Expectation and Variance of ARG

Etienne Pardoux

Université de Provence (Marseille, France)

Majid Salamat

Université de Provence (Marseille, France)

and

Sharif University of Technology (Tehran, Iran)

Abstract

The goal of this paper is to give formulas for the expectation and variance of the height and length of the ancestral recombinaison graph (ARG). The first formula is known, see e. g. [6], the others seem to be new. We obtain in particular (see Theorem 4.1 below) 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.

Key words:
Wright–Fisher model, Coalescent, Recombination, Ancestral Recombination Graph.

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 [1] (which generalizes the Wright–Fisher model) or by Moran's model [7], and time is scaled

Email addresses: pardoux@cmi.univ-mrs.fr (Etienne Pardoux), majid.salamat@gmail.com (Majid Salamat)

by a factor $1/N$, then under very mild assumptions on the model, the genealogy of the above sample, looking backward in time, is described in the limit $N \to \infty$ by Kingman's n–coalescent [5].

If we forget about the exact genealogy (i. e. about who is the brother, cousin of whom ?), Kingman's n–coalescent is a death process $\{X_t,\ t \geq 0\}$, which is the number of lineages ancestral to the sample which are alive at time t, starting from $X_0 = n$, and ending at state 1 at the random time $T_1 = \inf\{t > 0,\ X_t = 1\}$, when the Most Recent Common Ancestor is found. Each death happens at a time when two lineages ancestral to the sample find a common ancestor. The waiting time S_k in state k is exponential with parameter $k(k-1)/2$, the S_k being independent for different k. Clearly $T_1 = S_n + S_{n-1} + \cdots + S_2$.

Let us now account for recombinations. At rate $\theta/2$ 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, abbreviated ARG.

Births happen at rate $\theta X_t/2$, 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, one easily shows that $T_1 = \inf\{t > 0,\ X_t = 1\}$ is finite a. s. We refer to [4], [2], [3] and [10] for more complete introductions and descriptions of Kingman's coalescent and the ARG.

Now we define the height of the ARG as $H = T_1 = \inf\{t, X_t = 0\}$ and the lenght of the ARG as $L = \int_0^{T_1} X_t dt$.

The aim of this paper is to compute the first two moments of the height and length of the ARG. While the formula for the expectation of the height of the ARG is not new (see [10], and [6] where the replacement of Kingman's coalescent by a graph models selection rather than recombination), we believe that our three other formulas are new. We in particular obtain a very simple formula which expresses the expectation of the length of the ARG as a linear combination of that of Kingman's coalescent, and of the expectation of 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, T_1 is the time when

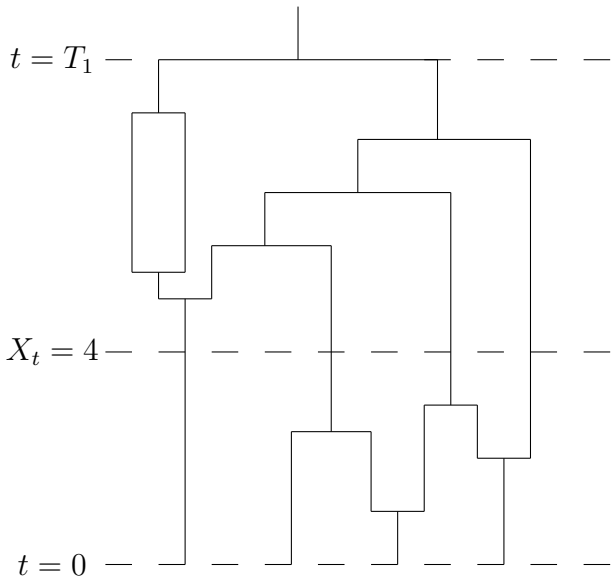

Figure 1: ARG

the so–called Ultimate Ancestor (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.

The fours sections of this paper gives formulas for respectively the expectation and variance of the height of the ARG, the expectation and variance of the length of the ARG.

We write H_n (resp. L_n) for the height (resp. the length) of the ARG with n leaves.

