UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO
PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

THE BOLTHAUSEN-SZNITMAN COALESCENT: GENERAL THEORY AND APPLICATIONS IN POPULATION GENETICS

QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN CIENCIAS

PRESENTA:
ALEJANDRO HERNANDEZ WENCES

DIRECTOR DE TESIS:
ARNO SIRI-JEGOUSSE
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS

CIUDAD UNIVERSITARIA , CD. MX.
OCTUBRE 2017

The Bolthausen-Sznitman Coalescent: General Theory and Applications in Population Genetics

Introduction

The general theory of coalescent processes aims to provide a rigorous mathematical framework that can be used to model natural phenomena where a collection of particles may fuse together and form new particles as the system evolves over time. It has a variety of applications in distinct disciplines such as Physics and Biology. In the biological realm, particularly in the field of Population Genetics, it is used to model the parental relations of a given population as we track the ancestry of the individuals backwards in time, thus leading to the construction of a genealogical tree. In this interpretation the fusion of particles occurs at the time when a set of individuals meet a common ancestor in the past. Once we have a suitable coalescent model that describes the evolution of a particular population we can use it to study questions of biological relevance such as determining the time needed to reach the last common ancestor, the expected genetic diversity for neutral positions of the genome, or whether natural selection has played an important role in the evolution of the population.

The Bolthausen-Sznitman coalescent (BSC) is a well known example of a simple coalescent process that contemplates the fusion of multiple particles in a single event; it was first introduced in the study of spin glasses in physics but was rapidly adopted for the study of genealogical trees. It has been described as a limit process for the genealogies of different population evolution models, including models where the reproductive success of the individuals is determined by a fitness function (i.e. the population is under the pressure of natural selection).

From a mathematical perspective, coalescent processes are Markov processes that take values on the space of partitions of $\mathbb{N}$. In Chapter 1 we rigorously define this space and the random variables that take values on it. We then define exchangeable random partitions and prove a fundamental representation theorem, Kingman's representation, which is reminiscent of

de Finetti's theorem for exchangeable random sequences. In Chapter 2 we first define the coagulation operator, which is the operator on the space of partitions that will allow us to formalize the idea of 'fusing particles'. We will then formally define exchangeable coalescent processes in the most general setting and provide a basic construction based on Poisson point processes. Finally, we end Chapter 2 with a representation theorem for coagulation rates which allows for a characterization of coalescent processes in terms of measures on the simplex. In Chapter 3 we leave the general setting to focus on the particular case of simple coalescents. We provide an alternative construction for simple coalescents due to Pitman [12] and a simpler proof for the characterization of its coagulation rates. Then we formally define the Bolthausen-Sznitman coalescent and describe a recent construction in terms of random recursive trees. We then use this construction to prove an asymptotic result on the time to absorption of the BSC. Lastly, we spend the rest of the chapter to present a recently developed coupling technique that has proven to be very prolific in the study of different functionals for the BSC such as the total branch length, the total internal and external branch lengths, and the total number of jumps. As an example we present the proof published in the original work [4] where a weak limit law for the number of jumps in the BSC was obtained. Finally, in Chapter 4 we define the Site Frequency Spectrum (SFS) and interpret it in biological terms as the principal measure of the genetic diversity present in a population. We then prove a new result that gives an explicit expression for the expected SFS of the BSC for frequencies greater than $1/2$, and an upper bound for frequencies below $1/2$. To conclude, we show an application of the SFS as a model selection tool in the study of a population evolution model that contemplates the effect of natural selection but that has escaped a rigorous treatment due to its complexity.

Contents

Chapter 1

Random Partitions

1.1 Partitions

In this section we will define the basic mathematical structures that will help us represent and study coalescent processes.

Definition 1.1.1. Let B be a subset of $\mathbb{N}$ and π be a countable collection of subsets of B. We call π a *partition* of B if:

- $B_i \cap B_j = \emptyset \qquad$ for all B_i and B_j in π
- $\cup_{i \geq 1} B_i = B$.

We call $\{B \subset \mathbb{N} : B \in \pi\}$ the *blocks* of π.

If we denote the block of π that contains k by $\pi(k)$, we can define an equivalence relation in $\mathbb{N}$ by:

$$i \overset{\pi}{\sim} j \iff \pi(i) = \pi(j).$$

Conversely, given an equivalence relation in $\mathbb{N}$ we can define a partition π whose blocks are the corresponding equivalence classes. Given a set $A \subset B$ we define the *restriction* of π to A as:

$$\pi\big|_A \coloneqq \{B \cap A : B \in \pi\}.$$

Also, given a set $B \subset \mathbb{N}$ we will denote the collection of all partitions of B by $\mathscr{P}_B$. We will typically work with the sets $\{1, \ldots, n\}$ so we will denote them by $[n]$, and write $\mathscr{P}_n$ instead of $\mathscr{P}_{[n]}$, $\mathscr{P}_\infty$ instead of $\mathscr{P}_{\mathbb{N}}$, and $\pi\big|_n$ instead of $\pi\big|_{[n]}$.

Definition 1.1.2. A sequence of partitions $\{\pi_n\}_{n\in\mathbb{N}}$ with $\pi_n \in \mathscr{P}_n$ is *compatible* if $\pi_n\big|_k = \pi_k$ for all $k \leq n$, $n \in \mathbb{N}$.

Lemma 1.1.1. A sequence of partitions $\{\pi_n\}_{n\in\mathbb{N}}$ is compatible if and only if there exists a partition $\pi_\infty \in \mathscr{P}_\infty$ such that $\pi_\infty\big|_n = \pi_n$ for all $n \in \mathbb{N}$.

Proof. The reverse implication is readily seen. For the forward implication we set $q_1 = 1$ and consider $\{\pi_n(q_1)\}_{n\in\mathbb{N}}$. Note that this is an increasing sequence of sets since $\{\pi_n\}_{n\in\mathbb{N}}$ is compatible. Define:

$$B_{q_1} = \bigcup_{\mathbb{N}} \pi_n(q_1)$$

and let $q_2 = \min\{n \in \mathbb{N} : n \notin B_{q_1}\}$. Following the same procedure define recursively:

$$B_{q_k} = \cup_{\mathbb{N}} \pi_n(q_k), \quad q_{k+1} = \min\{n \in \mathbb{N} : n \notin B_{q_k}\}$$

and let $\pi_\infty = \{B_{q_k} : k \in \mathbb{N}\}$. It is easily seen that π_∞ is a partition of $\mathbb{N}$; also, for fixed k, if $q_j > k$ we have that $B_{q_j}\big|_k = \emptyset$ whereas if $q_j \leq k$ we have by the compatibility of $\{\pi_n\}_{n\in\mathbb{N}}$:

$$B_{q_j}\big|_{[k]} = \bigcup_{n=1}^{k} \pi_n(q_j) = \pi_k(q_j).$$

So $\pi_\infty\big|_k = \pi_k$ for all $k \in \mathbb{N}$, which finishes the proof. □

When working with partitions, particularly with compatible sequences of partitions or, equivalently, with partitions in $\mathscr{P}_\infty$, it is often helpful to consider the *partition tree*. The partition tree is a labeled tree whose nodes at level n are labeled with the set $\mathscr{P}_n$, and such that the children of $\pi \in \mathscr{P}_n$ are all the nodes in level $n+1$ with labels in $\mathscr{P}_{n+1}$ whose restriction to n is equal to π. Thus, the last common ancestor of the partition tree is labeled $\{\{1\}\}$, its descendants are labeled $\{\{1,2\}\}$ and $\{\{1\},\{2\}\}$, the descendants of, say, $\{\{1,2\}\}$, are the nodes with labels $\{\{1,2\},\{3\}\}$ and $\{\{1,2,3\}\}$, and so forth (see Figure 1.1). The above lemma tells us that to each infinite path starting at node $\{\{1\}\}$ and descending through the partition tree by choosing one children at a time corresponds to a unique partition π in $\mathscr{P}_\infty$ and vice versa, thus, we can think of $\mathscr{P}_\infty$ as the leaves of the partition tree.

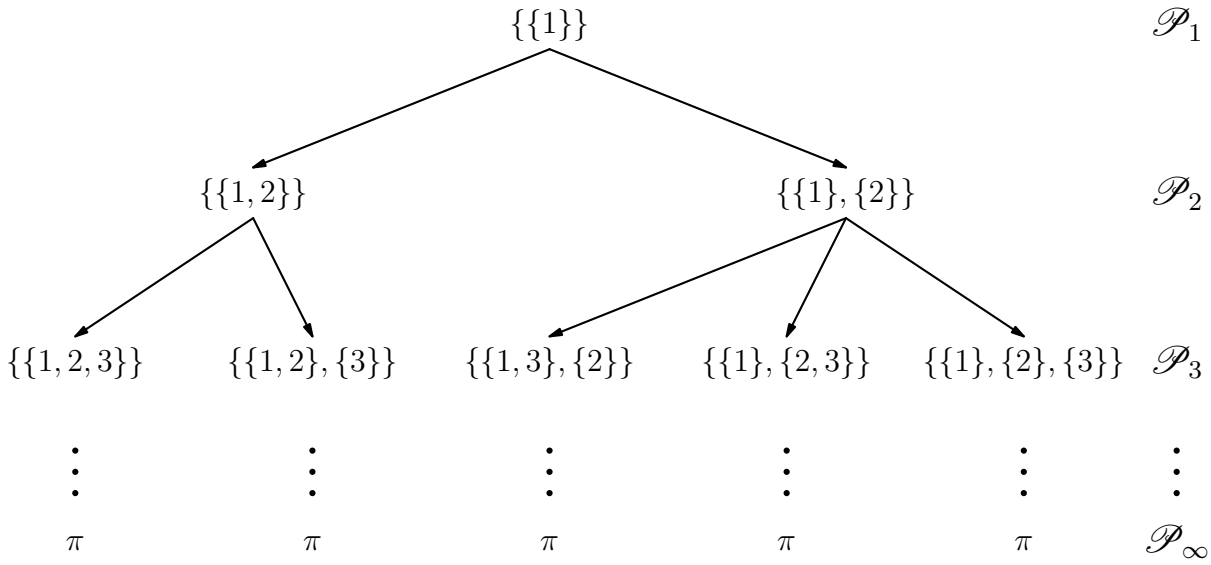

Figure 1.1: Partition Tree

Definition 1.1.3. Let B be any subset of $\mathbb{N}$, and π be any partition of $\mathbb{N}$.

- We say that a set B has an *asymptotic frequency* if the following limit exists:
$$|B| \coloneqq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_B(i)$$
where $\mathbb{1}_B$ is the usual indicator function for B.

- We say that π has *asymptotic frequencies* if π is such that $|B|$ exists for all $B \in \pi$. In this case we define $|\pi|^{\downarrow}$ as the sequence of asymptotic frequencies of π written in decreasing order, and write:
$$|\pi|^{\downarrow} = (|\pi|_1^{\downarrow}, \cdots).$$
Note that by definition we have $\sum |\pi|_i^{\downarrow} \leq 1$.

1.2 Exchangeable Random Partitions

Let us formalize the measurable space in which we will define random partitions. We will consider the set $\mathscr{P}_\infty$ and endow it with a distance function which will allow us to define a Borel σ-algebra.

Definition 1.2.1. We define a distance δ in $\mathscr{P}_\infty$ by:
$$\delta(\pi_1, \pi_2) = 1/\max\left\{n \in \mathbb{N} : \pi_1\big|_n = \pi_2\big|_n\right\}.$$

It is again useful to use the partition tree in order to visualize the closed balls given by δ. Fix a partition $\pi \in \mathscr{P}_\infty$ and consider the subtree given by all the possible paths starting from $\pi\big|_n$ and descending through all the levels of the tree; then, the corresponding partitions in $\mathscr{P}_\infty$ (leaves) associated with these paths form the closed ball $\{\pi' \in \mathscr{P}_\infty : \delta(\pi, \pi') \leq 1/n\}$. We will use subtrees such as the one just described later in the text so let us make a general definition.

Definition 1.2.2. Let $\pi \in \mathscr{P}_n$ and $m \geq n$. We will denote by $\mathscr{P}_m(\pi)$ the set of all partitions $\pi' \in \mathscr{P}_m$ such that $\pi'\big|_n = \pi$, or, in other words, the set of all descendants of π at level m.

Observe that in the discussion above we have

$$\mathscr{P}_\infty(\pi\big|_n) = \{\pi' \in \mathscr{P}_\infty : \delta(\pi, \pi') \leq 1/n\},$$

or, equivalently:

$$\mathscr{P}_\infty(\pi\big|_n) = \Big\{\pi' \in \mathscr{P}_\infty : \delta(\pi, \pi') < \frac{1}{n-1}\Big\}.$$

Theorem 1.2.1. $(\mathscr{P}_\infty, \delta)$ is a compact metric space.

Proof. Let $\{\pi_n\}_n^\infty$ be a sequence of partitions in $\mathscr{P}_\infty$ and let $\pi^1 \coloneqq \{1\}$. There exists a partition $\pi^2 \in \mathscr{P}_2$ such that $\pi\big|_2 = \pi^2$ for an infinite number of partitions π in the sequence $\{\pi_n\}_n^\infty$, pick one of these and denote it by $\hat{\pi}_1$. Then, in a recursive way, for every $k \in \mathbb{N}$ we can choose a partition $\pi^k \in \mathscr{P}_n$ such that $\pi^k\big|_j = \pi^j$ for all $j \leq k$ and a partition $\hat{\pi}_{k-1} \in \{\pi_n\}_n^\infty$ with $\hat{\pi}_{k-1}\big|_k = \pi^k$. The sequence of partitions $\{\pi^k\}_k^\infty$ is compatible so there exists a partition $\pi^\infty \in \mathscr{P}_\infty$ such that $\pi^\infty\big|_k = \pi^k$ for every $k \in \mathbb{N}$ and, by construction, $\hat{\pi}_k \xrightarrow{\delta} \pi^\infty$. □

Definition 1.2.3. Let $\mathscr{F}$ be the Borel σ-algebra in $\mathscr{P}_\infty$ induced by δ. A *random partition* Π is a random element of $(\mathscr{P}_\infty, \mathscr{F})$.

We will only be concerned with a particular type of random partitions: the exchangeable random partitions. We will see that this type of partitions have a nice representation reminiscent of the one assured by de Finetti's theorem for exchangeable random sequences. Similar to the context of de Finetti's theorem in which permutations of random sequences are defined, let us first define permutations of partitions.

Definition 1.2.4. A *finite permutation* of $\mathbb{N}$ is a bijective function $\sigma : \mathbb{N} \to \mathbb{N}$ with the property that there exists and integer M such that $\sigma(j) = j$ for all $j \geq M$.

Definition 1.2.5. Let σ be a permutation of $\mathbb{N}$ and π be a partition in $\mathscr{P}_\infty$. We define $\sigma(\pi)$ the *permutation of π given by σ* to be the partition:

$$\sigma(\pi) \coloneqq \big\{\sigma^{-1}(B) : B \in \pi\big\}.$$

Note that the blocks of $\sigma(\pi)$ are given by the inverse images of the blocks of π under σ and not by $\{\sigma(B) : B \in \pi\}$, actually, one should not expect that if $i \overset{\pi}{\sim} j$ then $\sigma(i) \overset{\sigma(\pi)}{\sim} \sigma(j)$, but rather that $\sigma^{-1}(i) \overset{\sigma(\pi)}{\sim} \sigma^{-1}(j)$.

Proposition 1.2.2. Let σ be a permutation of $\mathbb{N}$, then the map $\pi \mapsto \sigma(\pi)$ is continuous and, thus, measurable.

Proof. By definition there is an $N \in \mathbb{N}$ such that for all $j \geq N$ we have $\sigma(j) = j$. If $\delta(\pi, \pi') \leq 1/M$ with $M \geq N$ we have that $\pi\big|_M = \pi'\big|_M$ and $\sigma(\pi\big|_M) = \sigma(\pi'\big|_M)$. Since $\sigma(\pi\big|_M) = \sigma(\pi)\big|_M$ and $\sigma(\pi'\big|_M) = \sigma(\pi')\big|_M$ we see that $\delta(\sigma(\pi), \sigma(\pi')) \leq 1/M$. □

Definition 1.2.6. Let Π be a random partition. We say that Π is an *exchangeable random partition* if for every permutation σ we have:

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

Note that in particular for all $A \in \mathscr{F}$ we have:

$$\mathbb{P}(\Pi \in A) = \mathbb{P}(\sigma(\Pi) \in A) = \mathbb{P}(\Pi \in \sigma^{-1}(A)).$$

Notice that if π has asymptotic frequencies then $|\pi|^{\downarrow} = |\sigma(\pi)|^{\downarrow}$ since for every block B of π the block $\sigma^{-1}(B)$ of $\sigma(\pi)$ has the same asymptotic frequency as B (σ permutes only a finite number of elements of $\mathbb{N}$). The definition of exchangeable random partitions can intuitively be interpreted by saying that the only thing that determines $\mathbb{P}(\Pi = \pi)$ are the asymptotic frequencies of π and not the particular composition of its blocks. We will see that, indeed, exchangeable random partitions do have asymptotic frequencies almost surely; also, in the following section we will give a construction of exchangeable random partitions whose only input is a sequence of "asymptotic frequencies". This construction will turn out to be exhaustive, that is, all exchangeable random partitions can be derived from it (in a sense specified further ahead) thus evincing that the intuition given above is correct.

1.3 Mass Partitions

Definition 1.3.1. Let $\rho = (\rho_1, \rho_2, \cdots)$ be a sequence of real numbers in $[0, 1]$ such that:

- $\rho_i \geq \rho_j$ for all $i \geq j$
- $\sum_{i=0}^{\infty} \rho_i \leq 1$.

Such a sequence is called a *mass partition*. Let $\mathscr{P}_{[0,1]}$ be the space of all mass partitions and endow it with the supremum norm in ℓ_1 and its corresponding Borel σ-algebra $\mathscr{M}$.

Given a mass partition ρ define ρ_0 as:

$$\rho_0 \coloneqq 1 - \sum_{i=1}^{\infty} \rho_i.$$

We call ρ_0 the *dust* of ρ. We say that ρ is *proper* if $\rho_0 = 0$ and *improper* otherwise. We will interpret a mass partition as the sequence of strictly positive asymptotic frequencies of a partition π, and ρ_0 as the asymptotic frequency of the set formed by the union of all the blocks of π whose asymptotic frequency is equal to zero (thus justifying the term "dust").

Given a mass partition $\rho = (\rho_1, \rho_2, \cdots)$ we can construct a countable collection of open intervals $\{I_i\}_{i\in\mathbb{N}}$ such that:

- $I_i \bigcap I_j = \emptyset$ for all $i \neq j$
- $\lambda(I_i) = \rho_i$ for all $i \in \mathbb{N}$
- $\rho_0 = \lambda\big([0,1] \setminus \bigcup I_i\big)$

with λ being the Lebesgue measure on $[0,1]$. We call such a collection of intervals an *interval representation* of ρ. Conversely, given an open set $\mathcal{O}$ in $[0,1]$ we can find a countable collection of open intervals $\{I_i\}_{i\in\mathbb{N}}$ such that:

- $\bigcup_{\mathbb{N}} I_i = \mathcal{O}$
- $\sum_{\mathbb{N}} \lambda(I_i) = \lambda(\mathcal{O}) \leq 1$
- $\lambda(I_i) \geq \lambda(I_j)$ for all $i \geq j$

so we can construct a mass partition ρ given by $(\lambda(I_1), \lambda(I_2), \cdots)$. We will now use an interval representation $\{I_i\}_{i\in\mathbb{N}}$ of a mass partition ρ in order to construct an exchangeable random partition Π. Let A_0 be:

$$A_0 = [0,1] \setminus \bigcup I_i.$$

Also, consider a sequence numbers $\{u_i\}_{i\in\mathbb{N}} \in [0,1]^{\mathbb{N}}$ and construct π by defining the blocks

$$B_k = \{j \in \mathbb{N} : u_j \in I_k\},\ k \in \mathbb{N}$$

and setting

$$\pi = \{B_k : k \in \mathbb{N}\} \bigcup \{\{j\} : u_j \in A_0\}.$$

In other words, all the indices $j \in \mathbb{N}$ such that u_j falls in A_0 become singletons whereas all the indices such that u_j falls in I_k constitute the block B_k, for all $k \in \mathbb{N}$. Clearly this procedure generates a partition $\pi \in \mathscr{P}_\infty$ for each sequence $(u_1, u_2, \cdots)$, that is, we have a map $[0,1]^{\mathbb{N}} \xrightarrow{h} \mathscr{P}_\infty$ which is easily seen to be measurable; thus, if $\{U_i\}_{i\in\mathbb{N}}$ is a sequence of independent uniformly distributed random variables then $\Pi \coloneqq h(U_1, U_2, \cdots)$ is a random partition. In order to see that Π is exchangeable we just need to note that for every permutation σ we have:

$$(U_1, U_2, \cdots) \overset{\mathrm{d}}{=} (U_{\sigma(1)}, U_{\sigma(2)}, \cdots)$$

since $\{U_i\}_{i\in\mathbb{N}}$ are independent and identically distributed. Also, note that if

$$(u_1, u_2, \cdots) \overset{h}{\mapsto} \pi$$

then

$$(u_{\sigma(1)}, u_{\sigma(2)}, \cdots) \overset{h}{\mapsto} \sigma(\pi)$$

and thus $\Pi \overset{d}{=} \sigma(\Pi)$, which is the condition for being an exchangeable random partition.

Observe that if $\rho = (\rho_1, \rho_2, \cdots)$ then the law of large numbers ensures that $|B_k| = \lambda(I_k) = \rho_k$ for all $k \in \mathbb{N}$. Also, if

$$B_0 \coloneqq \{j : U_j \in A_0\}$$

then $|B_0| = \rho_0$ since $\rho_0 = 1 - \lambda(\bigcup I_i) = \lambda(A_0)$.

The construction described above is called the *paint-box* construction [10]. In the next section we will show that any exchangeable random partition can be constructed by setting an appropriate distribution on $\mathscr{P}_{[0,1]}$ and performing a paint-box process with a randomly chosen mass partition ρ according to this distribution. More precisely, if for any fixed $\rho \in \mathscr{P}_{[0,1]}$ we denote by ϱ_ρ the probability measure on $\mathscr{P}_\infty$ induced by the corresponding paint-box construction, for every exchangeable partition Π there exists a distribution $Q(d\rho)$ on $\mathscr{P}_{[0,1]}$ such that:

$$\mathbb{P}(\Pi \in \cdot) = \int_{\mathscr{P}_{[0,1]}} \varrho_\rho(\cdot) Q(d\rho).$$

This result is called the Kingman's Representation for exchangeable random partitions.

1.4 Kingman's Representation

The goal of this section is to prove Kingman's representation theorem for exchangeable random partitions. We will first use de Finetti's theorem to show that exchangeable random partitions have asymptotic frequencies almost surely. We will then show that the blocks of an exchangeable random partition can either have a positive asymptotic frequency or have a single element (i.e. blocks can either be singletons or have an infinite number of elements). Finally these two results will help us prove Kingman's representation.

Definition 1.4.1. A sequence of random variables $\{X_i\}_{i=1}^{\infty}$ is said to be *exchangeable* if for any finite collection of integers $\ell_1, \cdots, \ell_k$ and any permutation σ of these integers we have: $(X_{\sigma(\ell_1)}, \cdots, X_{\sigma(\ell_k)}) \overset{d}{=} (X_{\ell_1}, \cdots, X_{\ell_k})$.

Theorem 1.4.1 (de Finetti)**.** Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of exchangeable random variables defined on a probability space $(\Omega, \Sigma, \mathbb{P})$ with values in $\mathbb{R}$. Then:

- For every $A \in \mathscr{B}(\mathbb{R})$ the limit

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(X_i)$$

  exists almost surely.

- If we define $\mu(\omega, A) \coloneqq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(X_i(\omega))$ then μ is a probability kernel from Ω to $\mathbb{R}$.

- For every finite collection of indices $\{j_1, \cdots, j_k\}$ and Borel sets $\{A_1, \cdots, A_k\}$ we have:

$$\mathbb{P}(X_{j_1} \in A_1, \cdots, X_{j_k} \in A_k) = \int_{\Omega} \quad \prod_{i=1}^{k} \mu(\omega, A_i) \quad \mathbb{P}(d\omega).$$

  Observe that $\prod_{i=1}^{k} \mu(\omega, A_i)$ is $\mathcal{G}$-measurable where $\mathcal{G}$ is the σ-algebra generated by the random variables $\{\mu(\cdot, A)\}_{A\in\mathscr{B}(\mathbb{R})}$. Thus, in particular, $\{X_i\}_{i\in\mathbb{N}}$ are conditionally independent given μ.

The proof of this theorem has been extensively reviewed elsewhere [8][3].

Lemma 1.4.2. Let Π be an exchangeable random partition, then Π has asymptotic frequencies almost surely.

Proof. Fix an index $j \in \mathbb{N}$ and for all $i \neq j$ define the random variable:

$$\delta_j^{\Pi}(i) = \begin{cases} 1 \text{ if } \Pi(i) = \Pi(j) \\ 0 \text{ otherwise,} \end{cases}$$

then $\{\delta_j^{\Pi}(i)\}_{i \neq j}$ is an exchangeable random sequence. Indeed, notice that the exchangeability of Π ensures that for every permutation σ of $\mathbb{N}$ with $\sigma(j) = j$, and any collection of zero-one digits $d_1, \cdots, d_k$ we have:

$$\begin{aligned} \mathbb{P}(\delta_j^{\Pi}(i_1) = d_1, \cdots, \delta_j^{\Pi}(i_k) = d_k) &= \mathbb{P}(\delta_j^{\sigma(\Pi)}(i_1) = d_1, \cdots, \delta_j^{\sigma(\Pi)}(i_k) = d_k) \\ &= \mathbb{P}(\delta_j^{\Pi}(\sigma(i_1)) = d_1, \cdots, \delta_j^{\Pi}(\sigma(i_k)) = d_k) \end{aligned}$$

where the second equality holds since $\sigma(j) = j$ and, therefore, $\sigma(\Pi)(j) = \sigma^{-1}(\Pi(j))$, and $\sigma(\Pi)(i) = \sigma(\Pi)(j)$ if and only if $\Pi(\sigma(i)) = \Pi(j)$. By de Finetti's theorem, the limit:

$$\begin{aligned} |\Pi(j)| &= \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \delta_j^{\Pi}(i) \\ &= \lim_{n \to \infty} \frac{1}{n} \sum_{i \neq j}^{n} \delta_j^{\Pi}(i) \end{aligned}$$

exists almost surely. Since the choice of j was arbitrary we conclude that Π has asymptotic frequencies almost surely. □

Lemma 1.4.3. If Π is an exchangeable random partition then the blocks of Π are either singletons or have an infinite number of elements almost surely.

Proof. We will prove this by contradiction. Denote by $\#\Pi(j)$ the cardinality of the block $\Pi(j)$. Let $j \in \mathbb{N}$ and assume that $\mathbb{P}(\#\Pi(j) = m) > 0$ for some $m \in \mathbb{N}, m > 1$. We will construct a probability measure $\mathbb{P}'$ on $\mathbb{N}^{m-1}$ which will turn out to be "uniform", thus leading to a contradiction. For every (m-1)-tuple of integers $(n_1, \cdots, n_{m-1})$ define:

$$\mathbb{P}'(n_1, \cdots, n_{m-1}) = \frac{\mathbb{P}(\Pi(j) = \{j, n_1, \cdots, n_{m-1}\})}{\mathbb{P}(\#\Pi(j) = m)}.$$

$\mathbb{P}$' is easily checked to be a probability measure. Now, if $(\ell_1, \cdots, \ell_{m-1})$ is another collection of integers and σ is a permutation such that:

$$\begin{aligned} \sigma(n_1) &= \ell_1 \\ &\vdots \\ \sigma(n_{m-1}) &= \ell_{m-1} \\ \sigma(k) &= k \text{ for all } k \notin \{n_1, \cdots, n_{m-1}\} \end{aligned}$$

then, by the exchangeability of Π, it follows that

$$\begin{aligned} \mathbb{P}'(n_1, \cdots, n_{m-1}) &= \frac{\mathbb{P}(\Pi(j) = \{j, n_1, \cdots, n_{m-1}\})}{\mathbb{P}(\#\Pi(j) = m)} \\ &= \frac{\mathbb{P}(\sigma(\Pi)(j) = \{j, n_1, \cdots, n_{m-1}\})}{\mathbb{P}(\#\Pi(j) = m)} \\ &= \frac{\mathbb{P}(\Pi(j) = \{j, \sigma(n_1), \cdots, \sigma(n_{m-1})\})}{\mathbb{P}(\#\Pi(j) = m)} \\ &= \frac{\mathbb{P}(\Pi(j) = \{j, \ell_1, \cdots, \ell_{m-1}\})}{\mathbb{P}(\#\Pi(j) = m)} \\ &= \mathbb{P}'(\ell_1, \cdots, \ell_{m-1}). \end{aligned}$$

Since this is true for any collection of integers $(\ell_1, \cdots, \ell_{m-1})$ it follows that all the elements of $\mathbb{N}^{m-1}$ have the same probability under $\mathbb{P}'$, which is impossible since $\mathbb{P}'$ is a probability measure and $\mathbb{N}^{m-1}$ is an infinite set. Since the choice of j and m was arbitrary, for all j and $m > 1$ in $\mathbb{N}$ we have:

$$\mathbb{P}(\#\Pi(j) = m) = 0.$$

□

The preceding lemmas tell us a lot about the structure of exchangeable random partitions, they say that if Π is an exchangeable random partition then, almost surely, it takes values in a set that is much smaller than all of $\mathscr{P}_\infty$, particularly we may assume that Π takes values on the measurable set

$$\{\pi \in \mathscr{P}_\infty : \forall B \in \pi,\ |B| \text{ exists and } (|B| = 0 \iff B \text{ is a singleton })\}.$$

Moreover, using the same techniques as in the lemmas above, it is easy to show that, with probability one, B_0 has an asymptotic frequency and is

either empty or has an infinite number of elements. Thus, for every $\omega \in \Omega$ we may construct a block $B_0(\omega)$ as the union of all the singletons of $\Pi(\omega)$. Also, we may arrange the blocks of $\Pi(\omega)$ that have strictly positive asymptotic frequencies by decreasing size thus constructing a sequence of blocks $(B_1(\omega), B_2(\omega), \cdots,)$ such that $|B_i(\omega)| \geq |B_j(\omega)|$ whenever $i \leq j$. The only ambiguity comes up in the case where multiple blocks of a partition have the same asymptotic frequency, leading to multiple admissible orderings of its blocks; however, this is easily fixed by arbitrarily picking one of the possible arrangements. In practice we will want to work with measurable functions and sets, so we will break ties in asymptotic frequencies by ordering the blocks by their least elements. Formally, we will construct the sequence of blocks $\{B_i\}_{i=1}^{\infty}$ in a way such that

- $\Pi = \{B_1, B_2, \cdots\} \bigcup \{\{k\} : k \in B_0\}$
- $|B_i| \geq |B_j|$ whenever $i \leq j$
- $\min\{k : k \in B_i\} \leq \min\{k : k \in B_j\}$ whenever $i \leq j$ and $|B_i| = |B_j|$.

Definition 1.4.2. Let Π be an exchangeable partition and fix any $j \in \mathbb{N}$. Define $b_j^{\Pi} : \Omega \to \mathbb{N} \cup \{0\}$ by

$$b_j^{\Pi}(\omega) = i \quad \text{if } j \in B_i(\omega).$$

That is, $b_j^{\Pi} = i$ if j is in the ith block of $(B_0, B_1, \cdots)$, where $(B_0, B_1, \cdots)$ is constructed from Π as explained before.

It is easily seen that Π can be recovered from $\{b_j^{\Pi}\}_{j=1}^{\infty}$. With this in hand we can finally prove Kingman's Representation Theorem.

Theorem 1.4.4 (Kingman's Representation)**.** Let Π be an exchangeable random partition defined on a probability space $(\Omega, \Sigma, \mathbb{P})$. Then, there exists a probability measure Q on $\mathscr{P}_{[0,1]}$ such that

$$\mathbb{P}(\Pi \in A) = \int_{\mathscr{P}_{[0,1]}} \varrho_\rho(A)\ Q(d\rho) \quad \forall A \in \mathscr{F},$$

where we use the notation ϱ_ρ for the probability measure on $\mathscr{P}_\infty$ induced by the paint-box construction directed by ρ introduced in Section 1.3.

Proof. We first characterize the law of an exchangeable random partition $\hat{\Pi}$ when it is given by the paint-box construction for a fixed mass partition $\rho = (\rho_1, \rho_2, \cdots)$. If $\{I_n\}_{n\in\mathbb{N}}$ is an interval representation of ρ and if $A_0 = [0,1] \setminus \bigcup I_n$, then for any collection of integers $(\ell_1, \cdots, \ell_k)$ in $\mathbb{N} \cup \{0\}$ we have:

$$\begin{aligned}\varrho_\rho(b^{\hat{\Pi}}_{j_1} = \ell_1, \cdots, b^{\hat{\Pi}}_{j_k} = \ell_k) &= \prod_{i=1}^{k} \mathbb{P}(U_{j_1} \in I_{\ell_i}) \\ &= \prod_{i=1}^{k} \lambda(I_{\ell_i}) \\ &= \prod_{i=1}^{k} |B_{\ell_i}|\end{aligned}$$

where the last equality follows from an application of the law of large numbers. Now, returning to the general exchangeable random partition Π, we note that $\{b^{\Pi}_n\}_{n\in\mathbb{N}}$ is an exchangeable random sequence since for any permutation σ of $\{j_1, \cdots, j_k\}$ we have:

$$\begin{aligned}\mathbb{P}(b^{\Pi}_{j_1} = \ell_1, \cdots, b^{\Pi}_{j_k} = \ell_k) &= \mathbb{P}(b^{\sigma(\Pi)}_{\sigma^{-1}(j_1)} = \ell_1, \cdots, b^{\sigma(\Pi)}_{\sigma^{-1}(j_k)} = \ell_k) \\ &= \mathbb{P}(b^{\Pi}_{\sigma^{-1}(j_1)} = \ell_1, \cdots, b^{\Pi}_{\sigma^{-1}(j_k)} = \ell_k)\end{aligned}$$

where we used that

$$\left\{\omega \in \Omega : b^{\Pi}_{j_1} = \ell_1, \cdots, b^{\Pi}_{j_k} = \ell_k\right\} = \left\{\omega \in \Omega : b^{\sigma(\Pi)}_{\sigma^{-1}(j_1)} = \ell_1, \cdots, b^{\sigma(\Pi)}_{\sigma^{-1}(j_k)} = \ell_k\right\}$$

for the first equality, and the exchangeability of Π for the second. Thus, by de Finetti's theorem we have:

$$\mathbb{P}(b^{\Pi}_{j_1} = \ell_1, \cdots, b^{\Pi}_{j_k} = \ell_k) = \int_\Omega \prod_{i=1}^{k} \mu(\omega, \ell_i)\ \mathbb{P}(d\omega)$$

where $\mu(\omega, \ell) = \lim_{n\to\infty} \frac{1}{n} \sum^n \delta_\ell(b^{\Pi}_i) = |B_\ell|$, $\ell \in \mathbb{N} \cup \{0\}$. Let $\mathcal{G}$ be the σ-algebra generated by μ, and let $\rho(\Pi)$ be the random variable in $\mathscr{P}_{[0,1]}$ defined by $\rho(\Pi) \coloneqq (|B_1|, |B_2|, \cdots)$; then the σ-algebra generated by $\rho(\Pi)$ is identically $\mathcal{G}$ since $\rho(\Pi) = (\mu(\cdot, 1), \mu(\cdot, 2), \cdots)$. Let Q be the probability measure induced on $\mathscr{P}_{[0,1]}$ by $\rho(\Pi)$, and $\hat{\Pi}$ be the random partition given by

the paintbox construction with mass partition $\rho(\Pi)$, then, since

$$\prod_{i=1}^{k} \mu(\cdot, \ell_i) = \prod_{i=1}^{k} |B_{\ell_i}| = \varrho_{\rho(\Pi)}(b_{j_1}^{\hat{\Pi}} = \ell_1, \cdots, b_{j_k}^{\hat{\Pi}} = \ell_k),$$

substituting in the integral above we have:

$$\begin{aligned}
\mathbb{P}(b_{j_1}^{\Pi} = \ell_1, \cdots, b_{j_k}^{\Pi} = \ell_k) &= \int_{\Omega} \prod_{i=1}^{k} \mu(\omega, \ell_k)\ \mathbb{P}(d\omega) \\
&= \int_{\Omega} \varrho_{\rho(\Pi)}(b_{j_1}^{\hat{\Pi}} = \ell_1, \cdots, b_{j_k}^{\hat{\Pi}} = \ell_k)\ \mathbb{P}(d\omega) \\
&= \int_{\mathscr{P}_{[0,1]}} \varrho_{\rho}(b_{j_1}^{\hat{\Pi}} = \ell_1, \cdots, b_{j_k}^{\hat{\Pi}} = \ell_k) Q(d\rho)
\end{aligned}$$

□

Chapter 2

Coalescent Processes

2.1 Coagulation

Exchangeable coalescents are going to be defined as a stochastic process in continuous time, with values in $\mathscr{P}_\infty$. The evolution of these processes will be determined by a binary operator defined in $\mathscr{P}_\infty$, the coagulation operator, whose increments will be stationary over time. In order to define the coagulation operator Coag we first need to introduce the ordering of the blocks of a partition π by their least element. That is, for a partition π we construct an ordered sequence of blocks $(\pi_1, \pi_2, \cdots)$ such that:

- $\pi_j \in \pi$ for all $j \in \mathbb{N}$
- $\pi = \bigcup_{i=1}^{\infty} \{\pi_i\}$
- $\min\{k : k \in \pi_j\} \geq \min\{k : k \in \pi_i\}$ for all $j \geq i$ in $\mathbb{N}$.

From now on, when we refer to the kth block of π we mean the kth block under the order just described. Also, we will sometimes use the notation $[\pi]_k$ instead of π_k to make emphasis that we are referring to the kth block of π, specially when the notation for the partition within the brackets is large.

Definition 2.1.1. Let A be any set, we define $\#A$ to be the cardinality of A. In particular, if π is a partition, $\#\pi$ gives the number of blocks of π.

Definition 2.1.2 (Coagulation)**.** Let $\pi' \in \mathscr{P}_m$ with $m \in \mathbb{N} \cup \{\infty\}$. Then, for any partition π such that $\#\pi \leq m$ the pair (π, π') is called an *admissible*

pair, and we define the coagulation of π and π' as:

$$\mathrm{Coag}(\pi, \pi') = (\hat{\pi}_1, \hat{\pi}_2, \cdots),$$

where $\hat{\pi}_k$ is given by

$$\hat{\pi}_k \coloneqq \bigcup_{j \in \pi'_k} \pi_j,$$

where π_j is set to $\emptyset$ if $j > \#\pi$.

We now show that $(\mathscr{P}_m, \mathrm{Coag})$ is a monoid. Indeed, note that if $\pi = \{\{1\}, \cdots, \{m\}\}$, then $\mathrm{Coag}(\pi, \pi') = \pi'$ and $\pi' = \mathrm{Coag}(\pi', \pi)$ whenever (π, π') and (π', π) are admissible pairs. In other words, the partition $\{\{1\}, \cdots, \{m\}\}$ is the neutral element of $\mathscr{P}_m$ of the coagulation operator. For this reason we define $\mathbf{0}_m \coloneqq (\{1\}, \cdots, \{m\})$.

Lemma 2.1.1 (Associativity)**.** Let (π, π') and (π', π'') be admissible pairs. We have:

$$\mathrm{Coag}(\pi, \mathrm{Coag}(\pi', \pi'')) = \mathrm{Coag}(\mathrm{Coag}(\pi, \pi'), \pi'').$$

Proof. Let $\hat{\pi} = \mathrm{Coag}(\pi, \pi') = (\hat{\pi}_1, \cdots)$ and $\tilde{\pi} = \mathrm{Coag}(\pi', \pi'') = (\tilde{\pi}_1, \cdots)$. By definition we have that

$$[\mathrm{Coag}(\pi, \mathrm{Coag}(\pi', \pi''))]_k = \bigcup_{j \in \tilde{\pi}_k} \pi_j \quad \text{where} \quad \tilde{\pi}_k = \bigcup_{i \in \pi''_k} \pi'_i$$

and, thus

$$\begin{aligned}
[\mathrm{Coag}(\pi, \mathrm{Coag}(\pi', \pi''))]_k &= \bigcup_{i \in \pi''_k} \bigcup_{j \in \pi'_i} \pi_j \\
&= \bigcup_{i \in \pi''_k} [\mathrm{Coag}(\pi, \pi')]_i \\
&= [\mathrm{Coag}(\mathrm{Coag}(\pi, \pi'), \pi'')]_k.
\end{aligned}$$

□

Also, another important property of the coagulation operator which is easily seen is that:

$$\mathrm{Coag}(\pi, \pi')\big|_n = \mathrm{Coag}(\pi\big|_n, \pi') = \mathrm{Coag}(\pi\big|_n, \pi'\big|_n) \tag{2.1}$$

for any $n \in \mathbb{N}$.

Theorem 2.1.2. Let Π and Π' be two independent exchangeable random partitions. Then $\hat{\Pi} = \mathrm{Coag}(\Pi, \Pi')$ is an exchangeable partition.

Proof. Let σ be any permutation, then the blocks of $\hat{\Pi}$ are given by $\{\sigma^{-1}(\hat{\Pi}_k)\}_{k\in\mathbb{N}}$ where for each $k \in \mathbb{N}$ we have:

$$\begin{aligned}\sigma^{-1}(\hat{\Pi}_k) &= \sigma^{-1}\left(\bigcup_{j\in\Pi'_k} \Pi_j\right)\\ &= \bigcup_{j\in\Pi'_k} \sigma^{-1}\big(\Pi_j\big).\end{aligned}$$

Notice that in general we should not expect that $\sigma^{-1}(\Pi_j) = [\sigma(\Pi)]_j$ so it is not true that $\sigma(\hat{\Pi}) = \mathrm{Coag}(\sigma(\Pi), \Pi')$. However, we can define a map from $\mathscr{P}_\infty$ to the set of all permutations, $\pi \mapsto \widehat{\sigma_\pi}$, where $\widehat{\sigma_\pi}$ is given by:

$$\sigma^{-1}(\pi_j) = [\sigma(\pi)]_{\widehat{\sigma_\pi}(j)} \ \ \forall\, j \in \mathbb{N}.$$

Since there exists an integer M such that $\sigma(k) = k$ for all $k \geq M$ it follows that $\sigma^{-1}(\pi_k) = \pi_k$ for all $k \geq M$, therefore $\widehat{\sigma_\pi}(k) = k$ for all $k \geq M$ which proves that $\widehat{\sigma_\pi}$ is indeed a permutation. Furthermore, since the latter is true for every partition π, the map just described takes values on the set of permutations of the first $m-1$ integers, which is finite. Now let $\widehat{\sigma_\Pi}$ be its composition with Π. Then $\widehat{\sigma_\Pi}$ is independent of Π' since Π is, and $\widehat{\sigma_\Pi}$ induces a discrete probability measure on the set of all possible permutations. Moreover, for every $k \in \mathbb{N}$ we have:

$$\begin{aligned}\sigma^{-1}\big(\hat{\Pi}_k\big) &= \bigcup_{j\in\Pi'_k} \sigma^{-1}(\Pi_j)\\ &= \bigcup_{j\in\Pi'_k} [\sigma(\Pi)]_{\widehat{\sigma_\Pi}(j)}\\ &= \bigcup_{j\in\widehat{\sigma_\Pi}^{-1}\left(\Pi'_k\right)} [\sigma(\Pi)]_j.\end{aligned}$$

For every $k \in \mathbb{N}$ there is a unique $\ell \in \mathbb{N}$ such that $[\widehat{\sigma_\Pi}(\Pi')]_\ell = \sigma_\Pi^{-1}\big(\Pi'_k\big)$ and viceversa, thus

$$\sigma(\hat{\Pi}) = \mathrm{Coag}(\sigma(\Pi), \widehat{\sigma_\Pi}(\Pi')).$$

Let $\mathcal{A}$ be the finite range of $\widehat{\sigma_\Pi}$, then, by the independence of $(\Pi, \widehat{\sigma_\Pi})$ and Π', and the exchangeability of Π and Π', for any measurable sets $A, B \in \mathscr{P}_\infty$ we have

$$\begin{aligned}
\mathbb{P}\Big(\sigma(\Pi) \in A \bigcap \widehat{\sigma_\Pi}(\Pi') \in B\Big) &= \sum_{\sigma' \in \mathcal{A}} \mathbb{P}\Big(\sigma(\Pi) \in A \bigcap \widehat{\sigma_\Pi} = \sigma' \bigcap \sigma'(\Pi') \in B\Big) \\
&= \sum_{\sigma' \in \mathcal{A}} \mathbb{P}\Big(\sigma(\Pi) \in A \bigcap \widehat{\sigma_\Pi} = \sigma'\Big)\mathbb{P}\Big(\sigma'(\Pi') \in B\Big) \\
&= \sum_{\sigma' \in \mathcal{A}} \mathbb{P}\Big(\sigma(\Pi) \in A \bigcap \widehat{\sigma_\Pi} = \sigma'\Big)\mathbb{P}\Big(\Pi' \in B\Big) \\
&= \mathbb{P}\Big(\sigma(\Pi) \in A\Big)\mathbb{P}\Big(\Pi' \in B\Big) \\
&= \mathbb{P}\Big(\Pi \in A\Big)\mathbb{P}\Big(\Pi' \in B\Big) \\
&= \mathbb{P}\Big(\Pi \in A \bigcap \Pi' \in B\Big)
\end{aligned}$$

thus proving that $(\sigma(\Pi), \widehat{\sigma_\Pi}(\Pi')) \overset{d}{=} (\Pi, \Pi')$. Finally, it follows that:

$$\sigma\big(\operatorname{Coag}(\Pi, \Pi')\big) = \operatorname{Coag}(\sigma(\Pi), \widehat{\sigma_\Pi}(\Pi')) \overset{d}{=} \operatorname{Coag}(\Pi, \Pi').$$

□

Definition 2.1.3. Let $\{\pi^i\}_{i=1}^m$ be a collection of admissible partitions. We define

$$\operatorname*{CO}_{i=1}^{m} \pi^i$$

to be the partition given by the recurrence:

$$\begin{aligned}
\operatorname*{CO}_{i=1}^{2} \pi^i &\coloneqq \operatorname{Coag}(\pi^1, \pi^2) \\
\operatorname*{CO}_{i=1}^{m+1} \pi^i &\coloneqq \operatorname{Coag}\Big(\operatorname*{CO}_{i=1}^{m} \pi^i, \pi^{m+1}\Big).
\end{aligned}$$

To end this section we note that if $\{\Pi^i\}_{i=1}^m$ is a collection of independent exchangeable partitions then, using the preceding theorem in an induction argument, we see that $\operatorname{CO}_{i=1}^m \Pi_i$ is also an exchangeable partition.

2.2 Exchangeable Coalescents

We now follow the lines of Jean Bertoin [2] in order to define exchangeable coalescent processes.

Definition 2.2.1. Let $\Pi = (\Pi(t), t > 0)$ be a Markov process in continuous time with values in $\mathscr{P}_m$ for some $m \in \mathbb{N} \cup \{\infty\}$. Π is an *exchangeable coalescent* if $\Pi(0)$ is an exchangeable partition and the transition kernels of Π satisfy:

$$\mathbb{P}(\Pi(t+h) \in A \mid \Pi(t) = \pi) = \mathbb{P}(\mathrm{Coag}(\pi, \widetilde{\Pi}_h) \in A)$$

where A is any measurable set in $\mathscr{P}_m$ and $\widetilde{\Pi}_h$ is an exchangeable random partition that depends only on h. We call the collection $(\widetilde{\Pi}_h)_{h\in\mathbb{R}^+}$ the *stationary increments* of Π. Also, if $\Pi(0) = \mathbf{0}_m$ we call Π a *standard exchangeable coalescent*.

Because the values of $(\Pi(t), t > 0)$ are determined by the stationary increments $(\widetilde{\Pi}_h)_{h\in\mathbb{R}^+}$ in a way that resembles the definition of Lévy processes, coalescent processes may be loosely interpreted as Lévy processes where the binary operation is Coag in the set $\mathscr{P}_n$, instead of the usual sum operation in $\mathbb{R}$.

Lemma 2.2.1. If Π is a stadard exchangeable coalescent, then for all $h > 0$ we have:

$$\widetilde{\Pi}_h \overset{d}{=} \Pi(h).$$

Proof. Since $\Pi(0) = (\{1\}, \{2\}, \cdots)$, for any measurable set A we have:

$$\begin{aligned}\mathbb{P}(\Pi(h) \in A) &= \mathbb{P}(\mathrm{Coag}(\Pi(0), \widetilde{\Pi}_h) \in A) \\ &= \mathbb{P}(\widetilde{\Pi}_h \in A).\end{aligned}$$

□

If Π is an exchangeable coalescent with values in $\mathscr{P}_m$ then $(\mathrm{Coag}(\pi, \Pi(t)), t > 0)$ is also an exchangeable coalescent whenever $\pi \in \mathscr{P}_m$. In particular, if Π is standard then $\mathrm{Coag}(\pi, \Pi(0)) = \pi$, so $(\mathrm{Coag}(\pi, \Pi(t)), t > 0)$ is an exchangeable coalescent that starts at π, and whose probability kernels are determined by the stationary increments $\{\Pi(h)\}_{h\in\mathbb{R}^+}$. For this reason we will only consider standard coalescents from now on.

The following two results tell us that if Π is an exchangeable coalescent with values in $\mathscr{P}_\infty$, then we can equivalently study the process Π or the collection of consistent processes $(\Pi\big|_n, n \in \mathbb{N})$. The latter is easier to study since $\Pi\big|_n$ is a Markov process that takes values on the finite space $\mathscr{P}_n$ and thus is entirely described by its finite array of jumping rates.

Lemma 2.2.2. If Π is an exchangeable coalescent in $\mathscr{P}_m$, $m \in \mathbb{N} \cup \{0\}$, with increments $\widetilde{\Pi}_h$, then, for every $n \leq m$, $\Pi\big|_n := \big(\Pi(t)\big|_n : t > 0\big)$ is again an exchangeable coalescent with stationary increments given by $\left\{\widetilde{\Pi}_h\Big|_n\right\}_{h\in\mathbb{R}^+}$.

Proof. For any measurable set A we have:

$$\mathbb{P}\Big(\Pi\big|_n(t+h) \in A \;\Big|\; \Pi\big|_n(t) = \pi\Big) = \frac{\mathbb{P}\Big(\Pi\big|_n(t+h) \in A \bigcap \Pi\big|_n(t) = \pi\Big)}{\mathbb{P}\Big(\Pi\big|_n(t) = \pi\Big)}.$$

Using definition 1.2.2 of the set $\mathscr{P}_m(\pi)$, and equation (2.1), we get:

$$\begin{aligned}
\mathbb{P}\Big(\Pi\big|_n(t+h) \in A \;\Big|\; \Pi\big|_n(t) = \pi\Big) &= \frac{\displaystyle\int_{\mathscr{P}_m(\pi)} \mathbb{P}\Big(\Pi\big|_n(t+h) \in A \;\Big|\; \Pi(t) = \pi'\Big)\mathbb{P}(d\pi')}{\mathbb{P}\Big(\Pi\big|_n(t) = \pi\Big)} \\
&= \frac{\displaystyle\int_{\mathscr{P}_m(\pi)} \mathbb{P}\Big(\mathrm{Coag}\left(\pi', \widetilde{\Pi}_h\right)\Big|_n \in A\Big)\mathbb{P}(d\pi')}{\mathbb{P}\Big(\Pi\big|_n(t) = \pi\Big)} \\
&= \frac{\displaystyle\int_{\mathscr{P}_m(\pi)} \mathbb{P}\Big(\mathrm{Coag}\left(\pi'\big|_n, \widetilde{\Pi}_h\Big|_n\right) \in A\Big)\mathbb{P}(d\pi')}{\mathbb{P}\Big(\Pi\big|_n(t) = \pi\Big)} \\
&= \frac{\mathbb{P}\Big(\mathrm{Coag}\left(\pi, \widetilde{\Pi}_h\Big|_n\right) \in A\Big)\displaystyle\int_{\mathscr{P}_m(\pi)} \mathbb{P}(d\pi')}{\mathbb{P}\Big(\Pi\big|_n(t) = \pi\Big)} \\
&= \mathbb{P}\Big(\mathrm{Coag}\left(\pi, \widetilde{\Pi}_h\Big|_n\right) \in A\Big)
\end{aligned}$$

where, for the fourth equality, we used that $\pi'\big|_n = \pi$ for every $\pi' \in \mathscr{P}_m(\pi)$.

Thus, since $\Pi\big|_n(t)$ and $\widetilde{\Pi}_h\Big|_n$ are both exchangeable partitions, $\Pi\big|_n$ is an exchangeable coalescent with the desired property. □

Theorem 2.2.3. If Π takes values on $\mathscr{P}_\infty$ and is such that $\Pi\big|_n$ is an exchangeable coalescent for every $n \in \mathbb{N}$, then Π is an exchangeable coalescent.

Proof. To prove this theorem we just note that for every $t > 0$ the collection of exchangeable partitions $\{\Pi\big|_n(t) : n \in \mathbb{N}\}$ is consistent and, therefore, $\Pi(t)$ is an exchangeable partition. Also, for any $A \in \mathscr{F}$, if $A\big|_n := \{\pi\big|_n : \pi \in A\}$, we have:

$$\Big\{\Pi(t+h) \in A \bigcap \Pi(t) = \pi\Big\} = \bigcap_{n\in\mathbb{N}} \Big\{\Pi\big|_n(t+h) \in A\big|_n \bigcap \Pi\big|_n(t) = \pi\big|_n\Big\}$$

and, similarly:

$$\Big\{ \mathrm{Coag}(\pi, \Pi(h)) \in A\Big\} = \bigcap_{n\in\mathbb{N}} \Big\{\mathrm{Coag}(\pi, \Pi(h))\big|_n \in A\big|_n\Big\}.$$

Therefore:

$$\begin{aligned}
\mathbb{P}(\Pi(t+h) \in A \mid \Pi(t) = \pi) &= \lim_{n\to\infty} \mathbb{P}\Big(\Pi\big|_n(t+h) \in A\big|_n \;\Big|\; \Pi\big|_n(t) = \pi\big|_n\Big) \\
&= \lim_{n\to\infty} \mathbb{P}\Big(\mathrm{Coag}(\pi\big|_n, \Pi\big|_n(h)) \in A\big|_n\Big) \\
&= \lim_{n\to\infty} \mathbb{P}\Big(\mathrm{Coag}(\pi, \Pi(h))\big|_n \in A\big|_n\Big) \\
&= \mathbb{P}(\mathrm{Coag}(\pi, \Pi(h)) \in A).
\end{aligned}$$

So Π is an exchangeable coalescent with increments $\Pi(t)$. □

Let Π be an exchangeable coalescent taking values in $\mathscr{P}_\infty$. Since $\Pi\big|_n$ takes values on the finite set $\mathscr{P}_n$ its trajectories are entirely determined by its jumping rates:

$$\alpha_n(\pi', \pi) = \lim_{t\to 0} \frac{1}{t}\mathbb{P}\Big(\Pi\big|_n(t) = \pi \;\Big|\; \Pi\big|_n(0) = \pi'\Big) \quad \text{with } \pi' \neq \pi \text{ in } \mathscr{P}_n.$$

If (π', π) is a pair of admissible partitions in $\mathscr{P}_n$ and π is distinct from $\mathbf{0}_n$,

we have:

$$\begin{aligned}\alpha_n(\pi', \mathrm{Coag}(\pi', \pi)) &= \lim_{t\to 0} \frac{1}{t}\mathbb{P}\Big(\Pi\big|_n(t) = \mathrm{Coag}(\pi', \pi) \;\Big|\; \Pi\big|_n(0) = \pi'\Big) \\ &= \lim_{t\to 0} \frac{1}{t}\mathbb{P}\big(\,\mathrm{Coag}(\pi', \Pi(t)) = \mathrm{Coag}(\pi', \pi)\big) \\ &= \lim_{t\to 0} \frac{1}{t}\mathbb{P}\big(\Pi\big|_n(t) = \pi\big) \\ &= \alpha_n(\mathbf{0}_n, \pi).\end{aligned}$$

In other words, $\alpha_n(\mathbf{0}_n, \pi)$ is the jumping rate of $\Pi\big|_n$ from π' to $\mathrm{Coag}(\pi', \pi)$. On the other hand, if π cannot be written in the form $\pi = \mathrm{Coag}(\pi', \pi'')$ for any $\pi'' \in \mathscr{P}_n$, then

$$\mathbb{P}\Big(\Pi\big|_n(t) = \pi \;\Big|\; \Pi\big|_n(0) = \pi'\Big) = 0$$

so $\alpha_n(\pi', \pi) = 0$. The last two results combined tell us that the set

$$\{\alpha_n(\mathbf{0}_n, \pi) : \pi \in \mathscr{P}_n \setminus \mathbf{0}_n, n \in \mathbb{N}\}$$

completely determines the trajectories of $\Pi\big|_m$ for every $m \in \mathbb{N}$, and, thus, also the trajectories of Π. To ease notation from now on we will write α_π instead of $\alpha_n(\mathbf{0}_n, \pi)$.

Theorem 2.2.4. The set $\{\alpha_\pi : \pi \in \mathscr{P}_n \setminus \mathbf{0}_n, n \in \mathbb{N}\}$ determines a unique measure μ on $\mathscr{P}_\infty$ such that $\mu(\mathbf{0}_\infty) = 0$, and

$$\mu(\mathscr{P}_\infty(\pi)) = \alpha_\pi$$

for every $\pi \in \mathscr{P}_n \setminus \mathbf{0}_n, n \in \mathbb{N}$. We call μ the *coagulation rate* of Π.

Proof. The idea of the proof is to use Charatheodory's extension theorem in order to construct a measure on $\mathscr{P}_\infty \setminus \mathbf{0}_\infty$ and then define $\mu(\mathbf{0}_\infty) \coloneqq 0$. Towards this, note that the set

$$\mathscr{S} \coloneqq \{\mathscr{P}_\infty(\pi) : \pi \in \mathscr{P}_n \setminus \mathbf{0}_n, \;\; n \in \mathbb{N}\}$$

is a semiring. Define a measure $\hat{\mu}$ in $\mathscr{S}$ by

$$\hat{\mu}(\mathscr{P}_\infty(\pi)) = \alpha_\pi.$$

Note that if $\pi \in \mathscr{P}_m$ and $n \geq m$, then

$$\mathscr{P}_\infty(\pi) = \bigcup_{\pi' \in \mathscr{P}_n(\pi)} \mathscr{P}_\infty(\pi').$$

Hence, in order to prove that $\hat{\mu}$ is finitely additive we need to verify that

$$\hat{\mu}(\mathscr{P}_\infty(\pi)) = \sum_{\pi' \in \mathscr{P}_n(\pi)} \hat{\mu}(\mathscr{P}_\infty(\pi')).$$

Observe that

$$\{\omega \in \Omega : \Pi\big|_m = \pi\} = \bigcup_{\pi' \in \mathscr{P}_n(\pi)} \{\omega \in \Omega : \Pi\big|_n = \pi'\},$$

thus, the rate at which $\mathbf{0}_m$ jumps to π, equals the rate at which $\mathbf{0}_n$ jumps to $\bigcup_{\pi' \in \mathscr{P}_n(\pi)} \{\Pi\big|_n = \pi'\}$. Since the sets on the last union are pairwise disjoint, the rate at which the latter occurs is

$$\sum_{\pi' \in \mathscr{P}_n(\pi)} \alpha_{\pi'} = \sum_{\pi' \in \mathscr{P}_n(\pi)} \hat{\mu}(\mathscr{P}_\infty(\pi'))$$

so $\hat{\mu}$ is finitely additive. To see that $\hat{\mu}$ is infinitely additive note that for any partition $\pi \in \mathscr{P}_n$ ($n \in \mathbb{N}$), $\mathscr{P}_\infty(\pi)$ is closed and, by theorem 1.2.1, compact. Therefore, if there exists a collection of partitions $\{\pi_i\}_i^\infty$ such that $\{\mathscr{P}_\infty(\pi_i)\}_{i=1}^\infty$ are pairwise disjoint and $\mathscr{P}_\infty(\pi) = \cup_{i=1}^\infty \mathscr{P}_\infty(\pi_i)$, then, since $\{\mathscr{P}_\infty(\pi_i)\}_{i=1}^\infty$ is an open cover of $\mathscr{P}_\infty(\pi)$, it must be the case that $\mathscr{P}_\infty(\pi_i) = \emptyset$ for all but finitely many $i \in \mathbb{N}$. By the finite additivity of $\hat{\mu}$ we then have the equality $\hat{\mu}(\mathscr{P}_\infty(\pi)) = \sum_{i=1}^\infty \hat{\mu}(\mathscr{P}_\infty(\pi_i))$. By Charatheodory's theorem, $\hat{\mu}$ can be uniquely extended to a measure μ on $\mathscr{P}_\infty \setminus \mathbf{0}_\infty$, and setting $\mu(\mathbf{0}_\infty) \coloneqq 0$ finishes the proof of the theorem. □

Remark. Note that $\mu(\mathscr{P}_\infty(\mathscr{P}_n \setminus \mathbf{0}_n)) < \infty$ for every $n \in \mathbb{N}$ and, therefore, μ is a σ-finite measure.

To end this section we note that if $\pi \neq \mathbf{0}_n$ then for every permutation σ we have

$$\begin{aligned}
\alpha_n(\mathbf{0}_n, \pi) &= \lim_{t \to 0} \frac{1}{t} \mathbb{P}(\Pi\big|_n(t) = \pi) \\
&= \lim_{t \to 0} \frac{1}{t} \mathbb{P}(\sigma(\Pi\big|_n(t)) = \pi) \\
&= \lim_{t \to 0} \frac{1}{t} \mathbb{P}(\Pi\big|_n(t) = \sigma^{-1}(\pi)) \\
&= \alpha_n(\mathbf{0}_n, \sigma^{-1}(\pi)).
\end{aligned}$$

Starting with $\sigma(\pi)$ instead of π above, we see that $\alpha_n(\mathbf{0}_n, \sigma(\pi)) = \alpha_n(\mathbf{0}_n, \pi)$ and, therefore, the coagulation rate μ is invariant under permutations. More precisely, for any permutation σ and any measurable set A we have:

$$\mu(A) = \mu(\sigma(A)).$$

2.3 Poissonian Construction

In this section we will construct an exchangeable coalescent Π on $\mathscr{P}_\infty$ from any measure μ that satisfies the three properties described for coagulation rates in the previous section, mainly:

- $\mu(\mathscr{P}_n \setminus \mathbf{0}_n) < \infty$ for all $n \in \mathbb{N}$
- $\mu(\sigma(A)) = \mu(A)$ for any measurable set A and permutation σ
- $\mu(\mathbf{0}_\infty) = 0$.

In other words, we will see that any measure μ satisfying these properties is the coagulation rate of an exchangeable coalescent. Let M be a Poisson random measure on $\mathbb{R}^+ \times \mathscr{P}_\infty$ with intensity $\lambda \otimes \mu$. For each $n \in \mathbb{N}$ define a random measure M_n on $\mathbb{R}^+ \times \mathscr{P}_n$ by:

$$M_n([0,t] \times \pi) \coloneqq M([0,t] \times \mathscr{P}_\infty(\pi)) \quad (\forall \pi \in \mathscr{P}_n)$$

and note that M_n is a Poisson random measure on $\mathbb{R}^+ \times \mathscr{P}_n$ with intensity $\lambda \otimes \mu_n$, where μ_n is the measure on $\mathscr{P}_n$ given by:

$$\mu_n(\pi) = \mu(\mathscr{P}_\infty(\pi)) \quad (\forall \pi \in \mathscr{P}_n).$$

Since $\mu_n(\mathscr{P}_n \setminus \mathbf{0}_n) < \infty$, M_n has a finite number of atoms in $[0,t] \times \mathscr{P}_n \setminus \mathbf{0}_n$ with probability one. Also, if (t_1, π^1) and (t_2, π^2) are two atoms of M_n in $[0,t] \times \mathscr{P}_n \setminus \mathbf{0}_n$, then $t_1 \neq t_2$ with probability one; that is, the atoms of M_n in $[0,t] \times \mathscr{P}_n \setminus \mathbf{0}_n$ occur at different times with probability one. Therefore, for every $n \in \mathbb{N}$ the atoms of M_n in $[0,t] \times \mathscr{P}_n \setminus \mathbf{0}_n$ can be ordered according to their first coordinate and we may define the sequence of random vectors $\{(T_i, \Pi_i)\}_{i \in \mathbb{N}}$ given by this ordering. Using the latter, we consider the process Π^n with values in $\mathscr{P}_n$ such that for every $t \geq 0$, $\Pi^n(t)$ is given by the ordered coagulation:

$$\Pi^n(t) = \underset{\{i : 0 < T_i \leq t\}}{\mathrm{CO}} \Pi_i.$$

From now on we will write $\mathrm{CO}_{0<T_i\leq t}\,\Pi_i$ instead of $\mathrm{CO}_{\{i:0<T_i\leq t\}}\,\Pi_i$.

Note that it is possible to construct the Poisson random measure M_n through the following procedure: let $\widehat{\mu_n} := \mu_n(\mathscr{P}_n \setminus \mathbf{0}_n)$ and consider a sequence of independent identically distributed random partitions $\{\Pi\}_{i\in\mathbb{N}}$ with values in $\mathscr{P}_n \setminus \mathbf{0}_n$ and law $\mu_n(\pi)/\widehat{\mu_n}$. Let P be an independent Poisson process in $\mathbb{R}^+$ with parameter $\widehat{\mu_n}$, and denote its jumping times by $\{T_i\}_{i=1}^{\infty}$. Finally, define the atoms of M_n in $\mathbb{R}^+ \times \mathscr{P}_n \setminus \mathbf{0}_n$ to be the points $\{(\Pi_i, T_i)\}$. To see that this indeed generates a Poisson random measure on $\mathbb{R}^+ \times \mathscr{P}_n \setminus \mathbf{0}_n$ with intensity $\lambda \otimes \mu_n$ we compute for any $\pi \in \mathscr{P}_n \setminus \mathbf{0}_n$ and $t_1, t_2 \in \mathbb{R}^+$:

$$\begin{aligned}
\mathbb{P}\big(M_n(\pi \times [t_1,t_2]) = k\big) &= \sum_{j=k}^{\infty} \mathbb{P}(M_n(\mathscr{P}_n \times [t_1,t_2]) = j)\mathbb{P}(M_n(\pi \times [t_1,t_2]) = k | M_n(\mathscr{P}_n \times [t_1,t_2]) = j) \\
&= \sum_{j=k}^{\infty} \mathbb{P}(P(t_2) - P(t_1) = j)\binom{j}{k}\Big(\frac{\mu_n(\pi)}{\widehat{\mu_n}}\Big)^k\Big(\frac{\widehat{\mu_n} - \mu_n(\pi)}{\widehat{\mu_n}}\Big)^{j-k} \\
&= \sum_{j=k}^{\infty} e^{-\widehat{\mu_n}(t_2-t_1)}\frac{[\widehat{\mu_n}(t_2-t_1)]^j}{j!}\binom{j}{k}\Big(\frac{\mu_n(\pi)}{\widehat{\mu_n}}\Big)^k\Big(\frac{\widehat{\mu_n} - \mu_n(\pi)}{\widehat{\mu_n}}\Big)^{j-k} \\
&= \frac{e^{-\widehat{\mu_n}(t_2-t_1)}[\mu_n(\pi)(t_2-t_1)]^k}{k!}\sum_{j=k}^{\infty}\frac{[(\widehat{\mu_n} - \mu_n(\pi))(t_2-t_1)]^{j-k}}{(j-k)!} \\
&= \frac{e^{-\widehat{\mu_n}(t_2-t_1)}[\mu_n(\pi)(t_2-t_1)]^k e^{(\widehat{\mu_n}-\mu_n(\pi))(t_2-t_1)}}{k!} \\
&= \frac{e^{-\mu_n(\pi)(t_2-t_1)}[\mu_n(\pi)(t_2-t_1)]^k}{k!}.
\end{aligned}$$

So $M_n([t_1,t_2]\times\pi)$ is Poisson distributed with parameter $\mu_n(\pi)(t_2-t_1)$. By a similar calculation it is easy to see that for any measurable sets A and B such that $A \cap B = \emptyset$, the random variables $M_n(A)$ and $M_n(B)$ are independent.

Theorem 2.3.1. The process Π^n with values in $\mathscr{P}_n$ given by:

$$\begin{aligned}
\Pi^n(0) &= \mathbf{0}_n \\
\Pi^n(t) &= \underset{0<T_i\leq t}{\mathrm{CO}}\,\Pi_i
\end{aligned}$$

is a standard exchangeable coalescent.

Proof. It is clear that if A is any measurable set and $t_1, \cdots, t_n \in [0,t]$ then:

$$\mathbb{P}\big(\Pi^n(t+h) \in A \,\big|\, \Pi^n(t)\big) = \mathbb{P}\big(\Pi^n(t+h) \in A \,\big|\, \Pi^n(t), \Pi^n(t_1), \cdots, \Pi^n(t_n)\big)$$

so Π^n is a Markov process. Now, consider the sequence of atoms $\{(T_i, \Pi_i)\}_{i\in\mathbb{N}}$ of M_n in $\mathbb{R}^+ \times \mathscr{P}_n \setminus \mathbf{0}_n$. Since μ is invariant under permutations then μ_n is also invariant under permutations. Thus, for every $i \in \mathbb{N}$ and any permutation σ we have $\mathbb{P}(\Pi_i = \pi) = \mathbb{P}(\Pi_i = \sigma(\pi))$, that is, Π_i is an exchangeable partition. Also, by the construction of the Poisson random measure M_n described above, we see that the random partitions $\{\Pi_i\}$ are independent and identically distributed. We also have that

$$M_n((0,h] \times \mathscr{P}_n \setminus \mathbf{0}_n) \stackrel{d}{=} M_n((t,t+h] \times \mathscr{P}_n \setminus \mathbf{0}_n)$$

since M_n is a Poisson random measure. Therefore

$$\underset{t<T_i\leq t+h}{\mathrm{CO}} \Pi_i \stackrel{d}{=} \underset{0<T_i\leq h}{\mathrm{CO}} \Pi_i.$$

Now, since $\Pi^n(t+h)$ is given by

$$\Pi^n(t+h) = \mathrm{Coag}\Big(\Pi^n(t), \underset{t<T_i\leq t+h}{\mathrm{CO}} \Pi_i\Big),$$

we only need to show that $\mathrm{CO}_{0<T_i\leq h}\, \Pi_i$ is an exchangeable partition. To prove this we note that for any finite collection of indices $J \subset \mathbb{N}$, the partition

$$\underset{i\in J}{\mathrm{CO}}\, \Pi_i$$

is exchangeable since the partitions $\{\Pi_i : i \in J\}$ are independent and exchangeable. Using the latter we compute:

$$\begin{aligned}
\mathbb{P}\Big(\underset{0<T_i\leq h}{\mathrm{CO}} \Pi_i = \pi\Big) &= \sum_{k=1}^{\infty} \mathbb{P}\Big(T_k \leq h < T_{k+1} \bigcap \overset{k}{\underset{i=1}{\mathrm{CO}}}\, \Pi_i = \pi\Big) \\
&= \sum_{k=1}^{\infty} \mathbb{P}\Big(T_k \leq h < T_{k+1} \bigcap \sigma\Big(\overset{k}{\underset{i=1}{\mathrm{CO}}}\, \Pi_i\Big) = \pi\Big) \\
&= \mathbb{P}\Big(\sigma\Big(\underset{0<T_i\leq h}{\mathrm{CO}} \Pi_i\Big) = \pi\Big).
\end{aligned}$$

□

Theorem 2.3.2. For any fixed $t > 0$, the sequence of partitions $\{\Pi^n(t)\}_{n\in\mathbb{N}}$ is consistent.

Proof. For any pair of integers $n > m$ let $\{(T_i, \Pi_i)\}_{i=1}^{\infty}$ be the atoms of M_n and $\{(T_{k_i}, \Pi_{k_i})\}_{i=1}^{\infty}$ be the subsequence of $\{(T_i, \Pi_i)\}_{i=1}^{\infty}$ such that $\Pi_i\big|_m \neq \mathbf{0}_n$. Then we note that the atoms of M_m are given by $\{(T_{k_i}, \Pi_{k_i}\big|_m)\}_{i=1}^{\infty}$ and:

$$\begin{aligned} \Pi^n\big|_m(t) &= \Big(\underset{0<T_i\leq t}{\mathrm{CO}} \Pi_i \Big)\Big|_m \\ &= \underset{0<T_i\leq t}{\mathrm{CO}} \Pi_i\big|_m \\ &= \underset{0<T_{k_i}\leq t}{\mathrm{CO}} \Pi_{k_i}\big|_m \\ &= \Pi^m(t). \end{aligned}$$

□

Finally, by Lemma 1.1.1 and Theorem 2.2.3, the exchangeable coalescents $\{\Pi^n\}_{n\in\mathbb{N}}$ determine a unique (in law) exchangeable coalescent Π in $\mathscr{P}_\infty$. Since Π^n is a Markov chain for every n, and since for every $\pi \in \mathscr{P}_n \setminus \mathbf{0}_n$ we have $\mathbb{P}(\Pi^n(T_1) = \pi) = \mu(\mathscr{P}_\infty(\pi))/\mu(\mathscr{P}_n \setminus \mathbf{0}_n)$, then it follows that

$$\alpha_\pi = \lim_{t\to 0} \frac{1}{t}\mathbb{P}\big(\Pi^n(t) = \pi\big) = \mu(\mathscr{P}_\infty(\pi)),$$

so Π has coagulation rate μ.

2.4 Representation of Coagulation Rates

In this section we will present an exhaustive way of constructing coagulation rates. In pursuance of this, let us first describe the two types of coagulation rates that will be the basis of our construction.

For each pair of integers i, j consider the partition $\pi_{i\sim j}$ given by the block $\{i, j\}$ and the singletons $\{\{k\} : k \neq i, k \neq j\}$. *Kingman's coagulation rate* μ_K is the measure on $\mathscr{P}_\infty$ given by atoms of size one at the points $\{\pi_{i\sim j} : 1 \leq i < j < \infty\}$. Note that the coalescent process determined by this coagulation rate evolves through coagulations of exactly two blocks at a time.

For the second type of coagulation rate consider any measure ν on $\mathscr{P}_{[0,1]}$ such that $\nu(\mathbf{0}) = 0$ and

$$\int_{\mathscr{P}_{[0,1]}} \sum_{i=1}^{\infty} \rho_i^2 \ \nu(d\rho) < \infty. \tag{2.2}$$

Then, using the measures $(\varrho_\rho,\ \rho \in \mathscr{P}_{[0,1]})$ given by the paintbox construction introduced in section 1.3, define the measure μ_ν on $\mathscr{P}_\infty$ by:

$$\mu_\nu(\cdot) \coloneqq \int_{\mathscr{P}_{[0,1]}} \varrho_\rho(\cdot)\ \ \nu(d\rho).$$

It follows that μ_ν is a coagulation rate. To see this notice that ϱ_ρ is invariant under permutations and, hence, μ_ν is also invariant under permutations. Also, $\mu_\nu(\mathbf{0}_\infty) = 0$ since $\nu(\mathbf{0}) = 0$. Finally, note that $\mu_\nu(\mathscr{P}_\infty(\mathscr{P}_n \setminus \mathbf{0}_n)) < \infty$ for every $n \in \mathbb{N}$, since for every $\pi \in \mathscr{P}_n \setminus \mathbf{0}_n$ there exists i, j such that $i \sim j$ and:

$$\begin{aligned}\mu_\nu(\mathscr{P}_\infty(\pi)) &\leq \mu_\nu\big(\mathscr{P}_\infty\big(\{\pi : i \sim j\}\big)\big)\\ &= \int_{\mathscr{P}_m} \sum_{i=1}^{\infty} \rho_i^2\ \ \nu(d\rho) < \infty.\end{aligned}$$

Since μ_K and μ_ν are coagulation rates it is easily seen that for every $c > 0$, the measure $\mu \coloneqq c\mu_k + \mu_\nu$ is again a coagulation rate. The following theorem states that all coagulation rates can be constructed in this way.

Theorem 2.4.1. Let μ be any coagulation rate. There exists a constant $c > 0$ and a measure ν in $\mathscr{P}_{[0,1]}$ that satisfies $\nu(\mathbf{0}) = 0$ and (2.2) such that

$$\mu = c\mu_K + \mu_\nu.$$

Proof. First, we construct a measure ν on $\mathscr{P}_{[0,1]}$ such that for every measurable set $A \subset \mathscr{P}_\infty$ we have:

$$\mu\big(A \bigcap |\pi|^\downarrow \neq \mathbf{0}\big) = \int_{\mathscr{P}_{[0,1]}} \varrho_\rho(A)\ \ \nu(d\rho).$$

For this, fix any $n \in \mathbb{N}$ and consider the measure μ_n on $\mathscr{P}_\infty$ given by:

$$\mu_n(A) \coloneqq \mu\big(A \bigcap \{\pi\big|_n \neq \mathbf{0}_n\}\big) \quad (\forall A \in \mathscr{F}).$$

Note that μ_n is finite and invariant under permutations. In particular, for any permutation σ such that $\sigma(k) = k$ for every $k \in \{1, \cdots, n\}$, and any measurable set A, we have $\mu_n(A) = \mu_n(\sigma(A))$. Therefore, if $\widehat{\mu_n}$ is the image measure of μ_n under the translation by n (denoted T^n) given by

$$\pi \overset{T^n}{\to} \big(i \sim j \iff i + n \overset{\pi}{\sim} j + n\big),$$

then $\widehat{\mu_n}$ is also invariant under permutations and finite. By Kingman's Representation, if ν_n is the image measure of $\widehat{\mu_n}$ under the map $\pi \to |\pi|^{\downarrow}$, we have

$$\widehat{\mu_n}(\cdot) = \int_{\mathscr{P}_{[0,1]}\setminus \mathbf{0}} \varrho_\rho(\cdot)\nu_n(d\rho)$$

and

$$\begin{aligned}(2.3)\qquad \int_{\mathscr{P}_{[0,1]}} \sum_{i=1}^{\infty} \rho_i \ \nu_n(d\rho) &= \widehat{\mu_n}\big(\{\pi : 1 \sim 2\}\big) \\ &= \mu_n\big(\{\pi : 1+n \sim 2+n\}\big) \\ &= \mu\Big(\{\pi : 1+n \sim 2+n\} \bigcap \mathscr{P}_\infty(\mathscr{P}_n \setminus \mathbf{0}_n)\Big) \\ &\leq \mu\big(\{\pi : 1 \sim 2\}\big) < \infty.\end{aligned}$$

Note that $|\pi|^{\downarrow} = |T^n(\pi)|^{\downarrow}$, so ν_n is also the image measure of μ_n under the map $\pi \to |\pi|^{\downarrow}$. Since $\mu_n \uparrow \mu$ it follows that $\nu_n \uparrow \hat{\nu}$ where $\hat{\nu}$ is the image measure of μ under the same map. By taking the limit as n goes to infinity in (2.3) we see that

$$\int_{\mathscr{P}_{[0,1]}} \sum_{i=1}^{\infty} \rho_i \ \hat{\nu}(d\rho) < \infty,$$

so if $\nu \coloneqq \mathbb{1}_{\{|\pi|^{\downarrow} \neq \mathbf{0}\}}\hat{\nu}$, then $\nu(\mathbf{0}) = 0$ and μ_ν is a coagulation rate. It remains to show that

$$\mu\big(A \bigcap |\pi|^{\downarrow} \neq \mathbf{0}\big) = \int_{\mathscr{P}_{[0,1]}} \varrho_\rho(A) \ \nu(d\rho),$$

or equivalently that

$$\mu_\nu = \mathbb{1}_{\{|\pi|^{\downarrow} \neq \mathbf{0}\}}\mu.$$

For this we compute for any $\pi' \in \mathscr{P}_k \setminus \mathbf{0}_k$ and any integer $n \geq k$:

$$\begin{aligned}\mu\Big(\mathscr{P}_\infty(\pi') \bigcap |\pi|^{\downarrow} \neq 0\Big) &= \mu_n\Big(\mathscr{P}_\infty(\pi') \bigcap |\pi|^{\downarrow} \neq 0\Big) \\ &= \mu_n\Big(\mathscr{P}_\infty(\pi') \bigcap |\pi|^{\downarrow} \neq 0 \bigcap \pi\big|_{k+1,\cdots,k+n} \neq \mathbf{0}_{\{k+1,\cdots,k+n\}}\Big)\end{aligned}$$

where the last equality holds since, by Kingman's representation, μ_n is supported on partitions of $\mathbb{N}$ whose blocks are singletons if and only if their asymptotic frequency is identically zero; therefore, we have:

$$\mu_n(\{|\pi|^{\downarrow} \neq 0\} \setminus \{\pi\big|_{k+1,\cdots,k+n} \neq \mathbf{0}_{\{k+1,\cdots,k+n\}}\}) = 0.$$

Now, for fixed n let σ be the permutation such that $1 \to n+1, \cdots, k \to n+k$ and $k+1 \to 1, \cdots, k+n \to n$. Then, by the exchangeability of μ and the definition of $\widehat{\mu_n}$, we have:

$$\begin{aligned}
\mu\Big(\mathscr{P}_\infty(\pi')\bigcap|\pi|^{\downarrow}\neq 0\Big) &= \mu_n\Big(\mathscr{P}_\infty(\sigma(\pi'))\bigcap|\pi|^{\downarrow}\neq 0\bigcap\sigma\big(\pi\big|_{k+1,\cdots,k+n}\neq \mathbf{0}_{\{k+1,\cdots,k+n\}}\big)\Big)\\
&= \mu_n\Big(\mathscr{P}_\infty(\sigma(\pi'))\bigcap|\pi|^{\downarrow}\neq 0\bigcap\pi\big|_n\neq\mathbf{0}_n\Big)\\
&= \widehat{\mu_n}\Big(\mathscr{P}_\infty(\pi')\bigcap|\pi|^{\downarrow}\neq 0\Big)\\
&= \int_{\mathscr{P}_{[0,1]}}\varrho_\rho(\mathscr{P}_\infty(\pi'))\ \ \nu(d\rho).
\end{aligned}$$

Since this holds for every $\pi' \in \mathscr{P}_k \setminus \mathbf{0}_k$ and $n \geq k$, we conclude that $\mu_\nu = \mathbb{1}_{\{|\pi|^{\downarrow}\neq\mathbf{0}\}}\mu$. Let us now characterize $\mathbb{1}_{\{|\pi|^{\downarrow}=\mathbf{0}\}}\mu$. Let $\tilde{\mu}$ be the restriction of μ to the set $\mathscr{P}_\infty(\{1,2\})\cap\{\pi\in\mathscr{P}_\infty : |\pi|^{\downarrow}=0\}$. Then, the image measure of $\tilde{\mu}$ under the translation by 2 is a finite exchangeable measure, with support on the set of partitions with asymptotic frequencies equal to zero and, thus, by Kingman's representation, is entirely concentrated on $\mathbf{0}_\infty$. Therefore, $\tilde{\mu}$ almost-everywhere, the block containing 1 and 2 can either be of the form $\{1,2\}$, or $\{1,2,j\}$ for some $j\in\mathbb{N}$. Since $\tilde{\mu}$ is finite and μ is exchangeable, using a similar argument as in Lemma 1.4.3 (i.e. constructing a 'uniform' finite measure on $\mathbb{N}$), we see that the block containing 1 and 2 is actually $\{1,2\}$ $\tilde{\mu}$-a.e. Thus $\tilde{\mu}$ is entirely concentrated on the partition $\pi_{1\sim 2}$ and, in fact, $\tilde{\mu}(\pi_{1\sim 2}) = \mu(\mathscr{P}_\infty(\{1,2\})\cap|\pi|^{\downarrow}=0)$. Let $c \coloneqq \mu(\mathscr{P}_\infty(1\sim 2)\cap|\pi|^{\downarrow}=0)$, then by the exchangeability of μ we have

$$\mathbb{1}_{\{|\pi|^{\downarrow}=\mathbf{0}\}}\mu = c\mu_K,$$

since $\mu(\mathscr{P}_\infty(\{1,2\})\cap|\pi|^{\downarrow}=0) = \mu(\mathscr{P}_\infty(\{i,j\})\cap|\pi|^{\downarrow}=0)$ for every pair of integers i,j. Thus, joining both equalities we get:

$$\mu = c\mu_K + \mu_\nu.$$

□

Chapter 3

The Bolthausen-Sznitman Coalescent

In this chapter we will first introduce the class of simple coalescents, provide an alternative construction of this class due to Pitman, and also alternative proof for the existence and characterization of their coagulation rates. Then we will specialize in a particular type of simple coalescent: the Bolthausen-Sniztman (BS) coalescent. We will describe an alternative construction of BSC based on random recursive trees, and a copuling method with random walks that will provide elegant proofs for the study of different functionals on the BS coalescent such as the total number of jumps.

3.1 Simple Coalescents

3.1.1 Definition

The poissonian construction given in Section 2.3 tells us that we can intuitively think of exchangeable coalescents as a process where one selects a number of time points $\{T_i\}_{i\in\mathbb{N}}$ according to a Poisson process on $\mathbb{R}^+$ and then for each time point T_i one picks an "increment" Π_i according to some distribution in $\mathscr{P}_\infty$. The coalescent process at time t is then constructed through the sequential coagulation prescribed by all the increments that occur before time t. Until now we have considered the general case in which the increments $\{\Pi_i\}_{i\in\mathbb{N}}$ may prescribe the simultaneous coalescence of multiple groups of blocks at the same time; we will now focus on processes whose

infinitesimal increments prescribe the coalescence of at most one group of blocks at a time. The following definitions make this intuition precise.

Definition 3.1.1. Let $\pi \in \mathscr{P}_\infty$ be a partition, and Π be an exchangeable coalescent.

- We say that π is a *simple partition* if all its blocks, except possibly one, are singletons.
- We say that Π is *simple* if its coagulation rate is supported on simple partitions.

We note that if μ is the coagulation rate of a simple coalescent then the image measure of μ on $\mathscr{P}_{[0,1]}$ under the map $\pi \to |\pi|^{\downarrow}$ is supported on mass partitions of the form $\rho = (p, 0, 0, \cdots)$, $p \in [0,1]$. Therefore, if ν is the measure on $\mathscr{P}_{[0,1]}$ such that $\mu = c\mu_K + \mu_\nu$, then ν is also supported on mass partitions of this form, so we can simplify ν to a measure Λ on $[0,1]$ by setting $\Lambda(A) = \nu(A \times \mathbf{0})$ for every borel set A in $[0,1]$. In this case we also have that $\Lambda(0) = 0$ and, by equation (2.2):

$$\int_{[0,1]} p^2 \ \Lambda(dp) < \infty.$$

The above discussion can also be read in reverse, that is, for any measure Λ in $[0,1]$ such that $\Lambda(0) = 0$ and $\int_{[0,1]} p^2 \ \Lambda(dp) < \infty$ we can construct a measure ν on $\mathscr{P}_{[0,1]}$ which corresponds to a simple coalescent. For this reason from now on we will write ν instead of Λ for the measure on $[0,1]$ associated to a simple coalescent and, by a slight abuse of notation, we will say that the coagulation rate μ of a simple coalescent is given by $\mu = c\mu_K + \mu_\nu$ where ν is a measure on $[0,1]$. Given the coagulation rate of a simple coalescent $\mu = c\mu_K + \mu_\nu$ we interpret c as the intensity with which increments coalesce pairs of blocks, and μ_ν as the intensity with which a proportion p of all the blocks is coalesced instead. Also, if we consider the restriction of Π to $\mathscr{P}_n$, then, for any simple partition $\pi \in \mathscr{P}_n \setminus \mathbf{0}_n$ such that its non-singleton block has k elements ($k \in [n] \setminus \{1\}$), we have

$$\alpha_\pi = c\mathbb{1}_{k=2} + \int_{[0,1]} p^k (1-p)^{n-k} \ \nu(dp);$$

so if we define $\alpha_{n,k} \coloneqq c\mathbb{1}_{k=2} + \int_{[0,1]} p^k(1-p)^{n-k} \ \nu(dp)$, then $\alpha_{n,k}$ gives the intensity with which any particular collection of k blocks coalesce whenever

there are $n \geq k$ blocks. Moreover, the intensities $\alpha_{n,k}$ satisfy the recursion

$$\alpha_{n,k} = \alpha_{n+1,k} + \alpha_{n+1,k+1} \tag{3.1}$$

since the rate at which a collection of k blocks coalesce when there are n blocks equals the rate, when there are $n+1$ blocks, at which they coalesce along with the $(n+1)th$ block plus the rate at which they coalesce excluding the $(n+1)th$ block; more precisely, if B is the non-singleton block of a simple partition $\pi \in \mathscr{P}_n$, and π', π'' are the simple partitions in $\mathscr{P}_{n+1}$ with non-singleton blocks B and $B \cup \{n+1\}$ respectively, then, by the additivity of the coagulation rate, we have:

$$\alpha_\pi = \alpha_{\pi'} + \alpha_{\pi''},$$

so (3.1) follows.

Furthermore, if A_k is the set of all simple partitions in $\mathscr{P}_n$ such that their non-singleton element has k elements ($k \in [n] \setminus \{1\}$), then

$$\begin{aligned} \lambda_{n,k} &:= \sum_{\pi \in A_k} \alpha_\pi \\ &= (\#A_k)\alpha_{n,k} \\ &= \binom{n}{k} \alpha_{n,k} \end{aligned}$$

gives the rate at which a coalescence of exactly k blocks occurs, whenever there are $n \geq k$ blocks; and

$$\alpha_n := \sum_{k=1}^{n} \lambda_{n,k}$$

gives the total coagulation rate.

3.1.2 Pitman's Construction of Simple Coalescents [12]

In Chapter 2 we described a characterization of coalescent processes via its coagulation rate μ on $\mathscr{P}_\infty$. In this chapter we describe Pitman's construction of simple coalescents whose only input is an array of nonnegative numbers $(\alpha_{n,k}, 2 \leq k \leq n < \infty)$ satisfying recursion (3.1). The construction is based on an application of Kolmogorov's consistency theorem. In pursue of this

we will first introduce another definition of the set $\mathscr{P}_\infty$ in terms of product spaces. By Lemma 1.1.1 there is a one to one correspondence between $\mathscr{P}_\infty$ and compatible sequences of partitions $(\pi^1, \pi^2, \cdots) \in \otimes_{k=1}^{\infty} \mathscr{P}_k$, so we may consider $\mathscr{P}_\infty$ as a subset of $\otimes_{k=1}^{\infty} \mathscr{P}_k$. Furthermore, we can endow $\mathscr{P}_\infty$ with the topology it inherits from the product topology in $\otimes_{k=1}^{\infty} \mathscr{P}_k$, where each $\mathscr{P}_k$ is endowed with the discrete topology. A coalescent Π is then a process that takes values in $\mathscr{P}_\infty$, with càdlàg paths, and such that for every pair $s < t \in \mathbb{R}^+$, $\Pi(t)$ is the result of a coagulation operation on $\Pi(s)$. On the other hand, coalescent processes in $\mathscr{P}_n$ can be constructed as Markov chains satisfying the same restriction for $\Pi(t)$ and $\Pi(s)$ $(s < t)$ as above.

Theorem 3.1.1. For each $n \in \mathbb{N}$ let Π^n be a coalescent processes in $\mathscr{P}_n$ with transition rates given by the rule that, whenever there are b blocks, any given collection of k blocks coalesce into a single block with intensity $\alpha_{b,k}$. Then, equation (3.1) holds if and only if for every pair of integers $m < n$, the processes Π^m and $\Pi^n\big|_m$ have the same distribution.

Proof. By the theory of Markov processes we need only verify that the transition rates of the two processes Π^m and $\Pi^n\big|_m$ coincide. This is done following the same lines as in the proof of the additivity of the measure $\hat{\mu}$ in Theorem 2.2.4. □

By the previous theorem, if the collection of nonnegative numbers $(\alpha_{n,k}, 2 \leq k \leq n < \infty)$ satisfies (3.1), then the corresponding collection of processes $\{\Pi^n\}_{n=1}^{\infty}$ generates a consistent family of probability measures $\{\mu_n\}_{n=1}^{\infty}$ on $\otimes_{k=1}^{n} \mathscr{P}_k^{\mathbb{R}^+}$ and, by Kolmogorov's consistency theorem, there exists a stochastic process Π on $\otimes_{k=1}^{\infty} \mathscr{P}_k^{\mathbb{R}^+}$ with finite dimensional distributions $\{\mu_n\}_{n=1}^{\infty}$. Since the probability measures μ_n are supported on compatible sequences of partitions (again by theorem 3.1.1), then Π takes values on $\mathscr{P}_\infty \subset \otimes_{k=1}^{\infty} \mathscr{P}_k$. Furthermore, since $\{\Pi^n\}_{n=1}^{\infty}$ are coalescent processes, it is easily seen that for any pair $s < t \in \mathbb{R}^+$, $\Pi(t)$ is the result of a coagulation operation on $\Pi(s)$, so Π is indeed a coalescent process.

Finally, the next result makes use of de Finetti's theorem to give an alternative proof for the existence of a finite nonnegative measure μ on $[0, 1]$ which characterizes the coagulation rate of a simple coalescent Π.

Theorem 3.1.2. There exists a bijection between collections of nonnegative numbers $(\alpha_{n,k}, 2 \leq k \leq n < \infty)$ that satisfy (3.1), and finite nonnegative

measures μ on $[0,1]$, that satisfy:

$$\alpha_{i+j+2,i+2} = \int_{[0,1]} x^i(1-x)^j \mu(dx). \tag{3.2}$$

Proof. Note that if $\alpha_{i+j+2,i+2} = \int_{[0,1]} x^i(1-x)^j\mu(dx)$ for some finite measure μ on $[0,1]$, then substituting the integrals in (3.1) we get:

$$\begin{aligned}
\alpha_{i+j+2+1,i+2} + \alpha_{i+j+2+1,i+2+1} &= \int_{[0,1]} x^i(1-x)^{j+1}\mu(dx) + \int_{[0,1]} x^{i+1}(1-x)^j\mu(dx) \\
&= \int_{[0,1]} (1-x)^j(x^i(1-x)+x^{i+1})\mu(dx) \\
&= \int_{[0,1]} (1-x)^j x^i \mu(dx) \\
&= \alpha_{i+j+2,i+2}
\end{aligned}$$

so (3.1) does hold.

Now, for the reverse implication we will make use of the preceding theorem and interpret the collection $(\alpha_{n,k}, 2 \le k \le n < \infty)$ as the transition rates of a simple coalescent Π. For each $n \ge 2$ let τ_n be the first jumping time of $\Pi\big|_n$. For $n \ge 2$ define the probability measure μ_n on $\{0,1\}^n$ by the following procedure: let $\Delta = (\delta_1, \cdots, \delta_n)$ be a n-tuple of zero-one digits and construct π_Δ as the partition given by putting all the indices i with $\delta_i = 1$ in a single block and leaving the rest to form singletons. Then define:

$$\mu_n(\Delta) = \mathbb{P}\bigg(\Pi\big|_n(\tau_n) = \pi_\Delta \bigg| \Pi\big|_2(\tau_n) = \{1,2\}\bigg).$$

We now show that if μ_1 is given by $\mu_1(\{1\}) = 1$, then $\{\mu_n\}_{n\ge1}$ is a consistent family of probability measures. Indeed, for any sequence $\Delta = (\delta_1, \cdots, \delta_n)$ in $\{0,1\}^n$ with $\delta_1 = \delta_2 = 1$ let $k = \sum_{i=1}^n \delta_i$ and note that by the additivity of the transition rates in Theorem 3.1.1 we have:

$$\begin{aligned}
\mu_n(\Delta) =& \frac{\alpha_{n,k}}{\alpha_2} \\
=& \frac{\alpha_{n+1,k+1} + \alpha_{n+1,k}}{\alpha_2} \\
=& \mu_{n+1}(\Delta \cup \delta_{n+1} = 1) + \mu_{k+1}(\Delta \cup \delta_{n+1} = 0) \\
=& \mu_{n+1}(\Delta \times \{0,1\}) \\
=& \mu_{n+1}\big|_{\{0,1\}^n}(\Delta);
\end{aligned}$$

whereas if $\delta_1 = 0$ or $\delta_2 = 0$ then $\mu_(\Delta) = 0 = \mu_{n+1}\big|_{\{0,1\}^n}(\Delta)$. Now define $X_1 = 1$ and for $k \geq 2$:

$$X_k \coloneqq \begin{cases} 1 & \text{if } k \text{ participates in the first coalescence event of } \Pi\big|_k \text{ along with 1 and 2.} \\ 0 & \text{otherwise.} \end{cases}$$

By the exchangeability of Π, and thus of $\Pi\big|_k$ for every $k \geq 2$, and conditional on the event $X_1 = X_2 = 1$, the sequence $\{X_k\}_{k=1}^{\infty}$ is exchangeable. Therefore by de Finetti's theorem there exists a probability measure $\hat{\mu}$ on $[0,1]$ such that, for any $\Delta = (\delta_1, \cdots, \delta_n)$ with $\delta_1 = \delta_2 = 1$, $i{+}2 = \sum \delta_k$, and $i{+}2{+}j = n$, we have:

$$\mathbb{P}(X_3 = \delta_3, \cdots, X_n = \delta_n \mid X_1 = X_2 = 1) = \int_{[0,1]} x^i (1-x)^j \hat{\mu}(dx).$$

On the other hand, by the consistency of the sequence $\{\mu_n\}_{n\geq 2}$ we can compute the same conditional probability using the coagulation rates of $\Pi\big|_{i+j+2}$, we have:

$$\begin{aligned} \frac{\alpha_{i+j+2,i+2}}{\alpha_{2,2}} &= \frac{\alpha_{i+j+2,i+2}}{\alpha_{i+j+2}} \frac{\alpha_{i+j+2}}{\alpha_{2,2}} \\ &= \frac{\mathbb{P}(X_1 = 1, X_2 = 2, X_3 = \delta_3, \cdots, X_{i+j+2} = \delta_{i+j+2})}{\mathbb{P}(X_1 = 1, X_2 = 1)} \\ &= \mathbb{P}(X_3 = \delta_3, \cdots, X_{i+j+2} = \delta_{i+j+2} | X_1 = X_2 = 1) \end{aligned}$$

Hence, if $\mu \coloneqq \alpha_{2,2}\hat{\mu}$, we obtain (3.2). □

3.1.3 The Bolthausen-Sznitman Coalescent

The Bolthausen-Sznitman (BS) coalescent is the simple coalescent given by the coagulation rate $\mu = \mu_\nu$ where $\nu(dx) = x^{-2}dx$, that is, the intensity of pair-wise coalescence is zero and the intensity of coalescing a proportion dx of all the blocks is given by $\nu(dx) = x^{-2}dx$. From the discussion in Section

3.1 we obtain the equations

$$\begin{aligned}\alpha_{n,k} &= \int_{[0,1]} p^k (1-p)^{n-k} \nu(dp) \\ &= \int_{[0,1]} p^{k-2} (1-p)^{n-k} dp \\ &= \frac{\Gamma(k-1)\Gamma(n-k+1)}{\Gamma(n)} \\ &= \frac{(k-2)!(n-k)!}{(n-1)!},\end{aligned}$$

and

$$\begin{aligned}\lambda_{n,k} &= \binom{n}{k} \frac{(k-2)!(n-k)!}{(n-1)!} \\ &= \frac{n!}{(n-k)!k!} \frac{(k-2)!(n-k)!}{(n-1)!} \\ &= \frac{n}{k(k-1)}.\end{aligned}$$

We also have:

$$\alpha_n = n - 1.$$

The BS coalescent is a particular case of the Beta coalescent. The Beta coalescent of parameter a, $0 < a \leq 2$, is the simple coalescent given by the coagulation rate $\mu = \mu_\nu$ where $\nu(dx) = x^{-2} Beta(2-a, a)dx$, and $Beta(2-a, a)$ is the kernel of the Beta distribution of parameters $(2 - alpha, alpha)$; that is:

$$\nu(dx) = x^{-2} \frac{1}{\Gamma(2-a)\Gamma(a)} x^{1-a}(1-x)^{a-1}.$$

Thus, the BS coalescent is the Beta coalescent of parameter 2.

3.2 Random Recursive Trees and the BSC

3.2.1 Construction of the BSC through Random Recursive Trees

In this section we will present a construction of the BS coalescent using random recursive trees due to Goldschmidt and Martin [1]. Let $L = \{B_1, \cdots, B_n\}$

be a partition of $[m]$ for some $m \in \mathbb{N}$, or a subset of such a partition, and endow L with the total order given by ordering the blocks of a partition by their least elements as described in Section 2.1. A *recursive tree* on n vertices with labels L is a tree rooted at $\{B_1\}$, and such that the paths from the root to the leaves are increasing according to the order of its labels. A *random recursive tree* on $\{B_1, \cdots, B_n\}$ is a tree chosen uniformly at random among the $(n-1)!$ possible recursive trees. We can construct a random recursive tree by fixing the root at the node with the label B_1 and adding the rest of the nodes sequentially, by uniformly choosing a parent from the already existing nodes. We also define a *cutting-merge* procedure on such trees which consists of selecting one edge uniformly at random (cutting), and merging all the nodes below the edge with the node above it (merge); thus obtaining a new recursive tree on a set of labels that constitute a new partition of $[n]$, or a subset of such a partition (see Figure 3.1).

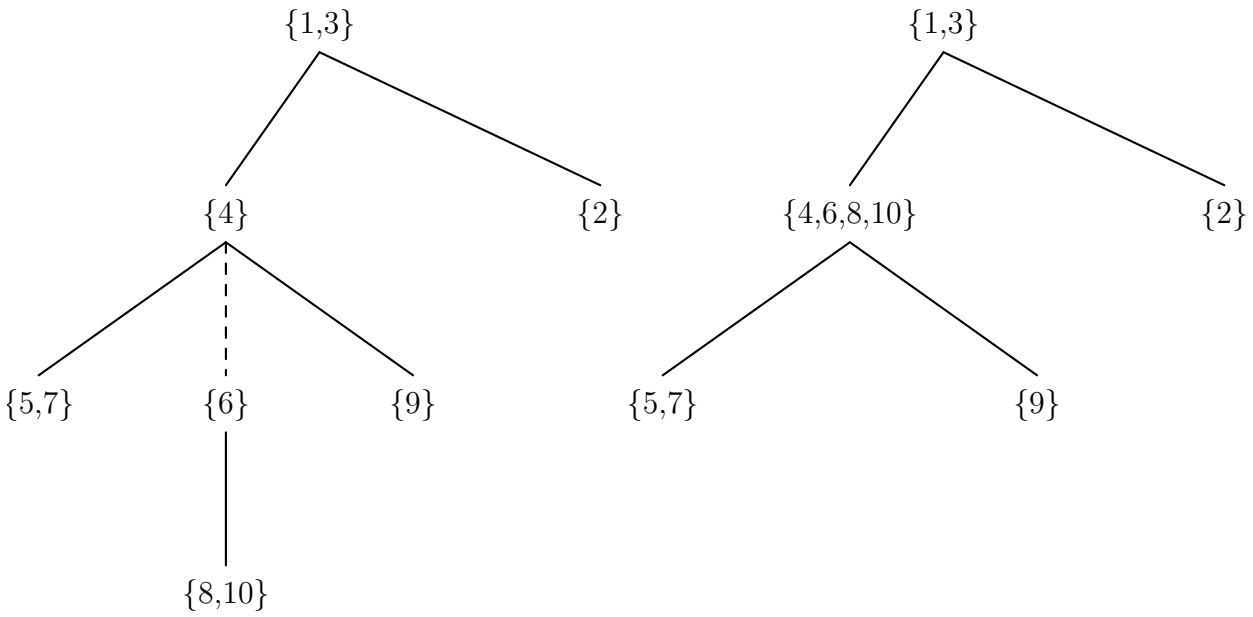

Figure 3.1: On the left, an example of a recursive tree whose labels constitute a partition of $\{1, \cdots, 10\}$ with 7 blocks. On the right, the resulting recursive tree after a cutting-merge procedure performed on the marked edge (dashed line) of the first tree.

Lemma 3.2.1. Let T be a random recursive tree on a set of n labels, then the tree resulting from performing a cutting-merge procedure on T is a random recursive tree on the resulting set of labels.

Proof. Let T be a random recursive tree on the set of labels $L \coloneqq \{B_1, \cdots, B_n\}$, L_k be a subset of L of size k, $\hat{B}_m$ be the minimal element of L_k, and

$$\hat{B} \coloneqq \bigcup_{B \in L_k} B.$$

We need to show that, conditionally on the event that the resulting tree after the cutting-merge procedure performed on T has labels $\{L \setminus L_k\} \cup \{\hat{B}\}$, the resulting tree is uniformly distributed on the set of all possible recursive trees on that label set. It is easy to see that the tree resulting from a cutting-merge procedure performed on a recursive tree is again a recursive tree. Now, fix any recursive tree T' on the set of labels $\{L \setminus L_k\} \cup \{\hat{B}\}$, then all the trees on the label set $\{B_1, \cdots, B_n\}$ that end in T' after a cutting-merge procedure can be constructed by replacing the label $\hat{B}$ with the label $\hat{B}_m$ on T', and picking a recursive tree T^* on the label set $L_k \setminus \{\hat{B}_m\}$ and joining T^* with T' by adding an edge between the root of T^* and the node with label $\hat{B}_m$ on T'. Since the number of possible recursive trees on the label set $L_k \setminus \{\hat{B}_m\}$ is $(k-2)!$, the total number of initial configurations of T that result on T' after a cutting-merge procedure is $(k-2)!$. On the other hand, all the configurations of T that result in a recursive tree on the label set $\{L \setminus L_k\} \cup \{\hat{B}\}$ can be constructed by picking two recursive trees T' and T^*, the former on the label set $\{L \setminus L_k\} \cup \{\hat{B}_m\}$ and the latter on $L_k \setminus \{\hat{B}_m\}$, and joining them by adding an edge between the root of T^* and the node with label $\hat{B}_m$ on T'. Thus, since there are $(n-k)!$ possible recursive trees on $\{L \setminus L_k\} \cup \{\hat{B}_m\}$, and $(k-2)!$ possible recursive trees on $L_k \setminus \{\hat{B}_m\}$, the total number of initial configurations of T that result in a recursive tree on the label set $\{L \setminus L_k\} \cup \{\hat{B}\}$ is $(n-k)!(k-2)!$. Finally, since T is uniformly distributed among all possible recursive trees on $\{B_1, \cdots, B_n\}$, the probability of obtaining any particular recursive tree on the label set $\{L \setminus L_k\} \cup \{\hat{B}\}$, conditioned on obtaining precisely this label set, is given by $\frac{(k-2)!}{(n-k)!(k-2)!} = \frac{1}{(n-k)!}$. Since this probability is the same for any such tree, we have proven that the resulting recursive tree after a cutting-merge procedure on T is again a random recursive tree on the resulting set of labels. □

Theorem 3.2.2. Let T be a random recursive tree with labels $\mathbf{0}_n$. Associate to each edge of T an independent exponential random variable with mean 1. The

exponential variables will be the time at which a cutting-merge procedure will occur at a particular edge, with the addition that all the labels on the subtree below that edge will be instantly added to the node above. If for each time we construct a partition $\Pi(t)$ of $[n]$ whose blocks are given by the labels on each of the nodes of T, then $\mathit{\Pi} = (\Pi(t), t > 0)$ is a BS coalescent restricted to $[n]$.

Proof. We need to show that the probability with which any particular set of k blocks coalesces whenever there are b blocks is $\frac{(k-2)!(b-k)!}{(b-1)!(b-1)}$, since in this case the coagulation rate of the selected blocks would be $\frac{(k-2)!(b-k)!}{(b-1)!}$ as in the BS coalescent. The latter follows from the fact that the total coagulation rate is $b-1$, since there are $b-1$ possible coagulations given by a cutting-merge procedure on a recursive tree with b blocks. Using a similar argument as in the preceding lemma we note that the probability that any set of k blocks coalesces when there are b blocks can be computed by counting the number of recursive trees on the label set $\pi = \{B_1, \cdots, B_b\}$ that result in the coalescence of the selected k blocks, denoted by L_k. This is the same as counting the number of recursive trees on $\{B_1, \cdots, B_b\}$ that result in a recursive tree on the set of labels $\{\{B_1, \cdots, B_b\} \backslash L_k\} \cup \hat{B}$ with $\hat{B}$ as in Lemma 3.2.1, where we also proved that this number is $(b-k)!(k-2)!$. Since there are $(b-1)!(b-1)$ possible ways of performing a cutting-merge procedure on the set of all recursive trees on $\{B_1, \cdots, B_b\}$ (i.e. there are $(b-1)!(b-1)$ recursive trees with a single marked edge), we see that the probability of coalescing the selected set of k blocks is

$$\frac{(b-k)!(k-2)!}{(b-1)!(b-1)},$$

so the coagulation rates do coincide with those of the BSC.

Finally, by Lemma 3.2.1, the recursive tree resulting from a cutting-merge procedure performed on T is again a recursive tree on the resulting set of labels and, therefore, the process is markovian. □

From the previous theorem we observe that to each random recursive tree on $\mathbf{0}_n$ with exponential variables associated to its edges corresponds a path of the BSC with values in $\mathscr{P}_n$. We can also define the coalescent process $\mathit{\Pi}$ in $\mathscr{P}_n$ of the previous theorem as follows: construct a random recursive tree T on $\mathbf{0}_n$ and associate to each edge an independent standard exponential random variable as before. For each $t > 0$ define $\mathit{\Pi}(t)$ to be

the random partition constructed from the tree resulting from performing a cutting-merge procedure on all the edges of T whose associated exponential variable is less than t. The resulting process has the same distribution as the process in Theorem 3.2.2. To see this let E_t be the collection of edges of T whose associated exponential variable is less than t, and let N_t be the collection of edges not contained in E_t and such that the path from the root to the node above does not contain any edge in E_t; then the path of Π after time t is entirely determined by the subsequent cutting-merge procedures performed on the edges in N_t since any cutting-merge procedure scheduled for time $t + s$, $s > 0$, on any other edge of T is ineffectual in terms of the resulting random partition $\Pi(t + s)$ (see Figure 3.2). Thus the next jump of Π after time t is determined in exactly the same way as in Theorem 3.2.2 and, therefore, Π is a BSC on $[n]$.

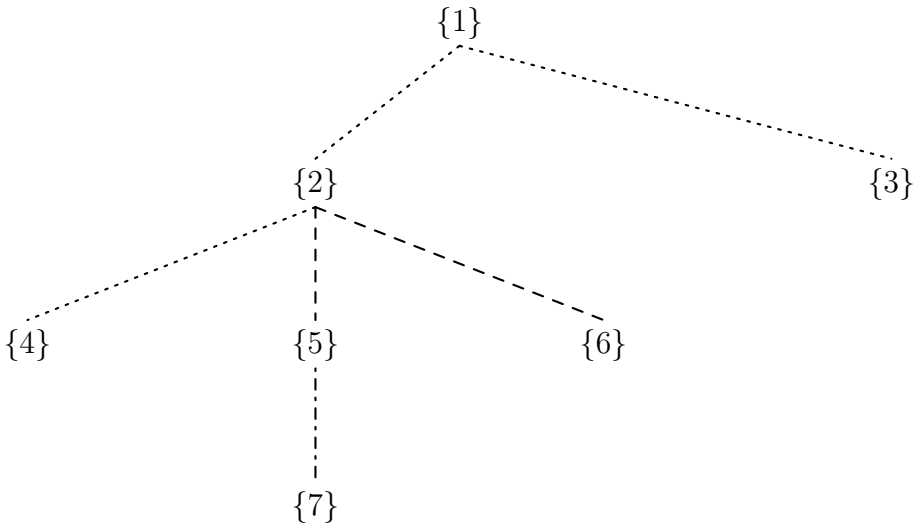

Figure 3.2: An example of a recursive tree on [7] with marked edges at time t. The dashed edges correspond to E_t, the dotted edges correspond to N_t, and the dash-dotted edge marks an ineffectual cutting-merge procedure scheduled after time t. It is clear that the behavior of Π after time t is determined by N_t in the same way as in Theorem 3.2.2

3.2.2 The Last Jump of the BSC [1]

In this section we will study the behavior as $n \to \infty$ of the distribution of J^n, the size of the last jump of the BSC restricted to $[n]$. For this we will use the representation of the BSC in terms of a random recursive tree T as described in the previous section. However, this time it will be helpful to consider the construction of T as a process in continuous time, which will allow us to make a connection with Yule processes. The Yule process is a birth process starting with one individual at time 0, and letting the individuals present at time t duplicate with intensity 1 independently from the rest of the population and from the previous generations, thus increasing the total population size by 1. Formally, a Yule process is a continuous time Markov chain on $\mathbb{N}$ starting at 1, and with transition rates $\alpha(n, n+1) = n$ and $\alpha(n, k) = 0$ for $k \neq n+1$. In order to construct a random recursive tree as a continuous-time process start with the root labeled $\{1\}$ at time 0, and think of the nodes of T as arriving sequentially, where node $n+1$ arrives and attaches to the existing n nodes with intensity 1 and in an independent manner. In particular we see that the total size of the tree is a Yule process and, moreover, the size of the growing subtree rooted at any given node is also a Yule process. Also, the number of children of $\{1\}$ is a Poisson process with rate 1. As explained in the previous section, we associate an independent standard exponential random variable to each of the edges of the arriving nodes. Thus, the arriving times of the children of $\{1\}$ along with their associated exponential variables behave as the atoms of a Poisson random measure on $\mathbb{R}^+ \times \mathbb{R}^+$ with intensity $dt \otimes e^{-x}dx$, where the first coordinate represents the arrival times of the children and the second coordinate represents the value of their associated exponential variable. Now, we may stop the construction of T at time t and obtain a recursive tree T_t on $\{1, \cdots, m\}$ where m is a random variable; to such tree corresponds a path of the BSC on $[m]$ as described in the previous section. The time of the last jump of the BSC thus constructed is given by the maximum of all the exponential variables associated to the children of $\{1\}$ that are present at time t, denote this maximum by E_t. Now let C_t be the child of $\{1\}$ associated to E_t and let $T^0_{C_t}$ be the subtree of T_t rooted at C_t. For $s > 0$ define $T^s_{C_t}$ to be the tree resulting from performing all the cutting-merge procedures on $T^0_{C_t}$ scheduled before time s according to the exponential variables of the edges of $T^0_{C_t}$. Then the size of the last jump (i.e. the number of nodes involved in the last jump) of the BSC constructed in this way, say J_t, is equal to 1 plus the number of nodes in $T^{E_t}_{C_t}$.

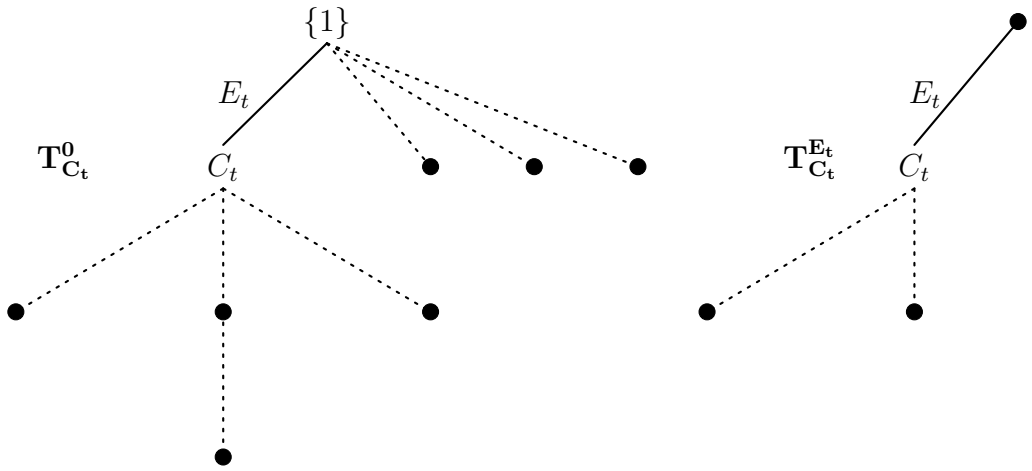

Figure 3.3: A schematic representation of the idea discussed in the main text. On the left we show the tree T_t with the node C_t, its associated exponential variable E_t, and its subtree $T^0_{C_t}$ marked. On the right we show the same tree after performing all the cutting-merge procedures scheduled before time E_t, i.e. the tree that describes the last jump.

We want to determine the distribution of J_t. As mentioned before, the size of $T^0_{C_t}$ grows as a Yule process, so if $0 \leq A_t \leq t$ is the arrival time of C_t to $\{1\}$, then the size of $T^0_{C_t}$ has distribution $Y(t-A_t)$, where Y is a Yule process. Similarly, for any $s \in \mathbb{R}^+$, the size of $T^s_{C_t}$ is distributed as $Y(e^{-s}(t-A_t))$. Indeed, more generally we have that if $\hat{Y}$ is the process constructed as a Yule process but with the addition that its increments are kept with probability p and discarded with probability $1-p$, $p \in (0,1)$, then $\hat{Y}(t) \stackrel{d}{=} Y(pt)$ for all $t > 0$. In our case the size of $T^{E_t}_{C_t}$ grows as a Yule process whose increments are kept with probability e^{-E_t}; therefore, J_t, which equals 1 plus the size of $T^{E_t}_{C_t}$, is distributed as $1 + Y(e^{-E_t}(t-A_t))$. Now, using the Poisson random measure described above for the arrival times of the children of $\{1\}$ and the values of their associated exponential variables, we see that A_t is uniformly distributed on $[0,t]$, and that the distribution function of E_t is given by:

$$\mathbb{P}(E_t \leq u) = e^{-te^{-u}} \tag{3.3}$$

which is the probability that the Poisson random measure on $[0,t] \times (u, \infty)$ is equal to zero. Thus, if U_1 and U_2 are two independent uniform random variables on $[0,1]$, then

$$A_t \stackrel{d}{=} tU_1$$

and

$$e^{-E_t} \stackrel{d}{=} \frac{-\log(U_2)}{t} \stackrel{d}{=} \frac{E}{t},$$

where E is a standard exponential random variable. Therefore J_t has distribution:

$$J_t \stackrel{d}{=} 1 + Y(e^{-E_t}(t - A_t)) \stackrel{d}{=} 1 + Y\left(\frac{E}{t} tU_1\right) = 1 + Y(EU_1)$$

where E and U_1 are independent.

Theorem 3.2.3. Let J^n be the size of the last jump of the BSC with n initial blocks. Then, as $n \to \infty$ we have:

$$J^n \stackrel{d}{\to} 1 + Y(EU)$$

where Y, E and U are independent; Y is a Yule process, E is a standard exponential variable, and U is a uniform variable.

Proof. In the above discussion we derived the distribution of J_t, the last jump of a BSC constructed from random a recursive tree, and whose number of initial blocks was random and distributed as $Y(t)$ with Y a Yule process. However, we want to study the distribution of the size of the last jump of the BSC starting with precisely $n \in \mathbb{N}$ blocks, and determine the limiting distribution as $n \to \infty$. In order to tackle this consider the stopping time $\tau_n := \inf\{t \in \mathbb{R}^+ : T_t \text{ has } n \text{ nodes}\}$, where T_t is as above. We first study the behavior of τ_n as $n \to \infty$ in order to establish deterministic lower and upper bounds for τ_n, say t_n^- and t_n^+, with high probability as $n \to \infty$. This in turn will allow us to approximate J_{τ_n} by $J_{t_n^-}$ and $J_{t_n^+}$. In pursue of this, note that the distribution of τ_n is given by:

$$\tau_n \stackrel{d}{=} E_1 + \frac{E_2}{2} + \cdots + \frac{E_n}{n}$$

where $\{E_j\}_{j \in \mathbb{N}}$ are independent standard exponential variables. Thus,

$$\mathbb{E}(\tau_n) = 1 + \frac{1}{2} + \cdots + \frac{1}{n-1} = \log n + \mathcal{O}(1),$$

and

$$V(\tau_n) = 1 + \frac{1}{2^2} + \cdots + \frac{1}{(n-1)^2}.$$

Therefore we have as $n \to \infty$:

$$\frac{\mathbb{E}((\tau_n - \log n)^2)}{\log n} = \frac{V(\tau_n) + \mathbb{E}((\tau_n - \mathbb{E}(\tau_n))\mathcal{O}(1)) + \mathcal{O}(1)}{\log n} \to 0,$$

and by Markov's inequality:

$$\mathbb{P}(|\tau_n - \log n| > (\log(n))^{1/2}) \to 0 \quad \text{as} \quad n \to \infty;$$

that is, if $t_n^- \coloneqq \log n - (\log n)^{1/2}$ and $t_n^+ \coloneqq \log n + (\log n)^{1/2}$, then:

$$\mathbb{P}(\tau_n \in (t_n^-, t_n^+)) \to 1 \quad \text{as} \quad n \to \infty. \tag{3.4}$$

Now, for C_t as above, we prove that

$$\mathbb{P}(C_{t_n^-} = C_{\tau_n} = C_{t_n^+}) \to 1 \quad \text{as} \quad n \to \infty.$$

For this, note that

$$\mathbb{P}(C_{t_n^-} = C_{\tau_n} = C_{t_n^+}) \geq \mathbb{P}(\tau_n \in (t_n^-, t_n^+), C_{t_n^-} = C_{t_n^+})$$

so, by (3.4), we need only see that

$$\mathbb{P}(C_{t_n^-} = C_{t_n^+}) \to 1 \quad \text{as} \quad n \to \infty.$$

The latter follows from the fact that $C_{t_n^-} \neq C_{t_n^+}$ if and only if $A_{t_n^+} \in (t_n^-, t_n^+)$, where we use the notation A_t for the arrival time of C_t to $\{1\}$ as before. Since $A_{t_n^+}$ is uniformily distributed on $[0, t_n^+]$, we have:

$$\mathbb{P}(A_{t_n^+} \in (t_n^-, t_n^+)) = \frac{t_n^+ - t_n^-}{t_n^+} \to 0 \quad \text{as} \quad n \to \infty.$$

Observe that on the set $G_n \coloneqq \{\tau_n \in (t_n^-, t_n^+), C_{t_n^-} = C_{t_n^+}\}$ we also have that $E_{t_n^-} = E_{\tau_n} = E_{t_n^+}$ almost surely, and, therefore, $J_{t_n^-} \leq J_{\tau_n} \leq J_{t_n^+}$. Thus, using the distributions of $J_{t_n^-}$ and $J_{t_n^+}$ derived above, we have:

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}(J_{\tau_n} \leq m) &= \lim_{n\to\infty} \mathbb{P}(J_{\tau_n} \leq m, G_n) \\
&\geq \lim_{n\to\infty} \mathbb{P}(J_{t_n^+} \leq m, G_n) \\
&= \lim_{n\to\infty} \mathbb{P}(J_{t_n^+} \leq m) \\
&= \mathbb{P}(1 + Y(EU) \leq m),
\end{aligned}$$

and, similarly:

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(J_{\tau_n} \leq m) &\leq \lim_{n\to\infty} \mathbb{P}(J_{t_n^-} \leq m) \\ &= \mathbb{P}(1 + Y(EU) \leq m);\end{aligned}$$

that is:

$$J_{\tau_n} \xrightarrow{d} 1 + Y(EU) \quad \text{as} \quad n \to \infty.$$

Finally, since $J_{\tau_n} \stackrel{d}{=} J^n$ we see that $J^n \xrightarrow{d} 1 + Y(EU)$ as $n \to \infty$. □

From the derived distribution of E_t and from equation (3.4) we can also determine the limit distribution of the time to absorption of the BSC on $[n]$ as $n \to \infty$; that is, the limit distribution of E_{τ_n} as $n \to \infty$.

Theorem 3.2.4. We have, as $n \to \infty$:

$$E_{\tau_n} - \log\log n \xrightarrow{d} -\log E,$$

where E is a standard exponential random variable.

Proof. From the definition of E_t, i.e. the maximum of the exponential variables associated to the children of $\{1\}$ present at time t, we see that if $s < t$ then $E_s \leq E_t$. Therefore, if t_n^-, t_n^+, and τ_n are as before, then from equation (3.4) we see that:

$$\mathbb{P}(E_{t_n^-} \leq E_{\tau_n} \leq E_{t_n^+}) \to 1 \quad \text{as} \quad n \to \infty$$

Now, from equation (3.3) we see that

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(E_{t_n^-} - \log\log n \leq x) &= \lim_{n\to\infty} \exp\Big(-\frac{t_n^-}{\log n} e^{-x}\Big) \\ &= \exp\big(- e^{-x}\big);\end{aligned}$$

that is,

$$E_{t_n^-} - \log\log n \xrightarrow{d} -\log\big(-\log U\big) \stackrel{d}{=} -\log E$$

where U is a uniform random variable and E is a standard exponential variable. Following the same lines for t_n^+ we also obtain

$$E_{t_n^+} - \log\log n \xrightarrow{d} -\log E.$$

Therefore:

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}(E_{\tau_n}-\log\log n\leq x) &= \lim_{n\to\infty}\mathbb{P}(E_{\tau_n}-\log\log n\leq x, E_{t_n^-}\leq E_{\tau_n})\\ &\geq \lim_{n\to\infty}\mathbb{P}(E_{t_n^-}-\log\log n\leq x)\\ &= \mathbb{P}(-\log E\leq x);\end{aligned}$$

and, similarly:

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}(E_{\tau_n}-\log\log n\leq x) &\leq \lim_{n\to\infty}\mathbb{P}(E_{t_n^+}-\log\log n\leq x)\\ &= \mathbb{P}(-\log E\leq x).\end{aligned}$$

□

3.3 Random Walks with Barrier and the BSC

Random walks with barrier have been extensively used in the study of the BSC, in particular in the study of the Markov process given by the evolution of the number of its blocks. In short, in [7] Iksanov and Möhle prove that for large n this process behaves as a random walk with barrier. In this section we present the techniques they used to prove this result and its application to the study of the total number of jumps of the BSC. This result and the techniques developed in [7] have also been used in the study of the total branch length of the BSC [4], and the length of external and internal branches [9].

3.3.1 Number of Jumps of the BSC [7]

Let X_n be the number of coalescence events that occur before arriving at the state $\{1,\cdots,n\}$ in a simple coalescent with values in $\mathscr{P}_n$; that is, the number of coalescence events until all blocks have coalesced. The aim of this section is to derive a weak law for X_n as $n\to\infty$ in the particular case of the BS coalescent. We will prove that

$$\frac{(\log n)^2}{n}X_n-\log n-\log\log n\xrightarrow{d}Z \quad \text{as} \quad n\to\infty$$

where Z is a stable random variable with index 1 and characteristic function $t\mapsto\exp\big(-\frac{\pi}{2}|t|+it\log|t|\big)$, $t\in\mathbb{R}$.

In pursue of this, we will consider the coupled process $(R^n_k, S_k)_{k=0}^{\infty}$, which depends on $n \in \mathbb{N}$, where $S \coloneqq (S_k)_{k=0}^{\infty}$ is a random walk with increments in $\mathbb{N}$, and $R^n \coloneqq (R^n_k)_{k=0}^{\infty}$ is the same random walk but with barrier n. That is, if $(\xi_k)_{k=1}^{\infty}$ is a collection of independent identically distributed random variables with values in $\mathbb{N}$, consider the coupled process:

$$\begin{aligned}(R^n_0, S_0) &\coloneqq (0,0),\\ (R^n_k, S_k) &\coloneqq (R^n_{k-1}, S_{k-1}) + \begin{cases}(\xi_k, \xi_k) & \text{if } \ R_{k-1} + \xi_k < n\\ (0, \xi_k) & \text{otherwise.}\end{cases}\end{aligned}$$

Also consider the measurable function on the canonical space $M : \mathbb{N}^{\mathbb{N}} \to \mathbb{N}$ given by $X \mapsto \#\{k \in \mathbb{N} : X_k \neq X_{k-1}\}$, and define the random variables $M_n \coloneqq M(R^n)$, i.e. the number of jumps of R^n. We will show that under a suitable choice for the distribution of ξ we have $M_n \overset{d}{=} X_n$. First, observe that M_n satisfies the distributional recursion:

$$\mathbb{P}(M_n = j) = \sum_{\ell=1}^{n-1} \frac{\mathbb{P}(\xi = \ell)}{1 - \mathbb{P}(\xi \geq n)} \mathbb{P}(M_{n-\ell} = j-1). \tag{3.5}$$

Indeed, since R^n is a Markov process, and $\tau \coloneqq \inf\{k \in \mathbb{N} : R^n_0 \neq R^n_k\}$ (the time of first jump) is a stopping time, by the strong Markov property we have:

$$\begin{aligned}\mathbb{P}(M_n = j) &= \mathbb{P}(R^n_\tau < n, M(\theta_\tau \circ R^n) = j-1)\\ &= \sum_{\ell=1}^{n-1} \mathbb{P}(R^n_\tau = \ell)\mathbb{P}(M(\theta_\tau \circ R^n) = j-1 \mid R^n_\tau = \ell).\end{aligned}$$

where θ is the usual shift operator on stochastic processes. Then, since

$$\begin{aligned}\mathbb{P}(M(\theta_\tau \circ R^n) = j-1 \mid R^n_\tau = \ell) &= \mathbb{P}(M(R^{n-\ell}) = j-1)\\ &= \mathbb{P}(M_{n-\ell} = j-1)\end{aligned}$$

and

$$\mathbb{P}(R_\tau = \ell) = \frac{\mathbb{P}(\xi = \ell)}{1 - \mathbb{P}(\xi \geq n)},$$

we obtain (3.5).

Lemma 3.3.1. If the distribution of the random variable ξ above is given by:

$$\mathbb{P}(\xi = \ell) = \frac{1}{\ell(\ell+1)},$$

then $X_n \overset{d}{=} M_n$.

Proof. To prove this we will see that both X_n and M_n satisfy the same distributional recurrence. By the recursion derived above for M_n, and substituting both $\mathbb{P}(\xi = \ell)$ and $\mathbb{P}(\xi \geq n)$, we have:

$$\begin{aligned}\mathbb{P}(M_n = j) &= \sum_{\ell=1}^{n-1} \frac{1}{\ell(\ell+1)(1-1/n)} \mathbb{P}(M_{n-\ell} = j-1) \\ &= \sum_{\ell=1}^{n-1} \frac{n}{\ell(\ell+1)(n-1)} \mathbb{P}(M_{n-\ell} = j-1).\end{aligned}$$

In order to derive the same recursion for X_n we construct the Markov chain $(Y_k)_{k=1}^{\infty}$ given by the difference in the number of blocks after each jump of the BS coalescent starting with n blocks. Using the coagulation rates of the BSC, $\lambda_{n,\ell+1} = \frac{n}{\ell(\ell+1)}$ and $\alpha_n = n-1$, we see that $\mathbb{P}(Y_1 = \ell) = \frac{\lambda_{n,\ell+1}}{\sum_{j=2}^{n} \lambda_{n,j}} = \frac{n}{\ell(\ell+1)(n-1)}$ and, by the Markov property at time 1, we have:

$$\begin{aligned}\mathbb{P}(X_n = j) &= \sum_{\ell=1}^{n-1} \mathbb{P}(Y_1 = \ell)\mathbb{P}(X_{n-\ell} = j-1) \\ &= \sum_{\ell=1}^{n-1} \frac{n}{\ell(\ell+1)(n-1)} \mathbb{P}(X_{n-\ell} = j-1).\end{aligned}$$

Hence M_n and X_n satisfy the same distributional recurrence. □

Now, in order to derive the distributional limit of M_n as $n \to \infty$ we will first need to study the distribution of the stopping time $N_n := \inf\{k \in \mathbb{N} : S_k \geq n\}$. The following lemmas describe the limiting behavior of N_n and S_{N_n} as $n \to \infty$; and the last theorem relates these two, along with M_n, in order to derive the desired weak limit for M_n.

Lemma 3.3.2. For N_n defined above, we have:

$$\frac{(\log n)^2}{n} N_n - \log n - \log\log n \xrightarrow{d} Z$$

where Z is a stable random variable with index 1 and characteristic function $t \mapsto \exp\big(-\frac{\pi}{2}|t| + it\log|t|\big)$, $t \in \mathbb{R}$.

Proof. By the theory of stable distributions [5] we have that:

$$\frac{S_n}{n} - \log n \xrightarrow{d} Z \quad \text{as} \quad n \to \infty,$$

where Z is a stable random variable with index 1 and characteristic function $t \mapsto \exp\big(-\frac{\pi}{2}|t| - it\log|t|\big)$, $t \in \mathbb{R}$. Also, if F_n is the distribution function of $\frac{S_n}{n} - \log n$, and F is the distribution function of Z, then, since F is continuous, $F_n(x) \overset{n\to\infty}{\longrightarrow} F(x)$ uniformly on $x \in \mathbb{R}$. For this reason, if $\{x_n\}_{n\in\mathbb{N}}$ is a sequence of real numbers that converges to x, then $F_n(x_n)$ converges to $F(x)$ as $n \to \infty$. Now, note that for any pair of integers k and n we have:

$$\mathbb{P}(N_k \le n) = \mathbb{P}(S_n \ge k) = \mathbb{P}\left(\frac{S_n}{n} - \log n \ge \frac{k}{n} - \log n\right) = 1 - F_n\left(\frac{k}{n} - \log n\right),$$

so if k and n are functions of each other, chosen in a way such that

$$\frac{k}{n} - \log n \to x \quad \text{as} \quad n \to \infty \tag{3.6}$$

(or, alternatively, as $k \to \infty$), then, by the uniform convergence mentioned above:

$$\lim_{k\to\infty} \mathbb{P}(N_k \le n) = \lim_{n\to\infty} 1 - F_n\left(\frac{k}{n} - \log n\right) = 1 - F(x),$$

On the other hand we have:

$$\mathbb{P}(N_k \le n) = \mathbb{P}\left(\frac{(\log k)^2}{k} N_k - \log k - \log\log k \le \frac{(\log k)^2}{k} n - \log k - \log\log k\right),$$

so if we prove that $\frac{(\log k)^2}{k} n - \log k - \log\log k$ converges to $-x$ as $k \to \infty$, then, taking the limit in the above equation and equating it with the previous computation of the same limit, we see that:

$$\lim_{k\to\infty} \mathbb{P}\left(\frac{(\log k)^2}{k} N_k - \log k - \log\log k \le -x\right) = 1 - F(x);$$

that is,

$$\frac{(\log k)^2}{k} N_k - \log k - \log\log k \xrightarrow{d} -Z \quad \text{as} \quad k \to \infty.$$

Thus, it only remains to show that $\frac{(\log k)^2}{k} n - \log k - \log\log k \to -x$ as $k \to \infty$. Note that from (3.6), for sufficiently large n we have:

$$n(x-\epsilon) + n\log n \leq k \leq n(x+\epsilon) + n\log n,$$

hence:

$$\lim_{n\to\infty} \frac{k}{n\log n} = 1 \tag{3.7}$$

and, therefore:

$$\lim_{n\to\infty} \log(k) - \log(n) - \log\log n = \lim_{n\to\infty} \log \frac{k}{n\log n} = 0. \tag{3.8}$$

Since $\frac{\log\log n}{\log n} \to 0$, and using the above limit, we have:

$$\lim_{n\to\infty} \frac{\log k}{\log n} - 1 = \lim_{n\to\infty} \frac{\log(k) - \log(n) - \log\log n}{\log n} = 0.$$

Thus $\lim_{n\to\infty} \frac{\log k}{\log n} = 1$, and

$$\lim_{n\to\infty} \log\log k - \log\log n = 0. \tag{3.9}$$

Now, from (3.6) and (3.8) we have:

$$\frac{k}{n} - \log k + \log\log n \to x,$$

and by (3.9), we get:

$$\log k \frac{n\log k - k}{n\log k} - \log\log k = -\left(\frac{k}{n} - \log k + \log\log k\right) \to -x. \tag{3.10}$$

Now, since $\lim_{n\to\infty} \frac{\log k}{\log n} = 1$, and using (3.7), we also have:

$$\lim_{n\to\infty} \frac{k}{n\log k} = \lim_{n\to\infty} \frac{k}{n\log k}\frac{\log n}{\log n} = 1.$$

Thus, multiplying (3.10) by $\frac{n\log k}{k}$, we obtain:

$$\log k \frac{n\log k - k}{k} - \frac{n\log k}{k}\log\log k \to -x,$$

and

$$\log k \frac{n\log k - k}{k} - \left(\frac{n\log k}{k} - 1\right)\log\log k - \log\log k \to -x. \tag{3.11}$$

Multiplying the above expression by $\frac{\log\log k}{\log k}$ we see that

$$\log\log k \frac{n\log k - k}{k} - \left(\frac{n\log k}{k} - 1\right)\frac{(\log\log k)^2}{\log k} - \frac{(\log\log k)^2}{\log k} \to 0, \tag{3.12}$$

but since

$$\frac{(\log\log k)^2}{\log k} \to 0,$$

and

$$\frac{k}{n\log k} \to 1,$$

the two terms on the right converge to zero and, hence:

$$\log\log k \frac{n\log k - k}{k} = \log\log k\left(\frac{n\log k}{k} - 1\right) \to 0.$$

Finally, substituting this in (3.11), we obtain:

$$\log k \frac{n\log k - k}{k} - \log\log k = \frac{(\log k)^2}{k} n - \log k - \log\log k \to -x,$$

which finishes the proof of the lemma. □

Lemma 3.3.3. As $n \to \infty$, $\frac{(\log n)^2}{n}(n - 1 - S_{N_n - 1})$ converges in probability to 0.

Proof. To prove this lemma we use the Corollary to Theorem 6 (arithmetic version) in [5] and see that

$$\frac{\log(n - 1 - S_{N_n-1})}{\log(n-1)} \xrightarrow{d} U \quad \text{as} \quad n \to \infty, \tag{3.13}$$

where U is a standard uniform random variable. To connect notation in our case we define $Y_n \coloneqq n - S_{N_{n+1}-1}$ and use $m_1(t) \coloneqq \log t$ instead of the truncated mean $m(t) \coloneqq \int_0^t \mathbb{P}(\xi > x)\, dx$ as explained in Remark 1 of the same theorem. We now use this result in order to prove that

$$\frac{(\log n)^2}{n}(n - 1 - S_{N_n-1}) \xrightarrow{P} 0.$$

Note that for any $\epsilon > 0$:

$$\mathbb{P}\left(\frac{(\log n)^2}{n}(n-1-S_{N_n-1}) > \epsilon\right) = \mathbb{P}\left(\frac{\log(n-1-S_{N_n-1})}{\log(n-1)} > \frac{\log\left(\epsilon\frac{n}{(\log n)^2}\right)}{\log(n-1)}\right)$$
$$= \mathbb{P}\left(\frac{\log(n-1-S_{N_n-1})}{\log(n-1)} > \frac{\log\epsilon + \log n - 2\log\log n}{\log(n-1)}\right).$$

Since

$$\frac{\log\epsilon + \log n - 2\log\log n}{\log(n-1)} \to 1 \quad \text{as} \quad n \to \infty,$$

for any $0 < \delta < 1$ and sufficiently large n, we have

$$\delta < \frac{\log\epsilon + \log n - 2\log\log n}{\log(n-1)}.$$

Thus, using the above estimate and (3.13), we see that

$$\mathbb{P}\left(\frac{(\log n)^2}{n}(n-1-S_{N_n-1}) > \epsilon\right) \leq \mathbb{P}\left(\frac{\log(n-1-S_{N_n-1})}{\log(n-1)} > \delta\right) \to 1-\delta.$$

Finally, since δ was arbitrary we get:

$$\mathbb{P}\left(\frac{(\log n)^2}{n}(n-1-S_{N_n-1}) > \epsilon\right) \to 0.$$

□

Theorem 3.3.4. We have, as n tends to infinity:

$$\frac{(\log n)^2}{n}(M_n - N_n) \xrightarrow{P} 0,$$

and

$$\frac{(\log n)^2}{n}M_n - \log n - \log\log n \xrightarrow{d} Z.$$

where Z is a stable random variable as in Lemma 3.3.2

Proof. Note that $R^n_{N_n-1} = S_{N_n-1}$ and, therefore, R^n has at most $N_n - 1 + (n-1) - S_{N_n-1}$ increments, otherwise it would eventually be greater than

$n-1$ which is impossible by definition. Thus, we have $0 \leq M_n \leq N_n - 1 + (n-1) - S_{N_n-1}$ and, since $N_n - 1 \leq M_n$,

$$0 \leq \frac{(\log n)^2}{n}(M_n - N_n + 1) \leq \frac{(\log n)^2}{n}((n-1) - S_{N_n-1}).$$

Since $\frac{(\log n)^2}{n}((n-1) - S_{N_n-1}) \xrightarrow{P} 0$ as $n \to \infty$, we conclude that $\frac{(\log n)^2}{n}(M_n - N_n) \xrightarrow{P} 0$. Finally, since

$$\frac{(\log n)^2}{n}M_n - \log n - \log\log n = \frac{(\log n)^2}{n}N_n - \log n - \log\log n + \frac{(\log n)^2}{n}(M_n - N_n),$$

by Lemma 3.3.2 and Slutsky's theorem we get

$$\frac{(\log n)^2}{n}M_n - \log n - \log\log n \xrightarrow{d} Z.$$

□

Chapter 4

Site Frequency Spectrum of the BSC

4.1 Definition of the SFS

In this chapter we will first define the Site Frequency Spectrum (SFS) of a coalescent process, then we will describe previous studies related to the SFS of the BSC, and then we will present a new result: the derivation of the expected value of the SFS for the BSC. Finally we will give a brief application of the use of the SFS for model selection in the case of population evolution studies.

The SFS of a coalescent process with n initial blocks, say $\mathit{\Pi} \coloneqq (\Pi(t), t \geq 0)$, is a collection of $n-1$ random variables constructed as follows: at time 0 the blocks of $\Pi(t)$ have zero marks, and at any time t the blocks of $\Pi(t)$ acquire a new mark at a constant rate θ. Then, for each integer $1 \leq b \leq n-1$ we count the number of marks that fell on blocks of size b from time 0 until the absorption time; we call these counts $SFS_{n,b}$, and the total number of marks (i.e. the sum of the latter) SFS_n. On the other hand, $SFS_{n,b}$ can also be interpreted, and even alternatively defined, in the following way: given a path of $\mathit{\Pi}$ we construct a genealogical tree in the natural way according to the evolution of its blocks, and using the exponential jumping times in order to determine the lengths of the branches, and throw points to the branches of the resulting tree according to a Poisson process of rate θ; these points are interpreted as neutral mutations that occur in the associated lineages and thus are inherited to all the individuals in generation 0 that are below each

mark. Each mutation is assumed to occur at a different place in the genome thus creating a new segregating site. Finally, for each integer $1 \leq b \leq n-1$ we count how many mutations (segregating sites) are shared by a proportion b/n of individuals in generation 0 (see Figure 4.1).

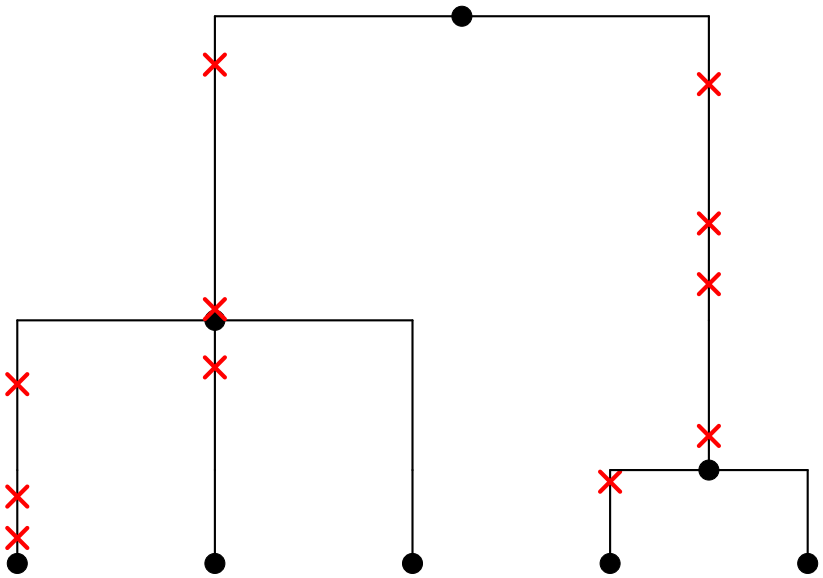

Figure 4.1: Schematic representation of a Poisson process on a genealogical tree. In this example we have: $SFS_{5,1} = 5$, $SFS_{5,2} = 4$, $SFS_{5,3} = 2$, and $SFS_{5,4} = 0$.

Let n be the number of initial blocks of a coalescent process Π. Consider the corresponding genealogical tree and for $1 \leq b \leq n-1$ define $\mathcal{L}_{n,b}$ as the total length of the branches associated to blocks of size b; note that, given $\mathcal{L}_{n,b}$, $SFS_{n,b}$ is Poisson distributed with parameter $\theta\mathcal{L}_{n,b}$. Similarly, if $\mathcal{L}_{n,T}$ is the total length of the tree then, given $\mathcal{L}_{n,T}$, the total number of mutations is also Poisson distributed with parameter $\theta\mathcal{L}_{n,T}$. In both cases we have:

$$\mathbb{E}[SFS_{n,b}] = \mathbb{E}[\mathbb{E}[SFS_{n,b}|\mathcal{L}_{n,b}]] = \theta\mathbb{E}[\mathcal{L}_{n,b}].$$

In the case of the BSC, an asymptotic limit for the number of mutations occurring in both internal and external branches has been derived in [9]. Note that the number of mutations occurring in external branches is the same as $SFS_{n,1}$ since, by definition, external branches are the branches of

the tree associated to blocks of size 1; similarly, the number of mutations occurring in internal branches is just $SFS_n - SFS_{n,1}$. In particular they show that as $n \to \infty$:

$$\frac{(\log n)^2}{n} SFS_{n,1} \to \theta$$

and

$$\frac{(\log n)^2}{n}\left(SFS_n - SFS_{n,1}\right) - \theta \log n - \theta \log\log n \xrightarrow{d} \theta(Z-1)$$

where Z is the stable random variable of index 1 that appears in Lemma 3.3.2. The method they use relies on representing the total internal branch length $\sum_{b=2}^{n-1} \mathcal{L}_{n,b}$ as:

$$\sum_{b=2}^{n-1} \mathcal{L}_{n,b} \overset{d}{=} \sum_{k=1}^{X_n - 1} C_{n,1}^{(k)} \frac{E_k}{C_n^{(k)} - 1}$$

where X_n is as in Section 3.3.1, $C_n^{(k)}$ and $C_{n,1}^{(k)}$ are the total number of blocks and the number of blocks of size 1, after k coalescence events; and $\{E_i\}_{i=1}^{X_n-1}$ is a collection of independent standard exponential random variables. Then they prove that $\frac{(\log n)^2}{n}\sum_{b=2}^{n-1} \mathcal{L}_{n,b} \xrightarrow{P} 1$ as $n \to \infty$ by using the coupling method introduced in [7] (and also described in Section 3.3.1) in order to study the evolution of the number of blocks of the BSC; and use this result in order to prove the limit for $SFS_{n,1}$. To prove the second limit they combine the limit for $SFS_{n,1}$ and the asymptotic result for SFS_n derived in [4], mainly that:

$$\frac{(\log n)^2}{n} SFS_n - \theta \log n - \theta \log\log n \xrightarrow{d} \theta Z$$

as $n \to \infty$.

On the other hand, in [1] the Allelic Frequency Spectrum (AFS) is studied for the BSC, the AFS is defined similarly to the SFS but with the difference that instead of considering all ancestral mutations for each individual in generation 0 only the most recent mutation is considered. Again the number of mutations shared by a proportion b/n ($1 \leq b \leq n-1$) of individuals in generation 0 is computed and denoted by $AFS_{n,b}$. For example in Figure 4.1 by ignoring ancestral mutations we see that $AFS_{5,1} = 3, AFS_{5,2} = 1, AFS_{5,3} = 1$ and $AFS_{5,4} = 0$. In [1] it is proved that, as $n \to \infty$:

$$\frac{\log n}{n} AFS_{n,1} \xrightarrow{P} \theta$$

and

$$\frac{(\log n)^2}{n} AFS_{n,b} \xrightarrow{P} \frac{\theta}{b(b-1)}.$$

Note that the SFS is deeply related to the branch lengths (jumping times) and the sizes and number of blocks of each coalescence event in a coalescent process; it can be regarded as a summary of these attributes that can be read directly from the population at time zero, which in applications is typically the only information available. Although the correspondence between the SFS and exchangeable coalescents may not be one to one, the SFS is often used in biology as a model selection tool for the study of the evolutionary history of a population when the only information available is a present day genotype sample of evolutionarily neutral positions of the genome. In particular, since the experimental tools used to measure the SFS of a given population produce reliable results for large frequencies of individuals but are strongly biased by noise for low frequencies, the SFS for large frequencies is of special importance in applications.

4.2 Derivation of $\mathbb{E}[SFS_{n,b}]$ for the BSC

In this section we will derive two new results: an explicit expression for $\mathbb{E}[SFS_{n,b}]$ when $\lfloor n/2 \rfloor < b \leq n-1$, and an upper bound for $\mathbb{E}[SFS_{n,b}]$ when $2 \leq b \leq \lfloor n/2 \rfloor$. To begin, define $C_{n,b}(t)$ as the number of blocks of $\Pi(t)$ with exactly b elements, and note that:

$$\begin{aligned}\mathcal{L}_{n,b} &= \sum_{k=1}^{\lfloor n/b \rfloor} k\lambda(\{t \in \mathbb{R}^+ : C_{n,b}(t) = k\}) \\ &= \int_0^\infty C_{n,b}(t)dt,\end{aligned}$$

where λ is the Lebesgue measure. Using Tonelli's theorem for interchanging the order of integration we have:

$$\begin{aligned}\mathbb{E}[SFS_{n,b}] &= \mathbb{E}[\theta \mathcal{L}_{n,b}] \\ &= \theta \mathbb{E}\left[\int_0^\infty C_{n,b}(t)dt\right] \\ &= \theta \int_0^\infty \mathbb{E}[C_{n,b}(t)]dt.\end{aligned}$$

So if we knew the distribution of the number blocks and block sizes at time t then we should be able to compute $\mathbb{E}[SFS_{n,b}]$. Also, from the expression above we may assume $\theta = 1$ since for other values one needs only multiply the derived expressions by θ. In the particular case of the BSC we will use the random tree construction described in Section 3.2.1 in order to compute $\mathbb{E}[C_{n,b}(t)]$. Although the same techniques can be used to compute $\mathbb{E}[C_{n,b}(t)]$ for any combination of n and b such that $2 \leq b \leq n-1$, in this work we will only derive the exact expression in the case where $\lfloor n/2 \rfloor < b \leq n-1$ (since this simplifies the computations considerably), and obtain an upper bound for $\mathbb{E}[C_{n,b}(t)]$ when $2 \leq b \leq \lfloor n/2 \rfloor$.

Theorem 4.2.1. For $\lfloor n/2 \rfloor < b \leq n-1$ we have:

$$\mathbb{E}[SFS_{n,b}] = \frac{n}{(n-b)b} \int_0^1 \left[\prod_{i=1}^{b-1} 1 - \frac{p}{i} \right] \left[\prod_{i=1}^{n-1-b} \frac{p}{i} + 1 \right] dp.$$

Proof. First, note that:

$$\mathbb{E}[C_{n,b}(t)] = \sum_{k=1}^{\lfloor n/b \rfloor} \mathbb{P}(C_{n,b}(t) \geq k),$$

and since $n/b < 2$ we have:

$$\mathbb{E}[C_{n,b}(t)] = \mathbb{P}(C_{n,b}(t) = 1).$$

Now, using the exchangeability of $\Pi(t)$, and the fact that $\Pi(t)$ cannot have more than one block of size b, we obtain:

$$\begin{aligned}
\mathbb{P}(C_{n,b}(t) = 1) &= \binom{n}{b} \mathbb{P}\left(\Pi\big|_b(t) = \{1, 2, \cdots, b\}, \bigcap_{i=b+1}^{n} \bigcap_{j=1}^{b} i \nsim j \right) \\
&= \binom{n}{b} \mathbb{P}(\Pi\big|_b(t) = \{1, 2, \cdots, b\}) \mathbb{P}\left(\bigcap_{i=b+1}^{n} \bigcap_{j=1}^{b} i \nsim j \,\middle|\, \Pi\big|_b(t) = \{1, 2, \cdots, b\} \right).
\end{aligned}$$

We now use the random tree construction for computing the latter probabilities. To that end we consider the arriving nodes with labels $\{2\}, \cdots, \{n\}$ and their associated exponential variables $E_2, \cdots, E_n$, and note that:

$$\begin{aligned}
\mathbb{P}(\Pi\big|_b(t) = \{1, 2, \cdots, b\}) =& \mathbb{P}(2 \sim 1) \mathbb{P}(3 \sim 1 \mid 2 \sim 1) \cdots \mathbb{P}(b \sim 1 \mid 1 \sim 2 \sim \cdots \sim b-1) \\
=& \mathbb{P}(E_2 \leq t) \prod_{i=3}^{b} (1 - \mathbb{P}(E_i > t, i \text{ attaches to } 1))
\end{aligned}$$

Also, using the same reasoning we find:

$$\begin{aligned}\mathbb{P}\Big(\bigcap_{i=b+1}^{n}\bigcap_{j=1}^{b} i \nsim j \Big| \Pi\big|_b(t)=\{1,2,\cdots,b\}\Big) =& \mathbb{P}\Big(\bigcap_{j=1}^{b} b+1 \nsim j \Big| \Pi\big|_b(t)=\{1,2,\cdots,b\}\Big)\times \\ & \prod_{i=b+2}^{n}\mathbb{P}\Big(\bigcap_{j=1}^{b} i\nsim j \Big| \Pi\big|_b(t)=\{1,2,\cdots,b\}, \bigcap_{k=b+1}^{i-1}\bigcap_{j=1}^{b} k\nsim j\Big) \\ =& \mathbb{P}(E_{b+1}>t, \text{b+1 attaches to 1})\times \\ & \prod_{i=b+2}^{n}\mathbb{P}\Big(\big\{E_i>t, \text{i attaches to 1}\big\}\bigcup \bigcup_{k=b+1}^{i-1} \text{i attaches to k}\Big) \\ =& \left(\frac{e^{-t}}{b}\right)\prod_{i=b+2}^{n}\left(\frac{e^{-t}}{i-1}+\frac{i-1-b}{i-1}\right) \\ =& \left(\frac{e^{-t}}{b}\right)\prod_{i=1}^{n-1-b}\left(\frac{e^{-t}}{b+i}+\frac{i}{b+i}\right) \\ =& \frac{e^{-t}}{b}\frac{(n-1-b)!}{\prod_{i=1}^{n-1-b} b+i}\prod_{i=1}^{n-1-b}\left(\frac{e^{-t}}{i}+1\right).\end{aligned}$$

where the term $\prod_{i=1}^{n-1-b}\left(\frac{e^{-t}}{i}+1\right)$ is set to 1 if $n-1-b\leq 0$. Thus, joining the last two expressions we get:

$$\begin{aligned}\mathbb{P}(C_{n,b}(t)=1) =& \binom{n}{b}\left[\prod_{i=1}^{b-1}\left(1-\frac{e^{-t}}{i}\right)\right]\left[\frac{(n-1-b)!}{\prod_{i=1}^{n-1-b} b+i}\frac{e^{-t}}{b}\prod_{i=1}^{n-1-b}\left(\frac{e^{-t}}{i}+1\right)\right] \\ =& \frac{n}{(n-b)b}e^{-t}\left[\prod_{i=1}^{b-1}1-\frac{e^{-t}}{i}\right]\left[\prod_{i=1}^{n-1-b}\frac{e^{-t}}{i}+1\right].\end{aligned}$$

Therefore, making the change of variable $dp=e^{-t}dt$, we obtain:

$$\begin{aligned}\mathbb{E}[SFS_{n,b}] =& \int_0^{\infty}\frac{n}{(n-b)b}e^{-t}\left[\prod_{i=1}^{b-1}1-\frac{e^{-t}}{i}\right]\left[\prod_{i=1}^{n-1-b}\frac{e^{-t}}{i}+1\right]dt \\ =& \frac{n}{(n-b)b}\int_0^1\left[\prod_{i=1}^{b-1}1-\frac{p}{i}\right]\left[\prod_{i=1}^{n-1-b}\frac{p}{i}+1\right]dp.\end{aligned}$$

□

Using the same rationale we can obtain an upper bound for $\mathbb{P}(C_{n,b}(t) \geq k)$ and any n, b, k with $kb \leq n$, which can be used to get an upper bound for $\mathbb{E}[SFS_{n,b}]$.

Theorem 4.2.2. For any n, b with $2 \leq b \leq n-1$, and k with $kb \leq n$, we have:

$$\int_0^\infty \mathbb{P}(C_{n,b}(t) \geq k)\, dt \leq \begin{cases} \frac{1}{(n-kb\vee 1)b^{k-1}} \int_0^1 p^{k-2} \left[\prod_{i=1}^{b-1} 1 - \frac{p}{i}\right]^k dp & \text{if } n - kb = 0. \\ \frac{n}{(n-kb\vee 1)b^k} \int_0^1 p^{k-1} \left[\prod_{i=1}^{b-1} 1 - \frac{p}{i}\right]^k dp & \text{if } n - kb = 1. \\ \frac{n}{(n-kb\vee 1)b^k} \int_0^1 p^{k-1} \left[\prod_{i=1}^{b-1} 1 - \frac{p}{i}\right]^k \left[\prod_{i=1}^{n-1-kb} 1 + \frac{pk}{i}\right] dp & \text{if } n - kb > 1. \end{cases}$$

Proof. Note that for any n, b, k with $kb \leq n$ we have:

$$\begin{aligned} \mathbb{P}(C_{n,b}(t) \geq k) \leq & \frac{n!}{k!(b!)^k(n-kb)!} \mathbb{P}\Bigg(\Pi\big|_{kb}(t) = \{1, 2, \cdots, b\}, \cdots, \{(k-1)b, \cdots, kb\}, \\ & \qquad \bigcap_{i=kb+1}^{n} \bigcap_{j=1}^{kb} i \nsim j\Bigg) \\ = & \frac{n!}{k!(b!)^k(n-kb)!} \mathbb{P}(\Pi\big|_{kb}(t) = \{1, 2, \cdots, b\}, \cdots, \{(k-1)b+1, \cdots, kb\}) \times \\ & \mathbb{P}\Bigg(\bigcap_{i=b+1}^{n} \bigcap_{j=1}^{b} i \nsim j \Bigg| \Pi\big|_{kb}(t) = \{1, 2, \cdots, b\}, \cdots, \{(k-1)b+1, \cdots, kb\}\Bigg). \end{aligned}$$

where $\frac{n!}{k!(b!)^k(n-kb)!}$ is just the number of ways in which we can construct k blocks of size b when there are n elements. Observe that the latter is just an overestimate since multiplying the probability

$$\mathbb{P}\Bigg(\Pi\big|_{kb}(t) = \{1, 2, \cdots, b\}, \cdots, \{(k-1)b, \cdots, kb\}, \bigcap_{i=kb+1}^{n} \bigcap_{j=1}^{kb} i \nsim j\Bigg)$$

by the number of combinations considered in the factor $\frac{n!}{k!(b!)^k(n-kb)!}$ does not account for the events where, apart from the k initial blocks of size b constructed with the kb initial nodes, the remaining nodes $\{kb+1, \cdots, n\}$ form

one or more additional blocks of size b; in these cases, some of the permutations considered in $\frac{n!}{k!(b!)^k(n-kb)!}$ do not actually produce a different partition, thus causing these events to be accounted for multiple times. Continuing with the computation and using the random recursive tree construction as before, we get:

$$\mathbb{P}(C_{n,b}(t) \geq k) \leq \begin{cases} \frac{1}{(n-kb\vee 1)b^{k-1}} e^{-t(k-1)} \left[\prod_{i=1}^{b-1} 1 - \frac{e^{-t}}{i} \right]^k & \text{if } n-kb=0. \\ \frac{n}{(n-kb\vee 1)b^{k}} e^{-tk} \left[\prod_{i=1}^{b-1} 1 - \frac{e^{-t}}{i} \right]^k & \text{if } n-kb=1. \\ \frac{n}{(n-kb\vee 1)b^{k}} e^{-tk} \left[\prod_{i=1}^{b-1} 1 - \frac{e^{-t}}{i} \right]^k \left[\prod_{i=1}^{n-1-kb} 1 + \frac{e^{-t}k}{i} \right] & \text{if } n-kb>1. \end{cases}$$

Thus, using the change of variable $dp = e^{-t}dt$ to integrate the latter expressions with respect to t from 0 to ∞ we get the desired result. □

Finally, since

$$\begin{aligned} \mathbb{E}[SFS_{n,b}] &= \int_0^\infty \mathbb{E}[C_{n,b}(t)]\ dt \\ &= \int_0^\infty \sum_{k=1}^{\lfloor n/b \rfloor} \mathbb{P}(C_{n,b}(t) \geq k)\ dt, \end{aligned}$$

we can use the upper bounds for $\int_0^\infty \mathbb{P}(C_{n,b}(t) \geq k)\, dt$ derived above to obtain an upper bound for $\mathbb{E}[SFS_{n,b}]$ by summing them over all the values of k such that $1 \leq k \leq \lfloor n/b \rfloor$. The expressions thus obtained are very easily computed using a computer program that integrates polynomials. In Figure 4.2 we show how these theoretic expressions compare to actual simulations of the SFS for the BSC. Note that for frequencies greater than 1/2 the theoretic expression for $SFS_{n,b}$ does match the simulations; on the other hand, for frequencies equal or below 1/2 the theoretic upper bound does lie above the simulated values; moreover, for frequencies approaching 1/2 from the left we see that the bound becomes tighter.

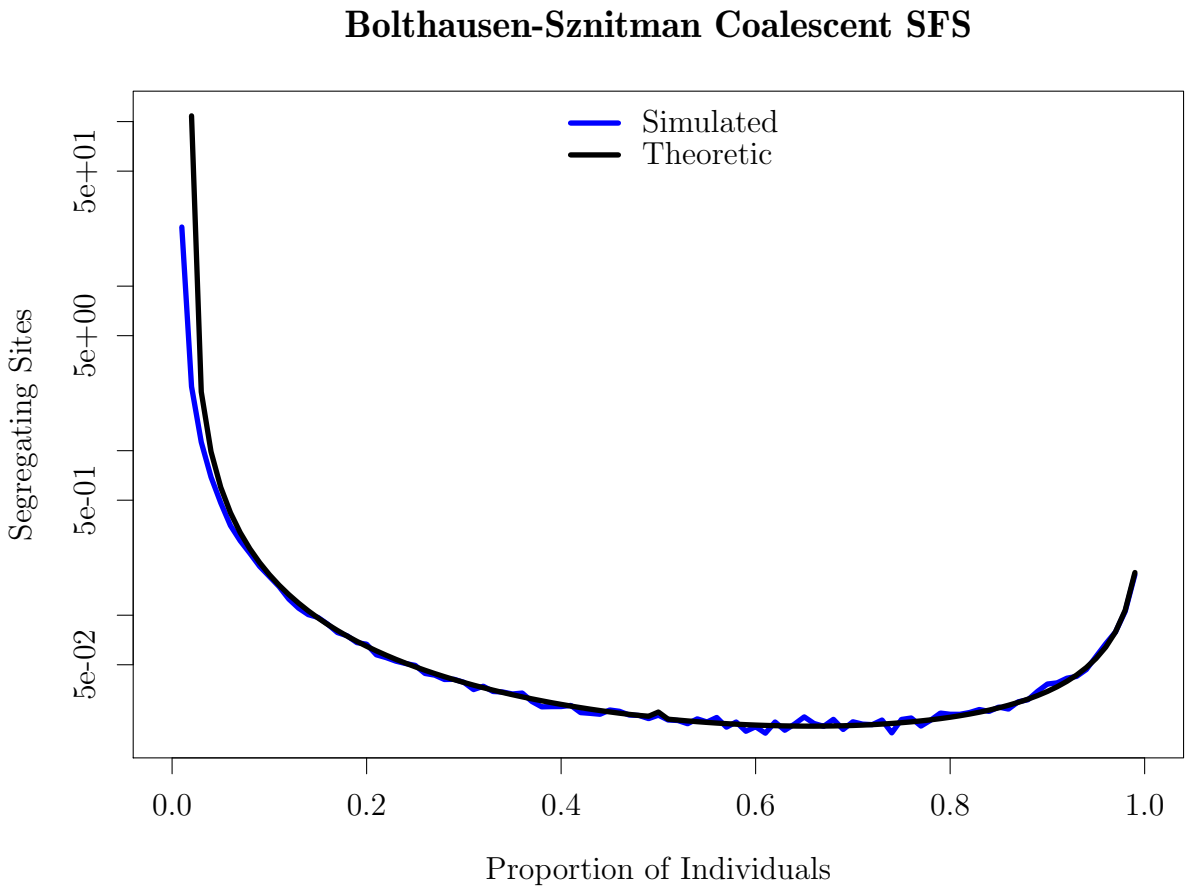

Figure 4.2: The expressions for the SFS with parameter $\theta = 10$ of the BSC derived above are compared against simulated data. For proportions less than or equal to $\lfloor n/2 \rfloor$ the obtained upper bound is plotted, for proportions greater than $\lfloor n/2 \rfloor$ the exact theoretic SFS is shown. The simulations are on initial populations with 100 individuals.

4.3 Population Evolution Models and the BSC

In this section we will describe the population evolution model introduced by Neher and Hallatschek [11]. This model has the property that the reproductive success of an individual is determined by a fitness function; that

is, it considers the effect of natural selection in the evolution of the population. The simplest models of population evolution consider a population of individuals evolving in discrete time, where each time point corresponds to a non-overlapping generation. The model then aims to describe the way in which generation $g+1$ is related to generation g by determining which individuals of generation $g+1$ are descendants of each individual in generation g (including the possibility that an individual in generation g may not have any descendancy) through a probabilistic process. The way in which this is typically done is by giving a distribution on the number of descendants for each individual in generation g, and choosing them uniformly at random from generation $g+1$. For example, one of the better studied models of population evolution is Cannings model; in this model the population size N remains fixed and the offspring sizes of every individual in any particular generation is given by an indepent copy of an exchangeable random vector $(\xi_1, \cdots, \xi_N)$ with the property that $\sum_{i=1}^{N} \xi_i = N$. Given a population evolution model we may construct a coalescent process with values in $\mathscr{P}_n$ by choosing n individuals from the last generation and tracking the associated genealogy backwards in time. As an example, in [13] and [14] respectively, Sagitov and Schweinsberg study two different Cannings models of population evolution where all the individuals have the same fitness and reproductive success (i.e. there is no natural selection); they give conditions under which the associated coalescent processes, after an appropriate time normalization, converge to the exchangeable coalescents described so far, in particular to the BSC.

We now describe Neher-Hallatschek's model and some of its properties. In this model we consider a population of N_g individuals at generation g, each with an associated fitness that determines its reproductive success. The fitness of individual j in generation g, $j \in [N_g]$, is represented by a real number s_j that determines the distribution of its offspring size, which is set to be Poisson distributed with parameter:

$$\lambda \coloneqq \exp\{s_j - \overline{S} + 1 - N_g/N\},$$

where $\overline{S} \coloneqq \frac{1}{N_g} \sum_{i=1}^{N_g} s_i$ is the mean population fitness, and N is the objective population size (typically $N = N_0$). The term $1 - N_g/N$ in the expresion above indeed ensures that N_g stays roughly at N as the population evolves over time. Once generation g reproduces, the fitness of each individual in generation $g+1$ is inherited from that of its parent and updated in a two step

process: first the individual is mutated according to a fixed probability μ, and then the difference in fitness (δ) caused by the mutation is sampled from a fixed distribution D with values in $\mathbb{R}$. That is, the fitness of individual j in generation $g+1$ is set to:

$$s_j \coloneqq s_p + m\delta,$$

where $m \sim \text{Bernoulli}(\mu)$, $\delta \sim D$, and s_p is the fitness of its parent. Finally we consider the associated coalescent process $\Pi^g_{n,N}$ given by selecting n individuals from generation g and tracking their genealogy backwards in time. Here we make emphasis on the starting generation g used for the construction of the coalescent process since, a priori, the distribution of family sizes and of parental relations is not constant across generations, i.e. the evolution of the population is not exchangeable over time.

This model is much more complicated than the well studied Cannings models of Sagitov and Schweinsberg, firstly because in contrast to the latter the population size at generation g is a random variable, and secondly, and most importantly, because the offspring distribution of the individuals in generation $g+1$ is not independent from that of their parents, since it depends on the inherited fitness; that is, there is a strong dependency of the reproductive success of individuals accross generations.

Here we will use simulations to observe some properties of this model, for example that the fitness distribution of the poupulation evolves as a traveling wave of Gaussian shape, and that the SFS associated to this model can be roughly approximated by that of the BSC.

As a case study we simulated populations with mutation probability $\mu = 1$ and Normally distributed mutational effects; in particular we simulated two populations with $\delta \sim Normal(\text{mean} = 0, \text{sd} = 0.1)$ and $\delta \sim Normal(\text{mean} = 0, \text{sd} = 0.1)$ respectively, and an objective/intial population size $N = N_0 = 10^5$. In Figure 4.3 we show that the centered fitness distribution converges to a Gaussian distribution as the population evolves over time, and that the fitness distribution evolves as a traveling wave of Gaussian shape. Note that even though in both cases the centered fitness distribution converges to a Normal distribution with zero mean and standard deviation $s = 0.39$ and $s = 0.08$ respectively, and, moreover, the offspring size distribution also converges to the distribution given by $X \mid f \sim Poisson(f), f \sim Normal(0, s)$, the evolution of the population cannot be approximated by that of a Schweinsberg population with these

offspring size distributions, the reason is that in the latter case the reproduction of any generation is independent from the rest, while in the Neher-Hallatschek model this requirement is not met. Indeed, by Theorem 4 in [14] the coalescent process associated to the Schweinsberg population with offspring distribution $X \mid f \sim Poi(f), f \sim N(0, s)$ is Kingman's coalescent, since, as shown below, the distribution of offspring size X has a finite second moment:

$$\begin{aligned}\mathbb{E}[X^2] &= \mathbb{E}[\mathbb{E}[X^2 \mid f]] \\ &= \mathbb{E}[e^{2f} + e^f] \\ &= e^{s^2} + e^{s^2/4} \\ &< \infty;\end{aligned}$$

whereas, as shown in Figure 4.4, the SFS of the Neher-Hallatschek population is much more similar, and can indeed be very well approximated by that of the BSC and not by that of Kingman's coalescent. The fact that the associated coalescent has jumps where multiple blocks coalesce can be intuitively explained by the following rationale: since the distribution of offspring sizes depends heavily on the fitness of each individual, and since the fitness is inherited, the population evolves through a series of selective sweeps where only the descendants of individuals with high fitness in early generations survive, thus after some generations have passed all the individuals in the population, or at least the grand majority, will be descendants of the few high fitness individuals in early generations. This has the effect that family sizes tend to be much larger than those of the populations associated to Kingman's coalescent and gives rise to events where multiple blocks coalesce. In Figure 4.4 we show the simulated SFS with rate $\theta = 1$ of the Neher-Hallatschek populations, where we have scaled the time of the associated coalescents according to:

1. Making the expected absorption time equal to that of BSC.
2. Making the expected tree length equal to that of the BSC.
3. Finding the factor that when multiplied to the SFS of the Neher-Hallatschek model, the sum of squared distances of the function $\log(SFS)$ between the BSC and the Neher-Hallatschek populations is minimized.
4. Considering each generation as a unit of time (i.e. no time normalization).

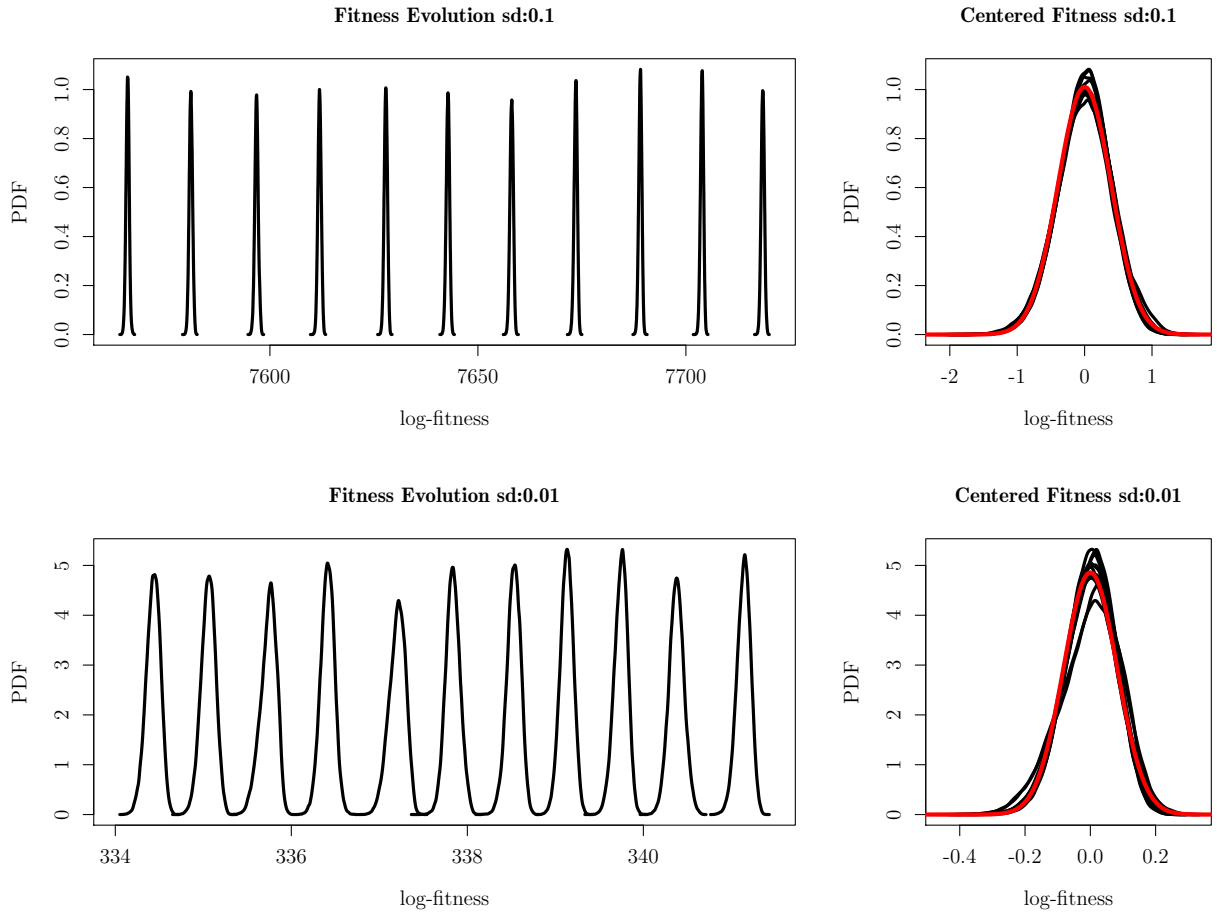

Figure 4.3: On the left we show the evolution of the fitness distribution for a single population with parameters $\mu = 1$, $\delta \sim Normal(\text{mean} = 0, \text{sd} = 0.1)$ and $\delta \sim Normal(\text{mean} = 0, \text{sd} = 0.01)$ respectively, and $N = N_0 = 10^5$. Only ten generations separated by 100 generations and starting from generation 5×10^5 are shown. On the right, the centered fitness distribution is shown for the same set of generations, algonside a Normal distribution, shown in red, with zero mean and standard deviation set to the mean standard deviation of the fitness distributions from generation 5×10^5 to generation 10^6, that is, set to 0.39 and 0.08 respectively.

Exploring these four normalizations is justified since the Neher-Hallatschek model has not been studied thoroughly and the appropriate time normalization has not been described yet. The resulting SFS are compared against the SFS of the Bolthausen-Sznitman and Kingman's coalescents. According to these graphs the SFS of the Neher-Hallatschek model can be roughly approximated by that of the BSC under a suitable time normalization.

However, the associated coalescents may not converge as processes as suggested by the difference in the distributions of the total number of jumps and T_2, the number of jumps until the coalescence of two randomly chosen individuals (i.e. the number of jumps until the first common ancestor), as shown in Figure 4.5. Finally, it is worth mentioning that apart from Cannings model of population evolution [13], which is a generalization of the Wright-Fisher model, and from the models based on supercritical Galton-Watson processes [14], the BSC has also been recently shown, in a rigorous way, to arise in other population models that contemplate the effect of natural selection which are similar in spirit to the Neher-Hallatschek model described in the present text [15].

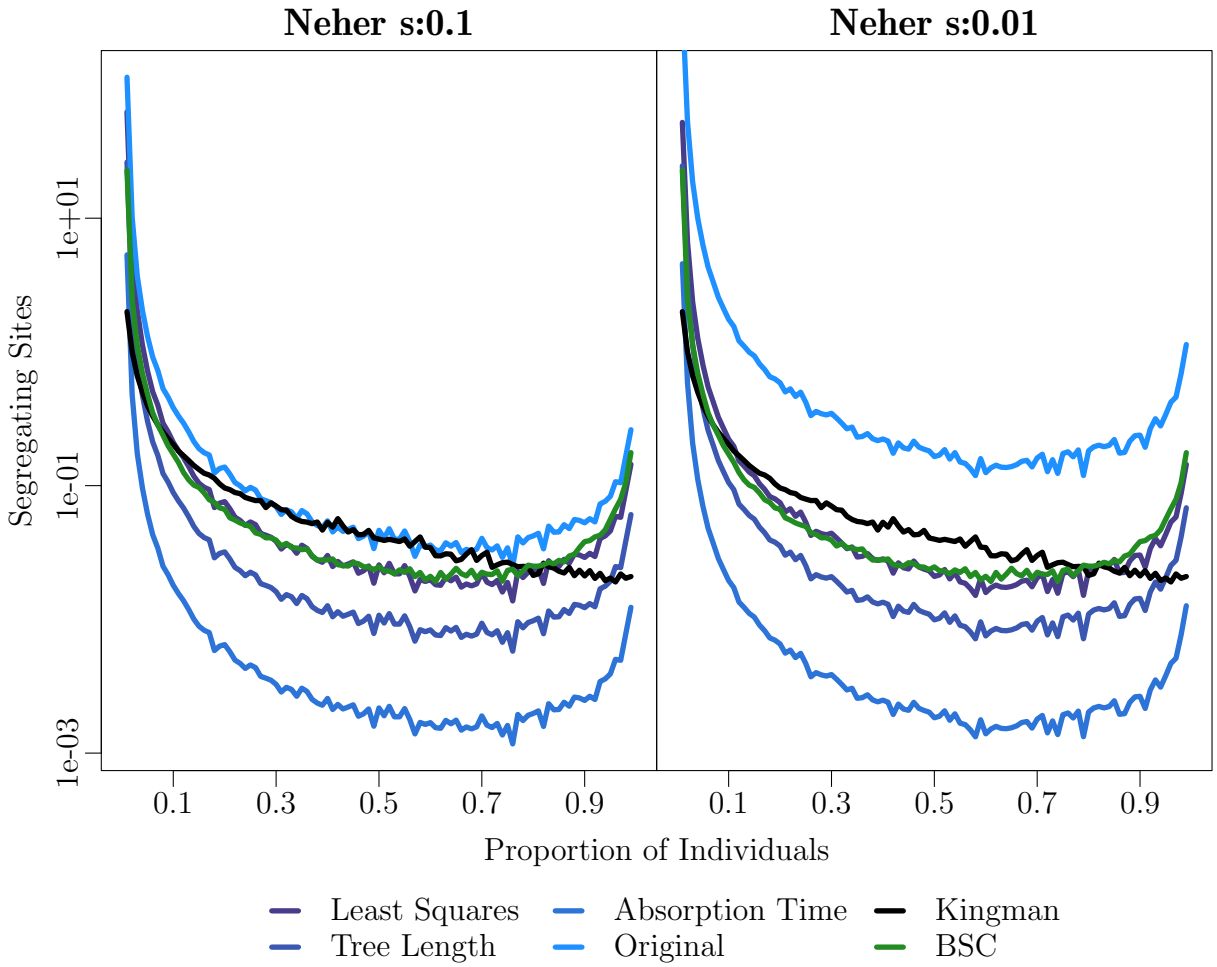

Figure 4.4: For both set of parameters ($s = 0.1$ and $s = 0.01$) 4000 populations where simulated and the mean SFS was obtained for each proportion of individuals. Then the SFS of the Neher populations were modified according to each of the time normalizations described in the main text and compared against that of the Bolthausen-Sznitman and Kingman's coalescents. The time normalization computed using the least squares method gives a very good approximation of the SFS for the BSC.

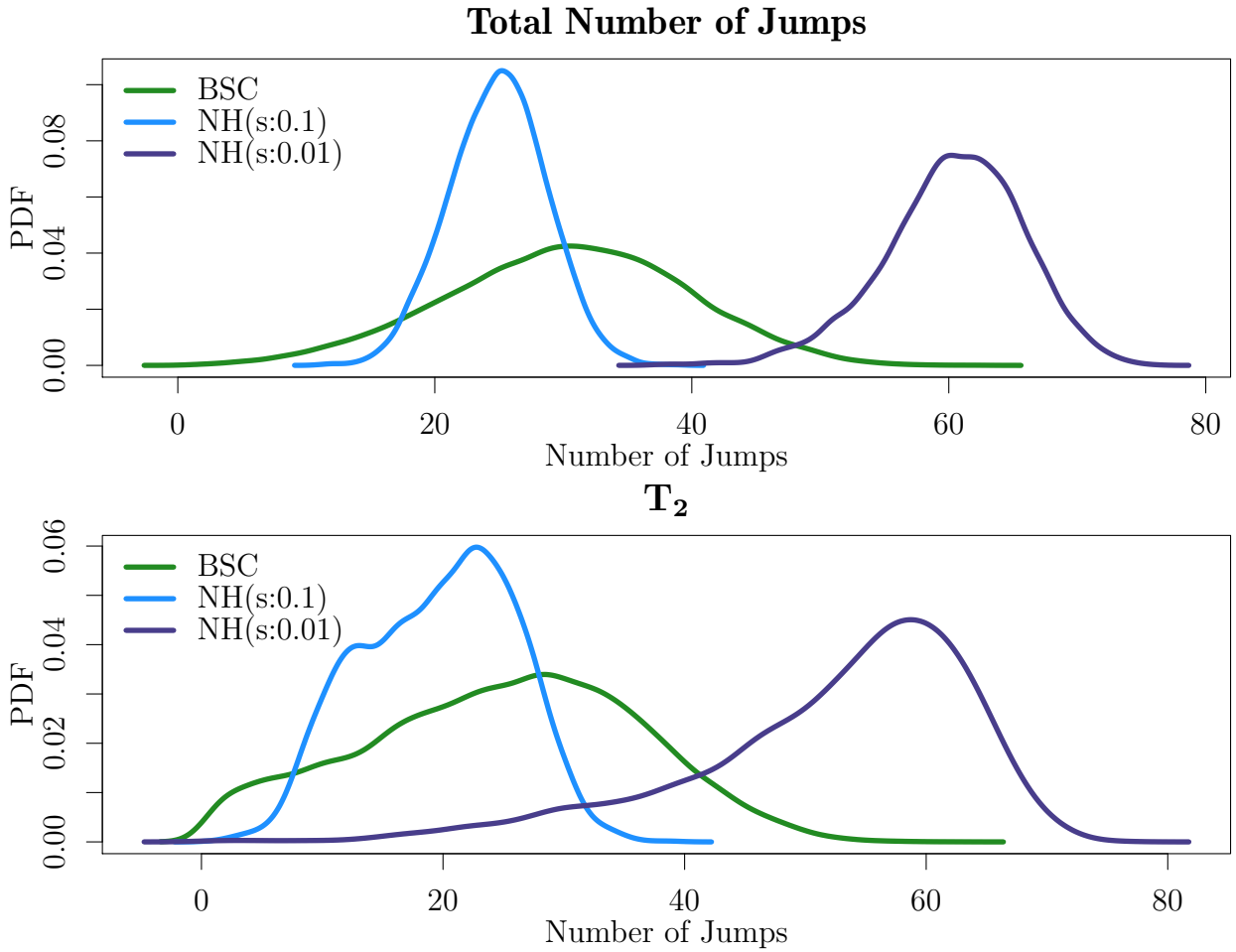

Figure 4.5: On the top, the empirical distribution of the total number of jumps is shown for the BSC and the two Neher-Hallatschek populations ($s = 0.1$ and $s = 0.01$). Below, the empirical distribution of the number of jumps needed to reach the first common ancestor of two randomly chosen individuals. The distributions for the BSC coalescent were estimated from 16000 simulations, while 4000 simulations where used to estimate those of the Neher-Hallatschek populations.

Bibliography

[1] A.-L. Basdevant and C. Goldschmidt Asymptotics of the allele frequency spectrum associated with the Bolthausen-Sznitman coalescent, *Electron. J. Probab.* **13** (2008), 486–512.

[2] J. Bertoin, *Random fragmentation and coagulation processes*, Cambridge University Press, (2006).

[3] P. Billingsley,*Probability and Measure*, 3rd edn, John Wiley & Sons, (1995)

[4] M. Drmota, A. Iksanov, M. Möhle and U Röler, Asymptotic results concerning the total branch length of the Bolthausen-Sznitman coalescent, *Stochastic Process. Appl.* **117** (2007), no. 10, 1404–1421.

[5] J. L. Geluk and L. de Haan, Stable probability distributions and their domains of attraction: a direct approach, *Probab. Math. Statist.* **20** (2000), no. 1, 169–188.

[6] C. Goldschmidt and J.B. Martin Random recursive trees and the Bolthausen-Sznitman coalescent, *Electron. J. Probab.* **10** (2005), 718–745.

[7] A. Iksanov and M. Möhle, A probabilistic proof of a weak limit law for the number of cuts needed to isolate the root of a random recursive tree, *Electron. Comm. Probab.* **12** (2007), 28–35.

[8] O. Kallenberg, *Foundations of modern probability*, 2nd edn, Springer, (2002).

[9] G. Kersting, J.C. Pardo and A. Siri-Jégousse, Total internal and external lengths of the Bolthausen-Sznitman coalescent, *J. Appl. Probab.* **51A** (2014), 73–86.

[10] Kingman, J. F. C., The representation of partition structures, *Math. Soc.* **18** (1978), 374–380.

[11] R.H. Neher and O. Hallatschek, Genealogies of rapidly adapting populations, *Proc. Nat. Acad. Sci. USA* **110** (2013), 437–442.

[12] J. Pitman, Coalescents with multiple collisions, *Ann. Probab.* **27** (1999), no. 4, 1870–1902.

[13] S. Sagitov, The general coalescent with asynchronous mergers of ancestral lines, *J. Appl. Probab.* **36** (1999), no. 4, 1116–1125.

[14] J. Schweinsberg, Coalescent processes obtained from supercritical Galton-Watson processes, *Stoch. Proce. Appl.* **106** (2003), no. 1, 107–139.

[15] J. Schweinsberg, Rigorous results for a population model with selection II: genealogy of the population, Preprint on *Arxiv*.