ALEA, Lat. Am. J. Probab. Math. Stat. **21**, 725–739 (2024)
DOI: 10.30757/ALEA.v21-28

On a Markov chain related to the individual lengths in the recursive construction of Kingman's coalescent

Linglong Yuan

Department of Mathematical Sciences, University of Liverpool, Liverpool, United Kingdom.
E-mail address: linglong.yuan@liverpool.ac.uk

Abstract. Kingman's coalescent is a widely used process to model sample genealogies in population genetics. Recently there have been studies on the inference of quantities related to the genealogy of additional individuals given a known sample. This paper explores the recursive (or sequential) construction which is a natural way of enlarging the sample size by adding individuals one after another to the sample genealogy via individual lineages to construct Kingman's coalescent. Although the process of successively added lineage lengths is not Markovian, we show that it contains a Markov chain which records the information of the successive largest lineage lengths and we prove a limit theorem for this Markov chain.

1. Introduction

The coalescent theory was introduced by Kingman (Kingman, 1982) and has since then become a standard framework to model sample genealogies. The Kingman's n-coalescent with $n \geq 1$, denoted by $\Pi^n = (\Pi^n(t))_{t\geq 0}$, is a continuous-time Markov process with state space $\mathcal{P}(n)$, the set of partitions of $[n] := \{1, 2, \cdots, n\}$. It starts at time 0 with the partition of singletons $\{\{1\}, \{2\}, \cdots, \{n\}\}$, and at any time, any two blocks merge into one at rate 1 independently. Eventually, the coalescent reaches the final state $\{\{1, 2, \cdots, n\}\}$, called the most recent common ancestor (MRCA), and stays there forever. We set by convention that $\Pi^1(t) = \{\{1\}\}$ for any $t \geq 0$.

We introduce further notations: if π is a partition of a set of integers, let $|\pi|$ be the number of blocks in π; let $\mathbb{Z}_+ = \{1, 2, \cdots\}$ and $\mathcal{P}(\infty)$ be the set of partitions of $\mathbb{Z}_+$.

The Kingman's n-coalescents are consistent: for any $m > n \geq 1$, if we consider the natural restriction of Π^m to the partitions in $\mathcal{P}(n)$, then the resulting new process has the same law as Π^n, thus independent of m. The consistency property will allow to construct the Kingman's (infinite) coalescent $\Pi^\infty = (\Pi^\infty(t))_{t\geq 0}$ which starts at time 0 with the set of singletons $\{\{1\}, \{2\}, \cdots\} \in \mathcal{P}(\infty)$, and the restriction of Π^∞ to $\mathcal{P}(n)$ has the same law as Π^n for all $n \geq 1$. This Π^∞ can be constructed using Kolmogorov's extension theorem, see Berestycki (2009, Proposition 2.1).

The process Π^∞ can also be constructed naturally by first giving Π^1, and conditionally on Π^n for $n \geq 1$, we use consistency property to construct Π^{n+1} by connecting individual $n + 1$ to Π^n at

Received by the editors January 4th, 2024; accepted February 28th, 2024.

2020 *Mathematics Subject Classification.* primary 60J90, 60B10, 60B12, secondary 37A30, 60J10.

Key words and phrases. Kingman's coalescent, recursive construction, sequential construction, lineage length, provisional external branch length, convergence of non-Markov processes.

a random time, see Figure 1.1. More precisely, given Π^n: at any time t, if individual $n+1$ has not been connected to Π^n, then the rate for it to be connected is equal to $|\Pi^n(t)|$; if the connection takes place at time t, then the individual will coalesce with a block chosen uniformly from $\Pi^n(t)$. We denote the connection time of individual $n+1$ by L_{n+1}. We shall also call it the *lineage length* of individual n+1. The construction just explained is the so-called *recursive (or sequential) construction* of Kingman's coalescent; see Dhersin et al. (2013, Section 5) for an introduction of recursive construction of Λ-coalescents for which Kingman's coalescent is a special case, see also Crane (2016, Section 3.4).