2. The expectation of the height of ARG

Let us first recall that (also this result is not new, we provide a proof which 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*

$$\mathbb{E}_\theta(H_n) = 2\left(1 - \frac{1}{n}\right) + 2\sum_{k=1}^{n-1} \frac{1}{k(k+1)} \frac{e^\theta}{\theta^{k+1}} \int_0^\theta t^{k+1} \exp(-t) dt.$$

Proof. Define, $U_n = \mathbb{E}_\theta(H_n)$. Let us write a recursion formula for the U_n's. The mean waiting time of X_t in state n is $\frac{2}{n(n+\theta-1)}$, the next state is $n+1$ with probability $\frac{\theta}{n+\theta-1}$, $n-1$ with probability $\frac{n-1}{n+\theta-1}$. Consequently

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

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

$$\begin{aligned} W_n &= (n-2)!\left(2\sum_{k=0}^{m-1}\frac{\theta^k}{(n+k)!} + \frac{\theta^m}{(n+m-2)!}W_{n+m}\right) \\ &= \frac{2(n-2)!}{\theta^n}\left(e^\theta - \sum_{k=0}^{n-1}\frac{\theta^k}{k!}\right) + \lim_{m\to\infty}\frac{(n-2)!\theta^m}{(n+m-2)!}W_{n+m}. \end{aligned}$$

On the other hand, we have

$$W_{n+m} = U_{n+m} - U_{n+m-1} = \mathbb{E}_\theta(H_{n+m}) - \mathbb{E}_\theta(H_{n+m-1}) := \mathbb{E}_\theta(T_{n+m-1})$$

where T_{n+m-1} is the time take by the Birth and Death process to reach the value $n+m-1$, starting from $n+m$. Let R_{n+m} be the number of recombinations which occur before the process reaches $n+m-1$, starting at state $n+m$. So for $k \geq 1$ we have

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

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 alive lineages. It is the "Catalan number" (see [9])

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

Conditionally upon $\{R_{n+m} = k\}$, $k \geq 0$, 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 of those events we obtain

$$\mathbb{E}_\theta(T_{n+m-1}\,|R_{n+m} = k) \leq \frac{2(2k+1)}{(n+m)(n+m-1)}.$$

Moreover $\mathbb{P}_\theta(R_n = 0) \leq 1$. Finally

$$\begin{aligned}\mathbb{E}_\theta(T_{n+m-1}) &= \sum_{k=0}^{\infty} \mathbb{E}_\theta(T_{n+m-1}\,|R_n = k)\,\mathbb{P}_\theta(R_{n+m} = k) \\ &\leq \frac{c}{(n+m)(n+m-1)} \sum_{k=0}^{\infty} \left(\frac{4\theta}{n+m-1}\right)^k \leq \frac{c'}{(n+m)(n+m-1)}.\end{aligned}$$

It is now easy to deduce that $U_{n+1} - U_n = 2\frac{(n-1)!}{\theta^{n+1}} \sum_{j=n+1}^{\infty} \frac{\theta^j}{j!}$ and consequently

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

We finally deduce the following formula for $\mathbb{E}_\theta(H_n) = U_n$.

$$\mathbb{E}_\theta(H_n) = 2\sum_{k=1}^{n-1}\sum_{j=0}^{\infty} \frac{(k-1)!}{(k+j+1)!}\theta^j = 2\left(1 - \frac{1}{n}\right) + 2\sum_{k=1}^{n-1} \frac{1}{k(k+1)} \frac{(k+1)!}{\theta^{k+1}} \sum_{\ell=k+2}^{\infty} \frac{\theta^\ell}{\ell!}$$

and the result finally follows from the following identity, which is easily checked by successive integrations by parts

$$e^\theta \int_0^\theta t^{k+1} \exp(-t)dt = (k+1)! \left(e^\theta - \sum_{\ell=0}^{k+1} \frac{\theta^\ell}{\ell!}\right).$$

