J. Appl. Prob. **46,** 669–689 (2009)
Printed in England

ON THE HEIGHT AND LENGTH OF THE ANCESTRAL RECOMBINATION GRAPH

ETIENNE PARDOUX,* ** *Université de Provence*
MAJID SALAMAT,* *** *Université de Provence and Sharif University of Technology*

Abstract

The goal of this paper is to provide formulae for the expectation and variance of the height and length of the ancestral recombination graph (ARG). While the formula for the expectation of the height is known (see, e.g. Krone and Neuhauser (1997)), the other formulae seem to be new. We obtain in particular (see Theorem 4.1) a very simple formula which expresses the expectation of the length of the ARG as a linear combination of the expectations of both the length of the coalescent tree and the height of the ARG. Finally, we study the speed at which the ARG comes down from infinity.

Keywords: Wright–Fisher model; coalescent; recombination; ancestral recombination graph

2000 Mathematics Subject Classification: Primary 60J27
Secondary 60G51; 92D10

1. Introduction and preliminaries

Consider a sample of size n from a population of fixed size N. If the genealogy of the population is described by Canning's model [2] (which generalizes the Wright–Fisher model) or by Moran's model [9] and time is scaled by a factor $1/N$, then, under very mild assumptions on the model, the genealogy of the above sample, looking backwards in time, is described in the limit $N \to \infty$ by Kingman's n-coalescent [7].

If we ignore the partitions (i.e. which genes coalesce at each coalescence event), Kingman's n-coalescent is a death process $\{X_t,\ t \geq 0\}$, where X_t is the number of lineages ancestral to the sample that are alive at time t, starting from $X_0 = n$ and ending at state 1 at the random time $\tau_1 = \inf\{t > 0,\ X_t = 1\}$, when the most recent common ancestor (MRCA) is found. Each death happens at a time when two lineages ancestral to the sample find a common ancestor. The waiting time T_k in state k is exponential with parameter $k(k-1)/2$, the various T_ks being mutually independent. Clearly, $\tau_1 = T_n + T_{n-1} + \cdots + T_2$.

Let us now account for recombinations. At rate $\rho/2$ along each branch of Kingman's coalescent tree, a recombination takes place between an individual from the sample and an individual from outside the sample. Now X_t is a birth-and-death process, since at each recombination the genome of an individual splits into two genomes of two different individuals. Kingman's coalescent tree is replaced by the ancestral recombination graph (ARG). The effect of recombination will be that the ancestral material to a specific DNA sequence comes from two DNA sequences in the parental generation, each of which also came from two different

Received 5 February 2009; revision received 17 May 2009.
* Postal address: LATP, UMR-CNRS 6632, Centre de Mathématiques et d'Informatique, 39 rue F. Joliot-Curie, F-13453, Marseille cedex 13, France.
** Email address: pardoux@cmi.univ-mrs.fr
*** Email address: majid.salamat@gmail.com

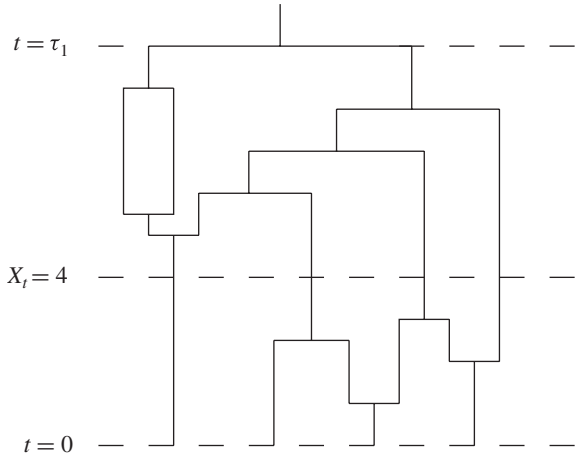

FIGURE 1: The ARG.

grandparents, etc (see Figure 1). In the generation before a sequence was created by a recombination, there would have been one more sequence carrying ancestral material then after. If we focus on a single point on the sequence, it will be inherited from one parent only; thus, the Wright–Fisher model with recombination reduces to the Wright–Fisher model without recombination for each point on the sequence, but different points on the sequence are correlated instances of the Wright–Fisher process without recombination. The tree relating the sequences in a single position is called the local tree of that position. Thus, the genealogy of the whole sequence can be seen as a collection of local trees, one for each position.

Births happen at rate $\rho X_t/2$, while deaths happen at rate $X_t(X_t - 1)/2$. Because the death rate is a quadratic function of X_t, while the birth rate is linear, we can easily show that $\tau_1 = \inf\{t > 0,\ X_t = 1\}$ is finite almost surely (a.s.). We refer the reader to [4]–[6] and [12, Chapter 10] for more complete introductions and descriptions of Kingman's coalescent and the ARG.

Now we define the height of the ARG as $H = \tau_1 = \inf\{t,\ X_t = 1\}$ and the length of the ARG as $L = \int_0^{\tau_1} X_t\,\mathrm{d}t$.

It does not seem possible to give formulae for the laws of H and L. In this paper we compute the first two moments of these random variables. While the formula for the expectation of the height of the ARG (Theorem 2.1) is not new (see [8], in which the analogue of Kingman's coalescent for models with selection rather than recombination was provided, and [12]), we believe that our three other formulae are new. In particular, we obtain a very simple formula which expresses the expectation of the length of the ARG as a linear combination of the expectations of the length of Kingman's coalescent and the height of the ARG.

Let us make precise the fact that we do not specify any model for the splitting of the ancestral genome during a recombination event. Consequently, we do not restrict the ARG to those branches which effectively contain genetic material ancestral to the sample. In other words, τ_1 is the time when the so-called ultimate ancestor (the ancestor of all branches of the ARG) is found, which may very well differ from the MRCA of all the genetic material ancestral to the sample.

Note that a model formally identical to our ARG has been introduced by Krone and Neuhauser [8] under the name of the *ancestral selection graph (ASG)* to model the genealogy

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

of a population where some of the individuals possess a selective advantage. In this model an increase in the sample size while going backwards in time corresponds to the fact that we do not know which branch we should follow, unless we know whether or not the individual we are following backwards in time possesses the selective advantage (this can be decided only when we follow the time forward, after having found the ultimate ancestor of the ASG). In the ASG, individuals follow one or the other branch depending upon whether or not they possess the selective advantage. In the ARG, a particular gene follows one or the other branch, depending upon whether it is located to the left or to the right of the recombination point. At any rate, our results also apply to the ASG.

The first four sections of this paper respectively give formulae for the expectation and variance of the height of the ARG, and the expectation and variance of the length of the ARG. In Sections 6 and 7 we respectively give formulae for the expectation and variance of the number of recombinations.

We write H_n and L_n for the height and the length, respectively, of the ARG with n leaves.

It follows from the formulae below that the expectation of H_n remains bounded as $n \to \infty$. Consequently, the ARG, like Kingman's coalescent, comes down from infinity, in the sense that we can define it with $X_0 = +\infty$, while $X_t < \infty$ for all $t > 0$. It is possible to describe the speed at which the ARG comes down from infinity, through a law of large numbers (LLN) and a central limit theorem (CLT). We show in Section 8 that the ARG satisfies the same LLN and CLT as Kingman's coalescent. This indicates that, asymptotically as $n \to \infty$, the number of recombination events that happen while X_t goes down from n to 1 is of order smaller than n. Nevertheless, the number of recombination events that happen while X_t goes down from $+\infty$ is a.s. infinite. See more on this at the end of Section 6.

In this paper, P_ρ, E_ρ, and var_ρ respectively stand for the probability, the expectation, and the variance in the model with recombination rate $\rho/2$. The case in which $\rho = 0$ corresponds to Kingman's coalescent (no recombination).

2. Expectation of the height of the ARG

Let us first recall the following result. This result is not new; see, e.g. [8] for a proof. We provide a proof since it is the model for some other proofs in this paper.