We are interested in the asymptotic behaviour of the process of lineage lengths (or connection times) of individuals in this construction. This is partly motivated by a recent work (Favaro et al., 2019) which studied the inference of quantities related to the genealogy of additional individuals given a known sample. The recursive construction provides a natural way of enlarging sample sizes, and could be a useful angle to investigate the genealogical relationship between known and new additional individuals. The idea of adding up small parts to construct the whole process can also be found in the measure division construction of Λ-coalescents (Yuan, 2014), see also Berestycki et al. (2008, Section 3).

Based on the definition of recursive construction, for any $n \geq 1$, we have

$$\mathbb{P}(L_{n+1} \geq t \,|\, \Pi^n) = \exp\left(-\int_0^t |\Pi^n(s)| ds\right), \quad \forall t \geq 0. \tag{1.1}$$

Since $\Pi^1(t) = \{1\}$ for any $t \geq 0$, we set $L_1 = \infty$ by convention. Note that in this construction, L_{n+1} is the external branch length of individual $n+1$ in Π^{n+1}. However as more individuals are added, the external branch length of individual $n+1$ will be shorter and shorter, see Figure 1.1. From this point of view, we call L_n the *provisional external branch length* (although for brevity we will still use *lineage length* later) of individual n, for $n \geq 1$. Here the case $n = 1$ is included for completeness.

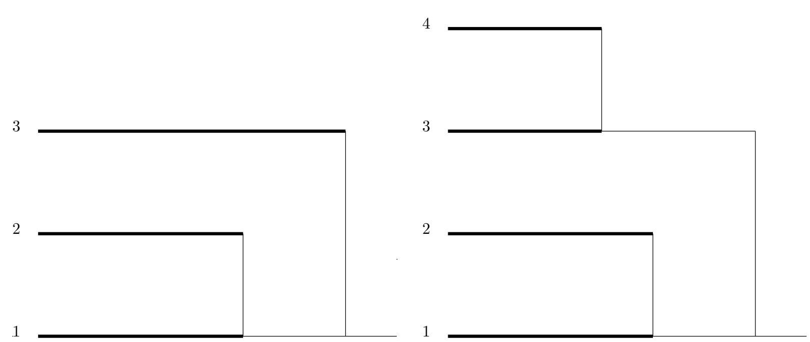

FIGURE 1.1. On the left, the recursive construction is up to individual 3, and on the right is up to individual 4. The bold segments are the external branches. On the left, individual 3 is just added and thus L_3 is the external branch length of individual 3. On the right, since individual 4 coalesced with individual 3, the external branch length of individual 3 becomes L_4 which is smaller than L_3.

Definition 1.1. In the recursive construction of Kingman's coalescent, we call the process

$$(L_n) := (L_n)_{n\geq 1}$$

provisional external branch length sequence (PEBLS) of Kingman's coalescent.

Given (L_n), we can construct Π^∞, following the description of recursive construction:

- for any $n \geq 2$, choose uniformly an element from $\{i : L_i \geq L_n, 1 \leq i \leq n-1\}$, say j;
- then merge individual n with the cluster containing j at time L_n.

The resulting process has the same law as Π^∞.

The recursive construction allows to build Π^∞ by sample size expansion. We can view integer n as individual n and also as time n. This is the main difference with the usual Kingman's coalescent which fixes the sample size (finite or infinite) first and evolves in (real) time. A direct consequence of the size-expansion point of view is that (L_n) is not a Markov chain, since to determine the law of L_n, we need to know not only L_{n-1}, but all L_i for $2 \le i \le n-1$, see (1.1). The main result of this paper is that, surprisingly, there is a Markov chain out of (L_n), see Theorem 2.1 in the next section.