□

Corollary 2.2. *For small $\theta > 0$*

$$\mathbb{E}_\theta(H_n) = 2(1 - \frac{1}{n}) + \frac{(n-1)(n+2)}{2n(n+1)}\theta + \frac{(n-1)(n^2+4n+6)}{9n(n+1)(n+2)}\theta^2 + O(\theta^3).$$

Corollary 2.3. *As $n \to \infty$*

$$\lim_{n\to\infty} \mathbb{E}_\theta(H_n) = \frac{2}{\theta} \int_0^\theta \frac{e^x - 1}{x} dx.$$

Proof.

$$\lim_{n\to\infty} \mathbb{E}_\theta(H_n) = 2\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \frac{\theta^j}{k(k+1)\cdots(k+j+1)} = \frac{2}{\theta}\sum_{j=1}^{\infty} \frac{\theta^j}{j \cdot j!} = \frac{2}{\theta}\int_0^\theta \frac{e^x - 1}{x} dx$$

where the second equality follows from

$$\sum_{k=1}^{\infty} \frac{1}{k(k+1)\cdots(k+j)} = \frac{1}{j \cdot j!}, \quad \forall j \geq 1. \tag{2.1}$$

See the Appendix for a proof. □

As Kingman's coalescent, the ARG "comes down from infinity", i.e. we can define an ARG with infinitely many leaves. We hope to discuss some questions related to that property in another publication.

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

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

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 [8].

Theorem 3.2. *The variance of the height of the ARG is given by*

$$\begin{aligned} Var_\theta(H_n) = \sum_{p=2}^{n} 4 \frac{{}_3F_3\ ([1, p, p+\theta-1], [p+\theta, p+1, p+1], \theta)}{(p+\theta-1)p^2(p-1)} \\ + \sum_{p=2}^{n} \sum_{k=1}^{\infty} \frac{4(p-2)!\theta^k}{(p+k-3)!(p+k+\theta-2)((p+k-1)^2-1)^2(p+k-1)^2} \\ \times \left(2(p+k-1) + \theta + \frac{(p+k+\theta-2)e^\theta}{\theta^{p+k}} \int_0^\theta t^{p+k} e^{-t} dt\right)^2. \end{aligned}$$

Proof.

$$H_n = S_n + H_{n-1}\mathbb{I}_{\{Coalescence\}} + H_{n+1}\mathbb{I}_{\{Recombination\}}$$

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 $H_{n-1}\mathbb{I}_{\{Coalescence\}}+H_{n+1}\mathbb{I}_{\{Recombination\}}$, hence

$$\mathrm{Var}_\theta(H_n) = \mathrm{Var}_\theta(S_n) + \mathrm{Var}_\theta(H_{n-1}\mathbb{I}_{\{Coalescence\}} + H_{n+1}\mathbb{I}_{\{Recombination\}})$$

Since moreover H_{n-1} and the event $\{Coalescence\}$ are independent, as well as H_{n+1} and the event $\{Recombination\}$,

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

But we have

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

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

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

Hence

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

where

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

It is easy to deduce the following recursion formula for Y_n

$$\begin{aligned}Y_n =& \sum_{k=1}^{m} \frac{4(n-2)!\theta^{k-1}}{(n+\theta+k-2)(n+k-1)^2(n+k-2)!} \\ &+ \sum_{k=1}^{m} \frac{(n-2)!\theta^{k-1}}{(n+k-3)!}A_{n+k-1} + \frac{(n-2)!\theta^m}{(n+m-2)!}Y_{n+m}.\end{aligned}$$

We have the identity (see the Appendix below)

$$A_n = \frac{4\theta}{(n+\theta-1)(n^2-1)^2n^2}\left(2n+\theta+\frac{(n+\theta-1)e^\theta}{\theta^{n+1}}\int_0^\theta t^{n+1}e^{-t}dt\right)^2 \tag{3.1}$$

from which we deduce that

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

Hence $\sum_{k=0}^{\infty}\frac{A_{k+2}\theta^k}{k!}$ converges for all θ. Now, by letting m tends to ∞, we have the following

$$\begin{aligned} Y_n = &\sum_{k=1}^{\infty}\frac{4(n-2)!\theta^{k-1}}{(n+\theta+k-2)(n+k-1)^2(n+k-2)!} \\ &+\sum_{k=1}^{\infty}\frac{(n-2)!\theta^{k-1}}{(n+k-3)!}A_{n+k-1}+\lim_{m\to\infty}\frac{(n-2)!\theta^m}{(n+m-2)!}Y_{n+m}. \end{aligned}$$

