

THE COALESCENT  
Lectures given at CIMPA–school  
St Louis, Sénégal, April 2010

Étienne Pardoux



# Contents

<b>1</b>	<b>Kingman’s coalescent</b>	<b>5</b>
1.1	The Wright–Fisher model . . . . .	5
1.2	Cannings’ model . . . . .	5
1.3	Looking backward in time . . . . .	6
1.4	Kingman’s coalescent . . . . .	7
1.5	The height and the length of Kingman’s coalescent . . . . .	11
1.6	The speed of cdi . . . . .	13
1.6.1	Proof of the strong law of large numbers . . . . .	14
1.6.2	Proof of the central limit theorem . . . . .	16
<b>2</b>	<b>Wright–Fisher</b>	<b>19</b>
2.1	The simplest Wright–Fisher model . . . . .	19
2.2	Wright–Fisher model with mutations . . . . .	23
2.3	Wright–Fisher model with selection . . . . .	24
2.4	Duality . . . . .	25
<b>3</b>	<b>Look–down</b>	<b>29</b>
3.1	Introduction . . . . .	29
3.2	The Moran model . . . . .	29
3.3	The look–down construction . . . . .	33
3.4	a. s. convergence . . . . .	35
<b>4</b>	<b>Infinitely many alleles</b>	<b>41</b>
4.1	Hoppe’s urn . . . . .	41
4.2	Ewens’ sampling formula . . . . .	46
<b>5</b>	<b>Infinitely many sites</b>	<b>51</b>
5.1	The number of segregating sites . . . . .	51

5.2	Pairwise mismatches . . . . .	52
5.3	Tajima's $D$ test statistics . . . . .	54
5.4	Two final remarks . . . . .	55
<b>6</b>	<b>Appendix</b>	<b>57</b>
6.1	Some elements of stochastic calculus . . . . .	57
6.2	Tightness in $D$ . . . . .	60
6.2.1	The space $D$ . . . . .	60
6.2.2	Compactness criterion in $D([0, \infty); \mathbb{R}^d)$ . . . . .	61
6.3	de Finetti's theorem . . . . .	62

# Chapter 1

## Kingman's coalescent

### 1.1 The Wright–Fisher model

Consider a population of fixed size  $N$ , which evolves in discrete generations. Each individual of generation  $k$  chooses his father uniformly among the individuals of the previous generation, independently of the choices of the other individuals.

Looking backward in time, if we sample  $n$  individuals in the present population, we want to describe at which generation any two of those had the same common ancestor, until we reach the most recent common ancestor of the sample.

### 1.2 Cannings' model

We can generalize the Wright–Fisher model as follows. Suppose at each generation, we label the  $N$  individuals randomly. For  $r \geq 0$ ,  $1 \leq i \leq N$ , let  $\nu_i^r$  denote the number of offsprings in generation  $r + 1$  of the  $i$ -th individual from generation  $r$ . Clearly those r. v.'s must satisfy the requirement that

$$\nu_1^r + \cdots + \nu_N^r = N.$$

Cannings' model stipulates moreover that

$$\nu^r, r \geq 0 \text{ are i. i. d. copies of } \nu,$$

and that the law of  $\nu$  is exchangeable, i. e.

$$(\nu_1, \dots, \nu_N) \simeq (\nu_{\pi(1)}, \dots, \nu_{\pi(N)}), \forall \pi \in S_N.$$

The above conditions imply that  $\mathbb{E}\nu_1 = 1$ . To avoid the trivial case where  $\mathbb{P}(\nu_1 = \dots = \nu_N = 1) = 1$ , we assume that  $\text{Var}(\nu_1) > 0$ . A particular case of Cannings' model is the Wright–Fisher model, in which  $\nu$  is multinomial.

### 1.3 Looking backward in time

Consider a population of fixed size  $N$ , which has been reproducing for ever according to Cannings' model. We sample  $n < N$  individuals from the present generation, and label them  $1, 2, \dots, n$ . For each  $r \geq 0$ , we introduce the equivalence relation on the set  $\{1, \dots, n\}$ :  $i \sim_r j$  if the individuals  $i$  and  $j$  have the same ancestor  $r$  generations back in the past. Denote this equivalence relation by  $R_r^{N,n}$ . For  $r \geq 0$ ,  $R_r^{N,n}$  is a random equivalence relation, which can be described by its associated equivalence classes, which is a random partition of  $(1, \dots, n)$ . Thus  $\{R_r^{N,n}; r \geq 0\}$  is a Markov chain with values in the set  $\mathcal{E}_n$  of the partitions of  $(1, \dots, n)$ , which starts from the trivial *finest* partition  $(\{1\}, \dots, \{n\})$ , and eventually reaches the *coarsest* partition consisting of the set  $\{1, \dots, n\}$  alone. We denote by  $P_{\xi, \eta}^{N,n}$  the transition matrix of that chain.

The probability that two individuals in today's population have the same ancestor in the previous generation is

$$c_N = \frac{\sum_{i=1}^N \mathbb{E} \left[ \binom{\nu_i}{2} \right]}{\binom{N}{2}} = \frac{\sum_{i=1}^N \mathbb{E}[\nu_i(\nu_i - 1)]}{N(N-1)} = \frac{\mathbb{E}[\nu_1(\nu_1 - 1)]}{N-1}.$$

Provided that  $c_N \rightarrow 0$  as  $N \rightarrow \infty$ , if  $r = t/c_N$ ,

$$\mathbb{P}(1 \not\sim_r 2) = (1 - c_N)^r \approx e^{-t}.$$

This suggests to consider

$$\mathcal{R}_t^{N,n} := R_{\lfloor t/c_N \rfloor}^{N,n}, \quad t \geq 0.$$

## 1.4 Kingman's coalescent

Let  $\{\mathcal{R}_t^n; t \geq 0\}$  be a continuous time  $\mathcal{E}_n$ -valued jump Markov process with the rate matrix given by (for  $\eta \neq \xi$ )

$$Q_{\xi\eta} = \begin{cases} 1 & \text{, if } \eta \text{ is obtained from } \xi \text{ by merging exactly two classes,} \\ 0 & \text{, otherwise.} \end{cases} \quad (1.4.1)$$

This is Kingman's  $n$  coalescent. In order for  $\mathcal{R}^{N,n}$  to converge to Kingman's coalescent, we certainly need that merges of 3 or more lineages are asymptotically negligible. The probability that three individuals in today's population have the same ancestor in the previous generation is

$$d_N := \frac{\sum_{i=1}^N \mathbb{E} \left[ \binom{\nu_i}{3} \right]}{\binom{N}{3}} = \frac{\mathbb{E}[\nu_1(\nu_1 - 1)(\nu_1 - 2)]}{(N - 1)(N - 2)}.$$

**Exercise 1.4.1.** Compute  $c_N$  and  $d_N$  in the Wright–Fisher model, as well as in the model where at each generation a common father of all individuals of the next generation is chosen uniformly in the present generation.

**Theorem 1.4.2.**  $\mathcal{R}^{N,n} \Rightarrow \mathcal{R}^n$  in  $D(\mathbb{R}_+; \mathcal{E}_n)$  iff, as  $N \rightarrow \infty$ , both

$$\begin{cases} c_N \rightarrow 0, \\ \frac{d_N}{c_N} \rightarrow 0. \end{cases} \quad (1.4.2)$$

**Remark 1.4.3. Non-constant population size** *This result assumes in an essential way that the size of the population is constant in time. What is the effect of modifying the population size? Assume (that is true in particular for the Wright–Fisher model) that  $\mathbb{E}[\nu_1(\nu_1 - 1)] \rightarrow c > 0$  as  $N \rightarrow \infty$ . In that case our theorem says roughly that for large  $N$ ,  $R_{Nt/c}^{N,n} \simeq \mathcal{R}_t^n$ . Then for any  $x > 0$ , we have similarly that  $R_{xNt/c}^{xN,n} \simeq \mathcal{R}_t^n$ . In other words,  $R_{Nt/c}^{xN,n} \simeq \mathcal{R}_{t/x}^n$ . This means that if we multiply the size of the population by a factor  $x$ , we should accelerate time by a factor  $1/x$ , or, what is exactly the same, multiply the pairwise coalescence rate by the factor  $1/x$ . This argument can be justified in the case of a varying population size. The rule is to multiply at each time  $t$  the pairwise coalescence rate by 1 over the “renormalized population size”.*

PROOF: The sufficiency will follow from the standard Lemma 1.4.4 below and the fact that (1.4.2) implies that

$$P_{\xi,\eta}^{N,n} = \delta_{\xi,\eta} + c_N Q_{\xi,\eta} + o(c_N),$$

where the error term is small, uniformly with respect to  $\xi, \eta \in \mathcal{E}_n$ . It follows from exchangeability that for any  $f : \{0, 1, \dots, N\} \rightarrow \mathbb{R}_+$ ,

$$\begin{aligned} (N-1)\mathbb{E}[\nu_2 f(\nu_1)] &= \sum_{j=2}^N \mathbb{E}[\nu_j f(\nu_1)] \\ &= \mathbb{E}[(N-\nu_1)f(\nu_1)] \\ &\leq N\mathbb{E}[f(\nu_1)], \end{aligned}$$

hence

$$\mathbb{E}[\nu_2 f(\nu_1)] \leq \frac{N}{N-1} \mathbb{E}[f(\nu_1)]. \quad (1.4.3)$$

From the Markov inequality and (1.4.2), with the notations  $(\nu)_2 = \nu(\nu-1)$ ,  $(\nu)_3 = \nu(\nu-1)(\nu-2)$ , if  $\varepsilon N \geq 2$ ,

$$\begin{aligned} \mathbb{P}(\nu_1 > \varepsilon N) &\leq \frac{\mathbb{E}[(\nu_1)_3]}{(\varepsilon N)_3} \\ &= \frac{o(N\mathbb{E}[(\nu_1)_2])}{\varepsilon^3 N^3}, \end{aligned}$$

consequently

$$\mathbb{P}(\nu_1 > \varepsilon N) \leq \varepsilon^{-3} o(c_N/N). \quad (1.4.4)$$

Next

$$\begin{aligned} \mathbb{E}[(\nu_1)_2(\nu_2)_2] &\leq \varepsilon N \mathbb{E}[(\nu_1)_2 \nu_2; \nu_2 \leq \varepsilon N] + N^2 \mathbb{E}[(\nu_1)_2; \nu_2 > \varepsilon N] \\ &\leq \varepsilon N \mathbb{E}[(\nu_1)_2 \nu_2] + N^3 \mathbb{E}[\nu_1; \nu_2 > \varepsilon N] \\ &\leq \varepsilon N \frac{N}{N-1} \mathbb{E}[(\nu_1)_2] + N^3 \frac{N}{N-1} \mathbb{P}(\nu_2 > \varepsilon N), \end{aligned}$$

where we have used (1.4.3) twice in the last inequality. Combining this with (1.4.4), we conclude that for all  $\varepsilon > 0$ ,

$$\begin{aligned} \limsup_{N \rightarrow \infty} \frac{\mathbb{E}[(\nu_1)_2(\nu_2)_2]}{N\mathbb{E}[(\nu_1)_2]} &\leq \varepsilon + \limsup_{N \rightarrow \infty} \frac{\mathbb{P}(\nu_1 > \varepsilon N)}{c_N/N} \\ &= \varepsilon. \end{aligned}$$

Let  $I_1, \dots, I_n$  denote the parents of  $n$  ordered randomly chosen individuals of a given generation. We have the following identities

$$\begin{aligned} \mathbb{P}(I_1 = I_2) &= c_N \\ \mathbb{P}(I_1 = I_2 = I_3) &= d_N \\ \mathbb{P}(I_1 = I_2 \neq I_3 = I_4) &= \frac{\sum_{1 \leq i < j \leq N} \mathbb{E} \left[ \binom{\nu_i}{2} \binom{\nu_j}{2} \right]}{\binom{N}{4}} \\ &= 3 \frac{\mathbb{E}[(\nu_1)_2(\nu_2)_2]}{(N-2)(N-3)}. \end{aligned}$$

Hence we deduce from the last estimate that

$$\lim_{N \rightarrow \infty} \frac{\mathbb{P}(I_1 = I_2 \neq I_3 = I_4)}{\mathbb{P}(I_1 = I_2)} = 0, \quad (1.4.5)$$

while (1.4.2) tells us that

$$\lim_{N \rightarrow \infty} \frac{\mathbb{P}(I_1 = I_2 = I_3)}{\mathbb{P}(I_1 = I_2)} = 0. \quad (1.4.6)$$

We now conclude, using (1.4.5) and (1.4.6). Let  $\xi = (C_{11}, C_{12}, C_2, \dots, C_a)$  and  $\eta = (C_1, C_2, \dots, C_a)$ , where  $C_1 = C_{11} \cup C_{12}$ . We have

$$\begin{aligned} &\mathbb{P}(I_1 = I_2) - \mathbb{P}(\{I_1 = I_2\} \cap \{\exists 3 \leq m \leq a+1; I_m = I_1\}) \\ &\quad - \mathbb{P}(\{I_1 = I_2\} \cap \{\exists 3 \leq \ell < m \leq a+1; I_\ell = I_m \neq I_1\}) \\ &\leq P_{\xi, \eta}^{N, n} \leq \mathbb{P}(I_1 = I_2). \end{aligned}$$

From (1.4.6),

$$\begin{aligned} \mathbb{P}(\{I_1 = I_2\} \cap \{\exists 3 \leq m \leq a+1; I_m = I_1\}) &\leq (a-1)\mathbb{P}(I_1 = I_2 = I_3) \\ &= o(\mathbb{P}(I_1 = I_2)), \end{aligned}$$

and from (1.4.5),

$$\begin{aligned} \mathbb{P}(\{I_1 = I_2\} \cap \{\exists 3 \leq \ell < m \leq a+1; I_\ell = I_m \neq I_1\}) &\leq \binom{a-1}{2} \mathbb{P}(I_1 = I_2 \neq I_3 = I_4) \\ &= o(\mathbb{P}(I_1 = I_2)) \end{aligned}$$

We have proved that for such a pair  $(\xi, \eta)$ ,  $P_{\xi, \eta}^{N, n} = c_N + o(c_N)$ . If  $\eta'$  is obtained from  $\xi$  by merging more than two classes, then there must be at least either a triple merger or two double mergers, hence from (1.4.6), (1.4.5),  $P_{\xi, \eta'}^{N, n} = o(c_N)$ . Finally, since  $|\mathcal{E}_n| < \infty$  and  $\sum_{\eta \in \mathcal{E}_n} P_{\xi, \eta}^{N, n} = 1$ ,

$$\begin{aligned} P_{\xi, \xi}^{N, n} &= 1 - \binom{|\xi|}{2} c_N + o(c_N) \\ &= 1 + Q_{\xi, \xi} c_N + o(c_N). \end{aligned}$$

□

It remains to prove :

**Lemma 1.4.4.** *Let  $E$  be a finite set and  $\{X_t, t \geq 0\}$  a continuous time  $E$ -valued jump Markov process, with generator  $Q = (Q_{x, y})_{x, y \in E}$ . Let for each  $N \in \mathbb{N}$   $X^N$  be a discrete time Markov chain with transition matrix satisfying*

$$P_{x, y}^N = \delta_{x, y} + c_N Q_{x, y} + o(c_N), \quad x, y \in E,$$

where  $c_N \rightarrow 0$ , as  $N \rightarrow \infty$ . Then whenever  $X_0^N \Rightarrow X_0$ ,

$$\{X_{[t/c_N]}^N, t \geq 0\} \Rightarrow \{X_t, t \geq 0\} \quad \text{in } D([0, +\infty); E).$$

PROOF: The fact that for any  $x, y \in E, s, t > 0$ ,

$$\mathbb{P}(X_{[(t+s)/c_N]}^N = y | X_{[t/c_N]}^N = x) \rightarrow \mathbb{P}(X_{t+s} = y | X_t = x),$$

together with the Markov property, implies the convergence of finite dimensional distributions. Indeed this follows easily from the fact that

$$\begin{aligned} \mathbb{P}(X_{[(t+s)/c_N]}^N = y | X_{[t/c_N]}^N = x) &= (P^N)_{xy}^{s/c_N} \\ &= (I + c_N Q + o(c_N))_{xy}^{s/c_N} \\ &= (e^{c_N Q} + o(c_N))_{xy}^{s/c_N} \\ &\rightarrow (e^{sQ})_{xy} \end{aligned}$$