Theorem 2.1. *The expectation of the height of the ARG for a sample of n individuals is given by*

$$\mathrm{E}_\rho(H_n) = 2\left(1 - \frac{1}{n}\right) + 2\sum_{k=1}^{n-1} \frac{1}{k(k+1)} \frac{\mathrm{e}^\rho}{\rho^{k+1}} \int_0^\rho t^{k+1} \mathrm{e}^{-t}\, \mathrm{d}t.$$

Note that the first term in this formula is well known to be $\mathrm{E}_0(H_n)$, the expectation of the height of Kingman's n-coalescent tree. The second term is thus the expectation of the additional height due to the recombinations.

Proof of Theorem 2.1. Define $U_n = \mathrm{E}_\rho(H_n)$. Clearly, $U_1 = 0$. Let us write a recursion formula for the U_ns. The mean waiting time of X_t in state n is $2/n(n+\rho-1)$, the next state is $n+1$ with probability $\rho/(n+\rho-1)$, and state $n-1$ has probability $(n-1)/(n+\rho-1)$. Consequently, for $n \geq 2$,

$$U_n = \frac{2}{n(n+\rho-1)} + \frac{\rho}{n+\rho-1} U_{n+1} + \frac{n-1}{n+\rho-1} U_{n-1}.$$

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

If we define $W_n = U_n - U_{n-1}$, we obtain the following relation:

$$W_n = (n-2)!\left(2\sum_{k=0}^{m-1}\frac{\rho^k}{(n+k)!} + \frac{\rho^m}{(n+m-2)!}W_{n+m}\right)$$
$$= \frac{2(n-2)!}{\rho^n}\left(\mathrm{e}^\rho - \sum_{k=0}^{n-1}\frac{\rho^k}{k!}\right) + \lim_{m\to\infty}\frac{(n-2)!\,\rho^m}{(n+m-2)!}W_{n+m}.$$

On the other hand, we have

$$W_{n+m} = U_{n+m} - U_{n+m-1} = \mathrm{E}_\rho(H_{n+m}) - \mathrm{E}_\rho(H_{n+m-1}) := \mathrm{E}_\rho(T_{n+m}),$$

where T_{n+m} is thought of as the time until the birth-and-death process started from $n+m$ reaches the value $n+m-1$. Let R_{n+m} be the number of recombinations that occur before the process reaches $n+m-1$, starting at state $n+m$. For $k \geq 1$, we have

$$\mathrm{P}_\rho(R_{n+m} = k) \leq a_k\left(\frac{\rho}{n+m-1}\right)^k,$$

where a_k is the number of distinct sequences of $k-1$ recombinations and $k-1$ coalescences which respect the constraint that there are always at least n lineages alive. The number a_k is the 'Catalan number' (see [11, pp. 172–173]), i.e.

$$a_k = \frac{1}{k+1}\binom{2k}{k} \sim \frac{4^k}{k^{3/2}\sqrt{\pi}}. \tag{2.1}$$

Conditionally upon $R_{n+m} = k$, there are k births and $k+1$ deaths until the process reaches the value $n-1$. Bounding the expectation of the time between two consecutive birth or death events we obtain

$$\mathrm{E}_\rho(T_{n+m} \mid R_{n+m} = k) \leq \frac{2(2k+1)}{(n+m)(n+m-1)}.$$

Moreover, $\mathrm{P}_\rho(R_n = 0) \leq 1$. Finally, provided that $n+m > 1+4\rho$,

$$\mathrm{E}_\rho(T_{n+m}) = \sum_{k=0}^{\infty}\mathrm{E}_\rho(T_{n+m} \mid R_n = k)\,\mathrm{P}_\rho(R_{n+m} = k)$$
$$\leq \frac{c}{(n+m)(n+m-1)}\sum_{k=0}^{\infty}\left(\frac{4\rho}{n+m-1}\right)^k$$
$$\leq \frac{c'}{(n+m)(n+m-1)}.$$

It is now easy to deduce that

$$U_{n+1} - U_n = 2\frac{(n-1)!}{\rho^{n+1}}\sum_{j=n+1}^{\infty}\frac{\rho^j}{j!},$$

and, consequently,

$$U_n = \sum_{k=1}^{n-1}(U_{k+1} - U_k) = 2\sum_{k=1}^{n-1}\frac{(k-1)!}{\rho^{k+1}}\sum_{j=k+1}^{\infty}\frac{\rho^j}{j!},$$

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

since $U_1 = 0$. We now deduce the following formula for $\mathrm{E}_\rho(H_n) = U_n$:

$$\mathrm{E}_\rho(H_n) = 2\sum_{k=1}^{n-1}\sum_{j=0}^{\infty}\frac{(k-1)!}{(k+j+1)!}\rho^j \tag{2.2}$$
$$= 2\left(1-\frac{1}{n}\right) + 2\sum_{k=1}^{n-1}\frac{1}{k(k+1)}\frac{(k+1)!}{\rho^{k+1}}\sum_{\ell=k+2}^{\infty}\frac{\rho^\ell}{\ell!},$$

and the result finally follows from the identity

$$\mathrm{e}^\rho\int_0^\rho t^{k+1}\mathrm{e}^{-t}\,\mathrm{d}t = (k+1)!\left(\mathrm{e}^\rho - \sum_{\ell=0}^{k+1}\frac{\rho^\ell}{\ell!}\right),$$

which is easily checked by successive integrations by parts.

Corollary 2.1. *For small $\rho > 0$,*

$$\mathrm{E}_\rho(H_n) = 2\left(1-\frac{1}{n}\right) + \frac{(n-1)(n+2)}{2n(n+1)}\rho + \frac{(n-1)(n^2+4n+6)}{9n(n+1)(n+2)}\rho^2 + O(\rho^3).$$

Corollary 2.2. *As $n \to \infty$,*

$$\lim_{n\to\infty}\mathrm{E}_\rho(H_n) = \frac{2}{\rho}\int_0^1 \frac{\mathrm{e}^{\rho x}-1}{x}\,\mathrm{d}x.$$

Proof. We have

$$\lim_{n\to\infty}\mathrm{E}_\rho(H_n) = 2\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\frac{\rho^j}{k(k+1)\cdots(k+j+1)}$$
$$= \frac{2}{\rho}\sum_{j=1}^{\infty}\frac{\rho^j}{j(j!)}$$
$$= \frac{2}{\rho}\int_0^\rho \frac{\mathrm{e}^x-1}{x}\,\mathrm{d}x,$$

where the second equality follows from

$$\sum_{k=1}^{\infty}\frac{1}{k(k+1)\cdots(k+j)} = \frac{1}{j(j!)} \quad \text{for all } j \geq 1. \tag{2.3}$$

See Appendix A for a proof of (2.3).

3. Variance of the height of the ARG

Definition 3.1. For all $p, q \in \mathbb{N}$, we define the hypergeometric function ${}_pF_q$ as a mapping from $\mathbb{R}_+^p \times \mathbb{R}_+^q \times \mathbb{R}$ into $\mathbb{R}$ as follows

$${}_pF_q([a_1,\ldots,a_p],[b_1,\ldots,b_q],z) = \sum_{r=0}^{\infty}\frac{(a_1)_r\cdots(a_p)_r}{(b_1)_r\cdots(b_q)_r}\frac{z^r}{r!},$$

where, for all $a \in \mathbb{R}$ and $r \in \mathbb{N}$,

$$(a)_r = a(a+1)\cdots(a+r-1).$$

For more on this subject, see [10, p. 90].

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

Theorem 3.1. *The variance of the height of the ARG is given by*

$$\operatorname{var}_\rho(H_n) = \sum_{p=2}^{n} 4\frac{{}_3F_3([1,\, p,\, p+\rho-1],\,[p+\rho,\, p+1,\, p+1],\,\rho)}{(p+\rho-1)p^2(p-1)} + \sum_{p=2}^{n}\sum_{k=1}^{\infty} \frac{4(p-2)!\,\rho^k}{(p+k-3)!\,(p+k+\rho-2)((p+k-1)^2-1)^2(p+k-1)^2} \times \left(2(p+k-1)+\rho+\frac{(p+k+\rho-2)\mathrm{e}^\rho}{\rho^{p+k}}\int_0^\rho t^{p+k}\mathrm{e}^{-t}\,\mathrm{d}t\right)^2.$$

Proof. See Appendix B.