It is easy to check that

$$\begin{aligned} &\sum_{k=1}^{\infty}\frac{4(n-2)!\theta^{k-1}}{(n+\theta+k-2)(n+k-1)^2(n+k-2)!} \\ &\qquad = 4\frac{{}_3F_3\left([1,n,n+\theta-1],[n+\theta,n+1,n+1],\theta\right)}{(n+\theta-1)n^2(n-1)}. \end{aligned}$$

We need to show that

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

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

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

By an argument similar to that in the proof of theorem 2, one can show that

$$\mathbb{E}_\theta(T_{n+m-1}^2)\le\frac{c'}{(n+m)^2(n+m-1)^2}, \tag{3.2}$$

Consequently

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

The theorem follows. □

4. Expectation of the length of the ARG

Theorem 4.1. *The expectation of the length of the ancestral recombinaison graph is given by*

$$\mathbb{E}_\theta(L_n) = \mathbb{E}_0(L_n) + \theta\ \mathbb{E}_\theta(H_n).$$

Proof. We define, $V_n = \mathbb{E}_\theta(L_n)$. The following recursion formula for V_n follows by considering the possible states after the first transition

$$V_n = \frac{2}{n+\theta-1} + \frac{\theta}{n+\theta-1}V_{n+1} + \frac{n-1}{n+\theta-1}V_{n-1}.$$

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

$$(n-1)(V_n - V_{n-1}) = 2 + \theta(V_{n+1} - V_n).$$

If we define $Z_n = V_n - V_{n-1}$, we obtain the following relation

$$Z_n = 2\sum_{k=1}^{m} \frac{\theta^{k-1}}{(n-1)n\cdots(n+k-2)} + \frac{\theta^m}{(n-1)n\cdots(n+m-2)}Z_{n+m}.$$

Hence

$$Z_n = 2\sum_{k=1}^{\infty} \frac{\theta^{k-1}}{(n-1)n\cdots(n+k-2)} + \lim_{m\to\infty}\frac{\theta^m}{(n-1)n\cdots(n+m-2)}Z_{n+m}.$$

On the other hand, we have

$$Z_{n+m} = V_{n+m} - V_{n+m-1} = \mathbb{E}_\theta(L_{n+m}) - \mathbb{E}_\theta(L_{n+m-1}) := \mathbb{E}_\theta(L_{n+m-1})$$

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

$$\mathbb{E}_\theta(L_{n+m-1}) \le \frac{c'}{(n+m)(n+m-1)}.$$

It is now easy to deduce that

$$\lim_{m\to\infty} \frac{\theta^m (n-2)!}{(n+m-2)!} Z_{n+m} = 0,\ \forall \theta \geq 0.$$

We can easily obtain the following relation

$$\begin{aligned} K_n = 2\sum_{k=1}^{\infty} & \frac{\theta^{k-1}}{(n-1)n\cdots(n+k-2)} \\ & + \lim_{m\to\infty} \frac{\theta^m}{(n-1)n\cdots(n+m-2)} \left(\mathbb{E}_0(L_{n+m}) - \mathbb{E}_0(L_{n+m-1})\right) \\ & + \lim_{m\to\infty} \frac{\theta^{m+1}}{(n-1)n\cdots(n+m-2)} \left(\mathbb{E}_\theta(H_{n+m}) - \mathbb{E}_\theta(H_{n+m-1})\right), \end{aligned}$$

Again the two limits on the right vanish. The result follows. □

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

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

we deduce from the last Theorem

Corollary 4.2. *For large n,*

$$\lim_{n\to\infty} \mathbb{E}_\theta(L_n) \sim 2\ln(n) + \frac{2}{\theta}\int_0^\theta \frac{e^x - 1}{x} dx.$$

