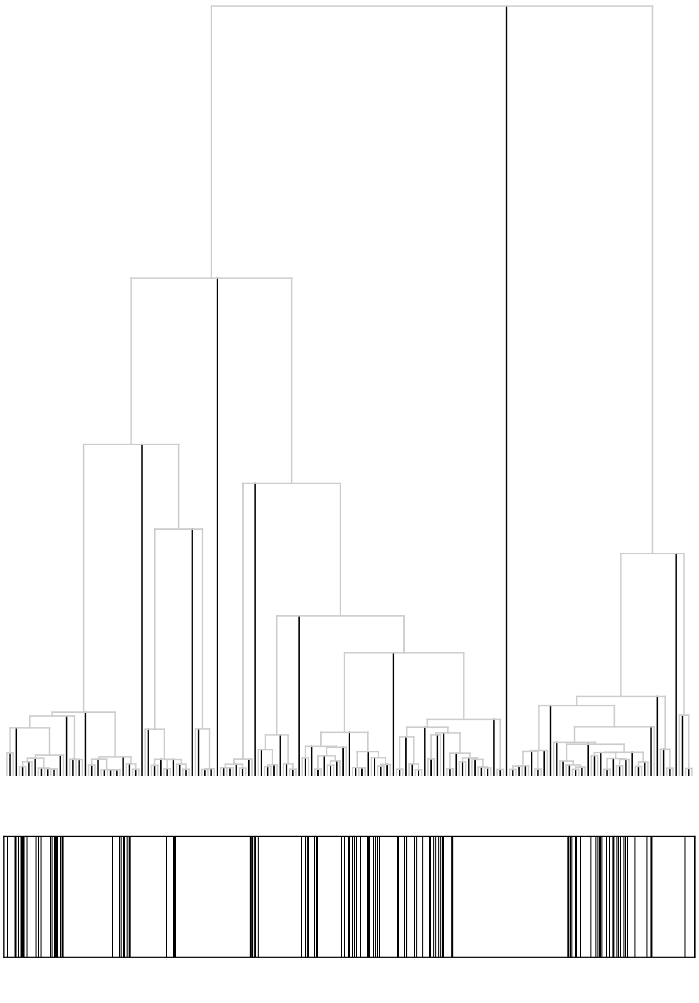


Figure 4.10: Simulation of the two-dimensional spatial Muller's ratchet. (a,b,c) State of the population at time $t = 50$, 100, 150 respectively. In each deme the number of mutations of the fittest individuals is represented. (d) Number of mutations at colonization time, see Section 4.2 for the definition. The parameters are $N = 300$, $\mu = 0.02$, $s = 0.02$, $r = 1$, $m = 0.1$, $B = 0$.

Chapter 5



5

A branching process with recombination

This chapter is work in progress with Amaury Lambert and Emmanuel Schertzer. Even if we provide rigorous proofs of the results stated in the introduction, the reader should be warned that these are only preliminary versions of them The notation can certainly be improved, the results strengthened and the proofs shortened.

Illustration. The top tree is a simulation of the limiting rescaled genealogy of the branching process with recombination, and the bottom picture is the location of the blocks of ancestral genome in the population corresponding to this genealogy, see the caption of Figure 5.2.

5.1 Introduction

5.1.1 Motivation

A large part of population genetics has been devoted to understanding the dynamics of the allele frequencies at one locus. Many results in this case are available, and allow to incorporate several evolutionary forces, such as selection, mutation, population structure, or spatial structure, see for instance [56]. When it comes to studying several loci, it is necessary to take into account recombination. Genetic recombination is any mechanism by which the offspring inherits a collection of alleles which not that of one of its parents. If we suppose that all alleles are on the same chromosome, the collection of alleles of an individual is called the *haplotype* of this individual. Without recombination, each individual would inherit the haplotype from one of its parents, and we would be back to the situation with one locus, where the locus is now the entire chromosome, and an allele is a haplotype.

In eukaryotes, an important recombination mechanism is the crossing-over, which is pictured in Figure 1.1 of Chapter 1. It is a mechanism by which an offspring can inherit a chromosome which is a mosaic of the parental chromosomes. A crossing-over corresponds to some location on the chromosome such that the genetic material on one side of the crossing-over originates from one parent, and that on the other side from the other parent. Thus, due to crossing-overs, all loci are not inherited from the same parent. The probability of inheriting a particular set of loci depends on the distance between them, and on their linear arrangement along the chromosome. This creates complex dependencies between the allelic frequencies in the population. It is possible to give an expression for the dynamics of these allelic frequencies, but in practice it becomes intractable even when considering a few number of loci. See for instance Chapter 3 in [52].

Instead of following the allelic frequencies at a finite number of loci, it is possible to envision the chromosome as a continuous line. Each crossing-over can be seen as a point on this line, called a junction [71]. The two parts of the chromosome separated by a junction have distinct evolutionary histories as they were inherited from different individuals. The set of all junctions splits the chromosome into segments called *identical by descent blocks* (IBD blocks, see for instance [38]) which have not experienced any recombination event, and thus correspond to a fragment of chromosome that existed in an ancestral population. In this work, we are interested into studying the dynamics of these junctions and IBD blocks in a constant-size, panmictic, neutral population. Obtaining a better understanding of the dynamics of the genomes in the population in this situation can serve as a null model, and help to identify deviations it. For instance, the length and location of the IBD blocks on the genome have already been leveraged to infer selection [187] or past demographic history [180, 146]. We will follow the genetic contribution of one focal individual in a branching population with recombination, similar to that considered in [7]. We think of this process as the branching approximation of a Wright-Fisher model with recombination. Let us now present the model that we consider.

5.1.2 Wright-Fisher and branching models with recombination

Wright-Fisher model. In order to motivate our object of study, let us consider first the following version of the Wright-Fisher model with recombination. Consider a population of fixed size N, where each individual carries a unique linear haploid chromosome of size R, represented by the segment $[0, R]$. Generations are non-overlapping, and generation $t + 1$ is built from generation t by, independently for each individual, realizing the following operations.

- Pick independently two parents uniformly at random from generation t, with resampling.

- With probability $1 - R/N$, no recombination occurs and the offspring inherits a copy of the chromosome of one of its parents, chosen with equal probability.

- With probability R/N, a recombination event occurs. We assume that each recombination event is made of a single crossing-over, so that both chromosomes are cut into two at the same location. This location is assumed to be uniformly distributed on $[0, R]$. The offspring inherits a copy of the portion

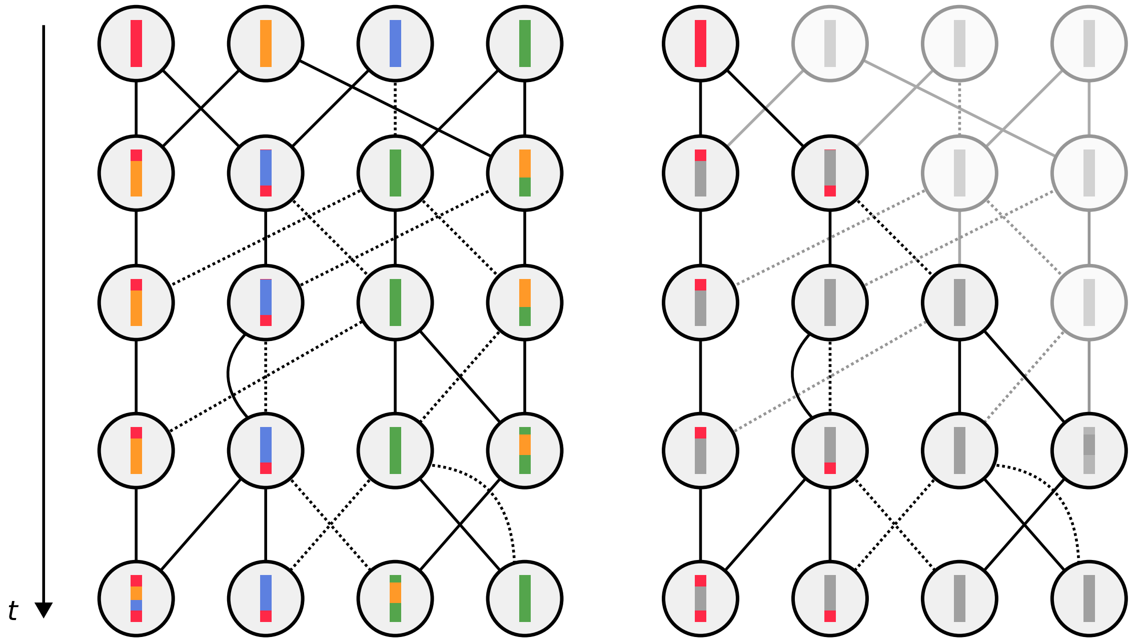


Figure 5.1: Left: Illustration of the Wright-Fisher model with recombination. Each individual is connected to each of its two parents with a line. This line is solid if the parent has transmitted some genetic material, and dotted if it has not transmitted any genetic material. The ancestral origin of the segments of chromosomes are indicated by the colors. Right: All individuals that are not in the pedigree of the left-most individual are represented with a lighter shade. Note that not all individuals in the pedigree carry genetic material from the ancestor, and that the formation of a chromosome with two intervals of genetic material requires the inbreeding of two individuals from the pedigree.

> of the chromosome to the left of this point from one of its parent, and a copy of the portion to the right of this point from its other parent. There are two possible such chromosomes, they are chosen with equal probability.

Suppose that, at $t = 0$, each chromosome is given a different color. Due to recombination, new chromosomes will be formed that are mosaics of the colors of the initial individuals, see Figure 5.1. Note that segments carrying the same color correspond to IBD segments. If, at some time, all individuals in the population share the same haplotype, that is, have the same mosaic of color on their chromosomes, then recombination cannot create new haplotypes anymore, and all subsequent individuals in the population will carry the same haplotype. We say that the haplotype has fixed. It is clear from the transition of the model that there will a.s. be a unique haplotype that reaches fixation, in finite time.

Our ultimate goal is to understand the distribution of colors along the chromosome that reaches fixation. As segments of the same color correspond to an IBD block, the distribution of colors along the fixed haplotype informs us on the identity by descent of the various loci of the fixed genome in the population.

Some results on this distribution have previously been derived in [139], under a large population, large chromosome size limit. Their approach relies on a description of the joint genealogy of all the loci on the chromosome, known as the ancestral recombination graph (ARG) [93]. An IBD block can be recovered from the ARG as the set of all loci that have the same ancestor at some large time. We will review their findings in the forthcoming Section 5.1.4. In this work we take a different approach and follow forward in time the long-term genetic contribution to the population of an initial focal individual. We anticipate that most of the genetic material of the focal individual will be lost within the first generations, where the blocks of ancestral genome are long, and recombination acts strongly on them. As long as the number of descendants of the focal individual is small compared to the total population size N, the fixed size constraint of the Wright-Fisher model is negligible, and the dynamics of the focal individual's progeny is well approximated by a branching process that we now introduce.

Branching model. The branching process that we consider is described in Definition 5.1. We start with an informal derivation of this branching approximation leading to this definition. Consider a focal ancestral individual at $t = 0$. Let us first consider the number of descendants of this individual, regardless of them carrying genetic material from that ancestor or not. Suppose that the ancestor has k descendants at generation t. For each individual at generation $t+1$, the probability that both its parents are descendants of the ancestral individual is $(k/N)^2$. As there are N individuals at generation $t+1$, the probability that at least one individual has its two parents among these k individuals vanishes. Thus, for large N, and as long as $k/N \ll 1$, there is no reproduction event among the descendants of the ancestral individual. We say that there is no *inbreeding*. Now, let X_i^N denote the number of individuals at generation $t+1$ that have one parent among the $N-k$ individuals that are not descendants of the focal individual, and whose other parent is i. It is clear that

$$(X_1^N, \dots, X_k^N) \overset{(\mathrm{d})}{=} \mathrm{Multinomial}(N; \tfrac{2(N-k)}{N^2}, \dots, \tfrac{2(N-k)}{N^2}).$$

Thus, the following convergence holds

$$(X_1^N, \dots, X_k^N) \longrightarrow (X_1, \dots, X_k),$$

where (X_i) are i.i.d. Poisson(2) random variables.

The previous informal derivation shows that, in the large N limit, the number of descendants of a focal individual at $t = 0$ converges to a Galton-Watson process with Poisson(2) offspring distribution. This total progeny is often referred to as the *pedigree* of the ancestor [37, 15]. However, not all individuals in the pedigree carry genetic material from the ancestor, see Figure 5.1. Understanding the genetic contribution of the ancestor requires us to superimpose the process of genetic transmission on top of the pedigree.

Recall that there is no inbreeding in the pedigree: exactly one of the parents of each individual in the pedigree at generation $t+1$ is in the pedigree at time t. An important consequence of this fact is that individuals in the pedigree can only carry *segments* of the ancestral genome, that is, the set of all loci the carry ancestral material is an interval. Forming a haplotype with two intervals of ancestral genetic material requires a recombination event between an individual with one interval, and the other with the other interval, see Figure 5.1. As there is no inbreeding in the pedigree in our approximation, such a recombination event does not occur. The segment of ancestral material carried by an individual in the pedigree will also be referred to as the *block* of ancestral genome. Let us consider the dynamics of the blocks along the pedigree.

Suppose that an individual has inherited a segment I of ancestral genetic material. It will give birth to a Poisson(2) distributed number of children. Each child has a probability $1 - R/N$ of experiencing no recombination. In this case, no ancestral genetic material is passed on to the child with probability $1/2$, otherwise the offspring inherits the entire segment I. If a recombination event occurs, and $|I|$ denotes the length of I, then with probability $1 - |I|/R$ the location of the crossing-over is outside of I, in which case we are back to the situation without recombination. If the crossing-over occurs within $I = [a, b]$, and U is uniform on $[a, b]$, then the offspring inherits the segment $[a, U]$ or $[U, b]$, with equal probability. Summarizing, each child inherits:

- no ancestral material with probability
$$\frac{1}{2}\left(1 - \frac{|I|}{N}\right);$$
- the whole block I with probability
$$\frac{1}{2}\left(1 - \frac{|I|}{N}\right);$$
- a recombined block $[a, U]$ or $[U, b]$, each with probability
$$\frac{1}{2}\frac{|I|}{N}.$$

Discarding all individuals that do not carry ancestral genetic material, and recalling that the number of offspring of each individual is a Poisson(2) variable, we obtain the following branching approximation of the Wright-Fisher model with recombination.

Definition 5.1. The branching process with recombination is a Markov branching process with values in the subintervals of $[0, R]$. An individual with block $I = [a, b]$ gives birth, independently of the rest of the population, to

- a Poisson($1 - |I|/N$) distributed number of children with block I;
- a Poisson($|I|/N$) distributed number of children with segment $[a, U]$, where U is uniformly distributed on $[a, b]$ and is independent for different children;
- a Poisson($|I|/N$) distributed number of children with segment $[U, b]$, where U is uniformly distributed on $[a, b]$ and is independent for different children. ○

In the previous definition, we could account for the individuals with no genetic material by assuming that an individual with chromosome I gives birth to an additional Poisson$(1-|I|/N)$ distributed number of individuals with block $\emptyset$. This branching process would correspond to the pedigree of the initial individual, and the interval carried by each individual gives the amount of genetic material inherited by this individual.

5.1.3 Main results

We study the branching process with recombination conditioned on having some remaining genetic materiel at generation σN. We have defined the block of ancestral genetic material of an individual v as an interval $I_v = [a, b]$. It will be more convenient to encode I_v as two reals: the block length $\rho_v = b-a$ and the block left endpoint $H_v = a$. We prove two main results in this chapter. The first one, see Theorem 5.2, provides the limit of the empirical distribution of the block lengths. The other one, see Theorem 5.4, gives the limit of the empirical distribution of the locations of the blocks on the genome. As we shall see, deriving this limit will require us to see the locations of these blocks as a random metric space, and thus our result provides the limit of the “geometry” of the blocks along the chromosome.

A k-spine decomposition theorem. Both the previous results are proved using a k-spine technique that we believe to be of independent interest. The broad idea of spinal decomposition results for branching Markov processes is to construct a tree with one distinguished lineage, the spine, and to connect the distribution of this tree to the original distribution of the branching process under consideration. Such spinal decomposition results have been derived for the case $k = 1$ in a variety of contexts including Galton-Watson trees [150], superprocesses [182], branching random walks [199], and CMJ processes [117].

One application of spinal decomposition theorems is the “many-to-one formula”. It is a general principle which states that the expectation of some quantity of interest, summed over all individuals at some generation, can be expressed in terms of the sole spine. Here, we want to study the empirical distribution of blocks at some large time using a moment technique. The many-to-one formula gives us access to the first order moment of this empirical distribution. In order to derive its moment of order k, we will need to construct a tree with k distinguished vertices, known as a k-spine. The analogous of the many-to-one formula for a k-spine has been dubbed the “many-to-few formula”. A general k-spine decomposition theorem and many-to-few formula has been derived in [103]. In this work, we derive another general k-spine decomposition theorem, closer in spirit to the 2-spine decomposition considered in [181], see Theorem 1.2. The link between our theorem and existing results is discussed in more details in Section 5.2. We use our version of the many-to-few formula to derive the following two results. As was already pointed out, we believe that our k-spine results is a quite general tool which could be used in many other contexts.

Distribution of the block sizes. Consider a branching process with recombination started from a chromosome of length R, with total population size N. Let us define the empirical distribution of block lengths at time t as a measure $\nu_t^{N,R}$ on $\mathbb{R}_+$ defined as

$$\nu_t^{N,R} = \sum_{|v|=t} \delta(\rho_v),$$

where the sum is over all individual alive at t. The following result, proved in Section 5.4.3, provides the large population, large chromosome size limit of ν_t.

Theorem 5.2. *Let $\bar{\nu}_t^{N,R}$ denote the empirical distribution of the block sizes of the branching process with recombination at generation t, conditioned on non-extinction at time σN, and started from a chromosome of size R in a population of size N. Then the following convergence holds in distribution for the weak topology,*

$$\lim_{R\to\infty} \lim_{N\to\infty} \frac{1}{N \log R} \bar{\nu}_{\lfloor N\sigma \rfloor}^{N,R} = Y_\sigma \mathscr{L}_\sigma,$$

where $Y_\sigma \sim \mathrm{Exponential}(1/\sigma)$, and $\mathscr{L}_\sigma$ denotes the $\mathrm{Exponential}(\sigma)$ distribution.

The previous result shows that the block lengths distribution converges to a deterministic measure $\mathscr{L}_\sigma$ with random total mass Y_σ. It is remarkable that we did not need to rescale the block lengths. Thus, even starting from a very large chromosome of size R, recombination acts so strongly that for any $\sigma > 0$, the size of the blocks in the population at time σN is of order 1. This also shows that the "interesting part" of the dynamics, where the blocks in the population are reduced from length R to length of order 1, occurs on a very short time-scale which is not captured in the natural time-scale of the process. See the discussion in Section 5.1.5 for a possible refinement of the Theorem 5.2 that would provide the distribution of the block lengths on this shorter time-scale.

Another consequence of Theorem 5.2 is that the total number of individuals, once properly renomarlized, converges to an exponential variable. This is reminiscent of Yaglom's exponential limit law for critical branching processes (see for instance [3], Section 9 of Chapter 1).

Remark 5.3. This result also shows that the total number of individuals at time $N\sigma$ is of order $N \log R \gg N$. Therefore we know that, by time $N\sigma$, some inbreeding will occur in the Wright-Fisher model so that the branching process approximation breaks down. The branching approximation should hold until a time of order $N/\log R$, in which case all individuals carry a block of length $\log R$. All proofs could be adapted to conditioning on survival until time $\sigma N/\log R$, as this would roughly amount to using that $R^\varepsilon \gg \log R$ rather than $R^\varepsilon \gg 1$, for any ε. We decided not to do so here as we are more interested in giving the general properties of the branching process rather than using it to prove results on the Wright-Fisher model. ○

Geometry of the blocks. Our result for the geometry of the blocks requires to see the population as a random metric measure space. More precisely, consider the branching process with recombination of size N and initial chromosome size R. Let $\mathcal{V}_t$ be the index set of all individuals alive at time t. (For simplicity we drop the N and R from the notation.) We can define a random metric measure space $(\mathcal{V}_t, d_\mathcal{V}, \mu_t)$ by defining

$$\forall u, v \in \mathcal{V}_t, \quad d_\mathcal{V}(u,v) = |H_u - H_v|$$

and setting $\mu_\mathcal{V}$ to be the counting measure on $\mathcal{V}_t$. The distance between two individuals is simply the usual distance between the left endpoint of their blocks.

The limiting geometry of the blocks will be expressed in terms of the Brownian coalescent point process (CPP) introduced in [1], that we now recall. Consider a Poisson point process $\mathscr{P}$ on $\mathbb{R}_+ \times \mathbb{R}_+$ with intensity measure

$$\mathrm{d}t \otimes \frac{1}{u^2} \mathrm{d}u.$$

Let Y_1 denote the first time when the second coordinate of an atom of $\mathscr{P}$ exceeds 1. (Note that $Y_1 \sim$ Exponential(1), and is independent of the restriction of $\mathscr{P}$ to $\mathbb{R}_+ \times [0,1]$.) Then $\mathscr{P}$ encodes a random ultrametric space defined as

$$\forall x < y \in [0, Y_1], \quad d(x,y) = \sup\{u : (t,u) \in \mathscr{P}, x \leq t \leq y\}.$$

This ultrametric space is naturally endowed with the Lebesgue measure on $[0, Y_1]$, defining a random ultrametric measure space. This random ultrametric space is actually an example of a comb metric space that were introduced in Chapter 2. It is the comb metric space associated to the comb function

$$\forall x \in [0, Y_1], \quad f(x) = \begin{cases} u & \text{if } (x,u) \in \mathscr{P}, \\ 0 & \text{else.} \end{cases}$$

It can be described in a pictorial way by thinking of each atom $(x,u) \in \mathscr{P}$ as a tooth of length u located at t, and defining $d(x,y)$ as the length of the largest tooth between x and y. See Figure 5.2 for an illustration.

Finally we need a notion of convergence for random metric measure spaces. We will endow the space of complete separable metric measure spaces with the Gromov-weak topology, introduced in [91] in the case where the measure is a probability measure, see [43] for an extension for general finite measures. For a general metric measure space (X, d, μ), let us define the following map, also called the n-distance matrix,

$$D_n \colon \begin{cases} X^n \to \mathbb{R}^{\binom{n}{2}} \\ (x_i) \mapsto (d(x_i, x_j)), \end{cases}$$

and define $\iota_n = \mu^{\otimes n} \circ D_n^{-1}$ as the push-forward measure of the product measure $\mu^{\otimes n}$ by the map D_n. A sequence (X^N, d^N, μ^N) converges in the Gromov-weak sense to (X, d, μ) if, for any $n \geq 1$, ι_n^N converges as $N \to \infty$ in the weak topology to ι_n,

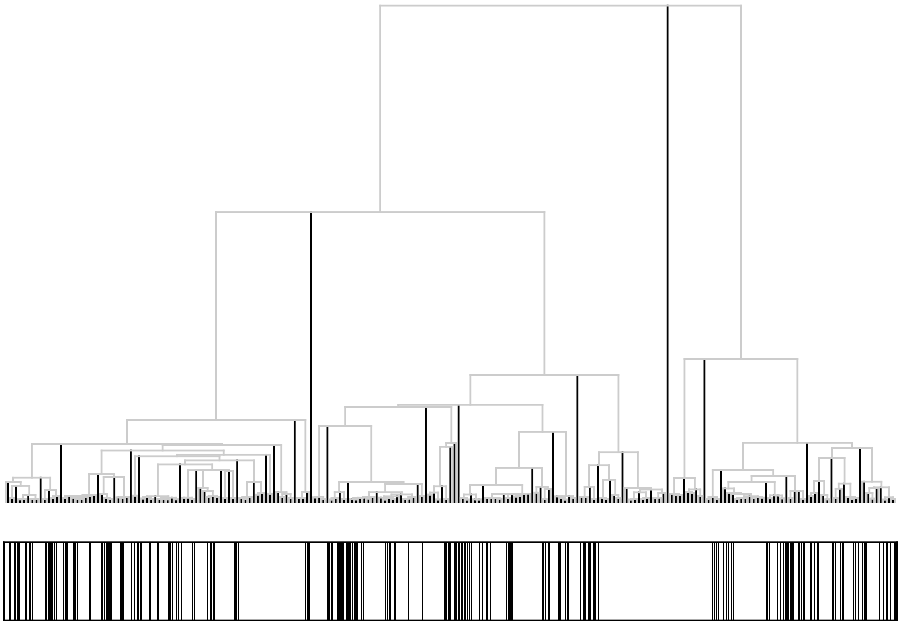


Figure 5.2: Top: simulation of a Brownian CPP. The black vertical lines represent to the atoms of $\mathscr{P}$, and the corresponding tree is pictured in grey. Bottom: geometry of the blocks of ancestral material corresponding to the top CPP. Each block is represented by a black stripe. The distance between two consecutive stripes is the logarithm of their distance on the chromosome. Note that this induces a strong deformation of the intuitive linear scale.

where ι_n^N, resp. ι_n, denotes the push-forward measure of μ^N, resp. μ, by D_n. The Gromov-weak topology naturally defines a notion of convergence in distribution for random metric measure spaces and we can show the following result.

Theorem 5.4. *Let $(\mathcal{V}_t, d_{\mathcal{V}}, \mu_{\mathcal{V}})$ represent the geometry of the blocks of a branching process with recombination in a population of size N, and initial chromosome size R. The following convergence holds in distribution in the Gromov-weak topology*

$$\lim_{R\to\infty}\lim_{N\to\infty}\Big(\mathcal{V}_{\lfloor N\sigma\rfloor}, \frac{\log(d_{\mathcal{V}}\vee 2)}{\log R}, \frac{\mu_{\mathcal{V}}}{N\log R}\Big) = \Big([0,Y_1], d, \sigma\,\mathrm{Leb}\Big),$$

where $([0,Y_1], d, \mathrm{Leb})$ is a Brownian CPP.

Remark 5.5. The metric $d_{\mathcal{V}}\vee 2$ is defined as

$$\forall u,v\in\mathcal{V}_t,\quad \mathbf{1}_{\{u\neq v\}}(d_{\mathcal{V}}(u,v)\vee 2).$$

This modification of $d_{\mathcal{V}}$ ensures that it remains a distance after taking the logarithm. ○

The previous result is proved in Section 5.5.2, and is illustrated in Figure 5.2. It shows that, in the limit, individuals at time σN can be identified with the leaf-set $[0,Y_1]$ of a Brownian CPP, and that the distance of the blocks of any two individuals x and y in the population is of order $R^{d(x,y)}$. Note that this distance does not depend on σ. The final time σ only impacts the total population size and the lengths of the blocks, but not their distribution along the genome. We also recover that the population size at time σN is of order $N \log R$.

Our proof of Theorem 5.4 also provides a genealogical interpretation to the limiting Brownian CPP. First, we show that the distance on the chromosome of two individuals in the population is of the same order as the length of the block of their most-recent common ancestor (MRCA). This can be explained intuitively as follows. At the time of the MRCA of two individuals u and v, the population can be separated into two parts: the family to which u belongs and that to which v belongs. By the branching property, these two families are independent. The final sizes of the blocks of u and v are of order one, and are negligible compared to that of the block of their MRCA. Thus, they appear as points on the block of their MRCA, and the location of these points are independent, so that their distance is of the same order as the length of the block of their ancestor. This idea is formalized in Proposition 5.24.

The second step of the proof is to show that the genealogy of the population is given by a Brownian CPP, after an appropriate time-change. Most of the branching events in the genealogy will occur within the short initial period where the chromosome is broken up from size R to size of order 1. Obtaining a non-degenerate genealogy requires us to time-change the process in order to explore this initial phase. A formal definition of this time-change requires further notation and is provided in Section 5.4.1, however it can be intuitively envisioned as a Lamperti transform [143] the branches of the tree.

Remark 5.6. Our two main results provide the convergence of the empirical measure of the block sizes and of the block locations. It would be natural to obtain the convergence of the joint empirical measure of these two quantity. A formalization of this result would require us to see the sizes of the blocks as marks attached to the individuals, and to use the framework of random marked metric measure spaces [44]. We prefer not to do so in this preliminary version of our work, even if the spinal techniques that we use should enable us to obtain a result in this direction. ○

5.1.4 Connection with the literature

A similar branching process was considered in [7], in a slightly more general setting. They allow for selection, and for more general offspring distributions than the Poisson distribution considered here. Our approach mostly depends on the fact that, conditional on the number of children, the recombination events are independent for different children. We believe that it could be extended to the more general case considered in [7].

Among other things, [7] provide a super-process approximation to the branching process with recombination, and derive an expression for the first two moments of the block length distribution at a given time. In our work, we have supposed that the recombination rate is of order R/N, and considered a large N limit first, and then a large R limit. Their result holds for a wider scaling of these two parameters. However, we believe that our results can also be adapted to obtain a joint limit, as long as $N \gg R$ (as in their work). The large N convergence step is essentially a convergence towards a Poisson point process, for which many precise convergence results exist. The expression that we obtain for the moments of the block lengths distribution coincides with that of [7] applied to the scaling that we consider. In this sense, Theorem 5.2 is an extension of their expression to all moments of the block length distribution.

Finally, they provide an expression for the probability of non-extinction that we will need. Under our scaling, their expression reduces to the following result.

Proposition 5.7 ([7])**.** *Let $p^{N,R}(t)$ be the probability of non-extinction of the branching process with recombination at generation t, started from one individual with chromosome of size R. Then, for any $\sigma > 0$, we have that*

$$\lim_{R\to\infty} \lim_{N\to\infty} \frac{N \log R}{R} p^{N,R}(\lfloor \sigma N \rfloor) = 1.$$

We now review some results from [139]. Their setting is similar to ours, but they consider a Moran model, and look backwards in time at the genealogy of the fixed haplotype. One of their main results gives the geometry of all loci of the fixed haplotype that are IBD with the left endpoint of the chromosome, that is, of all loci that are of the same color as the left endpoint.

More precisely, they prove the existence of a process valued in the partitions of $[0, R]$, the $\mathbb{R}^+$-partitioning process. This process admits a unique stationary distribution, and if Π_{eq} denotes a partition distributed according to this stationary distribution, the blocks of Π_{eq} are distributed as the colors of the fixed haplotype in a Moran model with recombination. Let us define the following random measure

$$\forall x \leq y \leq 1, \quad \vartheta^R([x,y]) = \int_{R^x}^{R^y} \mathbf{1}_{\{0 \sim_{\Pi_{\mathrm{eq}}} u\}} \, \mathrm{d}u$$

that encodes the loci that are IBD with the left endpoint of the chromosome, on a logarithmic scale. Then they proved the following result.

Theorem 5.8 ([139])**.** *The following convergence holds in distribution for the weak topology*

$$\lim_{R\to\infty} \frac{1}{\log R} \vartheta^R = \sum_i y_i \delta(x_i)$$

where $((x_i, y_i);\, i \geq 1)$ are the atoms of a Poisson point process on $\mathbb{R}_+ \times \mathbb{R}_+$ with intensity measure

$$\frac{1}{x^2} e^{-y/x} \, \mathrm{d}x \, \mathrm{d}y.$$

It was already noticed in [139] that the limit of ϑ^R in the previous theorem can be constructed from a Brownian CPP as follows. Let $\widetilde{\vartheta}$ be the random measure such that

$$\forall a \leq Y_1, \quad \widetilde{\vartheta}([0,a]) = \mathrm{Leb}(\{x \leq Y_1 : d(0,x) \leq a\}),$$

then $\widetilde{\vartheta} = \sum_i y_i \delta(x_i)$ where $((x_i, y_i);\, i \geq 1)$ are the atoms of a Poisson point process with intensity

$$\frac{1}{x^2} e^{-y/x} \, \mathrm{d}x \, \mathrm{d}y.$$

Theorem 5.4 shows that, in the limit, the population can be identified with the random metric measure space $([0, Y_1], d, \sigma \, \mathrm{Leb})$, where $d(x, y)$ corresponds to the limit of the logarithm of the chromosomic distance between the blocks of x and y. Thus, if in the discrete branching process U is an individual chosen uniformly in the population at generation σN, and we denote by

$$\vartheta^{N,R} = \sum_{|v| = \lfloor \sigma N \rfloor} \rho_v \delta\Big(\frac{\log |H_U - H_v|}{\log R} \Big)$$

the empirical measure of the chromosomic distance between the block of U and the other blocks in the population, Theorem 5.4 and Theorem 5.2 strongly suggest that

$$\frac{1}{N \log R} \vartheta^{N,R} \longrightarrow \widetilde{\vartheta},$$

so that we recover Theorem 5.8. Note that Theorem 5.8 only provides the geometry of the blocks "seen from the left endpoint of the chromosome". Blocks on the genome are aggregated into clusters that correspond to the atoms of the Poisson point process, and their result does not give the finer description of the geometry of these clusters. Here we have established the convergence in the Gromov-weak sense, so that we proved that the geometry of each cluster is given by a Brownian CPP. In that sense, Theorem 5.4 is an extension of Theorem 5.8, in the branching process framework. We also have provided a clear genealogical interpretation to this result.

Remark 5.9. The fact that we recover the same geometry on the chromosome for the IBD blocks in the fixed haplotype and for the surviving blocks in the branching process can be quite puzzling. In the former case, all the blocks belong to the *same* chromosome, carried by all individuals in the population. In the latter case, each block is carried by a unique individual, and all blocks lie on different chromosomes. After the branching phase, there will be a logistic phase where inbreeding will form haplotypes with more than one ancestral block. The fact that the distribution of the blocks along the genome is the same at the end of the branching phase and at fixation suggests that all portions of the ancestral chromosome that make it until the end of the branching phase will be combined on the same haplotype that will reach fixation. ○

5.1.5 Future directions, outline

Size of the largest block. Here, we have used the spinal decomposition theorem to compute the moments of the empirical distribution of the block sizes, using a many-to-few formula. This approach provides the behavior of the "bulk" of the population, but does not give any information about the extrema of the process, that is, about the size of the largest block. Another application of the spinal decomposition theorem in the branching random walk literature is to provide the location of the minimum of the walk, see for instance Chapter 5 in [199]. It could be interesting to see if these techniques could be adapted to provide an expression for the size of the largest block.

Initial fast recombining phase. As mentioned above, looking at the process at time σN completely misses the initial phase where the blocks break from size R to size of order 1. In order to investigate this initial phase, we consider the process until time $\sigma(R)N$, where

$$\sigma(R) = \frac{1}{R^{1-\sigma}}.$$

Recall that ν_t stands for the empirical measure of block lengths at generation t. Let us denote by

$$\widetilde{\nu}_\sigma(\mathrm{d}x) = \frac{R^{1-\sigma}}{N \log R} \nu_{N\sigma(R)}\Big(\frac{\mathrm{d}x}{R^{1-\sigma}} \Big)$$

the renormalized empirical measure at time $\sigma(R)N$. Note that we have rescaled the lengths of the blocks by $R^{1-\sigma}$. Let us now provide some heuristic arguments that suggest that the process $(\widetilde{\nu}_\sigma;\, s \in [0,1])$ is approximated in the limit by

$$\Big(Z_\sigma \mathscr{L}_1(\mathrm{d}x);\, \sigma \in [0,1] \Big) \tag{5.1}$$

where $\mathscr{L}_1$ is the distribution of an Exponential(1) variable, and $(Z_t;\, t \geq 0)$ solves

$$Z_0 = 0, \quad \mathrm{d}Z_t = \mathrm{d}t + \sqrt{Z_t}\, \mathrm{d}B_t$$

where $(B_t;\, t \geq 0)$ is a standard Brownian motion. The solution to the previous equation is a Doob harmonic transform of the critical Feller diffusion, and corresponds to the so-called Q-process of the Feller diffusion, that is, to the Feller diffusion "conditioned on never going extinct".

First, using the expression of [7] for the survival probability, we obtain that

$$\lim_{R\to\infty} \lim_{N\to\infty} \frac{N \log R}{R} p^{N,R}(N\sigma(R)) = \frac{1}{\sigma}.$$

We recover the usual $1/\sigma$ decay of the survival probability at time σ of a critical Feller diffusion.

Second, the Brownian CPP stopped at its first atom above level $h > 0$ corresponds to the genealogy of a critical Feller diffusion, conditioned on survival until time h. The calculation of Section 5.4 that proves that the genealogy of the population is a Brownian CPP stopped above level 1 can be readily adapted to consider the population until time $N\sigma(R)$. They show that the limiting genealogy of the population at time $N\sigma(R)$ is a Brownian CPP, stopped at its first atom above level σ.

Finally, an adaptation of those calculation would also prove that

$$\widetilde{\nu}_\sigma \longrightarrow Y_\sigma \mathscr{L}_1,$$

where Y_σ is an Exponential(σ) variable. This is the one-dimensional marginal of (5.1).

Therefore, those heuristic arguments suggest that both the survival probability, the one-dimensional marginals, and the genealogy of the population at time $N/R^{1-\sigma}$ converge to that of (5.1).

Outline. The paper is laid out as follows. In Section 5.2 we prove our spinal decomposition theorem for a general class of branching Markov processes. In Section 5.3 we apply these results to the branching process with recombination, and provide the large population size limit of its k-spine. Most of the proofs are contained in Section 5.4. We start by providing all the estimates required for the convergence of the block size distribution, and prove Theorem 5.2 in Section 5.4.3. Finally, Section 5.5 contains the proofs of the results on the genealogy of the branching process with recombination, and on the geometry of the blocks on the chromosome.

5.2 A spinal decomposition theorem

Our strategy to prove Theorem 5.2 and Theorem 5.4 is to use the method of moments. In this section, we derive an expression for the k-th factorial moment of the empirical measure of the branching process with recombination in terms of a tree with k leaves, that we call a k-spine. We first derive this expression for a large class of branching Markov processes, and then carry out the calculations in the special case of the branching process with recombination.

The general setting that we consider is the following. Let E be a Polish space. We consider a population process where each individual u is endowed with a random variable $X_u \in E$ that gives its location. The population starts from one individual located at $x_0 \in E$. Then, at each generation, individuals reproduce independently from each other. We suppose that, conditional on $X_u = x$, the location of the offspring of u is given by the atoms of a random point process $\xi(x)$. The distribution of the branching process thus depends on the location x_0 of the initial individual, and on the family of point processes $(\xi(x);\, x \in E)$. We denote this distribution by $\mathbf{P}_{x_0}$, see Section 5.2.1 for a more formal construction.

We say that a function $H \colon E \to \mathbb{R}_+$ is (positive) *harmonic* if

$$\forall x \in E, \quad \mathbb{E}\Big[\langle H, \xi(x) \rangle\Big] = H(x), \tag{5.2}$$

where we have used the notation $\langle f, \mu \rangle$ for the integral of f against the measure μ. Note that H is harmonic iff the following process is a martingale

$$\forall t \geq 0, \quad Z_t^{(1)} = \sum_{|u|=t} H(X_u),$$

where the sum is taken over all individual at generation t, with the convention that the sum is 0 if there are less than k individuals at generation N. Fix $k \geq 1$, some generation N, and let H be a harmonic function. Let us assume that the following variable has a finite expectation under $\mathbf{P}_{x_0}$,

$$Z_N^{(k)} := \sum_{\substack{u^1 \neq \cdots \neq u^k \\ |u^i|=N}} \prod_{i=1}^{k} H(X_{u^i})$$

where the sum is taken over all k-tuples of distinct individuals at generation N. We can then define a new probability $\mathbf{Q}_{x_0}^{k,N}$ by prescribing that

$$\frac{\mathrm{d}\mathbf{Q}_{x_0}^{k,N}}{\mathrm{d}\mathbf{P}_{x_0}^{N}} \propto \sum_{\substack{u^1 \neq \cdots \neq u^k \\ |u^i|=N}} \prod_{i=1}^{k} H(X_{u^i}), \tag{5.3}$$

where $\mathbf{P}_{x_0}^{N}$ is distribution of the first N generations of the branching Markov process. The aim of the current section is to provide a k-spine construction of the probability measure $\mathbf{Q}_{x_0}^{k,N}$, that is, to give a construction of $\mathbf{Q}_{x_0}^{k,N}$ in terms of a tree with k distinguished leaves.

Let us comment the definition of $\mathbf{Q}_{x_0}^{k,N}$. For $k = 1$, as we have assumed that H is harmonic, $(Z_N^{(1)}; N \geq 0)$ is a martingale, and thus $\mathbf{Q}_{x_0}^{1,N}$ is a martingale change of measure of the initial law $\mathbf{P}_{x_0}$. We recover the classical framework of spinal decomposition, see Chapter 4 of [199] for a nice account in the case of branching random walks, or [117] for a spinal decomposition for CMJ processes. If $k > 1$, $(Z_N^{(k)}; N \geq 0)$ is no longer a martingale. However, the change of measure $Z_N^{(k)}$ is rather elementary. If we forget about space and consider a plain Galton-Watson process, $Z_N^{(k)}$ is simply the k-th factorial moment of the population size at generation N. In general, $Z_N^{(k)}$ is directly related to the k-th moment of the empirical distribution of the locations of the individuals at generation N.

Compare this to the k-spine derived in [103]. The spinal decomposition theorem that they obtain involves a martingale change of measure, so that the k-spine they introduce is a Markov process. However, their change of measure is more involved than ours. For instance, their many-to-few formula, see their Lemma 1, depends on the whole genealogical structure of the k-spine, and not only on the leaves of the k-spine. In that sense, our spinal decomposition result is closer in spirit to the 2-spine decomposition proposed in [181], where the change of measure involved the second factorial moment of a Galton-Watson process. Actually, setting $k = 2$ and forgetting about space, we recover their 2-spine construction.

The rest of this section is laid out as follows. Section 5.2.1 contains the formal definition of the class of branching process that we consider. The k-spine is constructed in Section 5.2.2, and the spinal decomposition theorem, that is, that the k-spine tree built in Section 5.2.2 corresponds to the change of measure (5.3), is proved in Section 5.2.3 along with our many-to-few formula. Finally, Section 5.2.4 is devoted to a Palm measure construction of the k-spine.

5.2.1 Preliminaries and notation

We try to follow as much as possible the notation in [199]. A realization of the branching process is envisioned as a random tree, where each vertex of the tree is equipped with a mark that corresponds to the location in space of that individual. Let us denote by

$$\Omega \coloneqq \bigcup_{n=0}^{\infty} \mathbb{N}^n$$

the set of finite words with alphabet $\mathbb{N}$, that correspond to all individuals in the population. For $u \in \Omega$, we will denote by $|u|$ the *generation* of u, defined as the length of the vector u. Moreover, we denote by u_i the i-th coordinate of u, for $i \leq |u|$. If $u = (u_1, \dots, u_n)$ and $v = (v_1, \dots, v_m)$, we define

$$uv \coloneqq (u_1, \dots, u_n, v_1, \dots, v_m), \quad \overleftarrow{u} \coloneqq (u_1, \dots, u_{n-1})$$

to be the concatenation of u and v, and the parent of u respectively. Finally, define $\Omega^* = \Omega \setminus \{\emptyset\}$.

A branching Markov process is defined as a random subset T of Ω, as well as a collection $(X_u)_{u \in T}$ of E-valued random variables that encode the location of the individuals in the population. The distribution $\mathbf{P}_{x_0}$ of T and $(X_u)_{u \in T}$ is constructed out of an element $x_0 \in E$, giving the location of the initial individual $\emptyset$, and a collection of point processes $(\xi(x);\, x \in E)$ that encodes the reproduction events in the population.

More precisely if, for $n \geq 0$,

$$G_n = T \cap \{u : |u| = n\}$$

denotes the n-th generations of T, then conditional on $(G_1, \dots, G_n)$ and on the locations $(X_u;\, u \in G_n)$, G_{n+1} is constructed as follows. Let $(\xi_u;\, u \in G_n)$ be independent point processes such that $\xi_u \sim \xi(X_u)$. Then define

$$G_{n+1} = \bigcup_{u \in G_n} \{ui : i \leq |\xi_u|\},$$

where $|\xi_u|$ denotes the total mass of ξ_u. Moreover, suppose that for each $u \in G_n$, the atoms of ξ_u are uniformly labeled from 1 to $|\xi_u|$, and let $X_i(\xi_u)$ denote the location of i-th atom of ξ_u. Then for $i \leq |\xi_u|$, we set $X_{ui} = X_i(\xi_u)$. We let $\mathbf{P}_{x_0}$ be the distribution of the pair $[T, (X_u;\, u \in T)]$ constructed this way.

In words, an individual u alive at generation n gives birth to $|\xi_u|$ new individuals, independently of the rest of the population. The location of the newborns are given by X_i, where $(X_1, \dots, X_{|\xi_u|})$ denotes the location of the atoms of ξ_u.

Finally, we will need the notation T^w for the subtree of T attached at the vertex w:

$$T^w \coloneqq \{u \in T : \exists v \in \Omega^*,\, u = wv\}.$$

5.2.2 Construction of the k-spine

In the previous section, T denotes a (possibly) infinite subset of Ω. From now on, we fix a generation N and only consider the dynamics of the population up to this generation. To ease the notation, the dependence in N will be made implicit and we still denote by T the restriction of T to the first N generations. We now construct a tree $\widehat{T}_k$ with vertices locations $(\widehat{X}_u;\, u \in \widehat{T}_k)$ and k distinguished vertices $(V^1, \dots, V^k)$ at generation N. We start by constructing $\widehat{T}_1$ and V^1.

The 1-spine tree. The distribution of $\widehat{T}_1$ is that of a branching process where, at each generation, there is exactly one distinguished individual, which is marked. All unmarked individuals give birth according to the original collection of point processes $(\xi(x);\, x \in E)$. The marked individual gives birth according to a modified family of point processes $(\widehat{\xi}(x);\, x \in E)$, where $\widehat{\xi}(x)$ is such that for any functional F,

$$\mathbb{E}\Big[F(\widehat{\xi}(x))\Big] = \frac{1}{H(x)}\mathbb{E}\Big[\sum_{y \in \xi(x)} H(y) F(\xi(x))\Big]$$

for a fixed harmonic function H. (Note that this defines a probability distribution since H is harmonic.) All children but one of the marked particles are unmarked. Conditional on $\widehat{\xi} = \sum_i x_i$, the marked children is chosen to be the atom x_i of $\widehat{\xi}$ with probability proportional to $H(x_i)$. (Note that by construction of $\widehat{\xi}$, $\mathbb{P}(|\widehat{\xi}| > 0) = 1$, so that this is a well-defined procedure.)

As mentioned previously, this construction coincides with the classical spinal tree which has been derived in many contexts [199, 117]. Note that the 1-spine tree can be built in a Markovian way for all generations. This is essentially a consequence of the fact that the underlying change of measure is a martingale change of measure. We define $\widehat{T}_1$ as the restriction to the first N generations of the previous branching process, and V^1 as the unique marked vertex at generation N.

The k-spine tree. We now build a tree with k distinguished vertices by induction. Suppose that $\widehat{T}_k$, $(\widehat{X}_u;\, u \in \widehat{T}_k)$ and $(V^1, \dots, V^k)$ have been defined. For any vertex $u \in \Omega$ with $|u| = n$, let

$$[\![\emptyset, u]\!] = \bigcup_{t=1}^{n} (u_1, \dots, u_t)$$

denote the path from the root to u. Define

$$\mathbf{S}_k^i = \bigcup_{j=1}^{i} [\![\emptyset, V^i]\!]$$

as the subtree spanned by $(V^1, \dots, V^i)$ and let $\mathbf{S}_k = \mathbf{S}_k^k$. Further define the set of all children of $\mathbf{S}_k$ that do not belong to $\mathbf{S}_k$ as

$$B_k = \{v \in \widehat{T}_k : v = ui, u \in \mathbf{S}_k, i \geq 0\} \setminus \mathbf{S}_k.$$

The set B_k can be thought of as the set of “dangling ends” attached to the spine $\mathbf{S}_k$.