It remains to prove tightness in  $D([0, \infty); E)$ . This follows essentially from the fact that the probability that  $X^N$  jumps more than once in an interval of length  $\delta$  is of the order  $o(\delta)$ , uniformly in  $N$ . We skip the details. □

Let  $\{\mathcal{R}_t^n; t \geq 0\}$  start from the trivial partition of  $(1, \dots, n)$ . For  $2 \leq k \leq n$ , let  $T_k$  denote the length of the time interval during which there are

$k$  branches alive. From the Markov property of the coalescent, and the form of the generator, we deduce that

$$\begin{aligned} T_n, T_{n-1}, \dots, T_2 &\text{ are independent,} \\ T_k &\simeq \mathcal{E}_{\text{xp}} \left( \binom{k}{2} \right), \quad 2 \leq k \leq n, \end{aligned}$$

and consequently the expected time till the Most Recent Common Ancestor in the sample is

$$\begin{aligned} \sum_{k=2}^n \frac{2}{k(k-1)} &= 2 \sum_{k=2}^n \left( \frac{1}{k-1} - \frac{1}{k} \right) \\ &= 2 \left( 1 - \frac{1}{n} \right). \end{aligned}$$

For  $n' > n$ , denote by  $d_n$  the restriction to  $\mathcal{E}_n$  of an element of  $\mathcal{E}_{n'}$ . Kingman's  $n$ -coalescents have the consistency property that

$$d_n \left( \{\mathcal{R}_t^{n'}, t \geq 0\} \right) \simeq \{\mathcal{R}_t^n, t \geq 0\}.$$

This, together with the fact that  $\sum_{k \geq 2} T_k < \infty$  a. s., since the series of the expectations converges, allows us to define Kingman's coalescent  $\{\mathcal{R}_t, t \geq 0\}$  as the limit  $\lim_{n \rightarrow \infty} \{\mathcal{R}_t^n, t \geq 0\}$ . It is readily seen that Kingman's coalescent *comes down from infinity*, in the sense that, while  $\mathcal{R}_0$  is the trivial partition of  $\mathbb{N}^*$ , hence  $|\mathcal{R}_0| = \infty$ ,  $|\mathcal{R}_t| < \infty$ ,  $\forall t > 0$ .

## 1.5 The height and the length of Kingman's coalescent

The *height* of Kingman's  $n$ -coalescent is the r. v.

$$H_n = \sum_{k=2}^n T_k,$$

where the  $T_k$  are as above. This prescribes the law of  $H_n$ , which does not obey any simple formula. Note that

$$\mathbb{E}(H_n) = 2 \left( 1 - \frac{1}{n} \right), \quad \text{Var}(H_n) = \sum_{k=2}^n \frac{4}{k^2(k-1)^2}.$$

$\mathbb{E}(H_n) \rightarrow 2$  as  $n \rightarrow \infty$ , and  $\sup_n \text{Var}(H_n) < \infty$ .

The *length* of Kingman's  $n$ -coalescent (i. e. the sum of the lengths of the branches of this tree) is the r. v.

$$L_n = \sum_{k=2}^n kT_k = \sum_{k=2}^n U_k,$$

where the  $U_k$  are independent,  $U_k$  is an  $\text{Exp}((k-1)/2)$  r. v. The distribution function of  $L_n$  is given by

**Proposition 1.5.1.** *For all  $x \geq 0$ ,*

$$\mathbb{P}(L_n \leq x) = (1 - e^{-x/2})^{n-1}.$$

This Proposition follows from the fact that the law of  $L_n$  is that of the sup over  $n-1$  i. i. d.  $\text{Exp}(1/2)$  r. v.'s, which is a consequence of the

**Proposition 1.5.2.** *Let  $V_1, V_2, \dots, V_n$  be i. i. d.  $\text{Exp}(\lambda)$  r. v.'s, and  $V_{(1)} < V_{(2)} < \dots < V_{(n)}$  denote the same random sequence, but arranged in increasing order. Then  $V_{(1)}, V_{(2)} - V_{(1)}, \dots, V_{(n)} - V_{(n-1)}$  are independent exponential r. v.'s with respective parameters  $n\lambda, (n-1)\lambda, \dots, \lambda$ .*

PROOF: For any Borel measurable function  $f : \mathbb{R}_+^n \rightarrow \mathbb{R}_+$ ,

$$\begin{aligned} & \mathbb{E}f(V_{(1)}, V_{(2)} - V_{(1)}, \dots, V_{(n)} - V_{(n-1)}) \\ &= n! \mathbb{E}[f(V_1, V_2 - V_1, \dots, V_n - V_{n-1}); V_1 < V_2 < \dots < V_n] \\ &= n! \int_{0 < x_1 < x_2 < \dots < x_n} f(x_1, x_2 - x_1, \dots, x_n - x_{n-1}) \lambda^n e^{-\lambda \sum_{k=1}^n x_k} dx_1 dx_2 \dots dx_n \\ &= \prod_{k=1}^n (k\lambda) \int_0^\infty \dots \int_0^\infty f(y_1, y_2, \dots, y_n) \prod_{k=1}^n e^{-k\lambda y_{n+1-k}} dy_1 dy_2 \dots dy_n. \end{aligned}$$

The result follows. □

**Exercise 1.5.3.** *A Yule tree of rate  $\lambda$  is a random tree which develops as follows. Let  $T_k, k \geq 1$  be independent r. v.'s,  $T_k$  being exponential with parameter  $\lambda k$ . For  $0 \leq t < T_1$ , the tree has a unique branch issued from the root. At time  $T_1$  this branch splits into 2. For  $T_1 \leq t < T_1 + T_2$ , there are two branches. At time  $T_1 + T_2$ , we choose one of the two branches with equal probability, and that branch splits into 2, etc...*

Deduce from Proposition 1.5.2 that the law of the number  $Y_t$  of branches of the tree at time  $t$  is geometric with parameter  $e^{-\lambda t}$ , in the sense that for  $k \geq 0$ ,

$$\mathbb{P}(Y_t \geq k) = (1 - e^{-\lambda t})^k.$$

## 1.6 The speed at which Kingman's coalescent comes down from infinity

Consider the Kingman coalescent  $\{\mathcal{R}_t, t \geq 0\}$  starting from the trivial partition of  $\mathbb{N}^*$ . Let  $R_t = |\mathcal{R}_t|$ ,  $t \geq 0$ . Let  $\{T_n, n \geq 2\}$  be a sequence of independent r. v.'s, the law of  $T_n$  being the exponential law with parameter  $\binom{n}{2}$ . We let

$$S_n = \sum_{k=n+1}^{\infty} T_k, \quad n \geq 1.$$

Now the process  $\{R_t, t \geq 0\}$  can be represented as follows.

$$R_t = \sum_{n=1}^{\infty} n \mathbf{1}_{\{S_n \leq t < S_{n-1}\}}.$$

We know that  $R_t \rightarrow \infty$ , as  $t \rightarrow 0$ . We state two results, which give a precise information, as to the speed at which  $R_t$  diverges, as  $t \rightarrow 0$ . We first state a strong law of large numbers

**Theorem 1.6.1.** *As  $t \rightarrow 0$ ,*

$$\frac{tR_t}{2} \rightarrow 1 \quad a. s.$$

We next have a central limit theorem

**Theorem 1.6.2.** *As  $t \rightarrow 0$ ,*

$$\sqrt{\frac{6}{t}} \left( \frac{tR_t}{2} - 1 \right) \Rightarrow N(0, 1).$$

**Remark 1.6.3.** *As we will see in the proof, the behaviour of  $R_t$  as  $t \rightarrow 0$  is intimately connected to the behaviour of  $S_n$ , as  $n \rightarrow \infty$ . But while in the classical asymptotic results of probability theory we add more and more random variable as  $n \rightarrow \infty$ , here as  $n$  increases,  $S_n$  is the sum of less and less random variables (but always an infinite number of those).*

### 1.6.1 Proof of the strong law of large numbers

We first need to compute some moments of  $S_n$ .

**Lemma 1.6.4.** *We have*

$$\mathbb{E}(S_n) = \frac{2}{n} \tag{1.6.1}$$

$$\text{Var}(S_n) = \sum_{k=n}^{\infty} \frac{4}{k^2(k+1)^2} \tag{1.6.2}$$

$$\mathbb{E}(|S_n - \mathbb{E}S_n|^4) \leq \frac{c}{n^6}, \tag{1.6.3}$$

where  $c$  is a universal constant. Moreover

$$n^3 \text{Var}(S_n) \rightarrow \frac{4}{3}, \text{ as } n \rightarrow \infty. \tag{1.6.4}$$

PROOF: (1.6.1) follows readily from

$$\begin{aligned} \mathbb{E}(S_n) &= \sum_{k=n+1}^{\infty} \frac{2}{k(k-1)} \\ &= \sum_{k=n}^{\infty} \left( \frac{2}{k} - \frac{2}{k+1} \right) \end{aligned}$$

Similarly

$$\begin{aligned} \text{Var}(S_n) &= \sum_{k=n+1}^{\infty} \text{Var}(T_k) \\ &= \sum_{k=n}^{\infty} \frac{4}{k^2(k+1)^2} \end{aligned}$$

This proves (1.6.2). Now (1.6.4) follows from

$$\begin{aligned} \frac{4}{3n^3} &= \int_n^{\infty} \frac{4}{x^4} dx \leq \sum_n^{\infty} \frac{4}{(k+1)^4} \leq \text{Var}(S_n) \\ &\leq \sum_n^{\infty} \frac{4}{k^4} \leq \int_{n-1}^{\infty} \frac{4}{x^4} dx = \frac{4}{3(n-1)^3}. \end{aligned}$$

We finally prove (1.6.3). Note that  $\mathbb{E}(|T_k - \mathbb{E}T_k|^4) = 2^4/k^4(k-1)^4$ . Moreover

$$\begin{aligned} \mathbb{E}(|S_n - \mathbb{E}S_n|^4) &= \mathbb{E} \sum_{k=n+1}^{\infty} |T_k - \mathbb{E}T_k|^4 + 6\mathbb{E} \sum_{n < k < \ell} |T_k - \mathbb{E}T_k|^2 |T_\ell - \mathbb{E}T_\ell|^2 \\ &= \sum_{k=n}^{\infty} \frac{2^4}{k^4(k+1)^4} + 4 \times 4! \sum_{n \leq k < \ell} \frac{1}{k^2(k+1)^2 \ell^2(\ell+1)^2} \\ &\leq \frac{2^4}{7(n-1)^7} + \frac{4 \times 4!}{3^2(n-1)^6}. \end{aligned}$$

□

Theorem 1.6.1 will follow from

**Proposition 1.6.5.** *As  $n \rightarrow \infty$ ,*

$$\frac{S_n}{\mathbb{E}S_n} \rightarrow 1 \quad a. s.$$

PROOF: The result follows from Borel–Cantelli’s lemma and the next estimate, where we make use of (1.6.3) and (1.6.1)

$$\mathbb{E} \left( \left| \frac{S_n - \mathbb{E}S_n}{\mathbb{E}S_n} \right|^4 \right) \leq c \frac{n^4}{n^6} = cn^{-2}.$$

□

PROOF OF THEOREM 1.6.1 All we need to show is that for all  $\varepsilon > 0$ ,

$$\mathbb{P} \left( \left\{ \limsup_{t \rightarrow 0} \left| \frac{tR_t}{2} - 1 \right| > \varepsilon \right\} \right) = 0.$$

But

$$\left\{ \limsup_{t \rightarrow 0} \left| \frac{tR_t}{2} - 1 \right| > \varepsilon \right\} \subset \limsup_{n \rightarrow \infty} A_n,$$

where

$$A_n = \left\{ \sup_{S_n \leq t < S_{n-1}} \left| \frac{tn}{2} - 1 \right| > \varepsilon \right\}$$

Now

$$\begin{aligned} A_n &\subset \left\{ \left| \frac{nS_n}{2} - 1 \right| > \varepsilon \right\} \cup \left\{ \left| \frac{nS_{n-1}}{2} - 1 \right| > \varepsilon \right\} \\ &\subset \left\{ \left| \frac{nS_n}{2} - 1 \right| > \varepsilon \right\} \cup \left\{ \left| \frac{(n-1)S_{n-1}}{2} - 1 \right| > \varepsilon/2 \right\}, \end{aligned}$$

as soon as  $(\varepsilon + 1)/n \leq \varepsilon/2$ . But it follows from Proposition 1.6.5 that

$$\mathbb{P}(\limsup_n A_n) = 0.$$

□

## 1.6.2 Proof of the central limit theorem

Define for each  $n \geq 1$  the r. v.

$$Z_n = \sqrt{3n} \frac{S_n - \mathbb{E}S_n}{\mathbb{E}S_n}.$$

Let us admit for a moment the

**Proposition 1.6.6.** *As  $n \rightarrow \infty$ ,*

$$Z_n \Rightarrow N(0, 1).$$

PROOF OF THEOREM 1.6.2 Define, for all  $t > 0$ ,

$$\tau(t) = \inf\{0 < s \leq t; R_s = R_t\}.$$

Proposition 1.6.6 tells us that, as  $t \rightarrow 0$ ,

$$\sqrt{3R_t} \left( \frac{\tau(t)R_t}{2} - 1 \right) \Rightarrow N(0, 1).$$

Combining with Theorem 1.6.1, we deduce that

$$\sqrt{\frac{6}{t}} \left( \frac{\tau(t)R_t}{2} - 1 \right) \Rightarrow N(0, 1).$$

It remains to show that

$$\frac{t - \tau(t)}{\sqrt{t}} R_t \rightarrow 0 \quad \text{a. s. as } t \rightarrow 0.$$

From Theorem 1.6.1, this is equivalent to

$$\frac{t - \tau(t)}{t^{3/2}} \rightarrow 0 \quad \text{a. s. as } t \rightarrow 0.$$

But

$$\limsup_{t \rightarrow 0} \frac{t - \tau(t)}{t^{3/2}} \leq \limsup_{n \rightarrow \infty} \frac{T_n}{S_n^{3/2}},$$

and from Proposition 1.6.5, the right hand side goes to zero if and only if  $n^{3/2}T_n \rightarrow 0$  as  $n \rightarrow \infty$ . We have that  $\mathbb{E}(|n^{3/2}T_n|^4) \leq cn^{-2}$ , hence from Bienaymé–Tchebychef and Borel–Cantelli,  $n^{3/2}T_n \rightarrow 0$  a. s. as  $n \rightarrow \infty$ , and the theorem is proved.  $\square$

We finally give the

**PROOF OF PROPOSITION 1.6.6** Let  $\varphi_n$  denote the characteristic function of the r. v.  $Z_n$ . If we let  $c_n = \sqrt{3n}$ ,  $a_n = \sqrt{3n^3}$ , we have  $Z_n = -c_n + a_n S_n/2$ , hence

$$\begin{aligned} \varphi_n(t) &= e^{-itc_n} \prod_{k=n+1}^{\infty} \mathbb{E} [e^{-ita_n T_k/2}] \\ &= e^{-itc_n} \prod_{k=n+1}^{\infty} \left( 1 - \frac{ita_n}{k(k-1)} \right)^{-1} \\ &= e^{-itc_n} \exp \left\{ \sum_{k=n+1}^{\infty} \log \left( 1 + i \frac{ta_n}{k(k-1)} - t^2 \frac{a_n^2}{k^2(k-1)^2} + o(a_n^3 k^{-6}) \right) \right\} \\ &= e^{-itc_n} \exp \left\{ \sum_{k=n+1}^{\infty} \left( i \frac{ta_n}{k(k-1)} - \frac{t^2}{2} \frac{a_n^2}{k^2(k-1)^2} + o(a_n^3 k^{-6}) \right) \right\} \\ &= e^{-itc_n} e^{ita_n/n} \exp \left( -\frac{t^2}{2} \sum_{k=n+1}^{\infty} \frac{3n^3}{k^2(k-1)^2} + o(n^{-1/2}) \right) \\ &\rightarrow \exp(-t^2/2), \end{aligned}$$

where we have used again the argument leading to (1.6.4). The result follows.