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 ancestral recombinaison graph is given by*

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

where

$$B_n = \frac{4\theta}{n^2(n-1)^2(n+\theta-1)} \left(2n - 1 + \frac{1}{n+1}(2n\theta + \theta^2 + \frac{(n+\theta-1)e^\theta}{\theta^n}\int_0^\theta t^{n+1}e^{-t}dt)\right)^2.$$

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

$$L_n = nS_n + L_{n-1}\mathbb{I}_{\{Coalescence\}} + L_{n+1}\mathbb{I}_{\{Recombination\}}.$$

It is easy to show that S_n is independent of $L_{n-1}\mathbb{I}_{\{Coalescence\}}+L_{n+1}\mathbb{I}_{\{Recombination\}}$, hence

$$\begin{aligned}\mathrm{Var}_\theta(L_n) - \mathrm{Var}_\theta(L_{n-1}) =& \frac{4}{(n+\theta-1)(n-1)} + \frac{\theta}{n-1}[\mathrm{Var}_\theta(L_{n+1}) - \mathrm{Var}_\theta(L_n)] \\ &+ \frac{\theta}{n+\theta-1}[\mathbb{E}_\theta(L_{n+1}) - \mathbb{E}_\theta(L_{n-1})]^2.\end{aligned}$$

But we have

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

Define $D_n := \mathrm{Var}_\theta(L_n) - \mathrm{Var}_\theta(L_{n-1})$, hence

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

where

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

Then

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

So we have the following whose proof is similar to that of (3.1)

$$B_n = \frac{4\theta}{n^2(n-1)^2(n+\theta-1)}\left(2n-1+\frac{1}{n+1}(2n\theta+\theta^2+\frac{(n+\theta-1)e^\theta}{\theta^n}\int_0^\theta t^{n+1}e^{-t}dt)\right)^2. \tag{5.1}$$

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

It is not hard to show that

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

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

$$X_{n+m} \leq (n+m+R_{n+m})T_{n+m-1}$$

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

$\mathrm{Var}_\theta(L_{n+m}) - \mathrm{Var}_\theta(L_{n+m-1}) = \mathrm{Var}_\theta(X_{n+m}) \leq 2(n+m)^2\mathbb{E}_\theta(T^2_{n+m-1}) + 2\mathbb{E}_\theta(R^2_{n+m}T^2_{n+m-1})$.

We deduce from (3.2) that for m large enough say $(m+n \geq 8\theta)$

$$(n+m)^2\mathbb{E}_\theta(T^2_{n+m-1}) \leq \frac{c_1}{(n+m-1)^2},$$

and also

$$\begin{aligned}\mathbb{E}_\theta(R^2_{n+m}T^2_{n+m-1}) &= \sum_{k=1}^{\infty} k^2\mathbb{E}_\theta(T^2_{n+m-1}\,|R_{n+m}=k)\,\mathbb{P}_\theta(R_{n+m}=k)\\ &\leq \frac{c_2}{(n+m)(n+m-1)}\sum_{k=1}^{\infty}(k+1)^2\left(\frac{4\theta}{n+m-1}\right)^k \leq \frac{c_2'}{(n+m)(n+m-1)}.\end{aligned}$$

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

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

Therefore we obtain the following relation

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

The Theorem follows. □

A. Proof of (2.1)**.**

We define

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

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

$$C_{j+1} = \frac{1}{(j+1)(j+1)!}, \forall j \geq 0.$$

On the other hand,

$$\sum_{j=1}^{\infty}\frac{\theta^{j-1}}{j!} = \frac{e^\theta - 1}{\theta} \text{ hence } \sum_{j=1}^{\infty}\frac{\theta^j}{j.j!} = \int_0^\theta \frac{e^x-1}{x}dx.$$