The $k+1$-spine will be defined under the assumption that

$$\mathbb{E}\Big[\sum_{u\in B_k} H(\widehat{X}_u)\Big] < \infty. \tag{5.4}$$

Let $\widetilde{T}_k$, $(\widetilde{X}_u;\ u \in \widetilde{T}_k)$ be a tree with k distinguished vertices $(\widetilde{V}^1, \dots, \widetilde{V}^k)$, such that the distribution of $[\widetilde{T}_k, (\widetilde{X}_u), (\widetilde{V}^i)]$ is that of $[\widehat{T}_k, (\widehat{X}_u), (V^i)]$, biased by $\sum_{u\in B_k} H(\widehat{X}_u)$. That is, $[\widetilde{T}_k, (\widetilde{X}_u), (\widetilde{V}^i)]$ is such that for any functionals F, (f_i), (g_u),

$$\mathbb{E}\Big[F(\widetilde{T}_k)\prod_{i=1}^{k} f_i(\widetilde{V}^i)\prod_{u\in\widetilde{T}_k} g_u(\widetilde{X}_u)\Big] = \frac{\mathbb{E}\Big[\sum_{u\in B_k} H(\widehat{X}_u)\cdot F(\widehat{T}_k)\prod_{i=1}^{k} f_i(V^i)\prod_{u\in\widehat{T}_k} g_u(\widehat{X}_u)\Big]}{\mathbb{E}\Big[\sum_{u\in B_k} H(\widehat{X}_u)\Big]}.$$

Conditional on $[\widetilde{T}_k, (\widetilde{X}_u), (\widetilde{V}^i)]$, let W be sampled in B_k in such a way that

$$\mathbb{P}\Big(W = u \mid \widetilde{T}_k, (\widetilde{X}_v), (\widetilde{V}^i)\Big) = \frac{H(\widetilde{X}_u)}{\sum_{v\in B_k} H(\widetilde{X}_v)}.$$

The tree $\widehat{T}_{k+1}$ is obtained by replacing the subtree rooted at W by an independent 1-spine tree, and the vertex V^{k+1} is defined to be the only marked vertex at generation N of this 1-spine tree. More formally, conditional on $\widetilde{X}_W$, let $[\widehat{T}', (\widehat{X}'_u), V']$ be an independent 1-spine tree, started at $\widehat{X}'_\emptyset = \widetilde{X}_W$, and stopped at generation $N - |W|$. Recall the notation $\widetilde{T}_k^W$ for the subtree attached at W:

$$\widetilde{T}_k^W = \Big\{u \in \widetilde{T}_k : \exists v \in \Omega^*, u = Wv\Big\}$$

and let

$$\widehat{T}_{k+1} = \Big(\widetilde{T}_k \setminus \widetilde{T}_k^W\Big) \cup \Big\{Wu : u \in \widehat{T}'\Big\}.$$

Moreover, we define

$$\forall u \in \widehat{T}', \quad \widehat{X}_{Wu} = \widehat{X}'_u,$$

and $\widehat{X}_u = \widetilde{X}_u$ for all other vertices. Finally, we define $V^{k+1} = WV'$ and $V^i = \widetilde{V}^i$ for $i \le k$.

Remark 5.10. The distribution of the k-spine is not “sampling consistent”: if $\widehat{T}_{k+1}$ and $(V^1, \dots, V^{k+1})$ denote the $k+1$-spine, the joint distribution of $\widehat{T}_{k+1}$ and $(V^1, \dots, V^k)$ is *not* that of the tree $\widehat{T}_k$ with distinguished vertices $(V^1, \dots, V^k)$. Therefore, we should be more careful in our notation and indicate the total number of distinguished vertices k when referring to a distinguished vertex V^i. As k will always be clear from the context, we choose to stick to the current notation to not make it heavier. ○

5.2.3 A many-to-few formula

We start with the following spinal decomposition result that connects the distribution of the 1-spine to that of the original branching process. The proof of this lemma is classical, see for instance [199], Theorem 4.3 for a proof in the case of branching random walks.

Lemma 5.11. *Let* $\mathbf{t}$ *be a tree of height* N*,* $(\varphi_u;\, u \in \mathbf{t})$ *be continuous bounded functionals, and* H *be a harmonic function in the sense of* (5.2)*. Then, for any* v *in* $\mathbf{t}$ *such that* $|v| = N$*,*

$$H(X_\varnothing)\mathbb{E}\Big[\mathbf{1}_{\widehat{T}_1=\mathbf{t}}\mathbf{1}_{V^1=v}\prod_{u\in\mathbf{t}}\varphi_u(\widehat{X}_u)\Big] = \mathbb{E}\Big[\mathbf{1}_{T=\mathbf{t}}H(X_v)\prod_{u\in\mathbf{t}}\varphi_u(X_u)\Big].$$

We now provide our k-spine decomposition result. Recall the notation $Z_t^{(k)}$ from the beginning of Section 5.2.

Theorem 5.12. *Fix* $k \geq 1$ *and a harmonic function* H*.*

(i) *If for any* $i \leq k$*,* $\mathbb{E}[Z_N^{(i)}] < \infty$*, then for any* $i < k$*, assumption* (5.4) *is fulfilled and*

$$\mathbb{E}\big[Z_N^{(k)}\big] = H(X_\varnothing)\prod_{i=1}^{k-1}\mathbb{E}\Big[\sum_{u\in B_i}H(\widehat{X}_u)\Big].$$

(ii) *Let* $\mathbf{t}$ *be a tree of height* N*, and* $(\varphi_u;\, u \in \mathbf{t})$ *be continuous bounded functions. Then, under the assumption of the previous point, for any distinct vertices* $v^1, \dots, v^k$ *in* $\mathbf{t}$ *such that* $|v^i| = N$*, we have*

$$\mathbb{E}\Big[\mathbf{1}_{\widehat{T}_k=\mathbf{t},V^1=v^1,\dots,V^k=v^k}\prod_{u\in\mathbf{t}}\varphi_u(\widehat{X}_u)\Big] = \frac{1}{\mathbb{E}\big[Z_N^{(k)}\big]}\mathbb{E}\Big[\mathbf{1}_{T=\mathbf{t}}\prod_{i=1}^{k}H(X_{v^i})\prod_{u\in\mathbf{t}}\varphi_u(X_u)\Big].$$

In particular marginal the distribution of $\widehat{T}_k$*, started from* x_0*, is* $\mathbf{Q}_{x_0}^{k,N}$*.*

Proof. We prove the result by induction. Fix some vertices $(v^1, \dots, v^{k+1})$ at generation N. By analogy with the construction of the k spine, let us define

$$\mathbf{s}_k = \bigcup_{i=1}^{k}[\![\varnothing, v^i]\!],$$

the subtree spanned by $(v^1, \dots, v^k)$. Define w as the oldest ancestor of v^{k+1} that does not belong to $\mathbf{s}_k$, that is, $w = (v_0^{k+1}, \dots, v_p^{k+1})$ where p is the unique generation such that $(v_0^{k+1}, \dots, v_{p-1}^{k+1}) \in \mathbf{s}_k$ but $(v_0^{k+1}, \dots, v_p^{k+1}) \notin \mathbf{s}_k$. Finally, let $\mathbf{t}_w$ be the subtree of $\mathbf{t}$ attached to w:

$$\mathbf{t}_w = \{u \in \Omega^* : wu \in \mathbf{t}\}$$

and define $w\mathbf{t}_w = \{wu : u \in \mathbf{t}_w\}$. Let also v' be the unique vertex so that $v^{k+1} = wv'$.

By the branching property, we have that

$$\begin{aligned}
&\mathbb{E}\Big[\mathbf{1}_{T=\mathbf{t}}\prod_{i=1}^{k+1}H(X_{v^i})\prod_{u\in\mathbf{t}}\varphi_u(X_u)\Big]\\
&\qquad=\mathbb{E}\Big[\mathbf{1}_{T\setminus T^w=\mathbf{t}\setminus(w\mathbf{t}_w)}\prod_{i=1}^{k}H(X_{v^i})\prod_{u\in\mathbf{t}\setminus w\mathbf{t}_w}\varphi_u(X_u)\\
&\qquad\qquad\times\mathbb{E}\Big[\mathbf{1}_{T'=\mathbf{t}_w}H(X'_{v'})\prod_{u\in\mathbf{t}_w}\varphi_{wu}(X'_u)\mid X'_\varnothing=X_w\Big]\Big],
\end{aligned}$$

where T' is an independent copy of T, started at X_w. Moreover, by Lemma 5.11,

$$\begin{aligned}
&\mathbb{E}\Big[\mathbf{1}_{T'=\mathbf{t}_w}H(X'_{v'})\prod_{u\in\mathbf{t}_w}\varphi_{wu}(X'_u)\mid X'_\varnothing=X_w\Big]\\
&\qquad=H(X_w)\mathbb{E}\Big[\mathbf{1}_{\widehat{T}'=\mathbf{t}_w}\mathbf{1}_{V'=v'}\prod_{u\in\mathbf{t}_w}\varphi_{wu}(\widehat{X}'_u)\mid\widehat{X}'_\varnothing=X_w\Big]
\end{aligned}$$

where $[\widehat{T}',(\widehat{X}'_u),V']$ is a 1-spine tree. Now, summing first over all v', then over all w shows that

$$\begin{aligned}
&\sum_{\substack{|v^{k+1}|=N\\ \forall i,\,v^{k+1}\neq v^i}}\mathbb{E}\Big[\mathbf{1}_{T=\mathbf{t}}\prod_{i=1}^{k+1}H(X_{v^i})\prod_{u\in\mathbf{t}}\varphi_u(X_u)\Big]\\
&\qquad=\mathbb{E}\Big[\mathbf{1}_{T_k\setminus T_k^w=\mathbf{t}\setminus(w\mathbf{t}_w)}\prod_{i=1}^{k}H(X_{v^i})\prod_{u\in\mathbf{t}\setminus w\mathbf{t}_w}\varphi_u(X_u)\\
&\qquad\qquad\times\sum_{w\in b_k}H(X_w)\times\mathbb{E}\Big[\mathbf{1}_{\widehat{T}'=\mathbf{t}_w}\prod_{u\in\mathbf{t}_w}\varphi_{wu}(\widehat{X}'_u)\mid\widehat{X}'_\varnothing=X_w\Big]\Big],
\end{aligned}$$

where b_k is defined as B_k, replacing $\mathbf{S}_k$ by $\mathbf{s}_k$. By induction, we can write

$$\begin{aligned}
&\sum_{\substack{|v^{k+1}|=N\\ \forall i,\,v^{k+1}\neq v^i}}\mathbb{E}\Big[\mathbf{1}_{T=\mathbf{t}}\prod_{i=1}^{k+1}H(X_{v^i})\prod_{u\in\mathbf{t}}\varphi_u(X_u)\Big]\\
&\qquad=\mathbb{E}\Big[Z_N^{(k)}\Big]\mathbb{E}\Big[\mathbf{1}_{\widehat{T}_k\setminus\widehat{T}_k^w=\mathbf{t}\setminus(w\mathbf{t}_w)}\mathbf{1}_{V^1=v^1,\dots,V^k=v^k}\prod_{u\in\mathbf{t}\setminus w\mathbf{t}_w}\varphi_u(\widehat{X}_u)\\
&\qquad\qquad\times\sum_{w\in b_k}H(\widehat{X}_w)\times\mathbb{E}\Big[\mathbf{1}_{\widehat{T}'=\mathbf{t}_w}\prod_{u\in\mathbf{t}_w}\varphi_{wu}(\widehat{X}'_u)\mid\widehat{X}'_\varnothing=\widehat{X}_w\Big]\Big].
\end{aligned}$$

Further setting $\varphi_u\equiv 1$ and summing first over all distinct $(v^1,\dots,v^k)$, then over all $\mathbf{t}$ proves that

$$\mathbb{E}\Big[Z_N^{(k+1)}\Big]=\mathbb{E}\Big[Z_N^{(k)}\Big]\mathbb{E}\Big[\sum_{w\in B_k}H(\widehat{X}_w)\Big]$$

yielding the first part of the result.

For the second part of the result, by induction and by definition of $[\widetilde{T}_k, (\widetilde{X}_u), (\widetilde{V}^i)]$ we have

$$\begin{aligned}
&\mathbb{E}\Big[\mathbf{1}_{T=\mathbf{t}} \prod_{i=1}^{k+1} H(X_{v^i}) \prod_{u\in\mathbf{t}} \varphi_u(X_u)\Big] \\
&\qquad = \mathbb{E}\Big[\mathbf{1}_{T\setminus T^w=\mathbf{t}\setminus(w\mathbf{t}_w)} \prod_{i=1}^{k} H(X_{v^i}) \prod_{u\in\mathbf{t}\setminus w\mathbf{t}_w} \varphi_u(X_u) \\
&\qquad\qquad \times H(X_w) \times \mathbb{E}\Big[\mathbf{1}_{\widehat{T}'=\mathbf{t}_w} H(\widehat{X}'_{v'}) \prod_{u\in\mathbf{t}_w} \varphi_{wu}(\widehat{X}'_u) \mid \widehat{X}'_\varnothing = X_w\Big]\Big], \\
&\qquad \propto \mathbb{E}\Big[\mathbf{1}_{\widehat{T}_k\setminus\widehat{T}^w_k=\mathbf{t}\setminus(w\mathbf{t}_w)} \mathbf{1}_{V^1=v^1,\dots,V^k=v^k} \prod_{u\in\mathbf{t}\setminus w\mathbf{t}_w} \varphi_u(\widehat{X}_u) \\
&\qquad\qquad \times H(\widehat{X}_w) \times \mathbb{E}\Big[\mathbf{1}_{\widehat{T}'^*=\mathbf{t}^*_w} \mathbf{1}_{V'=v'} \prod_{u\in\mathbf{t}^*_w} \varphi_{wu}(\widehat{X}'_u) \mid \widehat{X}'_\varnothing = \widehat{X}_w\Big]\Big], \\
&\qquad \propto \mathbb{E}\Big[\mathbf{1}_{\widetilde{T}_k\setminus\widetilde{T}^w_k=\mathbf{t}\setminus(w\mathbf{t}_w)} \mathbf{1}_{\widetilde{V}^1=v^1,\dots,\widetilde{V}^k=v^k} \prod_{u\in\mathbf{t}\setminus w\mathbf{t}_w} \varphi_u(\widetilde{X}_u) \\
&\qquad\qquad \times \frac{H(\widetilde{X}_w)}{\sum_{v\in b_k} H(\widetilde{X}_v)} \times \mathbb{E}\Big[\mathbf{1}_{\widehat{T}'=\mathbf{t}_w} \mathbf{1}_{V'=v'} \prod_{u\in\mathbf{t}_w} \varphi_{wu}(\widehat{X}'_u) \mid \widehat{X}'_\varnothing = \widetilde{X}_w\Big]\Big], \\
&\qquad = \mathbb{E}\Big[\mathbf{1}_{\widehat{T}_{k+1}=\mathbf{t}} \mathbf{1}_{V^1=v^1,\dots,V^{k+1}=v^{k+1}} \prod_{u\in\mathbf{t}} \varphi_u(\widehat{X}_u)\Big],
\end{aligned}$$

ending the proof. □

Our many-to-few formula is now a simple corollary of the previous spinal decomposition result. In the next result, we only consider functionals that depend on the locations of k vertices $(v^1, \dots, v^k)$ at generation n. It would also follow from the spinal decomposition theorem that we can express any functional of the subtree spanned by $(v^1, \dots, v^k)$ in terms of the k-spine only.

Corollary 5.13 (Many-to-few)**.** *For $k \geq 1$, let $(g_1, \dots, g_k)$ be continuous bounded functions. Then*

$$\mathbb{E}\Big[\sum_{\substack{v^1\neq\cdots\neq v^k \\ |v^i|=n}} \prod_{i=1}^{k} g_i(X_{v^i})\Big] = \mathbb{E}\big[Z_N^{(k)}\big]\ \mathbb{E}\Big[\prod_{i=1}^{k} g_i(\widehat{X}_{V^i}) H(\widehat{X}_{V^i})^{-1}\Big].$$

Proof. Consider a fixed $\mathbf{t}$ and $(v^1, \dots, v^k)$. An application of Theorem 5.12 with

$$\varphi_u = \begin{cases} g_i & \text{if } u = v^i \\ 1 & \text{else,} \end{cases}$$

shows that

$$\mathbb{E}\Big[\mathbf{1}_{T=\mathbf{t}} \prod_{i=1}^{k} g_i(X_{v^i})\Big] = \mathbb{E}\big[Z_N^{(k)}\big]\ \mathbb{E}\Big[\mathbf{1}_{\widehat{T}_k=\mathbf{t}} \mathbf{1}_{V^1=v^1,\dots,V^k=v^k} \prod_{i=1}^{k} g_i(\widehat{X}_{V^i}) H(\widehat{X}_{V^i})^{-1}\Big].$$

The result now follows by summing first over all vertices at generation n in $\mathbf{t}$, and then over all trees $\mathbf{t}$. □

5.2.4 Distribution of the k-spine

In the previous sections, we have defined jointly the tree $\widehat{T}_k$ with k distinguished vertices $(V^1, \dots, V^k)$, spanning the subtree $\mathbf{S}_k$ of $\widehat{T}_k$. The many-to-few formula shows that in order to understand the distribution of the subtree spanned by k individuals sampled uniformly in the population, it is sufficient to study the tree $\mathbf{S}_k$. The objective of this section is to construct $\mathbf{S}_k$ in an autonomous way, and give distribution of $\widehat{T}_k$ conditional on $\mathbf{S}_k$.

Let us define $\widehat{\xi}_u$ as the point process giving the position of the children of u in $\widehat{T}_k$, defined as

$$\widehat{\xi}_u = \sum_{ui \in \widehat{T}_k} \delta(\widehat{X}_{ui}).$$

A first, direct consequence of Theorem 5.12 is the following result.

Corollary 5.14. *Conditional on $\mathbf{S}_k$, $(\widehat{X}_u;\, u \in \mathbf{S}_k)$ and $(\widehat{\xi}_u;\, u \in \mathbf{S}_k, |u| < N)$, the tree $\widehat{T}_k$ is recovered by grafting for each $u \in B_k$ an independent subtree T_u that has the original distribution T started from $\widehat{X}_u$.*

The previous corollary shows that all what needs to be understood is the joint distribution of $\mathbf{S}_k$ and of the children of $\mathbf{S}_k$. The tree $\mathbf{S}_k$ has been defined as a subset of $\widehat{T}_k$. This encoding is not convenient for our purpose, as the labeling of individuals in $\mathbf{S}_k$ contains information about their number of siblings. Our first task is to re-encode $\mathbf{S}_k$ as a collection $\big((X^i_t), K^i, L^i;\, i \le k\big)$. In this encoding, we envision $\mathbf{S}_k$ as being built inductively by grafting the branch $\mathbf{S}_k \setminus \mathbf{S}_k^{k-1}$ on $\mathbf{S}_k$. The variable L^k is the length of that branch, $(X^k_t;\, t \le L^k)$ the individual's location along the branch, and K^k the label of branch onto which $\mathbf{S}_k \setminus \mathbf{S}_k^{k-1}$ is grafted. Let us give a formal definition of these quantities.

Recall that $\mathbf{S}_k^i$ stands for the subtree spanned by $(V^1, \dots, V^i)$. For $u \in \mathbf{S}_k$, let us define the label of u as

$$\kappa_u = \inf\{i \le k : u \in \mathbf{S}_k^i\}.$$

This label is the unique $i \le k$ such that $u \in \mathbf{S}_k^i \setminus \mathbf{S}_k^{i-1}$. Moreover, define W^i as the youngest individual in $\mathbf{S}_k^i \setminus \mathbf{S}_k^{i-1}$. Any $u \in \mathbf{S}_k^i \setminus \mathbf{S}_k^{i-1}$ can be written as $W^i v$ with $v \in \Omega$. Let us define $t_u = |v| + 1$ and $L^i = N - |W^i| + 1$. The variable L^i is the length of $\mathbf{S}_k^i \setminus \mathbf{S}_k^{i-1}$, and t_u is the generation of individual u, re-indexed in such a way that W^i belongs to the first generation. The map

$$u \in \mathbf{S}_k \setminus \{\emptyset\} \mapsto (\kappa_u, t_u) \in \{(i,t) : i \le k,\ 1 \le t \le L^i\}$$

is a bijection. For $i \le k$ and $t \le L^i$, we denote by $u^i_t \in \mathbf{S}_k$ the individual in the spine corresponding to (i,t), and set X^i_t to be the location of u^i_t, that is,

$$X^i_t = X_{W^i v},$$

where v is the unique element of Ω verifying $W^i v \in \mathbf{S}_k^i$, and $|v| = t-1$. Note that $X^i_1 = X_{W^i}$ is the location of the oldest vertex in $\mathbf{S}_k^i \setminus \mathbf{S}_k^{i-1}$ and $X^i_{L^i} = X_{V^i}$ is the location of the i-th marked leaf. Finally we define

$$K^i = \inf\{j < i : \overleftarrow{W}^i \in \mathbf{S}_k^j\}$$

to be the label of the parent of W^i, that is, the label of the vertex of $\mathbf{S}_k^{i-1}$ on which $\mathbf{S}_k^i \setminus \mathbf{S}_k^{i-1}$ is grafted.

Finally, for (i,t) such that $i \leq k$ and $t < L^i$, we define

$$\zeta_t^i = \sum_{u_t^i j \notin \mathbf{S}_k} \delta(X_{u_t^i j}), \quad \bar{\zeta}_t^i = \sum_{u_t^i j \in \mathbf{S}_k} \delta(X_{u_t^i j}),$$

to be the location of the children that do not belong to $\mathbf{S}_k$, and do belong to $\mathbf{S}_k$ respectively. The following result gives the distribution of (ζ_t^i) conditional on the spine. We restrict our attention to the case where the point processes $(\xi(x);\, x \in E)$ are Poisson point processes, as is the case for the branching process with recombination. However, we provide an alternative statement of this result in the general case, which makes use of the Palm measures of $(\xi(x);\, x \in E)$ is Section 5.A. The first point of the following result does not require the Poisson assumption.

Proposition 5.15. *Let us assume that all the point processes $(\xi(x);\, x \in E)$ are Poisson point processes. Then the following holds.*

(i) *For $k = 1$, the process $(X_t^1;\, t \geq 0)$ is a Markov process. Its transition is given by*

$$\mathbb{E}\Big[f(X_{t+1}^1) \mid X_t^1 = x\Big] = \frac{1}{H(x)} \int_E H(y) f(y)\, \mu(x, \mathrm{d}y),$$

for a continuous bounded function f, and where $\mu(x, \mathrm{d}y)$ is the intensity measure of the point process $\xi(x)$.

(ii) *For any $k \geq 1$, conditional on $(X^i, K^i, L^i;\, i \leq k)$, the collection of point processes $(\zeta_t^i;\, i \leq k,\, t < L^i)$ is independent. Moreover, ζ_t^i is distributed as $\xi(X_{u_t^i})$.*

Proof of Proposition 5.15. Let us first prove point (i) and point (ii) for $k = 1$, and then point (ii) for any k by induction. Recall the construction of the 1-spine and its siblings. The population start from one marked particle. At generation t, if X_t^1 denotes the location of the marked particle, the location of its offspring is $X_t^1 + \widehat{\xi}(X_t^1)$, where $\widehat{\xi}(X_t^1)$ is an independent point process, whose distribution is that of the original point process $\xi(X_t^1)$, biased by $\sum_{y \in \xi(X_t^1)} H(y)$. The marked particle is chosen among the atoms of $X_t^1 + \widehat{\xi}(X_t^1)$ in such a way that an atom located at y is chosen with probability proportional to $H(y)$. According to Lemma 5.26, this reproduction step can be achieved by first choosing the location X_{t+1}^1 of the marked particle so that conditional on $X_t^1 = x$, X_{t+1}^1 is distributed as $H(y)\mu(x, \mathrm{d}y)$. Then, conditional on $X_{t+1}^1 = y$ the siblings of the marked particles are distributed as $\xi^{!,y}(x)$, where $\xi^{!,y}(x)$ has the reduced Palm distribution of $\xi(x)$, conditional on having an atom at y. This yields the result for $k = 1$. It is a well-known fact that the reduced Palm distribution of a Poisson point process Φ conditioned on having an atom at any location x is the original distribution of Φ, see for instance Theorem 3.2.4 in [4]. Thus, the distribution of ζ^1_t is that of $\xi(X^1_t)$, and is independent of X^1_{t+1}, yielding the result for $k = 1$.

Let us now assume that (ii) holds for some $k \geq 1$. The $k+1$-spine is obtained by first biasing the k-spine by

$$\sum_{u \in B_k} H(X_u) = H(X^1_0) + \sum_{i=1}^{k} \sum_{t=1}^{L^i - 1} \langle H, \zeta^i_t \rangle, = \sum_{(i,t)} \langle H, \zeta^i_t \rangle,$$

and we denote by $(\widetilde{X}^i, \widetilde{K}^i, \widetilde{L}^i;\, i \leq k)$ and by $(\widetilde{\zeta}^i_t)$ the corresponding biased variables. Then, an atom belonging to one of the point processes $(\widetilde{\zeta}^i_t)$ is chosen in such a way that, if it is located at y it is chosen with probability proportional to $H(y)$. Then, an independent 1-spine is grafted to this atom. From now on, let us work conditional on $(\widetilde{X}^i, \widetilde{K}^i, \widetilde{L}^i;\, i \leq k)$. If (I, T) denote respectively the label and re-indexed generation of the parent of the chosen atom, then for any collection of bounded continuous functions $(F_{(j,s)})$, by induction hypothesis,

$$\begin{aligned}
\mathbb{E}\Big[\mathbf{1}_{(I,T)=(i,t)} \prod_{(j,s)} F_{(j,s)}(\widetilde{\zeta}^j_s)\Big] &= \mathbb{E}\Big[\frac{\langle H, \widetilde{\zeta}^i_t \rangle}{\sum_{(j,s)} \langle H, \widetilde{\zeta}^j_s \rangle} \prod_{(j,s)} F_{(j,s)}(\widetilde{\zeta}^j_s)\Big] \\
&\propto \mathbb{E}\Big[\langle H, \zeta^i_t \rangle \prod_{(j,s)} F_{(j,s)}(\zeta^j_s)\Big] \\
&= \prod_{(j,s) \neq (i,t)} \mathbb{E}\Big[F_{(j,s)}(\zeta^j_s)\Big] \mathbb{E}\Big[\langle H, \zeta^i_t \rangle F_{(i,t)}(\zeta^i_s)\Big].
\end{aligned}$$

This calculation shows that

$$\mathbb{P}\Big((I, T) = (i, t)\Big) \propto \mathbb{E}\Big[\langle H, \zeta^i_t \rangle\Big]$$

and that, conditional on $(I, T) = (i, t)$, the point processes $(\widetilde{\zeta}^j_s)$ remain independent, and $\widetilde{\zeta}^i_t$ has the distribution of ζ^i_t biased by $\langle H, \zeta^i_t \rangle$ while all other point processes have their original distribution, that is, are distributed as $(\xi(\widetilde{X}^j_s))$. Now, note that conditional on $(I, T) = (i, t)$, the selected atom of $\widetilde{\zeta}^i_t$, and is chosen with probability proportional to $H(y)$. Lemma 5.26 again tells us that, if ν denotes the intensity measure of ζ^i_t and X' is the location of the chosen atom, then its distribution is such that

$$\mathbb{E}\Big[f(X')\Big] \propto \int H(y) f(y)\, \nu(\mathrm{d}y).$$

Moreover, conditional on X', $\widetilde{\zeta}^i_t - \delta(X')$ has the reduced Palm distribution of ζ^i_t, conditional on having an atom located at X'. By induction, ζ^i_t is distributed as $\xi(X^i_t)$, and thus its reduced Palm distribution is again that of $\xi(X^i_t)$, yielding the result. □

As a corollary of Proposition 5.15, we have the following inductive procedure to build the sequence of spines autonomously. Let us assume that the k-spine $(X^i, K^i, L^i;\, i \leq k)$ has been constructed. The $k+1$-spine can then be constructed inductively as follows.

1. Let $((\widetilde{X}^i_t), \widetilde{K}^i, \widetilde{L}^i;\, i \leq k)$ have the distribution of $((X^i_t), K^i, L^i;\, i \leq k)$ biased by
$$\sum_{i=1}^{k} \sum_{t=1}^{L^i-1} \mathbb{E}\Big[\langle H, \zeta^i_t \rangle\Big] = \sum_{i=1}^{k} \sum_{t=1}^{L^i-1} H(X^i_t).$$
2. Choose (I, T) such that
$$\mathbb{P}\Big((I,T) = (i,t) \mid ((\widetilde{X}^i_t), \widetilde{K}^i, \widetilde{L}^i;\, i \leq k)\Big) \propto H(\widetilde{X}^i_t).$$
3. Conditional on $(I,T) = (i,t)$, let $L^{k+1} = \widetilde{L}^i - t$, $K^{k+1} = i$, and $(X^{k+1}_t;\, t \leq L^{k+1})$ be an independent 1-spine started from $X^{k+1}_0 = X^I_T$.

5.3 The infinite population size limit

We now apply our spinal decomposition theorem to the branching process with recombination. This short section contains the description of the k-spine of the branching process with recombination in the large N limit. Most of the work for proving our main theorems will be carried out in the forthcoming Section 5.4.

5.3.1 Limit of the 1-spine

Recall Definition 5.1, where the branching process with recombination is constructed as a Markov branching process valued in the subintervals of $[0, R]$. The key property that allows the study of this branching process is that the total amount of genetic material, that is, the sum of the lengths of the blocks, is a martingale. Rephrased in the terminology of Section 5.2, the function
$$H \colon I \mapsto |I|$$
is harmonic. This is checked by an easy calculation.

Let us write $(I^N_t;\, t \geq 0)$ for 1-spine of the branching process with recombination, in a population of size N. We know from Section 5.2.4 that it is a discrete-time Markov process valued in the intervals whose transitions are given by
$$\begin{aligned}\mathbb{E}\Big[F(I^N_{t+1}) \mid I^N_t = [a,b]\Big] &= \frac{1}{\rho} \int |I| F(I)\, \mu\Big(([a,b], \mathrm{d}I\Big) \\ &= \frac{1-\rho}{N} F([a,b]) + \frac{\rho}{N}\frac{1}{\rho^2} \int_0^\rho x F\Big([a, a+x]\Big)\, \mathrm{d}x + \frac{\rho}{N}\frac{1}{\rho^2} \int_0^\rho x F\Big([b-x, b]\Big)\, \mathrm{d}x \\ &= \frac{1-\rho}{N} F([a,b]) + \frac{\rho}{2N} \int_0^1 2u F\Big([a, a+\rho u]\Big)\, \mathrm{d}u + \frac{\rho}{2N} \int_0^1 2u F\Big([b-\rho u, b]\Big)\, \mathrm{d}u\end{aligned}$$
for any functional F, and where $\rho = b - a$ is the length of $[a,b]$.

Thus, the transition of the 1-spine can be described as follows. Conditional on $I^N_t = [a,b]$, and with $\rho = b - a$:

- with probability $1 - \rho/N$, no recombination occurs and $I^N_{t+1} = [a,b]$;

- with probability ρ/N, a recombination occurs, and either $I^N_{t+1} = [a, a + \rho U^*]$ or $I^N_{t+1} = [b - \rho U^*, b]$ with equal probability, where $U^* \sim 2x\,\mathrm{d}x$ has the size-biased distribution of a uniform variable on $[0, 1]$.

The following convergence result is straightforward from the previous description of the transitions of the 1-spine.

Lemma 5.16. *Let $(I^N_t;\, t \geq 0)$ be the 1-spine with population size N and initial chromosome size R. The following convergence holds in distribution for the Skorohod topology:*

$$\Big(I^N_{\lfloor Nt\rfloor};\, t \geq 0\Big) \longrightarrow \Big(I_t;\, t \geq 0\Big),$$

where $(I_t;\, t \geq 0)$ is a Markov process started at $I_0 = [0, R]$ and such that, conditional on $I_t = [a, b]$, it jumps to $[a, a + \rho U^*]$ at rate $\rho/2$ and to $[b - \rho U^*, b]$ at rate $\rho/2$, where $U^* \sim 2x\,\mathrm{d}x$ and $\rho = b - a$.

Proof. The result follows by noting that the inter-jump times of $(I^N_{\lfloor Nt\rfloor};\, t \geq 0)$ converge to exponential variables with the parameters corresponding to the transition rates described above. □

Let us denote by $\rho_t = |I_t|$ the length of the block of the limiting 1-spine at time t. A direct consequence of the previous result is that, in the limit, the process $(\rho_t;\, t \geq 0)$ is also a Markov process. Conditional on $\rho_t = \rho$, it jumps to ρU^* at rate ρ, where $U^* \sim 2x\,\mathrm{d}x$. It is important to note that $(\rho_t;\, t \geq 0)$ is a *self-similar* process, in the sense that for any $c > 0$,

$$(c\rho_t;\, t \geq 0) \overset{\text{(d)}}{=} (\rho'_{ct};\, t \geq 0),$$

where $(\rho'_t;\, t \geq 0)$ is a copy of $(\rho_t;\, t \geq 0)$ started from $c\rho_0$, see for instance [168].

Let us end this section by providing a Poissonian construction of the 1-spine $(I_t;\, t \geq 0)$. Let $\mathscr{Q}$ be a homogeneous Poisson point process on $\mathbb{R}_+ \times \mathbb{R}_+$ with rate 1. Let V be an independent uniform variable on $[0, R]$. At time t, the set

$$[0, R] \setminus \{x : (x, s) \in \mathscr{Q},\, s \leq t\}$$

is the union of finitely many subintervals of $[0, R]$. Let I_t be the subinterval to which V belongs. Then $(I_t;\, t \geq 0)$ is distributed as the 1-spine of the branching process with recombination.

5.3.2 Limit of the k-spine

We will now prove an analogous convergence result for the k-spine. Here and later in this work, we will make repeated use of the following elementary fact, that we isolate as a lemma.

Lemma 5.17. *Let (X_n) be a sequence of positive, integrable random variables, and (Y_n) a sequence of random variables in any topological space. Suppose that, for $n \geq 1$, Z_n has the distribution of Y_n, biased by X_n. Then, if*

$$(X_n, Y_n) \longrightarrow (X, Y)$$

in distribution, if (X_n) *is a uniformly integrable family, and if* $\mathbb{P}(X > 0) > 0$*, we have that*

$$Z_n \longrightarrow Z$$

in distribution, where Z *has the distribution of* Y*, biased by* X*.*

Proof. Let φ be any continuous bounded real function on the state space of (Y_n). Then

$$\mathbb{E}\Big[\varphi(Z_n)\Big] = \frac{\mathbb{E}[X_n \varphi(Y_n)]}{\mathbb{E}[X_n]} \longrightarrow \frac{\mathbb{E}[X\varphi(Y)]}{\mathbb{E}[X]}$$

where the convergence follows from the fact that both (X_n) and $(X_n\varphi(Y_n))$ are uniformly integrable families of random variables. □

The following result provides the large N limit of the k-spine, as well as the inductive construction of the k-spine analogous to that of Section 5.2.4.

Proposition 5.18. *We any* $k \geq 1$*, let*

$$\Big((I_n^{i,N}), K^{i,N}, L^{i,N};\, i \leq k\Big)$$

be the k*-spine with population size* N *and initial chromosome size* R*, constructed until time* $\lfloor \sigma N \rfloor$ *for* $\sigma > 0$*. Then there exists a collection of random variables* $((I_t^i), K^i, L^i)$ *such that the following convergence holds in distribution:*

$$\Big((I_{\lfloor tN \rfloor}^{i,N}), K^{i,N}, \frac{L^{i,N}}{N}\Big) \longrightarrow \Big((I_t^i), K^i, L^i\Big).$$

Moreover, writing $\rho_t^i = |I_t^i|$*, in the limit the distribution of the* $k{+}1$*-spine is obtained inductively from the* k*-spine by:*

1. *Letting* $\Big((\widetilde{I}_t^i), \widetilde{K}^i, \widetilde{L}^i;\, i \leq k\Big)$ *have the law of* $\Big((I_t^i), K^i, L^i;\, i \leq k\Big)$ *biased by*

$$\sum_{i=1}^{k} \int_0^{L^i} \rho_t^i \,\mathrm{d}t.$$

2. *Sampling* K^{k+1} *so that*

$$\mathbb{P}\Big(K^{k+1} = i \mid \Big((\widetilde{I}_t^i), \widetilde{K}^i, \widetilde{L}^i;\, i \leq k\Big)\Big) = \frac{\displaystyle\int_0^{\widetilde{L}^i} \widetilde{\rho}_t^i \,\mathrm{d}t}{\displaystyle\sum_{j=1}^{k} \int_0^{\widetilde{L}^j} \widetilde{\rho}_t^j \,\mathrm{d}t}.$$

3. *Drawing* L^{k+1} *according to*

$$\mathbb{E}\Big[f(L^{k+1}) \mid K^{k+1} = i\Big] \propto \int_0^{L^i} f(t)\widetilde{\rho}^i_{\widetilde{L}^i - t} \,\mathrm{d}t.$$

4. *Conditional on K^{k+1} and L^{k+1}, letting $(I^{k+1}_t;\, t \le L^{k+1})$ be an independent* 1*-spine started from*
$$I_0^{k+1} = \widetilde{I}^i_{L^{k+1} - \widetilde{L}^{K^{k+1}}}.$$

Proof. We prove the result by induction. The case $k = 1$ has been treated in Lemma 5.16.

Assume that the result holds for some k. Recalling that for the k-spine we consider the harmonic function $H \colon I \mapsto |I|$, the $k+1$-spine is obtained by first biasing the distribution of the k-spine by

$$\sum_{i=1}^{k} \sum_{t=1}^{L^{i,N}-1} \rho_n^{i,N},$$

which is equivalent to biasing it by

$$\sum_{i=1}^{k} \frac{1}{N} \sum_{n=1}^{L^{i,N}-1} \rho_n^{i,N}.$$

Moreover, the convergence of the $\Big((\rho^{i,N}_{\lfloor tN \rfloor}), L^{i,N}/N;\, i \le k\Big)$ implies the following convergence in distribution

$$\sum_{i=1}^{k} \frac{1}{N} \sum_{n=1}^{L^{i,N}-1} \rho_n^{i,N} \longrightarrow \sum_{i=1}^{k} \int_0^{L^i} \rho_t^i \, \mathrm{d}t.$$

Using the inequality
$$\sum_{i=1}^{k} \frac{1}{N} \sum_{n=1}^{L^{i,N}-1} \rho_n^{i,N} \le kR$$

to obtain uniform integrability, Lemma 5.17 shows that

$$\Big((\widetilde{\rho}^{i,N}_{\lfloor tN \rfloor}), \widetilde{K}^{i,N}, \frac{\widetilde{L}^{i,N}}{N};\, i \le k\Big) \longrightarrow \Big((\widetilde{\rho}^i_t), \widetilde{K}^i, \widetilde{L}^i;\, i \le k\Big),$$

where the limiting variables have the distribution of $((\rho^i_t), K^i, L^i;\, i \le k)$, biased by

$$\sum_{i=1}^{k} \int_0^{L^i} \rho_t^i \, \mathrm{d}t.$$

That $K^{k+1,N}$ and $L^{k+1,N}$ converge to their respective limits is now straightforward. The last statement of the proposition is again a consequence of Lemma 5.16. □

5.4 The infinite chromosome size limit

We now consider the recombination rate limit of the branching process with recombination. Again, we first derive the limiting behavior of the 1-spine, and then that of the k-spine by induction.

5.4.1 Limit of the 1-spine

Let $(\rho_t;\, t \geq 0)$ be the block length of the continuous-time 1-spine, started at some fixed R_0. Let us consider the following random time-change:

$$\forall t \geq 0, \quad \theta_t = \int_0^t \rho_u \,\mathrm{d}u, \quad \tau_t = \theta_t^{-1},$$

and the time-changed process

$$\forall t \geq 0, \quad S_t = \rho \circ \tau_t.$$

Then, $(S_t;\, t \geq 0)$ is a Markov process such that, conditional on $S_t = \rho$, (S_t) jumps at rate one to $U^* \rho$, where $U^* \sim 2x \,\mathrm{d}x$.

Recall that $(\rho_t;\, t \geq 0)$ is a self-similar process, so that the time-change τ_t is the well-known Lamperti transform of the process [143]. The time-changed process $(S_t;\, t \geq 0)$ is the exponential of a surbordinator.

The following proposition provides the limiting distribution of the 1-spine for large R.

Proposition 5.19. *Let us assume that, as $R \to \infty$,*

$$\frac{\log R_0}{\log R} \longrightarrow \gamma.$$

The following convergences hold in distribution, as $R \to \infty$.

(i) *For any fixed $t > 0$, we have*

$$\frac{1}{\log R} \int_0^t \rho_u \,\mathrm{d}u \longrightarrow 2\gamma,$$

(ii) *For any fixed $t > 0$,*

$$\rho_t \longrightarrow Y_t.$$

where Y_t is a $\mathrm{Gamma}(2, t)$ *variable.*

(iii) *Finally*

$$\Big(\frac{\log S_{u \log R}}{\log R};\, u \geq 0\Big) \longrightarrow \big(\gamma - u/2;\, u \geq 0\big),$$

in distribution in the Skorohod topology.

Proof. To prove the result, let us build (ρ_t) and (S_t) jointly as follows. Consider an i.i.d. sequence $(T_i)_{i \geq 0}$ of Exponential(1) variables with mean 1, and an independent i.i.d. sequence $(U_i^*)_{i \geq 1}$ of variables on $[0, 1]$ distributed as $2x \,\mathrm{d}x$. Let us define

$$\forall i \geq 0, \quad \rho_i = R_0 \prod_{j=1}^{i} U_i^*$$

with the convention that $\rho_0 = R_0$. Define

$$\forall t \geq 0, \quad N_t = \inf\Big\{ i \geq 0 : \sum_{j=0}^{i} \frac{T_i}{\rho_i} > t \Big\}, \quad \rho_t = \rho_{N_t}.$$

Then it should be clear that $(\rho_t;\, t \geq 0)$ is distributed as a 1-spine started from R_0. Let us further define

$$\forall t \geq 0, \quad M_t = \inf\Big\{ i \geq 0 : \sum_{j=0}^{i} T_i > t \Big\}, \quad S_t = \rho_{M_t}.$$

Again, it is readily checked that $(S_t;\, t \geq 0)$ is the time-changed version of the process $(\rho_t;\, t \geq 0)$.

Let us start with the convergence of $\int_0^t \rho_u \, \mathrm{d}u / \log R$. The following inequality holds a.s.

$$\sum_{i=0}^{N_t - 1} T_i \leq \int_0^t \rho_u \, \mathrm{d}u \leq \sum_{i=0}^{N_t} T_i$$

so that, by the law of large numbers, it is sufficient to show that $N_t / \log R$ converges to 2γ in probability.

To ease the notation, let us define, for $\varepsilon > 0$,

$$n_1 = \lfloor 2\gamma(1+\varepsilon)\log R \rfloor, \quad n_0 = \lfloor 2\gamma(1+\varepsilon/2)\log R \rfloor.$$

We have

$$\begin{aligned}
\mathbb{P}\Big(\frac{N_t}{\log R} \geq 2\gamma(1+\varepsilon) \Big) &= \mathbb{P}\Big(\sum_{j=0}^{n_1} \frac{T_i}{\rho_i} \leq t \Big) \leq \mathbb{P}\Big(\sum_{j=n_0}^{n_1} \frac{T_i}{\rho_i} \leq t \Big) \\
&\leq \mathbb{P}\Big(\frac{1}{\rho_{n_0}} \sum_{j=n_0}^{n_1} T_i \leq t \Big).
\end{aligned}$$

Moreover,

$$\begin{aligned}
\frac{\log \rho_{n_0}}{\log R} &= \frac{\log R_0}{\log R} + \frac{1}{\log R} \sum_{i=1}^{n_0} \log U_i^* \\
&= \frac{\log R_0}{\log R} + \frac{\lfloor 2\gamma(1+\varepsilon/2)\log R \rfloor}{\log R} \frac{1}{n_0} \sum_{i=1}^{n_0} \log U_i^* \\
&\longrightarrow -\gamma\varepsilon/2 \quad \text{a.s.}
\end{aligned}$$

so that $\rho_{n_0} \to 0$ a.s., where we have used that $\mathbb{E}[\log U_1^*] = -1/2$. Thus

$$\mathbb{P}\Big(\frac{N_t}{\log R} \geq 2\gamma(1+\varepsilon) \Big) \longrightarrow 0.$$

Similarly, by setting $n_2 = \lfloor 2\gamma(1-\varepsilon)\log R \rfloor$,

$$\mathbb{P}\Big(\frac{N_t}{\log R} \leq 2\gamma(1-\varepsilon) \Big) = \mathbb{P}\Big(\sum_{j=0}^{n_2} \frac{T_i}{\rho_i} \geq t \Big) \leq \mathbb{P}\Big(\frac{1}{\rho_{n_2}} \sum_{j=0}^{n_2} T_j \geq t \Big).$$

By the same computation as above and by the law of large numbers

$$\frac{\log \rho_{n_2}}{\log R} \longrightarrow \varepsilon, \quad \frac{1}{n_2}\sum_{j=0}^{n_2} T_j \longrightarrow 1, \quad \text{a.s.}$$

so that

$$\frac{1}{\rho_{n_2}}\sum_{j=0}^{n_2} T_j \longrightarrow 0 \quad \text{a.s.}$$

yielding the result.

Let us now turn to the proof of the convergence of ρ_t. Using the construction of $(\rho_t;\, t \geq 0)$ with a Poisson point process and a uniform variable V from Section 5.3.1, we see that ρ_t is the sum of the smallest atom on $[0, VR_0]$ of a Poisson point process with intensity $t\,\mathrm{Leb}$, and of the smallest atom of this Poisson point process on $[0, (1-V)R_0]$. It is straightforward that, as $R_0 \to \infty$, the location of these atoms converge to independent Exponential(t) variables, yielding the result.

Let us finally prove the last convergence. We have that

$$\log R_0 + \sum_{i=1}^{M_{\lfloor u \log R \rfloor}} \log U_i^* \leq \log S_{u \log R} \leq \log R_0 + \sum_{i=1}^{M_{\lfloor u \log R \rfloor}-1} \log U_i^*$$

Moreover, as $(S_t;\, t \geq 0)$ jumps at rate 1, and M_t is the number of jumps of $(S_t;\, t \geq 0)$ before time t, we have that

$$\frac{M_{\lfloor u \log R \rfloor}}{\log R} \longrightarrow u \qquad \text{a.s.}$$

This, with the previous inequality and the law of large numbers proves the finite-dimensional convergence of the process. That the finite-dimensional convergence can be reinforced to a convergence in the Skorohod space follows from the fact that the process is non-increasing, see for instance Theorem 3.37 of Chapter IV of [111]. □

In the following, we will apply Lemma 5.17 to deduce convergence results on the k-spine from the previous proposition. The uniform integrability condition required in Lemma 5.17 is provided by the next results.