# Chapter 2

## The Wright–Fisher diffusion

Consider a population of fixed size  $N$ , which evolves in discrete generations. Assume that each individual can be of two different types ( $a$  and  $A$ , say).

### 2.1 The simplest Wright–Fisher model

Consider first the case where those are neutral, i. e. there is no selective advantage attached to either of those two types, and there is no mutation.

Reproduction is random (and asexual). More precisely, we assume that each individual picks his parent uniformly from the previous generation (with replacement), and copy his type. Denote

$$Y_k^N := \text{number of type } A \text{ individuals in generation } k.$$

Clearly

$$\mathbb{P}(Y_{k+1}^N = i | Y_k^N = j) = C_N^i \left(\frac{j}{N}\right)^i \left(1 - \frac{j}{N}\right)^{N-i}.$$

From this, we see that  $\{Y_k^N, k \geq 0\}$  is both a finite state Markov chain, and a bounded martingale. Note that the two states 0 and  $N$  are absorbing, and all other states are transient. Consequently

$$Y_\infty^N = \lim_{k \rightarrow \infty} Y_k^N \in \{0, N\}.$$

Moreover

$$j = \mathbb{E}[Y_\infty^N | Y_0^N = j] = N\mathbb{P}(Y_\infty^N = N),$$

hence the probability of fixation of type  $A$  is its initial frequency  $j/N$ .

The next question is what can we say about the time we have to wait until the population is homogeneous (i. e.  $Y_k^N = 0$  or  $N$ ) ?

If we want to study this and other questions for large  $N$ , we should understand the behaviour of the proportions of both alleles in a population of infinite size. Define the following continuous time process :

$$X_t^N = N^{-1}Y_{[Nt]}^N, \quad t \geq 0.$$

This means that we consider the fraction of type  $A$ -individuals, and the time is a number of generations divided by the size of the population.

Let  $t \in \mathbb{N}/N$  and  $\Delta t = N^{-1}$ . It is not hard to check that

$$\mathbb{E}[X_{t+\Delta t}^N - X_t^N | X_t^N] = 0, \quad \mathbb{E}[(X_{t+\Delta t}^N - X_t^N)^2 | X_t^N] = X_t^N(1 - X_t^N)\Delta t.$$

We now want to let  $N \rightarrow \infty$ .

**Theorem 2.1.1.** *Suppose that  $X_0^N \Rightarrow X_0$ , as  $N \rightarrow \infty$ . Then  $X^N \Rightarrow X$  in  $D(\mathbb{R}_+; [0, 1])$ , where  $\{X_t, t \geq 0\}$  solves the SDE*

$$dX_t = \sqrt{X_t(1 - X_t)}dB_t, \quad t \geq 0.$$

PROOF: The idea is to prove that  $\forall f \in C^3([0, 1])$ , the process

$$M_t^f := f(X_t) - f(X_0) - \frac{1}{2} \int_0^t X_s(1 - X_s)f''(X_s)ds, \quad t \geq 0 \quad (2.1.1)$$

is a martingale (with respect to its own filtration).

It is known that this martingale problem has a unique solution (the SDE has a unique strong solution, see next chapter). Hence the theorem follows from the two following statements

1. the sequence  $\{X^N, N = 1, 2, \dots\}$  is tight;
2. any weak limit of a sub-sequence solves the above martingale problem.

PROOF OF 1. We shall use the tightness criterion given in Corollary 6.2.3.

For  $0 \leq i < j$ , let  $s = i/N$ ,  $t = j/N$ . We have

$$\begin{aligned}
\mathbb{E} [|X_t^N - X_s^N|^4] &= N^{-4} \mathbb{E} \left( \left[ \sum_{k=i}^{j-1} (Y_{k+1}^N - Y_k^N) \right]^4 \right) \\
&= N^{-4} \mathbb{E} \sum_{k_1, \dots, k_4=i}^{j-1} \prod_{\ell=1}^4 (Y_{k_\ell+1}^N - Y_{k_\ell}^N) \\
&= N^{-4} \left( \mathbb{E} \sum_{k=i}^{j-1} (Y_{k+1}^N - Y_k^N)^4 + 2 \mathbb{E} \sum_{i \leq k_1 < k_2 \leq j-1} (Y_{k_1+1}^N - Y_{k_1}^N)^2 (Y_{k_2+1}^N - Y_{k_2}^N)^2 \right. \\
&\quad + \mathbb{E} \sum_{i \leq k_1 < k_2 \leq j-1} (Y_{k_1+1}^N - Y_{k_1}^N) (Y_{k_2+1}^N - Y_{k_2}^N)^3 \\
&\quad \left. + 2 \mathbb{E} \sum_{i \leq k_1 < k_2 < k_3 \leq j-1} (Y_{k_1+1}^N - Y_{k_1}^N) (Y_{k_2+1}^N - Y_{k_2}^N) (Y_{k_3+1}^N - Y_{k_3}^N)^2 \right) \\
&\leq C \left[ \frac{j-i}{N^2} + \left( \frac{j-i}{N} \right)^2 \right] \\
&= C \left[ \frac{t-s}{N} + (t-s)^2 \right].
\end{aligned}$$

Indeed, we first note that

$$\begin{aligned}
\mathbb{E} [(Y_{k+1}^N - Y_k^N)^2] &= \mathbb{E} \{ \mathbb{E} [(Y_{k+1}^N - Y_k^N)^2 | Y_k^N] \} \\
&\leq N/4,
\end{aligned}$$

from which it follows that the second term above has the right size. Concerning the first term, we note that

$$\mathbb{E} [(Y_{k+1}^N - Y_k^N)^4] = \mathbb{E} \{ \mathbb{E} [(Y_{k+1}^N - Y_k^N)^4 | Y_k^N] \}$$

Conditionally upon  $Y_k^N = y$ ,  $Y_{k+1}^N$  follows the binomial law  $B(N, p)$  where  $p = y/N$ . But if  $Z_1, \dots, Z_n$  are Bernoulli with  $\mathbb{P}(Z_i = 1) = p$ , then

$$\begin{aligned}
\mathbb{E} \left( \left[ \sum_{i=1}^N (Z_i - p) \right]^4 \right) &= \mathbb{E} \sum_{i=1}^N (Z_i - p)^4 + 4 \mathbb{E} \sum_{1 \leq i < j \leq N} (Z_i - p)^2 (Z_j - p)^2 \\
&\leq 2N^2
\end{aligned}$$

Consequently

$$\mathbb{E} [(Y_{k+1}^N - Y_k^N)^4] \leq 2N^2,$$

and the first term is bounded by  $cN^{-1}(t-s) \leq c(t-s)^2$ , since  $1 \leq N(t-s)$ . Moreover we have

$$\begin{aligned} \mathbb{E} [(Y_{k+1}^N - Y_k^N)^3 | Y_k^N] &= Y_k^N \left(1 - \frac{Y_k^N}{N}\right) \left(1 - 2\frac{Y_k^N}{N}\right), \\ |\mathbb{E} [(Y_{k+1}^N - Y_k^N)^3 | Y_k^N]| &\leq N, \\ |\mathbb{E} [(Y_{k_1+1}^N - Y_{k_1}^N)(Y_{k_2+1}^N - Y_{k_2}^N)^3]| &\leq N^2. \end{aligned}$$

Consequently the third term is estimated exactly as the second one. It remains to consider the last term, which is bounded above by

$$\begin{aligned} N^{-4} \mathbb{E} \sum_{i < k < j} (Y_k^N - Y_i^N)^2 (Y_{k+1}^N - Y_k^N)^2 &\leq N^{-3} \sum_{i < k < j} \mathbb{E} [(Y_k^N - Y_i^N)^2] \\ &= N^{-3} \sum_{i < k < j} \sum_{i \leq \ell < k} \mathbb{E} [(Y_{\ell+1}^N - Y_\ell^N)^2] \\ &\leq N^{-2} \sum_{i < k < j} (k - i) \\ &\leq (t - s)^2. \end{aligned}$$

We have proved that

$$\limsup_{N \rightarrow \infty} \mathbb{E} [|X_t^N - X_s^N|^4] \leq C(t-s)^2.$$

PROOF OF 2. With the same notations as above, in particular  $\Delta t = N^{-1}$ ,

$$\begin{aligned} G^N f(x) &:= \mathbb{E} [f(X_{t+\Delta t}^N) - f(X_t^N) | X_t^N = x] \\ &= \mathbb{E} \left[ f \left( N^{-1} \sum_{i=1}^N Z_i \right) - f(x) \right] \\ &= \frac{1}{2} \mathbb{E} \left[ \left( N^{-1} \sum_{i=1}^N Z_i - x \right)^2 \right] f''(x) + \frac{1}{6} \mathbb{E} \left[ \left( N^{-1} \sum_{i=1}^N Z_i - x \right)^3 f'''(\xi) \right] \\ &= \frac{1}{2N} x(1-x) f''(x) + r_N(x), \end{aligned}$$

where  $r_N(x) = O(N^{-3/2})$ , since

$$\begin{aligned} \frac{1}{6} \mathbb{E} \left| \left( N^{-1} \sum_{i=1}^N Z_i - x \right)^3 f'''(\xi) \right| &\leq \frac{1}{6} \|f'''\|_\infty \sup_{0 \leq x \leq 1} \mathbb{E} \left( \left| N^{-1} \sum_{i=1}^N Z_i - x \right|^3 \right) \\ &\leq \frac{1}{6} \|f'''\|_\infty \sup_{0 \leq x \leq 1} \left( \mathbb{E} \left( \left| N^{-1} \sum_{i=1}^N Z_i - x \right|^4 \right) \right)^{3/4} \\ &= O(N^{-3/2}). \end{aligned}$$

Now it is easily seen that

$$\begin{aligned} M_t^N &:= f(X_t^N) - f(X_0^N) - \sum_{i=0}^{[Nt]-1} G^N f(X_{i/N}^N) \\ &= f(X_t^N) - f(X_0^N) - \frac{1}{2} \int_0^{[Nt]/N} X_s^N (1 - X_s^N) f''(X_s^N) ds - \int_0^{[Nt]/N} N r_N(X_s^N) ds \end{aligned}$$

is a bounded martingale (with respect to the natural filtration generated by the process  $X^N$ ).

This means that for any  $0 \leq s < t$  and any bounded and continuous function  $\varphi : D([0, s]; [0, 1]) \rightarrow \mathbb{R}$ ,

$$\mathbb{E} [M_t^N \varphi((X_r^N)_{0 \leq r \leq s})] = \mathbb{E} [M_s^N \varphi((X_r^N)_{0 \leq r \leq s})].$$

Taking the limit along any converging subsequence, we get that any limit point  $X$  satisfies (2.1.1).  $\square$

## 2.2 Wright–Fisher model with mutations

Assume that mutation converts at birth an  $A$ -type to an  $a$ -type with probability  $\alpha_1$ , and converts an  $a$ -type to an  $A$ -type with probability  $\alpha_0$ . Here

$$\mathbb{P}(Y_{k+1}^N = i | Y_k^N = j) = C_N^i p_j^i (1 - p_j)^{N-i},$$

where

$$p_j = \frac{j(1 - \alpha_1) + (N - j)\alpha_0}{N}.$$

We now want to let  $N \rightarrow \infty$ . Assume that  $\alpha_0$  and  $\alpha_1$  are of the form

$$\alpha_1 = \gamma_1/N, \quad \alpha_0 = \gamma_0/N, \quad \text{where } \gamma_1 > 0, \quad \gamma_0 > 0 \text{ are fixed,}$$

and define again the continuous time process

$$X_t^N = N^{-1}Y_{[Nt]}^N, \quad t \geq 0.$$

$$\begin{aligned} \mathbb{E}[X_{t+\Delta t}^N - X_t^N | X_t^N] &= [-\gamma_1 X_t^N + \gamma_0(1 - X_t^N)]\Delta t, \\ \mathbb{E}[(X_{t+\Delta t}^N - X_t^N)^2 | X_t^N] &= X_t^N(1 - X_t^N)\Delta t + O(\Delta t^2). \end{aligned}$$

As  $N \rightarrow \infty$ ,  $X^N \Rightarrow X$ , where  $\{X_t, t \geq 0\}$  solves the SDE

$$dX_t = \gamma_0(1 - X_t)dt - \gamma_1 X_t dt + \sqrt{X_t(1 - X_t)}dB_t, \quad t \geq 0.$$

## 2.3 Wright–Fisher model with selection

Assume that type  $A$  is selectively superior to type  $a$ . Then

$$\mathbb{P}(Y_{k+1}^N = i | Y_k^N = j) = C_N^i p_j^i (1 - p_j)^{N-i},$$

where

$$p_j = \frac{j(1+s)}{j(1+s) + N - j}.$$

If we want to combine mutations and selection, we choose

$$p_j = \frac{(1+s)[j(1-\alpha_1) + (N-j)\alpha_2]}{(1+s)[j(1-\alpha_1) + (N-j)\alpha_2] + j\alpha_1 + (N-j)(1-\beta)}.$$

We again want to let  $N \rightarrow \infty$ . Let  $\alpha_1 = 0$ ,  $\alpha_0 = 0$ , and  $s = \beta/N$ , with  $\beta > 0$ . We define the continuous time process

$$X_t^N = N^{-1}Y_{[Nt]}^N, \quad t \geq 0.$$

$$\begin{aligned} \mathbb{E}[X_{t+\Delta t}^N - X_t^N | X_t^N] &= \beta X_t^N(1 - X_t^N)\Delta t + 0(\Delta t^2), \\ \mathbb{E}[(X_{t+\Delta t}^N - X_t^N)^2 | X_t^N] &= X_t^N(1 - X_t^N)\Delta t + 0(\Delta t^2). \end{aligned}$$

As  $N \rightarrow \infty$ ,  $X^N \Rightarrow X$ , where  $\{X_t, t \geq 0\}$  solves the SDE

$$dX_t = \beta X_t(1 - X_t)dt + \sqrt{X_t(1 - X_t)}dB_t, \quad t \geq 0.$$

Suppose now that  $\alpha_0 = \alpha_1 = \gamma/N$ . Then

$$\begin{aligned} \mathbb{E}[X_{t+\Delta t}^N - X_t^N | X_t^N] &= \beta X_t^N(1 - X_t^N)\Delta t + \gamma(1 - 2X_t^N)\Delta t + o(\Delta t^2), \\ \mathbb{E}[(X_{t+\Delta t}^N - X_t^N)^2 | X_t^N] &= X_t^N(1 - X_t^N)\Delta t + o(\Delta t^2). \end{aligned}$$

Then as  $N \rightarrow \infty$ ,  $X^N \Rightarrow X$ , where  $\{X_t, t \geq 0\}$  solves the SDE

$$dX_t = [\beta X_t(1 - X_t) + \gamma(1 - 2X_t)]dt + \sqrt{X_t(1 - X_t)}dB_t, \quad t \geq 0.$$

## 2.4 Duality between Kingman's coalescent and Wright–Fisher's diffusion

We associate to Kingman's coalescent again the process  $\{R_t, t \geq 0\}$  defined by  $R_t = |\mathcal{R}_t|$ .  $\{R_t, t \geq 0\}$  is a pure death process on  $\mathbb{N}^*$ , with transition from  $n$  to  $n - 1$  happening at rate  $\binom{n}{2}$ . Consider moreover  $\{X_t, t \geq 0\}$  a Wright–Fisher diffusion, i. e. the solution of the SDE

$$dX_t = \sqrt{X_t(1 - X_t)}dB_t, \quad t \geq 0; \quad X_0 = x,$$

where  $0 < x < 1$ .