Note that it can be shown that $\operatorname{var}_\rho(H_n) \le c(\rho) < \infty$ for all $n \ge 2$, where $c(\rho) = \frac{2}{45}\pi^4(\mathrm{e}^\rho + 4\mathrm{e}^{2\rho}(\mathrm{e}^\rho - 1))$.

4. Expectation of the length of the ARG

We now state and prove a very simple formula for the expectation of the length of the ARG.

Theorem 4.1. *The expectation of the length of the ARG is given by*

$$\mathrm{E}_\rho(L_n) = \mathrm{E}_0(L_n) + \rho\,\mathrm{E}_\rho(H_n).$$

Proof. See Appendix C.

Recalling that (in the case in which $\rho = 0$, the ARG reduces to Kingman's coalescent)

$$\mathrm{E}_0(L_n) = 2\left(1+\cdots+\frac{1}{n-1}\right),$$

we deduce the following result from Theorem 4.1.

Corollary 4.1. *For large n,*

$$\lim_{n\to\infty}\mathrm{E}_\rho(L_n) \sim 2\ln(n) + \frac{2}{\rho}\int_0^\rho \frac{\mathrm{e}^x - 1}{x}\,\mathrm{d}x.$$

We note that the additional length produced by the recombinations is bounded in mean as $n \to \infty$.

5. Variance of the length of the ARG

Theorem 5.1. *The variance of the length of the ARG is given by*

$$\operatorname{var}_\rho(L_n) = \sum_{p=2}^{n}\left(4\frac{{}_2F_2([1,\, p+\rho-1],\,[p+\rho,\, p],\,\rho)}{(p+\rho-1)(p-1)} + \sum_{k=1}^{\infty}\frac{(p-2)!\,\rho^{k-1}}{(p+k-3)!}B_{p+k-1}\right),$$

where

$$B_n = \frac{4\rho}{n^2(n-1)^2(n+\rho-1)}\left(2n-1+\frac{2n\rho+\rho^2}{n+1}+\frac{(n+\rho-1)\mathrm{e}^\rho}{(n+1)\rho^n}\int_0^\rho t^{n+1}\mathrm{e}^{-t}\,\mathrm{d}t\right)^2.$$

Proof. See Appendix D.

It can be shown that $\operatorname{var}_\rho(L_n) \le c'(\rho)$ for all $n \ge 2$, where $c'(\rho) \le \frac{2}{3}\pi^2\mathrm{e}^\rho(4\mathrm{e}^\rho + 1)$.

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

6. Expectation of the number of recombinations

As before, we denote by R_n the number of recombinations that happen before the process $\{X_t,\ t \geq 0\}$ reaches the value $n-1$, starting from $X_0 = n$.

Theorem 6.1. *The expectation of R_n is given by*

$$\mathrm{E}_\rho(R_n) = \rho \int_0^1 s^{n-2} \mathrm{e}^{\rho(1-s)}\, \mathrm{d}s.$$

Proof. See Appendix E.

Theorem 6.2. *Let $R(n)$ denote the total number of recombination events in the sample of size n before X_t reaches the value* 1. *We have the identity*

$$\mathrm{E}_\rho(L_n) = \frac{2}{\rho}\,\mathrm{E}_\rho(R(n)).$$

Proof. Starting from identity (2.2), we have

$$\begin{aligned}
\mathrm{E}_\rho(H_n) &= 2\sum_{k=1}^{n-1} \frac{(k-1)!}{\rho^{k+1}} \sum_{j=k+1}^{\infty} \frac{\rho^j}{j!} \\
&= \frac{2}{\rho^2} \sum_{k=1}^{n-1} \frac{(k-1)!}{\rho^{k-1}} \sum_{j=k}^{\infty} \frac{\rho^j}{j!} - \sum_{k=1}^{n-1} \frac{2}{k\rho} \\
&= \frac{2}{\rho^2} \sum_{k=1}^{n-1} \mathrm{E}_\rho(R_{k+1}) - \frac{1}{\rho}\,\mathrm{E}_0(L_n) \\
&= \frac{2}{\rho^2}\,\mathrm{E}_\rho(R(n)) - \frac{1}{\rho}\,\mathrm{E}_0(L_n).
\end{aligned}$$

The result now follows from Theorem 4.1.

Remark 6.1. Note that

$$\frac{\rho}{n-1} < \mathrm{E}_\rho(R_n) < \frac{\rho \mathrm{e}^\rho}{n-1}.$$

This is consistent with

$$\frac{\rho}{n+\rho-1} = \mathrm{P}_\rho(R_n \geq 1) \leq \mathrm{E}_\rho(R_n).$$

Since the R_ns are mutually independent and $\sum_n \mathrm{P}_\rho(R_n \geq 1) = +\infty$, it follows from the Borel–Cantelli lemma that, a.s., infinitely many recombination events occur while the ARG comes down from infinity.

On the other hand, the expectation of the total number of recombination events that occur while X_t goes down from n to 1 equals

$$\sum_{k=2}^{n} \mathrm{E}_\rho(R_k) = \rho \int_0^1 \frac{1-s^{n-1}}{1-s} \mathrm{e}^{\rho(1-s)}\, \mathrm{d}s.$$

This grows, up to a multiplicative constant, like $\rho \ln(n-1)$, while the number of coalescence events grows like n.

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

7. Variance of the number of recombinations

Theorem 7.1. *The variance of the number of recombinations is given by*

$$\mathrm{var}_\rho(R_n) = \frac{\rho^{n-2}}{(n-2)!}\sum_{i=0}^{\infty}\frac{(n+i-1)!}{\rho^{n-i-3}}\prod_{k=0}^{i}\frac{1}{(n+k+\rho-1)^2-\rho(n-1)}.$$

Proof. See Appendix F.

8. The speed at which the ARG comes down from infinity

We have

$$\begin{aligned}\mathrm{E}_\rho(H_n) &= 2\sum_{k=1}^{n-1}\sum_{j=0}^{\infty}\frac{(k-1)!}{(k+j+1)!}\rho^j\\ &= 2\sum_{k=1}^{n-1}\frac{1}{k(k+1)}\sum_{j=0}^{\infty}\frac{\rho^j}{(k+2)\cdots(k+j+1)}\\ &\le 2\sum_{k=1}^{n-1}\frac{1}{k(k+1)}\sum_{j=0}^{\infty}\frac{\rho^j}{j!}\\ &= 2\mathrm{e}^\rho\left(1-\frac{1}{n}\right).\end{aligned}$$

So, for fixed ρ, $\mathrm{E}_\rho(H_n) \le 2\mathrm{e}^\rho$ for all $n \ge 2$. Consequently, $H_\infty = \lim_{n\to\infty} H_n$ is finite a.s. We can then clearly define the population size $\{X_t,\ 0 < t \le \tau_1\}$, where again $\tau_1 = \inf\{t > 0,\ X_t = 1\}$, in such a way that $X_0 = +\infty$, while $X_t < \infty$ for all $t > 0$. Here, as in the introduction, X_t is a birth-and-death process with birth rate $\rho X_t/2$ and death rate $X_t(X_t-1)/2$. Indeed, if we let $\{X_t^n,\ 0 < t \le \tau_1\}$ denote the same process satisfying the initial condition $X_0^n = n$, then we can show that $X_{\cdot\wedge\tau_1} = \lim_{n\to\infty} X^n_{\cdot\wedge\tau_1}$ exists, where the limit is a weak limit for the Skorokhod topology of $D_E[0,+\infty)$, with $E = \{0,1,2,\ldots\}\cup\{+\infty\}$, following the arguments in [3].

The speed at which the ARG comes down from infinity is described by the following result, which contains both an LLN and a CLT.