Lemma 5.20. *For any $p \geq 1$, $\sigma > 0$,*

$$\sup_{R>0} \frac{1}{(\log R)^p} \mathbb{E}\Big[\Big(\int_0^\sigma \rho_t \,\mathrm{d}t\Big)^p \mid \rho_0 = R\Big] < \infty.$$

Proof. In the following computation let us assume that $R_0 = R$ without mentioning it to ease the notation. We have that

$$\begin{aligned}\frac{1}{(\log R)^p}\mathbb{E}\Big[\Big(\int_0^\sigma \rho_t \,\mathrm{d}t\Big)^p\Big] &= \frac{1}{(\log R)^p}\int_{[0,\sigma]^p} \mathbb{E}\big[\rho_{t_1}\dots\rho_{t_p}\big]\,\mathrm{d}t_1\dots\mathrm{d}t_p \\ &= \frac{p}{(\log R)^p}\int_0^\sigma \mathbb{E}\Big[\rho_{t_1}\int_{[t_1,\sigma]^{p-1}} \rho_{t_2}\dots\rho_{t_p}\,\mathrm{d}t_2\dots\mathrm{d}t_p\Big]\,\mathrm{d}t_1\end{aligned}$$

$$= \frac{p}{(\log R)^p} \int_0^\sigma \mathbb{E}\Big[\rho_{t_1} \mathbb{E}\Big[\int_{[t_1,\sigma]^{p-1}} \rho_{t_2} \dots \rho_{t_p} \, \mathrm{d}t_2 \dots \mathrm{d}t_p \ \mid \ \rho_0 = \rho_{t_1}\Big]\Big] \, \mathrm{d}t_1$$
$$< p \sup_{R>0} \Big\{ \frac{1}{(\log R)^{p-1}} \mathbb{E}\Big[\Big(\int_0^\sigma \rho_t \, \mathrm{d}t \Big)^{p-1} \Big] \Big\} \frac{1}{\log R} \int_0^\sigma \mathbb{E}\big[\rho_t\big] \, \mathrm{d}t$$

so that, by induction, we are left to showing that

$$\sup_{R>0} \frac{1}{\log R} \int_0^\sigma \mathbb{E}\big[\rho_t\big] \, \mathrm{d}t < \infty.$$

If $\rho_0 = R$, according to the Poisson construction of $(\rho_t;\, t \geq 0)$ described in Section 5.3.1, ρ_t is obtained by throwing a Poisson(Rt) distributed number of uniform variables on $[0, R]$, breaking $[0, R]$ is subintervals, and picking one of these subintervals in a size-biased way. Therefore, ρ_t/R has a density w.r.t. the Lebesgue measure on $[0, 1]$ given by

$$\sum_{k\geq 1} e^{-Rt} \frac{(Rt)^k}{k!} k(k+1)x(1-x)^{k-1} = xRte^{-Rt} \sum_{k\geq 0} \frac{(Rt)^k}{k!} (k+2)(1-x)^k$$
$$= xRte^{-xRt}(Rt(1-x)+2).$$

Moreover, ρ_t has an atom at R of mass e^{-Rt}. Therefore

$$\mathbb{E}\big[\rho_t\big] = Re^{-Rt} + \int_0^1 (xR)^2 te^{-xRt}(Rt(1-x)+2) \, \mathrm{d}x$$
$$\leq Re^{-Rt} + \Big(\frac{1}{t} + \frac{1}{Rt^2} \Big) \int_0^{Rt} u^2 e^{-u} \, \mathrm{d}u$$
$$= Re^{-Rt} + 2\Big(\frac{1}{t} + \frac{1}{Rt^2} \Big) \Big(1 - e^{-Rt}\Big(1 + Rt + \frac{(Rt)^2}{2}\Big)\Big).$$

Thus

$$\int_0^\sigma \mathbb{E}\big[\rho_t\big] \, \mathrm{d}t \leq 1 + 2 \int_0^{R\sigma} \Big(\frac{1}{v} + \frac{1}{v^2} \Big) \Big(1 - e^{-v}\Big(1 + v + \frac{v^2}{2}\Big)\Big) \, \mathrm{d}v,$$

from which is directly follows that

$$\limsup_{R\to\infty} \frac{1}{\log R} \int_0^\sigma \mathbb{E}\big[\rho_t\big] \, \mathrm{d}t < \infty$$

yielding the result. □

Corollary 5.21. *Let* $\big((\rho_t^i), K^i, L^i;\, i \leq k\big)$ *be the* k*-spine. We have*

$$\sup_{R>0} \frac{1}{(\log R)^p} \mathbb{E}\Big[\Big(\sum_{i=1}^k \int_0^{L^i} \rho_t^i \, \mathrm{d}t \Big)^p \ \mid \ \rho_0^1 = R \Big] < \infty.$$

Proof. Recall that $(\rho_t^k;\, t \leq L^k)$ is a 1-spine started from ρ_0^k, and that, conditional on ρ_0^k it is independent of $((\rho_t^i), K^i, L^i;\, i < k)$. By self-similarity of the 1-spine, we can assume that

$$\rho_t^k = \frac{\rho_0^k}{R} \rho'_{(\rho_0^k t)/R}$$

where $(\rho'_t;\, t \geq 0)$ is a 1-spine, started from R and is independent of the variables $((\rho^i_t), K^i, L^i;\, i < k)$. Thus we have that

$$\int_0^{L^k} \rho^k_t \,\mathrm{d}t = \frac{R}{\rho^k_0} \int_0^{(L^k \rho^k_0)/R} \rho^k_{(Rt)/\rho^k_0} \,\mathrm{d}t \leq \int_0^{\sigma} \rho'_t \,\mathrm{d}t,$$

and

$$\sum_{i=1}^{k} \int_0^{L^i} \rho^i_t \,\mathrm{d}t \leq \sum_{i=1}^{k-1} \int_0^{L^i} \rho^i_t \,\mathrm{d}t + \int_0^{\sigma} \rho'_t \,\mathrm{d}t.$$

The proof is ended by a straightforward induction using Lemma 5.20. □

5.4.2 Limit of the k-spine

We now study the large R limit of the k-spine. Before stating our result, let us show a lemma that we will need. In the large R limit, under the appropriate time-change, the branching times of the k-spine will be uniformly distributed on $[0, 2]$. The following simple lemma shows that this property is preserved under the inductive step that constructs the $k+1$-spine from the k-spine.

Lemma 5.22. *Let $(U_2, \dots, U_k)$ be i.i.d. uniform variables on $[0, 2]$, and let $U_1 = 2$. Let $(\widetilde{U}_1, \dots, \widetilde{U}_k)$ have the distribution of $(U_1, \dots, U_k)$, biased by $U_1 + \dots + U_k$, let K be such that*

$$\mathbb{P}(K = i \mid \widetilde{U}_1, \dots \widetilde{U}_k) \propto \widetilde{U}_i,$$

and $\widetilde{U}_{k+1} = V \widetilde{U}_K$, where V is an independent uniform variable on $[0, 1]$. Then the order statistics of $(\widetilde{U}_2, \dots, \widetilde{U}_{k+1})$ is that of k i.i.d. uniform variables on $[0, 2]$.

Proof. Let $i \geq 1$, and φ_j be bounded continuous functions. We have that

$$\begin{aligned} \mathbb{E}\Big[\prod_{j=1}^{k+1} \varphi_j(\widetilde{U}_j) \mathbf{1}_{K=i}\Big] &= \mathbb{E}\Big[\prod_{j=1}^{k} \varphi_j(\widetilde{U}_j) \mathbf{1}_{K=i} \frac{1}{\widetilde{U}_i} \int_0^{\widetilde{U}_i} \varphi_{k+1}(x) \,\mathrm{d}x\Big] \\ &= \mathbb{E}\Big[\prod_{j=1}^{k} \varphi_j(\widetilde{U}_j) \frac{1}{\sum_{j=1}^k \widetilde{U}_j} \int_0^{\widetilde{U}_i} \varphi_{k+1}(x) \,\mathrm{d}x\Big] \\ &\propto \mathbb{E}\Big[\varphi_i(U_i) \int_0^{U_i} \varphi_{k+1}(x) \,\mathrm{d}x\Big] \prod_{j \neq i} \mathbb{E}[\varphi_j(U_j)]. \end{aligned}$$

This shows that, for $i \geq 2$, conditional on $K = i$, $(\widetilde{U}_i, \widetilde{U}_{k+1})$ are distributed as the order statistics of two uniform variables on $[0, 2]$, independent of $(\widetilde{U}_j;\, j \neq i)$ that are i.i.d. uniform variables on $[0, 2]$. For $i = 1$, the previous calculation shows that, conditional on $K = 1$, $(\widetilde{U}_2, \dots, \widetilde{U}_{k+1})$ are i.i.d. uniform random variables on $[0, 2]$. Therefore, the result holds conditional on $K = i$ for any i proving the lemma. □

Let us now introduce the time-changes and notation for the convergence of the k-spine. Let $((\rho^i_t), K^i, L^i;\, i \leq k)$ be the k-spine, started from a block of length R, and stopped at time σ. For any $i \leq k$, we consider the random time-change

$$\forall t \leq L^i, \quad \theta^i_t = \int_0^t \rho^i_u \,\mathrm{d}u, \quad \bar{L}^i = \theta^i_{L^i} = \int_0^{L^i} \rho^i_u \,\mathrm{d}u.$$

We denote by τ^i its inverse, and let

$$\forall t \leq \bar{L}^i, \quad S^i_t = \rho^i \circ \tau^i_t$$

be the corresponding time-changed process. The following result is the k-spine analogous of Proposition 5.19.

Proposition 5.23. *Fix $k \geq 1$, and let $((S^i_t), K^i, \bar{L}^i; i \leq k)$ be the time-changed k-spine, started at R until time σ. Then, the following convergences hold jointly in distribution as $R \to \infty$.*

(i) *We have*

$$\Big(\frac{\bar{L}^1}{\log R}, \dots, \frac{\bar{L}^k}{\log R}\Big) \longrightarrow \big(U^1, \dots, U^k\big),$$

where $U^1 = 2$, and the order statistics of $(U^2, \dots, U^k)$ are that of i.i.d. uniform variables on $[0, 2]$.

(ii) *For any fixed $t > 0$, we have*

$$(\rho^1_t, \dots, \rho^k_t) \longrightarrow (Y^1_t, \dots, Y^k_t)$$

where $(Y^1_t, \dots, Y^k_t)$ are i.i.d., distributed as Gamma$(2, t)$ *variables, and independent of $(U^1, \dots, U^k)$.*

(iii) *Moreover, for each $i \leq k$,*

$$\Big(\frac{\log S^i_{t \log R}}{\log R}; t \leq \frac{\bar{L}^i}{\log R}\Big) \longrightarrow \Big(\frac{U^i - t}{2}; t \leq U^i\Big)$$

in distribution for the uniform topology.

(iv) *Finally,*

$$(L^1, \dots, L^k) \longrightarrow (\sigma, \dots, \sigma).$$

Proof. We proceed by induction. For $k = 1$, we have

$$\bar{L}^1 = \int_0^\sigma \rho^1_t \,\mathrm{d}t,$$

so that (i), (ii) and (iii) follow from the corresponding points in Proposition 5.19 with $\gamma = 1$. By definition we have $L^1 = \sigma$.

Suppose that all convergences occur jointly for some k. Then

$$\frac{\bar{L}^1 + \dots + \bar{L}^k}{\log R} \longrightarrow U^1 + \dots + U^k$$

in distribution. Thus, provided that the latter variables are uniformly integrable, by Lemma 5.17, all convergences in the statement of the result also hold for the versions of the variables biased by

$$\frac{\bar{L}^1 + \dots + \bar{L}^k}{\log R}$$

to the limiting variables, biased by

$$U^1 + \cdots + U^k.$$

The required uniform integrability follows from Corollary 5.21.

With an abuse of notation, we still denote by $((S^i_t), \bar{L}^i, K^i;\ i \leq k)$ the biased variables, and by $(U^1, \dots, U^k)$ the limiting variables biased by $(U^1, \dots, U^k)$. It follows from

$$\mathbb{P}\Big(K^{k+1} = i \mid (S^i_t), \bar{L}^i, K^i;\ i \leq k\Big) = \frac{\bar{L}^i}{\bar{L}^1 + \cdots + \bar{L}^k} \longrightarrow \frac{U^i}{U^1 + \cdots + U^k}.$$

that K^{k+1} converges in distribution to some random index K^{k+1}_∞ such that

$$\mathbb{P}\Big(K^{k+1}_\infty = i \mid U^1, \dots, U^k\Big) = \frac{U^i}{U^1 + \cdots + U^k}.$$

Moreover recall that, conditional on $K^{k+1} = i$, L^{k+1} is chosen so that

$$\mathbb{E}\Big[\varphi(L^{k+1}) \mid K^{k+1} = i\Big] \propto \int_0^{L^i} \rho^i_{L^i - t} \varphi(t) \, \mathrm{d}t$$

With the random time-change, we have that

$$\mathbb{E}\Big[\varphi(\bar{L}^{k+1}) \mid K^{k+1} = i\Big] \propto \int_0^{L^i} \varphi(\theta_t) \rho^i_{L^i - t} \, \mathrm{d}t = \int_0^{\bar{L}^i} \varphi(t) \, \mathrm{d}t,$$

so that, conditional on $K^{k+1} = i$, $\bar{L}^{k+1}$ is uniformly distributed on $[0, \bar{L}^i]$. Therefore, $\bar{L}^{k+1} / \log R$ converges to some limit U^{k+1}, and according to Lemma 5.22, the order statistics of $(U^2, \dots, U^{k+1})$ are that of i.i.d. uniform variables on $[0, 2]$. This proves (i).

Conditional on $K^{k+1} = i$, $(S^{k+1}_t;\ t \leq \bar{L}^{k+1})$ is obtained out of an independent 1-spine started at $S^i_{\bar{L}^i - \bar{L}^{k+1}}$. The joint convergence of

$$\frac{\bar{L}^i - \bar{L}^{k+1}}{\log R} \longrightarrow \widetilde{U}^i - \widetilde{U}^{k+1}$$

and of

$$\Big(\frac{\log S^i_{t \log R}}{\log R};\ t \leq \frac{\bar{L}^i}{\log R}\Big) \longrightarrow \Big(\frac{\widetilde{U}^i - t}{2};\ t \leq \widetilde{U}^i\Big)$$

proves that

$$\frac{\log S^{k+1}_0}{\log R} = \frac{\log S^i_{\bar{L}^i - \bar{L}^{k+1}}}{\log R} \longrightarrow \frac{\widetilde{U}^{k+1}}{2},$$

so that, by Proposition 5.19,

$$\Big(\frac{\log S^{k+1}_{t \log R}}{\log R};\ t \leq \frac{\bar{L}^{k+1}}{\log R}\Big) \longrightarrow \Big(\frac{\widetilde{U}^{k+1} - t}{2};\ t \leq \widetilde{U}^{k+1}\Big).$$

This proves (iii).

Let us prove the last two points. For any i we have

$$\frac{1}{\log R}\int_0^{\bar{L}^i} \rho_t^i \,\mathrm{d}t \longrightarrow U^i$$

whereas

$$\int_\varepsilon^{\bar{L}^i} \rho_t^i \,\mathrm{d}t \leq \sigma \rho_\varepsilon^i \longrightarrow \sigma Y_\varepsilon$$

with $Y_\varepsilon \sim \mathrm{Gamma}(2, \varepsilon)$. Thus, conditional on $K^{k+1} = i$, as $L^i - L^{k+1}$ is sampled in $[0, L^i]$ with a density proportional to ρ_t^i, we have that

$$\mathbb{P}(L^i - L^{k+1} \geq \varepsilon \mid K^{k+1} = i) = \frac{\displaystyle\int_\varepsilon^{L^i} \rho_t^i \,\mathrm{d}t}{\displaystyle\int_0^{L^i} \rho_t^i \,\mathrm{d}t} \leq \frac{\sigma Y_\varepsilon}{\displaystyle\int_0^{L^i} \rho_t^i \,\mathrm{d}t} \longrightarrow 0.$$

Thus, by induction,

$$L^{k+1} \longrightarrow \sigma,$$

and by Proposition 5.19, we directly obtain that for all $t > 0$,

$$\rho_t^{k+1} \longrightarrow Y_t$$

in distribution. Our proof of is now complete. □

5.4.3 Proof of Theorem 5.2

We are now ready to prove the convergence of the empirical measure of the block lengths.

Proof of Theorem 5.2. Let us denote by ν_t the empirical distribution of block length at time t, with population size parameter N and recombination rate R. For simplicity, the dependence in N and R is not taken into account in the notation. Then, if $M_t = \langle 1, \nu_t \rangle$ is the total population size at time t,

$$\begin{aligned}
\frac{1}{N^k}\mathbb{E}\Big[\langle f, \nu_t \rangle^k \mid M_t > 0\Big] &= \frac{1}{Np^{N,R}(t)}\frac{1}{N^{k-1}}\mathbb{E}\Big[\langle f, \nu_t \rangle^k\Big] \\
&= \frac{1}{Np^{N,R}(t)}\frac{1}{N^{k-1}}\mathbb{E}\Big[\sum_{|v^i|=t} f(\rho_{v^1}) \dots f(\rho_{v^k})\Big],
\end{aligned}$$

where $p^{N,R}(t)$ denotes the extinction probability at generation t. As, on the set of non-extinction at time $N\sigma$, the number of individuals is of order N, we have that

$$\mathbb{E}\Big[\sum_{|v^i|=\lfloor N\sigma \rfloor} f(\rho_{v^1}) \dots f(\rho_{v^k})\Big] \sim \mathbb{E}\Big[\sum_{\substack{|v^i|=\lfloor N\sigma \rfloor \\ v^1 \neq \dots \neq v^k}} f(\rho_{v^1}) \dots f(\rho_{v^k})\Big].$$

Moreover, by the many-to-few formula of Corollary 5.13 and by point (i) of Theorem 5.12, we have that

$$\mathbb{E}\Big[\sum_{\substack{|v^i|=\lfloor N\sigma\rfloor \\ v^1\neq\ldots\neq v^k}} f(\rho_{v^1})\ldots f(\rho_{v^k})\Big] = R\mathbb{E}\Big[\frac{f(\rho_{V^1})}{\rho_{V^1}}\ldots\frac{f(\rho_{V^k})}{\rho_{V^k}}\Big]\prod_{i=1}^{k-1}\mathbb{E}\Big[\sum_{j=1}^{i}\sum_{n=1}^{\lfloor N\sigma\rfloor}\rho_{V_n^j}\Big]$$

so that

$$\lim_{N\to\infty}\frac{1}{N^{k-1}}\mathbb{E}\Big[\sum_{\substack{|v^i|=\lfloor N\sigma\rfloor \\ v^1\neq\ldots\neq v^k}} f(\rho_{v^1})\ldots f(\rho_{v^k})\Big] = R\mathbb{E}\Big[\frac{f(\rho^1_{L^1})}{\rho^1_{L^1}}\ldots\frac{f(\rho^k_{L^k})}{\rho^k_{L^k}}\Big]\prod_{i=1}^{k-1}\mathbb{E}\Big[\sum_{j=1}^{i}\int_0^{L^j}\rho^j_t\,\mathrm{d}t\Big]$$

where $(\rho^1_t,\ldots,\rho^k_t)$ are the block length of the continuous-time k-spine. The last convergence requires some uniform integrability on the various variables. We claim that the required property follow from $\rho^i_t\leq R$ and from the Poissonian construction of the 1-spine. The behavior near 0 of ρ^i_t is that of the sum of two independent exponential variables, so that the integrability of $1/\rho^i_t$ is not problematic.

Let us now take the limit $R\to\infty$. By Proposition 5.23,

$$\lim_{R\to\infty}\frac{1}{\log R}\mathbb{E}\Big[\sum_{j=1}^{i}\int_0^{L^i}\rho^j_t\,\mathrm{d}t\Big] = \mathbb{E}\Big[U^1+\cdots+U^i\Big] = i+1,$$

and

$$\lim_{R\to\infty}\mathbb{E}\Big[\frac{f(\rho^1_{L^1})}{\rho^1_{L^1}}\ldots\frac{f(\rho^k_{L^k})}{\rho^k_{L^k}}\Big] = \mathbb{E}\Big[\frac{f(Y^1)}{Y^1}\ldots\frac{f(Y^k)}{Y^k}\Big]$$

where $(Y^1,\ldots,Y^k)$ are independent Gamma$(2,\sigma)$ variables. Thus

$$\mathbb{E}\Big[\frac{f(Y^1)}{Y^1}\ldots\frac{f(Y^k)}{Y^k}\Big] = \sigma^k\mathbb{E}\Big[f(Z)\Big]^k$$

where $Z\sim$ Exponential(σ). (Again, the uniform integrability follows either from Corollary 5.21 or from the Poissonian construction.)

Putting pieces together, and using Proposition 5.7, for $k\geq 1$, we have that

$$\begin{aligned}
&\lim_{R\to\infty}\lim_{N\to\infty}\frac{1}{(N\log R)^k}\mathbb{E}\Big[\langle f,\nu_t\rangle^k \mid M_t>0\Big]\\
&\quad=\lim_{R\to\infty}\Big(\lim_{N\to\infty}\frac{R}{p^R(\sigma)N\log R}\Big)\mathbb{E}\Big[\frac{f(\rho^1_\sigma)}{\rho^1_\sigma}\ldots\frac{f(\rho^k_\sigma)}{\rho^1_\sigma}\Big]\prod_{i=1}^{k-1}\mathbb{E}\Big[\frac{1}{\log R}\sum_{j=1}^{i}\int_0^{\sigma}\rho^j_t\,\mathrm{d}t\Big]\\
&\quad=\sigma^k\mathbb{E}\Big[f(Z)\Big]^k k!\\
&\quad=\mathbb{E}\Big[f(Z)\Big]^k\mathbb{E}\Big[Y^k\Big]
\end{aligned}$$

where $Y\sim$ Exponential$(1/\sigma)$.

Thus, by the method of moments, this proves that

$$\lim_{R\to\infty}\lim_{N\to\infty}\frac{1}{N\log R}\langle f,\nu_{\lfloor N\sigma\rfloor}\rangle = Y\langle f,Z\rangle$$

in distribution, conditional on $M_{\lfloor N\sigma \rfloor} > 0$, where Y is an Exponential$(1/\sigma)$ variable and Z an Exponential(σ) variable. This entails the convergence of the conditioned renormalized empirical measure in the weak topology to the random measure $Y\mathscr{L}(Z)$ and proves the result, see for instance [121], Theorem 4.11. □

5.5 Geometry of the blocks on the chromosome

In this section, we prove Theorem 5.4 which provides the limit of the locations on the chromosome of the blocks in the population, in the Gromov-weak sense. We start by recalling some convergence facts about this topology, in order to identify what needs to be proved, and to motivate the remainder of this section.

Let (X, d, μ) be a random metric space. Recall the definition of the (random) measure ι_n as the push-forward measure of $\mu^{\otimes n}$ by the map D_n. According to Lemma 2.7 in [43], in our case, proving the convergence in distribution in the Gromov-weak topology amounts to proving the convergence of

$$\mathbb{E}\Big[\langle F, \iota_n \rangle\Big] = \mathbb{E}\Big[\mu(X)^n F\Big(\big(d(Y_i, Y_j);\, i,j \leq n\big)\Big)\Big]$$

for any bounded continuous function $F\colon \mathbb{R}_+^{\binom{n}{2}} \to \mathbb{R}$, where $(Y_1, \dots, Y_n)$ are i.i.d. variables in X sampled according to $\mu/\mu(X)$. (Note that this is a consequence of the fact that the total mass of a Brownian CPP is exponentially distributed, and that this distribution is characterized by its moments.)

Let $(\mathcal{V}_t, d_{\mathcal{V}}, \mu_{\mathcal{V}})$ be the population at generation t, viewed as a random metric space. Then, for a continuous bounded function F, the previous functionals of the random metric space can be written as

$$\begin{aligned}\mathbb{E}\Big[\langle F, \iota_n \rangle\Big] &= \mathbb{E}\Big[\sum_{\substack{v^1,\dots,v^n\\ |v^i|=t}} F\Big(\big(d(v^i, v^j);\, i,j \leq n\big)\Big)\Big]\\ &= \mathbb{E}\Big[\sum_{\substack{v^1,\dots,v^n\\ |v^i|=t}} F\Big(\big(|H_{v^i} - H_{v^j}|;\, i,j \leq n\big)\Big)\Big].\end{aligned}$$

Moreover, by our many-to-few formula, the previous quantity can be directly expressed in terms of the locations of the blocks of the distinguished vertices of the k-spine. In Section 5.5.1 we start by showing that, in the limit, the distance on the chromosomes of the blocks of the k-spine are that of their time to the MRCA. Then, in Section 5.5.2, we express these times to the MRCA in terms of the Brownian CPP, and complete the proof of Theorem 5.4.

5.5.1 Convergence of the block distance

Recall the notation $(I^i_t;\, i \leq k,\, t \leq L^i)$ for the blocks carried by the k-spine. For each t let H^i_t be the left endpoint of I^i_t. We now consider the large recombination

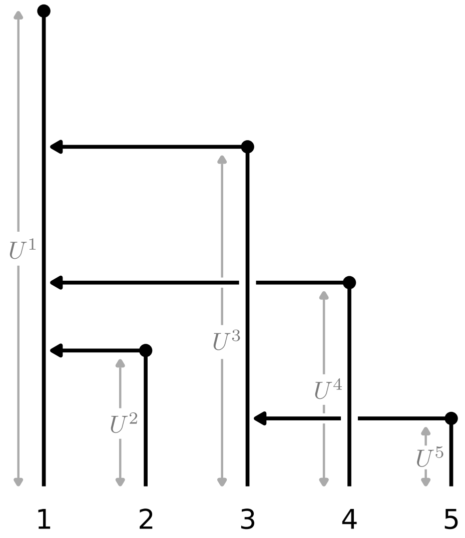


Figure 5.3: Construction of the tree encoded by $(U^1, \dots, U^5)$ and $(K^1_\infty, \dots, K^5_\infty)$. In this tree, $(K^1_\infty, \dots, K^5_\infty) = (0, 1, 1, 1, 3)$.

rate limit of the distance matrix of the tips of the spine, defined as

$$\Big(|H^i_{L^i} - H^j_{L^j}|;\, i, j \leq k \Big).$$

We will show that, as $R \to \infty$, the distance between the locations of the blocks of any two branches of the spine is of the same order as the block length of their most-recent common ancestor (MRCA). Let us start by giving a definition of that MRCA.

We can build a tree out of the vectors $(U^1, \dots, U^k)$ and $(K^1_\infty, \dots, K^k_\infty)$ as follows. Start from a branch of length $U^1 = 2$, with the root of the tree at one endpoint, and the other endpoint labeled 1. At step i, add an external branch of length U^i with label i to the tree. Graft one of its endpoint on the branch with label K^i_∞, in such a way that the other endpoint lies at distance 2 from the root. See Figure 5.3 for a graphical illustration. The time to the MRCA between i and j, $T^{i,j}$, is then defined as half the distance between the tips of the branches with label i and j.

Proposition 5.24. *As $R \to \infty$, the following convergence holds in distribution*

$$\Big(\frac{\log |H^i_{L^i} - H^j_{L^j}|}{\log R};\, i, j \leq k \Big) \longrightarrow \Big(\frac{T^{i,j}}{2};\, i, j \leq k \Big),$$

where $T^{i,j}$ is the time to the MRCA of i and j in the limiting tree encoded by the variables $(U^1, \dots, U^k)$ and $(K^1_\infty, \dots, K^k_\infty)$.

Proof. Again, we work by induction. For $k = 1$ there is nothing to prove. Suppose that the convergence holds for some $k \geq 1$. Again, by Lemma 5.17, we have convergence of the distribution of

$$\Big(\frac{\log |H^i_{L^i} - H^j_{L^j}|}{\log R};\, i, j \leq k \Big),$$

biased by $L^1 + \cdots + L^k$ to the distribution of

$$\Big(\frac{T^{i,j}}{2};\, i,j \leq k\Big),$$

biased by $U^1 + \cdots + U^k$. Therefore, if, with a slight abuse of notation, we still denote by $\big((H^i_t);\, i \leq k,\, t \leq L^i\big)$ and $(T^{i,j};\, i,j \leq k)$ the block locations and times to the MRCA of the first k branches of the $k+1$-spine, we might still assume that the convergence holds.

Let K^{k+1}_∞ be the parent of the $k+1$-branch. Let us first work conditional on $K^{k+1}_\infty = i$, and show that

$$\frac{\log|H^{k+1}_{L^{k+1}} - H^i_{L^i}|}{\log R} \longrightarrow \frac{T^{k+1,i}}{2}.$$

Recall that, for a fixed R, $(I^{k+1}_t;\, t \leq L^{k+1})$ is distributed as an independent 1-spine, started from $(I^i_{L^i - L^{k+1}})$. Recall that this 1-spine can be constructed from an independent homogeneous Poisson point process $\mathscr{Q}$ on $\mathbb{R}_+ \times \mathbb{R}_+$ and a uniform variable V on $[0,1]$ as follows. Let I_t be the subinterval of

$$I^i_{L^i - L^{k+1}} \setminus \{x : (x,s) \in \mathscr{Q}, s \leq t\}$$

to which $H^i_{L^i - L^{k+1}} + \rho^i_{L^i - L^{k+1}} V$ belongs. Then H^{k+1}_t can be defined as the left endpoint of I_t, and ρ^{k+1}_t as its length.

Moreover, we know from Proposition 5.23 that $\rho^{k+1}_{L^{k+1}}$ converges to a $\mathrm{Gamma}(2,\sigma)$ variable, while the variable $\rho^i_{L^i - L^{k+1}}$ converges to $+\infty$. Thus

$$\frac{H^{k+1}_{L^{k+1}} - H^i_{L^i - L^{k+1}}}{\rho^i_{L^i - L^{k+1}}} \longrightarrow V$$

in probability, that is, in the limit the size of the interval $I^{k+1}_{L^{k+1}}$ is negligible compared to that of its ancestor. As V is independent of the k-spine, we have that

$$\frac{|H^{k+1}_{L^{k+1}} - H^i_{L^i}|}{\rho^i_{L^i - L^{k+1}}} \longrightarrow \mathrm{Uniform}([0,1]),$$

in distribution. Thus, it follows from

$$\frac{\log \rho^i_{L^i - L^{k+1}}}{\log R} \longrightarrow \frac{U^{k+1}}{2},$$

that

$$\frac{\log|H^{k+1}_{L^{k+1}} - H^i_{L^i}|}{\log R} \longrightarrow \frac{U^{k+1}}{2} = \frac{T^{k+1,i}}{2}$$

in distribution.

Up to using Skorohod representation theorem, see for instance Theorem 6.7 in [25], we now assume that all the previous converges hold almost surely. By the construction of the time to the MRCA, as $K_\infty^{k+1} = i$, for $j \neq i$ we have

$$T^{k+1,j} = \begin{cases} T^{i,j} & \text{if } T^{k+1,i} < T^{i,j} \\ T^{i,k+1} & \text{if } T^{k+1,i} > T^{i,j}. \end{cases}$$

(This follows from the fact that the tree encoded by $(T^{i,j})$ is ultrametric.) Now, using that

$$\Big| |H^{k+1}_{L^{k+1}} - H^i_{L^i}| - |H^i_{L^i} - H^j_{L^j}| \Big| \le |H^{k+1}_{L^{k+1}} - H^j_{L^j}| \le |H^{k+1}_{L^{k+1}} - H^i_{L^i}| + |H^i_{L^i} - H^j_{L^j}|,$$

and that

$$\frac{\log|H^{k+1}_{L^{k+1}} - H^i_{L^i}|}{\log|H^j_{L^j} - H^i_{L^i}|} \longrightarrow \frac{T^{k+1,i}}{T^{i,j}}$$

we see that:

- if $T^{k+1,i} < T^{i,j}$, then
$$\frac{\log|H^{k+1}_{L^{k+1}} - H^j_{L^j}|}{\log R} \longrightarrow \frac{T^{i,j}}{2} = \frac{T^{k+1,j}}{2};$$

- if $T^{i,j} < T^{k+1,i}$,
$$\frac{\log|H^{k+1}_{L^{k+1}} - H^j_{L^j}|}{\log R} \longrightarrow \frac{T^{k+1,i}}{2} = \frac{T^{k+1,j}}{2}.$$

This proves the result conditional on $K^{k+1} = i$, for all $i \le k$, and thus ends the proof. □

5.5.2 Proof of Theorem 5.4

Now that we have connected the locations of the blocks of the k-spine to the times to the MRCA, it remains to be shown that the tree structure of the k-spine is that of a Brownian CPP.

Note that we have two encodings of the tree structure of the k-spine. First, the vector $(T^{i,j}; i,j \le k)$ encodes the ultrametric structure of the tree, that is, the tree distance between the leaves. Second, the vectors $(U^1, \dots, U^k)$ and $(K^1_\infty, \dots, K^k_\infty)$ represent respectively the length of the branch leading to leaf i, and the label of the branch to which it is grafted. The correspondence between the two representations is not one-to-one: there can be distinct vectors $(U^1, \dots, U^k)$ and $(K^1_\infty, \dots, K^k_\infty)$ with the same underlying ultrametric space. As we are only interested into showing that the ultrametric structure of the k-spine is that of a biased Brownian CPP, we can (and will) consider different ways of recovering branch lengths out of the Brownian CPP.

Recall the definition of the Brownian CPP $([0, Y_1], d, \mathrm{Leb})$ out of a Poisson point process $\mathscr{Q}$ with intensity

$$\mathrm{d}t \otimes \frac{1}{x^2}\,\mathrm{d}x.$$

Conditional on Y_1, let $(V^i)_{i \geq 1}$ be i.i.d. uniform variables on $[0, Y_1]$. We define a random ultrametric on $\mathbb{N}$ as follows

$$\forall i, j, \quad d_{\mathbb{N}}(i, j) = d(V^i, V^j).$$

This random ultrametric is called the distance matrix of the Brownian CPP.

We now define two ways of recovering trees whose underlying ultrametric space is d out of the Brownian CPP. For $i \leq k$, let $\overleftarrow{K}^i$ be the label of the closest variable among $(V^1, \dots, V^k)$ to the left of V^i, and let $\overleftarrow{V}^i = V^{\overleftarrow{K}^i}$ be the location of this variable. If V^i is the left-most variable, set $V^i = 0$, and assign an arbitrary value to K^i. Then, if

$$\overleftarrow{U}^i = \sup\{x : (x, t) \in \mathscr{Q},\, t \in [\overleftarrow{V}^i, V^i]\},$$

we call the tree with branch lengths $(\overleftarrow{U}^1, \dots, \overleftarrow{U}^k)$ and ancestors $(\overleftarrow{K}^1, \dots, \overleftarrow{K}^k)$ the *left-oriented tree* associated to the Brownian CPP. Similarly, we define $\overrightarrow{K}^i$ as the label of the closest variable among $(V^1, \dots, V^k)$ to the right of V^i, and $\overrightarrow{V}^i = V^{\overrightarrow{K}^i}$, with the convention that $\overrightarrow{V}^i = Y_1$ if V^i is the right-most variable. We set

$$\overrightarrow{U}^i = \sup\{x : (x, t) \in \mathscr{Q},\, t \in [V^i, \overrightarrow{V}^i]\},$$

and call the tree with branch lenghts $(\overrightarrow{U}^1, \dots \overrightarrow{U}^k)$ and ancestors $(\overrightarrow{K}^1, \dots, \overrightarrow{K}^k)$ the *right-oriented tree* associated to the Brownian CPP.

The left-oriented (resp. right-oriented) tree can be obtained from $\mathscr{Q}$ and from $(V^1, \dots, V^k)$ in the following pictorial way. Recall that $\mathscr{Q}$ can be seen as a set of teeth on $[0, Y_1]$. The variables $(V^1, \dots, V^k)$ break $[0, Y_1]$ into $k+1$ subintervals. Remove all teeth from $\mathscr{Q}$ but the largest tooth in each of the $k-1$ inner subintervals, that is, those intervals with both endpoints in $(0, Y_1)$. Add a tooth of length 1 at 0 (resp. Y_1). Each remaining tooth represents a branch of the tree, and its parent is defined as the first tooth encountered by an arrow sent from the top of the tooth towards the left (resp. the right). The label of a branch is the label of the closest variable to the right (resp. to the left) of the stem of the tooth. See Figure 5.4 for an illustration of this procedure.

Finally, we say that $(\widehat{Y}_1, \widehat{d}_{\mathbb{N}})$ is a k-biased Brownian CPP if is has the distribution of $(Y_1, d_{\mathbb{N}})$ biased by Y_1^k. Then $\widehat{Y}_1$ has a Gamma$(k+1, 1)$ distribution, and conditional on $\widehat{Y}_1$, the ultrametric $\widehat{d}_{\mathbb{N}}$ is obtained out of an i.i.d. sequence $(V^i;\, i \geq 1)$ of uniform variables on $[0, \widehat{Y}_1]$ in the exact same way as in the unbiased Brownian CPP. Moreover note that, as $\widehat{Y}_1$ is Gamma$(k+1, 1)$ distributed, $(V^1, \dots, V^k)$ break $[0, \widehat{Y}_1]$ in $k+1$ subintervals, and the length of these subintervals are independent Exponential(1) variables. Moreover, a simple calculation shows that the size of the largest tooth in those subintervals are independent uniform variables on $[0, 1]$. This is the key property that we use to show the following result.

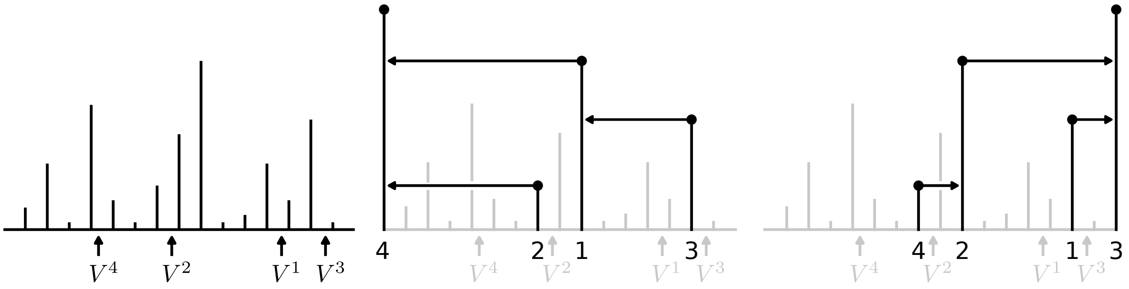


Figure 5.4: Orientation of a CPP tree. The CPP is represented on the right panel. The other two panels represent the left and right orientation of this CPP, respectively.

Proposition 5.25. *For $k \geq 1$, let $(T^{i,j};\, i,j \leq k)$ be the time to the MRCA in the tree obtained from $(U^1, \dots, U^k)$ and $(K^1_\infty, \dots, K^k_\infty)$. Then*

$$\Big(\widehat{d}_{\mathbb{N}}(i,j);\, i,j \leq k\Big) \overset{\text{(d)}}{=} \Big(\frac{T^{i,j}}{2};\, i,j \leq k\Big),$$

where $(\widehat{d}_{\mathbb{N}}(i,j);\, i,j \leq k)$ is the distance matrix of a k-biased Brownian CPP.

Proof. Let $\widehat{Y}_1$ be the total mass of a $k+1$-biased Brownian CPP, and $(V^1, \dots, V^{k+1})$ be $k+1$ uniform variables on $[0, \widehat{Y}_1]$. We build some branch lengths $(U^1_{\mathscr{Q}}, \dots, U^{k+1}_{\mathscr{Q}})$ and ancestors $(K^1_{\mathscr{Q}}, \dots, K^{k+1}_{\mathscr{Q}})$ by choosing the orientation of the $k+1$-biased CPP tree according to the location of V^{k+1}. More precisely:

- on the event $\{\forall i \leq k;\, V^{k+1} < V^i\}$, define
$$\begin{aligned} \forall i \leq k,\ (U^i_{\mathscr{Q}}, K^i_{\mathscr{Q}}) &= (\overleftarrow{U}^i, \overleftarrow{K}^i) \\ (U^{k+1}_{\mathscr{Q}}, K^{k+1}_{\mathscr{Q}}) &= (\overrightarrow{U}^{k+1}, \overrightarrow{K}^{k+1}); \end{aligned} \tag{5.5}$$

- on the event $\{\forall i \leq k;\, V^{k+1} > V^i\}$, define
$$\begin{aligned} \forall i \leq k,\ (U^i_{\mathscr{Q}}, K^i_{\mathscr{Q}}) &= (\overrightarrow{U}^i, \overrightarrow{K}^i) \\ (U^{k+1}_{\mathscr{Q}}, K^{k+1}_{\mathscr{Q}}) &= (\overleftarrow{U}^{k+1}, \overleftarrow{K}^{k+1}); \end{aligned} \tag{5.6}$$

- on the contrary event:
  - on $\{\overleftarrow{U}^{k+1} > \overrightarrow{U}^{k+1}\}$, define
$$(U^i_{\mathscr{Q}}, K^i_{\mathscr{Q}}) = \begin{cases} (\overrightarrow{U}^{k+1}, \overrightarrow{K}^{k+1}) & \text{if } i = k+1, \\ (\overleftarrow{U}^{k+1}, k+1) & \text{if } i = \overrightarrow{K}^{k+1}, \\ (\overleftarrow{U}^i, \overleftarrow{K}^i) & \text{else;} \end{cases} \tag{5.7}$$
  - on $\{\overleftarrow{U}^{k+1} < \overrightarrow{U}^{k+1}\}$, define
$$(U^i_{\mathscr{Q}}, K^i_{\mathscr{Q}}) = \begin{cases} (\overleftarrow{U}^{k+1}, \overleftarrow{K}^{k+1}) & \text{if } i = k+1, \\ (\overrightarrow{U}^{k+1}, k+1) & \text{if } i = \overleftarrow{K}^{k+1}, \\ (\overrightarrow{U}^i, \overrightarrow{K}^i) & \text{else.} \end{cases} \tag{5.8}$$

Let us go through all the cases separately. In equation (5.5), as conditional on the event $\{\forall i;\, V^{k+1} < V^i\}$, the variable $\widehat{Y}_1 - V^{k+1}$ has a Gamma$(k+1, 1)$ distribution, and $(V^1, \dots, V^k)$ are i.i.d. uniform on the interval $[V^{k+1}, \widehat{Y}_1]$, the subtree spanned by the first k leaves is distributed as the left-oriented tree of a k-biased Brownian CPP. Moreover, $U_{\mathscr{2}}^{k+1}$ is uniformly distributed on $[0, 1]$, and $K_{\mathscr{2}}^{k+1}$ is the label of the unique branch of length 1 in the subtree. A similar conclusion holds for equation (5.6), so that with probability $2/(k+1)$ the length of the $k+1$-th branch is uniformly distributed on $[0, 1]$, and it is grafted on the unique branch of length 2 of the tree with k leaves.

In equation (5.7), we consider the left-oriented tree associated to the Brownian CPP, except that we have swapped the branch lengths of the leaves $k+1$ and $\overrightarrow{K}^{k+1}$. (Note that this does not change the underlying ultrametric structure of the tree.) The distribution of the subtree spanned by the first k leaves is that of the left-oriented tree of the k-biased Brownian CPP, except that the branch length of $\overrightarrow{K}^{k+1}$ is the maximum of two independent uniform variables. As $\overrightarrow{K}^{k+1}$ is uniform in $\{1, \dots, k\}$, the tree on $k+1$ leaves is thus obtained by size-biasing the length of a uniformly chosen branch $\overrightarrow{K}^{k+1}$, and grafting the branch $k+1$ uniformly on it. A similar conclusion holds for equation (5.8).

Overall, we have just shown that the ultrametric structure of a sample of size $k+1$ in the $k+1$-biased Brownian CPP is obtain from that of the k sample by:

- with probability $2/(k+1)$, choosing the unique branch of size 1, and grafting branch $k+1$ uniformly on it;
- with probability $1/(k+1)$ for each other branch, size-biasing its length and grafting branch $k+1$ uniformly on it.

As this is the inductive step used to build the $k+1$-spine out of the k-spine, and as $U^1/2 = 1$, the proof is complete. □

The proof of Theorem 5.4 can now be completed.

Proof of Theorem 5.4. To ease the notation, let us define

$$D(v^1, \dots, v^k) = \Big(\frac{\log|H_{v^i} - H_{v^j}|}{\log R};\, i, j \le k\Big),$$

the distance matrix of $(v^1, \dots, v^k)$. Let $F\colon \mathbb{R}_+^{\binom{n}{2}} \to \mathbb{R}$ be a continuous bounded function and fix $\sigma > 0$. According to Lemma 2.7 in [43], the result is proved if we can show that

$$\lim_{R\to\infty} \lim_{N\to\infty} \frac{1}{N \log R} \mathbb{E}\Big[\sum_{\substack{v^1,\dots,v^k \\ |v^i| = \lfloor \sigma N \rfloor}} F\Big(D(v^1, \dots, v^k)\Big)\Big] = \mathbb{E}\Big[(\sigma Y_1)^k F\Big((d_{\mathbb{N}}(i,j);\, i, j \le k)\Big)\Big],$$

where $([0, Y_1], d, \mathrm{Leb})$ is a Brownian CPP, and $d_{\mathbb{N}}$ is its distance matrix. (The set of such functionals is convergence determining for the Gromov-weak topology, as the mass of the Brownian CPP is characterized by its moments.)

Reasoning along the same lines as in the proof of Theorem 5.2, we have that

$$\begin{aligned}\mathbb{E}\Big[\sum_{|v^i|=\lfloor\sigma N\rfloor} F\Big(D(v^1,\dots,v^k)\Big)\Big] &= \frac{1}{p^{N,R}(\lfloor\sigma N\rfloor)}\mathbb{E}\Big[\sum_{|v^i|=\lfloor\sigma N\rfloor} F\Big(D(v^1,\dots,v^k)\Big)\Big]\\ &\sim \frac{1}{p^{N,R}(\lfloor\sigma N\rfloor)}\mathbb{E}\Big[\sum_{\substack{|v^i|=\lfloor\sigma N\rfloor\\ v^1\neq\dots\neq v^k}} F\Big(D(v^1,\dots,v^k)\Big)\Big].\end{aligned}$$

By Corollary 5.13,

$$\begin{aligned}&\lim_{N\to\infty}\frac{1}{N^{k-1}}\mathbb{E}\Big[\sum_{\substack{|v^i|=\lfloor\sigma N\rfloor\\ v^1\neq\dots\neq v^k}} F\Big((v^1,\dots,v^k)\Big)\Big]\\ &\qquad= R\lim_{N\to\infty}\frac{1}{N^{k-1}}\mathbb{E}\Big[\prod_{i=1}^{k}\frac{1}{\rho_{V^i}}F\Big(D(V^1,\dots,V^k)\Big)\Big]\prod_{i=1}^{k-1}\mathbb{E}\Big[\sum_{j=1}^{i}\sum_{t=1}^{L^j}\rho_{V^j_t}\Big]\\ &\qquad= R\mathbb{E}\Big[\prod_{i=1}^{k}\frac{1}{\rho^i_{L^i}}F\Big(\frac{\log|H^i_{L^i}-H^j_{L^j}|}{\log R};\, i,j\Big)\Big]\prod_{i=1}^{k-1}\mathbb{E}\Big[\sum_{j=1}^{i}\int_0^{L^j}\rho^j_t\,\mathrm{d}t\Big].\end{aligned}$$

Moreover,

$$\frac{1}{(\log R)^{k-1}}\prod_{i=1}^{k-1}\mathbb{E}\Big[\sum_{j=1}^{i}\int_0^{L^j}\rho^j_t\,\mathrm{d}t\Big]\longrightarrow k!$$

and by Proposition 5.23, Proposition 5.24, and Proposition 5.25,

$$\mathbb{E}\Big[\prod_{i=1}^{k}\frac{1}{\rho^i_{L^i}}F\Big(\frac{\log|H^i_{L^i}-H^j_{L^j}|}{\log R};\, i,j\Big)\Big]\longrightarrow \sigma^k\mathbb{E}\Big[F\Big(\widehat{d}_{\mathbb{N}}(i,j);\, i,j\Big)\Big]$$

where $\widehat{d}_{\mathbb{N}}$ is the ultrametric obtained by sampling uniformly in the k-biased Brownian CPP. Moreover, if Y_1 denotes the total mass of a Brownian CPP, and $d_{\mathbb{N}}(i,j)$ the ultrametric obtained by sampling in the (unbiased) Brownian CPP, we have

$$\mathbb{E}\Big[F\Big(\widehat{d}(i,j);\, i,j\Big)\Big] = \frac{1}{k!}\mathbb{E}\Big[Y_1^k F\Big(d(i,j);\, i,j\Big)\Big].$$

Therefore, we have that

$$\lim_{R\to\infty}\lim_{N\to\infty}\frac{1}{(N\log R)^k}\mathbb{E}\Big[\sum_{|v^i|=\lfloor\sigma N\rfloor}F(D(v^1,\dots,v^k))\Big] = \mathbb{E}\Big[(\sigma Y_1)^k F\Big(d_{\mathbb{N}}(i,j);\, i,j\Big)\Big]$$

ending the proof. □

References for Chapter 5

[1] D. Aldous and L. Popovic. A critical branching process model for biodiversity. *Advances in Applied Probability* **37** (2005), 1094–1115.

[3] K. B. Athreya and P. E. Ney. *Branching Processes*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag Berlin Heidelberg, 1972.