B. Proof of (3.1)**.**

$$\begin{aligned}
&\sum_{j=0}^{\infty} \frac{\theta^j}{k(k+1)\cdots(k+j+1)} \\
&\quad = \frac{1}{k(k+1)} + \frac{1}{k(k+1)}\left[\frac{\theta}{k+2} + \frac{\theta^2}{(k+2)(k+3)} + \frac{\theta^3}{(k+2)(k+3)(k+4)} + \cdots\right] \\
&\quad = \frac{1}{k(k+1)} + \frac{1}{k(k+1)}\frac{e^\theta}{\theta^{k+1}} \sum_{j=0}^{\infty} (-1)^j \frac{\theta^{k+j+2}}{j!(k+j+2)} \\
&\quad = \frac{1}{k(k+1)} + \frac{1}{k(k+1)}\frac{e^\theta}{\theta^{k+1}} \int_0^\theta t^{k+1}\exp(-t)dt.
\end{aligned}$$

The second identity follows from

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

where the coefficients are given by $a_l = \frac{(-1)^l}{(l-2)!(j-l+1)!}$. By using the above relation we obtain

$$A_k = \frac{4\theta}{k+\theta-1}\left(\sum_{j=0}^{\infty} \frac{2k}{(k-1)k\cdots(k+j+1)}\theta^j + \sum_{j=0}^{\infty} \frac{j}{(k-1)k\cdots(k+j+1)}\theta^j\right)^2.$$

The first term in the right can be written as

$$\frac{2k}{k-1}\sum_{j=0}^{\infty} \frac{\theta^j}{k(k+1)\cdots(k+j+1)} = \frac{2}{k^2-1}\left(1 + \frac{e^\theta}{\theta^{k+1}}\int_0^\theta t^{k+1}e^{-t}dt\right),$$

and also

$$\sum_{j=0}^{\infty} \frac{\theta^j}{(k-1)k\cdots(k+j+1)} = \frac{2}{k(k^2-1)}\left(1 + \frac{e^\theta}{\theta^{k+1}}\int_0^\theta t^{k+1}e^{-t}dt\right).$$

Differentiating with respect to θ and multipling by θ and combinig this identity with that we obtained, we deduce (3.1)

$$\sum_{j=0}^{\infty} \frac{j\theta^j}{(k-1)k\cdots(k+j+1)} = \frac{\theta}{k(k^2-1)}\left(1 + \frac{(\theta-k-1)e^\theta}{\theta^{k+2}}\int_0^\theta t^{k+1}e^{-t}dt\right).$$

References

[1] C. Cannings, *The latent roots of certain Markov chains arising in genetics: A new approach. I. Haploid models* , Adv. Appl. Prob. 1974, 6: 260290.

[2] R. C. Griffiths and P. Marjoram, *Ancestral inference from samples of DNA sequences with recombination* , J. Comput. Biol., 1996, 3, 479-502.

[3] R. C. Griffiths and P. Marjoram, An ancestral recombination graph, in *Progress in population genetics and human evolution*, IMA Volumes in Mathematics and its applications **87**, P. Donnelly and S. Tavaré eds., 257–270, Springer–Verlag, 1997.

[4] J. Hein, M. Schierup and C. Wiuf, *Gene genealogies, variation and evolution : a primer in coalescent theory*, Oxford university press, 2004.

[5] J.F.C, Kingman, *The coalescent*, Stoch. Proc. Applns. 1982, 13: 235248.

[6] Stephen M. Krone and Claudia Neuhauser, *Ancestral Processes with Selection*, Theoretical Population Biology, 1997, 51, 210-237.

[7] P.A. Moran, *A general theory of the distribution of gene frequencies. I. Nonoverlapping generations*, Proc. R. Soc. Lond. B Biol. Sci., 1958, 149(934), 113116.

[8] L.J Slater, *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge 1966.

[9] R.P. Stanley, *Enumerative combinatorics*, Vol 2, Cambridge Univ. Press, 1999.

[10] S. Tavaré, *Ancestral Inference in Population Genetics*, Proceedings of Saint Flour Summer School in Probability and Statistics 2001, Springer, Lecture Notes in Mathematics.