Theorem 8.1. *For all $\rho \ge 0$, as $t \to 0$,*

$$\frac{tX_t}{2} \to 1 \quad \mathrm{P}_\rho\text{-a.s.}$$

and, moreover, under P_ρ,

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

This theorem says in a sense that X_t is asymptotically $\mathcal{N}(2/t, 2/3t)$ as $t \to 0$. This result does not depend on ρ. It is the same for $\rho > 0$ and $\rho = 0$. This means that the number, R_n, of recombinations that happen before X_t reaches 1, starting with $X_t = n$, is of order smaller than n, as already pointed out at the end of Remark 6.1. Again, denote by T_n the time taken by the process X_t to reach the value $n-1$, starting with $X_t = n$, and define

$$V_n = \sum_{k=n+1}^{\infty} T_k,$$

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

which is the time taken by the process X_t to reach the value n, starting from $X_0 = +\infty$. Clearly,

$$\sum_{n=1}^{\infty} n\,\mathbf{1}_{\{V_n \le t < V_{n-1}\}} \le X_t \le \sum_{n=1}^{\infty} (n + R_n)\,\mathbf{1}_{\{V_n \le t < V_{n-1}\}}.$$

Theorem 8.1 follows from the next result.

Proposition 8.1. *For all $\rho \ge 0$, as $n \to \infty$,*

$$\frac{nV_n}{2} \to 1 \quad \mathrm{P}_\rho\text{-a.s.} \tag{8.1}$$

and, moreover, under P_ρ,

$$\sqrt{3n}\left(\frac{nV_n}{2} - 1\right) \Rightarrow \mathcal{N}(0, 1). \tag{8.2}$$

Proposition 8.1 in turn follows from the next result.

Proposition 8.2. *For all $\rho > 0$,*

$$\mathrm{E}_\rho(|V_n - \mathrm{E}_\rho(V_n)|^4) \le \frac{c(\rho)}{n^6}, \tag{8.3}$$

$$\mathrm{E}_\rho(V_n) = \frac{2}{n} + O\left(\frac{1}{n^2}\right), \tag{8.4}$$

$$n^3 \operatorname{var}_\rho(V_n) \to \tfrac{4}{3} \quad \text{as } n \to \infty. \tag{8.5}$$

Note that the only difference in the statements of Proposition 8.2 between the $\rho > 0$ and $\rho = 0$ cases is that, for the $\rho = 0$ case, (8.4) reads $\mathrm{E}_0(V_n) = 2/n$.

Aldous [1] stated Theorem 8.1 for the case in which $\rho = 0$ (no recombination). The proofs of Theorem 8.1 and Propositions 8.1 and 8.2, in reversed order, will be the subject of the next three subsections.

8.1. Proof of Proposition 8.2

Proof of (8.3). Recall that

$$\mathrm{P}_\rho(R_n = k) \le a_k \left(\frac{\rho}{n-1}\right)^k,$$

where a_k is the Catalan number given by (2.1). So

$$\begin{aligned}
\mathrm{E}_\rho(|V_n - \mathrm{E}_\rho(V_n)|^4) &= \mathrm{E}_\rho\left(\sum_{k=n+1}^{\infty} |T_n - \mathrm{E}_\rho(T_k)|^4\right) \\
&\quad + 6\,\mathrm{E}_\rho\left(\sum_{n<k<l} |T_k - \mathrm{E}_\rho(T_k)|^2 |T_l - \mathrm{E}_\rho(T_l)|^2\right) \\
&= \sum_{k=n+1}^{\infty} \mathrm{E}_\rho(|T_n - \mathrm{E}_\rho(T_k)|^4) \\
&\quad + 6 \sum_{n<k<l} \mathrm{E}_\rho(|T_k - \mathrm{E}_\rho(T_k)|^2)\,\mathrm{E}_\rho(|T_l - \mathrm{E}_\rho(T_l)|^2).
\end{aligned}$$

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

So we have to estimate both $\mathrm{E}_\rho(|T_k - \mathrm{E}_\rho(T_k)|^4)$ and $\mathrm{E}_\rho(|T_k - \mathrm{E}_\rho(T_k)|^2)$. It is not hard to prove that

$$\mathrm{E}_\rho(T_k^2 \mid R_k = m) \le \frac{2^3(2m+1)^2}{(k+1)^2k^2}.$$

Hence,

$$\mathrm{E}_\rho(|T_k - \mathrm{E}_\rho(T_k)|^2) \le \mathrm{E}_\rho(T_k^2) \le \frac{2^3}{(k+1)^2k^2}\left(1 + \frac{c'\rho}{k}\right).$$

By a similar argument,

$$\mathrm{E}_\rho(|T_k - \mathrm{E}_\rho(T_k)|^4) \le \mathrm{E}_\rho(T_k^4) \le \sum_{l=1}^{\infty} \mathrm{E}_\rho(T_k^4 \mid R_k = l) a_l \left(\frac{\rho}{k}\right)^l + \frac{2^5}{k^4(k+1)^4},$$

and standard arguments lead to

$$\mathrm{E}_\rho(T_k^4 \mid R_k = m) \le \frac{2^5(2m+1)^4}{k^4(k+1)^4}.$$

Now we have

$$\mathrm{E}_\rho(|T_k - \mathrm{E}_\rho(T_k)|^4) \le \frac{2^5}{k^4(k+1)^4}\left(1 + \sum_{l=1}^{\infty}(2l+1)^4\left(\frac{4\rho}{k}\right)^l\right).$$

It is easy to show that, for $k > 8\rho$,

$$\sum_{l=1}^{\infty}(l+1)^4\left(\frac{4\rho}{k}\right)^l \le 32\frac{4\rho}{k}. \tag{8.6}$$

Hence,

$$\mathrm{E}_\rho(|T_k - \mathrm{E}_\rho(T_k)|^4) \le \frac{2^5}{k^4(k+1)^4}\left(1 + 32\frac{4\rho}{k}\right). \tag{8.7}$$

Now, by combining (8.6) and (8.7) with the last identity of the previous page, we obtain

$$\begin{aligned}
\mathrm{E}_\rho(|V_n - \mathrm{E}_\rho(V_n)|^4) &\le \sum_{k=n+1}^{\infty} \frac{2^5}{k^4(k+1)^4}\left(1 + \frac{c''\rho}{k}\right) \\
&\quad + 6\sum_{n<k<l}^{\infty} \frac{2^3}{k^2(k+1)^2}\left(1 + \frac{c'\rho}{k}\right)\frac{2^3}{l^2(l+1)^2}\left(1 + \frac{c'''\rho}{l}\right) \\
&\le 2(1 + c''\rho)\sum_{k=n+1}^{\infty} \frac{2^4}{k^4(k+1)^4} + \sum_{n\le k<l} \frac{3\times 2^7(1+c'\rho)(1+c'''\rho)}{k^2(k+1)^2l^2(l+1)^2} \\
&\le \frac{2^5(1+c''\rho)}{7(n-1)^7} + \frac{2^7(1+c'\rho)(1+c'''\rho)}{3(n-1)^6}.
\end{aligned}$$

Proof of (8.4). Since $T_n = H_n - H_{n-1}$, we deduce from Theorem 2.1 that

$$\mathrm{E}_\rho(T_n) = 2\sum_{j=0}^{\infty} \frac{(n-2)!}{(n+j)!}\rho^j.$$

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

Then

$$\mathrm{E}_\rho(V_n) = \frac{2}{n} + 2\sum_{k=n+1}^{\infty}\sum_{j=1}^{\infty}\frac{(k-2)!}{(k+j)!}\rho^j = \frac{2}{n} + O\left(\frac{1}{n^2}\right).$$

Proof of (8.5). Since $H_n = T_n + H_{n-1}$, H_{n-1}, and T_n are independent,

$$\mathrm{var}_\rho(T_n) = \sum_{k=1}^{\infty}\frac{4(n-2)!\,\rho^{k-1}}{(n+\rho+k-2)(n+k-1)^2(n+k-2)!} + \sum_{k=1}^{\infty}\frac{(n-2)!\,\rho^k}{(n+k-3)!\,(n+k+\rho-2)}\left(\sum_{j=0}^{\infty}\frac{2(n+k-3)!\,(2n+2k+j-2)}{(n+k+j)!}\rho^j\right)^2.$$

It is easy to show that

$$\sum_{l=n+1}^{\infty}\left(\sum_{k=1}^{\infty}\frac{4(l-2)!\,\rho^{k-1}}{(l+\rho+k-2)(l+k-1)^2(l+k-2)!}\right) = \sum_{l=n}^{\infty}\frac{4}{(l+\rho)(l+1)^2 l!} + O\left(\frac{1}{n^4}\right)$$

and also that

$$\sum_{l=n+1}^{\infty}\sum_{k=1}^{\infty}\frac{(l-2)!\,\rho^k}{(l+k-3)!\,(l+k+\rho-2)}\left(\sum_{j=0}^{\infty}\frac{2(l+k-3)!\,(2l+2k+j-2)}{(l+k+j)!}\rho^i\right)^2 = O\left(\frac{1}{n^4}\right).$$

Hence,

$$\mathrm{var}_\rho(V_n) = \sum_{l=n}^{\infty}\frac{4}{(l+\rho)(l+1)^2 l} + O\left(\frac{1}{n^4}\right).$$

But

$$\frac{1}{3(n+\rho)^3} = \int_{n+1}^{\infty}\frac{\mathrm{d}x}{(x+\rho)^4} \le \sum_{l=n}^{\infty}\frac{1}{(l+\rho)(l+1)^2 l} \le \int_{n-1}^{\infty}\frac{\mathrm{d}x}{x^4} = \frac{1}{3(n-1)^3},$$

and the result follows.

8.2. Proof of Proposition 8.1

Relation (8.1) follows easily from (8.3), (8.4), and the Borel–Cantelli lemma. We now prove (8.2). It suffices to prove that the sequence

$$Z_n = \frac{\sqrt{3n^3}}{2}(V_n - \mathrm{E}_\rho(V_n))$$

converges in law to $\mathcal{N}(0, 1)$.

Let ϕ_n be the characteristic function of the random variable Z_n, let $c_n = \sqrt{3n^3}/2$, and let $\bar{T}_k = T_k - \mathrm{E}_\rho(T_k)$. For every $t \in \mathbb{R}$, the characteristic function of $\bar{T}_k$ satisfies

$$\Psi_{\bar{T}_k} = 1 - t^2\frac{c_n^2}{2}\mathrm{var}_\rho(\bar{T}_k) - \frac{\mathrm{i}c_n^3 t^3}{6}(\mathrm{E}_\rho((\bar{T}_n)^3) + \delta_k(t)),$$

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

where, for all $k \geq 1$, $\delta_k(t) \to 0$ as $t \to 0$, and $|\delta_k(t)| \leq 2\,\mathrm{E}_\rho(|\bar{T}_k|^3)$ for all $t \in \mathbb{R}$. We have

$$\begin{aligned}
\phi_n(t) &= \mathrm{E}_\rho\left(\exp\left(\mathrm{i}tc_n \sum_{k=n+1}^{\infty} \bar{T}_k\right)\right) \\
&= \prod_{k=n+1}^{\infty} \mathrm{E}_\rho(\exp(\mathrm{i}tc_n\bar{T}_k)) \\
&= \exp\left(\sum_{k=n+1}^{\infty} \log\left(1 - t^2\frac{c_n^2}{2}\,\mathrm{var}_\rho(\bar{T}_k) - \frac{\mathrm{i}c_n^3t^3}{6}(\mathrm{E}_\rho((\bar{T}_k)^3) + \delta_k(t))\right)\right) \\
&= \exp\left(\sum_{k=n+1}^{\infty} \left(-t^2\frac{c_n^2}{2}\,\mathrm{var}_\rho(\bar{T}_k) - \frac{\mathrm{i}c_n^3t^3}{6}(\mathrm{E}_\rho((\bar{T}_k)^3) + \delta_k(t))\right)\right) \\
&= \exp\left(-t^2\frac{c_n^2}{2}\,\mathrm{var}_\rho(V_n) - \sum_{k=n+1}^{\infty} \frac{\mathrm{i}c_n^3t^3}{6}(\mathrm{E}_\rho((\bar{T}_k)^3) + \delta_k(t))\right) \\
&= \exp\left(-t^2\frac{3n^3}{8}\,\mathrm{var}_\rho(V_n) + O\left(\frac{1}{n^{3/2}}\right)\right) \\
&\to \exp\left(-\frac{t^2}{2}\right) \quad \text{as } n \to \infty, \text{ using (8.5).}
\end{aligned}$$

The last equality above follows from

$$\mathrm{E}_\rho(|\bar{T}_k|^3) = \mathrm{E}_\rho(|T_k - \mathrm{E}_\rho(T_k)|^3) \leq (\mathrm{E}_\rho(|T_k - \mathrm{E}_\rho(T_k)|^4))^{3/4} = \frac{c}{k^6}\left(1 + O\left(\frac{1}{k}\right)\right).$$

8.3. Proof of Theorem 8.1

The idea is to use the relations $I_t \leq X_t \leq J_t$, where

$$I_t = \sum_{n=1}^{\infty} n\,\mathbf{1}_{\{V_n \leq t < V_{n-1}\}}, \qquad J_t = \sum_{n=1}^{\infty} (n + R_n)\,\mathbf{1}_{\{V_n \leq t < V_{n-1}\}}.$$

We first show the following result.

Lemma 8.1. *As $t \to 0$,*

$$\sqrt{t}(J_t - I_t) \to 0 \quad \mathrm{P}_\rho\text{-a.s.}$$

Proof. We note that, for all $\varepsilon > 0$,

$$\left\{\limsup_{t\to 0} \sqrt{t}(J_t - I_t) > \varepsilon\right\} \subset \limsup_n A_n,$$

where

$$A_n = \{\sqrt{V_{n-1}}R_n > \varepsilon\}.$$

But

$$\begin{aligned}
\mathrm{P}_\rho(A_n) &\leq \mathrm{P}_\rho(V_{n-1} > \varepsilon^2 n^{-1/4}) + \mathrm{P}_\rho(R_n > n^{1/8}) \\
&\leq \frac{\sqrt{n}}{\varepsilon^4}\,\mathrm{E}_\rho(V_{n-1}^2) + n^{-1/4}\,\mathrm{E}_\rho(R_n) \\
&\leq c(\varepsilon, \rho)n^{-3/2} + \rho\mathrm{e}^\rho n^{-9/8}.
\end{aligned}$$

Consequently, $\sum_n \mathrm{P}_\rho(A_n) < \infty$. The result follows.

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

It now remains to prove Theorem 8.1 with X_t replaced by I_t, i.e. we only have to verify that, as $t \to 0$,

$$\frac{tI_t}{2} \to 1 \quad \mathrm{P}_\rho\text{-a.s.} \tag{8.8}$$

and, moreover, that

$$\sqrt{\frac{6}{t}}\left(\frac{tI_t}{2} - 1\right) \Rightarrow \mathcal{N}(0,1) \quad \text{under } \mathrm{P}_\rho. \tag{8.9}$$

Let us first prove (8.8). We have

$$\left\{\limsup_{t\to 0}\left|\frac{tI_t}{2} - 1\right| > \varepsilon\right\} \subset \limsup_n B_n,$$

where

$$B_n = \left\{\sup_{V_n \le t < V_{n-1}}\left|\frac{tn}{2} - 1\right| > \varepsilon\right\}.$$

Consequently,

$$B_n \subset \left\{\left|\frac{nV_n}{2} - 1\right| > \varepsilon\right\} \cup \left\{\left|\frac{(n-1)V_{n-1}}{2} - 1\right| > \frac{\varepsilon}{2}\right\} \cup \{V_{n-1} > \varepsilon\}.$$

It follows from (8.1) that $\mathrm{P}_\rho(\limsup_n B_n) = 0$ provided that $\varepsilon > 0$. Hence, (8.8) is established.

Let us finally prove (8.9). For all $t > 0$, let

$$\tau(t) = \inf\{0 < s \le t,\ I_s = I_t\}.$$

It follows readily from (8.2) that the relation

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

Combining this with (8.8), we deduce that

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

Equation (8.9) will follow if we prove that

$$\frac{t - \tau(t)}{\sqrt{t}} I_t \to 0 \quad \text{in probability, as } t \to 0,$$

which from (8.8) is equivalent to

$$t^{-3/2}(t - \tau(t)) \to 0 \quad \text{in probability, as } t \to 0.$$

This is a consequence of

$$V_n^{-3/2} T_n \to 0 \quad \text{in probability, as } n \to \infty.$$

Since $nS_n \to 2$ a.s. as $n \to \infty$, it suffices to show that $n^{3/2}T_n$ tends to 0 in probability. But $\mathrm{E}_\rho(T_n) \le c/n^2$. The result follows.

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

Appendix A. Proof of (2.3)

We define

$$C_j := \sum_{k=1}^{\infty} \frac{1}{k(k+1)\cdots(k+j)}.$$

It is easy to show that $C_j - C_{j+1} = C_j - 1/(j+1)! + jC_{j+1}$, so

$$C_{j+1} = \frac{1}{(j+1)(j+1)!} \quad \text{for all } j \geq 0.$$

On the other hand,

$$\sum_{j=1}^{\infty} \frac{\rho^{j-1}}{j!} = \frac{\mathrm{e}^{\rho} - 1}{\rho}; \quad \text{hence} \quad \sum_{j=1}^{\infty} \frac{\rho^j}{j\,(j!)} = \int_0^{\rho} \frac{\mathrm{e}^x - 1}{x}\,\mathrm{d}x.$$

Appendix B. Proof of Theorem 3.1

Let

$$H_n = S_n + H_{n-1}\,\mathbf{1}_{\{\text{coalescence}\}} + H_{n+1}\,\mathbf{1}_{\{\text{recombination}\}},$$

where S_n is the time until the first jump, starting with n individuals. It is easy to show that S_n is independent of $H_{n-1}\,\mathbf{1}_{\{\text{coalescence}\}} + H_{n+1}\,\mathbf{1}_{\{\text{recombination}\}}$; hence,

$$\mathrm{var}_\rho(H_n) = \mathrm{var}_\rho(S_n) + \mathrm{var}_\rho(H_{n-1}\,\mathbf{1}_{\{\text{coalescence}\}} + H_{n+1}\,\mathbf{1}_{\{\text{recombination}\}}).$$

Moreover, since H_{n-1} and the event {coalescence} are independent, as well as H_{n+1} and the event {recombination},

$$\begin{aligned}\mathrm{var}_\rho(H_n) - \mathrm{var}_\rho(H_{n-1}) = {} & \frac{4}{(n+\rho-1)n^2(n-1)} + \frac{\rho}{n-1}(\mathrm{var}_\rho(H_{n+1}) - \mathrm{var}_\rho(H_n)) \\ & + \frac{\rho}{n+\rho-1}(\mathrm{E}_\rho(H_{n+1}) - \mathrm{E}_\rho(H_{n-1}))^2.\end{aligned}$$

But we have

$$\mathrm{E}_\rho(H_{n+1}) - \mathrm{E}_\rho(H_{n-1}) = \sum_{j=0}^{\infty} \frac{2(n-2)!\,(2n+j)}{(n+j+1)!}\rho^j.$$

If we now define $Y_n := \mathrm{var}_\rho(H_n) - \mathrm{var}_\rho(H_{n-1})$, we have

$$Y_n = \frac{4}{(n+\rho-1)n^2(n-1)} + \frac{\rho}{n-1}Y_{n+1} + \frac{\rho}{n+\rho-1}\left(\sum_{j=0}^{\infty} \frac{2(n-2)!\,(2n+j)}{(n+j+1)!}\rho^j\right)^2.$$

Hence,

$$Y_n = \frac{4}{(n+\rho-1)n^2(n-1)} + \frac{\rho}{n-1}Y_{n+1} + A_n,$$

where

$$A_n = \frac{\rho}{n+\rho-1}\left(\sum_{j=0}^{\infty} \frac{2(n-2)!\,(2n+j)}{(n+j+1)!}\rho^j\right)^2.$$

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

It is easy to deduce the following recursion formula for Y_n:

$$Y_n = \sum_{k=1}^{m} \frac{4(n-2)!\,\rho^{k-1}}{(n+\rho+k-2)(n+k-1)^2(n+k-2)!} + \sum_{k=1}^{m} \frac{(n-2)!\,\rho^{k-1}}{(n+k-3)!} A_{n+k-1} + \frac{(n-2)!\,\rho^m}{(n+m-2)!} Y_{n+m}.$$

But we have

$$A_n = \frac{4\rho}{n+\rho-1}\left(\sum_{j=0}^{\infty} \frac{2n}{(n-1)n\cdots(n+j+1)}\rho^j + \sum_{j=0}^{\infty} \frac{j}{(n-1)n\cdots(n+j+1)}\rho^j\right)^2. \tag{B.1}$$

We easily obtain

$$\begin{aligned}
&\sum_{j=0}^{\infty} \frac{\rho^j}{n(n+1)\cdots(n+j+1)} \\
&\quad = \frac{1}{n(n+1)} + \frac{1}{n(n+1)}\left(\frac{\rho}{n+2} + \frac{\rho^2}{(n+2)(n+3)} + \frac{\rho^3}{(n+2)(n+3)(n+4)} + \cdots\right) \\
&\quad = \frac{1}{n(n+1)} + \frac{1}{n(n+1)} \frac{\mathrm{e}^{\rho}}{\rho^{n+1}} \sum_{j=0}^{\infty} (-1)^j \frac{\rho^{n+j+2}}{j!\,(n+j+2)} \\
&\quad = \frac{1}{n(n+1)} + \frac{1}{n(n+1)} \frac{\mathrm{e}^{\rho}}{\rho^{n+1}} \int_0^{\rho} t^{n+1}\mathrm{e}^{-t}\,\mathrm{d}t.
\end{aligned}$$

The second equality follows from

$$\frac{1}{(n+2)(n+3)\cdots(n+j+1)} = \frac{a_2}{n+2} + \frac{a_3}{n+3} + \cdots + \frac{a_{j+1}}{n+j+1},$$

where the coefficients are given by $a_l = (-1)^l/(l-2)!\,(j-l+1)!$.

The first term on the right-hand side of (B.1) can be written as

$$\frac{2n}{n-1} \sum_{j=0}^{\infty} \frac{\rho^j}{n(n+1)\cdots(n+j+1)} = \frac{2}{n^2-1}\left(1 + \frac{\mathrm{e}^{\rho}}{\rho^{n+1}} \int_0^{\rho} t^{n+1}\mathrm{e}^{-t}\,\mathrm{d}t\right),$$

and also

$$\sum_{j=0}^{\infty} \frac{\rho^j}{(n-1)n\cdots(n+j+1)} = \frac{2}{n(n^2-1)}\left(1 + \frac{\mathrm{e}^{\rho}}{\rho^{n+1}} \int_0^{\rho} t^{n+1}\mathrm{e}^{-t}\,\mathrm{d}t\right).$$

Differentiating with respect to ρ and multiplying by ρ, we deduce that

$$\sum_{j=0}^{\infty} \frac{j\rho^j}{(n-1)n\cdots(n+j+1)} = \frac{\rho}{n(n^2-1)}\left(1 + \frac{(\rho-n-1)\mathrm{e}^{\rho}}{\rho^{n+2}} \int_0^{\rho} t^{n+1}\mathrm{e}^{-t}\,\mathrm{d}t\right).$$

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

So we have the following identity:

$$A_n = \frac{4\rho}{(n+\rho-1)(n^2-1)^2 n^2}\left(2n+\rho+\frac{(n+\rho-1)\mathrm{e}^{\rho}}{\rho^{n+1}}\int_0^{\rho} t^{n+1}\mathrm{e}^{-t}\,\mathrm{d}t\right)^2, \tag{B.2}$$

from which we deduce that

$$A_n \le \frac{16\rho}{n^2(n-1)^2(n+\rho-1)}\left(\sum_{j=0}^{\infty}\frac{\rho^j}{j!}\right)^2 \le 16\rho\mathrm{e}^{2\rho}.$$

Hence, $\sum_{k=0}^{\infty} A_{k+2}\rho^k/k!$ converges for all ρ.

Now, by letting m tend to ∞ we have

$$Y_n = \sum_{k=1}^{\infty}\frac{4(n-2)!\,\rho^{k-1}}{(n+\rho+k-2)(n+k-1)^2(n+k-2)!} + \sum_{k=1}^{\infty}\frac{(n-2)!\,\rho^{k-1}}{(n+k-3)!}A_{n+k-1} + \lim_{m\to\infty}\frac{(n-2)!\,\rho^m}{(n+m-2)!}Y_{n+m}.$$

It is easy to check that

$$\sum_{k=1}^{\infty}\frac{4(n-2)!\,\rho^{k-1}}{(n+\rho+k-2)(n+k-1)^2(n+k-2)!} = 4\frac{{}_3F_3([1,n,n+\rho-1],[n+\rho,n+1,n+1],\rho)}{(n+\rho-1)n^2(n-1)}.$$

We need to show that

$$\lim_{m\to\infty}\frac{(n-2)!\,\rho^m}{(n+m-2)!}Y_{n+m} = 0.$$

With the notation introduced in Section 2, we have $H_{n+m} = T_{n+m} + H_{n+m-1}$, and from the strong Markov property, T_{n+m} and H_{n+m-1} are independent. Consequently,

$$\mathrm{var}_{\rho}(H_{n+m}) - \mathrm{var}_{\rho}(H_{n+m-1}) = \mathrm{var}_{\rho}(T_{n+m}) \le \mathrm{E}_{\rho}(T_{n+m}^2).$$

By an argument similar to that used in the proof of Theorem 2.1, we can show that

$$\mathrm{E}_{\rho}(T_{n+m}^2) \le \frac{c'}{(n+m)^2(n+m-1)^2}. \tag{B.3}$$

Consequently,

$$\lim_{m\to\infty}\frac{(n-2)!\,\rho^m}{(n+m-2)!}(\mathrm{var}_{\rho}(H_{n+m}) - \mathrm{var}_{\rho}(H_{n+m-1})) = 0.$$

The theorem follows.

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

Appendix C. Proof of Theorem 4.1

Let $Q_n = \mathrm{E}_\rho(L_n)$. By considering the possible states after the first transition we obtain the recursion formula

$$Q_n = \frac{2}{n+\rho-1} + \frac{\rho}{n+\rho-1} Q_{n+1} + \frac{n-1}{n+\rho-1} Q_{n-1}.$$

It is easy to show that $F_n := \mathrm{E}_0(L_n) + \rho\, \mathrm{E}_\rho(H_n)$ satisfies the same recursion. So we have

$$(n-1)(Q_n - Q_{n-1}) = 2 + \rho(Q_{n+1} - Q_n).$$

If we define $M_n = Q_n - Q_{n-1}$, we obtain the relation

$$M_n = 2 \sum_{k=1}^{m} \frac{\rho^{k-1}}{(n-1)n \cdots (n+k-2)} + \frac{\rho^m}{(n-1)n \cdots (n+m-2)} M_{n+m}.$$

Hence,

$$M_n = 2 \sum_{k=1}^{\infty} \frac{\rho^{k-1}}{(n-1)n \cdots (n+k-2)} + \lim_{m\to\infty} \frac{\rho^m}{(n-1)n \cdots (n+m-2)} M_{n+m}.$$

On the other hand, we have

$$M_{n+m} = Q_{n+m} - Q_{n+m-1} = \mathrm{E}_\rho(L_{n+m}) - \mathrm{E}_\rho(L_{n+m-1}) := \mathrm{E}_\rho(L'_{n+m}).$$

Again, by conditioning upon the value of R_{n+m} we can show that

$$\mathrm{E}_\rho(L'_{n+m}) \le \frac{c'}{(n+m)(n+m-1)}.$$

It is now easy to deduce that

$$\lim_{m\to\infty} \frac{\rho^m (n-2)!}{(n+m-2)!} M_{n+m} = 0 \quad \text{for all } \rho \ge 0.$$

We can easily obtain the relation

$$\begin{aligned} F_n = {} & 2 \sum_{k=1}^{\infty} \frac{\rho^{k-1}}{(n-1)n \cdots (n+k-2)} \\ & + \lim_{m\to\infty} \frac{\rho^m}{(n-1)n \cdots (n+m-2)} (\mathrm{E}_0(L_{n+m}) - \mathrm{E}_0(L_{n+m-1})) \\ & + \lim_{m\to\infty} \frac{\rho^{m+1}}{(n-1)n \cdots (n+m-2)} (\mathrm{E}_\rho(H_{n+m}) - \mathrm{E}_\rho(H_{n+m-1})). \end{aligned}$$

Again, the two limits on the right-hand side vanish. The result follows.

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

Appendix D. Proof of Theorem 5.1

We have, for $n \geq 2$, with the same notation as in Section 3,

$$L_n = nS_n + L_{n-1}\,\mathbf{1}_{\{\text{coalescence}\}} + L_{n+1}\,\mathbf{1}_{\{\text{recombination}\}}.$$

It is easy to show that S_n is independent of $L_{n-1}\,\mathbf{1}_{\{\text{coalescence}\}} + L_{n+1}\,\mathbf{1}_{\{\text{recombination}\}}$; hence,

$$\begin{aligned}\operatorname{var}_\rho(L_n) - \operatorname{var}_\rho(L_{n-1}) = {} & \frac{4}{(n+\rho-1)(n-1)} + \frac{\rho}{n-1}(\operatorname{var}_\rho(L_{n+1}) - \operatorname{var}_\rho(L_n)) \\ & + \frac{\rho}{n+\rho-1}(\mathrm{E}_\rho(L_{n+1}) - \mathrm{E}_\rho(L_{n-1}))^2.\end{aligned}$$

But we have

$$\mathrm{E}_\rho(L_{n+1}) - \mathrm{E}_\rho(L_{n-1}) = \frac{4n-2}{n(n-1)} + \sum_{j=1}^{\infty} \frac{2(n-2)!\,(2n+j-1)}{(n+j)!}\rho^j.$$

Define $D_n := \operatorname{var}_\rho(L_n) - \operatorname{var}_\rho(L_{n-1})$; hence,

$$D_n = \frac{4}{(n+\rho-1)(n-1)} + \frac{\rho}{n-1}Z_{n+1} + B_n,$$

where

$$B_n = \frac{\rho}{n+\rho-1}\left(\frac{4n-2}{n(n-1)} + \sum_{j=1}^{\infty} \frac{2(n-2)!\,(2n+j-1)}{(n+j)!}\rho^j\right)^2.$$

Then

$$D_n = \sum_{k=1}^{m} \frac{4(n-2)!\,\rho^{k-1}}{(n+\rho+k-2)(n+k-2)!} + \sum_{k=1}^{m} \frac{(n-2)!\,\rho^{k-1}}{(n+k-3)!}B_{n+k-1} + \frac{(n-2)!\,\rho^m}{(n+m-2)!}Z_{n+m}.$$

Similarly to the proof of (B.2) we have

$$B_n = \frac{4\rho}{n^2(n-1)^2(n+\rho-1)}\left(2n-1+\frac{2n\rho+\rho^2}{n+1} + \frac{(n+\rho-1)\mathrm{e}^\rho}{(n+1)\rho^n}\int_0^\rho t^{n+1}\mathrm{e}^{-t}\,\mathrm{d}t\right)^2.$$

It is easy to show that $\sum_{k=1}^{\infty} B_{k+2}\rho^k/k!$ converges for all ρ.

It is not hard to show that

$$\sum_{k=1}^{\infty} \frac{4(n-2)!\,\rho^{k-1}}{(n+\rho+k-2)(n+k-2)!} = 4\frac{{}_2F_2([1, n+\rho-1], [n+\rho, n], \rho)}{(n-1)(n+\rho-1)}.$$

Similarly as in Section 3, $L_{n+m} = X_{n+m} + L_{n+m-1}$, where

$$X_{n+m} \leq (n+m+R_{n+m})T_{n+m},$$

again with X_{n+m} and L_{n+m-1} independent, so that

$$\begin{aligned}\operatorname{var}_\rho(L_{n+m}) - \operatorname{var}_\rho(L_{n+m-1}) &= \operatorname{var}_\rho(X_{n+m}) \\ &\leq 2(n+m)^2\,\mathrm{E}_\rho(T_{n+m}^2) + 2\,\mathrm{E}_\rho(R_{n+m}^2 T_{n+m}^2).\end{aligned}$$

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

We deduce from (B.3) that, for large enough m, say $(m+n \geq 8\rho)$,

$$(n+m)^2 \,\mathrm{E}_\rho(T_{n+m}^2) \leq \frac{c_1}{(n+m-1)^2},$$

and also

$$\begin{aligned}\mathrm{E}_\rho(R_{n+m}^2 T_{n+m}^2) &= \sum_{k=1}^{\infty} k^2 \,\mathrm{E}_\rho(T_{n+m}^2 \mid R_{n+m}=k)\,\mathrm{P}_\rho(R_{n+m}=k)\\ &\leq \frac{c_2}{(n+m)(n+m-1)} \sum_{k=1}^{\infty} (k+1)^2 \left(\frac{4\rho}{n+m-1}\right)^k\\ &\leq \frac{c_2'}{(n+m)(n+m-1)}.\end{aligned}$$

Consequently, for all $\rho \geq 0$, as $m \to \infty$,

$$\frac{(n-2)!\,\rho^m}{(n+m-2)!} D_{n+m} \to 0.$$

Therefore, we obtain the relation

$$\mathrm{var}_\rho(L_n) - \mathrm{var}_\rho(L_{n-1}) = 4\frac{{}_2F_2([1, n+\rho-1], [n+\rho, n], \rho)}{(n+\rho-1)(n-l)} + \sum_{k=1}^{\infty} \frac{(n-2)!\,\rho^{k-1}}{(n+k-3)!} B_{n+k-1}.$$

The theorem follows.

Appendix E. Proof of Theorem 6.1

We can obtain the following relation for R_n:

$$R_n = \xi_n(1 + R_n' + R_{n+1}'), \tag{E.1}$$

noting that (ξ_n, R_n', R_{n+1}') is a sequence of independent random variables, ξ_n is a Bernoulli $(\rho/(n+\rho-1))$ random variable, and R_n' and R_{n+1}' are copies of R_n and R_{n+1}, respectively. So we have

$$\mathrm{E}_\rho(R_n) = \frac{\rho}{n+\rho-1}(1 + \mathrm{E}_\rho(R_n) + \mathrm{E}_\rho(R_{n+1})).$$

We can easily deduce the following relation from the above recursion formula:

$$\mathrm{E}_\rho(R_n) = \sum_{k=1}^{m} \frac{(n-2)!\,\rho^k}{(n+k-2)!} + \frac{(n-2)!\,\rho^m}{(n+m-2)!}\,\mathrm{E}_\rho(R_{n+m}).$$

On the one hand, we have

$$\lim_{m\to\infty} \frac{(n-2)!\,\rho^m}{(n+m-2)!}\,\mathrm{E}_\rho(R_{n+m}) = 0,$$

because

$$\mathrm{E}_\rho(R_{n+m}) = \sum_{k=1}^{\infty} k\,\mathrm{P}_\rho(R_{n+m}=k) \leq \sum_{k=1}^{\infty} k a_k \left(\frac{\rho}{n+m-1}\right)^k \leq \frac{4\rho}{n+m-1-4\rho}$$

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

for $n+m-1 \geq 8\rho$, where again a_k is the Catalan number. On the other hand, it is easy to show that

$$\sum_{k=1}^{\infty} \frac{(n-2)!\,\rho^k}{(n+k-2)!} = \frac{\mathrm{e}^{\rho}}{\rho^{n-2}}(\Gamma(n-1) - \Gamma(n-1,\rho)),$$

where $\Gamma(a,x)$ is the incomplete gamma function defined as

$$\Gamma(a,x) = \int_x^{\infty} t^{a-1}\mathrm{e}^{-t}\,\mathrm{d}t.$$

Hence, for all ρ, we obtain

$$\begin{aligned}\mathrm{E}_{\rho}(R_n) &= \frac{\mathrm{e}^{\rho}}{\rho^{n-2}}(\Gamma(n-1) - \Gamma(n-1,\rho)) \\ &= \frac{\mathrm{e}^{\rho}}{\rho^{n-2}} \int_0^{\rho} t^{n-2}\mathrm{e}^{-t}\,\mathrm{d}t \\ &= \rho \int_0^1 s^{n-2}\mathrm{e}^{\rho(1-s)}\,\mathrm{d}s.\end{aligned}$$

The theorem follows.

Appendix F. Proof of Theorem 7.1

From the recursion formula (E.1) we deduce the following formula for the variance of R_n:

$$\begin{aligned}\mathrm{var}_{\rho}(R_n) = &\sum_{i=0}^{m}\prod_{k=0}^{i} \frac{\rho(n+k-1)}{(n+k+\rho-1)^2 - \rho(n-1)} \\ &+ \prod_{k=0}^{m} \frac{\rho(n+k-1)}{(n+k+\rho-1)^2 - \rho(n-1)}\,\mathrm{var}_{\rho}(R_{n+m}).\end{aligned}$$

We have

$$\mathrm{E}_{\rho}(R_{n+m}^2) = \sum_{k=1}^{\infty} k^2 \left(\frac{4\rho}{n+m-1}\right) \leq \sum_{k=1}^{\infty} k^2 a_k \left(\frac{4\rho}{n+m-1}\right) \leq \frac{4(n+m-1)\rho}{(n+m-1-4\rho)^2}$$

for $8\rho \leq n+m-1$. From this we deduce that

$$\mathrm{var}_{\rho}(R_{n+m}) \leq \frac{4(n+m-1)\rho}{(n+m-1-4\rho)^2}.$$

We can easily show that

$$\lim_{m\to\infty} \prod_{k=0}^{m} \frac{\rho(n+k-1)}{(n+k+\rho-1)^2 - \rho(n-1)}\,\mathrm{var}_{\rho}(R_{n+m}) = 0.$$

Hence,

$$\mathrm{var}_{\rho}(R_n) = \sum_{i=0}^{\infty}\prod_{k=0}^{i} \frac{\rho(n+k-1)}{(n+k+\rho-1)^2 - \rho(n-1)}.$$

The result follows after some algebraic simplifications. It is easy to show that

$$\mathrm{var}_{\rho}(R_n) = \frac{\rho(n-1)}{(n-1)^2 + \rho(n-1) + \rho^2} + O\left(\frac{1}{n^2}\right).$$

https://doi.org/10.1239/jap/1253279845 Published online by Cambridge University Press

Acknowledgement

This work was partially supported by the ANR MAEV under contract ANR-06-BLAN-0113.

References

[1] Aldous, D. J. (1999). Deterministic and stochastic models for coalescence (aggregation and coagulation): a review of the mean-field theory for probabilists. *Bernoulli* **5,** 3–48.
[2] Cannings, C. (1974). The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. *Adv. Appl. Prob.* **6,** 260–290.
[3] Donnelly, P. (1991). Weak convergence to a Markov chain with an entrance boundary: ancestral processes in population genetics. *Ann. Prob.* **19,** 1102–1117.
[4] Griffiths, R. C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3,** 479–502.
[5] Griffiths, R. C. and Marjoram, P. (1997). An ancestral recombination graph. In *Progress in Population Genetics and Human Evolution* (IMA Vol. Math. Appl. **87**), eds P. Donnelly and S. Tavaré, Springer, New York, pp. 257–270.
[6] Hein, J., Schierup, M. and Wiuf, C. (2004). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory.* Oxford University Press.
[7] Kingman, J. F. C. (1982). The coalescent. *Stoch. Process. Appl.* **13,** 235–248.
[8] Krone, S. M. and Neuhauser, C. (1997). Ancestral processes with selection. *Theoret. Pop. Biol.* **51,** 210–237.
[9] Moran, P. A. (1958). A general theory of the distribution of gene frequencies. II. Non-overlapping generations. *Proc. R. Soc. London B* **149,** 113–116.
[10] Slater, L. J. (1966). *Generalized Hypergeometric Functions.* Cambridge University Press.
[11] Stanley, R. P. (1999). *Enumerative Combinatorics*, Vol. 2. Cambridge University Press.
[12] Tavaré, S. (2004). Ancestral inference in population genetics. In *Lectures on Probability Theory and Statistics* (Lecture Notes Math. **1837**), Springer, Berlin, pp. 1–188.