**Proposition 2.4.1.** *The following duality relation holds*

$$\mathbb{E}[X_t^n | X_0 = x] = \mathbb{E}[x^{R_t} | R_0 = n], \quad t \geq 0. \quad (2.4.1)$$

PROOF: We fix  $n \geq 1$ . Define

$$u(t, x) = \mathbb{E}[x^{R_t} | R_0 = n].$$

Since  $\{R_t; t \geq 0\}$  is a Markov process with generator  $Q$  defined by

$$Qf(n) = \frac{n(n-1)}{2}[f(n-1) - f(n)],$$

for any  $f : \mathbb{N} \rightarrow \mathbb{R}_+$ ,

$$\mathcal{N}_t^f = f(R_t) - f(R_0) - \int_0^t \binom{R_s}{2} [f(R_s - 1) - f(R_s)] ds$$

is a martingale. Let us explicit the above identity for the particular choice  $f(n) = x^n$  :

$$\begin{aligned} x^{R_t} &= x^n + \int_0^t \frac{R_s(R_s - 1)}{2} [x^{R_s - 1} - x^{R_s}] ds + \mathcal{N}_t \\ &= x^n + \frac{x(1-x)}{2} \int_0^t R_s(R_s - 1) x^{R_s - 2} ds + \mathcal{N}_t. \end{aligned}$$

Writing that  $\mathbb{E}[\mathcal{N}_t | R_0 = n] = 0$ , we deduce that for each  $n \in \mathbb{N}$ ,

$$u(t, x) = u(0, x) + \frac{x(1-x)}{2} \int_0^t \frac{\partial^2 u}{\partial x^2}(s, x) ds.$$

This means that  $u$  solves the following linear parabolic PDE

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{x(1-x)}{2} \frac{\partial^2 u}{\partial x^2}(t, x) & t \geq 0, 0 < x < 1; \\ u(0, x) = x^n, \quad u(t, 0) = 0, \quad u(t, 1) = 1 \end{cases}$$

It is easily checked that  $x \rightarrow u(t, x)$  is smooth. We then may apply Itô's calculus to develop  $u(t-s, X_s)$ , which yields, since  $u$  solves the above PDE,

$$u(0, X_t) = u(t, x) + M_t,$$

where  $M_t$  is a zero-mean martingale. Taking the expectation in the last identity yields  $u(t, x) = \mathbb{E}_x[X_t^n]$ .  $\square$

We deduce from the above a simple proof of the uniqueness in law of the solution of the Wright–Fisher SDE.

**Corollary 2.4.2.** *The law of the solution  $\{X_t, t \geq 0\}$  of Wright–Fisher SDE is unique.*

PROOF: Since the solution is a homogeneous Markov process, it suffices to show that the transition probabilities are uniquely determined. But for all  $t > 0$ ,  $x \in [0, 1]$ , the conditional law of  $X_t$ , given that  $X_0 = x$  is determined by its moments, since  $X_t$  is a bounded r. v. The result then follows from Proposition 2.4.1.

**Remark 2.4.3.** *As  $t$  gets large, both terms of the identity (2.4.1) tend to  $x$ . The left hand side because it behaves for  $t$  large as  $\mathbb{P}(X_t = 1) \rightarrow x$ , and the right hand side since  $\mathbb{P}(R_t = 1) \rightarrow 1$ , as  $t \rightarrow \infty$ .*



# Chapter 3

## The look–down approach to Wright–Fisher and Moran models

### 3.1 Introduction

Here we shall first define an alternative to the Wright–Fisher model, namely the continuous–time Moran model. We shall then present the look–down construction due to Donnelly and Kurtz [6], and show that this particular version of the Moran model converges a. s., as the population size  $N$  tends to infinity, towards the Wright–Fisher diffusion.

### 3.2 The Moran model

Consider a population of fixed size  $N$ , which evolves in continuous time according to the following rule. For each ordered pair  $(i, j)$  with  $1 \leq i \neq j \leq N$ , at rate  $1/2N$  individual  $i$  gives birth to an individual who replaces individual  $j$ , independently of the other ordered pairs. This can be graphically represented as follows. For each ordered pair  $(i, j)$  we draw arrows from  $i$  to  $j$  at rate  $1/2N$ . If we denote by  $\mathcal{P}$  the set of ordered pairs of elements of the set  $\{1, \dots, N\}$ ,  $\mu$  the counting measure on  $\mathcal{P}$ , and  $\lambda$  the Lebesgue measure on  $\mathbb{R}_+$ , the arrows constitute a Poisson process on  $\mathcal{P} \times \mathbb{R}_+$  with intensity measure  $(2N)^{-1}\mu \times \lambda$ .

Consider the Harris diagram for the Moran model in Figure 3.1. Time

flows down. If we follow the diagram backward from the bottom to the top, and coalesce any pair of individuals whenever they find a common ancestor, we see that starting from the trivial partition

$$\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}, \{9\}\},$$

after the first arrow has been reached we get

$$\{\{1\}, \{2\}, \{3, 4\}, \{5\}, \{6\}, \{7\}, \{8\}, \{9\}\},$$

next

$$\{\{1\}, \{2\}, \{3, 4\}, \{5\}, \{6, 9\}, \{7\}, \{8\}\},$$

next

$$\{1, 3, 4\}, \{2\}, \{5\}, \{6, 9\}, \{7\}, \{8\}\},$$

the fourth arrow does not modify the partition, next

$$\{\{1, 3, 4, 5\}, \{2\}, \{6, 9\}, \{7\}, \{8\}\},$$

next

$$\{\{1, 3, 4, 5\}, \{2\}, \{6, 7, 9\}, \{8\}\},$$

the next arrow has no effect, then

$$\{\{1, 3, 4, 5, 8\}, \{2\}, \{6, 7, 9\}\}$$

and the last arrow (the first on from the top) has no effect.

It is not hard to see that the coalescent which is imbedded in the Moran model looked at backward in time is exactly Kingman's coalescent – here more precisely Kingman's  $N$ -coalescent.

Suppose now that as in the preceding chapter the population includes two types of individuals, type  $a$  and type  $A$ . Each offspring is of the same type as his parent, we do not consider mutations so far. Denote

$$Y_t^N = \text{number of type } A \text{ individuals at time } t.$$

Provided we specify the initial number of type  $A$  individuals, the above model completely specifies the law of  $\{Y_t^N, t \geq 0\}$ . We now introduce the *proportion* of type  $A$  individuals in rescaled time, namey

$$X_t^N = N^{-1}Y_{Nt}^N, \quad t \geq 0.$$

Note that in this new time scale, the above Poisson process has the intensity measure  $\frac{1}{2}\mu \times \lambda$ . We have, similarly as in Theorem 2.1.1,

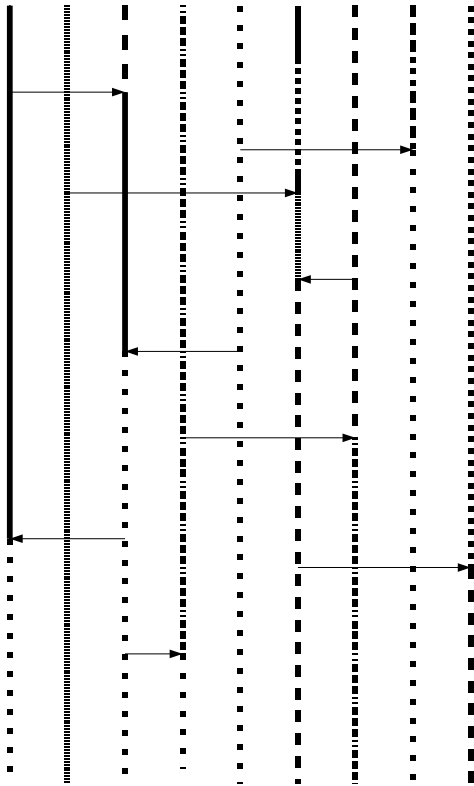


Figure 3.1: The Moran model

**Theorem 3.2.1.** *Suppose that  $X_0^N \Rightarrow X_0$ , as  $N \rightarrow \infty$ . Then  $X^N \Rightarrow X$  in  $D(\mathbb{R}_+; [0, 1])$ , where  $\{X_t, t \geq 0\}$  solves the SDE*

$$dX_t = \sqrt{X_t(1 - X_t)}dB_t, \quad t \geq 0. \quad (3.2.1)$$

PROOF:

As for Theorem 2.1.1, the proof goes through two steps.

PROOF OF TIGHTNESS We need to compute  $\mathbb{E}[|X_t^N - X_s^N|^4]$ , in order to apply Corollary 6.2.3. For  $0 \leq s < t$ , let  $Z_{s,t}$  denote the number of arrows in Harris' diagram between time  $Ns$  and time  $Nt$ .  $Z_{s,t}$  is a Poisson random variable with parameter  $N(N-1)(t-s)/2$ . We have

$$X_t^N - X_s^N = N^{-1} \sum_{i=1}^{Z_{s,t}} \xi_i$$

where the  $\xi_i$  are independent of  $Z_{s,t}$ ,

$$\xi_i = \mathbf{1}_{\{a \rightarrow_i A\}} - \mathbf{1}_{\{A \rightarrow_i a\}}$$

where  $\{a \rightarrow_i A\}$  is the event that the  $i$ -th arrow is drawn from an  $a$  individual to an  $A$  individual, and  $\{A \rightarrow_i a\}$  is the event that the  $i$ -th arrow is drawn from an  $A$  individual to an  $a$  individual. Conditioned upon  $Z_{s,t}$  and the past of  $X^N$  up to that jump time,  $\xi_i$  has zero first and third moments, second and fourth moments bounded by one. Consequently, arguing as in the proof of Theorem 2.1.1, we conclude that

$$\begin{aligned} \mathbb{E} \left[ |X_t^N - X_s^N|^4 \right] &= N^{-4} \mathbb{E} \left( \sum_{i=1}^{Z_{s,t}} \xi_i^4 + 2 \sum_{1 \leq i < j \leq Z_{s,t}} \xi_i^2 \xi_j^2 + \sum_{1 \leq i \leq Z_{s,t}} [\xi_1 + \dots + \xi_{i-1}]^2 \xi_i^2 \right) \\ &\leq N^{-4} \mathbb{E} \left( \sum_{i=1}^{Z_{s,t}} \xi_i^4 + 3 \sum_{1 \leq i < j \leq Z_{s,t}} \xi_i^2 \xi_j^2 \right) \\ &\leq N^{-4} \left( \mathbb{E} Z_{s,t} + \frac{3}{2} \mathbb{E} [Z_{s,t}(Z_{s,t} - 1)] \right) \\ &\leq \frac{t-s}{2N^2} + \frac{3}{8} (t-s)^2 \end{aligned}$$

IDENTIFICATION OF THE LIMIT Note that the process  $\{Z_t^N := Y_{Nt}^N, t \geq 0\}$  is a jump Markov process with values in the finite set  $\{0, 1, 2, \dots, N\}$ , which, when in state  $k$ , jumps to

1.  $k - 1$  at rate  $k(N - k)/2$ ,
2.  $k + 1$  at rate  $k(N - k)/2$ .

In other words if  $Q^N$  denotes the infinitesimal generator of this process,

$$Q^N f(Z_t^N) = Z_t^N(N - Z_t^N) \left[ \frac{f(Z_t^N + 1) + f(Z_t^N - 1)}{2} - f(Z_t^N) \right],$$

which implies that

$$\begin{aligned} \mathbb{E} [f(X_{t+\Delta t}^N) - f(X_t^N) | X_t^N = x] &= N^2 x(1 - x) \left[ \frac{f(x + \frac{1}{N}) + f(x - \frac{1}{N})}{2} - f(x) \right] \Delta t + o(\Delta t) \\ &= \frac{x(1 - x)}{2} f''(x) \Delta t + o(\Delta t), \end{aligned}$$

since from two applications of the order two Taylor expansion,

$$\frac{f(x + \frac{1}{N}) + f(x - \frac{1}{N})}{2} - f(x) = \frac{1}{2N^2} f''(x) + o(N^{-2}).$$

□

### 3.3 The look-down construction

The construction which we are going to present here is often called in the literature the *modified* look-down construction.

Let us again consider first the case where the size  $N$  of the population is finite and fixed. We redraw the Harris diagram of Moran's model, forbidding half of the arrows. We consider only arrows from left to right. Considering immediately the rescaled time, for each  $1 \leq i < j \leq N$ , we put arrows from  $i$  to  $j$  at rate 1 (twice the above  $1/2$ ). At such an arrow, the individual at level  $i$  puts a child at level  $j$ . Individuals previously at levels  $j, \dots, N - 1$  are shifted one level up; individual at site  $N$  dies.

Note that in this construction the level one individual is immortal, and the genealogy is not exchangeable.

However the partition at time  $t$  induced by the ancestors at time 0 is exchangeable, since going back each pair coalesces at rate 1.

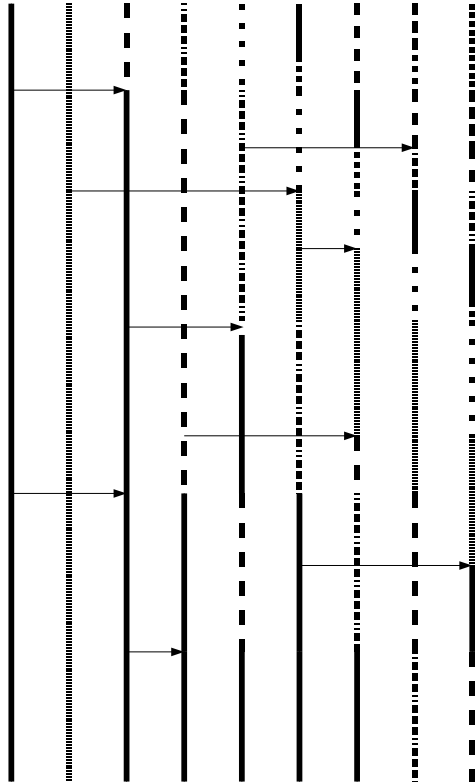


Figure 3.2: The look-down construction

Consider now the case where there are two types of individuals, type  $a$ , represented by black, and type  $A$ , represented by red. We want to choose the types of the  $N$  individuals at time 0 in an exchangeable way, with the constraint that the proportion of type red individuals is given. One possibility is to draw without replacement  $N$  balls from an urn where we have put  $k$  red balls and  $N - k$  black balls. At each draw, each of the balls which remain in the urn has the same probability of being chosen.

It follows from the above considerations that at each time  $t > 0$ , the types of the  $N$  individuals are exchangeable.

### 3.4 a. s. convergence to the Wright–Fisher diffusion

Our goal now is to take the limit in the above quantities as  $N \rightarrow \infty$ . The look-down construction can be defined directly with an infinite population. The description is the same as above, except that we start with an infinite number of lines, and no individual dies any more.

Note that the possibility of doing the same construction for  $N = \infty$  is related to the fact that in any finite interval of time, if we restrict ourselves to the first  $N$  individuals, the evolution is determined by finitely many arrows. This would not be the case with the standard Moran model, which could not be described in the case  $N = \infty$ . Indeed in the Moran model with infinitely many individuals, there would be infinitely many arrows towards any individual  $i$ , in any time interval of positive length. This is a great advantage of the look-down construction.

Consider now the case of two types of individuals. Suppose that the initial colours of the various individuals at time  $t = 0$  are i. i. d., each red with probability  $x$ , black with probability  $1 - x$ . Define

$$\eta_t(k) = \begin{cases} 1, & \text{if the } k\text{-th individual is red at time } t; \\ 0, & \text{if the } k\text{-th individual is black at time } t. \end{cases}$$

$\{\eta_0(k), k \geq 1\}$  are i. i. d. Bernoulli random variables, while at each  $t > 0$ ,  $\{\eta_t(k), k \geq 1\}$  is an exchangeable sequence of  $\{0, 1\}$ -valued random variables. A celebrated theorem due to de Finetti (see Corollary 6.3.6 below) says that an exchangeable sequence of  $\{0, 1\}$ -valued r. v. is a mixture of i.

i. d. Bernoulli. Consequently the following limit exists a. s.

$$X_t = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \eta_t(i) = \lim_{N \rightarrow \infty} X_t^N. \quad (3.4.1)$$

Before stating the main theorem of this section, let us establish three auxiliary results which we shall need in its proof.

**Proposition 3.4.1.** *Let  $\{\xi_1, \xi_2, \dots\}$  be a countable exchangeable sequence of  $\{0, 1\}$ -valued r. v.'s and  $\mathcal{T}$  denote its tail  $\sigma$ -field. Let  $\mathcal{H}$  be some additional  $\sigma$ -algebra. If conditionally upon  $\mathcal{T} \vee \mathcal{H}$ , the r. v.'s are exchangeable, then conditionally upon  $\mathcal{T} \vee \mathcal{H}$  they are i. i. d.*

PROOF: Let  $f : \{0, 1\}^n \rightarrow \mathbb{R}$  be an arbitrary mapping. It follows from the assumption that

$$\begin{aligned} \mathbb{E}(f(\xi_1, \dots, \xi_n) | \mathcal{T} \vee \mathcal{H}) &= \mathbb{E}\left(N^{-1} \sum_{k=1}^N f(\xi_{(k-1)n+1}, \dots, \xi_{kn}) \middle| \mathcal{T} \vee \mathcal{H}\right) \\ &= \mathbb{E}(f(\xi_1, \dots, \xi_n) | \mathcal{T}), \end{aligned}$$

where the second equality follows from the fact that the quantity inside the previous conditional expectation converges a. s. to  $\mathbb{E}(f(\xi_1, \dots, \xi_n) | \mathcal{T})$  as  $N \rightarrow \infty$ , as a consequence of exchangeability and de Finetti's theorem (see Corollary 6.3.6 below).  $\square$

**Lemma 3.4.2.** *Let  $\{X_n, n \geq 1\}$  and  $X$  be real-valued random variables such that  $X_n \rightarrow X$  a. s. as  $n \rightarrow \infty$ , and  $A, B \in \mathcal{F}$ . If*

$$\mathbb{P}(A \cap C) = \mathbb{P}(B \cap C), \quad \forall C \in \sigma(X_n), \quad \forall n \geq 1,$$

then

$$\mathbb{P}(A \cap C) = \mathbb{P}(B \cap C), \quad \forall C \in \sigma(X).$$

PROOF: The assumption implies that for all  $f \in C_b(\mathbb{R})$ , all  $n \geq 1$ ,

$$\mathbb{E}[f(X_n); A] = \mathbb{E}[f(X_n); B],$$

from which we deduce by bounded convergence that

$$\mathbb{E}[f(X); A] = \mathbb{E}[f(X); B].$$

The result follows.  $\square$

Let  $S_n$  denote the group of permutations of  $\{1, 2, \dots, n\}$ . If  $\pi \in S_n$ ,  $a \in \{0, 1\}^n$ , we shall write  $\pi^*(a) = (a_{\pi(1)}, \dots, a_{\pi(n)})$ . Recall that a partition  $\mathcal{P}$  of  $\{1, \dots, n\}$  induces an equivalence relation, whose equivalence classes are the blocks of the partition. Hence we shall write  $i \simeq_{\mathcal{P}} j$  whenever  $i$  and  $j$  are in the same block of  $\mathcal{P}$ . Finally we write  $\#\mathcal{P}$  for the number of blocks of the partition  $\mathcal{P}$ .

**Proposition 3.4.3.** *For all  $n \geq 1$ ,  $0 < r < s$ ,  $a \in \{0, 1\}^n$ ,  $p$  such that  $0 \leq np \leq n$  is an integer,  $\pi \in S_n$ ,*

$$\mathbb{P}(\{\eta_s^n = a\} \cap \{X_r^n = p\}) = \mathbb{P}(\{\eta_s^n = \pi^*(a)\} \cap \{X_r^n = p\}).$$

PROOF: Denote by  $\mathcal{P}_a$  the set of partitions  $\mathcal{P}$  of  $\{1, 2, \dots, n\}$  which are such that  $i \simeq_{\mathcal{P}} j \Rightarrow a_i = a_j$ .

The arrows between time  $r$  and time  $s$  in the look-down construction pointing to levels between 2 and  $n$  prescribe in particular which individuals at time  $s$  have the same ancestor back at time  $r$ . This corresponds to a partition  $\{1, 2, \dots, n\}$  which is the result of the coalescent process backward from time  $s$  to time  $r$ .  $\{\text{coal}_s^r = \mathcal{P}\}$  is the event that that partition is  $\mathcal{P}$ . Suppose that  $\#\mathcal{P} = k$ . The look-down construction prescribes that the block containing 1 carries the type of the individual sitting on level 1 at time  $r$ , the block containing the smallest level not in that first block carries the type of the individual sitting on level 2 at time  $r$ , ... Thus the event

$$\{\eta_s^n = a\} \cap \{\text{coal}_s^r = \mathcal{P}\},$$

which is non empty iff  $\mathcal{P} \in \mathcal{P}_a$ , determines the values of  $\eta_r^n(1), \dots, \eta_r^n(k)$  if  $k = \#\mathcal{P}$ . There is a finite (possibly zero) number of possible values  $b$  for  $\eta_r^n$  which respect both the above condition and the restriction  $n^{-1} \sum_{i=1}^n b_i = p$ . We denote by  $\mathcal{A}_{r,s}(a, \mathcal{P}, p) \subset \{0, 1\}^n$  the set of those  $b$ 's. Note that this set is empty if the restriction  $n^{-1} \sum_{i=1}^n b_i = p$  contradicts the conditions  $\eta_s^n = a$  and  $\text{coal}_s^r = \mathcal{P}$ .

We then have

$$\{\eta_s^n = a\} \cap \{X_r^n = p\} = \bigcup_{\mathcal{P} \in \mathcal{P}_a} \bigcup_{b \in \mathcal{A}_{r,s}(a, \mathcal{P}, p)} \{\text{coal}_s^r = \mathcal{P}\} \cap \{\eta_r^n = b\},$$

and from the independence of  $\text{coal}_s^r$  and  $\eta_r^n$

$$\mathbb{P}(\eta_s^n = a, X_r^n = p) = \sum_{\mathcal{P} \in \mathcal{P}_a} \sum_{b \in \mathcal{A}_{r,s}(a, \mathcal{P}, p)} \mathbb{P}(\text{coal}_s^r = \mathcal{P}) \mathbb{P}(\eta_r^n = b).$$

Similarly, if  $\pi^*(\mathcal{P})$  is defined by  $i \simeq_{\mathcal{P}} j \Leftrightarrow \pi(i) \simeq_{\pi^*(\mathcal{P})} \pi(j)$ ,

$$\begin{aligned} \mathbb{P}(\eta_s^n = \pi^*(a), X_r^n = p) &= \sum_{\mathcal{P} \in \mathcal{P}_{\pi^*(a)}} \sum_{b \in \mathcal{A}_{r,s}(\pi^*(a), \mathcal{P}, p)} \mathbb{P}(\text{coal}_s^r = \mathcal{P}) \mathbb{P}(\eta_r^n = b) \\ &= \sum_{\mathcal{P} \in \mathcal{P}_a} \sum_{b \in \mathcal{A}_{r,s}(\pi^*(a), \pi^*(\mathcal{P}), p)} \mathbb{P}(\text{coal}_s^r = \pi^*(\mathcal{P})) \mathbb{P}(\eta_r^n = b), \end{aligned}$$

We now describe a one-to-one correspondence  $\rho_\pi$  between  $\mathcal{A}_{r,s}(a, \mathcal{P}, p)$  and  $\mathcal{A}_{r,s}(\pi^*(a), \pi^*(\mathcal{P}), p)$ . Suppose that  $\#\mathcal{P} = k$ . Then  $\#\pi^*(\mathcal{P}) = k$  as well. Let  $b \in \mathcal{A}_{r,s}(a, \mathcal{P}, p)$ . The values of  $b_j$ ,  $1 \leq j \leq k$  are specified by the pair  $(a, \mathcal{P})$ . We define  $b' = \rho_\pi(b)$  as follows.  $b'_1, \dots, b'_k$  are specified by  $(\pi^*(a), \pi^*(\mathcal{P}))$ . Note that the definitions of  $\pi^*(a)$  and  $\pi^*(\mathcal{P})$  imply that

$$(b'_1, \dots, b'_k) = (b_{\pi'(1)}, \dots, b_{\pi'(k)}), \quad \text{for some } \pi' \in S_k.$$

We complete the definition of  $b'$  by the conditions

$$b'_j = b_j, \quad k < j \leq n.$$

Clearly there exists  $\pi'' \in S_n$  such that  $b' = \pi''^*(b)$ .

Consequently

$$\begin{aligned} \sum_{b \in \mathcal{A}_{r,s}(\pi^*(a), \pi^*(\mathcal{P}), p)} \mathbb{P}(\text{coal}_s^r = \pi^*(\mathcal{P})) \mathbb{P}(\eta_r^n = b) \\ &= \sum_{b \in \mathcal{A}_{r,s}(a, \mathcal{P}, p)} \mathbb{P}(\text{coal}_s^r = \pi^*(\mathcal{P})) \mathbb{P}(\eta_r^n = \rho_\pi(b)) \\ &= \sum_{b \in \mathcal{A}_{r,s}(a, \mathcal{P}, p)} \mathbb{P}(\text{coal}_s^r = \mathcal{P}) \mathbb{P}(\eta_r^n = b), \end{aligned}$$

where the last identity follows from the fact that both  $\eta_r^n$  and  $\text{coal}_s^r$  are exchangeable. The result follows from the three identities proved above.  $\square$

We can now prove

**Theorem 3.4.4.** *The  $[0, 1]$ -valued process  $\{X_t, t \geq 0\}$  defined by (3.4.1) possesses a continuous modification which is a weak sense solution of the Wright-Fisher SDE, i. e. there exists a standard Brownian motion  $\{B_t, t \geq 0\}$  such that*

$$dX_t = \sqrt{X_t(1-X_t)} dB_t, \quad t \geq 0.$$

STEP 1 We first need to show that  $\{X_t, t \geq 0\}$  defined by (3.4.1) possesses a continuous modification. This will follow from a well-known Lemma due to Kolmogorov, if we show that  $\mathbb{E}[|X_t - X_s|^4] \leq c(t-s)^2$ , which follows from Fatou's Lemma and the fact that  $\mathbb{E}[|X_t^N - X_s^N|^4] \leq c(t-s)^2$ . This last inequality can be shown exactly as for the Moran model (see the proof of Theorem 3.2.1).

STEP 2 We now want to show that  $\{X_t, t \geq 0\}$  is a Markov process. We know that conditionally upon  $X_s = x$ , the  $\eta_s(k)$  are i. i. d. Bernoulli with parameter  $x$ . Now for any  $t > s$ ,  $X_t$  depends only upon the  $\eta_s(k)$  and the arrows which are drawn between time  $s$  and time  $t$ , which are independent from  $\{X_r, 0 \leq r \leq s\}$ . So all we need to show is that conditionally upon  $\sigma(X_r, 0 \leq r \leq s)$ , the  $\eta_s(k)$  are i. i. d. In view of Proposition 3.4.1, it suffices to prove that conditionally upon  $\sigma(X_r, 0 \leq r \leq s)$ , the  $\eta_s(k)$  are exchangeable. This will follow from the fact that the same is true conditionally upon  $\sigma(X_{r_1}, \dots, X_{r_k}, X_s)$  for all  $k \geq 1, 0 \leq r_1 < r_2 < \dots < r_k < s$ .

We first prove this in the case  $k = 1$ . Write  $\eta_s^n = (\eta_s(1), \dots, \eta_s(n))$ . All we have to show is that for all  $n \geq 1, a \in \{0, 1\}^n, \pi \in S_n$ , if  $\pi^*(a) = (a_{\pi(1)}, \dots, a_{\pi(n)})$ ,  $A_r \in \sigma(X_r)$  and  $A_s \in \sigma(X_s)$ ,

$$\mathbb{P}(\{\eta_s^n = a\} \cap A_r \cap A_s) = \mathbb{P}(\{\eta_s^n = \pi^*(a)\} \cap A_r \cap A_s). \quad (3.4.2)$$

In view of Lemma 3.4.2, a sufficient condition for (3.4.2) is that for all  $m \geq n, p, q > 0$  such that  $0 \leq mp, mq \leq m$  are integers,

$$\mathbb{P}(\{\eta_s^n = a\} \cap \{X_r^m = p, X_s^m = q\}) = \mathbb{P}(\{\eta_s^n = \pi^*(a)\} \cap \{X_r^m = p, X_s^m = q\}),$$

and clearly it suffices to prove that result for  $n = m$ , which is done in Proposition 3.4.3.

A similar proof shows that the  $\eta_s(k)$  are conditionally exchangeable given  $\sigma(X_{r_1}, \dots, X_{r_k}, X_s)$ . The Markov property of the process  $\{X_t, t \geq 0\}$  is established.

STEP 3 It remains to show that the process  $\{X_t, t \geq 0\}$  has the right transition probability, which will follow (see Proposition 2.4.1) from the fact that for all  $n \geq 1, x \in [0, 1]$ ,

$$\mathbb{E}_x[X_t^n] = \mathbb{E}_n[x^{R_t}].$$

For all  $n \geq 1$ ,

$$X_t^n = \mathbb{P}(\eta_t(1) = \dots = \eta_t(n) = 1 | X_t),$$

consequently

$$\begin{aligned}\mathbb{E}_x[X_t^n] &= \mathbb{E}_x[\mathbb{P}(\eta_t(1) = \cdots = \eta_t(n) = 1|X_t)] \\ &= \mathbb{P}_x(\eta_t(1) = \cdots = \eta_t(n) = 1) \\ &= \mathbb{P}_x(\text{the ancestors at time } 0 \text{ of } 1, \dots, n \text{ are red}) \\ &= \mathbb{E}_n[x^{R_t}],\end{aligned}$$

where  $\{R_t, t \geq 0\}$  is a pure death continuous-time process, which jumps from  $k$  to  $k - 1$  at rate  $k(k - 1)/2$ .  $\square$

# Chapter 4

## Mutations : the infinitely many alleles model

Suppose now that mutations arise on each branch of the coalescence tree, according to a Poisson process with parameter  $\theta/2$ , see Figure 4.1. Assume that each mutation gives birth to a new type, different for all the others. For instance we may assume that the different types are i. i. d. r. v.'s following the uniform law on  $[0, 1]$ . We want to record the different types in a sample drawn at present time, we can as well “kill” the lineages which hit a mutation while going backward in time, which changes Figure 4.1 into Figure 4.2, which we can as well change into Figure 4.3. The killed coalescent can be produced by the following procedure : *Any pair of active classes merges at rate 1, any active class is killed at rate  $\theta/2$ .* When a class is killed, all its elements are assigned the same (different from all other classes) type. Finish when there are no classes left. Note that we add a mutation at the root of the tree.

### 4.1 Hoppe’s urn

Assume that there are  $k$  active classes in the killed coalescent described above. Then the probability that the next (backward in time) event is a

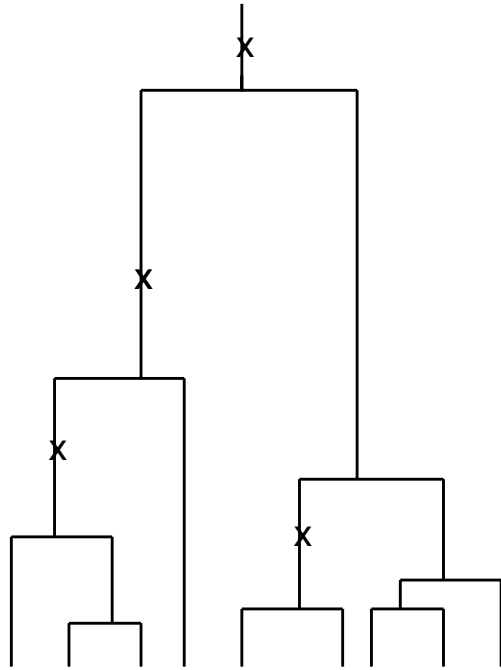


Figure 4.1: The coalescent with mutations

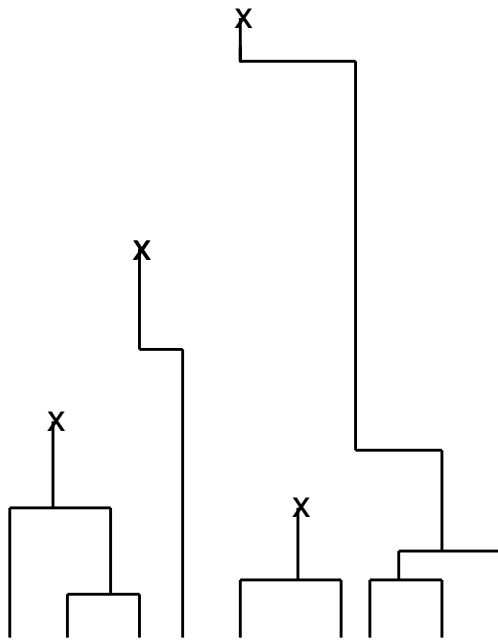


Figure 4.2: The lineages are killed above the mutations

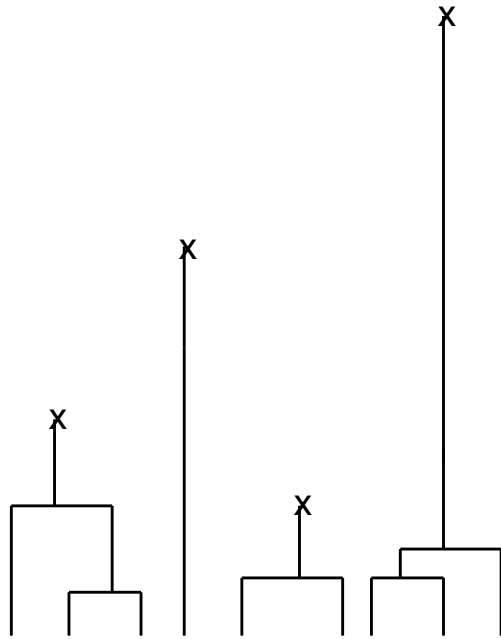


Figure 4.3: Equivalent to Figure 4.2

coalescence is

$$\frac{\binom{k}{2}}{\binom{k}{2} + k\frac{\theta}{2}} = \frac{k-1}{k-1+\theta},$$

and the probability that that event is a mutation (i. e. a killing) is

$$\frac{k\frac{\theta}{2}}{\binom{k}{2} + k\frac{\theta}{2}} = \frac{\theta}{k-1+\theta}.$$

Moreover, given the type of event, all possible coalescence (resp. mutations) are equally likely. The history of a sample of size  $n$  is described by  $n$  events  $e_n, e_{n-1}, \dots, e_1 \in \{\mathbf{coal}, \mathbf{mut}\}$ . Note that the event  $e_k$  happens just before (forward in time)  $k$  lineages are active, and each of those events corresponds backward in time to the reduction by one of the number of active lineages. The probability to observe a particular sequence is thus

$$\frac{\prod_{k=1}^n (\theta \mathbf{1}_{\{e_k=\mathbf{mut}\}} + (k-1) \mathbf{1}_{\{e_k=\mathbf{coal}\}})}{\prod_{k=1}^n (k-1+\theta)}. \quad (4.1.1)$$

Hoppe [8] noted that one can generate this sequence *forward in time* using the following urn model.

*Hoppe's urn model.* We start with an urn containing one unique black ball of mass  $\theta$ . At each step, a ball is drawn from the urn, with probability proportional to its mass. If the drawn ball is black return it to the urn, together with a ball of mass 1, of a new, not previously used, colour; if the drawn ball is coloured, return it together with another ball of mass 1 of the same colour.

At the  $k$ -th step, there are  $k$  balls, more precisely  $k-1$  coloured balls, plus the black (so called *mutation*) ball. The probability to pick the black ball is thus  $\theta/(k-1+\theta)$  while the probability to pick a coloured ball is  $(k-1)/(k-1+\theta)$ . If we define

$$e_k = \begin{cases} \mathbf{mut}, & \text{if in the } k\text{-step the black ball is drawn,} \\ \mathbf{coal}, & \text{otherwise.} \end{cases}$$

Clearly the probability to observe a particular sequence  $(e_1, \dots, e_n)$  is given by (4.1.1). Moreover, given that  $e_k = \mathbf{coal}$ , each of the  $k-1$  present coloured balls is equally likely to be picked.

Consequently, the distribution of the family sizes generated by the  $n$  coloured balls in Hoppe's urn after  $n$  steps is the same as the one induced by the  $n$ -coalescent in the infinitely-many-alleles mutation model.

## 4.2 Ewens' sampling formula

**Theorem 4.2.1.** *Let  $b_1, \dots, b_n \in \mathbb{N}$  be such that  $\sum_{j=1}^n j b_j = n$ . The probability of observing  $b_j$  different types, each with  $j$  representatives, ( $j = 1, \dots, n$ ) in a sample of size  $n$  is given by (here  $k = \sum_{j=1}^n b_j$ )*

$$\frac{n!}{1^{b_1} 2^{b_2} \dots n^{b_n}} \cdot \frac{1}{b_1! b_2! \dots b_n!} \cdot \frac{\theta^k}{\theta(\theta+1) \dots (\theta+n-1)}. \quad (4.2.1)$$

PROOF: We shall prove that the distribution of the type spectrum  $(B_1, \dots, B_n)$  in a sample of size  $n$  is the product of the measures  $\text{Poi}(\theta/j)$ ,  $j = 1, \dots, n$ , conditioned on  $\sum_{j=1}^n j B_j = n$ .

We start from the statement at the very end of the previous section. Now we describe another way of constructing the output of Hoppe's urn after  $n$  steps. Consider on  $\mathbb{R}_+$  a Poisson process of immigrants with parameter  $\theta$ . Each new immigrant starts immediately upon arrival to develop a Yule tree of parameter 1 (that is a new branch appears after a waiting time which is exponential with parameter 1, .. when they are  $k$  branches alive, a  $k+1$ -st appears after a waiting time which is exponential with parameter  $k$ , etc., the successive waiting times being mutually independent and independent of the time of arrival of the founder of the tree), and moreover the various trees are mutually independent. It follows from Exercise 1.5.3 that the number of branches of a Yule tree at time  $t$  (the tree being started at time 0) is geometric with parameter  $e^{-t}$ .

We can describe this model as follows. Consider the Markov process  $\{Y_t, t \geq 0\}$  with values in the subset  $E$  of  $\mathbb{N}^\infty$  consisting of those sequences whose only a finite number of components are non zero.  $Y_0 = (0, 0, \dots)$ , immigrants enter at rate  $\theta$ , the first immigrant creates the first tree, the second immigrant creates the second tree, etc... Each tree is a Yule tree, with develops independently of the other trees and of the arrivals of new immigrants.  $Y_t = (Y_t^1, Y_t^2, \dots)$ , where  $Y_t^k$  denotes the number of branches at time  $t$  of the  $k$ -th Yule tree. Define

$$|Y_t| = \sum_{k \geq 1} Y_t^k$$

the total number of branches of all the trees at time  $t$ .  $|Y_t|$  is a birth Markov process, the waiting time for the next birth when  $|Y_t| = k$  being exponential with parameter  $\theta + k$ .

It is easily seen that what we have just constructed is exactly a continuous time embedding of Hoppe's urn (just compute at each time  $s < t$  what is the probability that the next event is the arrival of a new immigrant, or the appearance of a new branch on an existing tree). Hence the output of Hoppe's urn has the same law as the set  $Y_t$  of Yule trees which this construction produces, if we look at it at the time when  $|Y_t|$  reaches the value  $n$ . It follows from Exercise 4.2.2 below that this law is the same as the law of  $Y_t$ , conditioned upon  $|Y_t| = n$ , for all  $t > 0$ .

This continuous time model can be considered as a Poisson process on  $[0, t] \times \mathbb{N}$ , with the intensity measure  $\theta ds \times \mathcal{G}(e^{-(t-s)})$ , where for  $0 < p < 1$ ,  $\mathcal{G}(p)$  denotes the geometric measure of parameter  $p$ . A point of this Poisson process is a pair  $(s, j)$ , where  $s \in [0, t]$  and  $j \geq 1$ . The point  $(s, j)$  corresponds to an immigrant which has appeared at time  $s$ , and whose associated Yule tree at time  $t$  has exactly  $j$  branches. Now for  $j \geq 1$ , let  $Z_j(t)$  denote the number of points of the above Poisson process whose second component equals  $j$ . It follows from well-known properties of Poisson processes that the  $Z_j(t)$ ,  $j \geq 1$  are mutually independent r. v.'s, the law of  $Z_j(t)$  being Poisson with parameter

$$\int_0^t \theta e^{-(t-s)} (1 - e^{-(t-s)})^{j-1} ds = \frac{\theta}{j} (1 - e^{-t})^j.$$

The above arguments show that the probability of observing  $b_j$  different types, each with  $j$  representatives, ( $j = 1, \dots, n$ ) in a sample of size  $n$  equals

$$\mathbb{P} \left( Z_1(t) = b_1, \dots, Z_n(t) = b_n \left| \sum_{j=1}^n j Z_j(t) = n \right. \right).$$

This is true for any  $t > 0$ . We can as well let  $t \rightarrow \infty$ , and we deduce that the same probability equals

$$\mathbb{P} \left( Z_1 = b_1, \dots, Z_n = b_n \left| \sum_{j=1}^n j Z_j = n \right. \right),$$

where  $Z_1, \dots, Z_n$  are independent, and for each  $1 \leq j \leq n$ , the law of  $Z_j$  is

Poisson with parameter  $\theta/j$ . This quantity is equal to

$$C(n, \theta) \prod_{j=1}^n e^{-\theta/j} \frac{(\theta/j)^{b_j}}{b_j!},$$

where the normalization constant satisfies

$$C(n, \theta)^{-1} = \mathbb{P} \left( \sum_{j=1}^n jB_j = n \right).$$

The result is proved, provided we check that

$$C(n, \theta) = \frac{n! \exp[\theta \sum_{j=1}^n 1/j]}{\theta(\theta+1) \cdots (\theta+n-1)}.$$

This will be done below in Lemma 4.2.3. Note however that we have already identified the Ewens sampling formula up to a normalization constant.  $\square$

**Exercise 4.2.2.** Let  $\{X_t, t \geq 0\}$  be a continuous time jump–Markov process, which takes values in a countable set  $E$ . Let  $T_0 = 0$  and  $T_n, n \geq 1$  denote the  $n$ -th jump time of  $X_t$ . Let  $\{Z_n, n \geq 0\}$  denote the associated embedded Markov chain, i. e.  $Z_0 = X_0$ , and for all  $n \geq 1$ ,  $Z_n = X_{T_n}$ . We know that there exists a function  $q : E \rightarrow (0, \infty)$  such that for each  $n \geq 0$ , the law of  $T_{n+1} - T_n$  is exponential with parameter  $q(Z_n)$ . Suppose that there exists a function  $h : \mathbb{N} \rightarrow (0, \infty)$  such that  $q(Z_n) = h(n), n \geq 0$ . Conclude that the sequences  $\{T_n, n \geq 1\}$  and  $\{Z_n, n \geq 1\}$  are mutually independent. Why is this last property not true in general ?

Apply this result to the process  $\{Y_t, t \geq 0\}$  from the previous proof. Show that the condition on  $q$  is satisfied here with  $h(n) = \theta + n$ . Prove that for all  $t > 0$ , the law of  $Z_n$  equals the conditional law of  $Y_t$ , given that  $|Y_t| = n$ .

We finally prove the

**Lemma 4.2.3.** If  $B_1, \dots, B_n$  are independent, each  $B_j$  being Poisson with parameter  $\theta/j$ , then

$$\mathbb{P} \left( \sum_{j=1}^n jB_j = n \right) = \frac{\theta(\theta+1) \cdots (\theta+n-1)}{n! \exp[\theta \sum_{j=1}^n 1/j]}.$$

PROOF: The left hand side of the identity to be established equals

$$\sum_{k_1, \dots, k_n; \sum j k_j = n} e^{-\theta/j} (\theta/j)^{k_j} / k_j! = \exp[-\theta \sum_{j=1}^n 1/j] \sum_k \alpha(n, k) \theta^k,$$

where

$$\alpha(n, k) = \sum_{k_1, \dots, k_n; \sum k_j = k, \sum j k_j = n} \left( \prod_{j=1}^n j^{k_j} k_j! \right)^{-1}.$$

It remains to show that

$$\theta(\theta + 1) \cdots (\theta + n - 1) = n! \sum_{k=1}^n \alpha(n, k) \theta^k.$$

Let  $s(n, k) = n! \alpha(n, k)$ . Splitting the last factor in the above left hand side into  $\theta$  plus  $n - 1$ , we deduce that

$$s(n, k) = s(n - 1, k - 1) + (n - 1) s(n - 1, k).$$

This shows that  $s(n, k)$  can be interpreted as the number of permutations of  $\{1, \dots, n\}$  which contain exactly  $k$  cycles. Now that number is given by

$$\begin{aligned} s(n, k) &= \sum_{k_1, \dots, k_n; \sum k_j = k, \sum j k_j = n} \frac{n!}{\prod_{j=1}^n (j k_j)!} \times \prod_{j=1}^n \left( \frac{(j k_j)!}{(j!)^{k_j} k_j!} [(j - 1)!]^{k_j} \right) \\ &= n! \sum_{k_1, \dots, k_n; \sum k_j = k, \sum j k_j = n} \prod_{j=1}^n \frac{1}{j^{k_j} k_j!}. \end{aligned}$$

Indeed in the above formula,

$$\frac{n!}{\prod_{j=1}^n (j k_j)!}$$

is the number of possibilities of choosing the elements for the cycles of size  $j$ ,  $j$  varying from 1 to  $n$ ,

$$\frac{(j k_j)!}{(j!)^{k_j} k_j!}$$

is the number of ways in which one can distribute the  $j k_j$  elements in the  $k_j$  cycles of size  $j$ , and

$$[(j - 1)!]^{k_j}$$

is the number of different possible orderings of the elements in the  $k_j$  cycles of size  $j$ .  $\square$

We now define  $K_n$  to be the number of different types observed in a sample of size  $n$ , or equivalently the number of different colours in Hoppe's urn after  $n$  steps. Then

$$K_n = X_1 + \cdots + X_n,$$

where

$$X_k = \mathbf{1}_{A_k}, \quad A_k = \{\text{the black ball is drawn at the } k\text{-th step}\},$$

consequently the events  $A_1, \dots, A_n$  are independent, with  $\mathbb{P}(A_k) = \theta/(\theta + k - 1)$ ,  $1 \leq k \leq n$ . Consequently

$$\begin{aligned} \mathbb{E}K_n &= \sum_{i=1}^n \frac{\theta}{\theta + i - 1} \simeq \theta \log(n), \\ \text{Var}(K_n) &= \sum_{i=1}^n \frac{\theta}{\theta + i - 1} \cdot \frac{i - 1}{\theta + i - 1} \simeq \theta \log(n), \\ \frac{K_n - \mathbb{E}K_n}{\sqrt{\text{Var}(K_n)}} &\Rightarrow N(0, 1), \quad \text{as } n \rightarrow \infty. \end{aligned}$$

**Exercise 4.2.4.** *Prove the last assertion, via a characteristic function computation.*

# Chapter 5

## Mutations : the infinitely many sites model

We now assume that each new mutation hits a new site, different from the sites hit by all other mutations. This is a reasonable assumption if the genomes under consideration are huge. A mathematical idealized model of the infinitely many sites model is to assume that the various mutations are i. i. d. random variables, all uniform on the interval  $[0, 1]$ . Again mutations arise according to a Poisson process along the branches of Kingman's coalescent tree, with intensity  $\theta/2$ .