2. Main results

Let M_j be the j-th largest length in $(L_n)_{n\ge 2}$. Note that $M_1 < \infty$ almost surely as Kingman's coalescent comes down from infinity (i.e. $|\Pi^\infty(t)| < \infty$ for all $t > 0$, almost surely). Let A_1 be the arrival time of $M_1 : L_{A_1} = M_1$. Let $R_1 = 1$. For any $i \ge 2$, define A_i, R_i by

$$A_i = \arg\max_j \{L_j : j > A_{i-1}\}, \quad M_{R_i} = L_{A_i}. \tag{2.1}$$

In other words,

- the largest length in $(L_n)_{n\ge 2}$ has index A_1;
- for any $i \ge 2$, the largest among $\{L_n : n > A_{i-1}\}$ has index A_i;
- moreover, the length with index A_i is the R_i-th largest among $(L_n)_{n\ge 2}$.

Thus, $(A) = (A_i)_{i\ge 1}$ records the arrival times of successive largest lengths in $(L_n)_{n\ge 2}$, and $(R) = (R_i)_{i\ge 1}$ records the rankings of these lengths in $(L_n)_{n\ge 2}$. By definition we have

$$1 < A_i < A_{i+1}, 1 \le R_i < R_{i+1}, \qquad \text{for any } i \ge 1. \tag{2.2}$$

It turns out that (R, A) is a Markov chain despite that (L_n) is non-Markov.

Theorem 2.1. *The process (R, A) is a Markov chain such that*

(1) $A_i - R_i \ge 1$, *for any* $i \ge 1$;
(2) $\mathbb{P}(A_1 = n) = \frac{2}{n(n+1)}$ *for any* $n \ge 2$, *and* $R_1 = 1$;
(3) *For any* $i \ge 1$ *we have* $R_i + 1 \le R_{i+1} \le A_i$ *and for any* $1 \le x \le A_i - R_i$,

$$\mathbb{P}(R_{i+1} \ge R_i + x \,|\, R_i, A_i) = \frac{\binom{2A_i-1}{A_i-R_i-x}}{\binom{2A_i-1}{A_i-R_i-1}}, \tag{2.3}$$

and for any $y \ge 1$,

$$\mathbb{P}(A_{i+1} \ge A_i + y \,|\, R_i, A_i, R_{i+1}) = \frac{A_i + R_{i+1}}{A_i + R_{i+1} + y - 1}. \tag{2.4}$$

To analyse the asymptotic behaviour of (R, A), we study the convergence of the processes below

$$\mathcal{W}^{(n)} := \left(\left(\frac{R^2_{n+1+i}}{A_{n+1+i}}, \frac{A_{n+i}}{A_{n+1+i}} \right) \right)_{i\ge 1}, \quad n \ge 0,$$

as $n \to \infty$. Note that the above processes are not Markov for any n. To state the convergence result, we need some more notations. Let Exp(1) be the exponential distribution with parameter 1 and Uni(0, 1) the uniform distribution on (0, 1). Let $(\eta_1, \eta_2, \cdots)$ be i.i.d. random variables with common law Uni(0, 1). Let $\xi_0 \ge 0$ be a random variable independent of $(\eta_1, \eta_2, \cdots)$. Inductively, define

$$\xi_i := (\xi_{i-1} + X_i)\eta_i, \quad i \ge 1 \tag{2.5}$$

where $X_i \sim \text{Exp}(1)$ and is independent of $\{\xi_0, \xi_1, \cdots, \xi_{i-1}, \eta_1, \eta_2, \cdots\}$. Denote $\mathcal{W} := ((\xi_i, \eta_i))_{i\ge 1}$. We use $\Longrightarrow$ to denote the weak convergence in finite dimensional distributions. Then we have the following asymptotic result for $\mathcal{W}^{(n)}$.

Theorem 2.2. *If $\xi_0 \sim Exp(1)$, then $\mathcal{W}^{(n)} \overset{n\to\infty}{\Longrightarrow} \mathcal{W}$ which is a stationary Markov chain with*

- *$\xi_i \sim Exp(1)$ for any $i \geq 1$,*
- *and the density function of (ξ_1, η_1) given by*

$$\frac{\mathrm{d}^2}{\mathrm{d}s\mathrm{d}t}\mathbb{P}(\xi_1 \leq s, \eta_1 \leq t) = \frac{s}{t^2}e^{-\frac{s}{t}}, \quad s \geq 0, 0 < t < 1. \tag{2.6}$$

The remaining part of the paper will be devoted to proofs. We will prove Theorem 2.1 in Section 3 and Theorem 2.2 in Section 4.

3. Proof of Theorem 2.1

3.1. *Aldous's construction.* Aldous (1999, Section 4.2) introduced a construction of Kingman's coalescent, see also Berestycki (2009, Theorem 2.2). The idea can be dated back to the seminal paper of Kingman (Kingman, 1982). This construction will play a key role in the proof of Theorem 2.1.

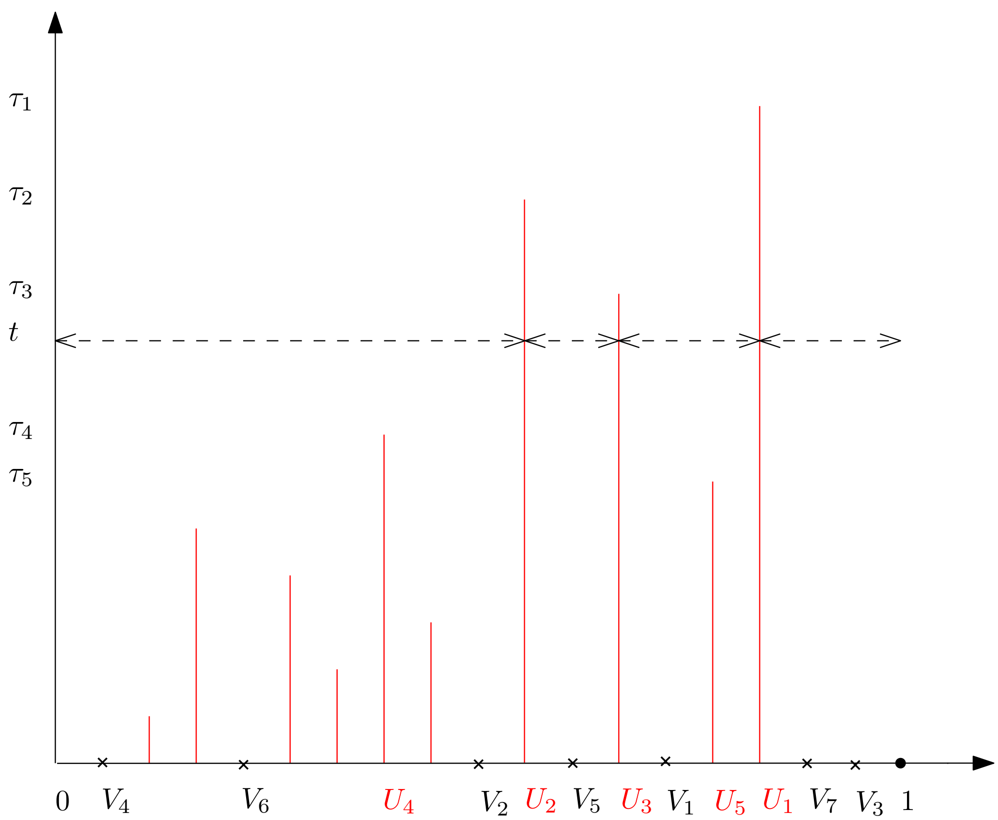

FIGURE 3.2. Aldous's construction of Kingman's coalescent. The vertical axis is the time axis. The stick i is at U_i with height τ_i. The V_i's (the crosses) are the locations of individuals. For the 7 individuals in the figure, the partition at time t is $\{\{2,4,6\},\{5\},\{1\},\{3,7\}\}$.

Let $(\zeta_i)_{i\geq 2}$ be independent random variables such that $\zeta_i \sim \text{Exp}\left(\binom{i}{2}\right)$. Define a random sequence $0 < \cdots < \tau_3 < \tau_2 < \tau_1 < \infty$ by $\tau_j = \sum_{k=j+1}^{\infty}\zeta_k$. Let $(U_1, U_2, \cdots, V_1, V_2, \cdots)$ be i.i.d. random variables with common law Uni(0,1) and independent of $(\tau_j)_{j\geq 1}$. Note that almost surely random variables in U_i's and V_i's are different from each other. Define a function $T : (0,1) \mapsto [0,\infty)$ such that $T(U_j) = \tau_j$ for any $j \geq 1$, and $T(u) = 0$ if $u \notin \{U_1, U_2, \cdots\}$. We call the vertical line from

(U_i, τ_i) down to $(U_i, 0)$ the stick i, see Figure 3.2. Then for any $t \geq 0$, we define a partition of $\mathbb{Z}_+$ such that i, j are in the same block if and only if

$$t \geq \sup_{\text{any } u \in (0,1) \text{ between } V_i, V_j} T(u).$$

The resulting process takes values in $\mathcal{P}(\infty)$ for any $t \geq 0$, and has the same law as Π^∞. In this construction, we call V_i the location of individual i, and U_i the location of stick i. Note that this construction applies to finitely many individuals as a natural restriction, see an example of 7 individuals in Figure 3.2.

Recall the sequence (M_j) introduced at the beginning of Section 2. Using Aldous's construction, it is clear that $M_j = \tau_j$ for any $j \geq 1$. We shall from now on only use the notation (τ_j) as further discussions are based on Aldous's construction.

3.2. *Identify (L_n) from Aldous's construction.* Aldous's construction gives a realisation of Π^∞, and thus we can use it to identify (L_n) and (R, A) in the recursive construction. We first deal with (L_n). We start with this question: find k such that $L_k = \tau_1$. Recall $U_1, U_2, \cdots, V_1, V_2, \cdots$ in Aldous's construction.

Lemma 3.1. *Let k be the smallest integer such that U_1 is between V_k and V_1. Then $L_k = \tau_1$.*

Proof: The stick 1 at U_1 splits $(0, 1)$ into two subintervals. WLOG, assume that $V_1, V_2, \cdots, V_{k-1}$ are on $(0, U_1)$ and V_k is on $(U_1, 1)$. By Aldous's construction, since the stick 1 at U_1 is of length τ_1 and is the largest among all sticks, all these k individuals will merge into one cluster at time τ_1, and the $k-1$ individuals on $(0, U_1)$ merge into one cluster at a time strictly smaller than τ_1. Then the only possibility is that individual k is connected to the coalescent process of the first $k-1$ individuals at time τ_1. Then the lemma is proved. □

Next we determine L_n for any $n \geq 2$ (recall $L_1 = \infty$).

Corollary 3.2. *Let $n \geq 2$. Let k be the smallest integer such that V_n is single without any others from $\{V_1, V_2, \cdots, V_n\}$ that is between U_k and some U_j for $1 \leq j < k$ or between U_k and 0 or between U_k and 1. Then $L_n = \tau_k$.*

Proof: The first $k-1$ sticks divide $(0, 1)$ into k subintervals. Individual n is located with some other individuals on one of the subintervals, say (a, b), where a, b are distinct elements in $\{U_1, U_2, \cdots, U_{k-1}, 0, 1\}$. The arrival of stick k at U_k will separate individual n from others on (a, b). This implies $L_n = \tau_k$, following a similar reasoning as in Lemma 3.1. Then the proof is finished. □

We present three more corollaries which will be used for the identification of (R, A). The first one finds the value of n such that $L_n = \tau_k$ for $k \geq 1$, generalising Lemma 3.1.

Corollary 3.3. *Let $k \geq 1$. Let X be the closest element to U_k from the left in $\{U_1, U_2, \cdots, U_{k-1}, 0, 1\}$ and Y from the right. Let n be the smallest integer such that the interval (X, U_k) contains at least one element from $\{V_1, V_2, \cdots, V_n\}$ and the same for (U_k, Y). Then $L_n = \tau_k$.*

Proof: For $k = 1$, it is a restatement of Lemma 3.1. For $k \geq 2$, we apply Corollary 3.2 to obtain that for such a unique n we have $L_n = \tau_k$. Then the proof is finished. □

The second one presents a special scenario where an upper bound for n in the above corollary can be given.

Corollary 3.4. *Let $s > 1$ and consider $V_1, V_2, \cdots, V_s$. Assume that for $k \geq 1$, U_k is neighbour to some V_i, V_j for $1 \leq i \neq j \leq s$ (i.e. U_k is between V_i and V_j; there exists no other U_a or V_b between V_i, V_j for $1 \leq a \leq k-1, 1 \leq b \leq s$). Then the individual n with $L_n = \tau_k$ must have $n \leq s$.*

Proof: WLOG, assume $V_i < U_k < V_j$. Let X be the closest element to U_k from the left in $\{U_1, U_2, \cdots, U_{k-1}, 0, 1\}$ and Y from the right. Then we have

$$X < V_i < U_k < V_j < Y.$$

Then by Corollary 3.3, the individual n such that $L_n = \tau_k$ must have $n \le i \vee j \le s$. Then the proof is finished. □

The last one finds a special scenario where a lower bound for n in Corollary 3.3 can be given.

Corollary 3.5. *Let $m \ge 1, k \ge 1$. If there is no element from $\{V_1, V_2, \cdots, V_m\}$ that is located between U_k and some U_j for $1 \le j < k$, or between U_k and 0, or between U_k and 1, then the individual n with $L_n = \tau_k$ must have $n > m$.*

Proof: To find n such that $L_n = \tau_k$, we need to consider more $V's$ so that the condition in Corollary 3.3 is satisfied. Therefore $n > m$. □

3.3. *Identify (R, A) from Aldous's construction.* Now we provide an algorithm to identify (R, A) based on Corollary 3.3.

Corollary 3.6. *We have $R_1 = 1$ by definition and the value of A_1 is given by Lemma 3.1 or Corollary 3.3: A_1 is the smallest integer such that U_1 is between V_{A_1} and V_1. We have $L_{A_1} = \tau_1$.*

In general, given (R_i, A_i) for some $i \ge 1$, we perform the following loop to obtain (R_{i+1}, A_{i+1}).

- *for $j \ge R_i + 1$*
- *find n using Corollary 3.3 such that $L_n = \tau_j$. If $n < A_i$, then continue the loop with $j = j+1$; otherwise (i.e. $n > A_i$) let $A_{i+1} = n$ and $R_{i+1} = j$ (hence $L_{A_{i+1}} = \tau_{R_{i+1}}$), and get out of the loop.*

Proof: We only need to check how the algorithm produces (R_{i+1}, A_{i+1}) given (R_i, A_i). We count $R_i + 1, R_i + 2, R_i + 3, \cdots$ until the first integer k such that the lineage length of rank k belongs to an individual, say m, which is larger than A_i. Then we have found $R_{i+1} = k, A_{i+1} = m$, based on the definition of (R, A). This is a restatement of the algorithm in the corollary and thus the proof is finished. □

It is clear that Corollary 3.6 is like a shell with the core being Corollary 3.3. We will provide another way of identifying (R, A) so that in Corollary 3.6 we do not need to find the exact value of n to know that $n < A_i$ (thanks to Corollary 3.4) and there is a natural way of determining n if $n > A_i$ (thanks to Corollary 3.5). The proof of Theorem 2.1 will reply on this new approach that we introduce in the next section.

3.4. *A simpler way of identifying (R, A) from Aldous's construction.* Note that Aldous's construction has two systems of notation: U_i's and V_i's, and also sticks and individuals. Stick i is at U_i and individual i is at V_i. Once U_i's and V_i's are given, the sticks and individuals are automatically planted at their locations. To facilitate the description of the simpler way of identification, we will plant the sticks and individuals successively in the order of $U_1, U_2, \cdots$ and $V_1, V_2, \cdots$ respectively. The implementation of the identification can be decomposed into two steps.

Step 1: plant the stick 1 at U_1. Plant individuals $1, 2, 3, \cdots$ successively on $(0, 1)$ at locations $V_1, V_2, V_3, \cdots$ until the first integer n ($n \ge 2$) such that U_1 is between V_n and V_1. We know from Lemma 3.1 that $A_1 = n$ and $R_1 = 1$.

Step 2: we proceed by induction to show the transition from (R_i, A_i) to (R_{i+1}, A_{i+1}) for any $i \ge 1$. Assume the following is true:

- We know R_i already, and after planting individual, say h, we obtain $A_i = h$;
- $U_1, U_2, \cdots, U_{R_i}$ and $V_1, V_2, \cdots, V_{A_i}$ are planted locations;

- Moreover, every U_l for $1 \leq l \leq R_i$ is neighbour to some V_k and V_j for $1 \leq k \neq j \leq A_i$ (i.e. U_l is between V_k, V_j; there exists no other U_s or V_t between V_k, V_j for $1 \leq s \leq R_i, 1 \leq t \leq A_i$).

Note that the notion of neighbour here is based on the planted locations. The above assumptions hold for $i = 1$, see Step 1. To obtain (R_{i+1}, A_{i+1}), we do the following.

- Finding R_{i+1}. Plant remaining sticks successively until the first stick n at U_n, which is not neighbour to some V_k and V_j for $1 \leq k \neq j \leq A_i$. Then we have found $R_{i+1} = n$. Note that in this case, there are three possibilities (and only one of them can happen):
  (1) U_n is neighbour to some U_l and V_j for $1 \leq l \leq n-1, 1 \leq j \leq A_i$;
  (2) U_n is neighbour to 0 and some V_j for $1 \leq j \leq A_i$;
  (3) U_n is neighbour to 1 and some V_j for $1 \leq j \leq A_i$.
- Finding A_{i+1}. Next we plant the remaining individuals successively until the first individual m such that V_m is neighbour to U_n and U_l or U_n and 0 or U_n and 1 depending on which one of the three possibilities happened in the step above finding R_{i+1}. Then we have found $A_{i+1} = m$. Moreover, the assumptions made at the beginning of Step 2 hold for (R_{i+1}, A_{i+1}).

Proposition 3.7. *The above procedure identifies* (R, A).

Proof: Starting from (R_i, A_i), before finding R_{i+1}, we are in the scenario described in Corollary 3.4. That is, for any $R_i + 1 \leq n < R_{i+1}$, we have $L_s = \tau_n$ for some $s \leq A_i$. The searching of R_{i+1} stops at n when we are in the scenario described in Corollary 3.5 as more individuals need to be planted successively until finding the individual (which is A_{i+1}) that has the lineage length τ_n (we use Corollary 3.3 to see that this individual has length τ_n) and thus $R_{i+1} = n$. The proof is finished. □

3.5. *Proof of Theorem 2.1.*

Proof of Theorem 2.1 - (1): In the identification procedure described in Section 3.4 for identifying (R, A), at any step i, we have planted exactly A_i individuals and R_i sticks, in such a way that each stick has to be neighbour to two individuals. Therefore necessarily $A_i \geq R_i + 1$. □

Proof of Theorem 2.1 - (2): To determine the law of A_1, we first recall the following lemma which is well known, see for instance Devroye (1986).

Lemma 3.8. *Assume there are* $k (k \geq 1)$ *i.i.d. uniform random variables on* $(0,1)$. *Then* $(0,1)$ *is cut into* $k+1$ *subintervals whose lengths are exchangeable and the vector of lengths follows the uniform distribution on a standard* k*-simplex. If we plant another independent uniform random variable on* $(0,1)$*, then it will enter one of the* $k+1$ *subintervals with equal probability. Conditioned on entering any subinterval, the resulting* $k+2$ *subintervals are again exchangeable and the new vector follows the uniform distribution on a standard* $(k+1)$*-simplex.*

Now we recall that $(U_1, U_2, \cdots, V_1, V_2, \cdots)$ are i.i.d. uniform on $(0,1)$. