UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

GENEALOGIES IN POPULATION MODELS WITH SEED BANK EFFECTS

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTORA EN CIENCIAS

PRESENTA:

LIZBETH PEÑALOZA VELASCO

DIRECTOR DE LA TESIS:

ARNO SIRI-JÉGOUSSE
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS

MIEMBROS DEL COMITÉ TUTOR:

GERÓNIMO URIBE BRAVO
INSTITUTO DE MATEMÁTICAS

ANDREAS E. KYPRIANOU
UNIVERSITY OF BATH

CIUDAD DE MÉXICO, 12 DE NOVIEMBRE DE 2021

Contents

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología (CONACyT-México) le agradezco la beca recibida durante mi doctorado, gracias a la cual este proyecto fue posible. Así como el apoyo brindado con número FC-2016-1946 para asistir a escuelas de verano en el extranjero.

Al Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), UNAM, con número IA103820, le doy gracias por la beca suplementaria que me brindó.

Al Posgrado en Ciencias Matemáticas y Especialización en Estadística Aplicada, le agradezco por los múltiples apoyos económicos que me dieron para asistir a congresos y escuelas, tanto nacionales como internacionales.

A la Universidad de Bath, U.K., por haberme recibido por 6 meses y brindarme apoyo económico durante esta estancia.

Ahora quiero Expresar mi profundo agradecimiento a mi tutor de tesis Dr. Arno Siri-Jégousse por sus enseñanzas, correcciones, consejos, paciencia, por siempre estarme motivando y sobre todo por saber guiarme en este camino. Al Dr. Adrián González Casanova Soberón le doy gracias por su colaboración, consejos, apoyo, enseñanzas y por compartirme sus conocimientos sobre el modelo seed bank.

Le quiero agradecer a mi comité tutorial, Dr. Gerónimo Uribe Bravo por sus consejos y apoyo, y al Dr. Andreas E. Kyprianou por darme la oportunidad de tener una estancia en la Universidad de Bath y realizar una colaboración en conjunto con el Dr. Tim Rogers.

Al resto de los miembros del jurado de correcciones de Tesis y examen de grado, Dra. Maite Isabel Wilke Berenguer, Dra. Julia Adela Palacios y a la Dra. María Clara Fittipaldi les doy gracias por su tiempo y dedicación para las correcciones y comentarios de este manuscrito, así como su participación en el jurado evaluador.

A continuación, le doy gracias a mis padres, Rogelio Peñaloza Bermúdez y Columba E. Velasco Colores, por darme la vida y las bases para llegar a cumplir este objetivo, por todos los valores que me inculcaron y algunos de ellos fueron el compromiso y la responsabilidad. Por enseñarme a tener la fuerza de levantarme cuando me he caído y seguir adelante. A mis hermanos, Sugey, Columba, Evelyn y Gustavo por

siempre estar conmigo, apoyarme y alentarme a perseguir mis sueños. A mis sobrinos y cuñados por formar parte de esta hermosa familia que siempre ha estado a mi lado.

A mis amigas, Yessica y Veroska por siempre estar a mi lado, escucharme y ser mis compañeras para relajarme y reír. A Ricardo por haber estado a mi lado al inicio de este proyecto, por haberme apoyado en muchos momentos y de diferentes formas, gracias por haber apoyado mis sueños. A mi amigo Alejandro por hacer más amenos los congresos con su compañía, por ser mi compañero de cubículo y por ayudarme a discutir mis demostraciones cuando yo ya no sabía por dónde ir.

A muchas más personas que conocí durante estos años del docotorado, que aunque no las mencione, fueron parte de este proyecto.

Resumen

En este trabajo hacemos un estudio de los modelos de genealogía de poblaciones, en particular hacemos un análisis a los modelos que tienen banco de semillas, es decir, son modelos que tienen individuos que se encuentran inactivos por algunas generaciones.

En el capítulo 1 se presentan los principales conceptos, modelos y técnicas que ayudarán a dar un mejor entendimiento de los resultados que se obtuvieron en esta tesis.

En el capítulo 2 se presenta un marco para la construcción simultanea del modelo seed-bank con distribución de saltos multigeneracionales y una ley de reproducción tipo Cannings que satisfaga una construcción paintbox, conjuntamente se dan algunos resultados límite hacia adelante y hacia atrás en el tiempo, es decir, damos algunas condiciones para la convergencia al coalescente de Kingman y estudiamos escenarios más allá de esta clase de coalescentes, en estos somos capaces de describir como el fenómeno de seed-bank débil reduce el tamaño típico de los eventos de coalescencia. También se presenta un resultado de dualidad. La principal técnica que se usa es construir una gráfica aleatoria que nos permita encajar el proceso ancestral y el proceso de frecuencia de ambos modelos, Cannings y seed-bank, simultáneamente y así estudiar la relación de dualidad. Con este resultado se cubre un hueco en el estudio de los modelos con seed bank, debido a que hasta el día de hoy no se había hecho un análisis, con mecanismos de reproducción más generales, tales como los basado en el modelo de Cannings.

En el capítulo 3 se hace un estudio del comportamiento asintótico de algunas funcionales del coalescente seed-bank, correspondiente a un modelo con seed-bank fuerte. Esto podría ser de utilidad para aplicaciones en genética. El principal resultado es la obtención de la longitud total del coalescente, donde obtuvimos que la longitud activa se comporta similar a la longitud del coalescente Kingman, lo que significa que no es posible distinguir entre el coalescente Kingman y el colescente seed-bank usando sólo la longitud. Lo que nuestros resultados muestran es que la mayoría de las mutaciones ocurren en la fase Kingman, es decir entre el tiempo cero y el tiempo de la primera reactivación, todo esto empezando con n plantan y cero semillas, y en esta parte del árbol, la parte dormida o inactiva es irrelevante. Para hacer una discriminación entre el coalescente Kingman y el seed-bank se necesitan resultados más finos como por ejemplo fórmulas de muestreo (Sampling formula). Al final del capítulo se presenta

una fórmula de muestreo usando la idea del proceso del restaurante chino. Como resultado obtuvimos una aproximación a la probabilidad de la frecuencia de bloques activos e inactivos al tiempo de la primera reactivación.

Un futuro trabajo es calcular la longitud del árbol del coalescente seed-bank con simultáneas activaciones y desactivaciones [10].

Summary

In this work we do a study of population genealogy models, in particular we analyze the models that have a seed bank, that is, they are models that have individuals that are inactive for some generations.

Chapter 1 presents the main concepts, models and techniques that will help to give a better understanding of the results obtained in this thesis.

Chapter 2 presents a framework for the simultaneous construction of the seed-bank model with distribution of multigenerational jumps and a Cannings-type reproduction law that satisfies a paintbox construction, together some limit results are given forwards and backwards in time, that is, we give some conditions for convergence to the Kingman coalescent and we study scenarios beyond this class of coalescers, in these we are able to describe how the weak seed-bank phenomenon reduces the typical size of the coalescence events. A duality result is also presented. The main technique used is to construct a random graph that allows us to fit the ancestral process and the frequency process of both models, Cannings and seed-bank, simultaneously and thus study the duality relationship. With this result, a gap is covered in the study of seed bank models, because until today an analysis had not been done, with more general reproduction mechanisms, such as those based on the Cannings model.

In chapter 3 a study of the asymptotic behavior of some functionalities of the seed-bank coalescent is made, corresponding to a model with a strong seed-bank. This could be useful for applications in genetics. The main result is obtaining the total length of the coalescent, where we obtained that the active length behaves similar to the length of the Kingman coalescent, which means that it is not possible to distinguish between the Kingman coalescent and the seed-bank coalescent using only the length. What our results show is that most mutations occur in the Kingman phase, that is, between time zero and the time of the first reactivation, all this starting with n plants and zero seeds, and in this part of the tree, the dormant or inactive part is irrelevant. To discriminate between the Kingman coalescent and the seed-bank coalescent, finer results are needed, such as Sampling formula. A sampling formula is presented at the end of the chapter using the idea of the Chinese restaurant process. As a result, we obtained an approximation to the probability of the frequency of active and inactive blocks at the time of the first reactivation.

A future possible work is to calculate the length of the seed-bank coalescing tree with simultaneous activations and deactivations [10].

Chapter 1

Introduction

1.1 Neutral models

1.1.1 Wright-Fisher Model and Wright-Fisher diffusion

Definition 1.1.1. Wright-Fisher Model. Consider a haploid (one parent) population and assume discrete and non-overlapping generations, in which each generation has a fixed number $N \in \mathbb{N}$ of individuals, and everybody has the same chance to reproduce, i.e., the model is neutral. The dynamics are as follows: each individual from generation $k+1$ chooses, independently and uniformly at random, its parent from generation k, for $k \in \mathbb{Z}$.

Now, suppose there are two different types of *alleles* in the population, A and a. Also, suppose that these alleles are neutral, i.e., the reproduction of one individual does not depend on its type. Furthermore, each individual copies the type of its parent. In the area of genetics, the samples are not the individuals themselves but their genetic material. A gene can have different forms, and each form is called allele [47]. For example, in Mendel's experiment, the gene that gives the color of peas has two alleles: one determines the color green, and the other determines the color yellow.

Let X_k^N denotes the number of individuals with allele A in generation k. The process $\{X_k^N\}_{k\geq 0}$ is a discrete time Markov chain, i.e., given the present state, the past does not matter to predict the future ([25], section 1.2). Moreover, given $X_k^N = i$, with $i \in \{0, 1, ..., N\}$, X_{k+1}^N has a binomial distribution,

$$\mathbb{P}(X_{k+1}^N = j | X_k^N = i) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}, \; j \in \{0, 1, ..., N\}. \tag{1.1}$$

We do not consider any mutation mechanism so, eventually, all individuals in the population will have the same type. Once one type is reached by the entire population

the proportion cannot change anymore. The process $\{X_k^N\}_{k\geq 0}$ has two absorbing states, 0 and N. Let $\tau = \inf\{k \geq 0 : X_k^N = 0 \text{ or } X_k^N = N\}$ be the first time that a whole generation has type A or a. For the allele A, the event $\{X_\tau^N = 0\}$ is called extinction, while the event $\{X_\tau^N = N\}$ is called fixation.

Proposition 1.1.2. *The Markov chain $\{X_k^N\}_{k\geq 0}$ is a (bounded) martingale and starting from i, the probability of fixation is i/N.*

The Markov chain $\{X_k^N\}_{k\geq 0}$ is a martingale due to its Binomial kernel. A proof can be found in [48], Proposition 1.1.1.1. Hence, its expectation is constant, and its conditional variance is

$$\text{Var}(X_{k+1}^N | X_k^N = i) \;\; = \;\; N\frac{i}{N}\left(1 - \frac{i}{N}\right). \tag{1.2}$$

Under the neutral Wright-Fisher model, there exists the possibility to lose genetic variability by pure chance. In other words, the change in allele frequencies is caused by the random variation in individual reproduction.

Now, we measure the time in units of N generations and we renormalize X^N as follows,

$$Y_t^N := \frac{X_{\lfloor Nt \rfloor}^N}{N},\ t \geq 0. \tag{1.3}$$

The process $\{Y_t^N\}_{t\geq 0}$ gives the A-allele frequency process (rescaled in time).

Now, suppose that the proportion of the population of type A at time zero is $Y_0^N = p$. Then, by taking time intervals of length $1/N$

$$\mathbb{E}[Y_{1/N}^N | Y_0^N = p] = \frac{1}{N}\mathbb{E}\left[X_1^N \middle| X_0^N = Np\right] = \frac{1}{N}Np = p \tag{1.4}$$

and the conditional variance is

$$\text{Var}(Y_{1/N}^N \big| Y_0^N = p) = \frac{1}{N^2}\mathbb{E}\left[\left(X_1^N - \mathbb{E}[X_1^N]\right)^2 \big| X_0 = Np\right] = \frac{p(1-p)}{N}. \tag{1.5}$$

Furthermore, $\mathbb{E}\left[\left(Y_{1/N}^N - p\right)^m \big| Y_0^N = p\right] = O\left(1/N^2\right)$ for all $m \geq 3$.

These computations provide some intuition for a diffusion approximation ([29], Chapter 10). The following theorem enunciates this formally.

Theorem 1.1.3. *[43] Let $\{Y_t^N\}_{t\geq 0}$ be the rescaled frequency process of individuals with A-allele in the Wright-Fisher model, started from $Y_0^N = p \in [0,1]$. If Y_0^N converges in distribution to a random variable Y_0, then*

$$\{Y_t^N\}_{t\geq 0} \Rightarrow \{Y_t\}_{t\geq 0}$$

(weakly on the Skorohod space of càdlàg functions with values in $[0,1]$*), where* Y *is the strong solution to the following stochastic differential equation:*

$$dY_t = \sqrt{Y_t(1-Y_t)}dB_t; Y_0 = p \tag{1.6}$$

where B *is a standard Brownian motion. The process* $\{Y_t\}_{t\geq 0}$ *is the so-called Wright-Fisher diffusion.*

Before giving a sketch of the theorem's proof, we introduce the definition of *the infinitesimal generator*.

Definition 1.1.4. ([29] Chapter 1, Section 1 and Chapter 4) Let $\{Y_t\}_{t\geq 0}$ be a Markov process with state space $\mathbb{R}$. For a function $f : \mathbb{R} \to \mathbb{R}$ with $||f|| = \sup_{y\in\mathbb{R}} |f(y)| < \infty$ and $y \in \mathbb{R}$, the infinitesimal generator is the linear operator $\mathcal{A}$ defined by

$$\mathcal{A}f(y) = \lim_{t\to 0} \frac{\mathbb{E}[f(Y_t) - f(Y_0)|Y_0 = y]}{t}. \tag{1.7}$$

The domain $\mathcal{D}(\mathcal{A})$ of $\mathcal{A}$ is the subspace of all f for which this limit exists.

To prove Theorem 1.1.3, we need the generator of Y^N to converge to the generator of the Wright-Fisher diffusion. Then, we can use Theorem 1.1 in Chapter 10 in [29]. Following the idea in [27], Chapter 2, we study how $\mathbb{E}[f(Y_t^N)]$ behaves over time, where $f : [0,1] \to \mathbb{R}$ is a *nice* function. Intuitively, if $t = 1/N$ in the process Y^N and N converges to infinity, the limit is the generator of Y, that is, by equation (1.7)

$$\mathcal{A}f(p) = \lim_{N\longrightarrow\infty} \frac{\mathbb{E}[f(Y_{1/N}^N)\big|Y_0^N = p] - \mathbb{E}[f(p)]}{1/N}.$$

By Taylor's theorem, and equations (1.4) and (1.5), there exists a constant C such that

$$\begin{aligned}
&\mathbb{E}[f(Y_{1/N}^N)\big|Y_0^N = p] \\
&= \mathbb{E}\left[f(p) + f'(p)(Y_{1/N}^N - p) + \frac{f''(p)}{2}\left(Y_{1/N}^N - p\right)^2 + C\left(Y_{1/N}^N - p\right)^3 \big|Y_0^n = p\right] \\
&= f(p) + \frac{f''(p)}{2}\frac{p(1-p)}{N} + O\left(\frac{1}{N^2}\right),
\end{aligned}$$

and then,

$$N\left(\mathbb{E}[f(Y_{1/N}^N)\big|Y_0^N = p] - \mathbb{E}[f(p)]\right) = \frac{f''(p)}{2}p(1-p) + O\left(\frac{1}{N}\right).$$

Thus, the generator of the limit process is

$$\mathcal{A}f(p) = \frac{p(1-p)}{2}f''(p). \tag{1.8}$$

This is the generator of the Wright-Fisher diffusion, which is the strong solution of the stochastic differential equation (1.6) (see Chapter 8 and Chapter 10 in [29]).

1.1.2 The Kingman coalescent

Consider a population with the Wright-Fisher model dynamics (the types of alleles do not matter). Then, take a sample of size two from the population of size N at generation k, with k fixed. The probability that these individuals have the same parent at the previous generation, $k-1$, is N^{-1}. Since the choice of the parents is made independently at each generation, the time to the most recent common ancestor (MRCA) of this sample has a geometric distribution with parameter N^{-1}. It implies that the expected number of generations to find the MRCA in a sample of size two is N. Usually, we are interested in large populations, so this suggests measuring time in units of size N, which corresponds to the time rescaling used in equation (1.3). Under this time rescaling, the distribution of the time to the MRCA converges to an exponential distribution with parameter one, that is, for $t \geq 0$

$$\mathbb{P}(\text{Time to the MRCA} > \lfloor tN \rfloor) = \left(1 - \frac{1}{N}\right)^{\lfloor tN \rfloor} \to e^{-t} \text{ as } N \to \infty.$$

Next, we consider a sample of size $2 \leq n \leq N$ at generation k. The probability that three individuals have the same parent one generation back is of order N^{-2}, and the probability that two have the same parent is $\binom{n}{2}N^{-1}$. Hence, it is more likely that two individuals have the same parent than three (or more). Following the same idea as for a sample of size two (we measure the generations in units of size N), the distribution of the time that any two ancestral lineages merge converges to an exponential distribution with parameter $\binom{n}{2}$. When this merger happens, there remain $n-1$ lineages because the first merger only affects two lines that were chosen uniformly at random, and again the distribution of the time to a new coalescence converges to an exponential with parameter $\binom{n-1}{2}$ and so on ([27], Chapter 2).

The family tree of these n individuals taken in a fixed generation k can be described by a sequence of equivalence relations [44], $\{W_g\}_{g\geq 0}$, on $[n] = \{1, 2, ..., n\}$, such that i and j, individuals of generation k, are in the same element (block) of W_g if they have a common ancestor at generation $k-g$. Each block of W_g corresponds to an individual at generation $k-g$. Suppose two individuals at generation $k-g$ choose the same parent at $k-g-1$. In that case, it implies that their corresponding equivalence classes of W_g are merged in W_{g+1}; otherwise, the blocks are not merged. Then, $\{W_g\}_{g\geq 0}$ is a discrete time partition-valued Markov chain. The next theorem says that this Markov chain converges in distribution with a specific time scale.

Theorem 1.1.5. *[44]. The process $\{W_{\lfloor Nt \rfloor}\}_{t\geq 0}$, converges in distribution, as $N \to \infty$, to $\{K_t^n\}_{t\geq 0}$. The limit of this genealogical process is called the Kingman n-coalescent.*

Before giving the formal definition of the Kingman n-coalescent, we introduce some notation. Let $\mathcal{P}_n$ be the set of partitions of $[n]$ and let $\mathcal{P}$ be the set of partitions on $\mathbb{N}$.

Definition 1.1.6. The Kingman coalescent, [45, 44]. A n-coalescent is a continuous time Markov chain, $\{K_t^n\}_{t\geq 0}$, with values in the state space $\mathcal{P}_n$, and with the following transition rates: let $\pi_1, \pi_2 \in \mathcal{P}_n$,

$$\pi_1 \mapsto \pi_2 \text{ at rate } \begin{cases} 1 \text{ if } \pi_1 \prec \pi_2, \\ 0 \text{ otherwise.} \end{cases}$$

The notation $\pi_1 \prec \pi_2$ means that π_2 is obtained from π_1 by merging exactly two blocks of π_1. The initial partition, K_0^n, is almost surely the trivial partition in singletons.

Kingman's n-coalescent satisfies the consistency property: let $m < n$ and denote by ρ^m the restriction map to $\mathcal{P}_m$, then $\{\rho^m \circ K_t^n\}_{t\geq 0} \overset{d}{=} \{K_t^m\}_{t\geq 0}$. Thus, by Kolmogorov's extension theorem, there exists a unique continuous time Markov chain $\{K_t\}_{t\geq 0}$, with state space $\mathcal{P}$, such that its restriction to $\mathcal{P}_n$ is equal in distribution to the n-coalescent, for each n. This projective limit is called the *Kingman coalescent*.

Observe that in Kingman's coalescent, two blocks of K_t merge independently of their size and the value of K_t for every $t \geq 0$. This is due to the property of *exchangeability* of the coalescent, meaning that the law of K_t is invariant under any permutation of $\mathbb{N}$ with finite support. This property also holds for the n-coalescent ([2], Chapter 2).

Another property of the Kingman coalescent is that the expected time of the MRCA of an (infinite) sample is finite and equals two. More precisely, let T_n be the time to the MRCA of a sample of n individuals and let τ_i be the amount of time during which there are i lineages. Observe that τ_i has an exponential distribution with parameter $\binom{i}{2}$ for $2 \leq i \leq n$, and that $T_n = \tau_2 + \tau_3 + \cdots + \tau_n$, then

$$\mathbb{E}[T_n] = \sum_{i=2}^{n} \mathbb{E}[\tau_i] = \sum_{i=2}^{n} \frac{1}{\binom{i}{2}} = 2\sum_{i=2}^{n} \frac{1}{i-1} - \frac{1}{i} = 2\left(1 - \frac{1}{n}\right) \xrightarrow[n\to\infty]{} 2 \tag{1.9}$$

Observe that $\mathbb{E}[\tau_2] = 1$, which means that the expected time of the final coalescence is half of the expected time to the MRCA of a sample of infinite size. Furthermore, the Kingman coalescent *comes down from infinity*: while starting with an infinite number of blocks, instantaneously, its number of blocks is finite, almost surely.

Theorem 1.1.7. *Let $\{N_t\}_{t\geq 0}$ be the block counting process of the Kingman coalescent (starting from an infinite number of singletons). Then almost surely, for every $t > 0$, N_t is finite.*

This is intuitively clear from the comments above. The reader can see a proof in [2], Chapter 2.

Observe that the process N is a death process with jumps of size one. The process N goes from state n to state $n-1$ at rate $\binom{n}{2}$, so by Definition 1.1.4, the generator of N is

$$\mathcal{G}f(n) = \binom{n}{2}(f(n-1) - f(n)) \tag{1.10}$$

1.1.3 Duality

In [58], the concept of *duality* defined in [53] is used to give a relation between the Wright-Fisher diffusion and the Kingman coalescent. This relation is a coupling of the forward process, $\{Y_t\}_{t\geq 0}$, which describes the evolution of the population forward in time, and the backward process, $\{N_t\}_{t\geq 0}$, which is the block counting process associated to the Kingman coalescent $\{K_t\}_{t\geq 0}$. The next theorem is the formalization of this but let us first introduce the definition of *duality*.

Definition 1.1.8. (Definition 1.1 in [55]) The process $\{V_t\}_{t\geq 0}$ is dual to the process $\{W_t\}_{t\geq 0}$ with respect to F, a bounded measurable function on $E_1 \times E_2$, where E_1 is the state space of V and E_2 is the state space of W, if

$$\mathbb{E}_v[F(V_t, w)] = \mathbb{E}_w[F(v, W_t)] \tag{1.11}$$

for all $v \in E_1$, $w \in E_2$ and $t \geq 0$. Here $\mathbb{E}_v$ denotes the expectation given that $V_0 = v$ and $\mathbb{E}_w$ denotes the expectation given that $W_0 = w$.

Theorem 1.1.9. *[58] Let $\{Y_t\}_{t\geq 0}$ be the Wright-Fisher diffusion starting at $Y_0 = p \in (0, 1)$ and let $\{N_t\}_{t\geq 0}$ be the block counting process of the Kingman coalescent starting at $N_0 = n \in \mathbb{N}$. Then, the process $\{Y_t\}_{t\geq 0}$ is moment dual to the process $\{N_t\}_{t\geq 0}$, that is*

$$\mathbb{E}_p[Y_t^n] = \mathbb{E}_n[p^{N_t}]. \tag{1.12}$$

In this case $F(p, n) = p^n$. Observe that $\mathbb{E}_p[Y_t^n]$ is the n-th moment of Y_t, thus (1.12) characterizes the moments of Y.

The proof of this result can be easily obtained thanks to Proposition 1.2 in [40], saying that the duality has to be verified only at the level of generators. More precisely, let $\mathcal{A}$ be the generator of $\{Y_t\}_{t\geq 0}$ as in (1.8) and let $\mathcal{G}$ be the generator of $\{N_t\}_{t\geq 0}$ as in (1.10). If $F(p, \cdot)$ and $\mathbb{E}_n[F(p, \cdot)]$ are in $\mathcal{D}(\mathcal{G})$ for all $p \in [0, 1]$ and if $F(\cdot, n)$ and $\mathbb{E}_p[F(\cdot, n)]$ are in $\mathcal{D}(\mathcal{A})$ for all $n \in \mathbb{N}$, and if

$$\mathcal{A}F(\cdot, n)(p) = \mathcal{G}F(p, \cdot)(n), \tag{1.13}$$

then, $\{Y_t\}_{t\geq 0}$ and $\{N_t\}_{t\geq 0}$ are duals with respect to F. Now, observe that

$$\mathcal{A}F(\cdot, n)(p) = \frac{1}{2}p(1-p)\frac{\partial^2 F}{\partial p^2} = \frac{1}{2}p(1-p)n(n-1)p^{n-2}$$

and

$$\mathcal{G}F(p, \cdot)(n) = \frac{1}{2}n(n-1)(p^{n-1} - p^n) = \frac{1}{2}n(n-1)p^{n-1}(1-p).$$

This proves the duality of the generators and hence of the processes.

There is a natural interpretation of this duality result. Observe that, on one hand, $\mathbb{E}_p[Y_t^n]$ is the probability that n individuals, chosen at random in the infinite population at time t, have the A-allele given that at time zero the population has a fraction p of A-allele individuals. On the other hand, $\mathbb{E}_n[p^{N_t}]$ is the probability that the N_t ancestors of the same sample of size n (chosen at time t) carry the A-allele t units of time in the past, because at time zero we know that $Y_0 = p$. The fact that those two values are equal means that one model stands for the backward in time version of the other. The duality theorem gives us two forms to calculate the moments of the marginals of the Wright-Fisher diffusion. Commonly it is easier to work with $\mathbb{E}_n[p^{N_t}]$ than $\mathbb{E}_p[Y_t^n]$.

This method can be extended in discrete time thanks to the same interpretation. Using the notation of Sections 1.1.1 and 1.1.2. the probability that n individuals chosen at random at generation g have the A-allele is $\mathbb{E}_p[(Y_g^N)^n]$ for $p \in \{\frac{1}{N}, \frac{2}{N}, ..., \frac{N}{N}\}$. Going g generations backward in time, this expectation is equal to $\mathbb{E}_n[p^{W_g}]$. This gives an easy intuition for a duality result that is harder to obtain trough computations than in the continuous case [34].

1.1.4 Cannings' model

We now give the definition of a more general model with discrete generations and fixed size of population, which is the key of *Möhle's lemma*, and is known as *Cannings' model* (thanks to works of Cannings [14],[15]). Recall that a random vector $(\nu_1, \nu_2, \ldots, \nu_N)$ is exchangeable if its law is invariant under permutations of its labels. More explicitly, for any permutation $\sigma = (\sigma(1), \ldots, \sigma(N))$ of $\{1, ..., N\}$,

$$(\nu_1, \nu_2, \ldots, \nu_N) \overset{d}{=} (\nu_{\sigma(1)}, \nu_{\sigma(2)}, \ldots, \nu_{\sigma(N)}).$$

Definition 1.1.10. Cannings' model. Consider a population of fixed size N, where N is a positive integer. The population is haploid and neutral. Each individual in each generation is randomly labeled from 1 to N. Generation $g+1$ is generated by the children of individuals from generation g. Each individual in generation $g+1$ chooses its parent from generation g randomly and independently according to an exchangeable random vector $(\nu_1, \nu_2, \ldots, \nu_N)$ such that $\sum_{i=1}^N \nu_i = N$. Intuitively, the random variable ν_i indicates the law of the number of offspring of the ith individual at each generation.

We are going to analyze the genealogical process as in the Wright-Fisher model. First, we define

$$c_N = \mathbb{E}\left[\frac{\nu_1(\nu_1 - 1)}{N - 1}\right] \tag{1.14}$$

which is the probability for two randomly chosen individuals at the same generation of having the same parent. To see this, denote this event by A and observe that

$$\begin{aligned}\mathbb{P}(A) =& \mathbb{E}[\mathbb{P}(A|(\nu_1, \nu_2, \ldots, \nu_N))] \\ =& \mathbb{E}\left[\sum_{i=1}^{N} \frac{\nu_i(\nu_i - 1)}{N(N-1)}\right] \\ =& \mathbb{E}\left[\frac{\nu_1(\nu_1 - 1)}{N-1}\right]\end{aligned}$$

where the last equality is obtained thanks to exchangeability. Note that $c_N = 0$ if and only if $\nu_i = 1$ a.s., for all $i \in [N]$. It is assumed in the sequel that $c_N > 0$ for all N. The value c_N is called the coalescent probability.

The Wright-Fisher model is a particular case of Cannings' model when the vector $(\nu_1, \nu_2, \ldots, \nu_N)$ has a multinomial distribution with parameters N and $p_i = 1/N$ for all $i \in [N]$. In this case $c_N = 1/N$.

Following the same analysis that we realized in Section 1.1.2, the law of the time to the MRCA in the Cannings' model for a sample of size two is geometric with parameter c_N, and its expectation is $1/c_N$, it suggests accelerating the time by $1/c_N$ to obtain a non degenerate limit as $N \to \infty$. Assume that $c_N \to 0$. Then, for $t \geq 0$,

$$\mathbb{P}(\text{ Time to the MRCA} > \lfloor t/c_N \rfloor) = (1 - c_N)^{\lfloor t/c_N \rfloor} \to e^{-t}.$$

Now, we consider a sample of size three at a given generation, the probability that these three individuals have the same parent one generation before is,

$$d_N = \mathbb{E}\left[\frac{\nu_1(\nu_1 - 1)(\nu_1 - 2)}{(N-1)(N-2)}\right].$$

Let $\sim_r$ be the equivalence relation on $[n]$, defined by $i \sim_r j$ if and only if the ith and jth individuals of a sample of size $n \leq N$ have a common ancestor r generations backward in time. Let $\{C_g\}_{g \geq 0}$ be the discrete Markov chain with values in the state space $\mathcal{P}_n$, the set of all equivalence relations on $[n]$ (it is constructed similar to the process $\{W_g\}_{g \geq 0}$ in Section 1.1.2) and let $D_{\mathcal{P}_n}[0, \infty)$ be the space of cádlág functions from $[0, \infty)$ to $\mathcal{P}_n$ together with the Skorokhod topology.

Theorem 1.1.11. *[57]. a) The time scaled ancestral process $\{C_{\lfloor t/c_N \rfloor}\}_{t \geq 0}$ converges in distribution, in $D_{\mathcal{P}_n}([0, \infty))$, to the Kingman n-coalescent $\{K_t^n\}_{t \geq 0}$ if and only if*

$$\lim_{N \to \infty} \frac{d_N}{c_N} = 0. \tag{1.15}$$

This means that the Kingman n-coalescent appears in the limit, as N tends to infinity if and only if triple mergers of ancestral lineages are asymptotically negligible in comparison with binary mergers [56].

b) The limit (1.15) implies that,

$$\lim_{N\to\infty} c_N = 0 \ \ and \ \ \lim_{N\to\infty} \frac{\mathbb{E}[\nu_1(\nu_1-1)\nu_2(\nu_2-1)]}{N^2 c_N} = 0. \tag{1.16}$$

The last limit means that, as N tends to infinity, simultaneous coalescences of ancestral lineages are asymptotically negligible in comparison with binary mergers.

This result is, actually, part of a much more general work found in [61], and we will discuss it later. Measuring time in units of size $1/c_N$ and assuming that (1.15) holds, we can adapt Theorem 1.1.3 and Theorem 1.1.9 to the case of large populations evolving according to Cannings' model, see Section 2.2 in [29].

When $\lim_{N\to\infty} d_N/c_N \neq 0$, we can have multiple collisions; that is, many mergers can occur to obtain one single line or simultaneous multiple collisions, which means that many mergers can happen to get one single line. Many of these can occur at the same time.

In [63], Sagitov introduced coalescents with multiple collisions as limits of ancestral processes of a population evolving according to the Cannings' model. Sagitov used a proper time-scale factor for their asymptotic analysis, similar to what we present in this section. See also [59] for a simpler formulation.

In [62], Pitman, independently, introduced and studied coalescents with multiple collisions as Markov and exchangeable coagulating processes (and without considering their connection to Cannings' model), and called this class of processes Λ-coalescents. They also appear in [22].

Theorem 1.1.12. *[62]. Let $(\lambda_{b,k}, 2 \leq k \leq b < \infty)$ be an array of non-negative real numbers. There exist for each $\pi \in \mathcal{P}$ a $\mathcal{P}$-valued coalescent Π with $\Pi_0 = \pi$, whose restriction Π^n to $[n]$ is for each n a Markov chain such that, when Π^n_t has b blocks, each k-tuple of blocks of Π^n_t is merging to form a single block at rate $\lambda_{b,k}$, if and only if*

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

for some non-negative and finite measure Λ on Borel subset of $[0,1]$. We call the $\mathcal{P}$-valued Markov process induced by a finite measure Λ on $[0,1]$ the Λ-coalescent.

For $\Lambda = \delta_0$, the point mass at zero, we obtain the Kingman coalescent. In [65], Schweinsberg showed that the Λ-coalescent comes down from infinity if and only if

$$\sum_{b=2}^{\infty} \gamma_b^{-1} < \infty$$

where $\gamma_b = \sum_{k=2}^{b} (k-1)\binom{b}{k}\lambda_{b,k}$ is the rate at which the number of blocks is decreasing. More generally, in [60], Möhle and Sagitov established the complete picture of limit genealogies of Cannings' model. A general coalescent structure allowing simultaneous and multiple collisions of ancestral lines is obtained. The central condition of their main result requires the existence of the limits

$$\lim_{N\to\infty} \frac{\mathbb{E}[(\nu_1)_{k_1} \cdots (\nu_j)_{k_j}]}{N^{k_1+\cdots+k_j-j} c_N} := \phi_j(k_1, \cdots, k_j)$$

for all $j \in \mathbb{N}$ and $k_1 \geq \cdots \geq k_j \geq 2$.

In [64], Schweinsberg introduced and studied the class of coalescents with simultaneous and multiple collisions thanks to exchangeability tools. The following definition describes the more general family of exchangeable coalescents.

Definition 1.1.13. [64], [3]. Let Ξ be a finite measure on the infinite simplex,

$$\Delta = \{(x_1, x_2, ...) : x_1 \geq x_2 \geq ... \geq 0\,, \sum_{i=1}^{\infty} x_i \leq 1\},$$

and write $\Xi = \Xi_0 + a\delta_0$ where Ξ_0 has no atom at zero, δ_0 is the unit mass at zero and $a \geq 0$. The Ξ-coalescent (with characteristic measure Ξ) is a process $\{\Pi_t\}_{t\geq 0}$ with values in $\mathcal{P}$ with the property that for each $n \in \mathbb{N}$, its restriction to $[n]$ is a $\mathcal{P}_n$-valued Markov chain. Let ξ be a partition with b blocks and let η be a partition obtained by merging disjoint groups of blocks of ξ, such that η has $r + s$ blocks in which s blocks remain unchanged and the other r blocks contain $k_1, k_2, ..., k_r \geq 2$ of the original blocks. Thus $b = \sum_{i=1}^{r} k_i + s$ and the rate of transition from ξ to η is $\lambda_{b;k_1,\cdots,k_r;s}$ equals

$$\lambda_{b;k_1,\cdots,k_r;s} = \int_{\Delta} \frac{\sum_{l=0}^{s} \sum_{i_1\neq\ldots\neq i_{r+l}} \binom{s}{l} x_{i_1}^{k_1} \cdots x_{i_r}^{k_r} x_{i_{r+1}} \cdots x_{i_{r+l}} \left(1 - \sum_{j=1}^{\infty} x_j\right)^{s-l}}{\sum_{j=1}^{\infty} x_j^2} \Xi_0(dx)$$

$$+ a\mathbf{1}_{\{r=1,k_1=2\}}.$$

The Λ-coalescent is a special case of the Ξ-coalescent when Ξ is concentrated on the subset of Δ consisting of the sequence $(x_1, x_2, \cdots)$ such that $x_i = 0$ for all $i \geq 2$. A

sufficient condition for a Ξ-coalescent to come down from infinity was also presented in [64].
As for the Wright-Fisher model, a general moment duality result, between Cannings' model and its ancestry process, was established in [34]. As $N \to \infty$, the moment duality relation still holds as the ancestry process converges to a Ξ-coalescent and Theorem 1.1.9 can be generalized. The associated forward in time process turns to a diffusion with jumps. More precisely, this process corresponds to the frequency of one type in a Ξ-Fleming-Viot process [7]. Previously, the duality relation for the Λ-coalescent was established in [4].

1.2 A weak seed bank model

In this section and the next one, we focus on two models that include some seed bank effects. These models were introduced in the last two decades ([41], [8], [9]) and they became an essential topic in population genetics. Indeed, phenomena of dormancy were observed in nature. Individuals can deactivate for some time and then awake. This seed bank effect can yield important modifications in the evolution of certain plant populations. For example, large fluctuations in the population size of *Linanthus parryae* in the Mojave desert is explained by some delay in seeds germination for many years [26]. A seed bank effect was also observed for eggs or bacteria [37]. The presence of seed banks dampens selective presence caused by environmental variables and, thus, increases the genetic variation (e.g., the number of mutations). Some effects of the seed bank were studied in [68], [50].

This section introduces one class of seed bank models, where the dormancy period is smaller than the diffusive time rescaling. Under this model, [71] observed that the relative allele frequencies within a sample are unchanged. The mean waiting times to

Figure 1.1: Linanthus parryae [1]

coalescence are similar to those of a population without a bank. This is known as the *weak seed bank effect*.

1.2.1 Definition

We first introduce a modification of the Wright-Fisher model introduced by Kaj, Krone, and Lascoux [41] (and extended in [8]) where the individuals can choose their parent at some random generation in the past.

Definition 1.2.1. Weak seed bank model, [41], [8]. Consider a haploid (neutral) population of fixed size $N \in \mathbb{N}$. Consider individuals $\nu = (i_\nu, k_\nu) \in V_N := \mathbb{Z} \times [N]$ where i_ν denotes the generation and k_ν the label among the N individuals alive at this generation. Let $A(\nu_0) = \{\nu_j\}_{j\geq 0}$ be the ancestral line of individual ν_0. In particular, $\{i_{\nu_j}\}_{j\geq 0}$ is a strictly decreasing sequence of generations, with i.i.d. decrements $\{i_{\nu_j} - i_{\nu_{j-1}}\}_{j\geq 1}$having distribution μ. The variables $\{k_{\nu_j}\}_{j\geq 0}$ are independent and uniformly distributed in $[N]$, independent of $\{i_{\nu_j}\}_{j\geq 0}$.

Observe that, if $\mu = \delta_1$, the classical Wright-Fisher model is recovered. In that case, all individuals choose their parent in the previous generation. In the reference [41], the weak seed bank model is introduced with μ having a finite support, while [8] extends this model to μ with infinite support and finite expectation.

1.2.2 Convergence of the genealogical process

Kaj et al. [41] and Blath et al. [8] showed that the limit of the rescaled ancestral process of a sample of size n converges, as N goes to infinity, to a time-changed Kingman n-coalescent. The coalescence rate should be slowed down due to the ancestors' structure, since two lineages jump among generations before they fall into a common generation, and one coalescence can occur.

In this introduction, we present the case when μ has finite support $[m]$. For the sake of simplicity, we consider the block counting process associated with the ancestral lineages. For this purpose, we see the population's state through a window consisting of m consecutive generations.

Definition 1.2.2. Ancestral process. Consider a sample of $n \leq N$ individuals that live between generation 0 and $m-1$ (backward in time). Let $A_{N,n}(k)$ be the number of most recent ancestors of the sample living between generation k and $k+m-1$, see Figure 1.2. The sample reaches its MRCA when $A_{N,n}(k) = 1$ for some $k \geq 0$.

The definition is generalized in [8] to a partition-valued process, and its convergence towards a time-changed Kingman coalescent is established.

Theorem 1.2.3. *[41],[8]. Suppose that* $\beta := 1/\mathbb{E}[i_{\nu_1} - i_{\nu_0}] > 0$. *As* $N \to \infty$, *the ancestral process* $\{A_{N,n}(\lfloor Nt/\beta^2 \rfloor)\}_{t\geq 0}$ *converges weakly in* $D_{[n]}[0,\infty)$ *to the block counting process of a Kingman* n*-coalescent.*

The key tool to prove this result relies on studying the time to the ancestor of two individuals. The complete picture is as follows.

Theorem 1.2.4. *[8]. Let* $\nu, \omega \in V_N$ *belonging to the same generation and let* τ *be the time to their MRCA,*

$$\tau := \inf\{i \geq 0 : A(\nu) \cap A(\omega) \cap (\{-i\} \times [N]) \neq \varnothing\}.$$

Suppose that the tails of μ *are of the form* $\mu(n, n+1, ...) = n^{\alpha}L(n)$ *for* $\alpha \in (0,\infty)$ *and a slowly varying function* L.

(i) If $\alpha \in (0, \frac{1}{2})$ *then* $\mathbb{P}(A(\nu) \cap A(\omega) \neq \varnothing) < 1$ *for all* $N \in \mathbb{N}$.

(ii) If $\alpha \in (\frac{1}{2}, 1)$ *then* $\mathbb{P}(A(\nu) \cap A(\omega) \neq \varnothing) = 1$ *and* $\mathbb{E}[\tau] = \infty$ *for all* $N \in \mathbb{N}$.

(iii) If $\alpha > 1$ *then* $\mathbb{P}(A(\nu) \cap A(\omega) \neq \varnothing) = 1$ *for all* $N \in \mathbb{N}$ *and* $\lim_{N\longrightarrow\infty} \mathbb{E}[\tau]/N = 1/\beta^2$.

The intuitions of the proof of Theorem 1.2.3 can be found in [41], when μ has finite support $[m]$ (satisfying case *(iii)*). As we mentioned before, the population's state can be seen through a window, which consists in m consecutive generations. We formulate this in terms of a space-time urn model. Each urn represents a generation and is labeled $0, 1, ...$ The urn with label 0 corresponds to the current generation, and the other urns correspond to the generations $1, 2, ...$ backward in time. Furthermore, each urn contains N cells, and the balls inside the urns represent the individuals. A m-window is a collection of m consecutive urns. The kth window consists in urns $(k, k+1, \ldots, k+m-1)$.

For $1 \leq n \leq N$, let

$$S_n := \left\{ (x_1, x_2, ..., x_m),\, x_i \in \mathbb{N} \cup \{0\},\, \sum_{i=1}^{m} x_i \leq n \right\}.$$

Let $X(0) = (X_1(0), ..., X_m(0)) \in S_n$ such that $X_i(0)$ is the number of balls in the $(i-1)$th urn at the 0th m-window. The starting configuration is such that $\sum_{i=1}^{m} X_i(0) = n$. The transition from the 0th m-window to the 1st m-window is made by relocating the $X_1(0)$ balls (the individuals in the 0th urn) independently to the urns $1, 2, \ldots, m$ according to the probabilities $\mu(1), \mu(2), \ldots, \mu(m)$, respectively. Then, they choose a cell into the urn uniformly. If a ball falls into an occupied cell or two balls fall in the same (empty) cell, a coalescence occurs, and thus, the total number of balls decreases

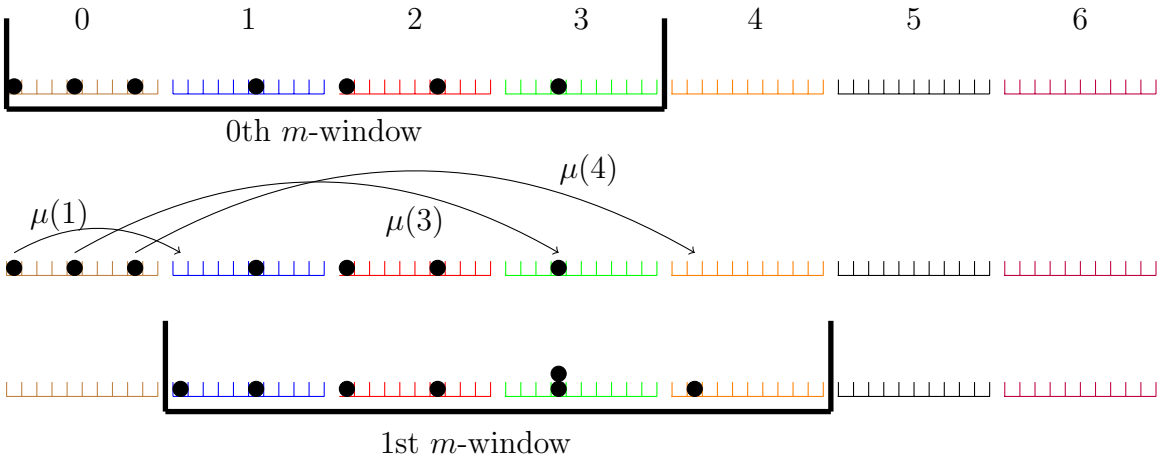

Figure 1.2: An example of jump from the 0th m-window to the 1st m-window with $m = 4$, $N = 10$ and $n = 7$. In this case $A_{N,n}(0) = 7$ and $A_{N,n}(1) = 6$. A coalescence is produced because one relocated ball falls into an occupied cell.

by one; that is, the balls merge into a single ball. The balls in urns $1, \ldots, m-1$ do not move, see Figure 1.2. Then, $X(k) = (X_1(k), ..., X_m(k))$, where $X_i(k)$ is the number of balls in urn $i + k - 1$, and the transition from the kth m-window to the $(k+1)$th m-window is made by relocating the balls in the kth urn to the urns $k+1, \ldots, k+m-1$ similar as we did before. Thus, $\{X(k)\}_{k\geq 0}$ is a discrete-time Markov chain with values in S_n. Observe that

$$A_{N,n}(k) = X_1(k) + ... + X_m(k).$$

Now, let $R(k+1) = (R_1(k+1), \ldots, R_m(k+1))$, where $R_i(k+1)$ is the number of relocated balls that fell at the ith urn of the $(k+1)$th m-window. Given $X_1(k)$, $R(k+1)$ has a multinomial distribution with parameters $X_1(k)$ and $\mu(1), \ldots, \mu(m)$. In particular

$$R_i(k+1) \sim Bin(X_1(k), \mu(i)). \tag{1.17}$$

We calculate the probability that a coalescence occurs after relocating balls. For this propose assume that there are b balls in a specific urn and r balls are relocated in this urn.

$$\begin{aligned}
\mathbb{P}(\text{no coalescence}) &= \left(1 - \frac{b}{N}\right)\left(1 - \frac{b+1}{N}\right)\ldots\left(1 - \frac{b+r-1}{N}\right) \\
&= 1 - \frac{1}{N}\sum_{i=0}^{r-1}(b+i) + O\left(\frac{1}{N^2}\right) \\
&= 1 - \frac{1}{N}\left(br + \binom{r}{2}\right) + O\left(\frac{1}{N^2}\right).
\end{aligned}$$

From here, observe that the coalescence of exactly two lineages occurs with a probability approximately equal to $(br + \binom{r}{2})/N$, and the coalescence of three or more lineages occurs with a probability $O(1/N^2)$. For N big enough, the event of a multiple coalescence will result negligible.

Now, denote the ith unit vector by e_i, set for $i \leq m-1$, $a_i = X_{i+1}(k)R_i(k+1) + \binom{R_i(k+1)}{2}$ and $\sigma(X_1(k), \ldots, X_m(k)) := (X_2(k), \ldots, X_m(k), 0)$. If we consider all the urns and condition on $X(k)$ and $R(k+1)$, the chain jumps from $X(k)$ to $X(k+1) = \sigma(X(k)) + R(k+1)$ with probability $1 - \frac{1}{N}\sum_i a_i(k) + O(1/N^2)$ (no coalescence), and jumps to $X(k+1) = \sigma(X(k)) + R(k+1) - e_i$ with probability $\frac{1}{N}a_i(k) + O(1/N^2)$ (one coalescence in the ith urn). In summary, the transitions of the block counting process, given $X(k)$ and $R(k+1)$, are such that

$$\begin{aligned}\mathbb{P}(A_{N,n}(k+1) - A_{N,n}(k) = -1 | X(k), R(k+1)) \\ = \frac{1}{N}\sum_{i=1}^{m-1} X_{i+1}(k)R_i(k+1) + \frac{1}{N}\sum_{i=1}^{m}\binom{R_i(k+1)}{2} + O\left(\frac{1}{N^2}\right).\end{aligned}$$

Then, using (1.17), we get

$$\begin{aligned}\mathbb{P}(A_{N,n}(k+1) - A_{N,n}(k) = -1 | X(k)) \\ = \mathbb{E}[\mathbb{P}(A_{N,n}(k+1) - A_{N,n}(k) = -1 | X(k), R(k+1)) | X(k)] \\ = X_1(k)\frac{1}{N}\sum_{i=1}^{m-1} X_{i+1}(k)\mu(i) + \frac{1}{N}\binom{X_1(k)}{2}\sum_{i=1}^{m}\mu^2(i) + O\left(\frac{1}{N^2}\right).\end{aligned}$$

Finally, define for $j \leq m$,

$$\beta_j = \frac{\sum_{i=j}^{m}\mu(i)}{\sum_{i=1}^{m} i\mu(i)}.$$

Observe that $\sum_{j=1}^{m}\beta_j = 1$ and $\beta_j = \beta_1\sum_{i=j}^{m}\mu(i)$, which implies that the probabilities β_j satisfy

$$\beta_j = \beta_{j+1} + \beta_1\mu(j),\ j = 1, ..., m-1$$

and $\beta_m = \beta_1\mu(m)$. It is also easy to prove that $\beta_1 = \beta$ in Theorem 1.2.3.

It is shown in [41] that

$$\mathbb{E}\left[X_1(0)\frac{1}{N}\sum_{i=1}^{m-1} X_{i+1}(0)\mu(i) + \frac{1}{N}\binom{X_1(0)}{2}\sum_{i=1}^{m}\mu^2(i)\right] = \beta^2\binom{n}{2}$$

Thus, the limit genealogies for this model, when time is multiplied by $1/\beta^2$, are given by the Kingman n-coalescent.

1.2.3 Duality and forward process

Since the limiting genealogies in the weak seed bank model are given by a time-changed Kingman coalescent, dual to a time-changed Wright-Fisher diffusion, it is clear that this diffusion should play the role of the forward frequency process. This observation appears in [8] but no convergence result is provided. In [46], some work is done in this direction considering a slightly different model. In chapter 2 we present an extended work on the weak seed bank model with a more general reproductive mechanism, such as Cannings model. Also we provide forward and backwards convergence results extending [41], [8] and [46] and we establish a duality result in the discrete case thanks to a random graph representation of the model.

1.3 A strong seed bank model

This section presents a seed bank model defined and studied by Blath, González Casanova, Kurt, and Wilke-Berenguer in 2016 [9]. It is a modification of the Wright-Fisher model where the seed bank age distribution is geometric. In this model, unlike the previous one, the evolution is thought like a model with two islands. One island represents the active population while the other island represents the dormant population, and there is migration between them, while the reproductions only happen in the active population. The size of the dormant population is of order N, and the time that individuals can be inactive is also of order N. We denominate these assumptions by the **strong seed bank effect**.

1.3.1 Definition of the model

Definition 1.3.1. Strong seed bank model, [9]. Consider a haploid population of fixed size N. Assume that the population additionally supports a seed bank of constant size M. The N *active* individuals are called *plants* and the M *dormant* individuals are called *seeds*. Let $0 \leq \varepsilon \leq 1$ such that $\lfloor \varepsilon N \rfloor \leq M$ and let $\delta := \varepsilon N/M$. The N plants from generation 0 produce $N - \lfloor \epsilon N \rfloor$ plants by multinomial sampling (as in the Wright-Fisher model) and $\lfloor \epsilon N \rfloor$ seeds in generation 1. Then, $\lfloor \delta M \rfloor = \lfloor \epsilon N \rfloor$ uniformly (without replacement) sampled seeds from the seed bank in generation 0 become plants in generation 1. Thus, generation 1 is again made of N plants and M seeds, see Figure 1.3. This random mechanism is then to be repeated independently to produce the next generations.

Observe that this model has, unlike [41], non-overlapping reproductions.

Suppose that there exist two constants $c, K \in (0, \infty)$ such that

$$\varepsilon = \frac{c}{N} \text{ and } M = \frac{N}{K}. \tag{1.18}$$

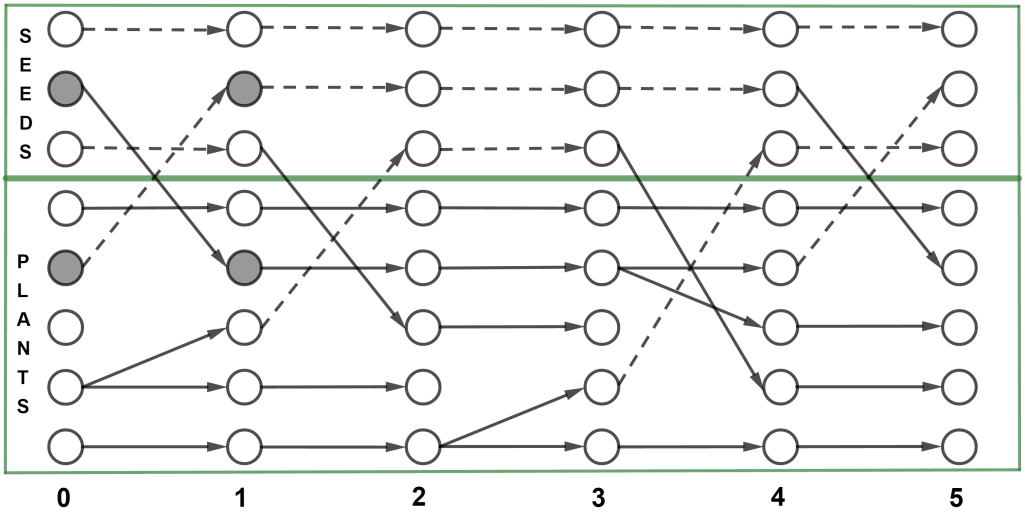

Figure 1.3: The discrete strong seed bank model. In this picture, $N = 5$, $M = 3$ and $\lfloor \varepsilon N \rfloor = \lfloor \delta M \rfloor = 1$, i.e., four plants are produced in each generation by active individuals, one seed germinates, and one new seed is produced.

In words, c corresponds to the number of seeds that become plants (activation), and to the number of plants that become seeds (deactivation) at each generation. Then, the seed bank age distribution is geometric with parameter

$$\delta = \frac{cK}{N}.$$

The parameter K is the relative size of the seed bank with respect to the active population [9].

1.3.2 The seed bank coalescent

The stochastic process that describes the limiting genealogy of a sample taken from the strong seed bank model is called the *seed bank n-coalescent*. We need to introduce some further notations before we can properly define this process.

The set of marked partitions $\mathcal{P}_n^{\{p,s\}}$, where s means seed and p means plant, is similar to the set of partitions of $[n]$, $\mathcal{P}_n$, but additionally each block of a partition $\pi \in \mathcal{P}_n^{\{p,s\}}$ has a flag which can be either p or s. For example, $\pi = \{1,2,3\}^p, \{4\}^s, \{5,6\}^s, \{7\}^p$ is an element of $\mathcal{P}_7^{\{p,s\}}$.

Definition 1.3.2. The seed bank coalescent, [9]. The seed bank n-coalescent $\{\Pi_t^n\}_{t\geq 0}$, with seed bank intensity $c > 0$ and relative seed bank size $1/K > 0$, is the

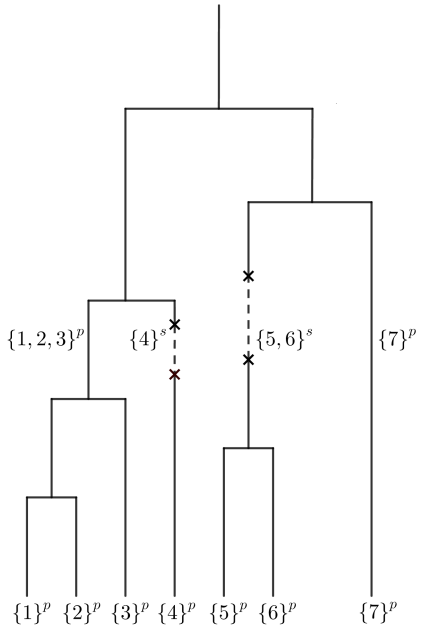

Figure 1.4: A possible realization of the seed bank 7-coalescent. Dotted lines indicate inactive individuals, and the crosses mean that an individual becomes a plant or a seed.

continuous time Markov chain with values in $\mathcal{P}_n^{\{p,s\}}$ having the next dynamics. Each pair of plant blocks merges at rate 1, independent of each other. Moreover, any block can change its flag from p to s at rate c, and vice versa at rate cK.

The seed bank coalescent, $\{\Pi(t)\}_{t\geq 0}$ is defined as the unique Markov process distributed as the projective limit, as n goes to infinity, of $\{\Pi_n(t)\}_{t\geq 0}$.

The block-counting process of the seed bank coalescent is the two-dimensional Markov chain $\{N_n(t), M_n(t)\}_{t\geq 0}$ with values in $(\mathbb{N}\cup\{0\})\times(\mathbb{N}\cup\{0\})$ and the following transition rates, for $t \geq 0$.

$$(N(t), M(t)) \text{ jumps from } (i,j) \text{ to } \begin{cases} (i-1, j), & \text{ar rate } \binom{i}{2} \quad \text{(coalescence)} \\ (i-1, j+1), & \text{at rate } ic \quad \text{(deactivation)} \\ (i+1, j-1), & \text{at rate } jcK \quad \text{(activation).} \end{cases}$$

Note that, for $t \geq 0$, $N(t)$ can have either an upward jump if a seed becomes a plant, or a downward jump if there is a coalescent event or a plant becomes a seed. Each jump has size one.

To see that the seed bank n-coalescent is the limit genealogy of a sample taken from the strong seed bank model, we can adapt the intuitions of section 1.1.2. Take a sample of size $n \leq N$ where, for simplicity, all individuals are plants from generation

zero. We go backward in time, and in each generation, we verify if there is an ancestor of the sample among the plants or the seeds. Let $\{\Pi_n^N(k)\}_{k\geq 0}$ denote the ancestral process with values in $\mathcal{P}_n^{\{p,s\}}$, where two individuals belong to the same block of $\Pi_n^N(k)$ if they have a common ancestor at generation $-k$ and the flag of the block indicates if the ancestor is a plant or a seed.

Observe that, in the seed bank model, only the plant lineages can coalesce. Then, by (1.18), the probability that a given block with flag p changes its flag by s at the next generation is $\varepsilon = c/N$. The probability that a given block with flag s changes its flag by p is $\delta = cK/N$, and the probability that two given blocks with flag p merge is $(1 - c/N)^2\, 1/N$. We start with n blocks, and the blocks' dynamics are independent. The probability of having simultaneous changes of flags, simultaneous or multiple coalescence is of order $1/N^2$ or smaller. A recent work of Blath, Gónzalez Casanova, Kurt, and Wilke-Berenguer, [10], presents the seed bank coalescent with simultaneous switching.

Theorem 1.3.3. *[9]. For any $n \in \mathbb{N}$, suppose that (1.18) holds. The ancestral process $\{\Pi_n^N(\lfloor Nt \rfloor)\}_{t\geq 0}$ converges weakly as $N \to \infty$ to the seed bank n-coalescent $\{\Pi_n(t)\}_{t\geq 0}$ starting with n plants.*

1.3.3 Properties of the seed bank coalescent

Unlike the Kingman coalescent, the seed bank coalescent does not come down from infinity.

We use the following notation: when the system starts with n plants and m seeds, we refer to the block-counting process as $(N_{(n,m)}, M_{(n,m)})$, and we simplify $(N_{(n,0)}, M_{(n,0)})$ to (N_n, M_n).

Theorem 1.3.4. *[9]. For any $(n, m) \in (\mathbb{N} \cup \{0\}) \times (\mathbb{N} \cup \{0\})$ such that $n + m$ is (countably) infinite, then*

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

If the system starts with an infinite number of plants and zero seeds, the number of deactivations is infinite a.s. This implies the presence of an infinite amount of lineages in the seed bank in all cases. To prove this, we can use a coupling where the new model does not see the reactivated individuals. Then if we let $Y_n(t)$ be the number of deactivations up to time $t \geq 0$, this random variable can be bounded by the sum of n independent Bernoulli random variables with respective parameters $cj/(\binom{j}{2} + cj), 1 \leq j \leq n$, the rate of deactivation over the rate of coalescence plus the rate of deactivation. This is then easy to show that for any $t > 0$ and any $k \in \mathbb{N}$,

$$\lim_{n\to\infty} \mathbb{P}(Y_n(t) < k) = 0.$$

This implies, together with other observations that,

$$\mathbb{P}(\forall t > 0 : Y_\infty(t) = \infty) = 1.$$

It means that there have been an infinite amount of movements to the seed bank, a.s. Finally, we can easily check that if we start with an infinite number of seeds, the number of seeds remains infinite for all $t > 0$ a.s.

The next property concerns the time to the most recent common ancestor (TMRCA)

$$\sigma_n = \inf\{t > 0 : N_n(t) + M_n(t) = 1\} = \inf\{t > 0 : N_n(t) = 1,\, M_n(t) = 0\}. \quad (1.19)$$

Theorem 1.3.5. *[9]. For all $c > 0$ and $K > 0$, the expectation of σ_n is bounded as follows.*

$$\mathbb{E}[\sigma_n] \asymp \log\log n,$$

where $\asymp$ means the weak asymptotic equivalence of sequences, that is

$$\liminf_{n\to\infty} \frac{\mathbb{E}[\sigma_n]}{\log\log n} > 0$$

and

$$\limsup_{n\to\infty} \frac{\mathbb{E}[\sigma_n]}{\log\log n} < \infty.$$

The intuition behind this result is that one seed has to become a plant before it is involved in a coalescence event. Thus the TMRCA of a sample of n plants is directed by the number of seeds and the time they take to coalesce. Thanks to the coupling with Bernoulli random variables, we can show that the number of individuals that visit the seed bank before they coalesce is asymptotically of order $\log n$, and as the rate of migration from the seed bank is linear, the time of reactivation of these seeds is of order $\log\log n$.

In Chapter 3 we present the asymptotic behavior of some relevant functionals of the seed bank tree, for example, the behavior of the first time that a plant becomes a seed, the first time that a seed becomes a plant, the number of plants and seeds at these times, and the total branch length of the tree, where we start with n plants and zero seeds.

1.3.4 Forward process and moment duality

As in Section 1.1.1, we suppose that there are two alleles, a and A, in the population and we denote by $\{U^N(k)\}_{k\geq 0}$ the A-allele frequency process of plants, that is, $U^N(k)$ is the number of plants with type A divided by N at generation k. Also, let $\{V^M(k)\}_{k\geq 0}$ be the A-allele frequency process of seeds. The two-dimensional process is a discrete-time Markov chain with values in $B^N \times C^M$, where

$$B^N = \left\{0, \frac{1}{N}, \frac{2}{N}, ..., 1\right\} \quad \text{and} \quad C^M = \left\{0, \frac{1}{M}, \frac{2}{M}, ..., 1\right\}.$$

Proposition 1.3.6. *[9]. Suppose that conditions (1.18) hold. Consider test functions $f \in C^3([0,1]^2)$. Consider a pair of elements $(u_N, v_M) \in B^N \times C^M$ converging to $(u,v) \in [0,1]^2$ Then, the discrete generator $\mathcal{A}^N$ of the frequency process $\{U^N(k), V^M(k)\}_{k\geq 0}$ has the following limit*

$$\lim_{N\to\infty} \mathcal{A}^N f(u_N, v_M) = \mathcal{A}f(u,v),$$

where $\mathcal{A}$ is defined by

$$\mathcal{A}f(u,v) := c(v-u)\frac{\partial f}{\partial u}(u,v) + cK(u-v)\frac{\partial f}{\partial v}(u,v) + \frac{1}{2}u(1-u)\frac{\partial f^2}{\partial u^2}(u,v). \quad (1.20)$$

The forward limit process can be reformulated with a system of SDEs.

Corollary 1.3.7. *[9]. Suppose that conditions of Proposition 1.3.6 hold. If $U^N(0) \to u$ a.s., and $V^M(0) \to v$ a.s., then*

$$\{U^N(\lfloor Nt \rfloor), V^M(\lfloor Nt \rfloor)\}_{t\geq 0} \Rightarrow \{U(t), V(t)\}_{t\geq 0} \quad (1.21)$$

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

$$\begin{aligned} dU(t) &= c(V(t) - U(t))dt + \sqrt{U(t)(1-U(t))}dB(t), \\ dV(t) &= cK(U(t) - V(t))dt, \end{aligned}$$

where B is standard Brownian motion.

Finally, we enunciate a duality result between the block counting process of the seed bank coalescent and the forward frequency process.

Theorem 1.3.8. *[9]. Let $(u,v) \in [0,1]^2$ and $(n,m) \in \mathbb{N} \cup \{0\} \times \mathbb{N} \cup \{0\}$. Denote by $\mathbb{E}_{u,v}$ the law of the forward frequency process given that $(U^N(0), V^N(0)) = (u,v)$, and by $\mathbb{E}_{n,m}$ the law of the block counting process of the seed bank coalescent given that $(N(0), M(0)) = (n,m)$. Then, for every $t \geq 0$*

$$\mathbb{E}_{u,v}[(U(t))^n (V(t))^m] = \mathbb{E}_{n,m}[u^{N(t)} v^{M(t)}]. \quad (1.22)$$

In this case, the duality of Definition 1.1.8 is obtained with the function

$$F(u,v;n,m) := u^n v^m.$$

The proof results from computations on generators, as for Theorem 1.1.9.

1.4 Inference techniques

1.4.1 Models for mutations

We have studied models where the offspring inherit the types of their parents. This section focuses on the population types, which can change through *mutations*.

All mutations are changes in the DNA sequence. Researchers can measure the mutation rate at several scales, for example, mutation across the entire genome (as the rate per genome per generation) or mutation in a gene (as the rate per *locus* per generation). Each gene's particular location on the chromosome is more formally called a genetic locus, [47].

Consider a Wright-Fisher model where individuals can mutate. Let $\hat{\mu}$ be the probability per individual per generation of a mutation for the locus under consideration. Suppose we observe only one ancestral lineage, and X denotes the number of generations until we see a mutation. This has a geometric distribution with parameter $\hat{\mu}$. Then, suppose that $N\hat{\mu} \sim \mu$ for some $\mu > 0$. Rescaling the time, we have for $t > 0$.

$$\mathbb{P}(X \leq \lfloor Nt \rfloor) = 1 - (1 - \hat{\mu})^{\lfloor Nt \rfloor} \longrightarrow 1 - e^{-\mu t} \text{ as } N \to \infty.$$

In the Wright-Fisher model, the limit genealogy of a sample of size n is given by the Kingman n-coalescent, and additionally, we add mutations in the ancestral lineages. Under these hypotheses, the probability of observing a coalescence and a mutation in the sample in a single generation is $O\left(1/N^2\right)$, disappearing in the rescaling.

A different point of view to add mutations in the Kingman coalescent is by throwing down an independent Poisson process of mutation points on each lineage with parameter μ. To ensure that the types in the sample are consistent with the pattern of mutation deriving from such a Poisson process, we must first assign a type to the MRCA. Then we go back through the coalescent tree assigning types to ancestral lineages ([27], Chapter 2, section 2.4) see figure 1.5.

In the next sections, we introduce some important models of mutation.

1.4.2 Infinite alleles model

We consider a haploid population of size N, with *parent independent mutation*. ([27] Chapter 2, section 2.4). In the parent independent mutation model, if an individual has a mutation that occurs at a constant rate per individual independent of the current type, her new type is chosen according to a probability distribution independent of the previous type. In the *infinite alleles model*, every time a mutation occurs, it generates a new, unique allele (type) that has never been seen before in the population. The

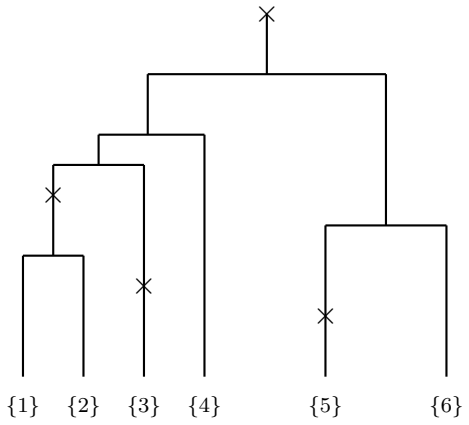

Figure 1.5: Mutations added to a Kingman 5-coalescent by superimposing an independent Poisson process of mutations on each branch. The crosses represent mutations. In this example, the individuals with label 4 and 6 have the same (ancestral) type.

infinite alleles model can be seen as the parent independent mutation model's limit when the number of alleles tends to infinity ([27] Chapter 2, section 2.4).

We are interested in the *allelic partition*, resulting when we identify individuals carrying the same type at the observed gene. The allelic partition is described through a vector called the *allele frequency spectrum*, which describes the number of different alleles with a given multiplicity. The Ewens' sampling formula gives the law of the allele frequency spectrum when the genealogies are given by the Kingman n-coalescent.

In the infinite alleles model, not all the mutations can be observed from a present sample. If two mutations occur in the same lineage, the ancient one is hidden by the new one. The genealogical process resulting is a coalescent with killing (or Kingman n-coalescent with freezing), see figure 1.6. The block-counting process of the Kingman coalescent with killing denoted by $(N_t^K, D_t)_{t\geq 0}$, is a continuous time Markov chain with values in $\mathbb{N}_0 \times \mathbb{N}_0$ and transition rates

$$\text{From } (i,j) \text{ to } \begin{cases} (i-1,j) \text{ at rate } \binom{i}{2} \text{ (coalescent)} \\ (i-1,j+1) \text{ at rate } i\mu \text{ (killing)} \end{cases}$$

The next result provides the entire distribution of the sample following the infinite allele model. To simplify notations, set $\theta = 2\mu$.

Theorem 1.4.1. ***Ewens' sampling formula****, [30]. Let a_i be the number of alleles present i times in a sample of size n. The distribution of the allele frequency spectrum is given by*

$$p(a_1, \ldots, a_n) = \frac{n!}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{j=1}^{n} \frac{\theta^{a_j}}{j^{a_j} a_j!}. \tag{1.23}$$

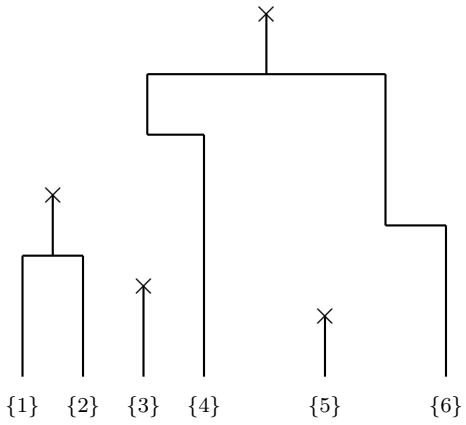

Figure 1.6: The Kingman coalescent with killing resulting from figure 1.5.

Now, in terms of the allelic partitions that the Kingman coalescent with killing generates, we can reformulate the last result.

Theorem 1.4.2. *([2] Chapter 2, section 2.3) Let* Π *be the allelic partition obtained from the Kingman coalescent and the infinite alleles model with mutation rate* $\mu = \theta/2$. *Then* Π *has the law of a Poisson-Dirichlet random partition with parameter* θ. *The allelic partition* Π *is such that his restriction to* $[n]$, *for all* $n \geq 1$, *has the distribution (1.23), which defines a consistent family of partitions as* n *increases.*

We can rewrite (1.23) as

$$p(a_1, \ldots, a_n) = \alpha_{\theta,n} \prod_{j=1}^{n} e^{-\theta/j} \frac{(\theta/j)^{a_j}}{a_j!},$$

where $\alpha_{\theta,n} = \frac{n! \exp\{\sum_{j=1}^{n} \theta/j\}}{\theta(\theta+1)\cdots(\theta+n-1)}$ is a normalization constant that depends on θ and n. Observe that the allele frequency spectrum has the distribution of independent Poisson random variables $W_1, \ldots, W_n$ with respective parameters θ/j, conditioned on the event $\sum_{j=1}^{n} iW_j = n$.

One manner to prove the Ewens' sampling formula is showing that Π can be constructed as a *Chinese Restaurant Process* with parameter θ.

Definition 1.4.3. Chinese Restaurant process ([2] Chapter 1, section 1.3). We consider a random partition process $\{\pi_n\}_n$ such that $\pi_n \in \mathcal{P}_n$ and being constructed by induction. Set $\pi_1 = \{1\}$. We build π_{n+1} from π_n by assigning a block to $(n+1)$. With probability $\dfrac{\theta}{\theta+n}$, $(n+1)$ stands in a new block. With probability $\dfrac{k}{\theta+n}$, $(n+1)$ is assigned to an existing block of size k, with $1 \leq k \leq n$. There is no problem in extending this process to $\pi \in \mathcal{P}$ such that its restriction to $[n]$ is π_n for all $n \geq 1$.

The name of the Chinese Restaurant process comes from the next interpretation, when $\theta = 1$. Consider an empty restaurant with round tables. A first customer arrives at the restaurant and sits by himself. When a new customer arrives, she decides, uniformly at random, between sitting at an empty table or sitting to the right of a customer already present in the restaurant. The partition obtained by this process, where each table represents a block, is that of the Chinese Restaurant process.

The random partition π obtained from the Chinese Restaurant process is a Poisson-Dirichlet random partition with parameter θ. In particular, π is exchangeable.

Returning to the idea of the proof of Theorem 1.4.2, define $0 < t_{n-1} < ... < t_1 < t_0$ to be the times at which we have an event (mutation or coalescence) in the Kingman n-coalescent with killing. Thinking as for the Chinese Restaurant process we label each customer with t_i. At time t_0, we add one lineage (it is the MRCA). During the time interval $(t_1, t_0]$, we have the partition $\pi_1 = \{1\}$. Suppose that π_k is the partition during time $(t_k, t_{k-1}]$. Then, given π_k, at time t_k we have a coalescence or a mutation. If the event was a mutation, $(k+1)$ will open a new block of the partition π_{k+1}. But if the event was a coalescent, $(k+1)$ will join one existing block, see figure 1.7. Suppose that blocks of π_k have size $n_1, ..., n_j$ for some $j \geq 1$ and $\sum_{i=1}^{j} n_i = k$. Observe that between time t_k and t_{k+1}, there are $k+1$ lineages. Then,

$$\mathbb{P}\,(\text{ new block } |\pi_k) = \frac{\frac{\theta(k+1)}{2}}{\binom{k+1}{2} + \frac{\theta(k+1)}{2}} = \frac{\theta}{k+\theta}$$

$$\mathbb{P}\,(k+1 \text{ joins the } j\text{th block } |\pi_k) = \frac{\binom{k+1}{2}}{\binom{k+1}{2} + \frac{\theta(k+1)}{2}} \frac{n_j}{k} = \frac{n_j}{k+\theta}$$

In the last probability, the first term corresponds to one coalescence, and the second is the probability of this coalescence occurs with a block of size n_j.

A large literature in population genetics is dedicated to the generalization of sampling formulas to other models. In Chapter 3, we present a sampling formula related to the seed bank coalescent, adapting the Chinese Restaurant process, with the difference that we start with a random number of tables.

1.4.3 Infinite sites model

Now, suppose that we observe a fixed chromosome. In the *infinite sites model*, every time a mutation occurs in a lineage, it affects a new, never touched before or after, site (locus) of the chromosome ([2] Chapter 2, section 2.3,[27] Chapter 2, section 2.4). Unlike the infinite alleles model, the infinite sites model keeps up variations in chromosome sites. In the first model, the individuals only carry information about the most recent mutation. In contrast, in the second model, they carry information

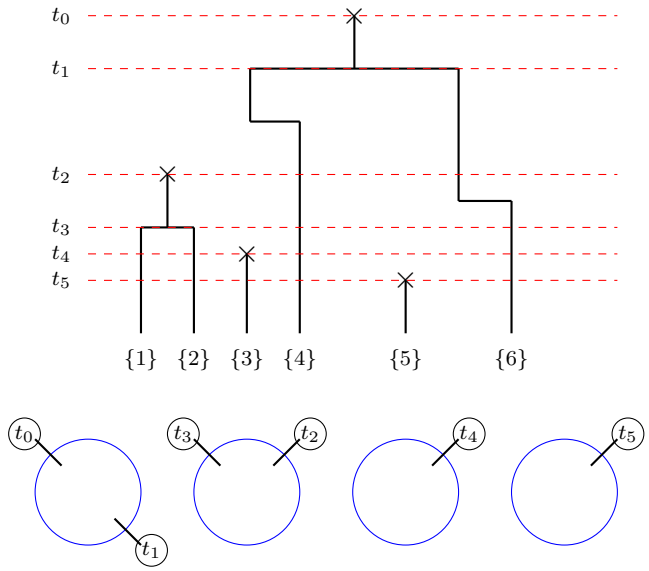

Figure 1.7: Representation of the Kingman coalescent with killing in terms of the Chinese Restaurant process. The blue circles represent the tables and the black circles represent the customers.

about all their ancestors' mutations because these are transmitted unchanged to all offspring and will be visible forever.

We still assume constant mutation rate, and given the coalescent tree, mutations fall on it according to a Poisson process with parameter $\theta/2$ per unit length. In this model, we are interested in the *site frequency spectrum*, SFS, it is a vector $(S_{n,1}, \ldots, S_{n,n-1})$ where $S_{n,i}$ is the number of sites at which exactly i individuals have a mutation.

Theorem 1.4.4. *([2] Chapter 2, section 2.3, [25] Chapter 2, section 2.1) Under the infinite sites model, the expected value of each component of the site frequency spectrum is*

$$\mathbb{E}[S_{n,i}] = \frac{\theta}{i} \tag{1.24}$$

As a straightforward consequence, the total number of sites at which a mutation occurs, S_n (also called the *number of segregating sites*), has the following expectation

$$\mathbb{E}[S_n] = \theta \sum_{i=1}^{n-1} \frac{1}{i}. \tag{1.25}$$

The latter result can be easily obtained by observing that, given the coalescent tree, S_n is a Poisson random variable with parameter $\frac{\theta}{2} L_n$, where L_n is the total length

of the Kingman n-coalescent, that is, the sum of the lengths of all the branches in the tree. The waiting times for a coalescence are independent and exponentially distributed with parameters $\binom{i}{2}, 2 \leq i \leq n$, so

$$L_n = \sum_{i=2}^{n} \frac{i}{\binom{i}{2}} E_i = \sum_{i=2}^{n} \frac{2}{i-1} E_i$$

where the E_i's are i.i.d. standard exponential random variables. Then,

$$\mathbb{E}[S_n] = \mathbb{E}[\mathbb{E}[S_n|L_n]] = \frac{\theta}{2} E[L_n] = \theta \sum_{i=1}^{n-1} \frac{1}{i}.$$

Actually, the asymptotics of L_n can be precised.

Theorem 1.4.5. *([27] Chapter 2, section 2.6) The variable $L_n/2$ is distributed as the maximum of $n-1$ i.i.d. standard exponential random variables. In particular*

$$\lim_{n\to\infty} \frac{L_n}{2} - \log n = Y \tag{1.26}$$

in distribution, where Y has a Gumbel distribution with density $f(y) = \exp\{-y - e^{-y}\}$.

As a consequence, we get the following almost sure convergence for S_n.

$$\lim_{n\to\infty} \frac{S_n}{\log n} = \theta. \tag{1.27}$$

This type of results is very interesting to obtain estimators of θ, since S_n is easily observable.

In the Kingman coalescent case, it is easy to obtain the expected value and the asymptotic of L_n. However, in other coalescents, obtain these results is more complicated. The main goal of Chapter 3 is to present similar results on the length of the seed bank coalescent.

Chapter 2

Seedbank Cannings Graphs: How dormancy alleviates random genetic drift

In the last two decades, Cannings models and their multiple merger genealogies ([63, 60, 66] and more recently [6, 31, 36]) as well as seed bank models based on individual dormancy ([41, 8, 9, 10]) have become significant topics in mathematical population genetics. One of the unifying themes in both modeling areas is that they arise from extensions of the Wright-Fisher model, and that classical population genetics forces such as genetic drift and selection are affected in important ways. While the theory, in particular on the side of seed bank models, is still incomplete, important progress has been made, in particular for Cannings models.

An important tool for the analysis of both models is given by moment duality for Markov processes. This technique establishes a mathematical relation between forward and backward in time processes. The celebrated duality between the Wright-Fisher diffusion and the Kingman colescent was gradually generalized to a wide class of neutral population genetics models, including some finite size discrete populations such as Cannings-type models [34]. In this case, the duality leads to asymptotic results for both forward frequency and genealogical processes. However, for seed bank models the duality tool still has to be set. It was established only in one limiting setting (with infinite population [9]). For finite population size systems (finite N), graphical constructions are highly useful, in particular if they allow the simultaneous construction of forward and backward processes (the most elegant tool here is certainly the look-down construction of Donnelly and Kurtz [20, 21], which even allows for nested approximating particle systems and also convergence results in the Cannings model case - for the seed bank model, look-down constructions are currently being developed but have not been published yet).

For seed bank models, approximating finite size graph-theoretic models have followed

two approaches. For the backward in time model of [41], which is based on the Wright-Fisher model with additional multi-generational jumps of (bounded) size, the system has been extended a) to geometric jump sizes of bounded expected range in [46] (which also provide some insight into the forward in time frequency diffusion), b) to the general finite expectation case in [8], and c) even to unbounded (heavy-tailed) jump sizes in [11]. A second modeling frame is given by an external modelling of the seed bank in terms of a "second island" (in the spirit of Wright's island models), effectively leading to geometric jump sizes on the evolutionary scale of order N (expectation scales with N). Here, forward and backward limits have been constructed, giving rise to the seed bank diffusion and the seed bank coalescent [9] (see more analysis and generalization in [35, 10] and an interesting connection with metapopulations in [49]).

Both modelling frames (generational jumps and second island) have their advantages and disadvantages. For the Wright-Fisher models with multi-generational jumps, one typically loses the Markov property. For the island version, one retains the Markov property, but then needs to investigate two-dimensional frequency processes, which in the limit are harder to analyze than one-dimensional diffusions, since e.g. the Feller theory is missing (this can in part be replaced by recent theory for polynomial diffusions [12]). Interestingly, it turns out that for the limiting frequency processes, both approaches are two sides of the same medal. The two-dimensional seed bank diffusion (corresponding to the island version) can be reformulated as a delay Stochastic Differential Equation (losing the Markov property), which then can be interpreted in terms of the approach of [41], see [12].

In none of the above papers, more general reproductive mechanisms, such as based on Cannings' models, have been analyzed. One motivation for this paper is to close this gap. We present an extended framework for the simultaneous construction of seed bank models with general multi-generational jumps distribution and Cannings-type reproductive laws satisfying a paintbox-construction. We are also able to provide forward and backward convergence results (extending [41], [46] and [8]) and to provide an explicit sampling duality, which is valid already in the finite individual models. Note that the interplay of general reproduction and seed banks with other evolutionary forces can be subtle, and we provide a frame for its analysis (also regarding the real-time embedding of coalescent-based estimates, see e.g. [10]).

In section 2.1 we construct a random graph that allows us to embed the ancestry and the frequency processes of both Cannings and seed bank models simultaneously and study the duality relation of the processes forward and backward in time. Furthermore, we analyze the scaling limits of the ancestral process in presence of skewed reproduction mechanisms and dormancy. We give conditions for convergence to the Kingman coalescent and study scenarios beyond this universality class, where we are able to describe how weak seed bank phenomena reduce the typical size coalescence events, when combining seed banks with Cannings models that would, in absence of the seed bank component, converge to a Λ or a Ξ coalescent. Section 2.2 uses the

moment duality to formally prove convergence of the frequency process to a Wright-Fisher diffusion. This intuitively clear result was missing in the literature, probably because the lack of the Markov property of the frequency process that makes usual techniques fail.

2.1 A random graph version of the model of Kaj, Krone and Lascoux

Consider a discrete-time haploid population of constant size $N \geq 1$ at each generation. The vertex set $V^N = \mathbb{Z} \times [N]$ represents the whole population. We denote the g-th generation by $V_g^N := \{v \in V^N : v = (g, k) \text{ for some } k \in [N]\}$. Set a probability measure $\mathcal{W}^N$ on the probability measures on $[N]$. Let $\{\bar{W}_g^N\}_{g\in\mathbb{Z}}$ be a sequence of independent $\mathcal{W}^N$-distributed random variables with $\bar{W}_g^N = \{W_v^N\}_{v\in V_g^N}$. Each variable W_v^N gives the reproductive weight of the individual v in the population graph. Also consider a sequence $\{m_N\}_{N\geq 1}$ of integers and set a probability measure μ^N on $[m_N]$. Let $\{J_v^N\}_{v\in V^N}$ be a collection of independent μ^N-distributed random variables. The variable J_v^N says how many generations ago individual v's father is living. Finally, set a collection of random variables in $[N]$, $\{U_v^N\}_{v\in V^N}$ such that U_v^N is the label of the father of v. Its conditional distribution is

$$\mathbb{P}(U_{(g,i)}^N = k | J_{(g,i)}^N = j, \{\bar{W}_g^N\}_{g\in\mathbb{Z}}) = W_{(g-j,k)}^N.$$

Definition 1. (The seed bank random di-graph) Consider the random set of directed edges

$$E^N = \{(v, (g - J_v^N, U_v^N)), \text{ for all } v = (g, i) \in V^N\}.$$

The *seed bank random di-graph* with parameters N, $\mathcal{W}^N$ and μ^N is given by $G^N := (V^N, E^N)$.

Two classical examples are

- the Kaj, Krone and Lascoux (KKL) seed bank graph [41], in this case μ^N has finite support $[m]$, i.e. $m_N = m$, and $\mathcal{W}^N = \delta_{(1/N,\dots,1/N)}$.
- the Cannings model with parameter $\mathcal{W}^N$ [14, 15, 60], in this case $\mu^N = \delta_1$.

For every $u, v \in V^N$ we denote by $\delta(u, v)$ the distance of u and v in the graph G^N, i.e. the number of vertices in a path from u to v or from v to u. Now let us define the ancestral process associated with this graph.

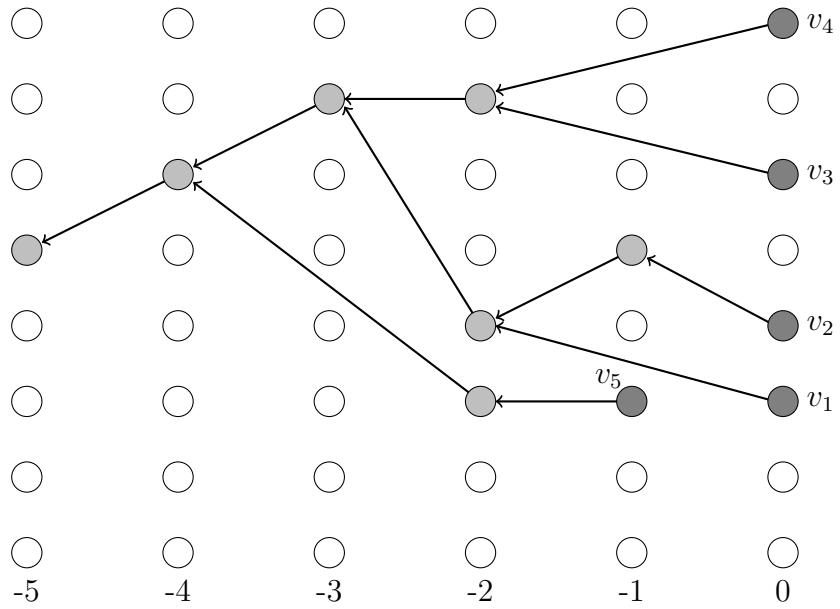

Figure 2.1: In this case $N = 8$ and $m_N = 2$. The gray circles represent the members of $S_0 = \{v_1, v_2, v_3, v_4, v_5\}$ where, for example, $v_2 = (0, 4)$ and $v_5 = (-1, 3)$. The light gray circles represent the ancestors of the sample. $\bar{A}_0^8 = \{4, 1\}$, $\bar{A}_1^8 = \{2, 2\}$, $\bar{A}_2^8 = \{3, 0\}$, $\bar{A}_3^8 = \{1, 1\}$, $\bar{A}_4^8 = \{1, 0\}$, $\bar{A}_5^8 = \{1, 0\}$.

Definition 2 (The ancestral process)**.** Fix a generation g_0 and S_{g_0} consisting in a sample of individuals living between generation g_0 and $g_0 - m_N + 1$, i.e. $S_{g_0} \subset \cup_{i=1}^{m_N} V_{g_0+1-i}^N$. For every $g \geq 0$, let $\mathcal{A}_g^N$ be the set composed by the most recent ancestors of the individuals of S_{g_0} that live at a generation $g_0 - g'$ for some $g' \geq g$, that is

$$\mathcal{A}_g^N = \{v \in \cup_{g'=g}^{\infty} V_{g_0-g'}^N : \exists u \in S_{g_0} \text{ such that } \delta(u, v) \leq \delta(u, v') \text{ for all } v' \in \cup_{g'=g}^{\infty} V_{g_0-g'}^N\}.$$

Define, for all $i \in [m_N]$,

$$A_g^{N,i} = |\mathcal{A}_g^N \cap V_{g_0-g+1-i}^N|$$

and $\bar{A}_g^N = (A_g^{N,1}, \ldots, A_g^{N,m_N})$. We call $\{\bar{A}_g^N\}_{g\geq 0}$ the ancestral process. In the sequel, we consider the initial configuration $S_{g_0}(\bar{n})$, for $\bar{n} = (n_1, \ldots, n_{m_N})$, such that $n_i \geq 0$ individuals are uniformly sampled (with repetition) from generation $g_0 + 1 - i$. We denote the law of the ancestral process of this sample by $\mathbb{P}_{\bar{n}}$. See Figure 2.1 for an illustration.

For simplicity, we suppose that $\sup\{i \geq 1 : n_i > 0\}$ does not depend on N. This model was introduced by Kaj et al. [41] directly, in the sense that they construct a random graph only implicitly. Our construction permits to provide a transparent relation between the ancestral process and the forward frequency process defined in Section 2.2. Observe that $\{\bar{A}_g^N\}_{g\geq 0}$ is a Markov chain.

Proposition 1 (Proof of Theorem 1 in [41]). *Let $M(n)$ be a multinomial random variable with parameters n and $\{\mu^N(i)\}_{i\geq 1}$. Also, for any $\bar{n} = (n_1, \ldots, n_{m_N}) \in [N]^{m_N}$, let $Z(\bar{n}) = (n_2, \ldots, n_{m_N}, 0) + M(n_1)$. Then, the transitions of $\{\bar{A}_g^N\}_{g\geq 0}$ can be written in terms of M and Z as follows.*

- $\mathbb{P}_{\bar{n}}(\bar{A}_1^N = Z(\bar{n})) = 1 - \sum_{i=1}^{\infty} \frac{1}{N}[\binom{n_1}{2}\mu^N(i)^2 + \mu^N(i)n_1n_{i+1}] + o(N^{-2})$
- $\mathbb{P}_{\bar{n}}(\bar{A}_1^N = Z(\bar{n}) - e_i) = \frac{1}{N}[\binom{n_1}{2}\mu^N(i)^2 + \mu^N(i)n_1n_{i+1}] + o(N^{-2})$

where e_i is the vector with the i-th coordinate equal to 1 and the others are equal to 0, for all $i \geq 1$.

Proof. We need to make two observations. First note that all the randomness in the transitions of the chain $\{\bar{A}_g^N\}_{g\geq 0}$ lies in what happens to the first coordinate. If for some $g \geq 0$, $\bar{A}_g^N = (0, n_2, \ldots, n_{m_N})$ it is easy to see that $\bar{A}_{g+1}^N = (n_2, \ldots, n_{m_N}, 0)$ almost surely. On the other hand, if $n_1 > 0$, the individuals that are in $\mathcal{A}_g^N \cap V_{g_0-g}^N$ cannot belong to $\mathcal{A}_{g+1}^N$, and then each of these individuals, if denoted by v, must be replaced by an individual which lives J_v^N generations in the past, that is

$$\mathbb{P}_{e_1}(\bar{A}_1^N = e_i) = \mu^N(i).$$

Further, if $n_1 > 1$, one needs to find n_1 new ancestors, but some of them could be the same due to some coalescence. The complete picture is as follows: for $i \geq 2$ and $j, k \geq 1$, and by denoting e_0 for the null vector,

$$\mathbb{P}_{2e_1+e_i}(\bar{A}_1^N = e_{i-1} + e_j + e_k) = \begin{cases} 2\mu^N(j)\mu^N(k) & \text{if } i-1 \neq j \neq k \\ (\mu^N(j))^2(1-1/N) & \text{if } i-1 \neq j, j = k \\ 2\mu^N(i-1)\mu^N(k)(1-1/N) & \text{if } i-1 = j, j \neq k \\ 2\mu^N(i-1)\mu^N(j)1/N & \text{if } i-1 \neq j, k = 0 \\ 2(\mu^N(i-1))^2 1/N(1-1/N) & \text{if } i-1 = j, k = 0 \\ (\mu^N(i-1))^2 1/N^2 & \text{if } j = k = 0 \end{cases}. \tag{2.1}$$

The proof follows easily after these observations. □

We now construct a less natural backward process which will be very useful when establishing its moment duality with the forward process in Section 2.2. We provide an illustrative example right after the definition.

Definition 3 (The window process). Fix a generation g_0, and $S_{g_0} \subset \cup_{i=1}^{m_N} V_{g_0+1-i}^N$. For $g \geq 1$, consider the equivalence relation on S_{g_0}, that we denote by $\sim_g$, such that $u \sim_g v$ if and only if they have a common ancestor at a generation between $g_0 - g + 1$ and g_0. Let $\pi_0 = \pi_1$ be the trivial partition made of the isolated elements of S_{g_0} (singletons) and let π_g be the partition induced by $\sim_g$ in the sample S_{g_0}. Let $\mathcal{B}_g^N$

be the set composed by the closest ancestors, living at a generation $g_0 - g'$ for some $g' \geq g$, of each of the blocks in π_g. Then, for $1 \leq i \leq m_N$, we define

$$B_g^{N,i} := |\mathcal{B}_g^N \cap V_{g_0-g+1-i}^N|$$

and $\bar{B}_g^N := (B_g^{N,1}, \ldots, B_g^{N,m_N})$. We call $\{\bar{B}_g^N\}_{g\geq 0}$ the window process. As for the ancestral process, we denote by $\mathbb{P}_{\bar{n}}$ the law of the window process generated from the initial sample $S_{g_0}(\bar{n})$.

Example 2.1.1. We illustrate the definition of window process by the realization pictured in figure 2.1. In this case $\pi_0 = \pi_1 = \pi_2 = \{\{v_1\}, \{v_2\}, \{v_3\}, \{v_4\}, \{v_5\}\}$. Observe that even if some individuals reach their common ancestor at generation -2, they remain isolated in π_2. Then $\pi_3 = \{\{v_1, v_2\}, \{v_3, v_4\}, \{v_5\}\}$, $\pi_4 = \{\{v_1, v_2, v_3, v_4\}, \{v_5\}\}$, $\pi_5 = \{\{v_1, v_2, v_3, v_4, v_5\}\}$. Hence, $\mathcal{B}_0^8 = \{v_1, v_2, v_3, v_4, v_5\}$ and, when moving some generations backwards, we get $\mathcal{B}_1^8 = \{v_5, (-1, 5), (-2, 4), (-2, 7), (-2, 7)\}$ and $\mathcal{B}_2^8 = \{(-2, 3), (-2, 4), (-2, 4), (-2, 7), (-2, 7)\}$. Observe that in $\mathcal{B}_2^8$ the ancestors $(-2, 4)$ and $(-2, 7)$ appear twice.
Also $\mathcal{B}_3^8 = \{(-3, 7), (-3, 7), (-4, 6)\}$, $\mathcal{B}_4^8 = \{(-4, 6), (-4, 6)\}$, $\mathcal{B}_5^8 = \{(-5, 5)\}$. Finally the values of the window process are $\bar{B}_0^8 = \{4, 1\}$, $\bar{B}_1^8 = \{2, 3\}$, $\bar{B}_2^8 = \{5, 0\}$, $\bar{B}_3^8 = \{2, 1\}$, $\bar{B}_4^8 = \{2, 0\}$, $\bar{B}_5^8 = \{1, 0\}$.

A more intuitive and graphical interpretation is the following: in the genealogical tree of the sample S_{g_0}, the variable $B_g^{N,1}$ gives the number of edges having an extremity at generation $g_0 - g$ (plus the number of individuals of S_{g_0} living at this generation). For other values of i, $B_g^{N,i}$ is the number of edges crossing generation $g_0 - g$ and having an extremity at generation $g_0 - g - i + 1$ (plus the number of individuals of S_{g_0} living at this generation). The window process and the ancestral process only differ in the time where we acknowledge a coalescence event. In the window process coalescence events only occur in the first coordinate, while in the ancestral process coalescence events may take place at every entry (see Figure 2.1).

The following equivalent (in law) definition of the window process allows us to compare it with the ancestral process. Let $C^N(n)$ be the number of ancestors after one generation of a sample of n individuals in a Cannings model with weights distributed as $\mathcal{W}^N$. As in Proposition 1, let $M(n)$ be a multinomial random variable with parameters n and $\{\mu^N(i)\}_{i\geq 1}$. Given $\bar{B}_{g-1}^N = \bar{n} = (n_1, \ldots, n_{m_N}) \in [N]^{m_N}$,

$$\bar{B}_g^N = (n_2, \ldots, n_{m_N}, 0) + M(C^N(n_1)).$$

in distribution. It is left to the reader to show that indeed both definitions are equivalent.

The process $\{\bar{B}_g^N\}_{g\geq 0}$ can be expressed in terms of a particle system. Fix N, μ^N and $\mathcal{W}^N$. Let $Y_g^N = (R_g^N, L_g^N)$ define a Markov chain with state space $\mathbb{N} \times [N]$ and transition probabilities, conditional on the weights $\bar{W}_g^N$,

$$\mathbb{P}\big((R_g^N, L_g^N) = (i, k) | \{\bar{W}_g^N\}_g, (R_{g-1}^N, L_{g-1}^N) = (1, j)\big) = W_{(g_0+1-g-i,k)}^N \mu^N(i)$$

for every $i \geq 1$, , $k \in [N]$, and

$$\mathbb{P}\big((R_g^N, L_g^N) = (i,k) | \{\bar{W}_g^N\}, (R_{g-1}^N, L_{g-1}^N) = (i+1, j)\big) = W^N_{(g_0-g,k)}$$

for every $i \geq 1$ and $k \in [N]$.

Proposition 2. *Set $n = \sum n_i$ to be the total size of the initial sample. For every $g \geq 0$, consider n independent realizations of Y_g^N, that we call $Y_g^{N,j} = (R_g^{N,j}, L_g^{N,j})$ for $1 \leq j \leq n$. Let $\sigma^{N,1} = \infty$ and*

$$\sigma^{N,j} = \inf \Big\{ g \geq 1 : Y_g^{N,j} = Y_g^{N,j'} = (1,k), \text{ for some } j' < j \text{ such that } \sigma^{N,j'} > g, \\ k \in [N] \Big\}.$$

For all $i \geq 1$, set $\sum_{j=1}^n 1_{\{R_0^{N,j}=i\}} = B_0^{N,i}$. Then, the i-th component $B_g^{N,i}$ of the random vector $\bar{B}_g^N$ is equal in distribution to $\sum_{j=1}^n 1_{\{R_g^{N,j}=i\}} 1_{\{\sigma^{N,j}>g\}}$, for all $g \geq 0$.

Proof. The proof consists in observing that $g_0 - R_g^{N,j} - g + 1.$ is equal in distribution to the generation of the most recent ancestor, living at a generation $g_0 - g'$ for some $g' \geq g$, of a fixed individual in the initial sample S_{g_0}. So we couple these two processes. At the particular times in which $R_g^{N,j} = 1$ (and thus a coalescence event can occur in the window process) we take $L_g^{N,j}$ to be the label of the closest ancestor. Then $\sigma^{N,j}$ corresponds to the generation at which individual j's ancestral lineage is involved into a coalescence event with the ancestral lineage of an individual of lower level. Under this coupling,

$$B_g^{N,i} = \sum_{j=1}^n 1_{\{R_g^{N,j}=i\}} 1_{\{\sigma^{N,j}>g\}}$$

almost surely. □

The chain $\{R_g^N, L_g^N\}_{g\geq 0}$ provides a very convenient coupling to the ancestral and the window processes, mainly because $\{R_g^N\}_{g\geq 0}$ has an invariant measure given by

$$\nu^N(i) = \frac{\mathbb{P}(J_v^N \geq i)}{\mathbb{E}[J_v^N]}.$$

To see this, just observe that the chain has two types of behaviours. Using the notation $\mathbb{P}_j(\cdot) = \mathbb{P}(\cdot | R_0^N = j)$, we have

(i) Deterministic transitions: if $j > 1$, then $\mathbb{P}_j(R_1^N = j-1) = 1$

(ii) Random transitions: for $j \geq 1$, $\mathbb{P}_1(R_1^N = j) = \mathbb{P}(J_v^N = j) = \mu^N(j)$.

Then,

$$\begin{aligned}\sum_{j=1}^{\infty}\mathbb{P}_j(R_1^N=i)\nu^N(j) &= \mathbb{P}_{i+1}(R_1^N=i)\nu^N(i+1)+\mathbb{P}_1(R_1^N=i)\nu^N(1)\\ &= \frac{\mathbb{P}(J_v^N\geq i+1)}{\mathbb{E}[J_v^N]}+\frac{\mathbb{P}(J_v^N=i)}{\mathbb{E}[J_v^N]}\\ &= \nu^N(i).\end{aligned}$$

The techniques that we will use to compare two Markov chains, in this case the rescaled window process and another chain which block-counting process converges to this of some coalescent, consist in using coupling concepts developed in [52]. Let us first recall the definition of mixing time (see page 55 of [52]). We denote $d^N(g)=\max_{j\in[m_N]}||\mathbb{P}_j(R_g^N\in\cdot)-\nu^N(\cdot)||_{TV}$ and the mixing time $\tau_N=\inf\{g>0: d^N(g)<1/4\}$.

The main theorem of [41] (proved for μ^N with finite support and extended to finite expectation in [8]) consists in showing that the L_1 norm of the ancestral process converges weakly to the block counting process of the Kingman coalescent under a constant time change. Here we extend this result to the window process and to some more general Cannings' mechanism.

Theorem 1 (Convergence of the window process I: Kingman limit)**.** Fix $\{\mu^N\}_{N\geq1}$ such that $\beta_N:=\mathbb{E}[J_v^N]<\infty$ and fix the distribution $\mathcal{W}^N$ on the N-dimensional vectors that sum to 1. Let τ_N be the mixing time of $\{R_g^N\}_{g\geq0}$, $c_N:=N\mathbb{E}[(W_{(1,1)}^N)^2]$ and $d_N:=N\mathbb{E}[(W_{(1,1)}^N)^3]$. Assume that $\mu^N(1)>0$ and that

$$c_N/\beta_N^2\to0,\quad N^{\varepsilon}\tau_N c_N\to0,\quad (1/4)^{N^{\varepsilon}}\beta_N^2\to0\quad\text{and}\quad d_N/(\beta_N c_N)\to0.$$

for some $\varepsilon>0$. Then, let $\{\bar{B}^N\}_{N\geq1}$ be the sequence of window processes with parameters N and μ^N and starting condition $\bar{B}_0^N=\bar{n}$ for all $N\in\mathbb{N}$ big enough. Then,

$$\{|\bar{B}^N_{\lfloor t\beta_N^2/c_N\rfloor}|\}_{t\geq0}\Rightarrow\{N_t^K\}_{t\geq0}\tag{2.2}$$

as $N\to\infty$, where $\{N_t^K\}_{t\geq0}$ stands for the block counting process of a Kingman coalescent.

Furthermore, suppose that ν^N converges to a measure ν as $N\to\infty$. Let $V^{t,K}$ be a (conditional) multinomial random variable with parameters N_t^K and ν. For any fixed time $t>0$, in distribution,

$$\lim_{N\to\infty}\bar{B}^N_{\lfloor t\beta_N^2/c_N\rfloor}=V^{t,K}.\tag{2.3}$$

Note that when $\beta_N \to \beta < \infty$ the third condition of Theorem 1 is automatically fulfilled. On the other side, when $\beta_N \to \infty$, then the fourth condition is always fulfilled because $d_N/c_N \leq 1$. The latter reflects the fact that a strong seed bank effect makes impossible the existence of multiple merges. The next two results discuss the interplay between weak seed banks and random genetic drift. Note also that the second condition implies that $\tau_N < \infty$ for all but finitely many N, meaning that the support of μ^N is finite for all but at much finitely many N.

Denote by Δ the infinite simplex on $[0,1]$. Let $F_\beta : \Delta \mapsto \Delta$ be such that for $A \in \Delta$, $F_\beta(A) = \{\bar{y}/\beta, \bar{y} \in A\}$. To any finite measure Ξ over Δ, we associate a finite measure Ξ^β defined by the rule

$$\Xi^\beta(F_\beta(A)) = \Xi(A)$$

for any borelian A on Δ. Observe that Ξ^β associates no weight on mass partitions $\bar{y} = (y_1, y_2, \dots)$ such that $\sum y_i > 1/\beta$.

Theorem 2 (Convergence of the window process II: Ξ limit)**.** Fix $\{\mu^N\}_{N\in\mathbb{N}}$ such that $\beta_N = \mathbb{E}[J_v^N] < \infty$ and fix the distribution $\mathcal{W}^N$. Assume that the ancestral process of a Cannings model driven by $\mathcal{W}^N$, that we denote by $\{C_g^N\}_{g\geq 0}$ is such that, as $N \to \infty$,

$$\{C^N_{\lfloor t/c_N \rfloor}\}_{t\geq 0} \Rightarrow \{N_t^\Xi\}_{t\geq 0}$$

where $\{N_t^\Xi\}_{t\geq 0}$ stands for the block counting process of a Ξ-coalescent. If $\beta_N \to \beta < \infty$, then

$$\{|\bar{B}^N_{\lfloor t/c_N \rfloor}|\}_{t\geq 0} \Rightarrow \{N_t^{\Xi^\beta}\}_{t\geq 0}. \tag{2.4}$$

Furthermore. Suppose that ν^N converges to a measure ν as $N \to \infty$. Let V^{t,Ξ^β} be a multinomial random variable with parameters ν and $N_t^{\Xi^\beta}$. For any fixed time $t > 0$, in distribution,

$$\lim_{N\to\infty} B^N_{\lfloor t/c_N \rfloor} = V^{t,\Xi^\beta}. \tag{2.5}$$

It is interesting that the seed bank effect sends the class of Λ-coalescents into itself.

Proof of Theorem 1. The proof consists in coupling the window process $\{\bar{B}_g^N\}_{g\geq 0}$ to a process which is "always in stationarity". Recall that the variables $Y_g^{N,j} = (R_g^{N,j}, L_g^{N,j})$ define the process that models the distance between g and the level of the ancestor of the j-th block induced by $\sim_g$.

If we suppose that $\{\bar{B}_g^N\}_{g\geq 0}$ starts a.s. with one lineage, i.e. $\bar{B}_0^N = e_k$ for some k, it is easy to verify that it has a stationary distribution $\bar{\nu}^N$ given by $\bar{\nu}^N(e_i) = \nu^N(i)$. Now, let $\underline{Y}_g^{N,j} = (\underline{R}_g^{N,j}, L_g^{N,j})$ where $\{\underline{R}_g^{N,j}\}_{g\geq 0}$ is a sequence of independent ν^N-distributed random variables. Let

$$\underline{\sigma}^{N,j} = \inf\Big\{ g \geq 1 : \underline{Y}_g^{N,j} = \underline{Y}_g^{N,j'} = (1,k), \text{ for some } j' < j \text{ such that } \underline{\sigma}^{N,j'} > g, \\ k \in [N]\Big\}.$$

Hence, we define an artificial window process by

$$\bar{Z}_g^N = (Z_g^{N,1}, \dots, Z_g^{N,m_N})$$

where, as $n = \sum n_i$,

$$Z_g^{N,i} = \sum_{j=1}^{n} 1_{\{\underline{R}_g^{N,j}=i\}} 1_{\{\underline{\sigma}^{N,j}\geq g\}}.$$

The process $\{|\bar{Z}_g^N|\}_{g\geq 0}$ is Markovian.

We now proceed in two steps to prove (2.2). First, we calculate the generator of $\{|\bar{Z}_g^N|\}_{g\geq 0}$ in order to discover its scaling limit. Let $f : \mathbb{N} \to \mathbb{R}$ be a bounded function. Then

$$\begin{aligned}\mathcal{G}^N f(n) &= \mathbb{E}[f(|\bar{Z}_1^N|) - f(n)] = \mathbb{P}(|\bar{Z}_1^N| = n-1)[f(n-1) - f(n)] \\ &\quad + O(\mathbb{P}(|\bar{Z}_1^N| = n-2)) \\ &= \binom{n}{2} \frac{c_N}{\beta_N^2}[f(n-1) - f(n)] + O(\frac{d_N}{\beta_N^3}).\end{aligned}$$

So we conclude that

$$\{|\bar{Z}^N_{\lfloor \beta_N^2 t/c_N \rfloor}|\}_{t\geq 0} \Rightarrow \{N_t^K\}_{t\geq 0}. \tag{2.6}$$

Second, let us couple $\{|\bar{Z}^N_{\lfloor \beta_N^2 t/c_N \rfloor}|\}_{t\geq 0}$ and $\{|\bar{B}^N_{\lfloor \beta_N^2 t/c_N \rfloor}|\}_{t\geq 0}$ to show that the same limit is true for the rescaled window process. The coupling consists in constructing for every $i \geq 1$ the random variable $(\underline{R}_{\rho_i}^{N,1}, \dots, \underline{R}_{\rho_i}^{N,m_N})$ as the optimal coupling of $(R_{\rho_i}^{N,1}, ..., R_{\rho_i}^{N,m_N})$ and the stationary distribution $(\nu^N)^{\otimes m_N}$, where the times $\{\rho_i\}_{i\geq 1}$ correspond to the times where the processes $\{|\bar{Z}^N_{\lfloor \beta_N^2 t/c_N \rfloor}|\}_{t\geq 0}$ and $\{|\bar{B}^N_{\lfloor t/c_N \rfloor}|\}_{t\geq 0}$ can jump. More precisely, if we denote for any $p, q \in [n]$, $\rho_k^{N,p,q} = \inf\{g > \rho_{k-1}^{N,p,q} : L_g^{N,p} = L_g^{N,q}\}$ (with $\rho_0^{N,p,q} = 0$), then $\rho_i = \inf\{g > \rho_{i-1} : g = \rho_k^{N,p,q}$ for some $p, q \in [n]$ and some $k \in \mathbb{N}\}$ (with $\rho_0 = 0$). Note that we do not precise the dependence on N in the notation. In our case, the probability that the coupling is successful

$$\begin{aligned} p_N &:= \inf_{\bar{n}\in[N]^{m_N}} \mathbb{P}_{\bar{n}}((\underline{R}_{\rho_1}^{N,1}, \dots, \underline{R}_{\rho_1}^{N,m_N}) = (R_{\rho_1}^{N,1}, \dots, R_{\rho_1}^{N,m_N})) \\ &= 1 - \sup_{\bar{n}\in[N]^{m_N}} \mathbb{P}_{\bar{n}}((\underline{R}_{\rho_1}^{N,1}, \dots, \underline{R}_{\rho_1}^{N,m_N}) \neq (R_{\rho_1}^{N,1}, \dots, R_{\rho_1}^{N,m_N})) \\ &= 1 - \sup_{\bar{n}\in[N]^{m_N}} \|\mathbb{P}_{\bar{n}}((R_{\rho_1}^{N,1}, \dots, R_{\rho_1}^{N,m_N}) = \cdot) - (\nu^N)^{\otimes m_N}(\cdot)\|_{TV}\end{aligned}$$

where $\mathbb{P}_{\bar{n}}$ stands for the law of $\{R_g^{N,1}, \dots, R_g^{N,m_N}\}_{g\geq 0}$ (or $\{\underline{R}_g^{N,1}, \dots, \underline{R}_g^{N,m_N}\}_{g\geq 0}$) starting at the state $\bar{n} \in [N]^{m_N}$ and where Proposition 4.7 in [52] is used for the last equality. To prove that $p_N \to 1$ when $N \to \infty$, take $\varepsilon > 0$ such that $N^\varepsilon \tau_N c_N \to 0$. The condition $\mu^N(1) > 0$ implies that, for any $i \geq 1$, the processes $\{R_g^{N,i}\}_{g\geq 0}$ are irreducible. So, by Theorem 4.9 in [52], we have

$$||\mathbb{P}_{\bar{n}}((R^{N,1}_{N^{\varepsilon}\tau_N},\dots,R^{N,m_N}_{N^{\varepsilon}\tau_N})=\cdot)-(\nu^N)^{\otimes m_N}(\cdot)||_{TV}<(1/4)^{N^{\varepsilon}}. \tag{2.7}$$

Then observe that, stochastically, $\rho_1 \geq \Gamma^N$ where Γ^N is a geometric random variable with parameter $n^2 c_N \geq \binom{n}{2} c_N$ and thus $\mathbb{P}(\rho_1 < N^{\varepsilon}\tau_N) \leq \mathbb{P}(\Gamma^N < N^{\varepsilon}\tau_N) \to 0$.

Let $T_1^N = \inf\{i \geq 1 : |\bar{Z}^N_{\rho_i}| = 1\}$. Observe that if $|\bar{Z}^N_0| = n$ (we will use the notation $\mathbb{P}_n$), stochastically

$$T_1^N \leq \sum_{i=1}^{n} G_i^N$$

where the G_i^N's are independent geometric random variables with parameter β_N^{-2}. We finish the proof noting that the trajectories of both processes are identical with overwhelming probability

$$\begin{aligned}
\mathbb{P}_n(\sup_t ||\bar{Z}^N_{\lfloor t\beta_N^2/c_N\rfloor}| - |\bar{B}^N_{\lfloor t\beta_N^2/c_N\rfloor}|| = 0) &= \mathbb{E}[p_N^{T_1^N}] && (2.8)\\
&\geq \sum_{i=1}^{\infty}(1-\frac{1}{\beta_N^2})^{i-1}\frac{1}{\beta_N^2}p_N^{ni} && (2.9)\\
&= \frac{p_N^n}{\beta_N^2}\frac{1}{1-(1-\frac{1}{\beta_N^2})p_N^n} \to 1 && (2.10)
\end{aligned}$$

which gives (2.2).

Finally, let us prove (2.3). Let $t > 0$ fixed, and suppose that ν^N converges to a measure ν. By, equation (2.6) we have that $\lim_{N\to\infty} |\bar{Z}^N_{\lfloor t\beta_N^2/c_N\rfloor}| = N_t^K$. Then, observe that $\bar{Z}_g^N$ has a multinomial distribution with parameters $|\bar{Z}_g^N|$ and ν^N. Thus, in distribution,

$$\lim_{N\to\infty} \bar{Z}^N_{\lfloor t\beta_N^2/c_N\rfloor} = V^{t,K}. \tag{2.11}$$

On the other hand, by (2.2), we have that

$$\lim_{N\to\infty} |\bar{B}^N_{\lfloor t\beta_N^2/c_N\rfloor}| = N_t^K.$$

For $t > 0$ fixed, let us couple $\bar{Z}^N_{\lfloor t\beta_N^2/c_N\rfloor}$ and $\bar{B}^N_{\lfloor t\beta_N^2/c_N\rfloor}$ to show that the limit (2.11) is the same for $\bar{B}^N_{\lfloor t\beta_N^2/c_N\rfloor}$. As we did before, the coupling consists in constructing the random variable $(\underline{R}^{N,1}_{\lfloor t\beta_N^2/c_N\rfloor},\dots,\underline{R}^{N,m_N}_{\lfloor t\beta_N^2/c_N\rfloor})$ as the optimal coupling of $(R^{N,1}_{\lfloor t\beta_N^2/c_N\rfloor},\dots,R^{N,m_N}_{\lfloor t\beta_N^2/c_N\rfloor})$. The probability that the coupling is successful

$$\begin{aligned}
\varrho_N &:= \inf_{\bar{n}\in[N]^{m_N}} \mathbb{P}_{\bar{n}}((\underline{R}^{N,1}_{\lfloor t\beta_N^2/c_N\rfloor},\dots,\underline{R}^{N,m_N}_{\lfloor t\beta_N^2/c_N\rfloor}) = (R^{N,1}_{\lfloor t\beta_N^2/c_N\rfloor},\dots,R^{N,m_N}_{\lfloor t\beta_N^2/c_N\rfloor}))\\
&= 1 - \sup_{\bar{n}\in[N]^{m_N}} \mathbb{P}_{\bar{n}}((\underline{R}^{N,1}_{\lfloor t\beta_N^2/c_N\rfloor},\dots,\underline{R}^{N,m_N}_{\lfloor t\beta_N^2/c_N\rfloor}) \neq (R^{N,1}_{\lfloor t\beta_N^2/c_N\rfloor},\dots,R^{N,m_N}_{\lfloor t\beta_N^2/c_N\rfloor}))\\
&= 1 - \sup_{\bar{n}\in[N]^{m_N}} \|\mathbb{P}_{\bar{n}}((R^{N,1}_{\lfloor t\beta_N^2/c_N\rfloor},\dots,R^{N,m_N}_{\lfloor t\beta_N^2/c_N\rfloor}) = \cdot) - (\nu^N)^{\otimes m_N}(\cdot)\|_{TV}
\end{aligned}$$

Take $\varepsilon > 0$ such that $N^\varepsilon \tau_N c_N \to 0$, and observe that $\beta_N > 0$. This implies that

$$\mathbb{P}(\lfloor t\beta_N^2/c_N \rfloor < N^\varepsilon \tau_N) \to 0 \text{ as } N \to \infty. \tag{2.12}$$

By, (2.7) and (2.12), we have that $\varrho_N \longrightarrow 1$. This gives (2.3).

□

Remark 1. Consider two processes, $\{\underline{R}_g^N\}_{g\geq 0}$ and $\{R_g^N\}_{g\geq 0}$, the first one starting with one particle in stationarity and the second one starting with one particle in state one. If we consider their Doebling coupling (which consists in letting them evolve according to their respective laws and in merging their paths when they meet for the first time, see [52], Section 5), their coupling time, T^N, is less than two with probability

$$\nu^N(1) + \sum_{i=1}^{m_N-1} \mu^N(i)\nu^N(i+1) \leq \nu^N(1) + \frac{1}{\beta_N}\sum_{i=1}^{m_N-1} \mu^N(i) = \frac{1}{\beta_N}(2 - \mu^N(m_N)).$$

As the process $\{R_g^N\}_{g\geq 0}$ visits the state one approximately every β_N steps we conclude that $\mathbb{P}(T^N > N^\varepsilon \beta_N^2) \to 0$ when $N \to \infty$. Since $\tau_N \leq \inf\{t \geq 0; \mathbb{P}(T^N > t) < 1/4\}$, we obtain that $\tau_N \leq \max(N^\varepsilon \beta_N^2, m_N)$ and that hypotheses of Theorem 1 can be relaxed to the following

$$c_N/\beta_n^2 \to 0, \quad N^\varepsilon \max(N^\varepsilon \beta_N^2, m_N)c_N \to 0, \quad (1/4)^{N^\varepsilon}\beta_N^2 \to 0 \quad \text{and} \quad d_N/(\beta_N c_N) \to 0$$

with the advantage that they are easier to verify.

Proof of Theorem 2. The proof of (2.4) is similar to that of (2.2) in Theorem 1. In the present case, let I_i denote the indicator of the event that $L_1^{N,i} = L_1^{N,j}$ for some $j \in [i-1]$. Note that $\{C^N_{\lfloor t/c_N \rfloor}\}_{t\geq 0}$ has generator

$$\mathcal{C}^N f(n) = c_N^{-1}\mathbb{E}[f(n - \sum_{i=1}^{n} I_i) - f(n)]$$

which by hypothesis converges to the generator of the block counting process of a Ξ-coalescent. Finally note that, using the same notation, the generator of the artificial (in stationarity) block counting process $\{\bar{Z}^N_{\lfloor t/c_N \rfloor}\}_{t\geq 0}$ is

$$\mathcal{C}^N f(n) = c_N^{-1}\mathbb{E}[f(n - \sum_{i=1}^{n} I_i 1_{\{\underline{R}_1^{N,i}=1\}}) - f(n)].$$

As $1_{\{\underline{R}_1^{N,i}=1\}}$ is a Bernoulli random variable with parameter tending to β^{-1} and independent of I_i, we conclude that

$$\{|\bar{B}^N_{\lfloor t/c_N \rfloor}|\}_{t\geq 0} \Rightarrow \{N_t^{\Xi^\beta}\}_{t\geq 0}.$$

The rest of the proof is identical to Theorem 1. □

2.2 The forward frequency process

In this section we introduce the forward frequency process associated to the weak seed bank graph, we establish duality results with the ancestral and window processes introduced in the previous section, and we establish some scaling limits results thanks to these tools.

Definition 4 (The frequency process)**.** Fix a generation g_0 and an initial sample $S_{g_0} \subset \cup_{i=1}^{m_N} V_{g_0+1-i}^N$, that we call the type A individuals. Hence, $\cup_{i=1}^{m_N} V_{g_0+1-i}^N \backslash S_{g_0}$ is the set of type a individuals. For $g \geq 0$, set (omitting again the dependence to S_{g_0})

$$X_g^{N,i} = \frac{1}{N} |\{v \in V_{g_0+g+1-i}^N : v \text{ is not connected to } u \text{ for some } u \in \cup_{i=1}^{m_N} V_{g_0+1-i}^N \backslash S_{g_0}\}|.$$

Then, define the process of the neutral frequency of type A individuals $\{\bar{X}_g^N\}_{g\geq 0}$, by

$$\bar{X}_g^N = (X_g^{N,1}, \ldots, X_g^{N,m_N}).$$

Set a vector $\bar{x} = (x_1, \ldots, x_{m_N}) \in ([N]/N)^{m_N}$. In the sequel, we suppose that the forward frequency process starts from a fraction x_1 of generation 0, a fraction x_2 of generation -1, and so on... We denote this sample by $S_0(\bar{x}) = \cup_{i=1}^{m_N} \cup_{k=1}^{x_i N} \{(1-i,k)\}$ and we denote the law of the frequency process starting from this configuration by $\mathbf{P}_{\bar{x}}$.

Again for simplicity, we suppose that $\sup\{i \geq 1 : x_i > 0\}$ does not depend on N.

Proposition 3. *Fix the parameters N, μ^N and $\mathcal{W}^N$ of the seed bank di-graph. The processes $\{\bar{X}_g^N\}_{g\geq 0}$ and $\{\bar{A}_g^N\}_{g\geq 0}$ are sampling duals: for every $g \geq 0$, we have $\mathbf{E}_{\bar{x}}[h^0(\bar{n}, \bar{X}_g^N)] = \mathbb{E}_{\bar{n}}[h^0(\bar{A}_g^N, \bar{x})]$ where $h^0(\bar{n}, \bar{x}) := \mathbb{P}_{\bar{n}}(\mathcal{A}_1^N \subset S_0(\bar{x}))$.*

Proof. Suppose that the ancestral process starts at generation $g+1$ from the sample $S_{g+1}(\bar{n})$, as in Definition 2. Also suppose that the frequency process starts at generation 0 from the sample $S_0(\bar{x})$, as in Definition 4. Introduce the functions

$$h^g(\bar{n}, \bar{x}) := \mathbb{P}_{\bar{n}}(\mathcal{A}_{g+1}^N \subset S_0(\bar{x})). \tag{2.13}$$

We can write $h^g(\bar{n}, \bar{x})$ in terms of the forward process by conditioning as follows.

$$\begin{aligned} h^g(\bar{n}, \bar{x}) &= \sum_{\bar{y} \in ([N]/N)^{m_N}} h^0(\bar{n}, \bar{y}) \mathbf{P}_{\bar{x}}(\bar{X}_g^N = \bar{y}) \\ &= \mathbf{E}_{\bar{x}}[h^0(\bar{n}, \bar{X}_g^N)]. \end{aligned}$$

At this point it should be clear that we can also condition according to the backward process.

$$\begin{aligned} h^g(\bar{n}, \bar{x}) &= \sum_{\bar{m} \in [N]^{m_N}} h^0(\bar{m}, \bar{x}) \mathbb{P}_{\bar{n}}(\bar{A}_g^N = \bar{m}) \\ &= \mathbb{E}_{\bar{n}}[h^0(\bar{A}_g^N, \bar{x})]. \end{aligned}$$

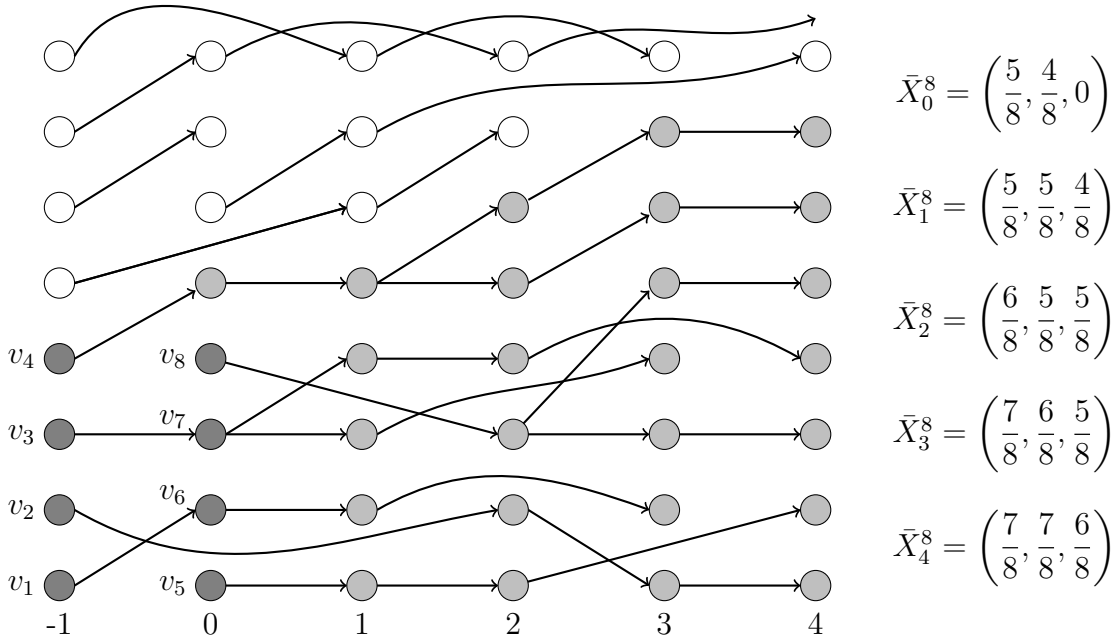

Figure 2.2: In this case $N = 8$, $m_N = 3$ and $\bar{x} = \left(\frac{4}{8}, \frac{4}{8}, 0\right)$. The gray circles represent the members of $S_0(\bar{x}) = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$ where, for example, $v_3 = (-1, 3)$ and $v_6 = (0, 2)$. The light gray circles represent the sample's offspring. It is useful to observe that $X_g^{N,i} = X_{g+1}^{N,i+1}$.

This implies that for all $\bar{x} \in ([N]/N)^{m_N}$, $\bar{n} \in [N]^{m_N}$ and $g \geq 1$,

$$\mathbb{E}_{\bar{n}}[h^0(\bar{A}_g^N, \bar{x})] = \mathbf{E}_{\bar{x}}[h^0(\bar{n}, \bar{X}_g^N)].$$

□

The sampling duality provides a relation between the forward and the ancestral processes. However, in this case the relation is not a moment duality. It is possible to write explicitly this sampling duality and to use it, but we will rather use the less natural window process which has the advantage of being precisely the moment dual of the forward process.

Proposition 4. *Fix the parameters N, μ^N and $\mathcal{W}^N$ of the seed bank di-graph. The window process $\{\bar{B}_g^N\}_{g\geq 0}$ and the forward frequency process $\{\bar{X}_g^N\}_{g\geq 0}$ are moment duals.*

Proof. We will construct a sampling duality that is exactly moment duality i.e duality with respect to the function $H : \mathbb{N}^{m_N} \times [0,1]^{m_N} \mapsto [0,1]$,

$$H(\bar{n}, \bar{x}) = \prod_{i=1}^{m_N} x_i^{n_i} \tag{2.14}$$

Fix $\bar{n} \in [N]^{m_N}$ and $\bar{x} \in ([N]/N)^{m_N}$, and set the samples $S_0(\bar{n})$ and $S_0(\bar{x})$ as in Definition 2 and Definition 4 (with $g_0 = 0$). Observe that $S_0(\bar{n}) = \{v = (1-i, U_{j,i}), i = 1, \ldots, m_N, j = 1, \ldots, n_i\}$ where the $U_{j,i}$'s form a family of independent uniformly distributed random variables with values in $[N]$. Then, we have

$$\tilde{h}(\bar{n}, \bar{x}) := \mathbb{P}(S_0(\bar{n}) \subset S_0(\bar{x})) = \prod_{i=1}^{m_N} \prod_{j=1}^{n_i} \mathbb{P}((1-i, U_{j,i}) \in S_0(\bar{x})) = \prod_{i=1}^{m_N} \prod_{j=1}^{n_i} x_i = H(\bar{n}, \bar{x}).$$

Now we prove sampling duality with respect to this function. As in the proof of Proposition 3, condition on $\bar{X}_g^N$ to obtain that $\mathbb{P}_{\bar{n}}(\mathcal{B}_g^N \in S_0(\bar{x})) = \mathbf{E}_{\bar{x}}[\tilde{h}(\bar{n}, \bar{X}_g^N)]$ and condition on $\bar{B}_g^N$ to obtain that $\mathbb{P}_{\bar{n}}(\mathcal{B}_g^N \in S_0(\bar{x})) = \mathbb{E}_{\bar{n}}[\tilde{h}(\bar{B}_g^N, \bar{x})]$. □

Now we are able to state an analogue of Theorem 1 for the dual process, using the moment duality.

Theorem 3 (Convergence of the forward frequency process)**.** Assume that $m_N \leq m < \infty$ for all $N \in \mathbb{N}$. Fix $\{\mathcal{W}^N\}_{N\geq 1}$ and $\{\mu^N\}_{N\geq 1}$ (and the associated stationary distribution ν^N) such that either the assumptions of Theorem 1 hold or the assumptions of Theorem 2 hold. Suppose that ν^N converges to a measure ν on $[m]$ as $N \to \infty$. Let $\{\bar{X}^N\}_{N\geq 1}$ be the sequence of frequency processes with parameters N, $\mathcal{W}^N$ and μ^N and starting condition $\bar{X}_0^N = (\lfloor Nx_1 \rfloor/N, \ldots, \lfloor Nx_m \rfloor/N)$ for some

$\bar{x} \in [0,1]^m$. Then,
i) Under that assumptions of Theorem 1 hold,

$$\{\bar{X}^N_{\lfloor t\beta_N^2/c_N \rfloor}\}_{t\geq 0} \Rightarrow \{\bar{X}_t\}_{t\geq 0}$$

where $\bar{X}_t$ is a vector with m identical coordinates X_t such that $X_0 = x_0 = \sum_{i=1}^m \nu(i)x_i$ a.s., and $\{X_t\}_{t\geq 0}$ is the Wright-Fisher diffusion (dual of $\{N_t^K\}_{t\geq 0}$).
ii) Under the assumptions of Theorem 2,

$$\{\bar{X}^N_{\lfloor t/c_N \rfloor}\}_{t\geq 0} \Rightarrow \{\bar{X}_t\}_{t\geq 0}$$

where $\bar{X}_t$ is a vector with m identical coordinates X_t such that $X_0 = x_0 = \sum_{i=1}^m \nu(i)x_i$ a.s., and $\{X_t\}_{t\geq 0}$ is moment dual of $\{N_t^{\Xi^\beta}\}_{t\geq 0}$.

Remark 2. The assumption of finite support for $\{\mu^N\}_{N\geq 1}$ seems to be more than a technical assumption. It is hard to believe that an asymptotically infinite dimensional sequence of processes would converge to an infinite dimensional processes with all entries being equal. A natural question is, what is the limit in this more general scenario?

Proof. We only write the details for case *i)*, case *ii)* follows identically. The proof is a consequence of Proposition 4, Theorem 1 and the moment problem. Let us abuse the notation and write $\bar{X}_0^N = \bar{x}$ for every N.

First, let us clarify the role of x_0. Recall that the process $\{X_t\}_{t\geq 0}$ is a martingale. In particular, its expectation remains constant. We claim that for every $i \in [m]$, $\lim_{N\to\infty} \mathbf{E}_{\bar{x}}[X^{N,i}_{\lfloor t\beta_N^2/c_N \rfloor}] = x_0$. To see this we use duality and convergence to stationarity of a single dual particle.

$$\lim_{N\to\infty} \mathbf{E}_{\bar{x}}[X^{N,i}_{\lfloor t\beta_N^2/c_N \rfloor}] = \lim_{N\to\infty} \mathbb{E}_{e_i}[\prod_{j=1}^m x_j^{B^{N,j}_{\lfloor t\tau_N\beta_N^2/(\tau_N c_N)\rfloor}}] = \sum_{i=1}^m x_i\nu(i) = x_0. \tag{2.15}$$

The first equality comes from duality. For the second equality, use the two first assumptions of Theorem 1 (resp. Theorem 2) to see that $\beta_N^2/(\tau_N c_N) \to \infty$ and thus that the process is in stationarity in the limit. The third equality follows from the fact that there is only one positive entry of the unitary vector $\bar{B}^N_{\lfloor t\tau_N\beta_N^2/(\tau_N c_N)\rfloor}$ and that the position of the entry with the one is ν-distributed in the limit.

Now let us study the limiting behavior of one coordinate. Let $n \geq 1$.

$$\begin{aligned}
\lim_{N\to\infty} \mathbf{E}_{\bar{x}}[(X^{N,1}_{\lfloor t\beta_N^2/c_N \rfloor})^n] &= \lim_{N\to\infty} \mathbf{E}_{\bar{x}}[H(n.e_1, \bar{X}^N_{\lfloor t\beta_N^2/c_N \rfloor})] \\
&= \lim_{N\to\infty} \mathbb{E}_{n.e_1}[H(\bar{B}^N_{\lfloor t\beta_N^2/c_N \rfloor}, \bar{x})] \\
&= \mathbb{E}_{n.e_1}[H(V^{t,K}, \bar{x})] \\
&= \mathbb{E}_n[x_0^{N_t^K}] \\
&= \mathbf{E}_{x_0}[X_t^n].
\end{aligned}$$

The third equality follows from (2.3) and in the fourth equality we used the same argument as for (2.15). This proves that all the moments of $X^{N,1}_{\lfloor t\beta_N^2/c_N\rfloor}$ converge to the moments of the Wright-Fisher diffusion.

Finally, we check that in the limit all the coordinates of $\bar{X}^N_{\lfloor t\beta_N^2/c_N\rfloor}$ must take the same value. To do this we will calculate the square of the difference of two arbitrary coordinates. Let $i, j \in [m]$. With te same arguments as before,

$$\begin{aligned}
\lim_{N\to\infty} \mathbf{E}_{\bar{x}}[(X^{N,i}_{\lfloor t\beta_N^2/c_N\rfloor} - X^{N,j}_{\lfloor t\beta_N^2/c_N\rfloor})^2] &= \lim_{N\to\infty} \Big[\mathbf{E}_{\bar{x}}[(X^{N,i}_{\lfloor t\beta_N^2/c_N\rfloor})^2] \\
&\quad +\mathbf{E}_{\bar{x}}[(X^{N,j}_{\lfloor t\beta_N^2/c_N\rfloor})^2] \\
&\quad -2\mathbf{E}_{\bar{x}}[X^{N,i}_{\lfloor t\beta_N^2/c_N\rfloor} X^{N,j}_{\lfloor t\beta_N^2/c_N\rfloor}]\Big] \\
&= \lim_{N\to\infty} \Big[\mathbf{E}_{\bar{x}}[H(2e_i, \bar{X}^N_{\lfloor t\beta_N^2/c_N\rfloor})] \\
&\quad +\mathbf{E}_{\bar{x}}[(H(2e_j, \bar{X}^N_{\lfloor t\beta_N^2/c_N\rfloor})] \\
&\quad -2\mathbf{E}_{\bar{x}}[H(e_i+e_j, \bar{X}^N_{\lfloor t\beta_N^2/c_N\rfloor})]\Big] \\
&= \lim_{N\to\infty} \Big[\mathbb{E}_{2e_i}[H(\bar{B}^N_{\lfloor t\beta_N^2/c_N\rfloor}, \bar{x})] \\
&\quad +\mathbb{E}_{2e_j}[(H(\bar{B}^N_{\lfloor t\beta_N^2/c_N\rfloor}, \bar{x})] \\
&\quad -2\mathbb{E}_{e_i+e_j}[H(\bar{B}^N_{\lfloor t\beta_N^2/c_N\rfloor}, \bar{x})]\Big] \\
&= \mathbb{E}_{2e_i}[H(V^{t,K}, \bar{x})] + \mathbb{E}_{2e_j}[(H(V^{t,K}, \bar{x})] \\
&\quad -2\mathbb{E}_{e_i+e_j}[H(\bar{V}^{t,K}, \bar{x})] \\
&= 0.
\end{aligned}$$

This ends the proof.

□

Chapter 3

The shape of a seed bank tree

In Section 1.3 we introduced the definition of the strong seed bank model and the associated seed bank coalescent. In this chapter we study the asymptotic behavior of some functionals of the seed bank tree, which shed light on connections between theoretical and applied population genetics.

Recall the definition of the model introduced in Definition 1.3.1. Consider a haploid population of fixed size N. Assume that the population additionally supports a seed bank of constant size M. The N *active* individuals are called *plants* and the M *dormant* individuals are called *seeds*. Let $0 \leq \varepsilon \leq 1$ such that $\lfloor \varepsilon N \rfloor \leq M$ and let $\delta := \varepsilon N/M$. The N plants from generation 0 produce $N - \lfloor \varepsilon N \rfloor$ plants by multinomial sampling (as in the Wright-Fisher model) and $\lfloor \varepsilon N \rfloor$ seeds in generation 1. Then, $\lfloor \delta M \rfloor = \lfloor \varepsilon N \rfloor$ uniformly (without replacement) sampled seeds from the seed bank in generation 0 become plants in generation 1. Thus, generation 1 is again made of N plants and M seeds, see Figure 3.1. This random mechanism is then to be repeated independently to produce the next generations.

As we mentioned in Section 1.3, the stochastic process that describes the limiting genealogy of a sample taken from the strong seed bank model is called the *seed bank n-coalescent.*

The seed bank coalescent is a structured coalescent with an active part, having the dynamics of a *Kingman coalescent*, and a dormant part where the lineages are frozen. Lineages can activate or deactivate at certain rates, see Figure 3.2 for an illustration.

As an illustration of the connections between theoretical and applied population genetics, there is a close relation between the shape of the tree of a sample of size n and the number of mutations observed in it. More precisely, suppose that mutations appear in the genealogy by simply superimposing a Poisson process on the ancestral lineages (as it is in the infinite sites model, see Chapter 1.4 in [25]). Then, the shape of the tree determines the distribution of the data obtained by DNA sequencing and

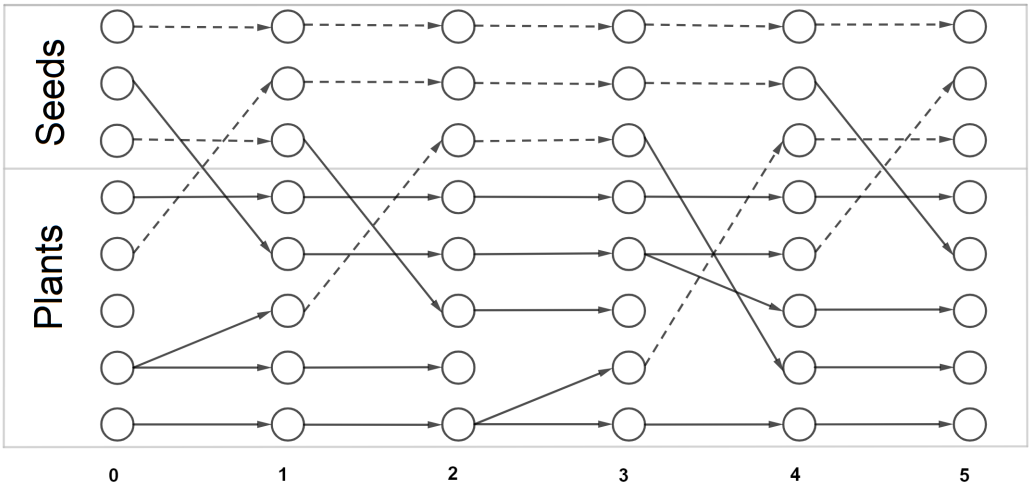

Figure 3.1: The discrete seed bank model. In this picture $N = 5$, $M = 3$ and $\lfloor \varepsilon N \rfloor = 1$, i.e., in each generation four plants are produced by active individuals, one seed germinates, and one plant creates one inactive individual.

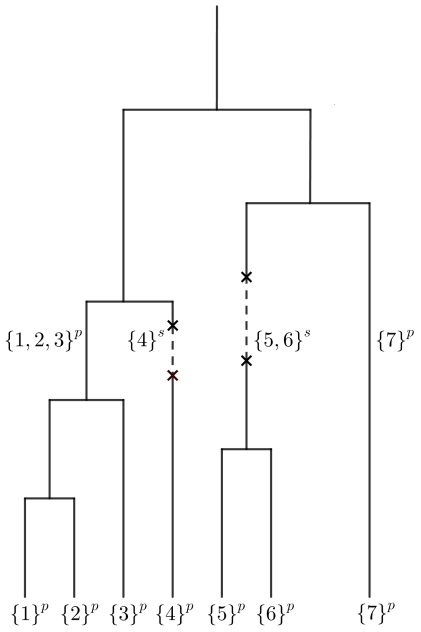

Figure 3.2: A possible realization of the seed-bank 7-coalescent. Dotted lines indicate inactive individuals and the crosses mean that there is a deactivation or a reactivation.

thus, it can be inferred from it. For example, conditionally on the total length of the Kingman coalescent, denoted by L_n, the number of mutations observed in the sample has Poisson distribution with parameter μL_n, where μ is the mutation rate. Thus, if we know the asymptotic behavior of the total length of the tree we can deduce the asymptotic behavior of the number of mutations. This is the key tool for obtaining a Watterson-type estimator for the mutation rate, see [25]. Not surprisingly, asymptotics of the total length of many classical coalescents have been studied, e.g. in [24, 17, 42, 19].

In [9], it was established that the time to the most recent common ancestor of a sample of size n in the seed bank coalescent is of order $\log\log n$. This is an important difference with the classical Kingman coalescent, whose height is finite. Our study establishes that the total length of the tree built from a sample of n plants and zero seeds is of the same order as that of the Kingman coalescent, behaving like $\log n$, but with a different multiplicative constant depending on the activation and deactivation parameters of the model. Moreover, we show that the total active length behaves like the total length of the Kingman coalescent. This means that it is not possible to distinguish between the null Kingman model and the alternative seed bank model using only the tree length unless the dormant individuals have the possibility to mutate while being in the seed bank. This is actually the case in the metapopulation framework described in [49]. This conclusion agrees with the main result in [54] where Maughan observed experimentally that a population of bacteria undergoing dormancy typically does not have significantly different number of mutations. Furthermore, our results offer new insights on the reason for this: most of the mutations occur in the *Kingman phase* (shortly before the leaves) and in this part of the ancestral tree, dormancy is irrelevant. On the other hand, populations suffering a significant amount of mutations while being in the dormant state would be expected to have a higher evolutionary rate. This remark together with [54] suggests that the mutations that occur to individuals in latent state are atypical. This is opposed to previous works suggesting that the normal rate of molecular evolution of bacteria with a seed bank is evidence that mutations affecting dormant individuals are frequent [54].

More theoretical and experimental work is needed to clarify the role of dormancy in the flow of evolution. Finer results, such as sampling formulas, can be derived to discriminate between both null and seed bank models. For the time being, we are able to describe the seed bank tree in detail as it undergoes different phases. It can be said that we describe the shape of the seed bank tree.

3.1 Main results

We study some relevant stopping times of the seed bank coalescent, leading to a complete description of the shape of the tree and explaining how long the genealogies

spend in successive dynamical phases, as is detailed precisely in Table 3.1 and Figure 3.3.

In definition 1.3.2 we introduced the seed bank n-coalescent. In this chapter, we are going to work, mainly, with the block-counting process of the seed bank n-coalescent that we mention again.

The block-counting process of the seed bank n-coalescent is the two-dimensional Markov chain $(N_n(t), M_n(t))_{t\geq 0}$ with values in $([n]\cup\{0\})\times([n]\cup\{0\})$ and the following transition rates, for $t \geq 0$.

$$(N_n(t), M_n(t)) \text{ jumps from } (i,j) \text{ to } \begin{cases} (i-1, j), & \text{at rate } \binom{i}{2} \quad \text{(coalescence)}, \\ (i-1, j+1), & \text{at rate } c_1 i \quad \text{(deactivation)}, \\ (i+1, j-1), & \text{at rate } c_2 j \quad \text{(activation)}. \end{cases}$$

In the sequel, we suppose that $N_n(0) = n$ and $M_n(0) = 0$.

For $i \in [n]$, we denote by τ_n^i the hitting time of the level i by the process N_n, i.e. $\tau_n^n = 0$ and

$$\tau_n^i = \inf\{t \geq 0 : N_n(t) = i\}. \tag{3.1}$$

Furthermore, let γ_n and θ_n be, respectively, the first time that some plant becomes a seed and the first time that some seed becomes a plant, i.e.

$$\gamma_n = \inf\{t > 0 : M_n(t-) < M_n(t)\} = \inf\{t > 0 : M_n(t) = 1\} \tag{3.2}$$

and

$$\theta_n = \inf\{t > 0 : M_n(t-) > M_n(t)\}. \tag{3.3}$$

Finally, denote by σ_n the time to the most recent common ancestor, already studied in [9],

$$\sigma_n = \inf\{t > 0 : N_n(t) + M_n(t) = 1\} = \inf\{t > 0 : N_n(t) = 1, M_n(t) = 0\}.$$

We first obtain asymptotic results on the random variables γ_n and θ_n and the size of the system at those times. The results obtained in Sections 3.2 and 3.3 can be summarized in Table 3.1 and Figure 3.3.

Observe that the rate of coalescence is quadratic with respect to the number of plants while the rate of deactivation (resp. the rate of activation) is linear with respect to the number of plants (resp. the number of seeds). From [9], we inferred that the number of seeds accumulated until time θ_n, $M_n(\theta_n)$, is of order $\log n$. Lemma 3.3.4 suggests that, until time $\tau_n^{\lfloor (\log n)^a \rfloor}$, for $a > 1/2$, the block-counting process $(N_n(t))_{t\geq 0}$ behaves similarly to that of the Kingman coalescent. However, at time $\tau_n^{\lfloor \sqrt{\log n} \rfloor}$, the system reaches a level of $\sqrt{\log n}$ plants and the times of decay are no longer close to those of the Kingman coalescent. Indeed, at this time, we claim that the number of

Stopping time (τ)	Asymptotics of τ	Asymptotics of $N_n(\tau)$	Asymptotics of $M_n(\tau)$
γ_n	$2(1-Y)/Yn$	Yn	1
θ_n	$T/\log n$	$Z\log n$	$2c_1\log n$
σ_n	$\log\log n$	1	0

Table 3.1: Summary of the asymptotic behavior of the functionals of the seed bank coalescent studied in this work. Here Y is a $Beta(2c_1, 1)$ distributed random variable, T is an exponential random variable with parameter $2c_1c_2$ and Z is a Fréchet random variable with shape parameter 1 and scale parameter $4c_1c_2$.

seeds is still of order $\log n$ and the coalescence events do not dominate any more the dynamics. The seed bank coalescent then enters into a mixed regime with coalescence and activation occurring at the same velocity.

In Section 4 we analyze the total length

$$L_n = A_n + I_n \tag{3.4}$$

where the *active length* is defined by

$$A_n = \int_0^{\sigma_n} N_n(t)dt \tag{3.5}$$

and the *inactive length* by

$$I_n = \int_0^{\sigma_n} M_n(t)dt. \tag{3.6}$$

Our main result is stated as follows.

Theorem 3.1.1. *Consider the seed bank coalescent starting with n plants and no seeds. Then,*

$$\lim_{n\to\infty} \frac{L_n}{\log n} = 2\left(1 + \frac{c_1}{c_2}\right)$$

in probability.

Interestingly, numerical techniques of [39] used to study the total length for fixed n show that the balance between active and inactive lengths is equally conserved for their expectations for any $n \geq 2$,

$$c_1\mathbb{E}[A_n] = c_2\mathbb{E}[I_n].$$

The behavior of both A_n and I_n is obtained by considering those variables before and after the time of the first activation θ_n. Hence, results of Section 3.3 are key tools for the forthcoming proofs. Theorem 3.1.1 also gives an immediate corollary on the number of active and inactive mutations on the seed bank tree.

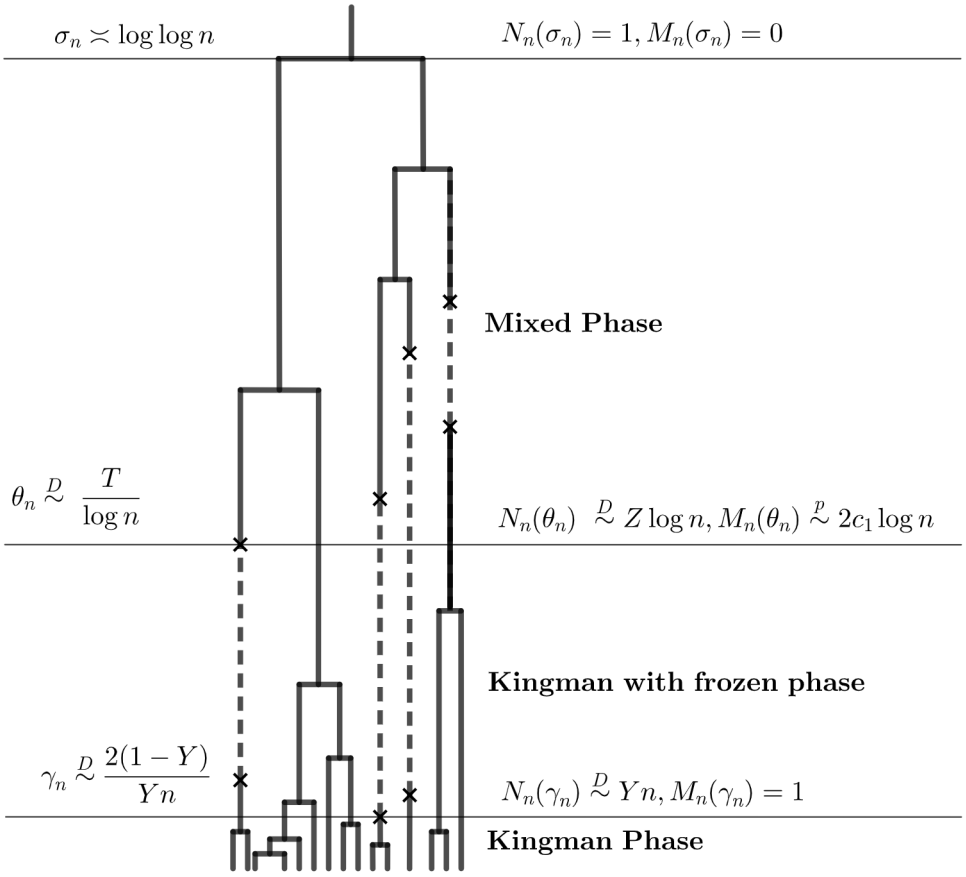

Figure 3.3: Summary of the asymptotic behavior of the functionals of the seed bank coalescent studied in this work. Here Y is a $Beta(2c_1, 1)$ distributed random variable, T is an exponential random variable with parameter $2c_1c_2$ and Z is a Fréchet random variable with shape parameter 1 and scale parameter $4c_1c_2$. The symbol $A_n \overset{p}{\sim} B_n$ means that $\frac{A_n}{B_n} \to 1$ in probability. The symbol $A_n \overset{\mathcal{D}}{\sim} XB_n$ means that $\frac{A_n}{B_n} \to X$ in distribution. The symbol $A_n \asymp B_n$ means that $C_1B_n \leq \mathbb{E}[A_n] \leq C_2B_n$ for some constants C_1, C_2.

Corollary 3.1.2. *Consider the seed bank coalescent starting with n plants and no seeds. Let S_n be the number of mutations in the seed bank tree. Let μ be the mutation rate for the active individuals and let κ be the mutation rate for the inactive individuals. Then*

$$\lim_{n\to\infty} \frac{S_n}{\log n} = 2\left(\mu + \kappa\frac{c_1}{c_2}\right)$$

in probability.

Finally, in Section 3.5, we establish a sampling formula inspired by Watterson's ideas

in [70], which helps us to understand the fine configuration of the blocks of a seed bank coalescent at given times.

3.2 The time of the first deactivation

We start with the study of γ_n, the time of the first deactivation defined in (3.2), and the size of the system at this time. Observe that, if $N_n(0) = n$ and $M_n(0) = 0$, there are $n - N_n(\gamma_n) - 1$ coalescence events until time γ_n and we can write

$$\gamma_n = \sum_{i=N_n(\gamma_n)+1}^{n} V_i$$

where the V_i's are independent exponential random variables with respective parameters $\binom{i}{2} + c_1 i$.

We start with an easy limit result on the variable $N_n(\gamma_n)$. Note that, in a classical Kingman coalescent with mutations appearing at rate c_1, the quantity $n - N_n(\gamma_n) - 1$ can also be interpreted as the number of coalescence events before the most recent mutation in the genealogy. Recent studies on the shape of coalescent trees at the time of the first mutation in a branch can be found in [33], with some direct applications to coalescent model selection [32].

Proposition 3.2.1. *Consider a seed bank coalescent starting with n plants and no seeds. Then,*

$$\lim_{n\to\infty} \frac{N_n(\gamma_n)}{n} = Y$$

in distribution, where $Y \sim Beta(2c_1, 1)$.

Proof. Let $z \in (0, 1)$. We have that

$$\begin{aligned} \mathbb{P}(N_n(\gamma_n) \leq zn) &= \prod_{i=\lfloor zn \rfloor + 1}^{n} \frac{\binom{i}{2}}{\binom{i}{2} + c_1 i} = \prod_{i=\lfloor zn \rfloor}^{n-1} \frac{i}{i + 2c_1} \\ &= \exp\left\{ - \sum_{i=\lfloor zn \rfloor}^{n-1} \log\left(1 + \frac{2c_1}{i}\right) \right\}. \end{aligned}$$

Using that $\log(1+x) \sim x$ near 0, we obtain

$$\begin{aligned}\mathbb{P}(N_n(\gamma_n) \leq zn) &\sim \exp\left\{-\sum_{i=\lfloor zn \rfloor}^{n-1} \frac{2c_1}{i}\right\} \\ &\sim \exp\left\{-2c_1 \log\left(\frac{1}{z}\right)\right\} \\ &= z^{2c_1}\end{aligned}$$

which is the distribution function of a $Beta(2c_1, 1)$ random variable. □

Now, let us establish the asymptotic behavior of the time of the first deactivation, γ_n.

Proposition 3.2.2. *Consider a seed bank coalescent starting with n plants and no seeds. Then,*

$$\lim_{n\to\infty} n\gamma_n = \Gamma := \frac{2(1-Y)}{Y} \tag{3.7}$$

in distribution, where Y is $Beta(2c_1, 1)$ distributed. The density function of Γ is

$$f_\Gamma(x) = c_1 \left(\frac{2}{2+x}\right)^{2c_1+1}$$

for $x \geq 0$. In particular, if $c_1 > 1/2$, then the expectation of Γ is finite

$$\mathbb{E}[\Gamma] = \frac{2}{2c_1 - 1}$$

and if $c_1 > 1$, the variance of Γ is finite

$$\mathrm{Var}(\Gamma) = \frac{4c_1}{(c_1-1)(2c_1-1)^2}.$$

Proof. Let $G_n(0) = 0$ and, for $t \in (0,1)$, define

$$G_n(t) = \sum_{i=\lfloor (1-t)n \rfloor + 1}^{n} V_i = \sum_{i=\lfloor (1-t)n \rfloor + 1}^{n} \frac{2e_i}{i(i-1+2c_1)},$$

where the e_i's are i.i.d standard exponential random variables. With this notation, we obtain $\gamma_n = G_n\left(1 - N_n(\gamma_n)/n\right)$.

We first show that, for any $t \in (0,1)$, we have

$$\lim_{n\to\infty} \left(nG_n(s)\right)_{s\leq t} = \left(\frac{2s}{1-s}\right)_{s\leq t} \tag{3.8}$$

in distribution, in the sense of weak convergence in the path space D[0,t]. To this aim, let us first establish that, for a fixed $t \in (0,1)$,

$$\lim_{n\to\infty} nG_n(t) = \frac{2t}{1-t} \tag{3.9}$$

in L^2. By definition, we have that

$$\begin{aligned}\mathbb{E}[nG_n(t)] &= \sum_{i=\lfloor(1-t)n\rfloor+1}^{n} \frac{2n}{i(i-1+2c_1)} \\ &\sim \frac{1}{n}\sum_{i=\lfloor(1-t)n\rfloor+1}^{n} \frac{2}{(i/n)^2}.\end{aligned}$$

By a Riemann sum argument, we obtain that

$$\mathbb{E}[nG_n(t)] \sim \int_{1-t}^{1} \frac{2}{x^2}dx = \frac{2t}{1-t}.$$

Now, by the independence of the random variables e_i,

$$\begin{aligned}\mathrm{Var}(nG_n(t)) &= \sum_{i=\lfloor(1-t)n\rfloor+1}^{n} \frac{4n^2}{i^2(i-1+2c_1)^2} \\ &\sim \sum_{i=\lfloor(1-t)n\rfloor+1}^{n} \frac{4n^2}{i^4}.\end{aligned}$$

Again, by a Riemann sum argument, we obtain that $\mathrm{Var}(nG_n(t))$ converges to 0 as $n \to \infty$. This gives (3.9).

To obtain (3.8) we follow the same steps as those of Proposition 6.1 in [18], with $\alpha = 2$. Then, the proof of (3.7) follows by adapting the alternative proof of Theorem 5.2 in [18], p. 1713, taking $\alpha = 2$ and the limit variable σ being $1-Y$ and $Beta(1, 2c_1)$ distributed.

The distribution function of Γ is given by

$$\begin{aligned}\mathbb{P}(\Gamma \le x) &= \mathbb{P}\left(Y \ge \frac{2}{2+x}\right) \\ &= 1 - \left(\frac{2}{2+x}\right)^{2c_1}\end{aligned}$$

for $x \ge 0$. We get the density by differentiating. The moments of Γ are obtained by computing

$$\mathbb{E}[\Gamma^k] = \int_0^\infty kx^{k-1}\mathbb{P}(\Gamma > x)dx = \int_0^\infty kx^{k-1}\left(\frac{2}{2+x}\right)^{2c_1} dx.$$

In particular, the kth moment is finite for $c_1 > k/2$. □

3.3 The time of the first activation

In this section we study θ_n, the first time that a seed becomes a plant, which we introduced in (3.3). We also provide some limit laws for $N_n(\theta_n)$ and $M_n(\theta_n)$. Observe that from time zero up to time θ_n only two types of events occur, either coalescence or deactivation. Recall the successive hitting times of the chain N_n, denoted by $(\tau_n^i)_{i=1}^n$ and defined in (3.1).

Proposition 3.3.1. *Consider a seed bank coalescent starting with n plants and no seeds. Then, the following asymptotics hold.*

$$\lim_{n\to\infty} \frac{N_n(\theta_n)}{\log n} = Z \tag{3.10}$$

in distribution, where Z is a Fréchet random variable with shape parameter 1 and scale parameter $4c_1c_2$, with distribution function $\mathbb{P}(Z \leq z) = \exp\{-4c_1c_2/z\}$. Also

$$\lim_{n\to\infty} \frac{M_n(\theta_n)}{\log n} = 2c_1 \tag{3.11}$$

in probability. Finally,

$$\lim_{n\to\infty} \log n\theta_n = T \tag{3.12}$$

in distribution, where T is an exponential random variable with parameter $2c_1c_2$.

The proof of (3.11) is obtained by combining Lemmas 3.3.2 and 3.3.5. The proof of (3.10) and (3.12) is obtained by combining Lemmas 3.3.3 and 3.3.6 which appear in the sequel. We get these results by coupling the seed bank coalescent with two simpler models.

The *coloured* seed bank coalescent (see Definition 4.2 in [9]) is a marked coalescent where additionally each element of $[n]$ has a flag indicating its color: white or blue. Movements and mergers of the blocks of the colored coalescent follow the same dynamics as those of the classical seed bank coalescent. Additionally, if a block activates, each individual inside this block gets the color blue. In other cases colors remain unchanged.

As in [9], we start with all individuals colored with white, so color blue only appears after a reactivation event, and we also use the notation $\underline{\mathrm{N}}_n(t)$ (resp. $\underline{\mathrm{M}}_n(t)$) for the number of white plants (resp. white seeds) at time t, starting with n (white) plants and zero seeds.

The notation for the reaching times of $\underline{\mathrm{N}}_n$ are $\underline{\tau}_n^n = 0$ and, for $i \in [n-1]$,

$$\underline{\tau}_n^i = \inf\{t > 0 : \underline{\mathrm{N}}_n(t) = i\}.$$

Note that, on the event $\{\tau_n^i < \theta_n\}$, we have $\underline{\tau}_n^i = \tau_n^i$ a.s., and in general the stochastic bound

$$\underline{\tau}_n^{i-1} - \underline{\tau}_n^i \leq_{st} \tau_n^{i-1} - \tau_n^i \tag{3.13}$$

holds.

This model is of particular use to prove that the number of seeds that "survive" up to moment θ_n is of order $\log n$. More precisely, as in [9], consider the independent Bernoulli random variables $B_n^i = \mathbf{1}_{\{\text{deactivation at } \underline{\tau}_n^i\}}$, for $i \in [n-1]$, with respective parameter

$$\begin{aligned} \mathbb{P}(B_n^i = 1) &= \frac{c_1(i+1)}{\binom{i+1}{2} + c_1(i+1)} \\ &= \frac{2c_1}{i + 2c_1}, \end{aligned} \tag{3.14}$$

independently of the number of seeds in the system. It is clear that, almost surely for any $t \geq 0$, $M_n(t) \leq \sum_{i=1}^{n-1} B_n^i$. This and Bienaymé-Chebyshev's inequality lead to the following straightforward result.

Lemma 3.3.2. *For any $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{t\geq 0} M_n(t) > 2c_1(1+\varepsilon)\log n\right) \leq \frac{1}{2c_1\varepsilon^2 \log n}. \tag{3.15}$$

In particular, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(M_n(\theta_n) \leq 2c_1(1+\varepsilon)\log n\right) = 1.$$

From now on, denote the upper bound $m_n := \lfloor 2c_1(1+\varepsilon)\log n \rfloor$ for $\varepsilon > 0$. We will use this notation in the following proofs.

The *bounded* seed bank coalescent is a modification of the original seed bank coalescent, where only m seeds can be accumulated in the bank. Thus, when the bank is full, a deactivating lineage disappears instead of moving to the bank. In our case, we start with n plants and m seeds (the bank is full from the beginning).

Denote by $\bar{N}_{n,m}(t)$ (resp. $\bar{M}_{n,m}(t)$) the number of plants (resp. seeds) at time t in the bounded coalescent starting with n plants and m seeds. The block-counting process of the bounded coalescent with parameters $c_1, c_2 > 0$ has the following transition rates. For $i \leq n$ and $j \leq m$,

$$(\bar{N}_{n,m}(t), \bar{M}_{n,m}(t)) \text{ jumps from } (i,j) \text{ to } \begin{cases} (i-1, j), & \text{at rate } \binom{i}{2} + c_1 i \mathbf{1}_{\{j=m\}}, \\ (i-1, j+1), & \text{at rate } c_1 i \mathbf{1}_{\{j<m\}}, \\ (i+1, j-1), & \text{at rate } c_2 j. \end{cases}$$

By coupling the seed bank coalescent with its bounded version, we obtain a lower bound for θ_n and an upper bound for $N_n(\theta_n)$.

Lemma 3.3.3. *Recall T and Z from Proposition 3.3.1. We have that*

$$\lim_{n\to\infty} \mathbb{P}\left(\theta_n \log n \le t\right) \le \mathbb{P}(T \le t) \tag{3.16}$$

and

$$\lim_{n\to\infty} \mathbb{P}\left(N_n(\theta_n) > z\log n\right) \le \mathbb{P}\left(Z > z\right). \tag{3.17}$$

Proof. On the event $\{M_n(\theta_n) \le m_n\}$, which occurs asymptotically with probability 1 by Lemma 3.3.2, the variable θ_n is bounded from below, stochastically, by the random variable $\bar{\theta}_{n,m_n}$ defined by

$$\bar{\theta}_{n,m_n} = \inf\{t \ge 0 : \bar{M}_{n,m_n}(t-) > \bar{M}_{n,m_n}(t)\}$$

which has an exponential distribution with parameter $c_2 m_n$. Then, for $t > 0$

$$\begin{aligned}
\mathbb{P}\left(\theta_n \log n \le t\right) &= \mathbb{P}\left(\theta_n \log n \le t, M_n(\theta_n) \le m_n\right) + o(1) \\
&\le \mathbb{P}\left(\bar{\theta}_{n,m_n} \log n \le t\right) + o(1) \\
&= 1 - \exp\left\{-t\frac{c_2\lfloor 2c_1(1+\varepsilon)\log n\rfloor}{\log n}\right\} + o(1).
\end{aligned}$$

So, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\theta_n \log n \le t\right) \le \mathbb{P}\left(T \le t(1+\varepsilon)\right). \tag{3.18}$$

This gives (3.16).

To prove (3.17), observe that, on the event $\{M_n(\theta_n) \le m_n\}$, the variable $N_n(\theta_n)$ is bounded from above, stochastically, by the random variable $\bar{N}_{n,m_n}(\bar{\theta}_{n,m_n})$. So,

$$\mathbb{P}\left(N_n(\theta_n) > z\log n\right) \le \mathbb{P}\left(\bar{N}_{n,m_n}(\bar{\theta}_{n,m_n}) > z\log n\right) + \mathbb{P}(M_n(\theta_n) > m_n). \tag{3.19}$$

Let us study the asymptotic of $\bar{N}_{n,m_n}(\bar{\theta}_{n,m_n})$. With similar arguments as used for Proposition 3.2.1, we have

$$\begin{aligned}
\mathbb{P}(\bar{N}_{n,m_n}(\bar{\theta}_{n,m_n}) \le z\log n) &= \prod_{i=\lfloor z\log n\rfloor+1}^{n} \frac{\binom{i}{2} + c_1 i}{\binom{i}{2} + c_1 i + c_2 m_n} \\
&= \exp\left\{-\sum_{i=\lfloor z\log n\rfloor+1}^{n} \log\left(1 + \frac{2c_2 m_n}{i(i-1+2c_1)}\right)\right\} \\
&\sim \exp\left\{-2c_2 m_n \sum_{i=\lfloor z\log n\rfloor+1}^{n} \frac{1}{i^2}\right\}.
\end{aligned}$$

By a Riemann sum argument, we know that

$$\lim_{n\to\infty} m_n \sum_{i=\lfloor z\log n\rfloor+1}^{n} \frac{1}{i^2} = 2c_1(1+\varepsilon)\int_z^\infty \frac{1}{x^2}dx = \frac{2c_1(1+\varepsilon)}{z}. \tag{3.20}$$

Since $\mathbb{P}(Z \leq z) = \exp\{-4c_1c_2/z\}$, we obtain, by taking the limits in (3.19), that

$$\lim_{n\to\infty} \mathbb{P}\left(N_n(\theta_n) > z\log n\right) \leq \mathbb{P}\left(Z > z/(1+\varepsilon)\right)$$

which implies (3.17). □

The bounded seed bank coalescent is also useful to bound from above the random variable $N_n(t)$, for any $t \geq 0$. Let $(K_n(t))_{t\geq 0}$ stand for the block-counting process of the Kingman coalescent starting with n lineages. Let $(\chi_i(t))_{i\geq 1}$ be a sequence of i.i.d. Bernoulli variables of parameter $1-\exp(-c_2t)$. Those variables are more easily understood as $\chi_i(t) = \mathbf{1}_{\{e_i<c_2t\}}$ where the e_i's are i.i.d. standard exponential variables. It is easy to convince oneself that, on the event $\{\sup_{t\geq 0} M_n(t) \leq m\}$, stochastically,

$$N_n(t) \leq K_n(t) + \sum_{i=1}^{m} \chi_i(t). \tag{3.21}$$

This follows because $K_n(t)$ bounds the number of blocks that have not been deactivated before time t and $\sum_{i=1}^{m}\chi_i(t)$ bounds the number of blocks that have already reactivated. Both processes are independent.

We now prove a useful lemma thanks to the two couplings introduced previously. To simplify the notations here and in the sequel, denote $\tau_n^{\lfloor(\log n)^a\rfloor}$ by $\tau_n^{(a)}$, for any $a > 0$.

Lemma 3.3.4. *For $a > b \geq 0$ such that $a + b > 1$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\tau_n^{(a)} \leq (\log n)^{-b}\right) = 1.$$

Proof. Denote $E_n = \{\sup_t M_n(t) \leq m_n\}$. We start by observing that

$$\mathbb{P}(\tau_n^{(a)} > (\log n)^{-b}) = \mathbb{P}(\tau_n^{(a)} > (\log n)^{-b}, E_n) + \mathbb{P}(\tau_n^{(a)} > (\log n)^{-b}, E_n^c)$$

From (3.15), we get that

$$\mathbb{P}(E_n^c) \leq \frac{1}{2c_1\varepsilon^2\log n}.$$

So it just remains to control the probability on the event E_n. Recall $(K_n(t))_{t\geq 0}$ and $(\chi_i(t))_{i\geq 1}$ from (3.21). Let $\omega_{n,a} = \inf\{t > 0 : K_n(t) = \lfloor\frac{1}{2}(\log n)^a\rfloor\}$. Observe that

$$\begin{aligned}
\{\tau_n^{(a)} > t, E_n\} &= \{N_n(t) > (\log n)^a, E_n\} \\
&\subset \{K_n(t) + \sum_{i=1}^{m_n}\chi_i(t) > (\log n)^a\} \\
&\subset \{K_n(t) > \frac{1}{2}(\log n)^a\} \cup \{\sum_{i=1}^{m_n}\chi_i(t) > \frac{1}{2}(\log n)^a\} \\
&= \{\omega_{n,a} > t\} \cup \{\sum_{i=1}^{m_n}\chi_i(t) > \frac{1}{2}(\log n)^a\}.
\end{aligned}$$

Taking $t = (\log n)^{-b}$, we obtain

$$\mathbb{P}(\tau_n^{(a)} > (\log n)^{-b}, E_n) \leq \mathbb{P}(\omega_{n,a} > (\log n)^{-b}) + \mathbb{P}(\sum_{i=1}^{m_n} \chi_i((\log n)^{-b}) > \frac{1}{2}(\log n)^a).$$

Observe that $\omega_{n,a}$ is the sum of independent exponential random variables with parameter $\binom{i}{2}$ for $\lfloor \frac{1}{2}(\log n)^a \rfloor + 1 \leq i \leq n$. Thus,

$$\mathbb{E}[\omega_{n,a}] = \sum_{i=\lfloor \frac{1}{2}(\log n)^a \rfloor + 1}^{n} \left[\frac{2}{(i-1)} - \frac{2}{i} \right] \leq 2\lfloor \frac{1}{2}(\log n)^a \rfloor.$$

So, Markov's inequality for $\omega_{n,a}$ gives

$$\mathbb{P}(\omega_{n,a} > (\log n)^{-b}) \leq C(\log n)^{b-a}$$

for some constant $C > 0$, which converges to 0 whenever $b < a$. On the other hand, Markov's inequality applied to a binomial random variable with parameters $\lfloor 2c_1(1+\varepsilon)\log n \rfloor$ and $1 - \exp(-c_2(\log n)^{-b})$ (whose expectation is of order $(\log n)^{1-b}$) leads to

$$\mathbb{P}(\sum_{i=1}^{m_n} \chi_i((\log n)^{-b}) > \frac{1}{2}(\log n)^a) \leq C(\log n)^{1-b-a}.$$

This quantity converges to 0 when $a + b > 1$. □

We now provide the lower bound for $M_n(\theta_n)$. This result, combined with Lemma 3.3.2 provides the convergence (3.11) in Proposition 3.3.1.

Lemma 3.3.5. *For any $\varepsilon > 0$ and $a > 1$,*

$$\lim_{n\to\infty} \mathbb{P}(M_n(\tau_n^{(a)}) > 2c_1(1-\varepsilon)\log n) = 1. \tag{3.22}$$

which implies that

$$\lim_{n\to\infty} \mathbb{P}(M_n(\theta_n) > 2c_1(1-\varepsilon)\log n) = 1. \tag{3.23}$$

Proof. Let us first note that (3.17) implies that

$$\lim_{n\to\infty} \mathbb{P}(N_n(\theta_n) < (\log n)^a) = 1,$$

which, thanks to the monotonicity of $(N_n(t))_{t\geq 0}$ until time θ_n, is equivalent to

$$\lim_{n\to\infty} \mathbb{P}\left(\theta_n > \tau_n^{(a)}\right) = 1.$$

Due to the monotonicity of $(M_n(t))_{t\geq 0}$ until time θ_n, (3.22) implies (3.23).

Now, on the event $\{\theta_n > \tau_n^{(a)}\}$, we have

$$M_n(\tau_n^{(a)}) = \sum_{i=\lfloor(\log n)^a\rfloor}^{n-1} B_n^i$$

almost surely, where the B_n^i's are the Bernoulli random variables introduced in (3.14). So,

$$\begin{aligned}\mathbb{P}(M_n(\tau_n^{(a)}) < 2c_1(1-\varepsilon)\log n) &= \mathbb{P}(M_n(\tau_n^{(a)}) < 2c_1(1-\varepsilon)\log n, \theta_n > \tau_n^{(a)}) + o(1)\\ &\leq \mathbb{P}\left(\sum_{i=\lfloor(\log n)^a\rfloor}^{n-1} B_n^i < 2c_1(1-\varepsilon)\log n\right) + o(1)\end{aligned}$$

The latter converges to 0 thanks to Bienaymé-Chebyshev's inequality. □

We are now able to end the overview of the system at time θ_n. The following result, combined with Lemma 3.3.3 provides the convergences (3.10) and (3.12) in Proposition 3.3.1.

Lemma 3.3.6. *Recall T and Z from Proposition 3.3.1. We have that*

$$\lim_{n\to\infty} \mathbb{P}\left(N_n(\theta_n) \leq z\log n\right) \leq \mathbb{P}\left(Z \leq z\right). \tag{3.24}$$

which implies that

$$\lim_{n\to\infty} \mathbb{P}\left(\theta_n \log n > t\right) \leq \mathbb{P}(T > t). \tag{3.25}$$

Proof. Fix $\varepsilon > 0$ and define $\widehat{m}_n := \lfloor 2c_1(1-\varepsilon)\log n\rfloor$. Also, denote $\hat{\tau}_n := \tau_n^{\lfloor z\log n\rfloor}$. First observe that

$$\mathbb{P}\left(N_n(\theta_n) \leq z\log n\right) = \mathbb{P}(\theta_n \geq \hat{\tau}_n)$$

So it is enough to prove that

$$\lim_{n\to\infty} \mathbb{P}(\theta_n \geq \hat{\tau}_n) \leq \mathbb{P}(Z \leq z). \tag{3.26}$$

For any $t \geq 0$, define $X(t)$ to be the number of reactivations until time t. Let $\mathcal{E}_i$ be an exponential random variable with parameter $c_2 i$, that can be understood as the minimum of i independent exponential random variables with parameter c_2. Then, for any $a > 1$,

$$\begin{aligned}\mathbb{P}(\theta_n \geq \hat{\tau}_n) &= \mathbb{P}(X(\hat{\tau}_n) = 0) = \mathbb{P}(X(\hat{\tau}_n) - X(\tau_n^{(a)}) = 0, X(\tau_n^{(a)}) = 0)\\ &\leq \mathbb{P}(X(\hat{\tau}_n) - X(\tau_n^{(a)}) = 0 \mid X(\tau_n^{(a)}) = 0)\\ &\leq \mathbb{P}(\mathcal{E}_{M_n(\tau_n^{(a)})} > \hat{\tau}_n - \tau_n^{(a)}).\end{aligned}$$

The latter inequality follows by observing that if there are no activations in the time interval $[\tau_n^{(a)}, \hat{\tau}_n]$, then none of the $M_n(\tau_n^{(a)})$ seeds present at time $\tau_n^{(a)}$ have activated. Hence,

$$\begin{aligned}
\mathbb{P}(\theta_n \geq \hat{\tau}_n) &\leq \mathbb{E}\left[e^{-c_2(\hat{\tau}_n - \tau_n^{(a)})M_n(\tau_n^{(a)})}\right] \\
&= \mathbb{E}\left[e^{-c_2(\hat{\tau}_n - \tau_n^{(a)})M_n(\tau_n^{(a)})}\mathbf{1}_{\{M_n(\tau_n^{(a)}) > \widehat{m}_n\}}\right] \\
&\quad + \mathbb{E}\left[e^{-c_2(\hat{\tau}_n - \tau_n^{(a)})M_n(\tau_n^{(a)})}\mathbf{1}_{\{M_n(\tau_n^{(a)}) \leq \widehat{m}_n\}}\right] \\
&\leq \mathbb{E}\left[e^{-c_2\widehat{m}_n(\hat{\tau}_n - \tau_n^{(a)})}\right] + \mathbb{P}(M_n(\tau_n^{(a)}) \leq \widehat{m}_n).
\end{aligned}$$

So, by denoting for simplicity $n_z = \lfloor z \log n \rfloor$ and $n_a = \lfloor (\log n)^a \rfloor$, and by (3.13), we obtain

$$\mathbb{P}(\theta_n \geq \hat{\tau}_n) \leq \mathbb{E}\left[e^{-c_2\widehat{m}_n \sum_{i=n_z+1}^{n_a}(\underline{\tau}_n^{i-1} - \underline{\tau}_n^{i})}\right] + \mathbb{P}(M_n(\tau_n^{(a)}) \leq \widehat{m}_n). \tag{3.27}$$

Since the variables $\underline{\tau}_n^{i-1} - \underline{\tau}_n^{i}$ are independent and exponentially distributed, we have

$$\begin{aligned}
\mathbb{E}\left[e^{-c_2\widehat{m}_n \sum_{i=n_z+1}^{n_a}(\underline{\tau}_n^{i-1} - \underline{\tau}_n^{i})}\right] &= \prod_{i=n_z+1}^{n_a} \frac{\binom{i}{2} + c_1 i}{\binom{i}{2} + c_1 i + c_2\widehat{m}_n} \\
&= \exp\left\{-\sum_{i=n_z+1}^{n_a} \log\left(1 + \frac{2c_2\widehat{m}_n}{i(i-1+2c_1)}\right)\right\}.
\end{aligned}$$

Now, we can use equivalences.

$$\mathbb{E}\left[e^{-c_2\widehat{m}_n \sum_{i=n_z+1}^{n_a}(\underline{\tau}_n^{i-1} - \underline{\tau}_n^{i})}\right] \sim \exp\left\{-\sum_{i=n_z+1}^{n_a} \frac{2c_2\widehat{m}_n}{i^2}\right\}$$

A similar limit as that given in (3.20) implies that

$$\lim_{n\to\infty} \mathbb{E}\left[e^{-c_2\widehat{m}_n \sum_{i=n_z+1}^{n_a}(\underline{\tau}_n^{i-1} - \underline{\tau}_n^{i})}\right] = e^{-\frac{4c_1c_2(1-\varepsilon)}{z}} = \mathbb{P}(Z \leq z/(1-\varepsilon)). \tag{3.28}$$

Plugging (3.28) and (3.22) into (3.27), and observing that the result is true for any $\varepsilon > 0$, we get (3.26).

A very similar path is followed to obtain (3.25). For some $t > 0$, let $t_n = t(\log n)^{-1}$ and for some $b > 1$, let $s_n = (\log n)^{-b}$. As before, we get

$$\begin{aligned}
\mathbb{P}\left(\theta_n \log n > t\right) &= \mathbb{P}\left(\theta_n > t_n\right) \\
&= \mathbb{P}\left(X\left(t_n\right) = 0\right) \\
&\leq e^{-c_2\widehat{m}_n(t_n - s_n)} + \mathbb{P}(M_n(s_n) \leq \widehat{m}_n),
\end{aligned}$$

The first term converges to $\mathbb{P}(T > t(1-\varepsilon))$ and the second to 0. To get the latter, first use (3.16) to see that

$$\lim_{n\to\infty} \mathbb{P}(\theta_n > s_n) = 1.$$

Then, just choose $a > b$ such that Lemma 3.3.4 holds, and use (3.22) Since the result is true for any $\varepsilon > 0$, we get (3.25). □

3.4 Branch Lengths

In this section, we study the total branch length L_n of the seed bank coalescent starting with n plants and no seeds as defined in (3.4) and prove Theorem 3.1.1 by combining upcoming Theorems 3.4.1 and 3.4.2.

3.4.1 The active length

Consider the active length defined in (3.5). We prove that this variable has the same (first-order) asymptotics as the total length of the Kingman coalescent.

Theorem 3.4.1. *Consider the seed bank coalescent starting with n plants and no seeds. Then,*

$$\lim_{n\to\infty} \frac{A_n}{\log n} = 2$$

in probability.

Proof. Recall the notation $\tau_n^{(a)} := \tau_n^{\lfloor (\log n)^a \rfloor}$ for some arbitrary fixed $a \in (1/2, 1)$. We divide A_n into the sum of two random variables

$$A_n = A_n^1 + A_n^2$$

where

$$A_n^1 = \int_0^{\tau_n^{(a)}} N_n(t)dt \quad \text{and} \quad A_n^2 = \int_{\tau_n^{(a)}}^{\sigma_n} N_n(t)dt.$$

We prove that $A_n^2/\log n$ converges to 0 in probability in part iii) later on.

To deal with A_n^1, we will divide it into the sum of two random variables

$$A_n^1 = A_n^{1,1} + A_n^{1,2}$$

where

$$A_n^{1,1} = \int_0^{\theta_n} N_n(t)dt \quad \text{and} \quad A_n^{1,2} = \int_{\theta_n}^{\tau_n^{(a)}} N_n(t)dt.$$

Here we have to work carefully since θ_n can be larger than $\tau_n^{(a)}$. However, observe that

$$\begin{aligned} \mathbb{P}\left(\theta_n \geq \tau_n^{(a)}\right) &= \mathbb{P}\left(N_n(\theta_n) \leq N_n(\tau_n^{(a)})\right) \\ &= \mathbb{P}\left(N_n(\theta_n) \leq \lfloor (\log n)^a \rfloor\right). \end{aligned}$$

By Proposition 3.3.1 we see that this probability converges to 0. So,

$$
\begin{aligned}
& \mathbb{P}\left(\Big|\frac{A_n^1}{\log n}-2\Big|>\varepsilon\right) \\
= \;& \mathbb{P}\left(\Big|\frac{A_n^1}{\log n}-2\Big|>\varepsilon, \theta_n<\tau_n^{(a)}\right)+\mathbb{P}\left(\Big|\frac{A_n^1}{\log n}-2\Big|>\varepsilon, \theta_n\geq\tau_n^{(a)}\right) \\
= \;& \mathbb{P}\left(\Big|\frac{A_n^1}{\log n}-2\Big|>\varepsilon, \theta_n<\tau_n^{(a)}\right)+o(1) \\
\leq \;& \mathbb{P}\left(\Big|\frac{A_n^{1,1}}{\log n}-2\Big|>\frac{\varepsilon}{2}, \theta_n<\tau_n^{(a)}\right)+\mathbb{P}\left(\Big|\frac{A_n^{1,2}}{\log n}\Big|>\frac{\varepsilon}{2}, \theta_n<\tau_n^{(a)}\right)+o(1).
\end{aligned}
$$

Parts i) and ii) are dedicated to show that these two terms converge to 0.

i) Let us first prove that

$$
\lim_{n\to\infty}\mathbb{P}\left(\Big|\frac{A_n^{1,1}}{\log n}-2\Big|>\frac{\varepsilon}{2}, \theta_n<\tau_n^{(a)}\right)=0 \tag{3.29}
$$

Observe that, between times 0 and θ_n, only coalescence or deactivation events occur. This implies that we can rewrite A_n^1 as follows,

$$
A_n^{1,1}=\sum_{i=N_n(\theta_n)+1}^{n} iE_i,
$$

where, given $(M_n(\tau_n^i))_{i=1}^n$, the E_i's are independent exponential random variables with respective parameters $\binom{i}{2}+c_1 i+c_2 M_n(\tau_n^i)$. Indeed, this parameter is that of the minimum of three exponential random variables, the first for coalescence, the second for deactivation, and the third for activation.

Let $h_n=\sum_{i=1}^{n-1}\frac{2}{i+2c_1}$. By proving that

$$
\mathbb{E}[|A_n^{1,1}-h_n|]=o(\log n),
$$

we get the desired result. Observe that the variable A_n^1 is stochastically bounded by the length of a Kingman coalescent with freezing ([70], [23] and see Section 1.3 in [25]), that is

$$
H_n=\sum_{i=2}^{n} iV_i,
$$

where the V_i's, as in Section 3.2, are independent exponential random variables with respective parameters $\binom{i}{2}+c_1 i$. This is true because the seeds "accelerate" the jump times. To be precise consider the following coupling. Let $V_i=\min\{E_i^{(c)}, E_i^{(d)}\}$ where $E_i^{(c)}$ is exponential with parameter $\binom{i}{2}$ and $E_i^{(d)}$ is exponential with parameter $c_1 i$. Now let $E_{i,m}^{(a)}$ be exponential with parameter $c_2 m$.

Construct a process $(\tilde{N}_n(t), \tilde{M}_n(t))_{t\geq 0}$, equal in distribution to $(N_n(t), M_n(t))_{t\geq 0}$ up to time θ_n, recursively, using these exponential random variables. This is,

$$(\tilde{N}_n(t), \tilde{M}_n(t)) \text{ jumps from } (i,m) \text{ to } \begin{cases} (i-1, m), & \text{if } \min\{E_i^{(c)}, E_i^{(d)}, E_{i,m}^{(a)}\} = E_i^{(c)} \\ (i-1, m+1), & \text{if } \min\{E_i^{(c)}, E_i^{(d)}, E_{i,m}^{(a)}\} = E_i^{(d)} \\ (0,0), & \text{otherwise .} \end{cases}$$

Here $(0,0)$ represents a cemetery state. Note that in distribution $(\tilde{N}_n(t), \tilde{M}_n(t)) = (N_n(t), M_n(t))1_{\{\theta_n < t\}}$. Thus, by writing $(\tilde{\tau}_n^i)_{i=1}^n$ for the successive jump times of the new process and $\tilde{r}_n = \sup\{i \geq 1 : \min\{E_i^{(c)}, E_i^{(d)}, E_{i,\tilde{M}_n(\tilde{\tau}_n^i)}^{(a)}\} = E_{i,\tilde{M}_n(\tilde{\tau}_n^i)}^{(a)}\}$, we obtain that

$$A_n^{1,1} = \sum_{i=\tilde{r}_n+1}^{n} iV_i \leq \sum_{i=2}^{n} iV_i = H_n,$$

where the first equality is in distribution and the others stand almost surely. The first equality is true because, although the V_i's are variables with the "wrong" parameter, they are not independent of $\tilde{r}_n$, and this dependence "accelerates" these exponential random variables. Hence,

$$\mathbb{E}[|A_n^{1,1} - h_n|] \leq \mathbb{E}[H_n - A_n^{1,1}] + \mathbb{E}[|H_n - h_n|].$$

The second term is bounded thanks to the L^1-convergence of sums of independent exponential variables. For the first term,

$$\begin{aligned} \mathbb{E}[H_n - A_n^{1,1}] &= \mathbb{E}\left[H_n - \mathbb{E}\left[A_n^{1,1} | N_n(\theta_n), (M_n(\tau_n^i))_{i\geq 1}\right]\right] \\ &= h_n - \mathbb{E}\left[\sum_{i=N_n(\theta_n)+1}^{n} \frac{2}{i - 1 + 2c_1 + \frac{2c_2 M_n(\tau_n^i)}{i}}\right] \\ &\leq h_n - \mathbb{E}\left[\sum_{i=N_n(\theta_n)+1}^{n} \frac{2}{i - 1 + 2c_1 + \frac{2c_2 \sup_t M_n(t)}{i}}\right]. \end{aligned}$$

Then, denote $a_n := \lfloor (\log n)^{1+\varepsilon_1} \rfloor$, for some $\varepsilon_1 > 0$, and recall the notation m_n from Section 3.3. Now, set the event

$$E_n = \{\sup_t M_n(t) \leq m_n, N_n(\theta_n) \leq a_n\}.$$

We obtain that

$$\begin{aligned}\mathbb{E}[H_n - A_n^{1,1}] &\leq h_n - \mathbb{E}\left[\mathbf{1}_{E_n} \sum_{i=N_n(\theta_n)+1}^{n} \frac{2}{i - 1 + 2c_1 + \frac{2c_2 \sup_t M_n(t)}{i}}\right] \\ &\leq h_n - \mathbb{P}(E_n) \sum_{i=a_n+1}^{n} \frac{2}{i - 1 + 2c_1 + \frac{2c_2 m_n}{i}} \\ &\leq h_n - \mathbb{P}(E_n) \sum_{i=a_n+1}^{n} \frac{2}{i - 1 + 2c_1 + \frac{2c_2 m_n}{a_n+1}}.\end{aligned}$$

Since $\frac{m_n}{a_n+1} \leq C(\log n)^{-\varepsilon_1}$ for some constant C and $\mathbb{P}(E_n)$ converges to 1 (thanks to Proposition 3.3.1), we get that

$$\mathbb{E}[H_n - A_n^{1,1}] = o(\log n).$$

The L^1-convergence is thus obtained. This implies (3.29).

ii) Let us now prove that

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{A_n^{1,2}}{\log n}\right| > \varepsilon, \theta_n < \tau_n^{(a)}\right) = 0. \tag{3.30}$$

It is clear that, on the event $\{\theta_n < \tau_n^{(a)}\}$,

$$A_n^{1,2} \leq \tau_n^{(a)}(N_n(\theta_n) + M_n(\theta_n)).$$

Combining Proposition 3.3.1 and Lemma 3.3.4 (choosing $b < a$), we obtain the result.

iii) Finally, let us prove that

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{A_n^{2}}{\log n}\right| > \varepsilon\right) = 0. \tag{3.31}$$

To this end, denote $U_0 = N_n(\tau_n^{(a)}) = \lfloor (\log n)^a \rfloor$ (by definition), $V_0 = M_n(\tau_n^{(a)})$ (which, by Lemma 3.3.2, is stochastically bounded by $2c_1(1+\varepsilon)\log n$), and, for any $k \geq 1$, U_k (resp. V_k) as the number of plants (resp. seeds) at the kth event after time $\tau_n^{(a)}$. Each event can be a coalescence, an activation or a deactivation. Note that the increments of U_k and V_k are in $\{-1, 1\}$. Let S_n be the number of jump times during the interval $(\tau_n^{(a)}, \sigma_n]$, i.e.

$$S_n = \inf\{k \geq 1 : U_k + V_k = 1\}.$$

With these notations, the active branch length on this time interval can be written as

$$A_n^2 = \sum_{k=0}^{S_n - 1} U_k E_k$$

where, conditional on U_k and V_k, the E_k's are independent exponential random variables with respective parameters $\binom{U_k}{2} + c_1 U_k + c_2 V_k$. So, we have

$$\mathbb{E}[A_n^2] = \mathbb{E}\left[\sum_{k=0}^{S_n-1} \frac{U_k}{\binom{U_k}{2} + c_1 U_k + c_2 V_k}\right].$$

Now define

$$D_n := |\{k \geq 0 : U_{k+1} - U_k = -1, V_{k+1} - V_k = 1\}|$$

as the number of deactivations during this time interval, and observe that

$$\mathbb{E}[D_n] = \mathbb{E}\left[\sum_{k=0}^{S_n-1} \frac{c_1 U_k}{\binom{U_k}{2} + c_1 U_k + c_2 V_k}\right].$$

This implies that

$$\mathbb{E}[A_n^2] = \frac{1}{c_1}\,\mathbb{E}[D_n].$$

So, it is enough to study the expectation of D_n. We decompose

$$D_n = \sum_{i=2}^{N_n(\tau_n^{(a)}) + M_n(\tau_n^{(a)})} D_n^i$$

where D_n^i is the number of deactivations occurring while the total number of lineages equals i, that is, $D_n^i := |\{k \geq 0 : U_{k+1} - U_k = -1, V_{k+1} - V_k = 1, U_k + V_k = i\}|$. We will bound $\mathbb{E}[D_n]$ thanks to the next model from Definition 4.9 of [9].

Let $(\widehat{N}_n(t), \widehat{M}_n(t))_{t\geq 0}$ having the same transitions as $(N_n(t), M_n(t))_{t\geq 0}$ whenever $\widehat{N}_n(t) \geq \sqrt{\widehat{N}_n(t) + \widehat{M}_n(t)}$. If not, coalescence events are not permitted. For any $i \geq 2$, by Lemma 4.10 of [9], $\mathbb{E}[D_n^i] \leq \mathbb{E}[\widehat{D}_n^i]$, where $\widehat{D}_n^i$ stands for the number of deactivations in this model while $\widehat{N}_n(t) + \widehat{M}_n(t) = i$. In what follows we will give an idea of why $\mathbb{E}[\widehat{D}_n^i] = O(i^{-1/2})$, implying that $\mathbb{E}[D_n] = O((\log n)^{1/2})$, and hence proving (3.31).

Details of the proof, which are unfortunately quite tedious, can be found inside the proof of Lemmas 4.10 and 4.11 of [9]. In the sequel, suppose that $c_1 = c_2 = 1$, for sake of simplicity.

Fix $i \geq 2$. The variables $\widehat{D}_n^i$ tends to take higher values when coalescences are not permitted, we focus on this case. Thus suppose that at time t, $\widehat{N}_n(t) + \widehat{M}_n(t)$ reaches i, with $\widehat{N}_n(t-) = \lfloor\sqrt{i}\rfloor + 1 \geq \sqrt{i+1}$. This means that $\widehat{N}_n(t) = \lfloor\sqrt{i}\rfloor \leq \sqrt{i}$. Reactivations are then needed to allow a new coalescence. Conditional on this configuration, the probability that $\widehat{D}_n^i$ equals 0 is equivalent to

$$\frac{i - \lfloor\sqrt{i}\rfloor}{i} \times \frac{\binom{\lfloor\sqrt{i}\rfloor}{2}}{\binom{\lfloor\sqrt{i}\rfloor}{2} + \lfloor\sqrt{i}\rfloor} \sim 1 - \frac{3}{\sqrt{i}} =: p_i.$$

This corresponds approximately to the probability of one reactivation, followed by one coalescence before one deactivation. So we have the following almost sure bound

$$\widehat{D}_n^i \leq \sum_{j=0}^{G^i-1} \Delta_j$$

where G^i is a geometric random variable with parameter p_i and the Δ_j's give the number of deactivations between each visit of the state $\lfloor\sqrt{i}\rfloor$. The time when coalescence is not allowed, is stochastically bounded from above by the time that a random walk that goes up one unit at rate $i-\sqrt{i}$ (rate at of a reactivation) and down at rate $\sqrt{i}$ (rate of a deactivation), started at zero, spends below level $\sqrt{i}$. The random walk has ballistic speed of order i. In particular, it reaches the level $\sqrt{i}$ after $\sqrt{i}/i = 1/\sqrt{i}$ units of time in average. During the period in which coalescence events are not allowed there are always less that $\sqrt{i}$ plants, each of which deactivates at rate $c_1(=1)$. Then, we conclude that, for any j,

$$\mathbb{E}[\Delta_j] \leq \frac{1}{\sqrt{i}} \cdot \sqrt{i} = 1$$

This uniform bound implies that

$$\mathbb{E}[\widehat{D}_n^i] \leq \mathbb{E}[G^i - 1]\mathbb{E}[\Delta_1] = O\left(\frac{1}{\sqrt{i}}\right),$$

since $\mathbb{E}[G^i - 1] \sim \frac{3}{\sqrt{i}}$. □

3.4.2 The inactive length

Consider the inactive length defined in (3.6).

Theorem 3.4.2. *Consider the seed bank coalescent starting with n plants and no seeds. Then,*

$$\lim_{n\to\infty} \frac{I_n}{\log n} = \frac{2c_1}{c_2}$$

in probability.

Proof. Divide I_n in two parts

$$I_n^1 = \int_0^{\theta_n} M_n(t)dt \quad \text{and} \quad I_n^2 = \int_{\theta_n}^{\sigma_n} M_n(t)dt.$$

It is easy to prove that $I_n^1/\log n$ converges to 0 in probability by observing that, almost surely,

$$I_n^1 \leq M_n(\theta_n) \cdot \theta_n,$$

and using Proposition 3.3.1.

To study I_n^2, we approximate it by the accumulated time for the $M_n(\theta_n)$ seeds to activate, namely

$$\tilde{I}_n^2 = \sum_{k=1}^{M_n(\theta_n)} \frac{e_k}{c_2}$$

where the e_k's are i.i.d. standard exponential random variables. The asymptotics of this random variable are easily obtained. First, by Proposition 3.3.1, we have that

$$M_n(\theta_n)/\log n \to 2c_1$$

in probability. Second, let $G_n(t) = \sum_{k=1}^{\lfloor t \log n \rfloor} \frac{e_k}{c_2}$,

$$\lim_{n\to\infty} \frac{G_n(t)}{\log n} = \frac{t}{c_2} \tag{3.32}$$

en L^2. Observe that $\tilde{I}_n^2 = G_n(M_n(\theta_n)/\log n)$. Following the same steps in the proof of Proposition 3.2.2. We obtain the desired result,

$$\lim_{n\to\infty} \frac{\tilde{I}_n^2}{\log n} = \frac{2c_1}{c_2}$$

in probability.

Finally, the difference between I_n^2 and $\tilde{I}_n^2$ can be bounded by $I_{N_n(\theta_n)} + I_{M_n(\theta_n)}$. Indeed, the variable $I_{N_n(\theta_n)}$ bounds the inactive length resulting from the plants present at time θ_n and the variable $I_{M_n(\theta_n)}$ bounds the inactive length resulting from the seeds present at time θ_n that activate and deactivate again. Its expectation is clearly of order $\log\log n$. This can be seen repeating the earlier arguments of this proof. □

3.5 Sampling formula

Consider the seed bank coalescent at time θ_n and go back, through the active part of the genealogical tree, until time zero when there are n active lineages and zero inactive lineages. During this period of time we observe $n - N_n(\theta_n)$ events divided into two types: branching inside one lineage (corresponding to a coalescence) and appearance of a new lineage (corresponding to a deactivation). When there are k lineages, the probability that a branching event occurs is

$$\frac{\binom{k+1}{2}}{\binom{k+1}{2} + c_1(k+1)} = \frac{k}{k+2c_1}$$

whereas the probability that a new lineage appears is $\frac{2c_1}{k+2c_1}$. This observation leads a connection with classical Hoppe's urn and the Chinese restaurant process (with parameter $2c_1$), which are the key tools to prove Ewens' sampling formula for the law of the allele frequency spectrum in the neutral model, see Chapter 1.3 in [25]. However, in our case, the initial configuration is made of a random number $N_n(\theta_n)$ of tables (old lineages) with one client in each. By applying results of [70], we can obtain a conditional sampling formula corresponding to observing a certain configuration of lineages that passed through the seed bank and lineages that did not deactivate (until time θ_n).

Now, let $k \leq n$ be a positive integer, we define the sets

$$A(k,n) = \left\{ a_i, b_i \geq 0, i \in [n] : \sum_{i=1}^{n} a_i = k \text{ and } \sum_{i=1}^{n} i(a_i + b_i) = n \right\}$$

and

$$\bar{A}(k,n) = \left\{ a_i \geq 0, i \in [n] : \sum_{i=1}^{n} a_i = k \text{ and } \sum_{i=1}^{n} i a_i \leq n \right\}.$$

From equation (3.3.2) in [70], we obtain the next theorem.

Theorem 3.5.1. *Let O_i be the number of "old" blocks of size i (i.e. active blocks of size i at time θ_n) and let R_i be the number of "recent" blocks of size i (i.e. inactive blocks of size i at time θ_n). Then*

$$\begin{aligned} &\mathbb{P}\left(O_1 = a_1, \ldots, O_n = a_n, R_1 = b_1, \ldots, R_n = b_n \mid N_n(\theta_n)\right) \\ &\overset{a.s.}{=} \frac{(n - N_n(\theta_n))! N_n(\theta_n)!}{(N_n(\theta_n) + 2c_1)_{(n-N_n(\theta_n))}} \prod_{i=1}^{n} \frac{1}{a_i!} \prod_{j=1}^{n} \frac{1}{b_j!} \left(\frac{2c_1}{j}\right)^{b_j}, \end{aligned} \tag{3.33}$$

with $(a_i, b_i)_{i\in[n]} \in A(N_n(\theta_n), n)$.

The notation $x_{(n)}$ stands for the ascending factorial, that is, $x_{(n)} = x(x+1)\ldots(x+n-1)$.

Remark 3.5.2. From the latter result and Proposition 3.3.1, we can obtain an ap-

proximate unconditioned sampling formula for large n.

$$\begin{aligned}
&\mathbb{P}\left(O_1 = a_1, \ldots, O_n = a_n, R_1 = b_1, \ldots, R_n = b_n\right) \\
&= \int_0^\infty \mathbb{P}\left(O_1 = a_1, \ldots, O_n = a_n, R_1 = b_1, \ldots, R_n = b_n | N_n(\theta_n) = \lfloor z \log n \rfloor\right) \times \\
&\quad \mathbb{P}(N_n(\theta_n) = \lfloor z \log n \rfloor) dz \\
&\sim \prod_{i=1}^n \frac{1}{a_i!} \prod_{j=1}^n \frac{1}{b_j!} \left(\frac{2c_1}{j}\right)^{b_j} \times \\
&\quad \int_0^\infty \frac{\Gamma(n - z\log n + 1)\Gamma(z\log n + 1)\Gamma(z\log n + 2c_1)}{\Gamma(n + 2c_1)} . \frac{4c_1c_2}{z^2} e^{-\frac{4c_1c_2}{z}} dz.
\end{aligned}$$

which does not depend on the non-observable variable $N_n(\theta_n)$. Observe that our sampling formula does not make a statement on allele frequencies, but on block frequencies of active and inactive blocks at time θ_n.

The variables O_i and R_i can be inferred if we are capable of deciding if a present individual has visited the seed bank or not. This seems hard, deactivation can be treated similarly as mutations form the mathematical point of view, but, as opposed to mutations, they don't leave tractable evidence. Furthermore, our result presents a snapshot of the partition at a random time, not at a deterministic one. For these reasons it seems too optimistic to believe that this study provides a possible method of estimating the parameters of the seed bank model.

From (3.33), we obtain the probability generating function of the old and recent blocks.

Corollary 3.5.3. *Let $O_1, \ldots O_n, R_1, \ldots, R_n$ be random variables with joint density given by (3.33). Then, their (conditional) probability generating function is*

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^n t_i^{O_i} \prod_{j=1}^n s_j^{R_j} | N_n(\theta_n)\right] &= \frac{(n - N_n(\theta_n))! N_n(\theta_n)!}{(N_n(\theta_n) + 2c_1)_{(n - N_n(\theta_n))}} \times \\
&\sum_{a_1, \ldots, a_n, b_1, \ldots, b_n \in A(N_n(\theta_n), n)} \prod_{i=1}^n \frac{(t_i)^{a_i}}{a_i!} \prod_{j=1}^n \frac{1}{b_j!} \left(\frac{2c_1 s_j}{j}\right)^{b_j}. \qquad (3.34)
\end{aligned}$$

Following the idea of Watterson [70], we use two artificial variables, $u \in (-1, 1)$ and $v \in (-1, 1)$. They will help us to rewrite (3.34) in a simpler way. First, observe that for $(a_i, b_i) \in A(k, n)$,

$$\prod_{i=1}^n (uv^i)^{a_i} \prod_{j=1}^n (v^j)^{b_j} = u^{\sum_{i=1}^n a_i} v^{\sum_{i=1}^n i(a_i + b_i)} = u^k v^n.$$

Now, let $c_{k,n}$ be the multiplying coefficient of $u^k v^n$ in $\exp\left\{\sum_{i=1}^{n} uv^i t_i + \sum_{j=1}^{\infty} \frac{2c_1}{j} s_j v^j\right\}$. We can rewrite (3.34) as

$$\mathbb{E}\left[\prod_{i=1}^{n} t_i^{O_i} \prod_{j=1}^{n} s_j^{R_j} | N_n(\theta_n)\right] = \frac{(n - N_n(\theta_n))! N_n(\theta_n)!}{(N_n(\theta_n) + 2c_1)_{(n-N_n(\theta_n))}} c_{N_n(\theta_n),n}. \tag{3.35}$$

From this relation, we obtain the probability generating function of the lineages that have not gone through the seed bank at time θ_n.

Corollary 3.5.4. *Let O_i be the number of "old" blocks of size i (i.e. active blocks of size i at time θ_n). Then, the joint probability generating function of $O_1, O_2, ..., O_n$ is*

$$\mathbb{E}\left[\prod_{i=1}^{n} t_i^{O_i} | N_n(\theta_n)\right] = \sum_{a_1,..,a_n \in \bar{A}(N_n(\theta_n),n)} \frac{N_n(\theta_n)!}{a_1! a_2! ... a_n!} t_1^{a_1} t_2^{a_2} \cdot \cdots t_n^{a_n} \frac{\binom{2c_1+n-z-1}{n-z}}{\binom{2c_1+n-1}{n-N_n\theta_n}} \tag{3.36}$$

where $z = \sum_{i=1}^{n} i a_i$.

Proof. First, we will write explicitly the term $c_{k,n}$ when $s_j = 1$ for all j. Observe that,

$$\begin{aligned}\exp\left\{\sum_{i=1}^{n} uv^i t_i + \sum_{j=1}^{\infty} \frac{2c_1}{j} v^j\right\} &= (1-v)^{-2c_1} \exp\left\{u \sum_{i=1}^{n} v^i t_i\right\} \\ &= (1-v)^{-2c_1} \sum_{k=0}^{\infty} \frac{\left[u \sum_{i=1}^{n} v^i t_i\right]^k}{k!}.\end{aligned}$$

It implies that the coefficient of u^k in the latter expression is

$$\frac{\left[\sum_{i=1}^{n} (v^i t_i)\right]^k}{k!} (1-v)^{-2c_1} = \frac{\left[\sum_{i=1}^{n} (v^i t_i)\right]^k}{k!} \left(\sum_{j=0}^{\infty} \binom{2c_1 + j - 1}{j} v^j\right).$$

Now, we need to find the coefficient of v^n in the latter expression. First, observe that

$$\left[\sum_{i=1}^{n} (v^i t_i)\right]^k = \sum_{a_1 + ... + a_n = k} \frac{k!}{a_1! a_2! ... a_n!} t_1^{a_1} t_2^{a_2} \cdot \cdots t_n^{a_n} v^z$$

where $z = \sum_{i=1}^{n} i a_i$. For $z \leq n$, the coefficient of v^{n-z} in the expression

$$\left(\sum_{j=0}^{\infty} \binom{2c_1 + j - 1}{j} v^j\right)$$

is $\binom{2c_1+n-z-1}{n-z}$. So,

$$c_{k,n} = \frac{1}{k!} \sum_{a_1,\ldots,a_n \in \bar{A}(k,n)} \frac{k!}{a_1!a_2!\ldots a_n!} t_1^{a_1} t_2^{a_2} \cdot \cdots \cdot t_n^{a_n} \binom{2c_1+n-z-1}{n-z}.$$

Thus, replacing $c_{N_n(\theta_n),n}$ and $s_j = 1$ for all j in (3.35) we have the result. □

From the previous corollary we obtain the joint distribution of the lineages which have not gone through the seed bank at time θ_n.

$$\mathbb{P}\left[O_1 = a_1, \ldots, O_n = a_n | N_n(\theta_n)\right] \overset{a.s.}{=} \frac{N_n(\theta_n)!}{a_1!a_2! \cdots a_n!} \frac{\binom{2c_1+n-z-1}{n-z}}{\binom{2c_1+n-1}{n-N_n(\theta_n)}}$$

when $a_1, \ldots, a_n \in \bar{A}(N_n(\theta_n), n)$.

Now, by taking $t_i = t^i$ and $s_j = 1$ for all $i, j \in [n]$ in (3.35), and finding the corresponding coefficient $c_{N_n(\theta_n),n}$, we obtain the conditional probability generating function of the number of lineages at time zero that has not been through the seed bank until time θ_n

$$\mathbb{E}\left[t^{\sum_{i=1}^n iO_i} | N_n(\theta_n)\right] = \sum_{z=N_n(\theta_n)}^{n} t^z \frac{\binom{2c_1+n-z-1}{n-z}\binom{z-1}{z-N_n(\theta_n)}}{\binom{2c_1+n-1}{n-N_n(\theta_n)}}.$$

Finally, from (3.35), by taking $t_i = 1$ for all $i \in [n]$, and from (3.36) we can find the conditional expectations of O_j and R_j for all $j = 1, 2, \ldots n - N_n(\theta_n)$,

$$\mathbb{E}\left(O_j | N_n(\theta_n)\right) = N_n(\theta_n) \frac{\binom{2c_1+n-j-1}{n-j-N_n(\theta_n)+1}}{\binom{2c_1+n-1}{n-N_n(\theta_n)}}$$

and

$$\mathbb{E}\left(R_j | N_n(\theta_n)\right) = \frac{2c_1}{j} \frac{\binom{2c_1+n-j-1}{n-j-N_n(\theta_n)}}{\binom{2c_1+n-1}{n-N_n(\theta_n)}}.$$

Bibliography

[1] `http://www.calflora.net/bloomingplants/parryslinanthus.html`.

[2] Berestycki, N.: *Recent progress in coalescent theory.* Ensaios Matematicos, 16, 2009.

[3] Bertoin, J.: *Random Fragmentation and Coagulation Processes.* Cambridge University Press, 2006, ISBN 0521867282.

[4] Bertoin, J. and Le Gall J.-F.: *Stochastic flows associated to coalescent processes.* Probability Theory and Related Fields, 126(2):261–288, 2003.

[5] Billingsley, P.: *Convergence of Probability Measures.* Wiley, New York, second edition edition, 1999, ISBN 978-0-471-19745-4.

[6] Birkner, M, Liu H., and Sturm A.: *Coalescent results for diploid exchangeable population models.* Electronic Journal of Probability, 23, 2018.

[7] Birkner, M., Blath J., Möhle M., Steinrücken M., and Tams J.: *A modified lookdown construction for the xi-fleming-viot process with mutation and populations with recurrent bottleneck.* ALEA, Latin America Journal of Probability and Mathematical Statistics, 6:25–61, 2009.

[8] Blath, J., González Casanova A., Kurt N., and Spanò D.: *The ancestral process of long-range seed bank models.* Journal of Applied Probability, 50(3):741–759, 2013.

[9] Blath, J., González Casanova A., Kurt N., and Wilke Berenguer M.: *A new coalescent for seed-bank models.* The Annals of Applied Probability, 26(2):857–891, 2016.

[10] Blath, J., González Casanova A., Kurt N., and Wilke Berenguer M.: *The seed bank coalescent with simultaneous switching.* Electronic Journal of Probability, 25(0), 2020.

[11] Blath, J., Eldon B., González Casanova A., and Kurt N.: *Genealogy of a wright-fisher model with strong seedbank component.* In *XI Symposium on Probability and Stochastic Processes*, pages 81–100. Springer, 2015.

[12] Blath, J., Buzzoni E., González Casanova A., and Wilke Berenguer M.: *Structural properties of the seed bank and the two island diffusion.* Journal of Mathematical Biology, 79(1):369–392, 2019.

[13] Blath, J., Buzzoni E., Koskela J., and Wilke Berenguer M.: *Statistical tools for seed bank detection.* Theoretical Population Biology, 132:1–15, 2020.

[14] Cannings, C.: *The latent roots of certain markov chains arising in genetics: A new approach, I. haploid models.* Advances in Applied Probability, 6(2):260–290, 1974.

[15] Cannings, C.: *The latent roots of certain markov chains arising in genetics: A new approach, II. further haploid models.* Advances in Applied Probability, 7(2):264–282, 1975.

[16] Coyne, J. A.: *Lack of genic similarity between two sibling species of drosophila as revealed by varied techniques.* Genetics, 84(3):593–607, 1976, ISSN 0016-6731. `https://www.genetics.org/content/84/3/593`.

[17] Delmas, J. F., Dhersin J.-S., and Siri–Jégousse A.: *Asymptotic results on the length of coalescent trees.* The Annals of Applied Probability, 18(3):997–1025, 2008.

[18] Dhersin, J. S., Freund F., Siri–Jégousse A., and Yuan L.: *On the length of an external branch in the beta-coalescent.* Stochastic Processes and their Applications, 123(5):1691–1715, 2013.

[19] Diehl, C. S. and Kersting G.: *Tree lengths for general Λ-coalescents and the asymptotic site frequency spectrum around the bolthausen–sznitman coalescent.* The Annals of Applied Probability, 29(5):2700–2743, 2019.

[20] Donnelly, P. and Kurtz T. G.: *A countable representation of the fleming-viot measure-valued diffusion.* The Annals of Probability, 24(2):698–742, 1996.

[21] Donnelly, P. and Kurtz T. G.: *Genealogical processes for fleming-viot models with selection and recombination.* The Annals of Applied Probability, 9(4):1091–1148, 1999.

[22] Donnelly, P. and Kurtz T. G.: *Particle representations for measure-valued population models.* The Annals of Probability, 27(1):166–205, 1999.

[23] Donnelly, P. and Tavaré S.: *The ages of alleles and a coalescent.* Advances in Applied Probability, 18(1):1–19, 1986.

[24] Drmota, M., Iksanov A., Möhle M., and Roesler U.: *Asymptotic results concerning the total branch length of the bolthausen–sznitman coalescent*. Stochastic Processes and their Applications, 117(10):1404–1421, 2007.

[25] Durrett, R.: *Probability Models for DNA Sequence Evolution*. Springer New York, 2008.

[26] Epling, C., Lewis H., and Ball F. M.: *The breeding group and seed storage: A study in population dynamics*. Evolution, 14(2):238–255, 1960. `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.1960.tb03082.x`.

[27] Etheridge, A.: *Some Mathematical Models from Population Genetics: École D'Été de Probabilités de Saint-Flour XXXIX-2009*. Number v. 2012 in *Lecture Notes in Mathematics*. Springer, 2011, ISBN 9783642166310.

[28] Etheridge, A.M. and Griffiths R.C.: *A coalescent dual process in a moran model with genic selection*. Theoretical Population Biology, 75(4):320–330, 2009.

[29] Ethier, S. N and Kurtz T. G.: *Markov processes: characterization and convergence*. Wiley, New York ; Chichester, 1986, ISBN 0471081868.

[30] Ewens, W.J.: *The sampling theory of selectively neutral alleles*. Theoretical Population Biology, 3(1):87–112, mar 1972.

[31] Freund, F.: *Cannings models, population size changes and multiple-merger coalescents*. Journal of mathematical biology, 80(5):1497–1521, 2020.

[32] Freund, F. and Siri–Jégousse A.: *The impact of genetic diversity statistics on model selection between coalescents*. Computational Statistics & Data Analysis, 156:107055, 2021.

[33] Freund, F. and Siri–Jégousse A.: *The minimal observable clade size of exchangeable coalescents*. Brazilian Journal of Probability and Statistics, 35(2):281–292, 2021.

[34] González Casanova, A. and Spanò D.: *Duality and fixation in Ξ-wright–fisher processes with frequency-dependent selection*. The Annals of Applied Probability, 28(1):250–284, 2018.

[35] González Casanova, A., Lizbeth Peñaloza, and Arno Siri-Jégousse: *The shape of a seed bank tree*. arXiv preprint arXiv:2001.04500. To appear in Journal of Applied Probability, 2020.

[36] González Casanova, A., Miró Pina V., and Siri–Jégousse A.: *The symmetric coalescent and wright-fisher models with bottlenecks*. arXiv preprint arXiv:1903.05642. To appear in The Annals of Applied Probability, 2019.

[37] González Casanova, A., Aguirre von Wobeser E., Espín G., Servín González L., Kurt N., Spanò D., Blath J., and Soberón Chávez G.: *Strong seed-bank effects in bacterial evolution.* Journal of Theoretical Biology, 356:62–70, 2014.

[38] Griffiths, R. C. and Spanó D.: *Diffusion processes and coalescent trees.* In Bingham, N. H. and C. M. Goldie (editors): *Probability and Mathematical Genetics*, pages 358–379. Cambridge University Press, 2010.

[39] Hobolth, A., Siri–Jégousse A., and Bladt M.: *Phase-type distributions in population genetics.* Theoretical Population Biology, 127:16–32, 2019.

[40] Jansen, S. and Kurt N.: *On the notion(s) of duality for markov processes.* Probability Surveys, 11(0):59–120, 2014.

[41] Kaj, I., Krone S. M., and Lascoux M.: *Coalescent theory for seed bank models.* Journal of Applied Probability, 38(2):285–300, 2001.

[42] Kersting, G.: *The asymptotic distribution of the length of beta-coalescent trees.* The Annals of Applied Probability, 22(5), 2012.

[43] Kimura, M.: *Diffusion models in population genetics.* Journal of Applied Probability, 1(2):177–232, 1964.

[44] Kingman, J. F. C.: *On the genealogy of large populations.* Journal of Applied Probability, 19(A):27–43, 1982.

[45] Kingman, J.F.C.: *The coalescent.* Stochastic Processes and their Applications, 13(3):235–248, 1982.

[46] Koopmann, B., Müller J., Tellier A., and Živković D.: *Fisher–wright model with deterministic seed bank and selection.* Theoretical Population Biology, 114:29–39, 2017.

[47] Krebs, J. E., Goldstein E. S., and Kilpatrick S. T: *Lewin's genes XII.* Jones & Bartlett Learning, 2018.

[48] Lambert, A.: *Population dynamics and random genealogies.* Stochastic Models, 24(sup1):45–163, 2008.

[49] Lambert, A. and Ma C.: *The coalescent in peripatric metapopulations.* Journal of Applied Probability, 52(2):538–557, 2015.

[50] Lennon, J. T. and Jones S. E.: *Microbial seed banks: the ecological and evolutionary implications of dormancy.* Nature Reviews Microbiology, 9(2):119–130, 2011.

[51] Lennon, J. T. and Jones S. E.: *Microbial seed banks: the ecological and evolutionary implications of dormancy.* Nature Reviews Microbiology, 9(2):119–130, 2011.

[52] Levin, D. A., Peres Y., and Wilmer E. L.: *Markov chains and mixing times.* `https://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf`, Missing.

[53] Liggett, T. M.: *Interacting Particle Systems.* Springer New York, 1985.

[54] Maughan, H.: *Rates of molecular evolution in bacteria are relatively constant despite spore dormancy.* Evolution, 61(2):280–288, 2007.

[55] Möhle, M.: *The concept of duality and applications to markov processes arising in neutral population genetics models.* Bernoulli, 5(5):761–777, 1999. `https://projecteuclid.org/euclid.bj/1171290398`.

[56] Möhle, M.: *Ancestral processes in population genetics—the coalescent.* Journal of Theoretical Biology, 204(4):629–638, 2000.

[57] Möhle, M.: *Total variation distances and rates of convergence for ancestral coalescent processes in exchangeable population models.* Advances in Applied Probability., 32(4):983–993, 2000.

[58] Möhle, M.: *Forward and backward diffusion approximations for haploid exchangeable population models.* Stochastic Processes and their Applications, 95(1):133–149, 2001.

[59] Möhle, M. and Sagitov S.: *A characterization of ancestral limit processes arising in haploid population genetics models.* Preprint. Johannes Gutenberg-Universität, Mainz, 1998.

[60] Möhle, M. and Sagitov S.: *A classification of coalescent processes for haploid exchangeable population models.* The Annals of Probability, 29(4):1547–1562, 2001.

[61] Möhle, M. and Sagitov S.: *Coalescent patterns in diploid exchangeable population models.* Journal of Mathematical Biology, 47(4):337–352, 2003.

[62] Pitman, J.: *Coalescents with multiple collisions.* The Annals of Probability, 27(4):1870–1902, 1999, ISSN 00911798. `http://www.jstor.org/stable/2652847`.

[63] Sagitov, S.: *The general coalescent with asynchronous mergers of ancestral lines.* Journal of Applied Probability, 36(4):1116–1125, 1999.

[64] Schweinsberg, J.: *Coalescents with simultaneous multiple collisions.* Electronic Journal of Probability, 5(12):50, 2000.

[65] Schweinsberg, J.: *A necessary and sufficient condition for the Λ-coalescent to come down from infinity.* Electronic Communications in Probability, 5(1):1–11, 2000.

[66] Schweinsberg, J.: *Coalescent processes obtained from supercritical galton–watson processes.* Stochastic Processes and their Applications, 106(1):107–139, 2003.

[67] Shoemaker, W. R. and Lennon J. T.: *Evolution with a seed bank: The population genetic consequences of microbial dormancy.* Evolutionary Applications, 11(1):60–75, 2018.

[68] Tellier, A., Laurent S. J. Y., Lainer H., Pavlidis P., and Stephan W.: *Inference of seed bank parameters in two wild tomato species using ecological and genetic data.* Proceedings of the National Academy of Sciences, 108(41):17052–17057, 2011.

[69] Underhill, P. A., Jin L., Lin A. A., Mehdi S. Q., Jenkins T., Vollrath D., Davis R. W., Cavalli Sforza L. L., and Oefner P. J.: *Detection of numerous y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography.* Genome Research, 7(10):996–1005, 1997.

[70] Watterson, G.A.: *Lines of descent and the coalescent.* Theoretical Population Biology, 26(1):77–92, 1984.

[71] Živković, D. and Tellier A.: *Germ banks affect the inference of past demographic events.* Molecular Ecology, 21(22):5434–5446, 2012.