[4] F. Baccelli, B. Błaszczyszyn, and M. Karray. *Random Measures, Point Processes, and Stochastic Geometry*. Inria, 2020.

[7] S. J. E. Baird, N. H. Barton, and A. M. Etheridge. The distribution of surviving blocks of an ancestral genome. *Theoretical Population Biology* **64** (2003), 451–471.

[15] N. H. Barton and A. M. Etheridge. The relation between reproductive value and genetic contribution. *Genetics* **188** (2011), 953–973.

[25] P. Billingsley. *Convergence of Probability Measures*. Second edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1999.

[37] J. T. Chang. Recent common ancestors of all present-day individuals. *Advances in Applied Probability* **31** (1999), 1002–1026.

[38] N. H. Chapman and E. A. Thompson. The effect of population history on the lengths of ancestral chromosome segments. *Genetics* **162** (2002), 449–458.

[43] A. Depperschmidt and A. Greven. Tree-valued Feller diffusion (2019). arXiv: 1904.02044.

[44] A. Depperschmidt, A. Greven, and P. Pfaffelhuber. Marked metric measure spaces. *Electronic Communications in Probability* **16** (2011), 174–188.

[52] R. Durrett. *Probability Models for DNA Sequence Evolution*. Second edition. Probability and its Applications. Springer, New York, NY, 2008.

[56] A. M. Etheridge. *Some Mathematical Models from Population Genetics. École d'Été de Probabilités de Saint-Flour XXXIX-2009*. Vol. 2012. Lecture Notes in Mathematics. Springer Science & Business Media, 2011.

[71] R. A. Fisher. A fuller theory of "junctions" in inbreeding. *Heredity* **8** (1954), 187–197.

[91] A. Greven, P. Pfaffelhuber, and A. Winter. Convergence in distribution of random metric measure spaces (Λ-coalescent measure trees). *Probability Theory and Related Fields* **145** (2009), 285–322.

[93] R. C. Griffiths and P. Marjoram. An ancestral recombination graph. *Progress in population genetics and human evolution*. Springer, 1997, 257–270.

[103] S. C. Harris and M. I. Roberts. The many-to-few lemma and multiple spines. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **53** (2017), 226–242.

[111] J. Jacod and A. N. Shiryaev. *Limit Theorems for Stochastic Processes.* Second edition. Grundlehren Der Mathematischen Wissenschaften. Springer-Verlag, 2003.

[117] P. Jagers and S. Sagitov. General branching processes in discrete time as random trees. *Bernoulli* **14** (2008), 949–962.

[121] O. Kallenberg. *Random Measures, Theory and Applications.* Probability Theory and Stochastic Modelling. Springer, Cham, 2017.

[139] A. Lambert, V. Miró Pina, and E. Schertzer. Chromosome Painting: how recombination mixes ancestral colors (2020). arXiv: `1807.09116`.

[143] J. Lamperti. Semi-stable Markov processes. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **22** (1972), 205–225.

[146] H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature* **475** (2011), 493–496.

[150] R. Lyons, R. Pemantle, and Y. Peres. Conceptual proofs of $L \log L$ criteria for mean behavior of branching processes. *Annals of Probability* **23** (1995), 1125–1138.

[168] J. C. Pardos and V. Rivero. Self-similar Markov processes. *Boletín de la Sociedad Matemática Mexicana* **19** (2013), 201–235.

[180] P. Ralph and G. Coop. The geography of recent genetic ancestry across Europe. *PLOS Biology* **11** (2013), 1–20.

[181] Y.-X. Ren, R. Song, and Z. Sun. A 2-spine decomposition of the critical Galton-Watson tree and a probabilistic proof of Yaglom's theorem. *Electronic Communications in Probability* **23** (2018), 12 pp.

[182] Y.-X. Ren, R. Song, and Z. Sun. Spine decompositions and limit theorems for a class of critical superprocesses. *Acta Applicandae Mathematicae* **165** (2020), 91–131.

[187] P. C. Sabeti, D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419** (2002), 832–837.

[199] Z. Shi. *Branching Random walks. École d'Été de Probabilités de Saint-Flour XLII-2012.* Vol. 2151. Lecture Notes in Mathematics. Springer, Cham, 2015.

Appendices for Chapter 5

5.A Palm measures

In this section, we recall some basic facts on the Palm measures of a point process, and provide an extension of Proposition 5.15 for general offspring point process. We follow the notation in [4] for the exposition of Palm measures.

Let Φ be a point process on some Polish space E. Let us denote by ν its intensity measure, defined as the unique measure on E such that

$$\int_E f(x)\,\nu(\mathrm{d}x) = \mathbb{E}\Big[\sum_{x\in\Phi} f(x)\Big],$$

for any bounded continuous function f. We now define a measure C on $E\times\mathcal{M}(E)$, where $\mathcal{M}(E)$ is the set of point measures on E, referred to as the Campbell measure of Φ. This measure is such that, for any continuous bounded function F, we have

$$\int_{E\times\mathcal{M}(E)} F(x,\mu)\,C(\mathrm{d}x,\mathrm{d}\mu) = \mathbb{E}\Big[\sum_{x\in\Phi} F(x,\Phi)\Big].$$

It is clear that the projection $C(\cdot,\mathcal{M}(E))$ of C to E is the intensity measure ν. By general disintegration results on finite measures, there exists a collection of random point processes $(\Phi^x;\, x\in E)$ such that

$$\int_{E\times\mathcal{M}(E)} F(x,\mu)\,C(\mathrm{d}x,\mathrm{d}\mu) = \int_E \mathbb{E}\Big[F(x,\Phi^x)\Big]\,\nu(\mathrm{d}x).$$

This family of point processes $(\Phi^x;\, x\in E)$ is called the family of *Palm measures* associated to Φ. It is not hard to see that for ν-almost every $x\in E$, we have $\mathbb{P}(x\in\Phi^x)=1$. Therefore, we can define for $x\in E$

$$\Phi^{!,x} = \Phi^x - \delta(x)$$

which is a point process for ν-almost every $x\in E$. The family of point processes $(\Phi^{!,x};\, x\in E)$ is called the family of *reduced Palm measures* associated to Φ.

The following result is a simple reformulation of the sampling procedure of the marked individuals in the 1-spine in terms of Palm measures.

Lemma 5.26. *Let Φ be a point process with intensity measure ν and H be such that*

$$\int_E H(x)\,\nu(\mathrm{d}x) < \infty.$$

Suppose that $\widetilde{\Phi}$ is distributed as

$$\mathbb{E}\Big[F(\widetilde{\Phi})\Big] \propto \mathbb{E}\Big[\sum_{x\in\Phi} H(x)F(\widetilde{\Phi})\Big]$$

and that, conditional on $\widetilde{\Phi}$, X is an atom of $\widetilde{\Phi}$ chosen with probability proportional to $H(x)$. Then X has distribution

$$\mathbb{E}\Big[f(X)\Big] \propto \int H(x)f(x)\,\nu(\mathrm{d}x),$$

and conditional on X, $\widetilde{\Phi} - \delta_X$ has the reduced Palm distribution of Φ conditional on having an atom at X.

Proof. We have

$$\begin{aligned}\mathbb{E}\Big[f(X)F(\widetilde{\Phi}-\delta_X)\Big] &= \mathbb{E}\Big[\frac{\sum_{x\in\widetilde{\Phi}} H(x)f(x)F(\widetilde{\Phi}-\delta_x)}{\sum_{x\in\widetilde{\Phi}} H(x)}\Big] \\ &= \frac{\mathbb{E}\Big[\sum_{x\in\Phi} H(x)f(x)F(\Phi-\delta_x)\Big]}{\int H(x)\,\nu(\mathrm{d}x)} \\ &= \frac{\int H(x)f(x)\mathbb{E}\Big[F\Big(\Phi^!_x\Big)\Big]\,\nu(\mathrm{d}x)}{\int H(x)\,\nu(\mathrm{d}x)},\end{aligned}$$

where $\Phi^!_x$ denotes the reduced Palm distribution of Φ with an atom at x. The first equality follows from the definition of X, the second from that of $\widetilde{\Phi}$ and the last from the Campbell-Little-Mecke formula, see for instance Corollary 3.1.14 in [4]. □

The Palm measure Φ^x corresponds, roughly speaking, to the distribution of Φ "conditioned on having an atom at x". We now introduce higher order Palm measures, which correspond the distribution of Φ, conditioned on having atoms located at ζ, where ζ is a finite point measure on E.

We proceed inductively on the number of atoms of ζ. If $\zeta = \delta_x$, we define

$$\Phi^{!,\zeta} = \Phi^{!,x}.$$

Moreover, if $x \in \zeta$, we define

$$\Phi^{!,\zeta} = \Big(\Phi^{!,\zeta-\delta_x}\Big)^{!,x}.$$

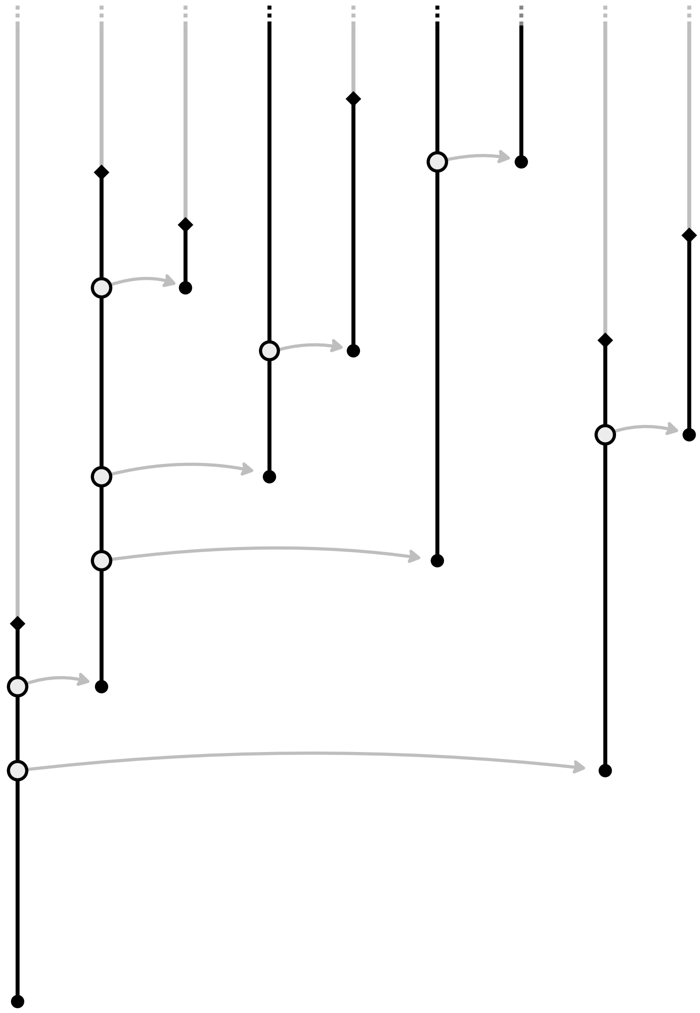
It can be checked that this definition does not depend on the order in which the atoms of ζ are considered. See Chapter 3 in [4] for a rigorous definition.

Recall the notation $(X^i, K^i, L^i;\, i \leq k)$ for the k-spine, and the notation ζ^i_t (resp. $\bar{\zeta}^i_t$) for the point process of children of X^i_t that do not belong (resp. belong) to the spine $\mathbf{S}_k$. Then we have the following reformulation of Proposition 5.15 in the general case. It can be proved along the same lines as Proposition 5.15.

Proposition 5.27. *Conditional on* $(X^i, K^i, L^i;\ i \leq k)$*, the collection of point processes* $(\zeta^i_t;\ i \leq k,\ t \leq L^i)$ *is independent. Moreover* ζ^i_t *has the reduced Palm distribution of the original point process* $\xi(X^i_t)$*, conditional on having atoms located at* $\bar{\zeta}^i_t$*.*

Branching models in epidemiology

Chapter 6



6

From individual-based epidemic models to McKendrick-von Foerster PDEs: A guide to modeling COVID-19 dynamics

This chapter has been submitted to *Theoretical Population Biology* [77], and I am listed as the first author of this project.

Illustration. Graphical representation of a Crump-Mode-Jagers branching process, which is the general framework from which our model is built.

6.1 Introduction

6.1.1 Challenges posed by complex diseases such as COVID-19

The transmission of pathogens between species is a global concern [17, 42]. As such zoonotic episodes are expected to become increasingly common in humans, it is critical to develop analytic tools that can quickly transform epidemiological observations into informed public policy in order to mitigate and control epidemics.

A novel coronavirus, SARS-CoV-2, has recently crossed the species barrier into humans and, within months, has rapidly spread to all corners of our planet [219]. The sheer scale of this pandemic has overburdened our medical infrastructure, caused fatalities estimated well into the hundreds of thousands, and shut down entire economies. Remarkably, the rapid spread of COVID-19 and its consequences can be attributed to the unique life cycle of a 30,000 base pair single-stranded virus. SARS-CoV-2 is an airborne pathogen transmitted by both symptomatic and asymptomatic carriers in close proximity to non-infected individuals. Milder COVID-19 symptoms include a dry cough, fever, and/or shortness of breath while more serious cases include respiratory failure and eventual death. With millions of infections and hundreds of thousands of documented deaths and recoveries, the COVID-19 pandemic is providing a wealth of independent estimates of important clinical characteristics that can help predict health outcomes specific for a country or region.

It quickly became understood that accurate descriptions of the life cycle of this disease needed to distinguish between several stages of the disease, referred to as compartments, depending on whether an infected individual is infectious or not, symptomatic or not, hospitalized, *etc.* However it remains unclear to what extent making precise predictions of the dynamics of such a complex disease requires to have a precise knowledge of clinical features such as incubation period, generation time, and duration times between infection, symptom establishment, hospitalization, recovery and death, to know how these durations correlate and what are the exact probabilities of transition between stages.

In this work, we consider a fully stochastic, generic epidemiological model with an arbitrary number of compartments, that encompasses life cycles of most complex diseases and that of COVID-19 in particular. We show how structuring the infected population by its infection age, i.e., time elapsed since infection, allows us to decouple dependencies between stages and to time. More specifically, when the population size is large enough, the joint evolution of all compartment sizes can be described by means of a linear, first-order partial differential equation (PDE) known as the McKendrick-von Foerster equation describing the number $n(t, a)$ of infecteds of (infection) age a at time t. The boundary condition at age 0 is driven by the infection rate from infecteds of age a *averaged* over all life cycles and the number of individuals of age a in compartment i at time t is obtained by thinning $n(t, a)$ by a factor $p(a, i)$ which is the probability of being in compartment i conditional on having age a, *averaged* over all life cycles.

In the case of COVID-19, we display a simple procedure to infer these parameters, some from the biological literature and most from time series of numbers of severe cases, hospitalized cases, discharged patients and deaths that can be applied easily to any regional or national dataset. We also allow for time inhomogeneity in the infection rate to account for temporary mitigation measures such as lockdowns or social distancing. We apply this procedure to French COVID-19 data from March to May 2020 and estimate various parameters of interest including R_0 in different phases of the epidemic (before, during, and after lockdown) and biological parameter values that we compare to empirical estimates.

6.1.2 Generic model assumptions

We consider a population model of the SIR fashion where each individual goes through successive stages, starting from stage S (susceptible) and ending in one of two states: R (recovered) or D (dead). Depending on disease complexity, the number of stages in this life cycle can vary. In the SARS-CoV-2 example, typical intermediate stages are A (asymptomatic) or P (presymptomatic), I (mild case) or C (severe case), H (hospitalized), U (intensive care unit). These stages are sometimes called compartments, types, classes, stages or simply states. See Figure 6.4 for an illustration.

We assume that S individuals are always in excess (branching assumption) and that each individual infects new S individuals at successive times of a random point process, one at a time. We further assume that upon infection, an S individual immediately changes state and never returns to state S (ruling out multiple infections in particular). More formally,

- The set of all possible states is denoted $\mathcal{S}$ and we consider that a stochastic process $(X(a);\ a \geq 0)$ with values in $\mathcal{S}$ gives the state of a typical individual of age a, where here *age* means *age of infection*, i.e., time elapsed since infection. For the sake of simplicity, we will assume that $\mathcal{S}$ is a discrete space.

- A random point measure $\mathcal{P}$ describes the times of secondary infections. Due to the previous assumptions, atoms of $\mathcal{P}$ all have mass 1 and only charge $(0, \infty)$.

- We can (and will) superimpose time heterogeneity to this process by means of a *suppression function* $(c(t);\ t \geq 0)$ valued in $[0, 1]$ thinning the infection process. More precisely, if t is a potential time of infection for individual x (i.e., t is an atom of its infection point process $\mathcal{P}_x$), we ignore the event with probability $1 - c(t)$. This suppression function can model the effect of vaccination, of density-dependence (i.e., relaxing the branching assumption due to an excess of removed or of deceased individuals), or of governmental mitigation measures (i.e., social distancing, lockdown).

The population is thus described by a *Crump-Mode-Jagers branching process*, where all individuals x are characterized by independent copies $(\mathcal{P}_x, X_x)$ of the pair $(\mathcal{P}, X)$ describing, respectively, the process of infection and the life cycle. In the branching process literature, X is often referred to as a *random characteristic* of individual x [162, 116, 205].

Remark 6.1. A typical infection measure is a Poisson Point Process with intensity λ killed at an independent exponential random variable with parameter γ. (By killing we mean that we erase all atoms of the PPP after the killing time.) This corresponds to the classical SIR process with rate of infection λ and recovery rate γ. More generally, we can construct a SEIR process (E for exposed) as follows. Let L be a random variable and consider ξ be a Poisson Point Process with an intensity measure $\lambda(x)\,\mathrm{d}x$. Then define $\mathcal{P}([a, b]) = \xi([(a - L)^+, (b - L)^+])$ so that no infection occurs during the incubation period $[0, L]$. ○

Remark 6.2. $\mathcal{P}$ and X are generally *not independent*. As a simple example, since (most) diseases cannot be transmitted by deceased individuals, no atom of $\mathcal{P}$ is allowed once X has reached the end-state D. In the same spirit, one could assume that the infection potential of a given individual is reduced once in the hospital and that individuals with many atoms in their infection process $\mathcal{P}$ (high infectiosity) are identified and isolated. ○

Remark 6.3. Classes such as P, I or H must sometimes be refined to account for additional structuring variables like general health condition, (real) age, spatial position or previous exposure to similar pathogen. Knowledge of such variables can help predict more accurately the outcome of the infection and parametrize more precisely the infection process. Regarding this last point, note that the assumptions in force here allow for any implicit or explicit structure provided that transmission from an individual of type i to an individual of type j does not depend on j (but may depend on i, as we have seen). Relaxing this assumption would result in describing the large population limit by a multidimensional PDE instead of a one-dimensional PDE (see Section 6.1.4). Also, note that ignoring structuring by a hidden variable such as spatial position or health condition can lead to difficulties in estimating sojourn times in each compartment (such as P, I or H) from clinical data, due to over- or under-representation in this compartment of subsets of individuals carrying certain values of the hidden variable. ○

6.1.3 Statement of results and outline

The stochastic epidemic models we consider here are fairly general and can exhibit quite complex dependencies (i) between states and time, due to the lack of any Markov-type assumption, (ii) between states, due to possibly hidden structuring variables impacting the life cycle, (iii) between state and infection rate, and (iv) between past and future infection events. The main message of this note is that despite this apparent complexity, most of this complexity vanishes when the size of the population is large. More specifically, we show that in the limit of large populations (obtained by starting from a large initial population or as a consequence of natural exponential growth), the population of infecteds structured by age (of the infection) can be described by a one-dimensional PDE that only depends on

(i) the average infection rate

$$\tau(\mathrm{d}a) \coloneqq \mathbb{E}(\mathcal{P}(\mathrm{d}a));$$

(ii) the one-dimensional marginals of the life-cycle process

$$p(a,i) \coloneqq \mathbb{P}(X(a)=i).$$

Large initial population. Let us start with N infected individuals and define the empirical measure μ_t^N describing the ages and types of infected individuals present at time t

$$\mu_t^N(\mathrm{d}a \times \{i\}) \coloneqq \sum_x 1_{\sigma_x<t}\delta_{(t-\sigma_x, X_x(t-\sigma_x))}(\mathrm{d}a\times\{i\}), \tag{6.1}$$

where the sum is taken over all individuals x having ever lived and σ_x denotes the birth (infection) time of x. According to our first result (Theorem 6.11), starting from N infected individuals at time 0 with i.i.d. infection ages with law g, we have the a.s. convergence

$$\lim_{N\to\infty} \frac{1}{N}\mu_t^N(\mathrm{d}a \times \{i\}) = n(t,a)p(a,i)\,\mathrm{d}a,$$

where $n(t,a)$ solves the McKendrick-von Foerster PDE

$$\begin{aligned} \forall t, a > 0, \quad \partial_t n(t,a) \ + \ \partial_a n(t,a) = 0 \\ \forall t > 0, \quad n(t,0) = c(t)\int_0^\infty \tau(\mathrm{d}a) n(t,a) \\ \forall a \geq 0, \quad n(0,a) = g(a). \end{aligned} \tag{6.2}$$

After lockdown. Our second result (TheoremTheorem 6.13) displays a similar, but more subtle, convergence in the case when the process is supercritical, where natural growth leads by itself to large population sizes. Let $Z(t)$ denote the total population size at time t and assume that $Z(0) = 1$, i.e., we start from *a single individual*. By a slight abuse of notation, denote by $\mu_t^{t_K}$ the empirical measure of ages and types as in (6.1), but under the assumption that the suppression function at time t is equal to $c(t - t_K)$ where c is equal to 1 for negative arguments, and t_K is some large, random time. We are motivated by modeling a situation where the infection is separated into two distinct phases:

(Phase 1) We let the epidemic develop until a certain random time t_K. For instance, t_K could be the time at which the number of recorded deaths exceeds a large threshold K. We assume no suppression before t_K;

(Phase 2) We let the suppression function vary after time t_K, e.g., due to mitigation measures and/or behavioral changes (i.e., lockdown phase).

Conditional on non-extinction, letting (t_K) be any sequence of stopping times such that $t_K \to \infty$ on the non-extinction event, we have the following convergence in probability

$$\lim_{K\to\infty} \frac{\mu_{t_K+t}^{t_K}}{Z(t_K)} = n(t,a)p(a,i)\,\mathrm{d}a.$$

Now $n(t,a)$ solves the McKendrick-von Foerster PDE with the same boundary condition as previously, but with initial condition $n(0,a) = \alpha e^{-\alpha a}$, where $\alpha > 0$ is the exponential growth rate of $Z(t)$ for $t \leq t_K$, also called Malthusian parameter. This second result can be seen as a refinement of limit theorems for exponentially growing populations counted with random characteristics, where here the characteristic of a typical individual is the number of her descendants in class i of age a, born at least s time units after her birth (summed over $s_x = t_K - \sigma_x$). In particular, taking $t = 0$ in the statement yields the convergence to the exponential stable age distribution decorated by the one-dimensional marginals p of X. The way we state the result nicely displays dependencies between characteristics corresponding to different t's.

To summarize: *(1) the macroscopic infection process is characterized by the sole intensity measure τ and dictates an explicit age structure of the population, and (2) the class structure is deduced by averaging the life-cycle process against the limiting age profile.* In order to validate our approach, we use those deterministic approximations to infer epidemiological parameters (R_0 before and during lockdown) from recent empirical observations, and show that our findings are in accordance with the current literature.

Outline. The paper is organized as follows. In Section 6.2, we study equation (6.2). After providing a precise description of the branching process that we are considering in Section 6.2.1, we give the definition of a weak solution to (6.2) in Section 6.2.2. Then, we give two probabilistic representations of these weak solutions. We first show in Section 6.2.3 that the weak solution to (6.2) corresponds to the first moment of the branching process that we are studying, when viewed as a random measure on the ages of infection. Second, Section 6.2.4 provides a construction of the weak solution using a dual genealogical process. The two laws of large numbers are proved in Section 6.3. Finally the inference in carried out in Section 6.4. Section 6.4.2 and Section 6.4.3 describe our choice of parametrization and the inference results are discussed in Section 6.4.4.

6.1.4 Natural extensions

Some of the assumptions underlying the previous models can be relaxed and our general framework can be adapted to more complex and realistic populations.

Contact matrix. So far, infectious individuals infect new individuals uniformly at random. In general, a contact matrix specifies the contact rate, depending on contact location (household, school, work, ...) or individual types of source and target (real age, susceptibility, ...) [34, 33]. More precisely, each individual now belongs to one class, and we denote by $\mathcal{C}$ the (finite) set of all classes. An individual in class $j \in \mathcal{C}$ is characterized by a multi-dimensional process $(X^j, \mathcal{P}^{j,1}, \dots, \mathcal{P}^{j,C})$ where the atoms of $\mathcal{P}^{j,k}$ provide the age at which this individual infects a new individual of type $k \in \mathcal{C}$, and $(X^j(a);\, a \geq 0)$ is a $\mathcal{S}$-valued process whose distribution can depend on j.

We define the mean contact matrix as

$$\forall j, k \in \mathcal{C}, \quad \tau^{jk}(\mathrm{d}u) := \mathbb{E}(\mathcal{P}^{j,k}(\mathrm{d}u)).$$

The population is again described by a multi-type Crump-Mode-Jagers branching process. Analogously to (6.1), $\mu_t^{\mathbf{N},j}$ denotes the empirical measure of j individuals starting with $\mathbf{N} = (N^j;\, j \in \mathcal{C})$ infected individuals at time $t = 0$. Assume that there exists a constant N such that for all $j \in \mathcal{C}$, $N^j/N \to y^j$ as $N \to \infty$. Under the usual technical conditions (the matrix τ is irreducible and Malthusian, $x \log x$ type condition, see [39]),

$$\forall j \in \mathcal{C}, \forall i \in \mathcal{S}, \quad \frac{1}{N^j} \mu_t^{\mathbf{N},j}(\mathrm{d}a \times \{i\}) \Longrightarrow n^j(t,a) \mathbb{P}(X^j(a) = i) \,\mathrm{d}a$$

where $(n^j;\, j \in \mathcal{C})$ satisfies the multidimensional McKendrick-von Foerster equation

$$\begin{aligned} &\partial_t n^j(t,a) + \partial_a n^j(t,a) = 0 \\ &\forall t > 0, \quad n^j(t,0) = \sum_{k\in\mathcal{C}} c^{kj}(t) \int_0^\infty \tau^{kj}(\mathrm{d}a) n^k(t,a), \\ &\forall a \geq 0, \quad n^j(0,a) = y^j g^j(a). \end{aligned} \tag{6.3}$$

where g^j describes the initial age profile of class j and $c^{jk}(t)$ is a matrix at time t generalizing the suppression function of the previous section. Following Theorem 6.13, it is natural to consider the initial condition

$$g^j(u) = \alpha \exp(-\alpha u) \varphi^j$$

where α is the Malthusian parameter, i.e. the unique α such that the largest eigenvalue of the matrix

$$\int \exp(-\alpha u)\, \tau(\mathrm{d}u)$$

is equal to 1 and $(\varphi^j;\, j \in \mathcal{C})$ is its Perron-Frobenius left eigenvector with (positive) entries summing up to 1. As in Theorem 6.13, such initial condition can be justified by starting with a single individual and let the population grow up to a large random time t_K conditional on non-extinction [39].

Shortage of susceptibles. Another natural extension consists in taking into account saturation of infected in the population. Start with a finite, but large population of size N with a fraction of infected individuals $x \in (0,1)$ with an age profile g at time $t = 0$. Here infection is only effective if the target individual is susceptible, which thins infection rate by the fraction of susceptibles in the population. At the limit, this saturation translates into a non-linear McKendrick-von Foerster equation

$$\begin{aligned} &\forall t, a > 0, \quad \partial_t n(t,a) + \partial_a n(t,a) = 0 \\ &\forall t > 0, \quad n(t,0) = S(t) c(t) \int_0^\infty \tau(\mathrm{d}a) n(t,a), \\ &\forall a \geq 0, \quad n(0,a) = x g(a), \end{aligned} \tag{6.4}$$

where we have defined

$$S(t) \coloneqq 1 - \int_0^\infty n(t,a)\, \mathrm{d}a,$$

and the limiting empirical measure is given by $n(t,a)p(a,j)\,\mathrm{d}a$. Convergence results to this limiting PDE are addressed by some of the present authors in Chapter 7.

6.1.5 Compartmental ODE models

An important special case of our model is when the process $(X(a);\, a \geq 0)$ is a Markov process and infections from individual x occur at a rate that only depends on the current state of x.

Under these assumptions, the McKendrick-von Foerster PDE reduces to a finite set of ODEs. Similar sets of ODEs have been widely used to model the SARS-CoV-2 epidemic [189, 185, 60, 46], and in that sense, *taking into account explicitly the (infection) age structure of the population allows us to incorporate all these models into the same general framework.*

More precisely, for $i \in \mathcal{S}$ define

$$\forall t \geq 0, \quad N_i(t) = \int_0^\infty n(t,a) p(a,i) \, \mathrm{d}a$$

to be the number of individuals in state i. We will assume that $(X(a);\, a \geq 0)$ is a Markov process with transitions $(\lambda_{ij};\, i,j \in \mathcal{S})$. Moreover, we suppose that conditional on $(X(a);\, a \geq 0)$, the infection process $\mathcal{P}_x$ from individual x is a Poisson point process with intensity rate τ_i when x is in state i. Then a direct computation shows the following result.

Proposition 6.4. *Suppose that $(X(a);\, a \geq 0)$ is a Markov process with transitions $(\lambda_{ij};\, i,j \in \mathcal{S})$, and that conditional on $(X(a);\, a \geq 0)$, $\mathcal{P}$ is a Poisson point process with intensity $(\tau_{X(a)};\, a \geq 0)$. Then, if $(n(t,a))$ denotes the solution to (6.4), $(N_i(t);\, t \geq 0)_{i \in \mathcal{S}}$ solves the following set of ODE:*

$$\forall i \in \mathcal{S}, \quad \dot{N}_i = \sum_{j \in \mathcal{S}} \lambda_{ji} N_j - N_i \sum_{j \in \mathcal{S}} \lambda_{ij} + \sum_{j \in \mathcal{S}} a_{ji} N_j S, \tag{6.5}$$

where $a_{ji} := \tau_j p(0,i)$ and $S := 1 - \sum_{i \in \mathcal{S}} N_i$.

Proof. Recall that

$$N_i(t) = \int_0^\infty n(t,a) p(a,i) \, \mathrm{d}a.$$

By differentiating both sides with respect to time we get

$$\begin{aligned} \dot{N}_i(t) &= \int_0^\infty \partial_t n(t,a) p(a,i) \, \mathrm{d}a = - \int_0^\infty \partial_a n(t,a) p(a,i) \, \mathrm{d}a \\ &= \int_0^\infty n(t,a) \partial_a p(a,i) \, \mathrm{d}a + n(t,0) p(0,i). \end{aligned}$$

By using the boundary condition and the fact that $(X(a);\, a \geq 0)$ is a Markov process, we obtain that

$$\begin{aligned} \dot{N}_i(t) &= \int_0^\infty n(t,a) \Big(\sum_{j \in \mathcal{S}} \lambda_{ji} p(a,j) - \lambda_{ij} p(a,i) \Big) \mathrm{d}a + S(t) p(0,i) \int_0^\infty n(t,a) \sum_{j \in \mathcal{S}} \tau_j p(a,j) \, \mathrm{d}a. \\ &= \sum_{j \in \mathcal{S}} \lambda_{ji} N_j(t) - N_j(t) \sum_{j \in \mathcal{S}} \lambda_{ij} + \sum_{j \in \mathcal{S}} a_{ji} S(t) N_j(t), \end{aligned}$$

which ends the proof. □

Conversely, let us consider from the start the system of ODEs (6.5). Here, λ_{ij} is interpreted as the rate at which a type i individual turns into type j and a_{ij} as the rate at which a type i individual gives birth to a type j individual. If the contact matrix with generic entries a_{ij} has rank 1 and non-negative entries, it can always be decomposed as $a_{ij} = \tau_i p(0, j)$ where $\tau_i \geq 0$ and $(p(0, j))$ is a probability vector. The vectors (τ_i) and $(p(0, j))$ can be recovered by

$$\tau_i = \sum_{j \in \mathcal{S}} a_{ij} \quad \text{and} \quad p(0, j) = \frac{a_{ij}}{\tau_i} \quad (\forall i, j).$$

Note that here $a_{ij} = \lambda(i)p(0, j)$, so that actually a type i individual gives birth at rate $\lambda(i)$ and her offspring has type independently distributed according to $p(0, \cdot)$. As a result, the contact matrix with generic entries a_{ij} has rank 1, and λ and $p(0, \cdot)$ can be recovered by

$$\lambda(i) = \sum_{j \in \mathcal{S}} a_{ij} \quad \text{and} \quad p(0, j) = \frac{a_{ij}}{\lambda(i)} \quad (\forall i, j).$$

Then one can define $(X(a);\, a \geq 0)$ as the $\mathcal{S}$-valued process with rates given by the matrix Λ (with diagonal entries $\lambda_{ii} = -\sum_{j \notin \mathcal{S}} \lambda_{ij}$) and initial distribution $(p(0, i))$. Denote as in the rest of the text

$$p(a, i) = \mathbb{P}(X(a) = i),$$

so that the row matrix $p(a, \cdot)$ can be computed as the product $p(0, \cdot) \exp(a\Lambda)$.

Now let us consider the solution $(N_i(t);\, t \geq 0)$ to (6.5) and assume there is some age distribution g (integrable but possibly not summing to 1) such that

$$N_i(0) = \int_0^\infty g(a) p(a, i) \,\mathrm{d}a. \tag{6.6}$$

Then by uniqueness of (N_i) and thanks to Proposition 6.4, for all $i \in \mathcal{S}$,

$$N_i(t) = \int_0^\infty n(t, a) p(a, i) \,\mathrm{d}a,$$

where n is the solution to the McKendrick-von Foerster PDE with initial condition g and boundary condition

$$n(t, 0) = \int_0^\infty \tau(a) n(t, a) \,\mathrm{d}a,$$

where $\tau(a) := \sum_{j \in \mathcal{S}} \tau_j p(a, j)$. This shows that the solution to any linear system of ODEs of the form (6.5) has a simple representation in terms of the solution to the McKendrick-von Foerster PDE decorated with types, provided there is a representation of the initial condition in the form (6.6). Note that this last property is not necessarily fulfilled. For example, if X is ergodic and started in its stationary probability distribution, then $p(a, i) = p(0, i)$ and (6.6) would only hold if $(n_i(0);\, i \in \mathcal{S})$ were proportional to $(p(0, i);\, i \in \mathcal{S})$.

If the matrix with generic entries a_{ij} does not have rank 1, one could derive a similar representation of the solutions to (6.5), but using the multi-dimensional version of the McKendrick-von Foerster equation (6.3).

6.1.6 Relation with previous works

Non-Markov epidemic models have already been investigated, see e.g. [198, 166, 9, 10]. The idea of representing a general branching population by its age structure has a rich history in probability theory [112, 65, 115, 114, 113, 101, 210, 69] and the connection with the McKendrick-von Foerster PDE has been acknowledged several times [65, 101]. In the latter two works, the authors allow for birth and death rates that may depend not only on abundances of each type, but also on the whole age structure of the population. This impressive level of generalization comes at the cost of assuming that the process describing the evolution of the empirical measure on ages and types is Markovian. In particular, birth and death rates are not allowed to depend on past individual birth events. The Markov property then allows the use of a generator for the empirical measure and with some extra finite second moment assumptions on the intensity measure, this approach allows the authors to obtain a law of large numbers and a central limit theorem.

We acknowledge that the current work is certainly not as mathematically challenging as the works alluded to above, and that some of our results are almost implicit in some of the previous works. However, we believe that our point of view (one-dimensional PDE decorated with types) does deserve to be highlighted in the current sanitary crisis since it provides a natural and efficient inference methodology. More than 70,000 publications related to the COVID-19 crisis have appeared since the onset of the pandemic, with many different modeling approaches. One of the modest aims of the present note is to convey the idea that individual-based stochastic models suggest a simple and tractable framework for tackling some of the complex features of the disease. Furthermore, since we ignore finite population effects, our proofs are quite elementary compared to [65, 101] and should be accessible to a much wider audience interested in such a modeling approach.

Finally, we already pointed out that the connection between branching processes and McKendrick-von Foerster PDE has been discussed in previous works. However, as far as we can tell, the duality result exposed in Section 6.2.4 is new and can presumably be extended to more general branching processes where birth and death rates are allowed to be frequency-dependent. In Chapter 7, some of the authors of the present work show that this duality result has a natural counterpart in a model with a finite but large population.

6.2 Two Feynman-Kac formulæ

6.2.1 Assumptions and notation

CMJ branching process with suppression. Recall that the infection process is modeled by a Crump-Mode-Jagers (CMJ) branching process [112, 162] with no death, starting from one individual called the progenitor (or root of the tree). Each individual x is characterized by an independent pair $(\mathcal{P}_x, X_x)$ embodying respectively the processes of secondary infection events from x and of types carried

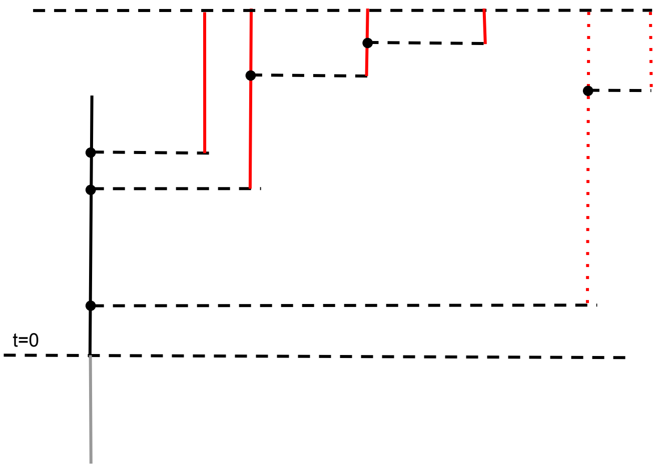


Figure 6.1: The initial individual $(\tilde{\mathcal{P}}, \tilde{X})$ is represented by a black segment. In Section 6.2.1, we assume that at time $t = 0$, the age of the initial individual (length of the grey segment) is distributed according to a probability density g. If a branching event is observed at time t (see e.g., black dots), the infection occurs with probability $c(t)$. In the CMJ, this amounts to prune the corresponding subtree with probability $c(t)$ (dotted red tree).

by x. Each pair $(\mathcal{P}_x, X_x)$ is a copy of the pair $(\mathcal{P}, X)$ with law $\mathcal{L}$, except when x is the root, where it is distributed as $(\tilde{\mathcal{P}}, \tilde{X})$ with law $\tilde{\mathcal{L}}$ (more on that below).

Also recall some infection events can be suppressed using a suppression function $(c(t);\, t \geq 0)$. Given a realization of the CMJ tree, for each branching point occurring at time t, we trim the tree by independently pruning the subtree stemming from it with probability $c(t)$. See Figure 6.1.

For simplicity, we will assume that the suppression function c is a piecewise right-continuous function, and that for any $t \geq 0$, the process $(X(a);\, a \geq 0)$ is a.s. continuous at t. Define the average infection measure (that is, the intensity measure of the point process $\mathcal{P}$) as

$$\tau(\mathrm{d}u) := \mathbb{E}(\mathcal{P})(\mathrm{d}u).$$

We assume that τ is absolutely continuous w.r.t. the Lebesgue measure in such a way that there exists a measurable non-negative function β such that

$$\tau(\mathrm{d}u) = \beta(u)\,\mathrm{d}u \quad \text{and} \quad R_0 := \int_0^\infty \beta(u)\,\mathrm{d}u < \infty. \tag{6.7}$$

We also assume that there exists a unique parameter $\alpha \in \mathbb{R}$ (the so-called Malthusian parameter of the CMJ process) such that

$$\int \exp(-\alpha u)\,\tau(\mathrm{d}u) = 1. \tag{6.8}$$

The parameter α can be either positive (supercritical) or negative (subcritical). Finally, we will also enforce the Kesten and Stigum criterium [165]

$$\mathbb{E}\Big(R_\alpha \log^+(R_\alpha)\Big) < \infty, \tag{6.9}$$

where

$$R_\alpha := \int_0^\infty \exp(-\alpha t)\,\mathcal{P}(\mathrm{d}t).$$

Initial shifted law. For any finite measure m on $\mathbb{R}_+$, we define $\theta_t \circ m$ as the measure shifted by t, i.e.,

$$\int_0^\infty f(u)\theta_t \circ m(\mathrm{d}u) = \int_t^\infty f(u-t)\,m(\mathrm{d}u),$$

where f is any measurable, bounded function f on $\mathbb{R}_+$. For any measurable function $F : \mathbb{R}_+ \to \mathcal{S}$, we similarly define $\theta_t \circ F$ by $\theta_t \circ F(u) = F(u+t)$. (We make a slight abuse of notation by using the same symbol for the shift operator on measures and functions.)

Let g be a probability density on $\mathbb{R}_+$. We now specify the law $\tilde{\mathcal{L}}_g$ of the pair $(\tilde{\mathcal{P}}, \tilde{X})$ characterizing the root. In order to connect the CMJ process with the McKendrick-von Foerster equation, we will focus on the case where $(\tilde{\mathcal{P}}, \tilde{X})$ is identical in law to $(\theta_A \circ \mathcal{P}, \theta_A \circ X)$, where A is a r.v. independent of $(\mathcal{P}, X)$ and distributed according to g. The distribution $\tilde{\mathcal{L}}_g$ will be referred to as the *shifted law.* In particular, we have

$$\tilde{\tau}(\mathrm{d}a) := \mathbb{E}(\tilde{\mathcal{P}}(\mathrm{d}a)) = \left(\int_0^\infty \beta(x+a)g(x)\,\mathrm{d}x\right)\mathrm{d}a.$$

Notation. We assume that individuals are indexed by the standard Ulam-Harris labeling. Namely, individuals are indexed in $I = \cup_n (\mathbb{N}^*)^n$. If $x \in I$, then xi (the concatenation of x and i) is interpreted as the i-th child of x. Children are ranked according to their birth time: $(x,1)$ is the oldest child of x, $(x,2)$ the second oldest, *etc.* (Since we assumed that τ has a density, there is no simultaneous births and the atoms of $\mathcal{P}$ are distinct a.s.) We denote by σ_x the date of birth of x with the convention that $\sigma_x = \infty$ if the individual is never born. For instance, if $\sigma_x < \infty$ and x has k children, then $\sigma_{xj} = \infty$ for $j > k$. Finally, $\emptyset$ will denote the root of the tree.

6.2.2 McKendrick-von Foerster PDE: Weak solutions

In this section, we consider the time-inhomogeneous, linear McKendrick-von Foerster PDE (6.2). The first line in (6.2) is the transport equation with unit velocity, i.e., ages of individuals increase at rate 1. The second line gives the number of newly infected individual (age 0) at time t.

In order to motivate our definition of weak solutions, we start by giving a formal resolution of the PDE using the method of characteristics. Fix $a > 0$. Let

$$A(t) = a - t$$

Then

$$\frac{\mathrm{d}}{\mathrm{d}s} n(t-s, A(s)) = -\partial_t n(t-s, A(s)) - \partial_a n(t-s, A(s)) = 0,$$

so that $s \mapsto n(t-s, a-s)$ is conserved along the characteristics, i.e.,

$$\forall s < a, \quad n(t,a) = n(t-s, a-s).$$

It follows that

$$n(t,a) = \begin{cases} g(a-t) & \text{when } a > t \\ b(t-a) & \text{when } a \leq t \end{cases} \tag{6.10}$$

where $b(t) = c(t) \int_0^\infty \tau(\mathrm{d}a) n(t,a)$. We now determine the function b. Injecting the previous expression into the "spatial" boundary condition of the PDE, we obtain a delayed equation for b: for every $t > 0$

$$b(t) = c(t) \int_0^t \tau(\mathrm{d}a) b(t-a) + c(t) \int_t^\infty \tau(\mathrm{d}a) g(a-t). \tag{6.11}$$

Lemma 6.5. *There exists a unique solution b to (6.11) which is locally integrable. Moreover, for any $\delta \geq 0$ such that $\delta > \alpha$ we have $b \in \mathcal{L}^{1,\delta}$, where $\mathcal{L}^{1,\delta}$ denotes the set of all functions $f\colon \mathbb{R}_+ \to \mathbb{R}$ such that $\|f\|_{L^{1,\delta}} := \int_0^\infty e^{-\delta t} |f(t)| \, \mathrm{d}t < \infty$.*

Proof. Fix $\delta > \alpha$ and denote by $L^{1,\delta}$ the space $\mathcal{L}^{1,\delta}$ quotiented by the relation $\sim_\delta$, where $f \sim_\delta g$ if $\|f - g\|_{L^{1,\delta}} = 0$. Then define the linear operator $\Phi\colon L^{1,\delta} \to L^{1,\delta}$ by

$$\Phi f \colon t \mapsto c(t) \int_0^t f(t-u) \beta(u) \, \mathrm{d}u.$$

Then we have

$$\begin{aligned} \|\Phi f\|_{L^{1,\delta}} &= \int_0^\infty e^{-\delta t} \Phi f(t) \, \mathrm{d}t = \int_0^\infty e^{-\delta t} c(t) \int_0^t f(t-u) \beta(u) \, \mathrm{d}u \, \mathrm{d}t \\ &= \int_0^\infty e^{-\delta u} f(u) \int_u^\infty \beta(t-u) e^{-\delta(t-u)} c(t) \, \mathrm{d}t \, \mathrm{d}u. \end{aligned}$$

Now using that

$$\int_u^\infty \beta(t-u) e^{-\delta(t-u)} c(t) \, \mathrm{d}t \leq \int_0^\infty \beta(t) e^{-\delta t} \, \mathrm{d}t < 1$$

we obtain that $\|\Phi\| < 1$. Define

$$\Psi := \mathrm{Id} - \Phi.$$

Then Ψ is invertible with inverse $\sum_{k \geq 0} \Phi^k$. Note that equation (6.11) can be written as

$$\Psi(b) = F,$$

where

$$F \colon t \mapsto c(t) \int_t^\infty \beta(a) g(t-a) \, \mathrm{d}a.$$

The proof ends noting that $F \in L^{1,\delta}$ as

$$\int_0^\infty e^{-\delta t} F(t) \, \mathrm{d}t \leq \int_0^\infty \int_t^\infty \beta(a) g(t-a) \, \mathrm{d}a \, \mathrm{d}t < \infty.$$

We have thus proved existence and uniqueness of the solution b to (6.11) in $L^{1,\delta}$. Now for any two functions b_1 and b_2 such that $b_1 \sim_\delta b_2$ and b_1 and b_2 both satisfy (6.11), we have $b_1 = b_2$ (i.e., there is a single element in the equivalence class of b). This shows uniqueness of the solution b to (6.11) in $\mathcal{L}^{1,\delta}$.

Since all elements of $L^{1,\delta}$ are locally integrable, this also shows the existence of a locally integrable solution to (6.11). Its uniqueness can be proved following the exact same reasoning as previously, replacing integrations on $[0, \infty)$ by integration on compact intervals. □

Definition 6.6. We say that $(n(t, a);\, a, t \geq 0)$ is the $\mathcal{L}^{1,\mathrm{loc}}$ weak solution to the McKendrick-von Foerster PDE with initial condition g if it satisfies the relation (6.10) where $(b(t);\, t \geq 0)$ is the unique locally integrable solution to (6.11) displayed in the previous lemma. ○

6.2.3 A forward Feynman-Kac formula

Define

$$Z(t) := \sum_x \mathbb{1}(\sigma_x \in (0, t]), \quad B(t) := \mathbb{E}\Big(Z(t)\Big)$$

where $Z(t)$ is interpreted as the number of infections between 0 and t. Recall that $R_0 := \int_0^\infty \beta(u)\,\mathrm{d}u < \infty$ guarantees that $B(t) < \infty$ for all $t \geq 0$. Finally, B is non-decreasing and we denote by $\mathrm{d}B$ the Stieljes measure associated to B.

Lemma 6.7. *There exists a locally integrable function* $(b(t);\, t \geq 0)$ *such that*

$$\mathrm{d}B(t) = b(t)\,\mathrm{d}t.$$

Further, b coincides with the unique locally integrable solution of the delayed equation (6.11).

Proof. The fact that $\mathrm{d}B$ has a density easily follows from the fact that τ has a density. The fact that $B(t) < \infty$ ensures that b is locally integrable.

Define $\bar{\mathcal{P}}_x$ the infection measure obtained from $\mathcal{P}_x$ after random thinning by the function $(c(t);\, t \geq 0)$. Namely, conditional on σ_x and the atoms $a_1 < a_2 < \cdots$ of $\mathcal{P}_x$, we remove independently each of the atoms with respective probabilities $1 - c(\sigma_x + a_1), 1 - c(\sigma_x + a_2) \ldots$, whereas the other atoms remain unchanged.

Fix $t > 0$. Let $k \leq n \in \mathbb{N}$. Define $\mathbb{T}^{k,n}(\mathcal{P}_x)$ as the measure obtained from $\mathcal{P}_x$ as follows. Conditional on the atoms $a_1 < a_2 < \cdots$ of $\mathcal{P}_x$, we remove independently each of the atoms with respective probabilities

$$1 - \max_{z \in (t\frac{k}{n}, t\frac{k+1}{n}]} c(z + a_1), 1 - \max_{z \in (t\frac{k}{n}, t\frac{k+1}{n}]} c(z + a_2) \cdots$$

and leave other atoms unchanged. Note that the thinning procedure is now independent of the starting time σ_x. Further, if $\sigma_x \in (t\frac{k}{n}, t\frac{k+1}{n}]$, the point measure $\mathbb{T}^{k,n}(\mathcal{P}_x)$ dominates $\bar{\mathcal{P}}_x$.

We decompose the births on $(0, t]$ into two parts: individuals stemming from the root $\emptyset$ and a second part from subsequent births. Using the fact that for every individual x, the (un-suppressed) random measure $\mathcal{P}_x$ is independent of its birth time σ_x (see second equality below), and setting $M(t) := \int_0^t \int_0^\infty g(a)\beta(a+u)c(u)\,\mathrm{d}a\,\mathrm{d}u$, we get

$$\begin{aligned}
B(t) &= \sum_{k=0}^{n-1}\sum_x \mathbb{E}\left(\mathbb{1}\left(\sigma_x \in (t\frac{k}{n}, t\frac{k+1}{n}]\right)\int_{[0,t-\sigma_x]} \bar{\mathcal{P}}_x(\mathrm{d}a)\right) \;+\; M(t) \\
&\leq \sum_{k=0}^{n-1}\sum_x \mathbb{E}\left(\mathbb{1}\left(\sigma_x \in (t\frac{k}{n}, t\frac{k+1}{n}]\right)\int_{[0,t-t\frac{k}{n}]} \mathrm{T}^{k,n}(\mathcal{P}_x)(\mathrm{d}a)\right) \;+\; M(t) \\
&= \sum_{k=0}^{n-1}\sum_x \mathbb{E}\left(\mathbb{1}\left(\sigma_x \in (t\frac{k}{n}, t\frac{k+1}{n}]\right)\right)\mathbb{E}\left(\int_{[0,t-t\frac{k}{n}]} \mathrm{T}^{k,n}(\mathcal{P})(\mathrm{d}a)\right) \;+\; M(t) \\
&= \sum_{k=0}^{n-1} \mathbb{E}\left(\sum_x \mathbb{1}\left(\sigma_x \in (t\frac{k}{n}, t\frac{k+1}{n}]\right)\right)\; \mathbb{E}\left(\int_{[0,t-t\frac{k}{n}]} \mathrm{T}^{k,n}(\mathcal{P})(\mathrm{d}a)\right) \;+\; M(t) \\
&= \sum_{k=0}^{n-1} \left(B(t\frac{k+1}{n}) - B(t\frac{k}{n})\right)\, \mathbb{E}\left(\int_{[0,t-t\frac{k}{n}]} \mathrm{T}^{k,n}(\mathcal{P})(\mathrm{d}a)\right) + \; M(t) \\
&= \sum_{k=0}^{n-1} \left(B(t\frac{k+1}{n}) - B(t\frac{k}{n})\right) \int_{[0,t-t\frac{k}{n}]} c^{k,n}(u)\beta(u)\,\mathrm{d}u \;+\; M(t).
\end{aligned}$$

with $c^{k,n}(y) = \max_{v\in(t\frac{k}{n},t\frac{k+1}{n}]} c(y+v)$. In particular, if $tk/n \to x$, and $x+y$ is a continuity point of c, we have $c^{k,n}(y) \to c(x+y)$. We will pass to the limit $n \to \infty$ in the latter inequality. Recall that c is bounded (and valued in $[0,1]$) and that c is right-continuous. The first term on the RHS can be written under the form

$$\sum_{k=0}^{n-1} \left(B(t\frac{k+1}{n}) - B(t\frac{k}{n})\right) \int_0^{t-[x]_n} c^{k,n}(u)\beta(u)\,\mathrm{d}u = \int_0^t f^n(x)\,\mathrm{d}B(x),$$

where

$$f^{(n)}(x) = \int_0^{t-[x]_n} \sup_{v\in([x]_n,\ [x_n]+\frac{t}{n}]} c(v+u)\beta(u)\,\mathrm{d}u \quad \text{and} \quad [x]_n = \frac{t}{n}\lfloor nx/t \rfloor.$$

We will now apply twice the bounded convergence theorem. On the one hand, for a fixed value of x, as $n \to \infty$

$$\mathbb{1}_{[0,t-[x]_n]}(u) \sup_{v\in([x]_n,[x]_n+\frac{t}{n}]} c(v+u)\beta(u) \to \mathbb{1}_{[0,t-x]}(u)c(x+u)\beta(u) \quad \text{Lebesgue a.e.}$$

Further, the latter term (i.e., the integrand in the integral defining f^n) is uniformly bounded by β and $\int_0^\infty \beta(u)du < \infty$. A first application of the bounded convergence theorem implies that for every x, as $n \to \infty$

$$f^n(x) \to \int_0^{t-x} c(x+u)\beta(u)\,\mathrm{d}u.$$

On the other hand, the uniform bound, $f^n(x) \leq R_0 = \int_0^\infty \beta(u)\,\mathrm{d}u$ for all x, n, allows us to again apply the bounded convergence theorem, so we get

$$B(t) \leq \int_0^t b(x)\,\mathrm{d}x \int_{[0,t-x]} c(x+u)\beta(u)\,\mathrm{d}u + \int_0^t \int_0^\infty g(a)\beta(a+u)c(u)\,\mathrm{d}a\,\mathrm{d}u.$$

By an analog argument, one can establish the same lower bound and strengthen the latter inequality into an equality. After a simple change of variable and interchanging the order of integration, this yields

$$B(t) = \int_0^t c(v) \int_0^v \beta(v-x)b(x)\,\mathrm{d}x\,\mathrm{d}v + \int_0^t \int_0^\infty g(a)\beta(a+u)c(u)\,\mathrm{d}a\,\mathrm{d}u.$$

Finally, differentiating with respect to t yields the desired result. □

Corollary 6.8 (Forward Feynman-Kac formula). *For every $t \geq 0$, define*

$$\bar{\mu}_t(\mathrm{d}a \times \{i\}) := n(t,a) \times \mathbb{P}(X(a) = i)\,\mathrm{d}a,$$

where n is the unique $\mathcal{L}^{1,loc}$ weak solution to the McKendrick-von Foerster PDE with initial condition g. Then

$$\bar{\mu}_t(\mathrm{d}a \times \{i\}) = \mathbb{E}\bigg(\sum_x 1_{\sigma_x < t} \delta_{(t-\sigma_x, X_x(t-\sigma_x))}(\mathrm{d}a \times \{i\}) \bigg) \tag{6.12}$$

where the expected value is taken with respect to a CMJ process starting with one individual with infection and life-process distributed according to the shifted law $\tilde{\mathcal{L}}_g$.

Proof. Define

$$\bar{\mu}'_t(\mathrm{d}a \times \{i\}) := \mathbb{E}\bigg(\sum_x 1_{\sigma_x < t} \delta_{(t-\sigma_x, X_x(t-\sigma_x))}(\mathrm{d}a \times \{i\}) \bigg)$$

We need to check that $\bar{\mu}'_t = \bar{\mu}_t$ on the space of finite measures. Let F be a measurable, non-negative, bounded, continuous function on $\mathbb{R}\times\mathbb{R}\times\mathcal{S}$ with compact support on $\mathbb{R}_+ \times \mathbb{R}_+ \times \mathcal{S}$. As in the previous lemma, we have

$$\begin{aligned}\int F(s,a,i)\bar{}\mu'_s(\mathrm{d}a,\mathrm{d}i)\,\mathrm{d}s &= \sum_{x\neq\emptyset} \mathbb{E}\bigg(\int F\left(s, s-\sigma_x, X_x(s-\sigma_x)\right) \mathbb{1}(\sigma_x < s)\,\mathrm{d}s \bigg) \\ &\quad + \int_0^\infty \int_0^\infty \mathbb{E}\left(F(t, t+a, X(t+a))\, g(a)\,\mathrm{d}a\,\mathrm{d}t.\right.\end{aligned}$$

Let (I) be the first term on the RHS. For every $n \in \mathbb{N}^*$

$$\begin{aligned}(I) &= \sum_{k\geq 0}\sum_x \mathbb{E}\bigg(\int F\left(s, s-\sigma_x, X_x(s-\sigma_x)\right) \mathbb{1}(\sigma_x > s, \sigma_x \in (\frac{k}{n}, \frac{k+1}{n}])\,\mathrm{d}s \bigg) \\ &\leq \sum_{k\geq 0}\sum_x \mathbb{E}\bigg(\int_0^s \max_{u\in(\frac{k}{n},\frac{k+1}{n}]} F\left(s, s-u, X_x(s-u)\right) \mathbb{1}(\sigma_x \in (\frac{k}{n}, \frac{k+1}{n}])\,\mathrm{d}s \bigg) \\ &= \sum_{k\geq 0}\sum_x \int_0^s \mathbb{E}\bigg(\max_{u\in(\frac{k}{n},\frac{k+1}{n}]} F\left(s, s-u, X(s-u)\right) \bigg)\, \mathbb{P}\bigg(\sigma_x \in (\frac{k}{n}, \frac{k+1}{n}] \bigg)\,\mathrm{d}s \\ &= \sum_{k\geq 0} \int_0^s \mathbb{E}\bigg(\max_{u\in(\frac{k}{n},\frac{k+1}{n}]} F\left(s, s-u, X(s-u)\right) \bigg) \bigg(B(\frac{k+1}{n}) - B(\frac{k}{n}) \bigg)\,\mathrm{d}s\end{aligned}$$

By reasoning along the same lines as in Lemma 6.7 (i.e., applying the bounded convergence theorem several times) and using the almost sure continuity at every fixed time of the process X, one can show that the RHS converges to

$$\int_{s=0}^{\infty}\int_{u=0}^{s}\mathbb{E}\Big(F\left(s,s-u,X(s-u)\right)\Big)b(u)\,\mathrm{d}u\,\mathrm{d}s$$

as $n\to\infty$ and thus

$$\begin{aligned}\int F(s,a,i)\bar{\mu}'_s(\mathrm{d}a,\mathrm{d}i)\,\mathrm{d}s \leq & \int_{s=0}^{\infty}\int_{u=0}^{s}\mathbb{E}\Big(F\left(s,s-u,X(s-u)\right)\Big)b(u)\,\mathrm{d}u\,\mathrm{d}s\\ &+\int_0^{\infty}\int_0^{\infty}\mathbb{E}\left(F(t,t+a,X(t+a))\,g(a)\,\mathrm{d}t.\right.\end{aligned}$$

By a similar argument, the inequality can be strengthened into an equality. Moreover we have

$$\begin{aligned}\int F(s,a,i)\,\bar{\mu}_t(\mathrm{d}a,\mathrm{d}i)\,\mathrm{d}s &= \int F(s,a,i)n(t,a)\mathbb{P}(X(a)\in\mathrm{d}i)\mathrm{d}a\,\mathrm{d}s\\ &=\int_0^a\int F(s,a,i)b(s-a)\mathbb{P}(X(a)\in\mathrm{d}i)\,\mathrm{d}a\,\mathrm{d}s\\ &\qquad+\int_a^{\infty}\int F(s,a,i)g(a-s)\mathbb{P}(X(a)\in\mathrm{d}i)\,\mathrm{d}a\,\mathrm{d}s\end{aligned}$$

so that

$$\int F(s,a,i)\,\bar{\mu}_t(\mathrm{d}a,\mathrm{d}i)=\int F(s,a,i)\,\bar{\mu}'_t(\mathrm{d}a,\mathrm{d}i).$$

Now take $F(s,a,i)=h(s)f(a,i)$ with h measurable, bounded, compact support on $\mathbb{R}_+$ and f bounded continuous. We get

$$\int h(s)\langle\bar{\mu}_s,f\rangle ds=\int h(s)\langle\bar{\mu}'_s,f\rangle\,\mathrm{d}s.$$

On the other hand it is easy to check that the two functions $s\to\langle\bar{\mu}_s,f\rangle$ and $s\to\langle\bar{\mu}'_s,f\rangle$ are both continuous. As a consequence, we have $\langle\bar{\mu}_s,f\rangle=\langle\bar{\mu}'_s,f\rangle$ for every test function f, concluding the proof. □

6.2.4 Dual CMJ process and backward Feynman-Kac formula

We end this section by making a connection between a dual process – interpreted as an *ancestral process* – and the (PDE) method of characteristics. In addition, this approach provides a probabilistic proof of uniqueness for the PDE.

Let $\mathcal{M}$ be any random point measure with intensity measure $\tau(\mathrm{d}u)$. Fix $a,T>0$. We now construct a dual process using the measure $\mathcal{M}$, which can be seen as a generalized Bellman-Harris branching process (individuals have a finite lifetimes, births only occur upon death). Let us first describe the process with no suppression (i.e., $c=1$).

- Start with a single particle at time $t = 0$. Assume that the residual lifetime of this original particle is a, so that this particle dies out at time a.

- As in a Bellman-Harris process, the number of offspring of an individual and their lifetime durations are independent of the parent's characteristics.

- Upon death, each individual x is endowed with an independent copy $\mathcal{M}_x$ of $\mathcal{M}$: the number of offspring of x is given by the number of atoms of $\mathcal{M}_x$ and their lifetime durations are given by the positions of the atoms in $\mathcal{M}_x$.

The dual process with suppression $c \neq 1$ can be coupled with the case $c = 1$. Given a realization of the process, if a branching occurs at time t, the children are killed independently with probability $c(T - t)$. (Note that as in the original CMJ process, suppression translates into trimming the dual tree.)

Remark 6.9. We note that there are as many dual processes as there are point processes with intensity measure τ. Here are a few natural choices:

1. $\mathcal{M} = \mathcal{P}$.

2. Let $\mathcal{M}$ be a Poisson Point Process with intensity measure $\tau(\mathrm{d}u)$. In this particular case, the dual process is a Bellman-Harris branching process (i.e., the offspring lifetime durations are independent conditional on offspring number). We note that $\tau(\mathrm{d}u)$ appears naturally when considering the ancestral spine of a critical CMJ, see e.g. [194]. The measure τ can be obtained by size-biasing $\mathcal{P}$ (i.e. biasing by the total mass of $\mathcal{P}$) and then recording the age of the individual at a uniformly chosen birth event. ○

Let $(Y_t;\, t \geq 0)$ be the stochastic process valued in $\cup_{n\in\mathbb{N}}\mathbb{R}_+^n$ recording the residual life-times at time t listed in increasing order, i.e. if $Y_t = (Y_t^{(1)}, \cdots, Y_t^{(n)})$ there are n particles alive at time t and $Y_t^{(k)}$ is the residual life-time of the k^{th}-individual with $Y_t^{(1)} < \cdots < Y_t^{(n)}$. (We assumed that τ has a density so that the residual lifetimes are distinct a.s.). In particular, the particle labelled 1 at any given time t will be the first to expire, and at death time $t + Y_t^{(1)}$ a random number of children is produced. We let $\dim(Y_T)$ denote the number of particules alive at time t, i.e., the dimension of the vector Y_t.

Finally, we will say that n is a right-continuous version to the McKendrick-von Foerster equation, if n is a $\mathcal{L}^{1,loc}$ weak solution and for every $T, x \geq 0$ the function $s \to n(T - s, x - s)$ is right-continuous on $[0, x]$.

Proposition 6.10 (Backward Feynman-Kac formula)**.** *Assume that the suppression function $(c(t);\, t \geq 0)$ is right-continuous. Then there is a unique right-continuous solution to the McKendrick-von Foerster equation, and for every $a \geq 0$, $T \geq 0$*

$$n(T, a) = \widehat{\mathbb{E}}_a\left(\sum_{i \leq \dim(Y_T)} g(Y_T^{(i)})\right) \tag{6.13}$$

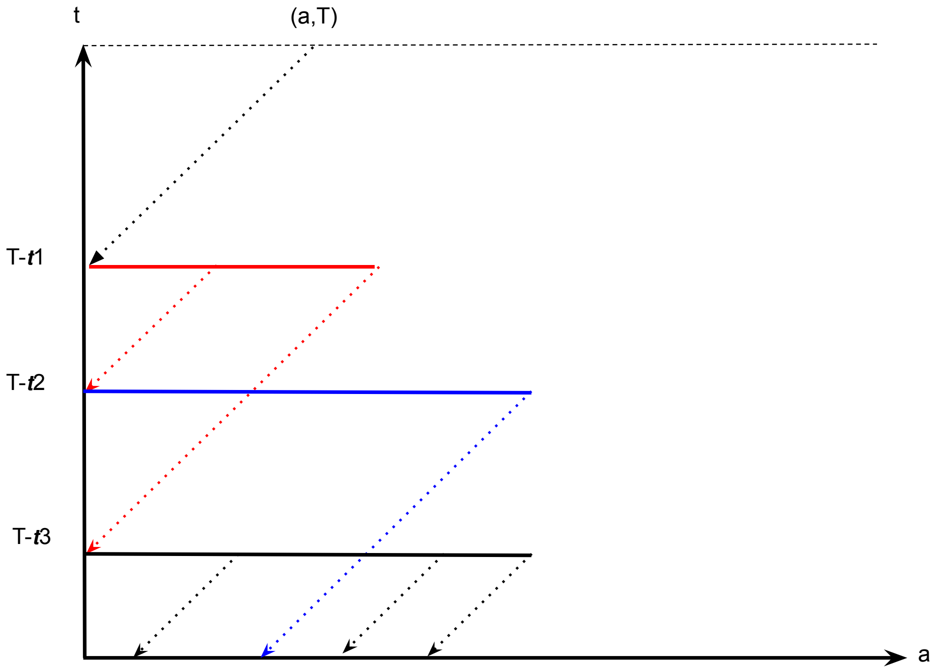


Figure 6.2: Graphical representation of the process $(Z_s;\ s \geq 0)$. We start with a single individual with residual lifetime a. In this picture, time flows downwards for the branching process. The residual lifetime of the initial individual decreases linearly until reaching 0 (this corresponds to time $T - t_1$ in our representation). At this time, the particle dies and produces 2 red particules. Residual lifetimes travel along the characteristics of the McKendrick-von Foerster PDE until reaching the spatial boundary condition $\{a = 0\}$ where a new branching occurs.

where $\widehat{\mathbb{E}}_a$ *is the distribution of the process* $(Y_t;\ t \geq 0)$ *starting with an individual with residual lifetime* a*.*

Proof. We first assume that there exists a right-continuous solution to the PDE. Let $t_1 < \cdots < t_k < \cdots$ be the successive branching times of the dual branching process. Since τ has a density, there is a single branching particle at the successive branching times $t_1, \cdots$. Define the càdlàg process

$$Z_s := \sum_{i \leq \dim(Y_s)} n(T - s, Y_s^{(i)})$$

See also Figure 6.2 for a pictorial representation of the process. It is plain from the definition that n is preserved along the characteristics of the PDE, i.e., that for every x the function $s \to n(T - s, x - s)$ remains constant on $[0, x)$. As a consequence, $(Z_s, s \geq 0)$ remains constant on every interval $[t_n, t_{n+1})$, with the convention $t_0 = 0$. Define $z_n := Z_{t_n}$ the value of the process $(Z_t;\ t \geq 0)$ at the n-th branching time. Let $(\mathcal{F}_n;\ n \in \mathbb{N})$ be the filtration induced by the process

$(z_n;\, n \in \mathbb{N})$. For every $n > 1$

$$\begin{aligned}\hat{\mathbb{E}}_a\left(z_n \mid \mathcal{F}_{n-1}\right) &= \sum_{2\le i\le \dim(z_n)} n(T-t_n, Y_{t_n}^{(i)}) + c(T-t_n)\int_0^\infty n(T-t_n,a)\,\tau(\mathrm{d}a) \\ &= \sum_{2\le i\le \dim(z_n)} n(T-t_n, Y_{t_n}^{(i)}) + n(T-t_n,0) = z_{n-1},\end{aligned}$$

where the second equality follows from the spatial boundary of the McKendrick-von Foerster equation. As already mentioned, the process $(Z_s;\, s\geq 0)$ is constant between two branching times. As a consequence, $(Z_s;\, s \geq 0)$ is a martingale (w.r.t. its natural filtration) so for every $s \geq 0$,

$$n(T,a) = \hat{\mathbb{E}}_a\left(\sum_{i\le \dim(Y_s)} n(T-s, Y_s^{(i)})\right).$$

Relation (6.13) follows by taking $s = T$ in the latter expression.

The proof ends by checking that when c is right-continuous, the RHS of (6.13) indeed is a right-continuous solution to the PDE. This elementary step is left to the interested reader. □

6.3 Two laws of large numbers

Theorem 6.11 (N individuals)**.** *Start with N individuals at time 0 with independent infection and life-processes distributed according to the initial shifted law $\tilde{\mathcal{L}}_g$. Define the empirical random measure for ages and types at time t*

$$\mu_t^N(\mathrm{d}a\times\{i\}) \coloneqq \sum_x 1_{\sigma_x<t}\delta_{(t-\sigma_x, X_x(t-\sigma_x))}(\mathrm{d}a\times\{i\}). \tag{6.14}$$

As in Corollary 6.8, let

$$\bar{\mu}_t(\mathrm{d}a\times\{i\}) = n(t,a)\times \mathbb{P}(X(a)=i)\,\mathrm{d}a$$

where n is the unique $\mathcal{L}^{1,\mathrm{loc}}$ weak solution to the McKendrick-von Foerster PDE with initial condition g. For every $t > 0$,

$$\frac{1}{N}\mu_t^N \xrightarrow[N\to\infty]{} \bar{\mu}_t \quad a.s.$$

where the convergence is meant in the weak topology.

Proof. We have

$$\mu_t^N(\mathrm{d}a\times\{i\}) = \frac{1}{N}\sum_{i=1}^N \mu_t^{1,(i)}(\mathrm{d}a\times\{i\}) \tag{6.15}$$

where $\{\mu_t^{1,(i)}\}$ are independent copies of μ_t^1. Let f be an arbitrary measurable and bounded function on $\mathbb{R}_+\times\mathcal{S}$. The L.L.N. combined with Corollary 6.8 implies that

$$\langle \frac{1}{N}\mu_t^N, f\rangle \to \langle \bar{\mu}_t, f\rangle \quad \text{a.s.}$$

which ends the proof. □

In the following, we are motivated by modeling a situation where the infection is separated into two distinct phases. We start from a single individual.

(Phase 1) We let the epidemic develop until a certain random time t_K. For instance, t_K could be the time at which the number of recorded deaths exceeds a large threshold K. We assume no suppression before t_K.

(Phase 2) We let the suppression function vary after time t_K, e.g., due to mitigation measures and/or behavioral changes (i.e., lockdown phase).

We will see in Theorem 6.13 below that the dynamics after time t_K converge to the same solution as in Theorem 6.11 but with an exponential initial age density ($g(x) = \alpha \exp(-\alpha x)$) and a (large) random number of initial infected individuals.

Let us now provide a more formal set-up. First ignore suppression and consider a plain CMJ process starting from one individual with shifted law $\tilde{\mathcal{L}}_g$. *Let us now assume that the Malthusian parameter α of the CMJ process is strictly positive (supercritical assumption).* We assume that g is chosen in such a way that there is a positive probability of non-extinction (to avoid trivialities).

Let $\mathcal{F}_t = \sigma\left(\{(\mathcal{P}_x, X_x) : x \in \bigcup_n \mathbb{N}^n,\ \sigma_x < t\}\right)$ be the σ-field generated by the observation of the infection and life-cycle processes of the individuals born before time t (including the root $\emptyset$). Let $\{t_K\}$ be a sequence of stopping times (w.r.t. $(\mathcal{F}_t;\, t \geq 0)$) with $t_K \to \infty$ a.s. on the non-extinction event.

Now we assume that the suppression function $c^K \equiv c$ of the CMJ depends on K and that $c^K(t) := C(t - t_K)$, where $(C(t);\, t \in \mathbb{R})$ is a piecewise continuous function in $[0, 1]$ such that $C(t) = 1$ for all $t \leq 0$. Finally, $\mu_t^{t_K}$ is again the empirical measure of ages and types (as defined in (6.14)) w.r.t. the suppression function c^K.

Example 6.12. Take

$$t_K = \inf\{t > 0 : \#\{x \in \cup_n \mathbb{N}^n : \sigma_x < t, X_x(t - \sigma_x) = D\} \geq K\},$$

i.e., t_K is the first time that the accumulated number of deaths reaches level K. Further take $C(t) = 1$ if $t \leq 0$ and $C(t) = r < 1$ if $t > 0$. This corresponds to a lockdown strategy where transmission is reduced by a factor r upon reaching K deaths. ○

Theorem 6.13 (One individual)**.** *Conditional on non-extinction*

- *There exists a r.v. W_∞ such that $W_\infty > 0$ a.s. and*

$$\sum_x \mathbb{1}(0 < \sigma_x < t_K) \exp(-\alpha t_K) \xrightarrow[t_K \to \infty]{} W_\infty,$$

almost surely and in L^1.

- *Fix $t \geq 0$. We have*

$$\exp(-\alpha t_K) \mu_{t_K + t}^{t_K} \underset{t_K \to \infty}{\Longrightarrow} W_\infty\ \bar{\mu}_t$$

in probability, where the convergence is meant in the weak topology, and

$$\bar{\mu}_t(\mathrm{d}a \times \{i\}) = n(t,a) \times \mathbb{P}(X(a) = i)\,\mathrm{d}a$$

with n the unique $\mathcal{L}^{1,\mathrm{loc}}$ weak solution to the McKendrick-von Foerster PDE with initial condition $g(a) = \alpha \exp(-\alpha a)$ and with suppression function given by $(C(t);\, t \geq 0)$.

Proof. We recall some basic facts about the method of random characteristics. We consider a plain CMJ with no death (no suppression function, no types). Every individual is characterized by an independent pair of random variables $(\mathcal{P}, \chi)$. As before, $\mathcal{P}$ is the infection measure recording the times of secondary infections. Now χ is a general stochastic process indexed by the age of the individual called a random characteristic with the convention $\chi(-a) = 0$ for $a \geq 0$. In great generality, $\mathcal{P}$ and χ may exhibit a non-trivial correlation. Define

$$Z^\chi(T) = \sum_x \chi_x(T - \sigma_x)$$

the branching process counted by the random characteristic χ at time T. We now recall one of the main results of Jagers and Nerman [116] (see also Theorem 5, Appendix A in [205]). On top of all the assumptions above, we make the two following extra assumptions.

(a) There exists $\alpha' < \alpha$ such that

$$\mathbb{E}\left(\sup_{T \geq 0} \exp(-\alpha' T)\chi(T)\right) < \infty. \tag{6.16}$$

(b) The map $T \to \mathbb{E}(\chi(T))$ is continuous a.e. With respect to the Lebesgue measure.

Then there exists a positive r.v. W_∞ (independent of the choice of χ) such that conditional on non-extinction

$$Z^\chi(T)\exp(-\alpha T) \longrightarrow W_\infty \int_0^\infty \alpha \exp(-\alpha t)\mathbb{E}(\chi(t))\,\mathrm{d}t$$

almost surely and in $L^1(\mathrm{d}x)$ as $T \to \infty$. (Note that the L_1 convergence holds thanks to the $x \log x$ condition (6.9).)

To illustrate the method, we recall that if we take $\chi(T) = \mathbb{1}_{\mathbb{R}_+}(T)$ then $Z^\chi(T)$ coincides with $B(T)$, the total number of births before time T. For this particular choice of (deterministic) characteristic, the two properties above are immediately satisfied (recall that $\alpha > 0$), so that conditional on non-extinction

$$\sum_x \mathbb{1}(0 < \sigma_x < u)\exp(-\alpha u) \longrightarrow W_\infty \tag{6.17}$$

almost surely and in $L^1(\mathrm{d}x)$ as $u \to \infty$. The a.s. convergence ensures that the first item of our theorem is satisfied.

Next, the second part of the theorem requires a choice of characteristic, called "general characteristic", that depends on the descendance of each extant individual at time t_K. Because we need to prove an a.s. convergence result, whereas limit theorems on branching processes counted with general characteristics only hold in distribution, we have to design by hand an individual characteristic that has the same distribution as the requested general characteristic.

In order to define our next random characteristics, we start with some definition. Let $(\mathcal{P}^*, X^*)$ be a pair of infection and life-cycle processes (that may or may not be identical to $(\mathcal{P}, X)$ in distribution). One can construct a collection $(\mathcal{Z}^{(\Delta)}; \Delta \geq 0)$ of $(\mathcal{P}^*, X^*)$-CMJ processes starting at respective times Δ and with a suppression mechanism C, i.e.,

- $\mathcal{Z}^{(\Delta)}$ is a $(\mathcal{P}^*, X^*)$-CMJ process *starting at time* Δ with one progenitor.
- The suppression function applied to infection events is C.
- The previous rule applies at time $t = \Delta$, that is, $\mathcal{Z}^{(\Delta)}$ is identically empty with probability $1 - C(\Delta)$.

For any $t \geq \Delta$, we let $\nu^{(\Delta)}(t)$ denote the empirical measure for ages and types of $\mathcal{Z}^{(\Delta)}(t)$. Finally, we define $\{(\nu_i^{(\Delta)}; \Delta \geq 0)\}_{i \in \mathbb{N}^*}$ as the collection made of independent copies of the collection $(\nu^{(\Delta)}; \Delta \geq 0)$.

We are now ready to construct our random characteristics by enlarging the initial CMJ process in the following way. Fix $t \in \mathbb{R}_+$ and f a bounded non-negative continuous function on $\mathbb{R} \times \mathcal{S}$ with compact support in $\mathbb{R}_+ \times \mathcal{S}$. Consider a typical individual $\emptyset$, with infection and life processes $(\mathcal{P}, X)$. Denote by (r_i) the atoms of $\mathcal{P}$ listed in increasing order. For any $T \geq 0$, define the individual random characteristic $\chi^{(t,f)}(T)$, by

$$\chi^{(t,f)}(T) := f(T + t, X(T + t)) + \sum_{i: r_i \in \mathcal{P} \cap [T, T+t]} \langle \nu_i^{(r_i - T)}(t), f \rangle,$$

See Figure 6.3 for a intuitive constructing of the random characteristics.

From now on, we assume that $(\mathcal{P}^*, X^*)$ is identical in law to $(\mathcal{P}, X)$, which implies the following two crucial facts.

(i) $\mathbb{E}(\int_0^\infty \chi^{(t,f)}(a) g(a) da) = \langle \bar{\mu}_t, f \rangle$ where $\bar{\mu}_t$ is defined as in Corollary 6.8 with initial condition g and suppression function C.

(ii) Let us now count our branching process by its random characteristic

$$Z^{\chi^{(t.f)}}(T) = \sum_x \chi_x^{(t,f)}(T - \sigma_x).$$

Since t_K is a stopping time with respect to the filtration $(\mathcal{F}_t; t \geq 0)$, the branching property implies that $Z^{\chi^{(t,f)}}(t_K)$ is identical in distribution to the process $\langle \mu_{t_K+t}^{t_K}, f \rangle$ where we remember that $\mu_t^{t_K}$ is the empirical measure w.r.t. the CMJ process with suppression function $c^K : t \mapsto C(t - t_K)$.

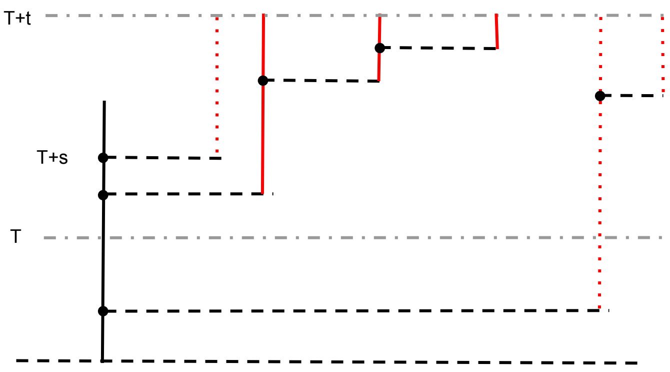


Figure 6.3: The individual x with characteristic $(\mathcal{P}_x, X_x)$ under consideration is represented by a black segment. We graft independent $(\mathcal{P}^*, X^*)$-CMJ processes with suppression function $t \mapsto C(t-T)$ to the atoms of $\mathcal{P}_x$ occurring at time $T+s$, independently with probability $C(s)$ if $s \geq 0$ and 0 if $s < 0$. We ignore all other atoms and their descendances (lower dotted red tree $s < 0$, upper dotted tree $s > 0$).

In order to apply the aforementioned result of Jagers and Nerman, we need to check that condition (a), (b) above are satisfied. Condition (b) easily follows from the fact that τ has a density. We now check the first condition. Consider a $(\mathcal{P}, X)$-CMJ process, assuming that the initial individual is un-shifted. For every s, let $v_s(\mathrm{d}a \times \{i\})$ the empirical measure of ages and types at time s. We assumed f to be non-negative and thus

$$\mathbb{E}\left(\sup_{s\in[0,t]} \langle v_s, f\rangle\right) \leq \|f\|_\infty B(t).$$

Let $a_1 < a_2 < \cdots$ be the atoms of $\mathcal{P}$ between T and $T+t$. The random characteristic under consideration is obtained by attaching independent (suppressed) CMJ processes between T and $T+t$ to the a_i's and by summing up the respective empirical measures at time $T+t-a_i$. By construction, $T \leq T+t-a_i \leq T+t$ so that

$$\begin{aligned}\mathbb{E}\left(\sup_{T\geq 0}\exp(-\alpha' T)\chi(T)\right) &\leq \mathbb{E}\left(\sup_{s\in[0,t]}\langle v_s, f\rangle\right)\mathbb{E}\left(\sup_{T\geq 0}\exp(-\alpha' T)\int_T^{T+t}\mathcal{P}(\mathrm{d}u)\right)\\ &= \|f\|_\infty B(t)\mathbb{E}\left(\sup_{T\geq 0}\exp(-\alpha' T)\int_T^{T+t}\mathcal{P}(\mathrm{d}u)\right)\\ &\leq R_0 B(t)\|f\|_\infty.\end{aligned}$$

This shows that property (a) is satisfied for any $0 \leq \alpha' < \alpha$. This shows that as $v \to \infty$, conditional on non-extinction

$$Z^{\chi^{(t,f)}}(v)\exp(-\alpha v) \to W_\infty \mathbb{E}\left(\int_0^\infty \alpha\exp(-\alpha u)\chi^{(t,f)}(u)\right) \quad \text{a.s.}$$

As already pointed out in (i),

$$\mathbb{E}\left(\int_0^\infty \alpha\exp(-\alpha u)\chi^{(t,f)}(u)\,\mathrm{d}u\right) = \langle\bar{\mu}_t, f\rangle$$

where $\bar{\mu}_t$ is defined as in Corollary 6.8 with initial condition $\alpha\exp(-\alpha t)$ and suppression function $(C(t);\, t \geq 0)$. Since the latter convergence is a.s. and $Z^{\chi^{(t,f)}}(t_K)$ is identical in law with $\langle\mu_{t_K+t}^{t_K}, f\rangle$ (see (ii) above), the result follows. □

6.4 Inference

6.4.1 Case study: the French COVID-19 epidemic

In this section, we illustrate how to use our framework to make inferences from macroscopic observables of the epidemic, e.g., incidence of positively tested patients, hospital or ICU (intensive care unit) admissions, deaths, etc. We show how to use those observables to extract the underlying age structure of the population, estimate model parameters and forecast the future of the epidemic.

We focused on a longitudinal case study in France. From March 18 2020, the French government has provided daily reports of the numbers of ICU and hospital admissions, of deaths, of discharged patients, and of occupied ICU and hospital beds. Moreover, several theoretical studies have already been conducted on the same dataset. This allowed us to fix the values of some crucial biological parameters that had already been estimated and to carry out a comparison with our method. We want to emphasize that the aim of this section is to provide a mathematical framework in which convergence results can be rigorously proved while remaining flexible enough for other applications. Our goal is not to provide new estimates of epidemiological parameters for France, as many robust estimates are already available. For instance we do not provide confidence intervals for our estimates, and neither do we conduct a sensibility analysis.

The remainder of the section is laid out as follows. In Section 6.4.2 we identify the mathematical quantities that impact the dynamics of the epidemic for large population sizes. Section 6.4.3 then presents the choice of distribution we made for these quantities and the parameters that need to be estimated. Finally, estimation of these parameters from the French incidence data is performed in Section 6.4.4. We start by fitting a simple model and then show how this model can be made more complex to account for more complex incidence data.

6.4.2 The model

As mentioned previously, the age structure of the population cannot be directly accessed. What is observed is a subset of the population with some characteristic of interest, for instance individuals that have been tested positively, deceased individuals or discharged patients. Recall that under the assumptions stated in Section 6.3, the number of individuals that are in a given state i at time t converges to

$$\int_0^\infty n(t,a)\mathbb{P}(X(a)=i)\,\mathrm{d}a,$$

where $(n(t,a))$ is the solution to the McKendrick-von Foerster equation. Note that the assumptions of Theorem 6.13 are in essence that the epidemic has been ongoing for a long enough time at the lockdown onset for the infected population to be large, which we assume to hold true for France as the number of infected individuals on March 16 2020 was on the order of tens of thousands of individuals.

The McKendrick-von Foerster equation is determined by two quantities: (i) the average infection measure τ defined Section 6.2.1 and (ii) an initial condition, which is of the form

$$\forall a \geq 0, \quad n(0,a) = W\alpha e^{-\alpha a}$$

for some initial number of infected individuals W and a parameter α which corresponds to the exponential growth rate of the epidemic before the lockdown onset.

Therefore, using Theorem 6.13 to obtain a theoretical prediction for some observables under consideration requires the knowledge of:

1. The intensity measure τ of secondary infections.
2. The initial number of infected W and the parameter α.
3. For each state i of interest the probability $\mathbb{P}(X(a)=i)$.

The next section exposes how we have parametrized these quantities.

6.4.3 Parametrization of the model

Average infection measure. Recall the definition of τ from Section 6.2.1. Let us further define

$$R_0 = \tau([0,\infty)), \quad \hat{\tau}(\mathrm{d}a) = \frac{\tau(\mathrm{d}a)}{\tau([0,\infty))},$$

so that τ can be expressed as

$$\tau(\mathrm{d}a) = R_0\,\hat{\tau}(\mathrm{d}a).$$

The total mass of τ, R_0, is the mean number of secondary infections induced by a single infected individual. Thus R_0 corresponds to the basic reproduction number of the epidemic, and we leave it as a parameter to infer. The epidemiological interpretation of the probability measure $\hat{\tau}$ is the following. Consider a large population of infected individuals. Then, as the size of that population goes to infinity, the distribution of the time between the infection of a uniformly sampled individual and the infection of its "parent" converges to $\hat{\tau}$. Therefore, $\hat{\tau}$ is the distribution of the so-called generation time of the epidemic, which has already been estimated in several previous studies. We used the estimation of [68], and assumed that $\hat{\tau}$ is a Weibull distribution, that is

$$\hat{\tau}(\mathrm{d}a) = \frac{k}{\lambda}\left(\frac{a}{\lambda}\right)^{k-1} e^{-(a/\lambda)^k}\,\mathrm{d}a, \tag{6.18}$$

where the values of the shape parameter k and scale parameter λ are recalled in Table 6.1.

Initial condition. The growth rate α is defined implicitly through equation (6.8). Moreover, we know that α corresponds to the exponential growth rate of *any* of the observables of the epidemic before the lockdown onset. We chose to estimate α from the cumulative number of deaths, which appeared to be more reliable than the number of positive tests as the number of tests that have been conducted in the early phase of the epidemic in France varied greatly. We simply estimated α using a linear regression on the logarithm of the number of deaths from March 7 to March 20, 2020. The estimated α as well as the corresponding R_0 pre-lockdown are shown in Table 6.1. The number of infected individuals at the time of lockdown was left as a parameter to infer.

Life-cycle. The last quantities that need to be defined are the one-dimensional marginals of the life-cycle process $(X(a);\, a \geq 0)$. From now on, we assume that the sequence of states visited by $(X(a);\, a \geq 0)$ is a Markov chain and that the sojourn times in each state is Gamma distributed.

More specifically, we suppose that we are given a global dispersion parameter $\gamma > 0$ and that for each state $i \in \mathcal{S}$, the time D_i spent in state i follows a Gamma distribution with expectation $t_i > 0$ and variance t_i/γ,

- $\mathbb{E}(D_i) = t_i$
- $\mathrm{Var}(D_i) = t_i/\gamma$

This assumption has the following consequence. Let $i_1, \dots, i_k$ be a possible sequence of states of $(X(a);\, a \geq 0)$, and denote by $T_{i_1}, \dots, T_{i_k}$ the respective entrance times in these states. Then conditional on $(X(a);\, a \geq 0)$ successively visiting the states $i_1, \dots, i_k$ in this order, we have

$$(T_{i_1}, \dots, T_{i_k}) \sim (0, Y_{t_1}, \dots, Y_{t_1 + \dots + t_{k-1}})$$

where $(Y_t)_{t \geq 0}$ is a Gamma process such that

$$Y_t \sim \frac{\gamma^{\gamma t}}{\Gamma(\gamma t)} a^{\gamma t - 1} e^{-\gamma a}\, \mathrm{d}a.$$

Another advantage of this parametrization is that the one-dimensional marginals of $(X(a);\, a \geq 0)$ can be computed efficiently, while only requiring one parameter for each state of interest, and a global dispersion parameter.

6.4.4 Fitting the model to incidence data

In this section, we fit six time series of the French epidemic: the number of ICU and hospital admissions, the number of deaths, the number of occupied ICU and hospital beds and the number of discharged patients. We provide two examples of

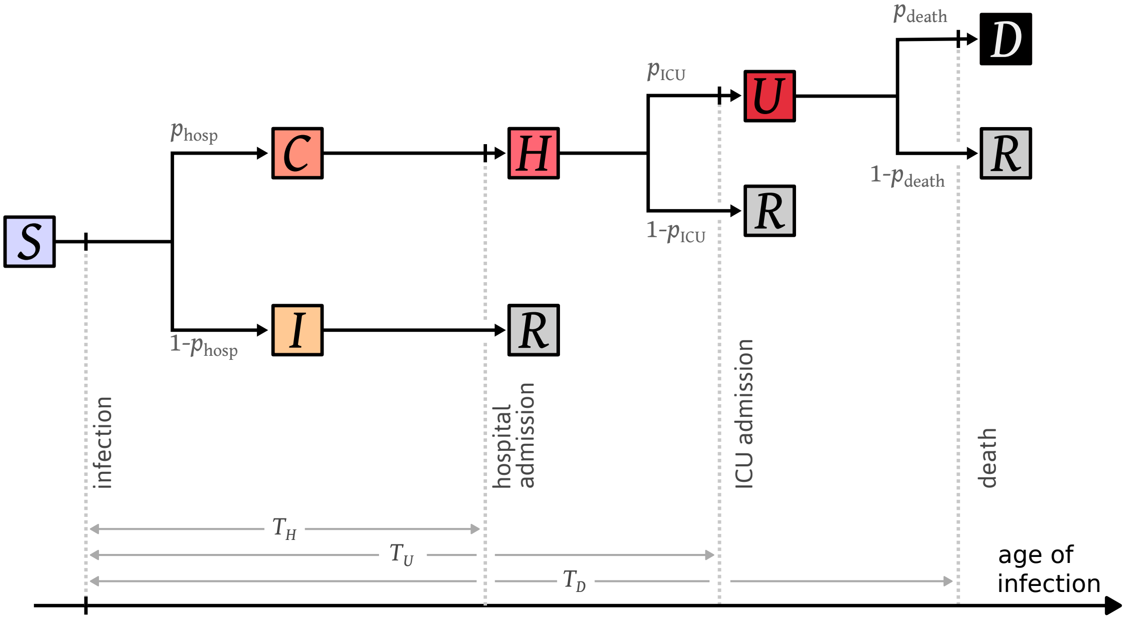


Figure 6.4: Illustration of the admission model.

Markov chains that we use to fit these data. The first Markov chain is only used to fit the first three curves, that we will call *incidence curves*, i.e., the daily number of ICU admissions, of hospital admissions and of deaths. With the parametrization of the previous section, the predictions of our model for these time series only involves the delay between infection and death, ICU, or hospital admissions. We do not need to estimate the time between hospitalization and discharge, which makes the model and the inference procedure easier. Then we show how this model can be made more complex to fit the last three curves, that we call *prevalence curves*. All parameter estimations were realized using the data from March 18 to May 11, 2020. This time frame corresponds to the lockdown period in France.

Fitting incidence data. In order to fit the incidence curves, we considered the simplest model that can account for these three time series. The model is illustrated in Figure 6.4.

Upon infection, with probability $1-p_{\text{hosp}}$, an individual develops a mild form of COVID-19 and is placed in state I, which encompasses all cases that do not require a hospitalization. With probability p_{hosp} the individual has a severe infection and is placed in state C. Individuals in state C are eventually hospitalized and moved to state H. Then, with probability p_{ICU} individuals in state H are admitted in ICU and move to state U. Otherwise they eventually recover and are discharged. Finally, individuals in state U die with probability p_{death}, or recover with probability $1-p_{\text{death}}$. In this model, only individuals in ICU may die.

As we are fitting the number of individuals that enter a state, and not the number of individuals that are currently in that state, we only need to track the times

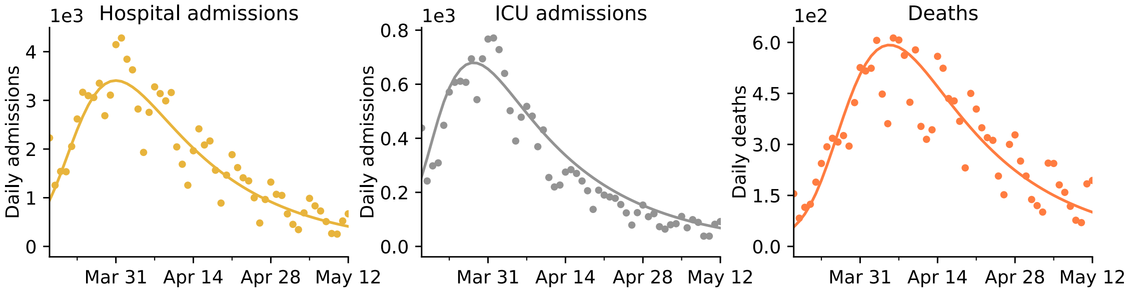


Figure 6.5: Best fit of the admission model. Solid line correspond to the number of hospital admissions, ICU admissions and deaths predicted by the admission model. The dots are the corresponding observed values.

T_H, T_U, and T_D elapsed between infection and hospital admission, ICU admission and death, respectively. We will refer to this model as the *admission model.*

Estimations for p_{hosp}, p_{ICU} and of death probability conditional on hospitalization (equal in our setting to $p_{\text{ICU}} \times p_{\text{death}}$) in France have already been conducted in [189]. We used these estimates and considered the values of p_{hosp}, p_{ICU} and p_{death} to be fixed. All other parameters were estimated using a maximum likelihood procedure which is described in Section 6.A. The parameter estimations are provided in Table 6.2, and the corresponding predicted values for the time series under consideration are displayed in Figure 6.5. Overall, our simple model seems to match the observed data. Note however that the model overestimates the number of ICU admissions in the second part of the lockdown. This is likely due to a temporal reduction in the ICU admission probability which has been reported in [189].

Our estimation of the basic reproduction number during the lockdown period is $R_0 = 0.745$. This suggests that lockdown has reduced the basic reproduction number by a factor 0.23 compared to the beginning of the epidemic. Moreover, we estimated that 9.85×10^5 infections have occurred in France before March 17th. Both these values are in line with previous estimates for France [202, 189].

We did not impose that $T_H < T_U$ in the inference procedure. Interestingly we found that the data are best explained by assuming that the mean of T_H is 14.4 days, whereas the mean of T_U is 11.4 days. This indicates that the delay between infection and hospital admission is shorter for individuals that end up in ICU, compared to the average time between infection and hospitalization. Therefore it would be more appropriate to allow individuals to have an admission to hospital delay that is different depending on whether they will end up in ICU or not, modeling the fact that they have a more severe form of the disease. We estimated the mean of T_D, the time between infection and death, to be 18.6 days. This estimate is lower than but consistent with previous estimates based on the study of individual-case data [219, 148, 214].

Fitting prevalence data. A first attempt to fit the prevalence curves could be to keep the admission model of Figure 6.4 and to estimate the time between hospital admission and discharge using the observed number of occupied ICU,

Notation	Description	Value	Source
α	Pre-lockdown exponential growth rate	0.315	E
R_{pre}	Basic reproduction number before lockdown	3.25	E
k	Shape parameter of the generation time	2.83	[68]
λ	Scale parameter of the generation time	5.67	[68]

Table 6.1: Parameter values common to both models. In the "Source" column, "E" indicates that the parameter has been estimated in the present work.

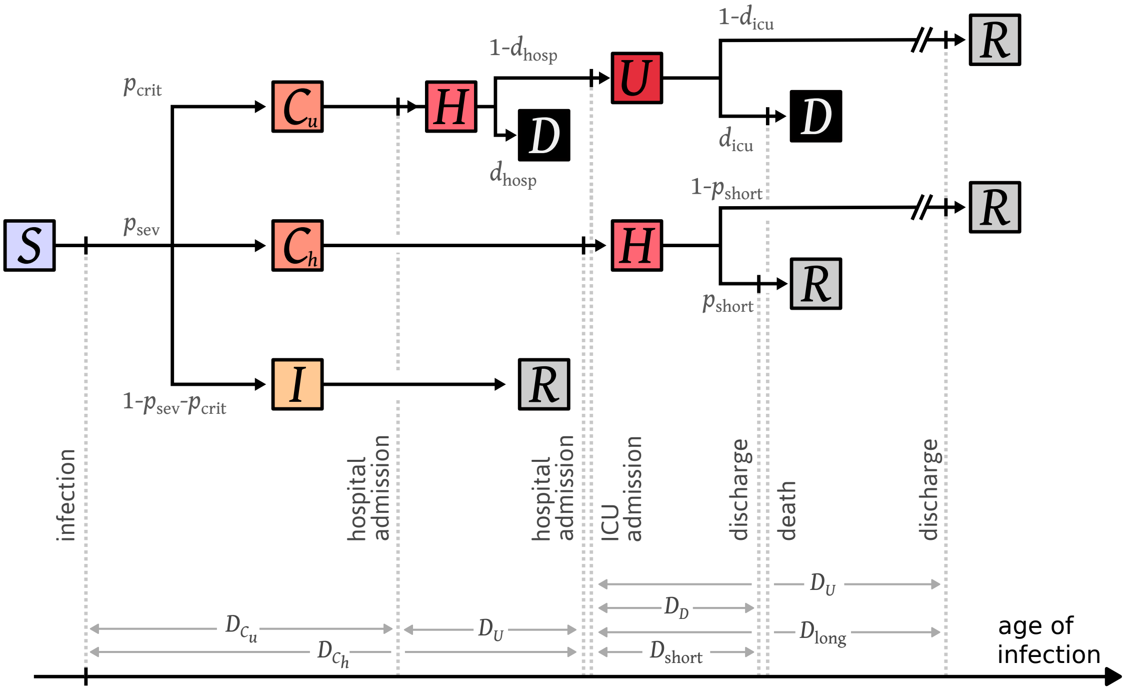


Figure 6.6: Illustration of the occupancy model

hospital beds, and discharged patients. However this only yields a poor fit of the data (see Section 6.B). We identified two main reasons for this discrepancy. First, we assumed that all individuals are admitted to ICU prior to death. Using the probability estimated in [189] then yields that the probability of dying conditional on being in ICU is 0.953. This value is unrealistically high, and we need to assume that a fraction of hospital deaths occur without going through the ICU. Second, under the admission model, the delay between hospital admission and discharge is almost unimodal. However, the observed number of occupied hospital beds rises fast but falls slowly. Such a shape cannot be easily accounted for by a unimodal distribution for the time spent in hospital.

Notation	Description	Value	Source
R_0	Basic reproduction number during lockdown	0.745	E
W	Total number of infections before March 17 2020	9.85×10^5	E
p_{hosp}	Probability of being hospitalized	0.036	[189]
p_{ICU}	Probability of entering ICU conditional on being at the hospital	0.19	[189]
$p_{\text{ICU}} \cdot p_{\text{death}}$	Death probability conditional on being hospitalized	0.181	[189]
T_H	Delay between infection and hospital admission	14.4 days	E
T_U	Delay between infection and ICU admission	11.4 days	E
T_D	Delay between infection and death	18.6 days	E
γ	Scale parameter common to all Gamma distributions	0.463	E

Table 6.2: Inferred parameter set for the admission model. The values indicated for the durations correspond to the means of the Gamma distributions. In the "Source" column, "E" indicates that the parameter has been estimated in the current work.

Taking into account the previous two points required us to make the model more complex. The resulting model, referred to as the *occupancy model*, is illustrated in Figure 6.6. We now consider that upon infection, individuals go to one of three states depending on the severity of their infection:

- The state C_u which gathers critical infections that lead to death or ICU admission. The probability of having a critical infection is denoted by p_{crit}.
- The state C_h which corresponds to severe infections that require a hospitalization but are not critical. Such infections occur with probability p_{sev}.
- The I state which consists of all mild infections that do not lead to a hospital admission, and occur with probability $1 - p_{\text{crit}} - p_{\text{sev}}$.

Individuals in state C_h are admitted to hospital after a duration D_{C_h}. Then, with probability p_{short} they are discharged after a duration D_{short}, while with probability $1 - p_{\text{short}}$ they are discharged after a duration D_{long}.

Critically infected individuals are admitted to hospital after a duration D_{C_u}. Upon arrival at hospital, they die immediately with probability d_{hosp}, or go to ICU

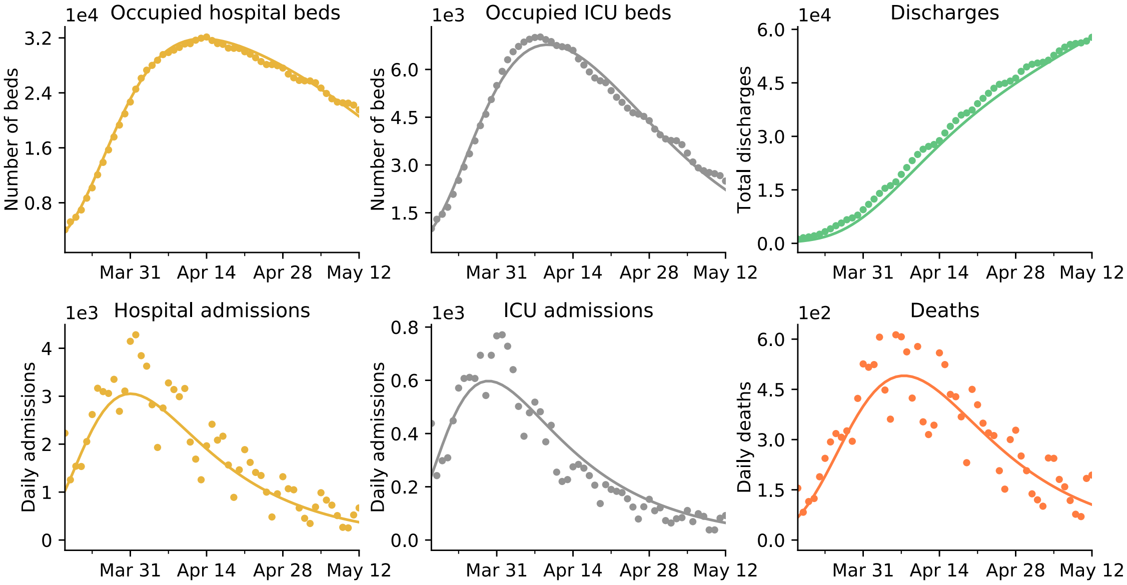


Figure 6.7: Best fit of the admission model. The solid lines correspond to the number of deaths, discharges, occupied ICU and hospital beds and ICU and hospital admissions predicted by the occupancy model. The dots are the corresponding observed values.

after a duration D_{H_u}. Individuals in ICU die with probability d_{ICU} after a delay D_D. Otherwise they are discharged after a stay of length D_U.

In our model, the probability of hospital admission is $p_{\text{crit}} + p_{\text{sev}}$, the probability of ICU admission is $p_{\text{crit}}(1 - d_{\text{hosp}})$ and that of death is $p_{\text{crit}}(d_{\text{hosp}} + (1 - d_{\text{hosp}})d_{\text{ICU}})$. We have fixed these three values to those estimated in [189], and we only had one remaining parameter out of 4 (p_{crit}, p_{sev}, d_{short}, d_{ICU}) to estimate from the data. We have fixed the time D_U to 1.5 days as estimated in [189]. All other parameters were estimated using the same likelihood method as previously, which is described in Section 6.A. The estimated parameter set is shown in Table 6.3, while Figure 6.7 shows the best-fitting model.

The estimated parameters provide a good fit of the six observed time series. Again, the model has a tendency to overestimate the ICU admissions in the second part of the lockdown, which has the same interpretation as before.

Under the occupancy model, we estimated that $R_0 = 0.734$, and $W = 9.52 \times 10^5$. These estimates are extremely close to those made with the admission model. The estimated mean time between infection and death or hospital, ICU admission are respectively 19.5 days, 13.7 days and 12.5 days. Again we see that these estimates in the more complex model are consistent with those of the simple model. The mean recovery time from hospital is 19.4 days for severe infections, and 28.2 days for critical infections. This yields an overall mean recovery time of 20.0 days. Finally, we estimated that the death probability conditional on being in ICU is 0.709. This yields that in our model a fraction 0.256 of all deaths occur shortly after hospital admission. This result is consistent with [189] that estimated that a fraction 0.15 of all deaths occurred within the first day after hospital admission. However, it has been reported in [192] that the death probability of ICU patients is 0.23. Our estimated value is thus unrealistically high. This indicates that there is a fraction of hospital deaths that occur without any ICU admission, and not quickly after hospital admission, that our model is not accounting for.

Our estimates, though they are not the key message of the present paper, can nevertheless draw attention to potential heterogeneities in the infected population. We estimated that the mean time between infection and ICU admission is shorter than that between infection and hospital admission. This suggests that the time between infection and severe symptom onset is shorter for critical infection, that lead to ICU admission, than for milder ones. Moreover, fitting the prevalence time series required to divide the hospital and death compartments in two subcompartments, indicating that the data are not well explained by a simple homogeneous model, as seen in Figure 6.8. Such heterogeneity could originate from underlying structuring variables, such as comorbidity or (real) age, that we are not accounting for. Many estimates of clinical features, such as the incubation period, are obtained from a pooled dataset that does not take heterogeneity in the population into account [5, 148, 144, 207, 22, 153, 46]. When estimating the total number of infected individuals using only a fraction of the detected cases, e.g., using the hospital admissions or deaths, it is interesting to keep in mind that the time periods estimated from pooled studies could be inaccurate for the fraction of infected individuals under consideration.

References for Chapter 6

[5] J. A. Backer, D. Klinkenberg, and J. Wallinga. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20-28 January 2020. *Eurosurveillance* **25** (2020).

[9] F. Ball. A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Advances in Applied Probability* **18** (1986), 289–310.

[10] A. D. Barbour. The duration of the closed stochastic epidemic. *Biometrika* **62** (1975), 477–482.

[17] R. Bauerfeind, A. von Graevenitz, P. Kimmig, H. G. Schiefer, T. Schwarz, W. Slenczka, and H. Zahner. *Zoonoses: infectious diseases transmissible from animals to humans.* Fourth edition. ASM Books. John Wiley & Sons, Ltd, 2015.

[22] Q. Bi, Y. Wu, S. Mei, C. Ye, X. Zou, Z. Zhang, X. Liu, L. Wei, S. A. Truelove, T. Zhang, W. Gao, C. Cheng, X. Tang, X. Wu, Y. Wu, B. Sun, S. Huang, Y. Sun, J. Zhang, T. Ma, J. Lessler, and T. Feng. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *The Lancet Infectious Diseases* **20** (2020), 911–919.

Notation	Description	Value	Source
R_0	Basic reproduction number during lockdown	0.734	E
W	Total number of infections before March 17 2020	9.52×10^5	E
$p_{\text{crit}} + p_{\text{sev}}$	Probability of being hospitalized	0.036	[189]
$\frac{p_{\text{crit}}(1-d_{\text{hosp}})}{p_{\text{crit}}+p_{\text{sev}}}$	Probability of entering ICU conditional on being at the hospital	0.19	[189]
$\frac{d_{\text{hosp}}+(1-d_{\text{hosp}})d_{\text{ICU}}}{1+p_{\text{sev}}/p_{\text{crit}}}$	Death probability conditional on being hospitalized	0.181	[189]
d_{ICU}	Probability of death conditional on being in ICU	0.709	E
p_{short}	Probability of a short stay at hospital	0.701	E
D_{C_h}	Delay between severe infection and hospital admission	14.5 days	E
D_{short}	Delay between hospital admission and quick discharge	7.36 days	E
D_{long}	Delay between hospital admission and slow discharge	47.5 days	E
D_{C_u}	Delay between critical infection and hospital admission	11.0 days	E
D_H	Delay between hospital admission and ICU admission	1.5 days	[189]
D_U	Delay between ICU admission and discharge	28.2 days	E
D_D	Delay between ICU admission and death	9.90 days	E
γ	Scale parameter common to all Gamma distributions	0.316	E

Table 6.3: Inferred parameter set for the occupancy model. The values indicated for the durations correspond to the means of the Gamma distributions. In the "Source" column, "E" indicates that the parameter has been estimated in the current work.

[33] T. Britton. Epidemics in heterogeneous communities: estimation of R_0 and secure vaccination coverage. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63** (2001), 705–715.

[34] T. Britton, F. Ball, and P. Trapman. The disease-induced herd immunity level for COVID-19 is substantially lower than the classical herd immunity level (2020). arXiv: 2005.03085.

[39] H. Cohn. Multitype finite mean supercritical age-dependent branching processes. *Journal of applied probability* **26** (1989), 398–403.

[42] D. H. Crawford. *Deadly Companions: How Microbes Shaped our History.* Second edition. Oxford University Press, 2018.

[46] R. Djidjou-Demasse, Y. Michalakis, M. Choisy, M. T. Sofonea, and S. Alizon. Optimal COVID-19 epidemic control until vaccine deployment. *medRxiv* (2020).

[60] T. Evgeniou, M. Fekom, A. Ovchinnikov, R. Porcher, C. Pouchol, and N. Vayatis. Epidemic models for personalised COVID-19 isolation and exit policies using clinical risk predictions. *medRxiv* (2020).

[65] J. Y. Fan, K. Hamza, P. Jagers, and F. C. Klebaner. Convergence of the age structure of general schemes of population processes. *Bernoulli* **26** (2020), 893–926.

[68] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, and C. Fraser. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368** (2020).

[69] R. Ferriere and V. C. Tran. Stochastic and deterministic models for age-structured populations with genetically variable traits. *ESAIM: Proceedings* **27** (2009), 289–310.

[77] F. Foutel-Rodier, F. Blanquart, P. Courau, P. Czuppon, J.-J. Duchamps, J. Gamblin, É. Kerdoncuff, R. Kulathinal, L. Régnier, L. Vuduc, A. Lambert, and E. Schertzer. From individual-based epidemic models to McKendrick-von Foerster PDEs: A guide to modeling and inferring COVID-19 dynamics (2020). arXiv: 2007.09622.

[101] K. Hamza, P. Jagers, and F. C. Klebaner. The age structure of population-dependent general branching processes in environments with a high carrying capacity. *Proceedings of the Steklov Institute of Mathematics* **282** (2013), 90–105.

[112] P. Jagers. *Branching processes with biological applications.* Wiley, 1975.

[113] P. Jagers and F. C. Klebaner. Population-size-dependent and age-dependent branching processes. *Stochastic Processes and their Applications* **87** (2000), 235–254.

[114] P. Jagers and F. C. Klebaner. Population-size-dependent, age-structured branching processes linger around their carrying capacity. *Journal of Applied Probability* **48** (2011), 249–260.

[115] P. Jagers and O. Nerman. Limit theorems for sums determined by branching and other exponentially growing processes. *Stochastic Processes and their Applications* **17** (1984), 47–71.

[116] P. Jagers and O. Nerman. The growth and composition of branching populations. *Advances in Applied Probability* **16** (1984), 221–259.

[144] S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, and J. Lessler. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of internal medicine* **172** (2020), 577–582.

[148] N. M. Linton, T. Kobayashi, Y. Yang, K. Hayashi, A. R. Akhmetzhanov, S.-m. Jung, B. Yuan, R. Kinoshita, and H. Nishiura. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine* **9** (2020), 538.

[153] C. Massonnaud, J. Roux, and P. Crépey. COVID-19: Forecasting short term hospital needs in France. *medRxiv* (2020).

[162] O. Nerman. On the convergence of supercritical general (C-M-J) branching processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **57** (1981), 365–395.

[165] P. Olofsson. The $x \log x$ condition for general branching processes. *Journal of applied probability* **35** (1998), 537–544.

[166] G. Pang and É. Pardoux. Functional limit theorems for non-Markovian epidemic models (2020). arXiv: `2003.03249`.

[185] L. Roques, É. K. Klein, J. Papaix, A. Sar, and S. Soubeyrand. Using early data to estimate the actual infection fatality ratio from COVID-19 in France. *Biology* **9** (2020), 97.

[189] H. Salje, C. Tran Kiem, N. Lefrancq, N. Courtejoie, P. Bosetti, J. Paireau, A. Andronico, N. Hozé, J. Richet, C.-L. Dubost, Y. Le Strat, J. Lessler, D. Levy-Bruhl, A. Fontanet, L. Opatowski, P.-Y. Boelle, and S. Cauchemez. Estimating the burden of SARS-CoV-2 in France. *Science* **369** (2020), 208–211.

[192] Santé Publique France. *COVID-19 : point épidémiologique du 4 juin 2020.* 2020. eprint: `https://www.santepubliquefrance.fr`.

[193] Santé Publique France. *Données hospitalières relatives à l'épidémie de COVID-19.* 2020. eprint: `https://www.data.gouv.fr/`. (accessed: 10.06.2020).

[194] E. Schertzer and F. Simatos. Height and contour processes of Crump-Mode-Jagers forests (I): General distribution and scaling limits in the case of short edges. *Electronic Journal of Probability* **23** (2018), 43 pp.

[198] T. Sellke. On the asymptotic distribution of the size of a stochastic epidemic. *Journal of Applied Probability* **20** (1983), 390–394.

[202] M. T. Sofonea, B. Reyné, B. Elie, R. Djidjou-Demasse, C. Selinger, Y. Michalakis, and S. Alizon. Epidemiological monitoring and control perspectives: application of a parsimonious modelling framework to the COVID-19 dynamics in France. *medRxiv* (2020).

[205] Z. Taïb. *Branching Processes and Neutral Evolution.* Vol. 93. Lecture Notes in Biomathematics. Springer Berlin Heidelberg, 1992.

[207] L. Tindale, M. Coombe, J. E. Stockdale, E. Garlock, W. Y. V. Lau, M. Saraswat, Y.-H. B. Lee, L. Zhang, D. Chen, J. Wallinga, and C. Colijn. Transmission interval estimates suggest pre-symptomatic spread of COVID-19. *medRxiv* (2020).

[210] V. C. Tran. Large population limit and time behaviour of a stochastic particle model describing an age-structured population. *ESAIM: Probability and Statistics* **12** (2008), 345–386.

[214] R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. T. Walker, H. Fu, A. Dighe, J. T. Griffin, M. Baguelin, S. Bhatia, A. Boonyasiri, A. Cori, Z. Cucunubá, R. FitzJohn, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, D. Laydon, G. Nedjati-Gilani, S. Riley, S. van Elsland, E. Volz, H. Wang, Y. Wang, X. Xi, C. A. Donnelly, A. C. Ghani, and N. M. Ferguson. Estimates of the severity of coronavirus disease 2019: A model-based analysis. *The Lancet Infectious Diseases* **20** (2020), 669–677.

[219] J. T. Wu, K. Leung, M. Bushman, N. Kishore, R. Niehus, P. M. de Salazar, B. J. Cowling, M. Lipsitch, and G. M. Leung. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nature Medicine* **26** (2020), 506–510.

Appendices for Chapter 6

6.A Maximum likelihood method

The incidence and prevalence data for France were taken from [193]. For a fixed set of parameters, the solution $n(t, a)$ to the McKendrick-von Foerster equation was solved numerically using a Euler scheme and spatial boundary condition making use of $\tau = R_0\hat{\tau}$ specified by (unknown, to be estimated) R_0 and $\hat{\tau}$ fixed as in (6.18). The predicted number of deaths, discharges, and ICU/hospital occupied beds were then computed numerically using

$$\forall t \geq 0, \quad n_i(t) = \int_0^\infty n(t, a) p(a, i) \, \mathrm{d}a,$$

where $n_i(t)$ is the size of the subpopulation in state i at time t and $p(a, i) = \mathbb{P}(X(a) = i)$, where X is the life process of a typical individual. The predicted incidence in state i between time t and s, denoted by $\tilde{n}_i(t, s)$, can be obtained using the expression

$$\tilde{n}_i(t, s) = n_i(s) - n_i(t) + \sum_j n_j(s) - n_j(t),$$

where the sums is taken over all states j such that the process $(X(a);\, a \geq 0)$ can reach state j after having visited state i. The predicted number of ICU/hospital admissions was computed using this expression.

We considered a Poisson likelihood. More precisely, given the predicted values displayed previously, we assumed that the observed values follow a Poisson distribution whose mean is the corresponding predicted values. We supposed that Poisson observations were independent among days, and among time series. This yields a product-form expression for the likelihood of the data. We then looked for the parameter set that maximizes this likelihood.

The maximum likelihood parameter set was obtained using the `minimize` function of the Python `scipy.optimize` module, using a Nelder-Mead algorithm. We selected as initial point of the optimization algorithm a set of parameters that were close to the existing estimates in the literature, or which seemed realistic if such estimates did not exist.

6.B Best fitting prevalence curves under admission model

Recall the admission model from Section 6.4.4. By adding two parameters to the model, one for the mean time between hospital admission and discharge, the

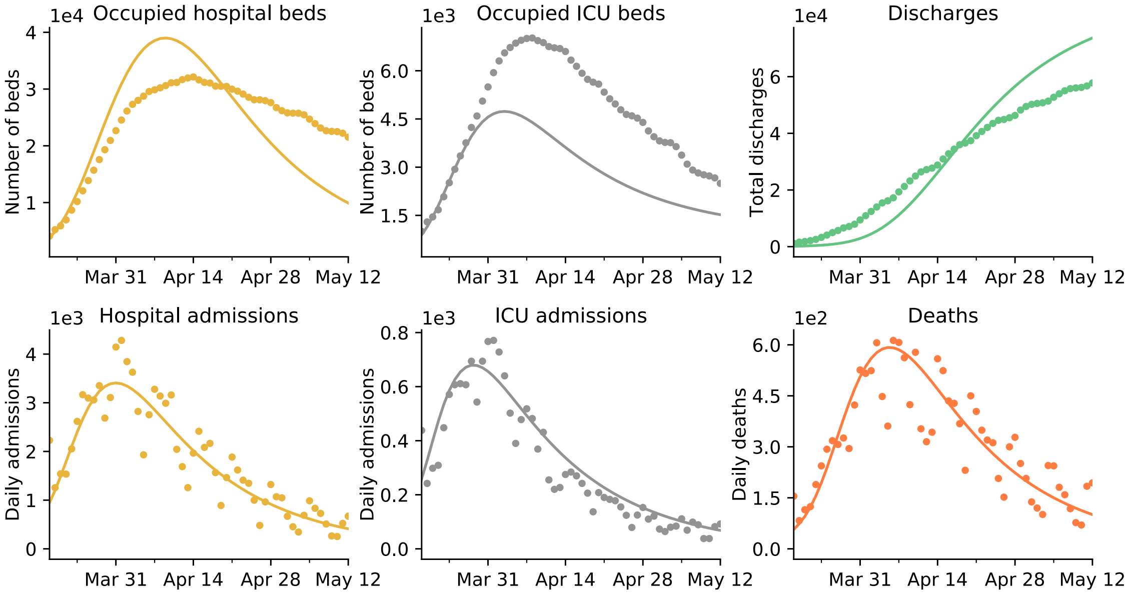
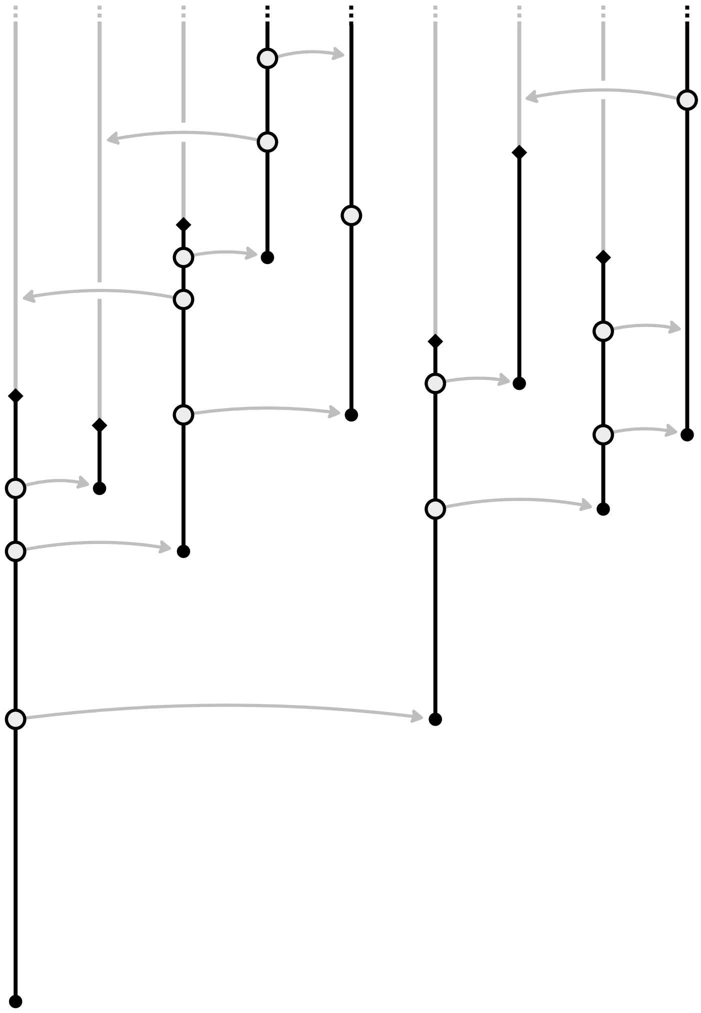


Figure 6.8: Best fit of the admission model for prevalence data.

other for the mean time between ICU admission and discharge, we can derive an expression for the likelihood of the prevalence and incidence time series under the admission model. The best-fitting values for these two parameters were obtained by maximizing the likelihood with all other parameters values fixed to those estimated in Table 6.2. The corresponding model is displayed in Figure 6.8.

Chapter 7



7

From individual-based epidemic models to McKendrick-von Foerster PDEs (II): The non-linear case

This chapter is work in progress with Jean-Jil Duchamps and Emmanuel Schertzer.

Illustration. Graphical representation of a Crump-Mode-Jagers process with saturation.

7.1 Introduction

7.1.1 General individual-based epidemic model

In this chapter, we study an extension of the general epidemiological framework that we introduced in [77] to model the COVID-19 epidemic. Let us briefly recall the main features of this model.

We consider a population made of susceptible individuals, that have never encountered the disease, and infected individuals. Each infected individual is supposed to belong to one *compartment*, that models the stage of the disease of this individual. Classical examples of compartments are the exposed compartment (E), where the individual is infected but not yet infectious, the infectious compartment (I), and the recovered compartment (R), once the individual has become immunized to the illness. In the case of the COVID-19 epidemic, it might be relevant to add a hospitalized (H) and an intensive care unit (U) compartment, as monitoring the number of individuals in these states is typically important for policy making. See Figure 7.1 for an example of compartmental model used for the COVID-19 epidemic. We denote by $\mathcal{S}$ the set of all compartments, and assume that $\mathcal{S}$ is finite.

We encode the compartment to which individual k belongs as a stochastic process $(X_k(a);\, a \geq 0)$ valued in $\mathcal{S}$, that we call the *life-cycle process*. The random variable $X_k(a)$ gives the compartment to which k belongs at *age of infection* a, that is, a unit of time after its infection. Moreover, individual k is endowed with a point

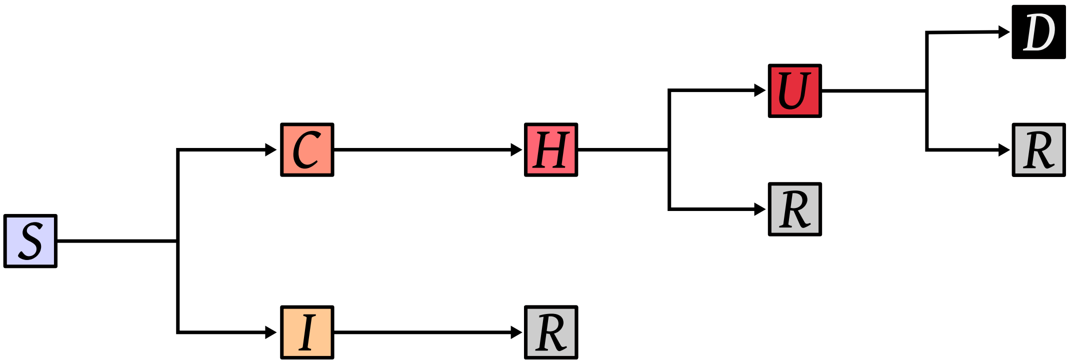


Figure 7.1: An example of compartmental model. The compartments are: S, susceptible; I, mildly infected; C, severely infected; H, hospitalized; U, admitted to intensive care unit; R, recovered; and D, dead.

measure $\mathcal{P}_k$ on $\mathbb{R}_+$ that we call the *infection point process*. The atoms of $\mathcal{P}_k$ encode the age at which k makes infectious contacts with the rest of the population. We think of the pair $(\mathcal{P}_k, X_k)$ as describing the course of the infection of individual k. We make the fundamental assumption that the pairs $(\mathcal{P}_k, X_k)$ are i.i.d. for distinct individuals in the population.

In [77] we assumed that susceptibles are in excess, and that any infectious contact leads to a new infection. The resulting population is then distributed as a Crump-Mode-Jagers (CMJ) process. In the current work, we consider an extension of this model that takes into account the saturation due to the finite pool of susceptibles in the population. More precisely, we consider a population of finite fixed size N. Each infectious contact is made with an individual uniformly chosen in this population, and it results in a new infection only if the targeted individual is susceptible. Finally, we model the impact of control measures, such as school closure, or national lockdown, with a *suppression function* $(c(t); t \geq 0)$. This suppression function is such that an infection occuring at time t is only effective with probability $c(t) \in [0,1]$. With probability $1 - c(t)$, the infection is removed. A formal description of this model is provided in Section 7.2.1

7.1.2 Convergence of the age structure

A standard way to study compartmental models is to consider the dynamics of the number of individuals in each compartment. If the underlying probabilistic model is Markovian, this typically gives rise to systems of ODEs of the SIR type in the large population size limit, see [35] for a recent account. Here, we will not keep track of the count of individuals in the various compartments, but we will rather be interested into the age structure of the population. Our main result is a law of large number for the age structure of population, which is the equivalent of Theorem 7 of [77] for our non-linear extension of the model.

We anticipate the notation of Section 7.2 and denote the empirical measure of ages and compartments in the population at time t as

$$\forall i \in \mathcal{S}, \quad \mu_t^N(\mathrm{d}a, \{i\}) = \sum_{\sigma_k \leq t} \mathbb{1}_{\{X_k(t-\sigma_k)=i\}} \delta_{t-\sigma_k}(\mathrm{d}a),$$

where σ_k is the birth time of individual k, and the sum runs over all infected individuals at time t. (Note that $t-\sigma_k$ is just the age of k at time t.) The measure μ_t^N encodes the ages and compartments of infected individuals at time t. The limiting distribution of μ_t^N will depend on the following two quantities:

- The intensity measure of the infection point process defined as
$$\tau(\mathrm{d}a) \coloneqq \mathbb{E}\Big[\mathcal{P}(\mathrm{d}a)\Big].$$
We assume that τ has a density w.r.t. the Lebesgue measure that we denote by $\tau(a)$ with a slight abuse of notation, and that $R_0 \coloneqq \tau([0,\infty)) < \infty$.
- The one-dimensional marginals of the life-cycle process, denoted by
$$\forall i \in \mathcal{S}, \, \forall a \geq 0, \quad p(a,i) \coloneqq \mathbb{P}\Big(X(a) = i\Big).$$

Let us also denote by

$$\forall t \geq 0, \, \forall i \in \mathcal{S}, \quad Y_t^N(i) = \#\{\text{individuals in } i \text{ at time } t\}.$$

We define $\mathcal{I}_0^N \subseteq [N]$ as the set of infected individuals, and denote by $|\mathcal{I}_0^N|$ the number of individuals in $\mathcal{I}_0^N$. We suppose that the ages of those individuals are i.i.d. with common distribution $g(a)\,\mathrm{d}a$ for some probability density g. See Section 7.2.1 for a formal description of this initial condition We can now state our main convergence result.

Theorem 7.1. *Assume that there exists $I_0 > 0$ such that*

$$\lim_{N\to\infty} \frac{1}{N} \, |\mathcal{I}_0^N| = I_0$$

in probability. Then, as $N \to \infty$, the following convergence holds in probability for the weak topology

$$\frac{1}{N}\mu_t^N(\mathrm{d}a, \{i\}) \longrightarrow n(t,a)p(a,i)\,\mathrm{d}a$$

where $(n(t,a);\, t,a \geq 0)$ is the solution to

$$\begin{aligned} \partial_t n(t,a) + \partial_a n(t,a) &= 0 \\ \forall t \geq 0, \; n(t,0) &= c(t)S(t)\int_0^\infty n(t,a)\,\tau(\mathrm{d}a) \\ \forall a \geq 0, \; n(0,a) &= I_0 g(a) \\ \forall t \geq 0, \; S(t) &= 1 - \int_0^\infty n(t,a)\,\mathrm{d}a. \end{aligned} \tag{7.1}$$

Moreover, for any $i \in \mathcal{S}$, we have

$$\Big(\frac{1}{N}Y_t^N(i);\, t \geq 0\Big) \longrightarrow \Big(\int_0^\infty n(t,a)p(a,i)\,\mathrm{d}a;\, t \geq 0\Big)$$

in probability in the Skorohod topology.

This result will follow from the more general Theorem 7.6, and is proved in Section 7.2.3. The definition of a solution to equation (7.1) is provided in Section 7.2.2. Theorem 7.1 states that the age structure of the population converges to a limiting non-linear PDE of the McKendrick-von Foerster type. Moreover, it also entails that the number of individuals in each compartment can be recovered by integrating the one-dimensional marginals $p(a, i)$ against the age structure.

There are two consequences of our result that we would like to emphasize. First, it shows that the macroscopic dynamics of the infected population is given by a universal equation, the McKendrick-von Foerster PDE, which does not depend on the distribution of the life-cycle process. In order to recover the number of individuals in each compartment, one needs to decorate this PDE with a life-cycle process. The expression that links the age structure to the individual counts in each compartment is elementary.

Second, our approach allows to identify the characteristics of the microscopic model that impact the large population size dynamics. Recall that X and $\mathcal{P}$ are *a priori* correlated in a complex fashion: in time, because X is not a Markov process, and $\mathcal{P}$ is not a homogeneous Poisson process; and between them, as X and $\mathcal{P}$ are not independent. However, in the limit, the only two parameters that impact the dynamics of the epidemic are the intensity measure τ and the one-dimensional marginals p. These parameters are “first order quantities” of X and $\mathcal{P}$ in the sense that they only involve the distribution of the respective processes at one point in time, and are not influenced by the aforementioned correlations. Moreover, τ is the intensity measure of the infection point process, “averaged” over all life-cycles. Therefore, there is no need to assess which compartments are the most infectious in order to compute τ. Finally, let us argue that τ and p are two quantities that can be accessed in real epidemics. Write the intensity τ as

$$\tau = R_0 \nu,$$

where

$$R_0 = \int_0^\infty \tau(\mathrm{d}a), \qquad \nu(\mathrm{d}a) = \frac{\tau(\mathrm{d}a)}{R_0}.$$

These two quantities have clear epidemiological interpretation:

- R_0 is the basic reproduction number, that is, the mean number of secondary infections induced by a single individual in an entirely susceptible population;
- ν is the distribution of the generation time, that is, the time between the infection of the source individual and that of the recipient individual in a typical infection pair [81].

The generation time distribution can be inferred from the time interval between the symptom onset of two individuals in an identified infectious contact, as in [68]. The basic reproduction number R_0 is typically a quantity that needs to be estimated from the course of the epidemic, and plays an important role in assessing the efficiency of and planning control measures. The one-dimensional marginals can be inferred using a compartmental model as in [77].

7.1.3 A genealogical dual to the delay equation

The McKendrick-von Foerster equation (7.1) can be reformulated in terms of a non-linear delay equation. If $(n(t,a);\, t, a \geq 0)$ denotes the solution to equation (7.1) with $c \equiv 1$, let us define the number of infections between time 0 and t as

$$B(t) = \int_0^t n(s,0)\,\mathrm{d}s = \int_0^t S(s) \int_0^\infty n(s,a)\,\tau(\mathrm{d}a)\mathrm{d}s.$$

Then we will derive in Section 7.2.2 that B solves the following non-linear delay equation:

$$B(t) = S_0\bigg(1 - \exp\bigg(- \int_0^t \tau(a) B(t-a)\,\mathrm{d}a - \int_0^\infty \int_0^t \tau(a+s) g(a)\,\mathrm{d}s\,\mathrm{d}a\bigg)\bigg), \quad (7.2)$$

where $S_0 = 1 - I_0$ is the initial number of susceptibles.

Our proof of Theorem 7.1 uses a genealogical approach, where we look backwards in time at the set of potential infectors of a focal individual. This approach leads to a genealogical dual to the delay equation that we think to be of independent interest. The dual is built out of the following branching process.

Recall that R_0 stands for the total mass of τ and $\nu = \tau / R_0$. We define a measure $\bar{\tau}$ on $\mathbb{R}_+$ such that for any continuous bounded function φ,

$$\int_0^\infty \varphi(u)\,\bar{\tau}(\mathrm{d}u) = \int_0^\infty g(a) \int_a^\infty \varphi(u-a) \tau(u)\,\mathrm{d}u\,\mathrm{d}a.$$

The measure $\bar{\tau}$ is the intensity measure of the infection point process of an individual with initial age distributed as g. Let us further set

$$\bar{R}_0 = \int_0^\infty \bar{\tau}(\mathrm{d}a), \quad \bar{\nu}(\mathrm{d}a) = \frac{\bar{\tau}(\mathrm{d}a)}{\bar{R}_0}.$$

The branching process is constructed as follows. Let us assume that individuals in the branching process are either marked or unmarked. Suppose that the population starts from a single unmarked individual. Then, at each generation, an unmarked individual produces:

- a Poisson$\big(S_0 R_0\big)$ distributed number of unmarked individuals;
- a Poisson$\big((1-S_0)\bar{R}_0\big)$ distributed number of marked individuals.

Unmarked individuals have no offspring. Draw an oriented edge from each individual towards its parent. Assign a weight independently to each edge, such that the weight of an edge originating from an unmarked individual is distributed as ν, and that of an edge coming from a marked individual is distributed as $\bar{\nu}$.

The previous branching process corresponds to the large population size limit of the set of potential infectors of a fixed individual. Marked individuals correspond to individuals that were initially infected. Each edge corresponds to an infectious contact in the population, and the weight of that edge is the age at which this contact occurs.

Let us denote by σ^∞ be the minimum of the lengths of the paths from a marked individual to the root in the previous tree. The length of a path is defined as the sum of the weights of the edges along the path. Then the following result connects the distribution of σ^∞ to the delay equation.

Proposition 7.2. *For any $t \geq 0$, define*

$$B(t) = S_0 \mathbb{P}(\sigma^\infty \leq t).$$

Then $(B(t);\, t \geq 0)$ solves the delay equation (7.2).

In Section 7.3.3, we will derive a similar dual for the delay equation with $c \not\equiv 1$. The previous proposition will be a special case of the more general Proposition 7.11.

7.1.4 Link with literature

The idea of considering an infection through its age structure dates back to at least the work of [125], who introduced the SIR model as a special case of a more general age-structured model. More generally, delay equations, which are an equivalent formulation of the McKendrick-von Foerster PDE, have been widely proposed as models for the spread of epidemics, see for instance [40, 217, 31, 32] Surprisingly, the convergence of probabilistic epidemic models towards the solution of a delay equation or a McKendrick-von Foerster PDE has received only little attention. Let us briefly review the works in this direction that we are aware of.

In [12], the authors studied an epidemic model extremely similar to the one under consideration here. The only difference with our model is that they do not explicitly model the compartments, and do not allow for a suppression function $(c(t);\, t \geq 0)$. They obtain a slightly different kind of law of large number. Instead of starting from a macroscopic fraction of infected individuals, they look at the epidemic started from one infected individual. They show that, after an appropriate time-shift so as to skip the long initial branching phase when there are few individuals, the number of susceptibles converges to a solution of the following delay equation

$$\forall t \in \mathbb{R}, \quad \dot{S}(t) = S(t) \int_0^\infty \tau(a) \dot{S}(t-a)\, \mathrm{d}a,$$

see their Theorem 2.10. Setting $B(t) = 1 - S(t)$ for the number of infecteds, the previous equation can be written in the following integral form,

$$\forall t \in \mathbb{R}, \quad B(t) = 1 - \exp\Big(- \int_0^\infty \tau(a) B(t-a)\, \mathrm{d}a \Big).$$

Compare this to equation (7.2). The limiting delay equation obtained in [12] is defined on the whole real line, and thus does not have an initial age profile at $t = 0$. Note also that this delay equation is stationary, in the sense that shifting a solution to the delay equation yields another solution. We became aware of [12] after having derived most of the results that are presented in this chapter. Even if our main result is very close to that of [12], let us highlight some important differences. First, [12] considered rather restrictive hypothesis on the infection point process τ, see Assumption 2 on the top of page 7. For instance, the case of the Markovian SIR model is not covered by these assumptions. Here, our only assumption is that τ has a finite mean, which is the minimal assumption to derive a law of large numbers. Second, even if the bulk of the work in this chapter is to prove the convergence of the age structure, and that further incorporating the life-cycle process $(X(a);\, a \geq 0)$ and the suppression function $(c(t);\, t \geq 0)$ is quite straightforward, these two extensions are very important from an application point of view. This is especially true in the case of the COVID-19 dynamics, where it is important to monitor many compartments, such as the number of ICU and hospital beds that are occupied, and where the transmission rate can vary greatly due to the enforcement of control measures. Finally, we believe that our graph convergence approach can be adapted to study more general epidemic models, see the discussion in the forthcoming Section 7.1.5.

In a recent work, [166] derived a functional law of large numbers and a functional central limit theorem for classical SIR-like models with general sojourn time distribution in each compartment. They also obtained similar results for extensions of these models incorporating spatial heterogeneity [167] and varying infectiosity [73], and applied these models to the COVID-19 epidemic in France [74]. The limiting equations that describe the dynamics of the density of individuals in each compartments are systems of so-called Volterra integral equations. These equations are particular cases of the more general McKendrick-von Foerster equation (7.1), when the infection point process is assumed to be a (inhomogeneous) Poisson point process restricted to the I state. However, their setting allows for more general initial conditions and the bulk of their works is dedicated to the proof of the various central limit theorems, which is out of reach with our current method.

Finally, there exists a rich literature on general age-dependent population processes, not necessarily related to epidemic models. Let us first mention the Crump-Mode-Jagers (CMJ) processes, where the birth times of the children are allowed to depend in a very general way on the age of the parent, see [112] or [205] for a more recent account. The only restriction is that individuals should reproduce independently from each other. By assuming that the age structure of the population is a Markov process, it is possible to release this assumption and consider a much wider class of age-dependent models. Using the framework introduced in [113, 114], [101, 65] derive respectively a law of large numbers and a central limit theorem for the age structure of a very general class of population models. The deterministic limit they obtain for the age structure corresponds to the McKendrick-von Foerster PDE that we have derived here. Even if our results are not trivially implied by [101, 65], as we do not make any Markov assumption, we believe that the main contribution of our work is to use explicitly the notion of age structure of a population in an epidemiological context, with a probabilistic framework which can be readily used for applications [77].

7.1.5 Future directions, outline

One initial infected. Theorem 7.1 provides the dynamics of the age structure of the population, when started from a macroscopic fraction of infecteds. This approach requires to prescribe an initial age profile, which is a quantity that cannot be easily observed. However, in many situations, the large number of infected individuals results from natural population growth from a small number of initial infecteds. In this case, the age profile of the population is shaped by the initial population growth, and it is natural to ask what this age profile should be.

The answer to that question is provided in [12], again under some rather drastic conditions on the infection point process. Recall that they proved the convergence of the number of susceptibles to a stationary version of the delay equation (7.2). Moreover, it was shown in [45] that this equation admits a unique positive nonincreasing solution, up to time-shift. Thus, for a given initial fraction of infecteds in the population, there is a unique initial age profile such that the solution of the McKendrick-von Foerster equation (7.1) coincides with the restriction of the delay equation to $\mathbb{R}_+$. This age profile should be the natural candidate in many situations.

From an application point of view, as we have already mentioned, it is interesting to explicitly model the compartments, and to introduce the suppression function $(c(t);\, t \geq 0)$. The branching approximation for the model that we consider here is the CMJ process that was studied in [77]. We know from general results [163] that the limiting age profile in a CMJ process is an exponential distribution, whose parameter is the Malthusian growth parameter of the CMJ. We believe that we could recover the result of [12] by plugging this exponential age profile as the initial condition of (7.1), and by letting the initial population size vanish. This broad idea that the large population size limit of a population process, when started from a few individuals, can be described as a deterministic dynamical system with a random initial condition that originates from the initial branching phase as already been considered in a variety of contexts, see for instance [11, 8].

More general dependence structure. It is important to derive epidemic models that account for time variations in the contact rate. Such variations can reflect the enforcement of control measures, such as the lockdown, school closure, mandatory masks, etc., but also reflect behavioral changes that occur during the course of the epidemic: some people avoid crowds, wash hands more often, etc. These variations are captured by the notion of instantaneous reproduction number [81, 40], R_t, which in our context can be written as

$$R_t := \frac{n(t,0)}{\int_0^\infty n(t,a)\tau(a)\,\mathrm{d}a} = R_0 c(t) S(t).$$

Therefore, in the present work, variations in R_t originate from the reduction of the number of susceptibles, and from the suppression function $(c(t);\, t \geq 0)$. From a modeling perspective this is not satisfying, as it is very likely that the variation of R_t is not a “pure” temporal effect, but rather the consequence of the dynamics of the epidemic itself. For example, control measures are typically enforced when some indicator (for instance, the number of deaths, or the number of cases) reaches a given threshold, and people's perception of the spread of the epidemic, which is probably one of the main driver of behavioral changes, is highly dependent on regular reports of similar indicators.

A possible way to model that is to introduce a more complex suppression function

$$C\colon \mathbb{R}_+ \times \mathcal{M}(\mathbb{R}_+ \times \mathcal{S}) \to [0,1],$$

where $\mathcal{M}(\mathbb{R}_+ \times \mathcal{S})$ is the space of positive measures on ages and compartments. An infection occurring at time t would then be effective with probability $C(t, \mu_t^N)$, where μ_t^N denotes the age and compartment empirical measure of the population at time t. We believe that, under some regularity conditions similar to Definition 3.1 of [101], the graph method that we develop here could be adapted to prove the convergence of this extension of our model to a McKendrick-von Foerster equation of the following form

$$\begin{aligned}
\partial_t n(t,a) + \partial_a n(t,a) &= 0 \\
\forall t \geq 0,\ n(t,0) &= C(t,\mu_t) S(t) \int_0^\infty n(t,a)\tau(a)\,\mathrm{d}a \\
\forall a \geq 0,\ n(0,a) &= I_0 g(a) \\
\forall t \geq 0,\ S(t) &= 1 - \int_0^\infty n(t,a)\,\mathrm{d}a \\
\mu_t(\mathrm{d}a, \{i\}) &= n(t,a) p(a,i)\,\mathrm{d}a.
\end{aligned}$$

In this situation, the instantaneous reproduction number is

$$R_t = R_0 C(t,\mu_t) S(t),$$

so that we have a more mechanistic interpretation of R_t, but it should not be possible to estimate the function C in practice.

Outline. The rest of this chapter is organized as follows. A formal description of the model is provided in Section 7.2.1, and the McKendrick-von Foerster PDE is studied in Section 7.2.2. Section 7.2.3 contains the statement of our main result, which states that the empirical distribution of infection times in the population converges to a delay equation.

The proof are carried out using a graph approach. The infection graph is introduced in Section 7.3.1 and Section 7.3.2. The last three sections, Section 7.3.3, Section 7.3.4, and Section 7.3.5 are dedicated to the proofs of the various convergence results.

7.2 Epidemiological model with saturation decorated with a life cycle

7.2.1 Description of the model

In the following, we will consider an epidemic model in which individuals' life trajectories are represented by independent stochastic processes. We distinguish between two types of individuals:

- Susceptible individuals that have never been infected before.
- Infected individuals that have been infected in the past. We emphasize that the meaning of infected is a bit broader than usual. For instance, a recovered or dead individual is considered as infected. To each infected individual, we associate an age. The age is the time elapsed since the beginning of the infection.

There are N individuals in the population. To each individual $x \in [N]$, we associate a pair of processes $(\mathcal{P}_x, X_x)$ describing respectively the process of secondary infections and the successive stages of the disease experienced by the focal individual x. More precisely:

- The *life-cycle process*, denoted by $(X_x(a);\, a \geq 0)$, is a random process valued in $\mathcal{S}$ where $X_x(a)$ is the stage of the disease (e.g., exposed, death, etc.) of x at age a.
- The *infection point process* $\mathcal{P}_x$ is a point measure describing the ages of potential infections.

Let us denote by $\mathcal{X}_x = (\mathcal{P}_x, X_x)$. We will always assume that $(\mathcal{X}_x;\, x \in [N])$ are i.i.d. and denote by $\mathcal{X} = (\mathcal{P}, X)$ their common distribution. The state space of $\mathcal{X}$ is denoted by $\mathscr{X}$.

Remark 7.3. Note that we allow for non-trivial correlation between the life-cycle and the infection process. Examples of such correlations can be that a deceased individual is not infectious anymore, a hospitalized individual may have a reduced potential of infection due to quarantine, etc. ○

We suppose that at $t = 0$, a subset $\mathcal{I}_0^N \subseteq [N]$ of the population is infected. For each $x \in \mathcal{I}_0^N$ we need to prescribe an age, or equivalently, an infection time. We assume that, conditional on $\mathcal{I}_0^N$, the ages of the initial individuals $(T_x;\, x \in \mathcal{I}_0^N)$ are i.i.d. with common distribution g. Let us denote by $(\sigma_x^N;\, x \in \mathcal{I}_0^N)$ the birth time of the initial infecteds, that is, $\sigma_x^N = -T_x$.

The epidemic now spreads as follows. Suppose that, at some time t_0, we have defined a set $\mathcal{I}_{t_0}^N \subseteq [N]$ of infected individuals at time t_0, and a vector $(\sigma_x^N;\, x \in \mathcal{I}_{t_0}^N)$ of infection times. Let t_1 be the first atom after t_0 of the point measure

$$\sum_{x \in \mathcal{I}_{t_0}^N} \sum_{a \in \mathcal{P}_x} \delta(\sigma_x + a).$$

If there is no such atom, the infection stops. Otherwise, let U be uniformly chosen in $[N]$, independent of the rest, it is the first individual that comes in contact with any of the infected individuals after time t_0. If $U \in \mathcal{I}^N_{t_0}$, then nothing happens, and we carry out the same procedure for the next atom t_2. If $U \notin \mathcal{I}^N_{t_0}$, then, with probability $1 - c(t_1)$, the infection is ineffective in which case nothing happens and we consider the next infection time t_2. Otherwise, set $\mathcal{I}^N_{t_1} = \mathcal{I}^N_{t_0} \cup \{U\}$ and $\sigma^N_U = t_1$, and continue the procedure as if starting from time t_1 with the initial infected set $\mathcal{I}^N_{t_1}$. This inductive procedure will be reformulated in terms of a graph in Section 7.3.1.

7.2.2 McKendrick-von Foerster PDE

In this section we provide our definition of the solution to the McKendrick-von Foerster equation (7.1). We start with a formal resolution of the PDE using the method of characteristics.

Let I_0 be the initial density of infected individuals and g the initial age profile of the population. First, note that if n is solution of the PDE, then for every pair (t, a) of non-negative numbers, $s \mapsto n(t - s, a - s)$ is constant on $(0, t \wedge a)$. This yields

$$\forall t, a \geq 0, \quad n(t,a) = \begin{cases} I_0 g(a-t) & \text{when } a > t \\ b(t-a) & \text{when } a \leq t, \end{cases} \tag{7.3}$$

with

$$\forall t \geq 0, \quad b(t) := n(t, 0)$$

is the number of new infections at time t. Moreover,

$$\begin{aligned} \dot{S}(t) &= -\int_0^\infty \partial_t n(t,a)\,\mathrm{d}a = \int_0^\infty \partial_a n(t,a)\,\mathrm{d}a \\ &= -b(t) = -c(t) S(t) \int_0^\infty \tau(a) n(t,a)\,\mathrm{d}a. \end{aligned}$$

As a result, we have

$$\begin{aligned} S(t) &= S(0) \exp\Bigg(- \int_0^t c(s) \int_0^\infty \tau(a) n(s,a)\,\mathrm{d}a\,\mathrm{d}s \Bigg) \\ &= S(0) \exp\Bigg(- \int_0^t c(s) \Big(\int_0^s \tau(a) b(s-a)\,\mathrm{d}a + I_0 \int_s^\infty \tau(a) g(a-s)\,\mathrm{d}a \Big)\,\mathrm{d}s \Bigg) \\ &= S(0) \exp\Bigg(- \int_0^t c(s) \Big(\int_0^s \tau(s-a) b(a)\,\mathrm{d}a + I_0 \int_0^\infty \tau(a+s) g(a)\,\mathrm{d}a \Big)\,\mathrm{d}s \Bigg) \end{aligned}$$

so necessarily

$$B(t) := \int_0^t b(s)\,\mathrm{d}s = S_0 - S(t)$$

solves the nonlinear delay equation

$$B(t) = S_0\Big[1 - \exp\Big(-\int_0^t c(s)\Big(\int_0^s \tau(s-a)B(\mathrm{d}a)\,\mathrm{d}a + I_0\int_0^\infty \tau(a+s)g(a)\,\mathrm{d}a\Big)\,\mathrm{d}s\Big)\Big] \tag{7.4}$$

where $B(\mathrm{d}a) = b(a)\,\mathrm{d}a$ is the Stieltjes measure associated to the nondecreasing map B. This motivates the following definition of a solution to the McKendrick-von Foerster equation.

Definition 7.4. We say that $(n(t,a);\, t, a \geq 0)$ is a weak solution to (7.1) if there exists a nonnegative function $(b(t);\, t \geq 0)$ such that:

(i) the functions n and b are related through (7.3);

(ii) the function $B(t) := \int_0^t b(s)\,\mathrm{d}s$ solves the delay equation (7.4). ○

If a nondecreasing function B satisfies (7.4), then we have the following inequality:

$$B(t+u) - B(t) \leq S(0)\int_t^{t+u} c(s)\Big(\int_0^s \tau(s-a)B(\mathrm{d}a) + I_0\int_0^\infty \tau(a+s)g(a)\,\mathrm{d}a\Big)\,\mathrm{d}s.$$

The previous inequality readily entails that B is absolutely continuous, and thus that we can find b such that $B(t) = \int_0^t b(s)\,\mathrm{d}s$. Therefore, existence and uniqueness of solutions to (7.1) reduce to existence and uniqueness of nondecreasing solutions to (7.4), which is provided by the following result.

Lemma 7.5. *There is a unique nondecreasing, nonnegative solution to* (7.4).

Proof. Let us denote by E the set of all nondecreasing, nonnegative, càdlàg functions on $[0,\infty)$. For $\gamma > \alpha \vee 0$, define

$$E_\gamma = \{f \in E : \int_0^\infty e^{-\gamma t} f(t)\,\mathrm{d}t < \infty\}.$$

We endow E_γ with the metric

$$d_\gamma(f,g) = \int_0^\infty e^{-\gamma t}|f(t) - g(t)|\,\mathrm{d}t$$

which makes (E_γ, d_γ) a complete metric space. As any solution to (7.4) is bounded and continuous, it is sufficient to show existence and uniqueness of the solution in E_γ.

We introduce the operator $\Phi\colon E_\gamma \to E_\gamma$ such that

$$\Phi f(t) = S_0\Big(1 - \exp\Big(-\int_0^t c(s)\Big(\int_0^s \tau(s-a)f(\mathrm{d}a)\Big)\,\mathrm{d}s \\ - I_0\int_0^\infty \Big(\int_0^t c(s)\tau(a+s)\,\mathrm{d}s\Big)g(a)\,\mathrm{d}a\Big)\Big),$$

where $f(\mathrm{d}a)$ denotes the Stieltjes measure associated to f. Note that $\Phi f \in E_\gamma$, since it is clear that Φf is bounded, continuous, nonnegative and nondecreasing. Let us show that Φ is a contraction. We have, for $f_1, f_2 \in E_\gamma$,

$$\begin{aligned} d_\gamma(\Phi f_1, \Phi f_2) &\leq S_0 \int_0^\infty e^{-\gamma t} \Big| \int_0^t c(s) \Big(\int_0^s \tau(s-a) f_1(\mathrm{d}a) - \int_0^s \tau(s-a) f_2(\mathrm{d}a) \Big) \,\mathrm{d}s \Big| \,\mathrm{d}t \\ &\leq \int_0^\infty e^{-\gamma t} \Big(\int_0^t \tau(s) |f_1(t-s) - f_2(t-s)| \,\mathrm{d}s \Big) \,\mathrm{d}t \\ &= d_\gamma(f_1, f_2) \int_0^\infty e^{-\gamma t} \tau(t) \,\mathrm{d}t. \end{aligned}$$

As $\gamma > \alpha$, we know that $\int_0^\infty e^{-\gamma t}\tau(t)\,\mathrm{d}t < 1$, showing that Φ is a contraction. The Banach fixed point theorem therefore shows that there exists a unique $B \in E_\gamma$ such that $\Phi B = B$, ending the proof. □

7.2.3 Main result

Rather than proving directly the convergence of the age and compartment empirical distribution of the population to the McKendrick-von Foerster equation, we will prove the convergence of the empirical measure of birth times in the population to the delay equation (7.4). Let us start by introducing some notation.

Recall that $\mathcal{X}_x = (\mathcal{P}_x, X_x) \in \mathscr{X}$ stands for the life-cycle and infection process of $x \in [N]$, and σ_x for the birth time of x. We define the empirical birth measure of the infection as the following measure on $\mathbb{R} \times \mathscr{X}$:

$$\lambda^N := \sum_{x \in [N]} \mathbb{1}_{\{\sigma_x < \infty\}} \delta(\sigma_x, \mathcal{X}_x).$$

The following result proves the convergence of λ^N to the solution of the delay equation (7.4).

Theorem 7.6. *Suppose that there exists $I_0 > 0$ such that*

$$\lim_{N\to\infty} \frac{1}{N} |\mathcal{I}_0^N| = I_0$$

in probability. Then as $N \to \infty$,

$$\frac{1}{N} \lambda^N \longrightarrow \widetilde{b}(t)\,\mathrm{d}t \otimes \mathbb{P}(\mathcal{X} \in \cdot),$$

where the convergence is in distribution for the weak topology and $\widetilde{b}\colon \mathbb{R} \to \mathbb{R}_+$ is the map defined by

$$\widetilde{b}(t) = \begin{cases} I_0 g(-t) & \text{if } t < 0 \\ b(t) = n(t,0) & \text{if } t \geq 0, \end{cases}$$

where $n(t,a)$ is the weak solution to (7.1).

Our proof of this result uses a random graph representation of the infection process which we introduce in Section 7.3. The proof is deferred until Section 7.3.4.

Recall the notation

$$\mu_t^N = \sum_{x \in [N]} \mathbb{1}_{\{\sigma_x \le t\}} \delta(t - \sigma_x, X_x(t - \sigma_x))$$

for the empirical distribution of ages and compartments at time t, and the notation

$$Y_t^N(i) = \sum_{x \in [N]} \mathbb{1}_{\{\sigma_x \le t, X_x(t-\sigma_x) = i\}} = \mu_t^N\big([0,\infty), \{i\}\big)$$

for the number of individuals in compartment i at time t. Note that μ_t^N can be written in terms of λ^N as follows

$$\int f(a,i)\, \mu_t^N(\mathrm{d}a, \mathrm{d}i) = \int \mathbb{1}_{\sigma \le t} f(t-\sigma, X_\xi(t-\sigma))\, \lambda^N(\mathrm{d}\sigma, \mathrm{d}\xi), \tag{7.5}$$

where $\xi = (\mathcal{P}_\xi, X_\xi)$ denotes a generic element of $\mathscr{X}$. Theorem 7.1 is now a direct consequence of Theorem 7.6.

Proof of Theorem 7.1. The convergence of μ_t^N for fixed $t \ge 0$ is immediate from that of λ^N by formula (7.5).

Because of the expressions of $Y^N(i)$ in terms of μ_t^N, identification of their limit is trivial. All there is to check is tightness of the processes. We will make the simplifying assumption that the underlying compartmental model has a “tree shape”. Without being too formal, we assume that for any two compartments $i, j \in \mathcal{S}$, if j can be accessed from i with positive probability, that is, if the event that we can find $s \le t$ such that $X(s) = i$ and $X(t) = j$ has positive probability, then i cannot be accessed from j. This assumption is not very restrictive, most natural compartmental models enjoy this “tree shape” property. Then, writing $i \preceq j$ if j can be accessed from i, the process

$$\sum_{j:\, i \preceq j} \frac{1}{N} Y_t^N(j),$$

is nondecreasing in time. Since the finite-dimensional marginals of this process converge towards a continuous limit, tightness follows easily, see for instance Theorem 3.37, Chapter VI of [111]. The tightness of $Y_t^N(i)/N$ follows by subtracting the previous processes in an appropriate way. □

7.3 A graph point of view of the infection

7.3.1 Infection graph

Recall the infection model defined in Section 7.2.1, and the notation $(\mathcal{P}_x;\ x \in [N])$ for the infection point processes, $\mathcal{I}_0^N$ for the set of initially infected individuals, and $(\sigma_x^N;\ x \in \mathcal{I}_0^N)$ for their birth time. Each atom of a point process $\mathcal{P}_x$ encodes an infectious contact, which is targeted to a uniformly chosen individual in the population. We now enrich the infection point processes by adding the information about the label of this target.

Formally, we define a collection $(\widehat{\mathcal{P}}_x;\ x \in [N])$ of point measures on $\mathbb{R}_+ \times [N]$ as follows.

- If $x \in \mathcal{I}_0^N$, conditional on $\mathcal{P}_x$, define
$$\widehat{\mathcal{P}}_x = \sum_{a \in \mathcal{P}} \mathbb{1}_{\{a+\sigma_x \geq 0\}} \delta(a+\sigma_x, U_{x,i})$$
where $(U_{x,i};\ i \in\ [N])$ are i.i.d. variables uniform on $[N]$, independent of all other variables. Note that all atoms before $t=0$ have been removed.

- If $x \notin \mathcal{I}_0^N$, conditional on $\mathcal{P}_x$, define
$$\widehat{\mathcal{P}}_x = \sum_{a \in \mathcal{P}} \delta(a, U_{x,i})$$
where again $(U_{x,i};\ i \in\ [N])$ are i.i.d. variables uniform on $[N]$, independent of all other variables.

We now build a graph out of the family $(\widehat{\mathcal{P}}_x;\ x \in [N])$ that records the potential infections in the population. For every $j \notin \mathcal{I}_0^N$, define the set of potential infectors
$$A_j^N = \{i \in [N] : (a,j) \in \widehat{\mathcal{P}}_i\}.$$
(Note that this set is possibly empty.) We give the following definition of the infection graph.

Definition 7.7. The *infection graph* built from the collection of point processes $(\widehat{\mathcal{P}}_x;\ x \in [N])$ is the random oriented weighted graph $\mathcal{G}^N = (V^N, E^N)$ with $V^N = [N]$ and
$$E^N = \bigcup_{i \in\ [N]} \ \bigsqcup_{(a,j) \in \widehat{\mathcal{P}}_i} (i,j),$$
where the second union is a disjoint union. Each edge e corresponds to an atom (a_e, j_e) of some point process $\widehat{\mathcal{P}}_{i_e}$. We define the weight of e to be a_e. ○

Remark 7.8. As the second union is a disjoint union, for every pair (i,j) we allow for multiple edges from i to j in the infection graph. ○

A path in $\mathcal{G}^N$ is a set of edges $\pi = (e_1, \dots, e_n)$ such that, $j_{e_k}\ = i_{e_{k+1}}$, with the notation (i_e, j_e) for the origin and target vertices of the edge e. The length of a path $|\pi|$ is defined as
$$|\pi| = \sum_{e_k \in \pi} a_{e_k}.$$
We say that π is a path from i to j if $i_{e_1} = i$ and $j_{e_n} = j$. A path in $\mathcal{G}^N$ from i to j corresponds to a potential infection chain between i and j. The length of the path is the length of time interval between the infection of i and that j.

If $c \equiv 1$, the graph $\mathcal{G}^N$ contains all information about the epidemic. In this case, the infection time of an individual x is given by the length of the shortest path from an initially infected individual $y \in \mathcal{I}_0^N$ to x. However, for general c, not all infections are effective, and some edges of $\mathcal{G}^N$ need to be removed. The following random procedure describes how the birth time of x can be recovered from a realization of the infection graph.

Procedure 7.9. Consider an initial infection graph $\mathcal{G}^N$ and a focal vertex $x \in [N]$. Let (π^k) be the set of all paths from a some vertex $y \in \mathcal{I}_0^N$ to x, ordered in such a way that

$$|\pi^1| < |\pi^2| < \dots$$

(Note that such an ordering always exists since we have assumed that τ has a density w.r.t. the Lebesgue measure.) We will assign to all edges in (π^k) a state, which can be marked or removed. Marked edges, resp. removed edges, correspond to infections that are known to be effective, resp. ineffective. Consider all the paths (π^k) in the previous order, starting from π^1.

At step k, let $(e_1^k, \dots, e_n^k)$ denote the edges of π^k. Examine successively these edges, starting from e_1^k. Suppose that e_p^k is being examined. It can be in one of three states:

- marked: then examine e_{p+1}^k;
- removed: then move to step $k+1$;
- unmarked: let $t = a_{e_1^k} + \dots + a_{e_p^k}$. With probability $c(t)$ mark e_p^k and examine e_{p+1}^k. Otherwise remove e_p^k and move to step $k+1$.

The procedure stops when all edge from a path π^{k_0} are marked. In this case set $\sigma_x^N = |\pi^{k_0}|$. If the procedure does not end, set $\sigma_x^N = \infty$, and the individual x is not infected. ○

Remark 7.10. (i) Even if the graph $\mathcal{G}^N$ is finite, the set of all paths from an infected individual to x can be infinite due to possible loops.

(ii) When $c \equiv 1$, Procedure 7.9 stops at $k_0 = 1$, provided that there is a path from a vertex $y \in \mathcal{I}_0^N$ to x. ○

It is not completely clear that the birth time σ_x defined from the previous procedure coincides with that defined from the description of the model of Section 7.2.1. To see this, note that when an edge (i, j) is examined there exists a path π of marked edges leading from an vertex in $\mathcal{I}_0^N$ to i, and π is the shortest such path, as we consider paths in increasing order of their lengths. Thus, as there are no shorter path of marked edges leading to i, the infection time of i is $|\pi|$. If a denotes the weight of the edge (i, j), then it is removed with probability $c(|\pi| + a)$, as in the construction from Section 7.2.1.

Our strategy to prove Theorem 7.6 is now the following. We will show that $\mathcal{G}^N$ converges, in some appropriate sense, to the random tree that has been described in

Section 7.1.3. We will then prove that the convergence of $\mathcal{G}^N$ in this sense implies that of σ_x^N to the dual σ^∞ of the delay equation. In the next section, we start by recalling a classical topology on graph that we will use.

7.3.2 Local topology on graphs

For $x \in [N]$, define $\mathcal{G}^N(x)$ as the subgraph induced by all the vertices y with an oriented path from y to x (including x itself). The graph $\mathcal{G}^N(x)$ is a pointed graph, with x as the reference vertex. Note that the edges of $\mathcal{G}^N(x)$ are endowed with weights as defined above. Finally, we distinguish between infected and susceptible vertices at time $t = 0$, by marking every infected vertices (that is, every $x \in \mathcal{I}_0^N$). Following the standard terminology of the literature, $\mathcal{G}^N(x)$ is a random element of $\mathscr{H}$, a pointed geometric oriented graph with marks. An element of $\mathscr{H}$ is characterized by four coordinates (V, E, w, m), respectively the set of vertices, the set of edges, w the weights on edges, $m \subseteq V$ the set of marked vertices. For $\mathcal{G}^N(x)$, note that $m = \mathcal{I}_0^N \cap V$.

We now equip $\mathscr{H}$ with a metric $d_{\mathscr{H}}$ so that $(\mathscr{H}, d_{\mathscr{H}})$ is a Polish space. A graph isomorphism φ between two *finite* rooted-marked graphs $G = (V, E, m)$ and $G' = (V', E', m')$ is a bijection from V to V' such that

(i) $(u, v) \in E$ iff $(u', v') \in E'$.

(ii) φ maps the reference vertex of G to the reference vertex in G'.

(iii) The map preserves the marking (i.e., m is mapped onto m').

By convention, we set $\min(\emptyset) = \infty$ in the following. Let $G_1 = (V_1, E_1, w_1, m_1)$, $G_2 = (V_2, E_2, w_2, m_2)$ be two finite elements of $\mathscr{H}$. Define

$$d(G_1, G_2) = \min\{1, \min_{\varphi} \max_{e:\, e \in E_1} |w_1(e) - w_2(\varphi(e))|\}$$

where the minimum is taken over all possible graph isomorphisms between the two graphs (in the sense prescribed above, that is, we only consider the isomorphisms preserving the root and the marking). In particular, if there is no isomorphism between G_1 and G_2, we set $d(G_1, G_2) = 1$.

For $\mathcal{G} \in \mathscr{H}$ and $y \in \mathcal{G}$, the topological distance to the reference vertex x is defined as

$$\inf\{n : \text{there exists a path } (y = x_1, \dots, x_n = x) \text{ in } \mathcal{G}\}.$$

For every $r \in \mathbb{N}^*$, we denote by $[\mathcal{G}]_r$, the subgraph induced by the vertices at a topological distance to the origin, that is, to the reference vertex, less than r. For two elements $\mathcal{G}_1, \mathcal{G}_2 \in \mathscr{H}$, we define the (pseudo-)distance $d_{\mathscr{H}}$ as follows

$$d_{\mathscr{H}}(\mathcal{G}_1, \mathcal{G}_2) \;=\; \sum_r 2^{-r} d([\mathcal{G}_1]_r, [\mathcal{G}_2]_r).$$

The metric $d_{\mathscr{H}}$ naturally induces a notion of local convergence on (equivalence classes of) $\mathscr{H}$. It can be proved that $(\mathscr{H}, d_{\mathscr{H}})$ is a Polish space.

7.3.3 A limiting random tree

Recall that we have define

$$\tau(a)\,\mathrm{d}a = \mathbb{E}\Big[\mathcal{P}(\mathrm{d}a)\Big], \quad R_0 = \int_0^\infty \tau(a)\,\mathrm{d}a.$$

Recall also that we have defined $\bar{\tau}$ so that for any f

$$\int_0^\infty f(s)\bar{\tau}(s)\,\mathrm{d}s = \int_0^\infty g(a)\int_a^\infty f(u-a)\tau(u)\,\mathrm{d}u\,\mathrm{d}a.$$

Note that $\bar{\tau}$ is the intensity measure of the shifted point process $\widehat{\mathcal{P}}_x$ of an infected individual $x \in \mathcal{I}_0^N$. Finally, recall the notation $\bar{R}_0$ for the total mass of $\bar{\tau}$ and the notation ν and $\bar{\nu}$ for the renormalization of τ and $\bar{\tau}$ to probability measures, respectively.

Recall the definition of the Poisson marked random tree of Section 7.1.3, which we denote by $\mathcal{H}$. The topological structure of $\mathcal{H}$ depends on the two positive real parameters S_0R_0, $(1-S_0)\bar{R}_0$, and the random weights are constructed from the two probability distribution ν, $\bar{\nu}$. The graph structure is given by a pointed-marked tree with Poisson offspring:

- Start from an unmarked root $\emptyset$.
- Unmarked nodes have independent Poisson(S_0R_0) unmarked offspring, and Poisson$((1-S_0)\bar{R}_0)$ marked offspring.
- Marked nodes have no offspring.

Edges of the tree are oriented towards the root. Every oriented edge (i,j) is assigned an independent age a_{ij}:

- If i is a marked node (and thus a leaf), a_{ij} is distributed according to $\bar{\nu}$.
- Otherwise, it is distributed according to ν.

The pair $(S_0R_0, (1-S_0)\bar{R}_0)$ will be referred to as the topological parameters of the random tree, whereas $(\nu, \bar{\nu})$ will be referred to as the weight parameters of the tree.

The random tree $\mathcal{H}$ corresponds to the local limit of the graph $\mathcal{G}^N(x)$. The birth time σ^∞ of the root is obtained by removing the edges of $\mathcal{H}$ that correspond to ineffective infections. Let us define σ^∞ as the random time obtained by applying Procedure 7.9 to $\mathcal{H}$. In order to apply Procedure 7.9, it is needed that the paths leading to the root can be ordered in increasing length. It is not *a priori* clear that this can be done for $\mathcal{H}$ since the tree is infinite, but it will follow from an argument that we give in the proof of Theorem 7.6. The following key result connects the distribution of σ^∞ to the delay equation.

Proposition 7.11. *Define*

$$\forall t \geq 0, \quad B(t) := S_0\mathbb{P}(\sigma^\infty \leq t).$$

Then B solves the delay equation (7.4).

Proof. As we have assumed that τ has a density w.r.t. the Lebesgue measure, it is clear that this also holds for the distribution of σ^∞. We denote its density by f. Let K, resp. $\bar{K}$, be the number of unmarked, resp. marked, children of the root of $\mathcal{H}$. Let $(\mathcal{H}_1, \dots, \mathcal{H}_N)$ denote the subtrees attached to the root $\varnothing$, and let $(\sigma_1^\infty, \dots, \sigma_K^\infty)$ be the birth times obtained by applying Procedure 7.9 to those subtrees. Moreover, let $(W_1, \dots, W_K)$ and $(\bar{W}_1, \dots, \bar{W}_{\bar{K}})$ be the weights of the edges starting from $\varnothing$ and leading to unmarked and marked children respectively. Conditional on these variables, let (B_i) and $(\bar{B}_i)$ be independent and such that

$$B_i \sim \text{Bernoulli}\Big(c(W_i + \sigma_i^\infty)\Big), \quad \bar{B}_i \sim \text{Bernoulli}\Big(c(\bar{W}_i)\Big).$$

When Procedure 7.9 is applied to a tree, it can be described recursively as follows. Apply Procedure 7.9 to all unmarked children of the focal vertex. This yields a vector $(\sigma_1^\infty, \dots, \sigma_K^\infty)$ of infection times for these children. Then, remove the edge leading to the i-th unmarked children with probability $c(\sigma_i^\infty + W_i)$, and that leading to the i-th marked children with probability $c(\bar{W}_i)$. The infection time of the focal individual is the minimum of the variables $(\sigma_i^\infty + W_i)$ and $(\bar{W}_i)$, for those i whose edge leading to the focal individuals has not been removed. In other words, the following holds

$$\sigma^\infty \overset{(\mathrm{d})}{=} \Big(\min_{1\le i\le K} \Big\{B_i(W_i + \sigma_i^\infty) + (1-B_i)\times\infty\Big\}\Big) \wedge \Big(\min_{1\le i\le \bar{K}} \Big\{\bar{B}_i\bar{W}_i + (1-\bar{B}_i)\times\infty\Big\}\Big),$$

with the convention $0\times\infty = 0$.

Define $G(t) = \mathbb{P}(\sigma^\infty > t)$. As by the branching property, conditional on K and $\bar{K}$, all previously introduced variables are independent, we have

$$\begin{aligned} G(t) &= \mathbb{E}\Big\{\Big(1 - \mathbb{E}\Big(c(\sigma^\infty + V)\mathbb{1}_{\{\sigma^\infty + V\le t\}}\Big)\Big)^K \Big(1 - \mathbb{E}\Big(c(\bar{V})\mathbb{1}_{\{\bar{V}\le t\}}\Big)\Big)^{\bar{K}}\Big\} \\ &= \mathbb{E}\Big\{\Big(1 - \int_0^t\int_0^{t-a} c(a+s)f(s)\,\mathrm{d}s\,\nu(\mathrm{d}a)\Big)^K \Big(1 - \int_0^t c(s)\,\bar{\nu}(\mathrm{d}s)\Big)^{\bar{K}}\Big\} \\ &= \exp\Big(- S_0 \int_0^t\int_0^{t-a} c(a+s)f(s)\tau(a)\,\mathrm{d}s\,\mathrm{d}a \\ &\qquad\qquad - I_0\int_0^t g(a)\int_a^\infty c(u-a)\tau(u)\,\mathrm{d}u\,\mathrm{d}a\Big), \end{aligned}$$

where, in the last equality, we have used the generating function of a Poisson distribution. It now follows that $B(t) = S_0(1 - G(t))$ satisfies (7.4). □

7.3.4 Convergence of the infection graph

Recall the infection graph $\mathcal{G}^N$ defined in Section 7.3.1, and the notation $\mathcal{X}_x = (\mathcal{P}_x, X_x)$. The following result proves the convergence of the local structure of $\mathcal{G}^N$ to the tree described in the previous section.

Proposition 7.12. *Let $x, y \notin \mathcal{I}_0^N$ with $x \neq y$. Then*

$$(\mathcal{G}^N(x), \mathcal{G}^N(y), \mathcal{X}_x, \mathcal{X}_y) \Longrightarrow \Big(\mathcal{G}, \mathcal{G}', \mathcal{X}, \mathcal{X}'\Big),$$

where all the limiting variables are independent; $\mathcal{X}$ and $\mathcal{X}'$ are identically distributed; $\mathcal{G}$ and $\mathcal{G}'$ are distributed as the aforementioned geometric Poisson marked tree with parameters $(S_0R_0, (1-S_0)\bar{R}_0)$ and $(\nu, \bar{\nu})$.

Proof. The proof is not so difficult with the right approach, but somewhat heavy in terms of notation. Consider $r \geq 1$ and T_1, T_2 two finite discrete, rooted, planar trees with height at most r and with marked subsets of their set of leaves. We separate $E_1 = E_1' \cup E_1''$, where E_1 is the set of edges of T_1, E_1'' is the set of edges that originate from a marked leaf, and $E_1' = E_1 \setminus E_1''$. Let B_1 be the set of unmarked vertices of T_1 that are at graph distance strictly less than r from the root. For any vertex $u \in B_1$, define n_u (resp. m_u) as the number of unmarked (resp. marked) offspring of u. Also, for any $e \in E_1$, let us fix a continuous bounded map $f_e \colon [0,\infty) \to [0,\infty)$. We define the analogous quantities $E_2 = E_2' \cup E_2''$ for T_2. Note that if $\mathcal{T}$ denotes a random Poisson marked tree with topological parameters $(S_0R_0, (1-S_0)\bar{R}_0)$ and weight parameters $(\nu, \bar{\nu})$, then the following functional F defined by

$$F(T_1, (f_e)_{e\in E_1}) \coloneqq \mathbb{E}\Bigg(\mathbb{1}([\mathcal{T}]_r = T_1) \prod_{e\in E_1} f_e(W_e)\Bigg),$$

where W_e denotes the weight of edge e in $\mathcal{T}$, can be explicitly computed. Indeed, by definition, it is immediate that

$$F(T_1, (f_e)_{e\in E_1}) = \prod_{u\in B_1} \Big(\mathrm{e}^{-S_0R_0-(1-S_0)\bar{R}_0} \frac{(S_0R_0)^{n_u}((1-S_0)\bar{R}_0)^{m_u}}{n_u! m_u!}\Big) \\ \prod_{e\in E_1'} \int f_e \,\mathrm{d}\nu \prod_{e\in E_1''} \int f_e \,\mathrm{d}\bar{\nu}.$$

Then, to rephrase the problem, we aim to show that for $x \neq y \in \mathcal{S}^N$, we have:

$$\mathbb{E}\Bigg(\mathbb{1}([G^N(x)]_r^{\circ,\mathrm{pl}} = T_1 \text{ and } [G^N(y)]_r^{\circ,\mathrm{pl}} = T_2)\Bigg[\prod_{e=(i,j)\in E_1\cup E_2} f_e(a_{ij})\Bigg] H_1(\mathcal{X}_x) H_2(\mathcal{X}_y)\Bigg) \\ \xrightarrow[N\to\infty]{} F(T_1, (f_e)_{e\in E_1}) F(T_2, (f_e)_{e\in E_2}) \mathbb{E}[H_1(\mathcal{X})]\mathbb{E}[H_2(\mathcal{X})],$$

where H_1 and H_2 and continuous bounded maps $\mathscr{X} \to \mathbb{R}$, and on the event that $[G^N(x)]_r$ is a tree, $[G^N(x)]_r^{\circ,\mathrm{pl}}$ denotes a random *planarization* of $[G^N(x)]_r$, built by forgetting the labels of $[G^N(x)]_r$, then choosing random uniform orders on the sets of children of all inner vertices.

To compute this, let us fix N large enough so that we can find a $[N]$-labelings of T_1 and T_2 in the following way. We choose some injective maps $\varphi_1 \colon T_1 \to [N]$ and $\varphi_2 \colon T_2 \to [N]$ such that

- for all $u \in T_1$, u is marked iff $\varphi_1(u) \in \mathcal{I}_0^N$,
- φ_1 maps the root of T_1 to x,

and where the analogous properties hold for φ_2. We further assume that the images of φ_1 and φ_2 are disjoint. Let us define $\tilde{T}_1$ and $\tilde{T}_2$ as the planar trees T_1 and T_2 whose vertices are labeled thanks to the corresponding maps φ_1 and φ_2. Let us identify the vertices $i \in \tilde{T}_1$ with their labels in $[N]$, and define for $i \in \tilde{T}_1$:

- $p(i)$ as the parent of i in $\tilde{T}_1$; if $i = x$, then for convenience let $p(i) := \dagger$ denote any element not in $[N]$. Note that the parent of i corresponds to an individual that i can infect.
- $e(i)$ as the edge $(i, p(i))$ in $\tilde{T}_1$.

We define the analogous $p(i), e(i)$ for $i \in \tilde{T}_2$. For $i \notin \tilde{T}_1 \cup \tilde{T}_2$, we define for convenience of notation $p(i) = e(i) = \dagger$. Recall the notation A_k^N for the set of ancestors of k in $\mathcal{G}^N$. Let us define the event

$$Q_i = \bigcap_{k \in B_1 \cup B_2 \setminus \{p(i)\}} \{i \notin A_k^N\}$$

where B_1, resp. B_2, denotes with an abuse of notation the set of unmarked vertices of $\tilde{T}_1$, resp. $\tilde{T}_2$, that are at graph distance strictly less than r. Finally, let $[G^N(x)]_r^{\mathrm{pl}}$ denote a uniform planarization of $[G^N(x)]_r$, where the vertices retain their original labels in $[N]$. With all these definitions, using the independence of $(\mathcal{P}_i;\, i \in [N])$ and the fact that

$$\Big\{[G^N(x)]_r^{\mathrm{pl}} = \tilde{T}_1 \text{ and } [G^N(y)]_r^{\mathrm{pl}} = \tilde{T}_2\Big\} \\ = \bigcap_{i \in (\tilde{T}_1 \cup \tilde{T}_2) \setminus \{x,y\}} \Big(\{i \in A_{p(i)}^N\} \cap Q_i\Big) \cap \bigcap_{i \notin (\tilde{T}_1 \cup \tilde{T}_2) \setminus \{x,y\}} Q_i,$$

it should be clear that

$$\mathbb{E}\Bigg(\mathbb{1}([G^N(x)]_r^{\mathrm{pl}} = \tilde{T}_1 \text{ and } [G^N(y)]_r^{\mathrm{pl}} = \tilde{T}_2)\Bigg[\prod_{e=(i,j) \in E_1 \cup E_2} f_e(a_{ij})\Bigg] H_1(\mathcal{X}_x) H_2(\mathcal{X}_y)\Bigg) \\ = \prod_{i \in (\tilde{T}_1 \cup \tilde{T}_2) \setminus \{x,y\}} \mathbb{E}\Big[f_{e(i)}(a_{ip(i)}) \mathbb{1}(\{i \in A_{p(i)}^N\} \cap Q_i)\Big] \times \prod_{i \notin (\tilde{T}_1 \cup \tilde{T}_2) \setminus \{x,y\}} \mathbb{P}(Q_i) \\ \times \mathbb{E}[H_1(\mathcal{X})] \mathbb{E}[H_2(\mathcal{X})] \times \prod_{u \in B_1} \frac{1}{n_u! m_u!} \prod_{u \in B_2} \frac{1}{n_u! m_u!}, \tag{7.6}$$

where the last term is simply the probability that the planarization of $[G^N(x)]_r$ and $[G^N(y)]_r$ matches that of $\tilde{T}_1$ and $\tilde{T}_2$.

Now let us compute each term separately. Recall the notation $\widehat{\mathcal{P}}_i$, which is the shifted infection point process for $i \in \mathcal{I}_0^N$, and the entire infection point process for $i \notin \mathcal{I}_0^N$. Starting from $i \notin (\tilde{T}_1 \cup \tilde{T}_2) \setminus \{x, y\}$, we have

$$\begin{aligned}\mathbb{P}(Q_i) &= \mathbb{E}\Big[\Big(1-\frac{n}{N}\Big)^{|\widehat{\mathcal{P}}_i|}\Big]\\ &= 1-\frac{n\mathbb{E}\big[|\widehat{\mathcal{P}}_i|\big]}{N}+\mathbb{E}\Big[\Big(1-\frac{n}{N}\Big)^{|\widehat{\mathcal{P}}_i|}-1+\frac{n|\widehat{\mathcal{P}}_i|}{N}\Big],\end{aligned}$$

with $n \coloneqq |B_1 \cup B_2|$. Noticing that $|1-(1-n/N)^{|\widehat{\mathcal{P}}_i|}| \leq n|\widehat{\mathcal{P}}_i|/N$, an application of dominated convergence shows that

$$\lim_{N\to\infty} N\mathbb{E}\Big[\Big(1-\frac{n}{N}\Big)^{|\widehat{\mathcal{P}}_i|}-1\Big] = n\mathbb{E}\big[|\widehat{\mathcal{P}}|_i\big]$$

so that

$$\mathbb{E}\Big[\Big(1-\frac{n}{N}\Big)^{|\widehat{\mathcal{P}}_i|}-1+\frac{n|\widehat{\mathcal{P}}_i|}{N}\Big] = o(N^{-1}). \tag{7.7}$$

Therefore, for $i \notin \mathcal{I}_0^N$, we have $\mathbb{P}(Q_i) = 1 - nR_0/N + o(N^{-1})$, and similarly, for $i \in \mathcal{I}_0^N$, we have $\mathbb{P}(Q_i) = 1 - n\bar{R}_0/N + o(N^{-1})$. This yields the following approximation for the product

$$\prod_{i\notin \tilde{T}_1\cup\tilde{T}_2\setminus\{x,y\}} \mathbb{P}(Q_i) = \mathbf{e}^{-n(S_0R_0+(1-S_0)\bar{R}_0)} + o(1).$$

For $i \in \tilde{T}_1 \cup \tilde{T}_2 \setminus \{x, y\}$, the argument is similar but slightly more complicated. First let us define $\tilde{Q}_i$ as the event

$$\tilde{Q}_i = \{i \text{ has no multiple edges to } p(i)\}.$$

We have

$$\mathbb{P}\Big(\{i \in A^N_{p(i)}\} \setminus \tilde{Q}_i\Big) = \mathbb{E}\Big[1-\Big(1-\frac{1}{N}\Big)^{|\widehat{\mathcal{P}}_i|} - |\widehat{\mathcal{P}}_i|\frac{1}{N}\Big(1-\frac{1}{N}\Big)^{|\widehat{\mathcal{P}}_i|-1}\Big].$$

Dominated convergence shows that

$$\lim_{N\to\infty} N\mathbb{E}\Big[|\widehat{\mathcal{P}}_i|\frac{1}{N}\Big(1-\frac{1}{N}\Big)^{|\widehat{\mathcal{P}}_i|-1}\Big] = \mathbb{E}\big[|\widehat{\mathcal{P}}_i|\big]$$

and combined with (7.7) with $n = 1$, this proves that

$$\mathbb{P}\Big(\{i \in A^N_{p(i)}\} \setminus \tilde{Q}_i\Big) = o(N^{-1})$$

Therefore if we show that

$$\mathbb{E}\Big[f_{e(i)}(a_{ip(i)})\mathbb{1}(\tilde{Q}_i \cap Q_i)\Big] = \frac{1}{N}C + o(N^{-1})$$

for some constant C, then we can simply plug this expression into (7.6). To show this, assume that $i \notin \mathcal{I}_0^N$ and check that by definition, the following holds:

$$\begin{aligned}\mathbb{E}\Big[f_{e(i)}(a_{ip(i)})\mathbb{1}(\{i \in A_{p(i)}^N\} \cap Q_i \cap \tilde{Q}_i\Big] \\ &= \mathbb{E}\bigg[\frac{1}{N}\Big(1 - \frac{n-1}{N}\Big)^{|\mathcal{P}_i|-1} \int f_{e(i)}(a)\,\mathrm{d}\mathcal{P}_i(\mathrm{d}a)\bigg] \\ &= \frac{1}{N}\int f_{e(i)}\,\mathrm{d}\tau + o(N^{-1}),\end{aligned}$$

where in the $o(N^{-1})$ terms, there are constants depending on the trees T_1 and T_2 and the maps (f_e). It is easy to check in the same way that if $i \in \mathcal{I}_0^N$, the same formula holds, with $\bar{\tau}$ instead of τ. Letting $n' := |T_1| + |T_2|$ and putting everything together into (7.6), we get

$$\begin{aligned}\mathbb{E}\bigg(\mathbb{1}([G^N(x)]_r^{\mathrm{pl}} = \tilde{T}_1 \text{ and } [G^N(y)]_r^{\mathrm{pl}} = \tilde{T}_2)\bigg[\prod_{e=(i,j)\in E_1\cup E_2} f_e(a_{ij})\bigg] H_1(\mathcal{X}_x)H_2(\mathcal{X}_y)\bigg) \\ = \frac{1}{N^{n'-2}}(\mathbf{e}^{-n(S_0R_0+(1-S_0)\bar{R}_0)}) \prod_{e\in E_1'\cup E_2'} R_0 \int f_e\,\mathrm{d}\nu \prod_{e\in E_1''\cup E_2''} \bar{R}_0 \int f_e\,\mathrm{d}\bar{\nu} \\ \times \mathbb{E}[H_1(\mathcal{X})]\mathbb{E}[H_2(\mathcal{X})] \times \prod_{u\in B_1} \frac{1}{n_u!m_u!} \prod_{u\in B_2} \frac{1}{n_u!m_u!} + o(N^{2-n'}).\end{aligned}$$

Notice that this approximation does not depend on the choice of the labelings φ_1 and φ_2. The number of unmarked, resp. marked, vertices is equivalent to S_0N, resp. $(1-S_0)N$, so that the number of maps φ_1 and φ_2 compatible with our assumptions is equivalent to

$$(S_0N)^{|E_1'|+|E_2'|}((1-S_0)N)^{|E_1''|+|E_2''|} = S_0^{|E_1'|+|E_2'|}(1-S_0)^{|E_1''|+|E_2''|}N^{n'-2},$$

where $n'-2$ is the number of edges in T_1 and T_2. It is readily checked that we get the correct approximation

$$\begin{aligned}\mathbb{E}\bigg(\mathbb{1}([G^N(x)]_r^{\mathrm{pl}} = \tilde{T}_1 \text{ and } [G^N(y)]_r^{\mathrm{pl}} = \tilde{T}_2)\bigg[\prod_{e=(i,j)\in E_1\cup E_2} f_e(a_{ij})\bigg] H_1(\mathcal{X}_x)H_2(\mathcal{X}_y)\bigg) \\ = F(T_1,(f_e)_{e\in E_1})F(T_2,(f_e)_{e\in E_2})\mathbb{E}[H_1(\mathcal{X})]\mathbb{E}[H_2(\mathcal{X})] + o(1),\end{aligned}$$

which ends the proof. □

7.3.5 Proof of Theorem 7.6

Corollary 7.13. *Let $x, y \notin \mathcal{I}_0^N$ with $x \neq y$, then*

$$(\sigma_x^N, \sigma_y^N, \mathcal{X}_x, \mathcal{X}_y) \Longrightarrow (\sigma^\infty, \tilde{\sigma}^\infty, \mathcal{X}, \mathcal{X}'),$$

where σ^∞ (resp., $\tilde{\sigma}^\infty$) are defined from $\mathcal{G}$ and $\mathcal{G}'$ (the limiting random variables in Proposition 7.12) according to Procedure 7.9.