### 5.1 The number of segregating sites

Let  $S_n$  denote the number of sites in the genome where the various individuals in the sample of size  $n$  do not coincide. This is the total number of sites hit by a mutation, i. e. the total number of mutations. Conditionally upon  $L_n$ ,  $S_n$  is Poisson with parameter  $\theta L_n/2$ . Consequently

$$\begin{aligned}\mathbb{E}S_n &= \mathbb{E}[\mathbb{E}(S_n|L_n)] \\ &= \frac{\theta}{2}\mathbb{E}L_n \\ &= \theta \sum_{j=1}^{n-1} \frac{1}{j}.\end{aligned}$$

Let  $a_n := \sum_{j=1}^{n-1} 1/j$ . Watterson's estimator of  $\theta$  is the unbiased estimator

$$\hat{\theta}_W = \frac{S_n}{a_n}.$$

Let us now compute the variance of  $\hat{\theta}_W$ . We have

$$S_n = \sum_{k=2}^n S_{n,k},$$

where  $S_{n,k}$  is the number of mutations which hit one of the ancestors of the sample, while there were  $k$  lineages ancestral to the sample. The  $S_{n,k}$ 's are independent, and if  $T_k$  is the duration of time during which there were  $k$  lineages active in the genealogy of the sample, the conditional law of  $S_{n,k}$ , given  $T_k$ , is Poisson with parameter  $\theta k T_k / 2$ . Now, with  $a_n$  defined as above and  $b_n = \sum_{j=1}^{n-1} j^{-2}$ ,

$$\begin{aligned} \text{Var}(S_n) &= \sum_{k=2}^n \text{Var}(S_{n,k}), \\ \mathbb{E}[S_{n,k}^2] &= \mathbb{E}[\mathbb{E}(S_{n,k}^2 | T_k)], \\ \mathbb{E}(S_{n,k}^2 | T_k) &= \left(\frac{\theta}{2} k T_k\right)^2 + \frac{\theta}{2} k T_k \\ \mathbb{E}[S_{n,k}^2] &= \frac{\theta}{k-1} + 2 \left(\frac{\theta}{k-1}\right)^2, \\ \mathbb{E}(S_{n,k}) &= \frac{\theta}{k-1} \\ \text{Var}(S_{n,k}) &= \frac{\theta}{k-1} + \left(\frac{\theta}{k-1}\right)^2, \\ \text{Var}(S_n) &= \theta a_n + \theta^2 b_n, \\ \text{Var}(\hat{\theta}_W) &= \frac{\theta}{a_n} + \theta^2 \frac{b_n}{a_n^2}. \end{aligned}$$

We see that  $\text{Var}(\hat{\theta}_W) \rightarrow 0$ , as  $n \rightarrow \infty$ .

## 5.2 Pairwise mismatches

For  $1 \leq i \neq j \leq n$ , let  $\Pi_{ij}$  denote the number of mismatches between the genome  $i$  and the genome  $j$ , which is the number of mutations which has hit

either  $i$  of  $j$ , but not both jointly. Tajima's estimator is

$$\hat{\theta}_T = \pi_n = \frac{2}{n(n-1)} \sum_{i < j} \Pi_{ij}.$$

We have

$$\begin{aligned} \mathbb{E}[\pi_n] &= \frac{2}{n(n-1)} \sum_{i < j} \mathbb{E}[\Pi_{ij}] \\ &= \mathbb{E}[\Pi_{12}] \\ &= \theta \mathbb{E}[T_2] \\ &= \theta, \end{aligned}$$

so that Tajima's estimator is unbiased. In order to compute the variance of  $\pi_n$ , let us first compute the law of  $\Pi_{12}$ .  $\Pi_{12}$  is the number of mutations on either branch 1 or 2, which happen before those two lineages coalesce. Following the lineages back in time, mutations on the two lineages happen at rate  $\theta$ , and coalescence comes at rate 1. Hence at any time before the coalescence, the next event is a mutation with probability  $\theta/(\theta + 1)$ . Consequently for  $k \geq 0$ ,

$$\mathbb{P}(\Pi_{12} = k) = \left( \frac{\theta}{\theta + 1} \right)^k \frac{1}{\theta + 1}.$$

This is a geometric distribution starting at 0 (sometimes called the "shifted geometric" distribution). Standard results yield

$$\mathbb{E}[\Pi_{12}] = \theta, \quad \text{Var}(\Pi_{12}) = \theta + \theta^2.$$

From this we deduce

**Lemma 5.2.1.** (*Tajima*) *We have*

$$\text{Var}(\pi_n) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2.$$

PROOF: Note that

$$\pi_n^2 = \frac{4}{n^2(n-1)^2} \sum_{i_1 < j_1} \sum_{i_2 < j_2} \Pi_{i_1 j_1} \Pi_{i_2 j_2}.$$

In this double sum, there are three types of terms

$\frac{n(n-1)}{2}$  terms with  $i_1 = i_2, j_1 = j_2$ ,

$n(n-1)(n-2)$  terms with  $i_1 = i_2, j_1 \neq j_2$ , or  $i_1 = j_2$ , or  $j_1 = i_2$ ,

$\frac{n(n-1)(n-2)(n-3)}{4}$  terms with  $\{i_1, j_1\} \cap \{i_2, j_2\} = \emptyset$ .

Define with distinct indices  $i, j, k, \ell$

$$\begin{aligned} U_2 &= \mathbb{E}(\Pi_{ij}^2) - \theta^2, \\ U_3 &= \mathbb{E}(\Pi_{ij}\Pi_{ik}) - \theta^2, \\ U_4 &= \mathbb{E}(\Pi_{ij}\Pi_{k\ell}) - \theta^2. \end{aligned}$$

With these notations we have

$$\text{Var}(\pi_n) = \frac{2}{n(n-1)} \left( U_2 + 2(n-2)U_3 + \frac{(n-2)(n-3)}{2}U_4 \right).$$

The above computations yield  $U_2 = \theta + \theta^2$ . Tajima's strategy consists in computing  $\text{Var}(\pi_3)$  and  $\text{Var}(\pi_4)$ , and use the last formula to deduce  $U_3$  and  $U_4$ . We refer the reader to Tajima's original paper (1983) or Durrett [7] for the details.  $\square$

We note that  $\text{Var}(\hat{\theta}_T) \rightarrow \frac{1}{3}\theta + \frac{2}{9}\theta^2$  as  $n \rightarrow \infty$ .

### 5.3 Tajima's $D$ test statistics

We have seen two unbiased estimates of the same parameter  $\theta$ . It is expected that the difference between those two estimates should be small. Tajima has introduced a normalized version of that difference, namely the quantity

$$D = \frac{\hat{\theta}_T - \hat{\theta}_W}{\sqrt{e_1 S_n + e_2 S_n(S_n - 1)}},$$

where

$$\begin{aligned} e_1 &= \frac{n+1}{3a_n(n-1)} - \frac{1}{a_n^2}, \\ e_2 &= \frac{1}{a_n^2 + b_n} \left( \frac{2(n^2 + n + 3)}{9n(n-1)} - \frac{n+2}{na_n} + \frac{b_n}{a_n^2} \right). \end{aligned}$$

The motivation for this choice of the denominator in the formula for  $D$  is that the variance of the numerator equals  $e_1\theta + e_2\theta^2$ , see [7].

Tajima showed that the distribution of  $D$  is close to a beta distribution.  $|D| \leq 2$  should be interpreted as the fact that the data confirm that the genealogy of our sample is well represented by Kingman's coalescent. Can should be deduce if  $D > 2$ , and if  $D < -2$  ?

Two extreme violations of Kingman's coalescent can be imagined. In the first one, all lineages diverged at an initial time, and evolved independently. In that case any single mutation is counted  $n - 1$  times in the sum of the  $\Pi_{ij}$ 's. Consequently

$$\hat{\theta}_T - \hat{\theta}_W = \frac{2}{n}S_n - \frac{S_n}{a_n} < 0$$

as soon as  $n > 2$ . Suppose now that all but the last coalescence have happened very near the present time, and that the two long branches of the tree support each  $n/2$  of the lineages. Then if we assume that all mutations happen of one of the two long branches,

$$\hat{\theta}_T - \hat{\theta}_W = \frac{n}{2(n-1)}S_n - \frac{S_n}{a_n} > 0.$$

Note that departure from Kingman's coalescent can be in particular the effect of variable population size, or selection.

## 5.4 Two final remarks

In these short notes, we have neglected two very important aspects of population genetics

**Remark 5.4.1. Selection** *So far we have assumed that all mutations are neutral, i. e. that there is no advantage nor disadvantage associated to them. In the case of selective mutations (i. e. mutations which gives a selective advantage – or disadvantage – to those who carry it), the coalescent process is modified by the mutation, or in other words there is an interaction between the process of mutations and the coalescent.*

**Remark 5.4.2. Recombinations** *One important aspect of the genetics of most species is recombinations. The rate of recombinations for human beings is higher than the rate of mutations.*

Going back to the MRCA, besides coalescence events, we have recombination events, which means that a genome splits into two parts, each one “recombining” with a complementary part from another genome. Since our sample is small compared to the total population size, we can assume that all recombinations are done with a genome which does not contain ancestral material to the sample. Taking into account recombinations means that Kingman’s coalescent tree should be replaced by an ancestral recombination graph. While there are  $k$  ancestral to the sample, recombinations happen at rate  $k\rho/2$ , while coalescences happen at rate  $k(k-1)/2$ . The number of ancestors to our sample follows a birth and death process, with birth rate  $k\rho/2$  and death rate  $k(k-1)/2$ . This is a bit simplified, since in that way we may follow lineages which do not contain any genomic material ancestral to the sample. At any rate, this process reaches eventually 1, which means that the MRCA of the sample has been found.

Another way of describing recombinations is to note that Kingman’s coalescent tree is different from one locus of the genome to another one. It is in fact possible to describe the evolution of the coalescent tree along the genome, see Leocard, Pardoux [11].

The Ewens sampling formula is still correct at any particular locus. The various allelic distributions at various loci are conditionally independent given the ancestral recombination graph, but their joint law is still unknown, except for very small samples.

Finally let us comment on the interaction between recombinations and selection. Suppose that an advantageous mutation appears at a particular locus (which we call below the “selective locus”) in one individual of the population. If that mutation happens to get fixed in the population, at the end of the period of fixation (called the selective sweep), all individuals carry that same allele at the advantageous locus. Because recombinations happen during the sweep, the alleles at neutral loci may differ among individuals in the population. However, if the sweep is rather short, a certain number of alleles at neutral loci close to the selective one are identical in all individuals of the population (and identical to the particular alleles which were carried by the individual who experienced the selective mutation). This is called “genetic hitchhiking”, and can be used to detect positive selection.

# Chapter 6

## Appendix

### 6.1 Some elements of stochastic calculus

In these lectures, we use stochastic calculus for two distinct classes of semimartingales. Since we treat almost only scalar-valued processes in this monograph, we present the necessary basic facts from stochastic calculus only in the scalar case.

The first class is the class of continuous semimartingales, whose martingale part is a stochastic integral with respect to Brownian motion. More precisely, let  $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$  be a probability space with the filtration  $\{\mathcal{F}_t, /t \geq 0\}$ , and  $\{B_t, t \geq 0\}$  be a  $\mathcal{F}_t$ -Brownian motion, that is a continuous  $\mathcal{F}_t$ -martingale, which is such that  $B_0 = 0$  and  $B_t^2 - t$  is also a  $\mathcal{F}_t$ -martingale. Suppose now that  $\{\psi_t, \varphi_t, t \geq 0\}$  are  $\mathcal{F}_t$ -progressively measurable processes (this means that for any  $t > 0$ ,  $(\omega, s) \rightarrow (\psi(\omega, s), \varphi(\omega, s))$  is  $\mathcal{F}_t \otimes \mathcal{B}([0, t])$  measurable from  $\Omega \times [0, t]$  into  $\mathbb{R}^d \times \mathbb{R}^{d \times k}$ ), such that for any  $T > 0$ ,

$$\int_0^T [|\psi_t| + |\varphi_t|^2] dt < \infty \quad \text{a. s.},$$

$X_0$  is an  $\mathcal{F}_0$ -measurable random variable, and

$$X_t = X_0 + \int_0^t \psi_s ds + \int_0^t \varphi_s dB_s, \quad t \text{ a. s.}$$

Then we have the Itô formula : for any  $f \in C^2(\mathbb{R})$ ,  $t \geq 0$ ,

$$f(X_t) = f(X_0) + \int_0^t \left[ f'(X_s) \psi_s + \frac{1}{2} f''(X_s) \varphi_s^2 \right] ds + \int_0^t f'(X_s) \varphi_s dB_s.$$

In the case where the process  $\int_0^t \psi_s ds$  is replaced by a more general continuous finite variation (denoted below FV) process  $V_t$ , the Itô formula reads

$$f(X_t) = f(X_0) + \int_0^t f'(X_s) dV_s + \frac{1}{2} \int_0^t f''(X_s) \varphi_s^2 ds + \int_0^t f'(X_s) \varphi_s dB_s,$$

where the first integral on the right is a Stieltjes integral. Let us write  $M_t = \int_0^t \varphi_s dB_s$  for the local martingale part of  $X_t$ . We have that

$$\langle M \rangle_t = [M]_t = \int_0^t \varphi_s^2 ds,$$

where the quadratic variation  $[M]$  of the continuous martingale  $M$  is defined as

$$[M]_t = M_t^2 - 2 \int_0^t M_{s-} dM_s \quad (6.1.1)$$

(the integral is written for the general case of a possibly discontinuous martingale) and the conditional quadratic variation  $\langle M \rangle$  of  $M$  is the unique predictable process such that  $[M]_t - \langle M \rangle_t$  is a martingale. Concerning the predictability property, it is sufficient to know that any progressively measurable and left-continuous process is predictable. In particular, if  $\{X_t, t \geq 0\}$  is progressively measurable and càdlàg, then  $\{X_{t-}, t \geq 0\}$  is predictable.

Let us now write Itô's formula in the case of a continuous semimartingale of the form

$$X_t = X_0 + V_t + M_t,$$

where  $\{V_t\}$  is a finite variation continuous process and  $\{M_t\}$  is a continuous local martingale. If  $f \in C^2(\mathbb{R})$ ,

$$f(X_t) = f(X_0) + \int_0^t f'(X_s) dV_s + \int_0^t f'(X_s) dM_s + \frac{1}{2} \int_0^t f''(X_s) d\langle M \rangle_s. \quad (6.1.2)$$

The second class of semimartingales which we need to use in this monograph is a class of discontinuous finite variation semimartingales. If  $\{X_t, t \geq 0\}$  is a finite variation right-continuous  $\mathbb{R}^d$ -valued process and  $f \in C^1(\mathbb{R})$ ,

$$f(X_t) = f(X_0) + \int_0^t f'(X_{s-}) dX_s + \sum_{0 \leq s \leq t} [f(X_s) - f(X_{s-}) - f'(X_{s-}) \Delta X_s],$$

where  $\Delta X_s = X_s - X_{s-}$ , and the above sum is over those  $s$  such that  $\Delta X_s \neq 0$ . The above formula follows by considering both the evolution of  $f(X_s)$

between the jumps of  $X_s$ , and the jumps of  $f(X_s)$  produced by those of  $X_s$ . Note that in the case  $f(x) = x^2$ , the above formula reduces to

$$(X_t)^2 = (X_0)^2 + 2 \int_0^t X_{s-} dX_s + \sum_{0 \leq s \leq t} (\Delta X_s)^2. \quad (6.1.3)$$

If  $X$  is the sum of a continuous FV process and a FV martingale  $\{M_t\}$ , then

$$\sum_{0 \leq s \leq t} (\Delta X_s)^2 = \sum_{0 \leq s \leq t} (\Delta M_s)^2 = [M]_t.$$

The last identity follows by comparing (6.1.1) and (6.1.3).

We use in this monograph several times the following simple result.

**Lemma 6.1.1.** *Suppose that  $X_0$  is  $\mathcal{F}_0$  measurable,  $\{\varphi_t\}$ ,  $\{\psi_t\}$  are progressively measurable,  $\{M_t\}$  and  $\{N_t\}$  are local martingales, such that a. s. for all  $t \geq 0$ ,*

$$\begin{aligned} X_t &= X_0 + \int_0^t \varphi_s ds + M_t, \\ (X_t)^2 &= (X_0)^2 + \int_0^t [2X_{s-}\varphi_s + \psi_s^2] ds + N_t. \end{aligned}$$

Then  $\langle M \rangle_t = \int_0^t \psi_s^2 ds$ ,  $t \geq 0$ .