Proof. It is sufficient to show that the distribution of σ_x^N, obtained from $\mathcal{G}^N$ out of Procedure 7.9, converges to that of σ^∞, obtained from Procedure 7.9 applied to the Poisson limiting tree $\mathcal{H}$. Up to using Skorohod's representation theorem, see Theorem 6.7 in [25], we might assume that $\mathcal{G}^N(x)$ converges a.s. to $\mathcal{H}$ is the topology defined in Section 7.3.2. In what follows, we work conditional on $\mathcal{G}^N$ and $\mathcal{H}$ and consider them as deterministic. It is now sufficient to prove that

$$\mathbb{P}\Big(\sigma_x^N \geq t\Big) \longrightarrow \mathbb{P}\Big(\sigma^\infty \geq t\Big).$$

First, let us show that the paths from a marked leaf to the root in $\mathcal{H}$ can be a.s. ordered in increasing order of their length. To see this, let $(Z_r;\, r \geq 0)$ be the process that records the ages of the unmarked vertices of $\mathcal{H}$, defined as

$$Z_r \coloneqq \sum_{\substack{u \in \mathcal{H},\, d(u,\emptyset)=r \\ u \text{ unmarked}}} \delta(|\pi_u|),$$

where π_u is the unique path connecting u to the root $\emptyset$. It is clear that $(Z_r)_{r\geq 0}$ is a branching random walk with Poisson offspring distribution, and it follows from general results that, conditional on non-extinction, its minimum drifts to ∞, see for instance Theorem 5.12 in [199]. As $\mathcal{H}$ is obtained by attaching independently to any unmarked vertex a Poisson$(I_0 \bar{R}_0)$ distributed number of marked leaves, this also shows that

$$\lim_{r\to\infty} \min_{\substack{u \in \mathcal{H} \\ d(u,\emptyset) > r}} |\pi_u| = \infty \tag{7.8}$$

where π_u is the unique path from u to the root. As there are only finitely many marked vertices in $[\mathcal{H}]_r$, they can be ordered such that their length is increasing.

Moreover, (7.8) also entails that, for a.e. realization of $\mathcal{H}$, we can find a large enough r such that for all $i \notin [\mathcal{H}]_r$, the path from i to the root has length larger that t. For N large enough, $[\mathcal{G}^N(x)]_r$ is isomorphic to $[\mathcal{H}]_r$, and the weights of the edges of $[\mathcal{G}^N(x)]_r$ converge to those of $[\mathcal{H}]_r$. Let S (resp. S^N) denote the number of steps before Procedure 7.9 applied to $\mathcal{H}$ (resp. $\mathcal{G}^N$) stops, and M (resp. M^N) the number of marked vertices in $[\mathcal{H}]_r$ (resp. $[\mathcal{G}^N]_r$). It should be clear that on $\{S > M\}$, we have $\sigma^\infty > t$, as all paths considered after step M have a length larger than t. It should also be clear that

$$\mathbb{P}\Big(\sigma_x^N \leq t;\, ^N \leq M^N\Big) \longrightarrow \mathbb{P}\Big(\sigma^\infty \leq t;\, \leq M\Big)$$

as this event can be expressed in terms of the tree topology of $[\mathcal{G}^N(x)]_r$ and the weights of the edges of $[\mathcal{G}^N(x)]_r$. (It is the probability that, in Procedure 7.9, we find a marked path from a marked vertex in $[\mathcal{G}^N(x)]_r$ to the root.) Therefore, this shows that

$$\mathbb{P}\Big(\sigma_x^N \leq t\Big) \longrightarrow \mathbb{P}\Big(\sigma^\infty \leq t\Big)$$

and ends the proof. □

We are now ready to prove Theorem 7.6.

Proof of Theorem 7.6. For the convergence of λ^N, note that as the ages of the initial infecteds are i.i.d. with distribution g, it follows from

$$\frac{|\mathcal{I}_0^N|}{N} \longrightarrow I_0$$

that

$$\mathbb{1}_{\sigma<0}\lambda^N(\mathrm{d}\sigma, \mathrm{d}\xi) \longrightarrow I_0 g(-\sigma)\,\mathrm{d}\sigma \otimes \mathbb{P}(\mathcal{X} \in \cdot).$$

Therefore it remains only to show that

$$\mathbb{1}_{\sigma\geq 0}\lambda^N(\mathrm{d}\sigma, \mathrm{d}\xi) \longrightarrow b(\sigma)\,\mathrm{d}\sigma \otimes \mathbb{P}(\mathcal{X} \in \cdot).$$

Since the limiting measure $b(\sigma)\,\mathrm{d}\sigma \otimes \mathbb{P}(\mathcal{X} \in \cdot)$ is deterministic, it is sufficient to check convergence in distribution of

$$\int \mathbb{1}_{\sigma\geq 0} f(\sigma)g(\xi)\,\lambda^N(\mathrm{d}\sigma, \mathrm{d}\xi) = \frac{1}{N}\sum_{x\in\mathcal{S}^N} f(\sigma_x^N)g(\mathcal{X}_x),$$

for any continuous bounded maps $f\colon \mathbb{R} \to \mathbb{R}$ and $g\colon \mathscr{X} \to \mathbb{R}$ (see for instance [121, Theorem 4.11]). First, taking the expectation in the previous display and using Corollary 7.13, we obtain

$$\Big(1 - \frac{|\mathcal{I}_0^N|}{N}\Big)\mathbb{E}[f(\sigma_1^N)]\mathbb{E}[g(\mathcal{X})] \xrightarrow[N\to\infty]{} S_0\mathbb{E}[f(\sigma^\infty)]\mathbb{E}[g(\mathcal{X})].$$

Now by definition of B, we can write

$$S_0\mathbb{E}[f(\sigma^\infty)] = \int f(\sigma)\,\mathrm{d}B(\sigma) = \int f(t)b(t)\,\mathrm{d}t.$$

Furthermore we have

$$\begin{aligned}\mathbb{E}\bigg(\frac{1}{N}\sum_{x\in\mathcal{S}^N} f(\sigma_x^N)g(\mathcal{X}_x)\bigg)^2 \\ &= \frac{1}{N^2}\sum_{x\notin\mathcal{I}_0^N}\mathbb{E}(f(\sigma_x^N)^2 g(\mathcal{X})^2] + \frac{1}{N^2}\sum_{x\neq y}\mathbb{E}\Big[f(\sigma_x^N)f(\sigma_y^N)g(\mathcal{X}_x)g(\mathcal{X}_y)\Big] \\ &\longrightarrow (S_0\mathbb{E}[f(\sigma^\infty)]\mathbb{E}[g(\mathcal{X})])^2\end{aligned}$$

again by Corollary 7.13. This concludes the proof. □

References for Chapter 7

[8] J. Baker, P. Chigansky, K. Hamza, and F. C. Klebaner. Persistence of small noise and random initial conditions. *Advances in Applied Probability* **50** (2018), 67–81.

[11] A. D. Barbour, P. Chigansky, and F. C. Klebaner. On the emergence of random initial conditions in fluid limits. *Journal of Applied Probability* **53** (2016), 1193–1205.

[12] A. D. Barbour and G. Reinert. Approximating the epidemic curve. *Electronic Journal of Probability* **18** (2013), 30 pp.

[25] P. Billingsley. *Convergence of Probability Measures*. Second edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1999.

[31] F. Brauer. The Kermack–McKendrick epidemic model revisited. *Mathematical Biosciences* **198** (2005), 119–131.

[32] F. Brauer and C. Castillo-Chavez. *Mathematical Models in Population Biology and Epidemiology*. Texts in Applied Mathematics. Springer, New Yord, 2012.

[35] T. Britton and É. Pardoux. *Stochastic epidemic models with inference*. Mathematical Biosciences Subseries. Springer International Publishing, 2019.

[40] A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology* **178** (2013), 1505–1512.

[45] O. Diekmann. Limiting behaviour in an epidemic model. *Nonlinear Analysis: Theory, Methods & Applications* **1** (1977), 459–470.

[65] J. Y. Fan, K. Hamza, P. Jagers, and F. C. Klebaner. Convergence of the age structure of general schemes of population processes. *Bernoulli* **26** (2020), 893–926.

[68] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, and C. Fraser. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368** (2020).

[73] R. Forien, G. Pang, and É. Pardoux. Epidemic models with varying infectivity (2020). arXiv: `2006.15377`.

[74] R. Forien, G. Pang, and É. Pardoux. Estimating the state of the COVID-19 epidemic in France using a non-Markovian model. *medRxiv* (2020).

[77] F. Foutel-Rodier, F. Blanquart, P. Courau, P. Czuppon, J.-J. Duchamps, J. Gamblin, É. Kerdoncuff, R. Kulathinal, L. Régnier, L. Vuduc, A. Lambert, and E. Schertzer. From individual-based epidemic models to McKendrick-von Foerster PDEs: A guide to modeling and inferring COVID-19 dynamics (2020). arXiv: `2007.09622`.

[81] C. Fraser. Estimating individual and household reproduction numbers in an emerging epidemic. *PLOS ONE* **2** (2007), 1–12.

[101] K. Hamza, P. Jagers, and F. C. Klebaner. The age structure of population-dependent general branching processes in environments with a high carrying capacity. *Proceedings of the Steklov Institute of Mathematics* **282** (2013), 90–105.

[111] J. Jacod and A. N. Shiryaev. *Limit Theorems for Stochastic Processes.* Second edition. Grundlehren Der Mathematischen Wissenschaften. Springer-Verlag, 2003.

[112] P. Jagers. *Branching processes with biological applications.* Wiley, 1975.

[113] P. Jagers and F. C. Klebaner. Population-size-dependent and age-dependent branching processes. *Stochastic Processes and their Applications* **87** (2000), 235–254.

[114] P. Jagers and F. C. Klebaner. Population-size-dependent, age-structured branching processes linger around their carrying capacity. *Journal of Applied Probability* **48** (2011), 249–260.

[121] O. Kallenberg. *Random Measures, Theory and Applications.* Probability Theory and Stochastic Modelling. Springer, Cham, 2017.

[125] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **115** (1927), 700–721.

[163] O. Nerman and P. Jagers. The stable doubly infinite pedigree process of supercritical branching populations. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **65** (1984), 445–460.

[166] G. Pang and É. Pardoux. Functional limit theorems for non-Markovian epidemic models (2020). arXiv: 2003.03249.

[167] G. Pang and É. Pardoux. Multi-patch epidemic models with general infectious periods (2020). arXiv: 2006.14412.

[199] Z. Shi. *Branching Random walks. École d'Été de Probabilités de Saint-Flour XLII-2012.* Vol. 2151. Lecture Notes in Mathematics. Springer, Cham, 2015.

[205] Z. Taïb. *Branching Processes and Neutral Evolution.* Vol. 93. Lecture Notes in Biomathematics. Springer Berlin Heidelberg, 1992.

[217] J. Wallinga and M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences* **274** (2007), 599–604.

References

Contribution

[23] F. Bienvenu, J.-J. Duchamps, and F. Foutel-Rodier. The Moran forest (2020). arXiv: 1906.08806.

[77] F. Foutel-Rodier, F. Blanquart, P. Courau, P. Czuppon, J.-J. Duchamps, J. Gamblin, É. Kerdoncuff, R. Kulathinal, L. Régnier, L. Vuduc, A. Lambert, and E. Schertzer. From individual-based epidemic models to McKendrick-von Foerster PDEs: A guide to modeling and inferring COVID-19 dynamics (2020). arXiv: 2007.09622.

[78] F. Foutel-Rodier and A. M. Etheridge. The spatial Muller's ratchet: Surfing of deleterious mutations during range expansion. *Theoretical Population Biology* **135** (2020), 19–31.

[79] F. Foutel-Rodier, A. Lambert, and E. Schertzer. Exchangeable coalescents, ultrametric spaces, nested interval-partitions: A unifying approach (2019). arXiv: 1807.05165.

[80] F. Foutel-Rodier, A. Lambert, and E. Schertzer. Kingman's coalescent with erosion. *Electronic Journal of Probability* **25** (2020), 33 pp.

Complete bibliography

[1] D. Aldous and L. Popovic. A critical branching process model for biodiversity. *Advances in Applied Probability* **37** (2005), 1094–1115.

[2] D. G. Aronson and H. F. Weinberger. Nonlinear diffusion in population genetics, combustion, and nerve pulse propagation. *Partial Differential Equations and Related Topics*. Springer Berlin Heidelberg, 1975, 5–49.

[3] K. B. Athreya and P. E. Ney. *Branching Processes*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag Berlin Heidelberg, 1972.

[4] F. Baccelli, B. Błaszczyszyn, and M. Karray. *Random Measures, Point Processes, and Stochastic Geometry*. Inria, 2020.

[5] J. A. Backer, D. Klinkenberg, and J. Wallinga. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20-28 January 2020. *Eurosurveillance* **25** (2020).

[6] Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, and M. Wang. Presumed asymptomatic carrier transmission of COVID-19. *JAMA* **323** (2020), 1406–1407.

[7] S. J. E. Baird, N. H. Barton, and A. M. Etheridge. The distribution of surviving blocks of an ancestral genome. *Theoretical Population Biology* **64** (2003), 451–471.

[8] J. Baker, P. Chigansky, K. Hamza, and F. C. Klebaner. Persistence of small noise and random initial conditions. *Advances in Applied Probability* **50** (2018), 67–81.

[9] F. Ball. A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Advances in Applied Probability* **18** (1986), 289–310.

[10] A. D. Barbour. The duration of the closed stochastic epidemic. *Biometrika* **62** (1975), 477–482.

[11] A. D. Barbour, P. Chigansky, and F. C. Klebaner. On the emergence of random initial conditions in fluid limits. *Journal of Applied Probability* **53** (2016), 1193–1205.

[12] A. D. Barbour and G. Reinert. Approximating the epidemic curve. *Electronic Journal of Probability* **18** (2013), 30 pp.

[13] N. H. Barton. The dynamics of hybrid zones. *Heredity* **43** (1979), 341–359.

[14] N. H. Barton, F. Depaulis, and A. M. Etheridge. Neutral evolution in spatially continuous populations. *Theoretical Population Biology* **61** (2002), 31–48.

[15] N. H. Barton and A. M. Etheridge. The relation between reproductive value and genetic contribution. *Genetics* **188** (2011), 953–973.

[16] N. H. Barton and M. Turelli. Spatial waves of advance with bistable dynamics: Cytoplasmic and genetic analogues of allee Effects. *The American Naturalist* **178** (2011), E48–E75.

[17] R. Bauerfeind, A. von Graevenitz, P. Kimmig, H. G. Schiefer, T. Schwarz, W. Slenczka, and H. Zahner. *Zoonoses: infectious diseases transmissible from animals to humans.* Fourth edition. ASM Books. John Wiley & Sons, Ltd, 2015.

[18] J. Berestycki. Exchangeable fragmentation-coalescence processes and their equilibrium measures. *Electronic Journal of Probability* **9** (2004), 770–824.

[19] J. Berestycki and N. Berestycki. Kingman's coalescent and Brownian motion. *ALEA, Latin American Journal of Probability and Mathematical Statistics* **6** (2009), 239–259.

[20] J. Bertoin. *Random Fragmentation and Coagulation Processes.* Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2006.

[21] J. Bertoin and J.-F. Le Gall. Stochastic flows associated to coalescent processes. *Probability Theory and Related Fields* **126** (2003), 261–288.

[22] Q. Bi, Y. Wu, S. Mei, C. Ye, X. Zou, Z. Zhang, X. Liu, L. Wei, S. A. Truelove, T. Zhang, W. Gao, C. Cheng, X. Tang, X. Wu, Y. Wu, B. Sun, S. Huang, Y. Sun, J. Zhang, T. Ma, J. Lessler, and T. Feng. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *The Lancet Infectious Diseases* **20** (2020), 911–919.

[23] F. Bienvenu, J.-J. Duchamps, and F. Foutel-Rodier. The Moran forest (2020). arXiv: `1906.08806`.

[24] P. Billingsley. *Probability and Measures.* Third edition. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1995.

[25] P. Billingsley. *Convergence of Probability Measures.* Second edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1999.

[26] M. Birkner, J. Blath, M. Capaldo, A. M. Etheridge, M. Möhle, J. Schweinsberg, and A. Wakolbinger. Alpha-stable branching and Beta-coalescents. *Electronic Journal of Probability* **10** (2005), 303–325.

[27] G. Birzu, O. Hallatschek, and K. S. Korolev. Fluctuations uncover a distinct class of traveling waves. *Proceedings of the National Academy of Sciences* **115** (2018), E3645–E3654.

[28] G. Birzu, S. Matin, O. Hallatschek, and K. S. Korolev. Genetic drift in range expansions is very sensitive to density feedback in dispersal and growth (2019). arXiv: 1903.11627.

[29] L. Bosshard, I. Dupanloup, O. Tenaillon, R. Bruggmann, M. Ackermann, S. Peischl, and L. Excoffier. Accumulation of deleterious mutations during bacterial range expansions. *Genetics* **207** (2017), 669–684.

[30] D. S. Boukal and L. Berec. Single-species models of the Allee effect: Extinction boundaries, sex ratios and mate encounters. *Journal of Theoretical Biology* **218** (2002), 375–394.

[31] F. Brauer. The Kermack–McKendrick epidemic model revisited. *Mathematical Biosciences* **198** (2005), 119–131.

[32] F. Brauer and C. Castillo-Chavez. *Mathematical Models in Population Biology and Epidemiology*. Texts in Applied Mathematics. Springer, New Yord, 2012.

[33] T. Britton. Epidemics in heterogeneous communities: estimation of R_0 and secure vaccination coverage. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63** (2001), 705–715.

[34] T. Britton, F. Ball, and P. Trapman. The disease-induced herd immunity level for COVID-19 is substantially lower than the classical herd immunity level (2020). arXiv: 2005.03085.

[35] T. Britton and É. Pardoux. *Stochastic epidemic models with inference*. Mathematical Biosciences Subseries. Springer International Publishing, 2019.

[36] C. Cannings. The Latent Roots of Certain Markov Chains Arising in Genetics: A New Approach, I. Haploid Models. *Advances in Applied Probability* **6** (1974), 260–290.

[37] J. T. Chang. Recent common ancestors of all present-day individuals. *Advances in Applied Probability* **31** (1999), 1002–1026.

[38] N. H. Chapman and E. A. Thompson. The effect of population history on the lengths of ancestral chromosome segments. *Genetics* **162** (2002), 449–458.

[39] H. Cohn. Multitype finite mean supercritical age-dependent branching processes. *Journal of applied probability* **26** (1989), 398–403.

[40] A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology* **178** (2013), 1505–1512.

[41] J. A. Coyne and H. A. Orr. Patterns of speciation in Drosophila. *Evolution* **43** (1989), 362–381.

[42] D. H. Crawford. *Deadly Companions: How Microbes Shaped our History.* Second edition. Oxford University Press, 2018.

[43] A. Depperschmidt and A. Greven. Tree-valued Feller diffusion (2019). arXiv: `1904.02044`.

[44] A. Depperschmidt, A. Greven, and P. Pfaffelhuber. Marked metric measure spaces. *Electronic Communications in Probability* **16** (2011), 174–188.

[45] O. Diekmann. Limiting behaviour in an epidemic model. *Nonlinear Analysis: Theory, Methods & Applications* **1** (1977), 459–470.

[46] R. Djidjou-Demasse, Y. Michalakis, M. Choisy, M. T. Sofonea, and S. Alizon. Optimal COVID-19 epidemic control until vaccine deployment. *medRxiv* (2020).

[47] R. Do, D. Balick, H. Li, I. Adzhubei, S. Sunyaev, and D. Reich. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nature Genetics* **47** (2015), 126–131.

[48] P. Donnelly and P. Joyce. Consistent ordered sampling distributions: Characterization and convergence. *Advances in Applied Probability* **23** (1991), 229–258.

[49] P. Donnelly and T. G. Kurtz. Particle representations for measure-valued population models. *Annals of Probability* **27** (1999), 166–205.

[50] T. Duquesne and C. Labbé. On the Eve property for CSBP. *Electronic Journal of Probability* **19** (2014), 31 pp.

[51] T. Duquesne and J.-F. Le Gall. *Random Trees, Lévy Processes and Spatial Branching Processes.* Astérisque, 2002.

[52] R. Durrett. *Probability Models for DNA Sequence Evolution.* Second edition. Probability and its Applications. Springer, New York, NY, 2008.

[53] R. Durrett and W.-T. Fan. Genealogies in expanding populations. *The Annals of Applied Probability* **26** (2016), 3456–3490.

[54] C. A. Edmonds, A. S. Lillie, and L. L. Cavalli-Sforza. Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences* **101** (2004), 975–979.

[55] A. M. Etheridge. *An Introduction to Superprocesses.* Vol. 20. University Lecture Series. American Mathematical Society, 2000.

[56] A. M. Etheridge. *Some Mathematical Models from Population Genetics. École d'Été de Probabilités de Saint-Flour XXXIX-2009.* Vol. 2012. Lecture Notes in Mathematics. Springer Science & Business Media, 2011.

[57] A. M. Etheridge, P. Pfaffelhuber, and A. Wakolbinger. How often does the ratchet click? Facts, heuristics, asymptotics. London Mathematical Society Lecture Note Series. Cambridge University Press, 2009, 365–390.

[58] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1986.

[59] S. Evans. *Probability and Real Trees. École d'Été de Probabilités de Saint-Flour XXXV-2005*. Vol. 1920. Lecture Notes in Mathematics. Springer, Berlin, Heidelberg, 2007.

[60] T. Evgeniou, M. Fekom, A. Ovchinnikov, R. Porcher, C. Pouchol, and N. Vayatis. Epidemic models for personalised COVID-19 isolation and exit policies using clinical risk predictions. *medRxiv* (2020).

[61] W. J. Ewens. *Mathematical Population Genetics. I. Theoretical Introduction*. Second edition. Interdisciplinary Applied Mathematics. Springer, New York, NY, 2004.

[62] L. Excoffier, M. Foll, and R. J. Petit. Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics* **40** (2009), 481–501.

[63] L. Excoffier and N. Ray. Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution* **23** (2008), 347–351.

[64] A. Eyre-Walker and P. D. Keightley. The distribution of fitness effects of new mutations. *Nature Reviews Genetics* **8** (2007), 610–618.

[65] J. Y. Fan, K. Hamza, P. Jagers, and F. C. Klebaner. Convergence of the age structure of general schemes of population processes. *Bernoulli* **26** (2020), 893–926.

[66] J. Fayard, É. K. Klein, and F. Lefèvre. Long distance dispersal and the fate of a gene from the colonization front. *Journal of Evolutionary Biology* **22** (2009), 2171–2182.

[67] J. Felsenstein. The evolutionary advantage of recombination. *Genetics* **78** (1974), 737–756.

[68] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, and C. Fraser. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368** (2020).

[69] R. Ferriere and V. C. Tran. Stochastic and deterministic models for age-structured populations with genetically variable traits. *ESAIM: Proceedings* **27** (2009), 289–310.

[70] R. A. Fisher. The Wave of advance of advantageous genes. *Annals of Eugenics* **7** (1937), 355–369.

[71] R. A. Fisher. A fuller theory of "junctions" in inbreeding. *Heredity* **8** (1954), 187–197.

[72] S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J. W. Eaton, M. Monod, P. N. Perez-Guzman, N. Schmit, L. Cilloni, K. E. C. Ainslie, M. Baguelin, A. Boonyasiri, O. Boyd, L. Cattarino, L. V. Cooper, Z. Cucunubá, G. Cuomo-Dannenburg, A. Dighe, B. Djaafara, I. Dorigatti, S. L. van Elsland, R. G. FitzJohn, K. A. M. Gaythorpe, L. Geidelberg, N. C. Grassly, W. D. Green, T. Hallett, A. Hamlet, W. Hinsley, B. Jeffrey, E. Knock, D. J. Laydon, G. Nedjati-Gilani, P. Nouvellet, K. V. Parag, I. Siveroni, H. A. Thompson, R. Verity, E. Volz, C. E. Walters, H. Wang, Y. Wang, O. J. Watson, P. Winskill, X. Xi, P. G. T. Walker, A. C. Ghani, C. A. Donnelly, S. Riley, M. A. C. Vollmer, N. M. Ferguson, L. C. Okell, S. Bhatt, and Imperial College COVID-19 Response Team. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584** (2020), 257–261.

[73] R. Forien, G. Pang, and É. Pardoux. Epidemic models with varying infectivity (2020). arXiv: 2006.15377.

[74] R. Forien, G. Pang, and É. Pardoux. Estimating the state of the COVID-19 epidemic in France using a non-Markovian model. *medRxiv* (2020).

[75] N. Forman. Exchangeable hierarchies and mass-structure of weighted real trees. *Electronic Journal of Probability* **25** (2020), 28 pp.

[76] N. Forman, C. Haulk, and J. Pitman. A representation of exchangeable hierarchies by sampling from random real trees. *Probability Theory and Related Fields* **172** (2018), 1–29.

[77] F. Foutel-Rodier, F. Blanquart, P. Courau, P. Czuppon, J.-J. Duchamps, J. Gamblin, É. Kerdoncuff, R. Kulathinal, L. Régnier, L. Vuduc, A. Lambert, and E. Schertzer. From individual-based epidemic models to McKendrick-von Foerster PDEs: A guide to modeling and inferring COVID-19 dynamics (2020). arXiv: 2007.09622.

[78] F. Foutel-Rodier and A. M. Etheridge. The spatial Muller's ratchet: Surfing of deleterious mutations during range expansion. *Theoretical Population Biology* **135** (2020), 19–31.

[79] F. Foutel-Rodier, A. Lambert, and E. Schertzer. Exchangeable coalescents, ultrametric spaces, nested interval-partitions: A unifying approach (2019). arXiv: 1807.05165.

[80] F. Foutel-Rodier, A. Lambert, and E. Schertzer. Kingman's coalescent with erosion. *Electronic Journal of Probability* **25** (2020), 33 pp.

[81] C. Fraser. Estimating individual and household reproduction numbers in an emerging epidemic. *PLOS ONE* **2** (2007), 1–12.

[82] D. H. Fremlin. Real-valued-measurable cardinals. *Set Theory of the Reals, Isreal Mathematical Conference Proceedings*. Vol. 6. 1993, 151–304.

[83] J. Garnier, T. Giletti, F. Hamel, and L. Roques. Inside dynamics of pulled and pushed fronts. *Journal de Mathématiques Pures et Appliquées* **98** (2012), 428–449.

[84] K. J. Gilbert, S. Peischl, and L. Excoffier. Mutation load dynamics during environmentally-driven range shifts. *PLOS Genetics* **14** (2018), 1–18.

[85] K. J. Gilbert, N. P. Sharp, A. L. Angert, G. L. Conte, J. A. Draghi, F. Guillaume, A. L. Hargreaves, R. Matthey-Doret, and M. C. Whitlock. Local adaptation interacts with expansion load during range expansion: Maladaptation reduces expansion load. *The American Naturalist* **189** (2017), 368–380.

[86] A. Gnedin. The representation of composition structures. *The Annals of Probability* **25** (1997), 1437–1450.

[87] S. C. González-Martínez, K. Ridout, and J. R. Pannell. Range expansion compromises adaptive evolution in an outcrossing plant. *Current Biology* **27** (2017), 2544–2551.e4.

[88] E. Graciá, F. Botella, J. D. Anadón, P. Edelaar, D. J. Harris, and A. Giménez. Surfing in tortoises? Empirical signs of genetic structuring owing to range expansion. *Biology Letters* **9** (2013), 20121091.

[89] M. Gralka, F. Stiewe, F. Farrell, W. Möbius, B. Waclaw, and O. Hallatschek. Allele surfing promotes microbial adaptation from standing variation. *Ecology Letters* **19** (2016), 889–898.

[90] B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303** (2004), 327–332.

[91] A. Greven, P. Pfaffelhuber, and A. Winter. Convergence in distribution of random metric measure spaces (Λ-coalescent measure trees). *Probability Theory and Related Fields* **145** (2009), 285–322.

[92] A. Greven, P. Pfaffelhuber, and A. Winter. Tree-valued resampling dynamics: Martingale problems and applications. *Probability Theory and Related Fields* **155** (2012), 899–838.

[93] R. C. Griffiths and P. Marjoram. An ancestral recombination graph. *Progress in population genetics and human evolution*. Springer, 1997, 257–270.

[94] M. Gromov. *Metric structures for Riemannian and non-Riemannian spaces*. Vol. 152. Progress in Mathematics. Birkhäuser Boston, 1999.

[95] S. Gufler. A representation for exchangeable coalescent trees and generalized tree-valued Fleming-Viot processes. *Electronic Journal of Probability* **23** (2018), 42 pp.

[96] K.-P. Hadeler and F. Rothe. Travelling fronts in nonlinear diffusion equations. *Journal of Mathematical Biology* **2** (1975), 251–263.

[97] J. Haigh. The accumulation of deleterious genes in a population—Muller's Ratchet. *Theoretical Population Biology* **14** (1978), 251–267.

[98] O. Hallatschek, P. Hersen, S. Ramanathan, and D. R. Nelson. Genetic drift at expanding frontiers promotes gene segregation. *Proceedings of the National Academy of Sciences* **104** (2007), 19926–19930.

[99] O. Hallatschek and D. R. Nelson. Gene surfing in expanding populations. *Theoretical Population Biology* **73** (2008), 158–170.

[100] O. Hallatschek and D. R. Nelson. Life at the front of an expanding population. *Evolution* **64** (2010), 193–206.

[101] K. Hamza, P. Jagers, and F. C. Klebaner. The age structure of population-dependent general branching processes in environments with a high carrying capacity. *Proceedings of the Steklov Institute of Mathematics* **282** (2013), 90–105.

[102] S. C. Harris, S. G. G. Johnston, and M. I. Roberts. The coalescent structure of continuous-time Galton–Watson trees. *Annals of Applied Probability* **30** (2020), 1368–1414.

[103] S. C. Harris and M. I. Roberts. The many-to-few lemma and multiple spines. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **53** (2017), 226–242.

[104] R. G. Harrison and E. L. Larson. Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity* **105** (2014), 795–809.

[105] A. Hastings. *Population Biology. Concepts and Models.* Springer-Verlag New York, 1997.

[106] X. He, E. H. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y. C. Lau, J. Y. Wong, Y. Guan, X. Tan, X. Mo, Y. Chen, B. Liao, W. Chen, F. Hu, Q. Zhang, M. Zhong, Y. Wu, L. Zhao, F. Zhang, B. J. Cowling, F. Li, and G. M. Leung. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine* **26** (2020), 672–675.

[107] T. A. Heath, J. P. Huelsenbeck, and T. Stadler. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences* **111** (2014), E2957–E2966.

[108] B. M. Henn, L. L. Cavalli-Sforza, and M. W. Feldman. The great human expansion. *Proceedings of the National Academy of Sciences* **109** (2012), 17758–17764.

[109] J. Hey. Using phylogenetic trees to study speciation and extinction. *Evolution* **46** (1992), 627–640.

[110] K. Higgins and M. Lynch. Metapopulation extinction caused by mutation accumulation. *Proceedings of the National Academy of Sciences* **98** (2001), 2928–2933.

[111] J. Jacod and A. N. Shiryaev. *Limit Theorems for Stochastic Processes*. Second edition. Grundlehren Der Mathematischen Wissenschaften. Springer-Verlag, 2003.

[112] P. Jagers. *Branching processes with biological applications*. Wiley, 1975.

[113] P. Jagers and F. C. Klebaner. Population-size-dependent and age-dependent branching processes. *Stochastic Processes and their Applications* **87** (2000), 235–254.

[114] P. Jagers and F. C. Klebaner. Population-size-dependent, age-structured branching processes linger around their carrying capacity. *Journal of Applied Probability* **48** (2011), 249–260.

[115] P. Jagers and O. Nerman. Limit theorems for sums determined by branching and other exponentially growing processes. *Stochastic Processes and their Applications* **17** (1984), 47–71.

[116] P. Jagers and O. Nerman. The growth and composition of branching populations. *Advances in Applied Probability* **16** (1984), 221–259.

[117] P. Jagers and S. Sagitov. General branching processes in discrete time as random trees. *Bernoulli* **14** (2008), 949–962.

[118] T. Jech. *Set Theory*. Third edition. Springer Monographs in Mathematics. Springer-Verlag Berlin Heidelberg, 2003.

[119] S. G. G. Johnston. The genealogy of Galton-Watson trees. *Electronic Journal of Probability* **24** (2019), 35 pp.

[120] O. Kallenberg. *Foundations of Modern Probability*. Second edition. Probability and its Applications. Springer-Verlag New York, 2002.

[121] O. Kallenberg. *Random Measures, Theory and Applications*. Probability Theory and Stochastic Modelling. Springer, Cham, 2017.

[122] J. Kelleher, A. M. Etheridge, and G. A. T. McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Computational Biology* **12** (2016), 1–22.

[123] J. G. Kemeny and J. L. Snell. *Finite Markov Chains*. Undergraduate Texts in Mathematics. Springer-Verlag New York, 1976.

[124] É. Kerdoncuff, A. Lambert, and G. Achaz. Testing for population decline using maximal linkage disequilibrium blocks. *Theoretical Population Biology* **134** (2020), 171–181.

[125] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **115** (1927), 700–721.

[126] G. Kersting, J. Schweinsberg, and A. Wakolbinger. The evolving beta coalescent. *Electronic Journal of Probability* **19** (2014), 27 pp.

[127] J. F. C. Kingman. The representation of partition structures. *Journal of the London Mathematical Society* **18** (1978), 374–380.

[128] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications* **13** (1982), 235–248.

[129] S. Klopfstein, M. Currat, and L. Excoffier. The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution* **23** (2006), 482–490.

[130] A. Kolmogoroff, I. Petrovsky, and N. Piscounoff. Études de l'équation avec croissance de la quantité de matière et son application à un problème biologique. *Moscow University Bulletin Of Mathematics* **1** (1937), 1–25.

[131] K. S. Korolev, M. J. I. Müller, N. Karahan, A. W. Murray, O. Hallatschek, and D. R. Nelson. Selective sweeps in growing microbial colonies. *Physical Biology* **9** (2012), 026008.

[132] A. M. Kramer, B. Dennis, A. M. Liebhold, and J. M. Drake. The evidence for Allee effects. *Population Ecology* **51** (2009), 341–354.

[133] C. Labbé. From flows of Λ-Fleming-Viot processes to lookdown processes via flows of partitions. *Electronic Journal of Probability* **19** (2014), 49 pp.

[134] A. Lambert. The branching process with logistic growth. *Annals of Applied Probability* **15** (2005), 1506–1535.

[135] A. Lambert. Population Dynamics and Random Genealogies. *Stochastic Models* **24** (2008), 45–163.

[136] A. Lambert. The contour of splitting trees is a Lévy process. *Annals of Probability* **38** (2010), 348–395.

[137] A. Lambert. Random ultrametric trees and applications. *ESAIM: Procs* **60** (2017), 70–89.

[138] A. Lambert. The coalescent of a sample from a binary branching process. *Theoretical Population Biology* **122** (2018), 30–35.

[139] A. Lambert, V. Miró Pina, and E. Schertzer. Chromosome Painting: how recombination mixes ancestral colors (2020). arXiv: `1807.09116`.

[140] A. Lambert and E. Schertzer. Recovering the Brownian coalescent point process from the Kingman coalescent by conditional sampling. *Bernoulli* **25** (2019), 148–173.

[141] A. Lambert and T. Stadler. Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. *Theoretical Population Biology* **90** (2013), 113–128.

[142] A. Lambert and G. Uribe Bravo. The comb representation of compact ultrametric spaces. *p-Adic Numbers, Ultrametric Analysis and Applications* **9** (2017), 22–38.

[143] J. Lamperti. Semi-stable Markov processes. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **22** (1972), 205–225.

[144] S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, and J. Lessler. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of internal medicine* **172** (2020), 577–582.

[145] J.-F. Le Gall. *Brownian Motion, Martingales, and Stochastic Calculus*. Graduate Texts in Mathematics. Springer International Publishing, 2016.

[146] H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature* **475** (2011), 493–496.

[147] Z. Li. *Measure-Valued Branching Markov Processes*. Probability and Its Applications. Springer-Verlag Berlin Heidelberg, 2011.

[148] N. M. Linton, T. Kobayashi, Y. Yang, K. Hayashi, A. R. Akhmetzhanov, S.-m. Jung, B. Yuan, R. Kinoshita, and H. Nishiura. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine* **9** (2020), 538.

[149] L. Loewe. Quantifying the genomic decay paradox due to Muller's ratchet in human mitochondrial DNA. *Genetical Research* **87** (2006), 133–159.

[150] R. Lyons, R. Pemantle, and Y. Peres. Conceptual proofs of $L \log L$ criteria for mean behavior of branching processes. *Annals of Probability* **23** (1995), 1125–1138.

[151] J. Mallet, N. Besansky, and M. W. Hahn. How reticulated are species? *BioEssays* **38** (2016), 140–149.

[152] J. Marin, G. Achaz, A. Crombach, and A. Lambert. The genomic view of diversification. *Journal of Evolutionary Biology* **33** (2020), 1387–1404.

[153] C. Massonnaud, J. Roux, and P. Crépey. COVID-19: Forecasting short term hospital needs in France. *medRxiv* (2020).

[154] G. A. T. McVean and N. J. Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360** (2005), 1387–1393.

[155] M. Möhle. Robustness results for the coalescent. *Journal of Applied Probability* **35** (1998), 438–447.

[156] M. Möhle. Weak convergence to the coalescent in neutral population models. *Journal of Applied Probability* **36** (1999), 446–460.

[157] M. Möhle and S. Sagitov. A classification of coalescent processes for haploid exchangeable population models. *Annals of Probability* **29** (2001), 1547–1562.

[158] P. A. P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society* **54** (1958), 60–71.

[159] H. Morlon, M. D. Potts, and J. B. Plotkin. Inferring the dynamics of diversification: A coalescent approach. *PLOS Biology* **8** (2010), 1–13.

[160] H. J. Muller. The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **1** (1964), 2–9.

[161] C. Müller and R. Tribe. Stochastic P.D.E.'s arising from the long range contact and long range voter processes. *Probability Theory and Related Fields* **102** (1995), 519–545.

[162] O. Nerman. On the convergence of supercritical general (C-M-J) branching processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **57** (1981), 365–395.

[163] O. Nerman and P. Jagers. The stable doubly infinite pedigree process of supercritical branching populations. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **65** (1984), 445–460.

[164] H. Nishiura, T. Kobayashi, T. Miyama, A. Suzuki, S.-m. Jung, K. Hayashi, R. Kinoshita, Y. Yang, B. Yuan, A. R. Akhmetzhanov, and N. M. Linton. Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). *International Journal of Infectious Diseases* **94** (2020), 154–155.

[165] P. Olofsson. The $x \log x$ condition for general branching processes. *Journal of applied probability* **35** (1998), 537–544.

[166] G. Pang and É. Pardoux. Functional limit theorems for non-Markovian epidemic models (2020). arXiv: 2003.03249.

[167] G. Pang and É. Pardoux. Multi-patch epidemic models with general infectious periods (2020). arXiv: 2006.14412.

[168] J. C. Pardos and V. Rivero. Self-similar Markov processes. *Boletín de la Sociedad Matemática Mexicana* **19** (2013), 201–235.

[169] É. Pardoux. *Probabilistic Models of Population Evolution. Scaling Limits, Genealogies and Interactions*. Stochastics in Biological Systems. Springer International Publishing, 2016.

[170] S. Peischl, I. Dupanloup, L. Bosshard, and L. Excoffier. Genetic surfing in human populations: from genes to genomes. *Current Opinion in Genetics & Development* **41** (2016), 53–61.

[171] S. Peischl, I. Dupanloup, A. Foucal, M. Jomphe, V. Bruat, J.-C. Grenier, A. Gouy, K. J. Gilbert, E. Gbeha, L. Bosshard, E. Hip-Ki, M. Agbessi, A. Hodgkinson, H. Vézina, P. Awadalla, and L. Excoffier. Relaxed selection during a recent human expansion. *Genetics* **208** (2018), 763–777.

[172] S. Peischl, I. Dupanloup, M. Kirkpatrick, and L. Excoffier. On the accumulation of deleterious mutations during range expansions. *Molecular Ecology* **22** (2013), 5972–5982.

[173] S. Peischl and L. Excoffier. Expansion load: recessive mutations and the role of standing genetic variation. *Molecular Ecology* **24** (2015), 2084–2094.

[174] S. Peischl, M. Kirkpatrick, and L. Excoffier. Expansion load and the evolutionary dynamics of a species range. *The American naturalist* **185** (2015), E81–E93.

[175] E. Pennisi. Shaking up the Tree of Life. *Science* **354** (2016), 817–821.

[176] P. Pfaffelhuber and A. Wakolbinger. The process of most recent common ancestors in an evolving coalescent. *Stochastic Processes and their Applications* **116** (2006), 1836–1859.

[177] P. Pfaffelhuber, A. Wakolbinger, and H. Weisshaupt. The tree length of an evolving coalescent. *Probability Theory and Related Fields* **151** (2011), 529–557.

[178] J. Pitman. Coalescents with multiple collisions. *The Annals of Probability* **27** (1999), 1870–1902.

[179] L. Popovic. Asymptotic genealogy of a critical branching process. *Annals of Applied Probability* **14** (2004), 2120–2148.

[180] P. Ralph and G. Coop. The geography of recent genetic ancestry across Europe. *PLOS Biology* **11** (2013), 1–20.

[181] Y.-X. Ren, R. Song, and Z. Sun. A 2-spine decomposition of the critical Galton-Watson tree and a probabilistic proof of Yaglom's theorem. *Electronic Communications in Probability* **23** (2018), 12 pp.

[182] Y.-X. Ren, R. Song, and Z. Sun. Spine decompositions and limit theorems for a class of critical superprocesses. *Acta Applicandae Mathematicae* **165** (2020), 91–131.

[183] C. Rogers and J. Pitman. Markov functions. *The Annals of Probability* **9** (1981), 573–582.

[184] L. Roques, J. Garnier, F. Hamel, and É. K. Klein. Allee effect promotes diversity in traveling waves of colonization. *Proceedings of the National Academy of Sciences* **109** (2012), 8828–8833.

[185] L. Roques, É. K. Klein, J. Papaix, A. Sar, and S. Soubeyrand. Using early data to estimate the actual infection fatality ratio from COVID-19 in France. *Biology* **9** (2020), 97.

[186] C. Roux, C. Fraïsse, J. Romiguier, Y. Anciaux, N. Galtier, and N. Bierne. Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLOS Biology* **14** (2016), 1–22.

[187] P. C. Sabeti, D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419** (2002), 832–837.

[188] S. Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability* **36** (1999), 1116–1125.

[189] H. Salje, C. Tran Kiem, N. Lefrancq, N. Courtejoie, P. Bosetti, J. Paireau, A. Andronico, N. Hozé, J. Richet, C.-L. Dubost, Y. Le Strat, J. Lessler, D. Levy-Bruhl, A. Fontanet, L. Opatowski, P.-Y. Boelle, and S. Cauchemez. Estimating the burden of SARS-CoV-2 in France. *Science* **369** (2020), 208–211.

[190] S. Sankararaman, S. Mallick, M. Dannemann, K. Prüfer, J. Kelso, S. Pääbo, N. Patterson, and D. Reich. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507** (2014), 354–357.

[191] S. Sankararaman, S. Mallick, N. Patterson, and D. Reich. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Current Biology* **26** (2016), 1241–1247.

[192] Santé Publique France. *COVID-19 : point épidémiologique du 4 juin 2020.* 2020. eprint: `https://www.santepubliquefrance.fr`.

[193] Santé Publique France. *Données hospitalières relatives à l'épidémie de COVID-19.* 2020. eprint: `https://www.data.gouv.fr/`. (accessed: 10.06.2020).

[194] E. Schertzer and F. Simatos. Height and contour processes of Crump-Mode-Jagers forests (I): General distribution and scaling limits in the case of short edges. *Electronic Journal of Probability* **23** (2018), 43 pp.

[195] J. Schweinsberg. Coalescents with Simultaneous Multiple Collisions. *Electronic Journal of Probability* **5** (2000), 50 pp.

[196] J. Schweinsberg. Coalescent processes obtained from supercritical Galton–Watson processes. *Stochastic Processes and their Applications* **106** (2003), 107–139.

[197] J. Schweinsberg. Dynamics of the evolving Bolthausen-Sznitman coalecent. *Electronic Journal of Probability* **17** (2012), 50pp.

[198] T. Sellke. On the asymptotic distribution of the size of a stochastic epidemic. *Journal of Applied Probability* **20** (1983), 390–394.

[199] Z. Shi. *Branching Random walks. École d'Été de Probabilités de Saint-Flour XLII-2012.* Vol. 2151. Lecture Notes in Mathematics. Springer, Cham, 2015.

[200] Y. B. Simons, M. C. Turchin, J. K. Pritchard, and G. Sella. The deleterious mutation load is insensitive to recent population history. *Nature Genetics* **46** (2014), 220–224.

[201] G. J. Slater, L. J. Harmon, and M. E. Alfaro. Integrating fossils with molecular phylogenies improves inference of trait evolution. *Evolution: International Journal of Organic Evolution* **66** (2012), 3931–3944.

[202] M. T. Sofonea, B. Reyné, B. Elie, R. Djidjou-Demasse, C. Selinger, Y. Michalakis, and S. Alizon. Epidemiological monitoring and control perspectives: application of a parsimonious modelling framework to the COVID-19 dynamics in France. *medRxiv* (2020).

[203] P. A. Stephens, W. J. Sutherland, and R. P. Freckleton. What is the Allee effect? *Oikos* **87** (1999), 185–190.

[204] A. N. Stokes. On two types of moving front in quasilinear diffusion. *Mathematical Biosciences* **31** (1976), 307–315.

[205] Z. Taïb. *Branching Processes and Neutral Evolution.* Vol. 93. Lecture Notes in Biomathematics. Springer Berlin Heidelberg, 1992.

[206] C. M. Taylor and A. Hastings. Allee effects in biological invasions. *Ecology Letters* **8** (2005), 895–908.

[207] L. Tindale, M. Coombe, J. E. Stockdale, E. Garlock, W. Y. V. Lau, M. Saraswat, Y.-H. B. Lee, L. Zhang, D. Chen, J. Wallinga, and C. Colijn. Transmission interval estimates suggest pre-symptomatic spread of COVID-19. *medRxiv* (2020).

[208] R. Tingley, M. Vallinoto, F. Sequeira, and M. R. Kearney. Realized niche shift during a global biological invasion. *Proceedings of the National Academy of Sciences* **111** (2014), 10233–10238.

[209] Z.-D. Tong, A. Tang, K.-F. Li, P. Li, H.-L. Wang, J.-P. Yi, Y.-L. Zhang, and J.-B. Yan. Potential presymptomatic transmission of SARS-CoV-2, Zhejiang province, China, 2020. *Emerging Infectious Disease journal* **26** (2020), 1052.

[210] V. C. Tran. Large population limit and time behaviour of a stochastic particle model describing an age-structured population. *ESAIM: Probability and Statistics* **12** (2008), 345–386.

[211] J. M. J. Travis, T. Münkemüller, O. J. Burton, A. Best, C. Dytham, and K. Johst. Deleterious mutations can surf to high densities on the wave front of an expanding population. *Molecular Biology and Evolution* **24** (2007), 2334–2343.

[212] M. C. Urban, B. L. Phillips, D. K. Skelly, and R. Shine. The cane toad's (*Chaunus* [*Bufo*] *marinus*) increasing ability to invade Australia is revealed by a dynamically updated range model. *Proceedings of the Royal Society B: Biological Sciences* **274** (2007), 1413–1419.

[213] W. Van Saarloos. Front propagation into unstable states. *Physics Report* **386** (2003), 29–222.

[214] R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. T. Walker, H. Fu, A. Dighe, J. T. Griffin, M. Baguelin, S. Bhatia, A. Boonyasiri, A. Cori, Z. Cucunubá, R. FitzJohn, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, D. Laydon, G. Nedjati-Gilani, S. Riley, S. van Elsland, E. Volz, H. Wang, Y. Wang, X. Xi, C. A. Donnelly, A. C. Ghani, and N. M. Ferguson. Estimates of the severity of coronavirus disease 2019: A model-based analysis. *The Lancet Infectious Diseases* **20** (2020), 669–677.

[215] E. M. Volz, K. Koelle, and T. Bedford. Viral phylodynamics. *PLoS computational biology* **9** (2013).

[216] J. Wakeley. *Coalescent theory. An introduction.* Roberts & Company Publishers, 2008.

[217] J. Wallinga and M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences* **274** (2007), 599–604.

[218] Y. Willi, M. Fracassetti, S. Zoller, and J. Van Buskirk. Accumulation of mutational load at the edges of a species range. *Molecular Biology and Evolution* **35** (2018), 781–791.

[219] J. T. Wu, K. Leung, M. Bushman, N. Kishore, R. Niehus, P. M. de Salazar, B. J. Cowling, M. Lipsitch, and G. M. Leung. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nature Medicine* **26** (2020), 506–510.