PROOF: We give the proof in the case where  $M$  is FV. The proof when  $M$  is continuous follows easily from (6.1.2). The general case, which we do not need in this monograph, follows from similar arguments using the general Itô formula, see [14]. From the first identity in the statement and (6.1.1),

$$(X_t)^2 = (X_0)^2 + 2 \int_0^t X_{s-}\varphi_s ds + 2 \int_0^t X_{s-} dM_s + [M]_t.$$

Comparing with the second identity of the statement, we deduce that

$$[M]_t - \int_0^t \psi_s^2 ds$$

is a local martingale. The result follows since,  $\{\int_0^t \psi_s^2 ds\}$  being progressively measurable and continuous, it is predictable.  $\square$

## 6.2 Tightness in $D$

### 6.2.1 The space $D$

We remind the reader that  $D([0, \infty); \mathbb{R}^d)$  (resp.  $D([0, T]; \mathbb{R}^d)$ ) denotes the vector space of functions from  $[0, \infty)$  (resp.  $[0, T]$ ) into  $\mathbb{R}^d$  which are right continuous and have left limits at every point  $t \geq 0$  (resp.  $0 \leq t \leq T$ ).

We equip the above space with the Skorohod topology, which we now describe. Let us indicate that  $x_n \rightarrow x$  in  $D([0, \infty); \mathbb{R}^d)$  iff there exists a sequence  $\{\lambda_n\}_{n \geq 1}$  of homeomorphisms of  $\mathbb{R}_+$  such that, uniformly on compact subsets of  $\mathbb{R}_+$ ,  $\lambda_n$  converges towards the identity mapping, and  $x_n \circ \lambda_n \rightarrow x$  as  $n \rightarrow \infty$ .

More precisely, the Skorohod topology on  $D([0, T]; \mathbb{R}^d)$  can be defined by the metric ( $\Lambda_T$  denotes the set of all increasing homeomorphisms of  $[0, T]$ )

$$\delta_T(x, y) = \inf_{\lambda \in \Lambda_T} \left[ \sup_{0 \leq t \leq T} |x(t) - y(\lambda(t))| + \sup_{0 \leq t \leq T} |t - \lambda(t)| + \sup_{0 \leq s < t \leq T} \left| \log \frac{\lambda(t) - \lambda(s)}{t - s} \right| \right].$$

For  $N > 0$ ,  $x \in D([0, \infty); \mathbb{R}^d)$ , let

$$\Pi_N(x)(t) := \begin{cases} x(t), & \text{if } 0 \leq t \leq N - 1; \\ (N - t)x(t), & \text{if } N - 1 \leq t \leq N; \\ 0, & \text{if } T > N. \end{cases}$$

The Skorohod topology on  $D([0, \infty); \mathbb{R}^d)$  can be defined by the metric

$$\delta(x, y) = \sum_{N=1}^{\infty} \frac{1}{2^N} [\delta_N(\Pi_N(x), \Pi_N(y)) \wedge 1].$$

Equipped with the Skohorod topology (i. e. with the corresponding metric), both  $D([0, T]; \mathbb{R}^d)$  and  $D([0, \infty); \mathbb{R}^d)$  are complete and separable. It is easily checked that a subset  $A \subset D([0, \infty); \mathbb{R}^d)$  is compact iff  $\Pi_N(A)$  is a compact subset of  $D([0, N]; \mathbb{R}^d)$  for all  $N \geq 1$ .

Note that  $C([0, \infty); \mathbb{R}^d)$  is a closed subset of  $D([0, \infty); \mathbb{R}^d)$  and moreover

**Lemma 6.2.1.** *Suppose  $\{x_n, n \geq 1\} \subset D([0, \infty))$  and  $x_n \rightarrow x$  for the Skorohod topology. If  $x$  is continuous, then  $x_n(t) \rightarrow x(t)$  locally uniformly.*

PROOF: The result follows from the local uniform continuity of  $t \rightarrow x(t)$ .  $\square$

### 6.2.2 Compactness criterion in $D([0, \infty); \mathbb{R}^d)$

We shall be satisfied with a criterion which implies at the same time that the limit is continuous. It is Theorem 15.5, possibly combined with Theorem 8.3, from Billingsley [3], which we recall for the convenience of the reader :

**Proposition 6.2.2.** *Let  $\{X_t^n, t \geq 0\}_{n \geq 1}$  be a sequence of random elements of  $D([0, \infty))$ . A sufficient condition for  $\{X^n\}$  to be tight is that the two conditions (i) and (ii) be satisfied :*

(i)  $\{X_0^n, n \geq 1\}$  is tight in  $\mathbb{R}$ ;

(ii) for any  $T, \varepsilon, \eta > 0$ , there exists  $\delta > 0$  such that

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left( \sup_{0 \leq s \leq t \leq T; t-s \leq \delta} |X_s^n - X_t^n| \geq \varepsilon \right) \leq \eta.$$

A sufficient condition for (ii) is

(ii') for any  $T, \varepsilon, \eta > 0$ , there exists  $\delta > 0$  such that for all  $0 \leq t \leq T$ ,

$$\limsup_{n \rightarrow \infty} \delta^{-1} \mathbb{P} \left( \sup_{t \leq s \leq t+\delta} |X_s^n - X_t^n| \geq \varepsilon \right) \leq \eta.$$

Moreover, if (i) and (ii) are satisfied, then any limit of a converging subsequence is a. s. continuous.

The argument which shows that (ii') implies (ii) is as follows. For any function  $x : [0, T] \rightarrow \mathbb{R}$ ,

$$\left\{ \sup_{0 \leq s \leq t \leq T; t-s \leq \delta} |x(s) - x(t)| \geq \varepsilon \right\} \subset \bigcup_{i=0}^{T/\delta} \left\{ \sup_{i\delta \leq s \leq (i+1)\delta} |x(s) - x(i\delta)| \geq \varepsilon/3 \right\}.$$

We can now state

**Corollary 6.2.3.** *A sufficient condition for a sequence  $\{M_t^n, t \geq 0\}$  of (possibly discontinuous) martingales to be tight is that both*

(j) the sequence  $\{M_0^n, n \geq 1\}_{n \geq 1}$  is tight;

(jj) for all  $T > 0$  there exist  $C(T)$  such that for all  $0 \leq s, t \leq T$ ,

$$\limsup_{n \rightarrow \infty} \mathbb{E} (|M_t^n - M_s^n|^4) \leq C(T)|t - s|^2.$$

PROOF: From Doob's inequality, there exists a constant  $C$  such that for each  $t > 0$ ,

$$P\left(\sup_{t \leq s \leq t+\delta} |M_s^n - M_t^n| \geq \varepsilon\right) \leq C \frac{\mathbb{E}(|M_{t+\delta}^n - M_t^n|^4)}{\varepsilon^4} \leq C'(T) \frac{\delta^2}{\varepsilon^4}.$$

Consequently (jj) implies (ii') in Proposition 6.2.2.  $\square$

### 6.3 de Finetti's theorem

A permutation  $\pi$  of the set  $\{1, 2, \dots\}$  is said to be finite if  $|\{i, \pi(i) \neq i\}| < \infty$ . Let us formulate the

**Definition 6.3.1.** *The countably infinite sequence  $\{X_n, n \geq 1\}$  is said to be exchangeable if for all finite permutation  $\pi$  of  $\{1, 2, \dots\}$ ,*

$$(X_1, X_2, \dots) \stackrel{\mathcal{L}}{=} (X_{\pi(1)}, X_{\pi(2)}, \dots).$$

It is not too hard to show that

**Lemma 6.3.2.** *Given a countably infinite sequence of r. v.'s  $\{X_1, X_2, \dots\}$ , the three following properties are equivalent*

1. *The sequence  $\{X_1, X_2, \dots\}$  is exchangeable.*
2. *For all  $n > 1$ ,*

$$(X_1, \dots, X_{n-1}, X_n, X_{n+1}, \dots) \stackrel{\mathcal{L}}{=} (X_n, \dots, X_{n-1}, X_1, X_{n+1}, \dots).$$

3. *For all sequence  $\{n_i, i \geq 1\}$  of distinct integers,*

$$(X_1, X_2, X_3, \dots) \stackrel{\mathcal{L}}{=} (X_{n_1}, X_{n_2}, X_{n_3}, \dots).$$

Let us recall the well-known "reversed martingale convergence theorem" (see e. g. [3])

**Theorem 6.3.3.** *Let  $\{\mathcal{G}_n, n \geq 1\}$  be a decreasing sequence of sub- $\sigma$ -fields of  $\mathcal{F}$ ,  $\mathcal{G} = \cap_n \mathcal{G}_n$ . Then for any integrable r. v.  $Z$ ,  $\mathbb{E}(Z|\mathcal{G}_n) \rightarrow \mathbb{E}(Z|\mathcal{G})$  a. s.*

We now prove an easy lemma

**Lemma 6.3.4.** *Let  $Y$  be a bounded r. v., and  $\mathcal{H} \subset \mathcal{G}$  be two sub- $\sigma$ -fields of  $\mathcal{F}$ . Then  $\mathbb{E} [\mathbb{E}(Y|\mathcal{H})^2] = \mathbb{E} [\mathbb{E}(Y|\mathcal{G})^2]$  (or a fortiori  $\mathbb{E}(Y|\mathcal{H}) \stackrel{\mathcal{L}}{=} \mathbb{E}(Y|\mathcal{G})$ ) implies that  $\mathbb{E}(Y|\mathcal{H}) = \mathbb{E}(Y|\mathcal{G})$  a. s.*

PROOF: The result follows readily from the identity

$$\mathbb{E} [(\mathbb{E}(Y|\mathcal{G}) - \mathbb{E}(Y|\mathcal{H}))^2] = \mathbb{E} [(\mathbb{E}(Y|\mathcal{G}))^2] - \mathbb{E} [(\mathbb{E}(Y|\mathcal{H}))^2].$$

□

We now state the celebrated de Finetti's theorem. Our proof follows one of the proofs given in [1]. See also [3] for the case of  $\{0, 1\}$ -valued r. v. 's.

**Theorem 6.3.5.** *An exchangeable (countably infinite) sequence  $\{X_n, n \geq 1\}$  of r. v.'s is a mixture of i. i. d. sequences, in the sense that conditionally upon  $\mathcal{T}$  (the tail  $\sigma$ -field of the sequence  $\{X_n\}$ ), the  $X_n$  are i. i. d.*

PROOF: For each  $n \geq 0$ , let  $\mathcal{G}_n := \sigma(X_{n+1}, X_{n+2}, \dots)$ , and let  $\mathcal{T} := \bigcap_n \mathcal{G}_n$  the tail  $\sigma$ -field. By exchangeability, for all  $n \geq 2$ ,

$$(X_1, X_2, X_3, \dots) \stackrel{\mathcal{L}}{=} (X_1, X_{n+1}, X_{n+2}, \dots).$$

Consequently for any bounded Borel measurable function  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ ,  $n \geq 2$ ,

$$\mathbb{E}(\varphi(X_1)|\mathcal{G}_1) \stackrel{\mathcal{L}}{=} \mathbb{E}(\varphi(X_1)|\mathcal{G}_n).$$

Theorem 6.3.3 implies that

$$\mathbb{E}(\varphi(X_1)|\mathcal{G}_n) \rightarrow \mathbb{E}(\varphi(X_1)|\mathcal{T}) \quad \text{a. s., as } n \rightarrow \infty.$$

We deduce that

$$\mathbb{E}(\varphi(X_1)|\mathcal{G}_1) \stackrel{\mathcal{L}}{=} \mathbb{E}(\varphi(X_1)|\mathcal{T}).$$

Now Lemma 6.3.4 implies that the equality holds a. s. This implies that

$$X_1 \quad \text{and} \quad \mathcal{G}_1 \quad \text{are conditionally independent given } \mathcal{T}.$$

The same argument applied to  $(X_n, X_{n+1}, \dots)$  says that for all  $n \geq 1$ ,

$$X_n \quad \text{and} \quad \mathcal{G}_n \quad \text{are conditionally independent given } \mathcal{T}.$$

This implies that the whole sequence  $\{X_n, n \geq 1\}$  is conditionally independent given  $\mathcal{T}$ . Now exchangeability says that for all  $n \geq 1$ ,

$$(X_1, X_{n+1}, X_{n+2}, \dots) \stackrel{\mathcal{L}}{=} (X_n, X_{n+1}, X_{n+2}, \dots).$$

So for the same  $\varphi$ 's as above,

$$\mathbb{E}(\varphi(X_1)|\mathcal{G}_n) = \mathbb{E}(\varphi(X_n)|\mathcal{G}_n) \quad \text{a. s.}$$

Taking the conditional expectation given  $\mathcal{T}$  yields

$$\mathbb{E}(\varphi(X_1)|\mathcal{T}) = \mathbb{E}(\varphi(X_n)|\mathcal{T}) \quad \text{a. s.}$$

Hence, conditionally upon  $\mathcal{T}$ , the  $X_n$  are also identically distributed.  $\square$

It follows from de Finetti's theorem that  $n^{-1} \sum_{k=1}^n X_k$  converges a. s. as  $n \rightarrow \infty$ . Indeed

$$\mathbb{E} \left( n^{-1} \sum_{k=1}^n X_k \text{ converges} \right) = \mathbb{E} \left[ \mathbb{P} \left( n^{-1} \sum_{k=1}^n X_k \text{ converges} | \mathcal{T} \right) \right] = 1.$$

We now deduce the

**Corollary 6.3.6.** *Let  $\{X_n, n \geq 1\}$  be an exchangeable (countably infinite) sequence of  $\{0, 1\}$ -valued r. v.'s. Then, conditionally upon*

$$\text{a. s. } \lim_n n^{-1} \sum_{k=1}^n X_k = x,$$

*the  $X_n$  are i. i. d. Bernoulli with parameter  $x$ .*

# Bibliography

- [1] David Aldous, Exchangeability and related topics, in *Ecole d'Ete St Flour 1983 Lecture Notes in Math.* **1117**, 1–198, Springer 1985.
- [2] Patrick Billingsley, *Convergence of probability measures*, Wiley 1968.
- [3] Patrick Billingsley, *Probability and measures*, 3d ed. Wiley 1995.
- [4] Matthias Birkner, Stochastic models from population biology, lecture notes for a course at TU Berlin, summer 2005 <http://www.wias-berlin.de/people/birkner/smpb-30.6.05.pdf>
- [5] L. Breiman : *Probability*, Addison–Wesley, 1968. New edition SIAM 1992.
- [6] P. Donnelly, T. Kurtz, A countable representation of the Fleming–Viot measure–valued diffusion, *Annals Probab.* **24**, 698–742, 1996.
- [7] Rick Durrett, *Probability models for DNA sequence evolution*, Probability and its applications, Springer 2002.
- [8] Fred Hoppe, Polya–like urns and the Ewens sampling formula, *J. Math. Biol.* **20**, 91–94, 1984.
- [9] J. F. C. Kingman, The coalescent, *Stoch. Proc. Appl.* **13**, 235–248, 1982.
- [10] Amaury Lambert, Population dynamics and random genealogies, *Stoch. Models* **24** 45–163.
- [11] Stéphanie Leocard, Etienne Pardoux, Evolution of the ancestral recombination graph along the genome in case of a selective sweep, *J. Math. Biology*, in press.

- [12] Russell Lyons, Yuval Peres, *Probability on trees and networks*, a book in progress, <http://mypage.iu.edu/~rdlyons/prbtree/prbtree.html>
- [13] Etienne Pardoux, *Processus de Markov et applications*, Dunod, 2007; Engl. translation *Markov processes and applications. Algorithms, networks, genome and finance*, Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester; Dunod, Paris, 2008.
- [14] Philip Protter, *Stochastic integration and differential equations*, 2nd Edition, Applications of Mathematics **21**, Springer, Berlin, 2004.
- [15] Daniel Revuz, Marc Yor, *Continuous martingales and Brownian motion*, 3rd Edition Springer 1999.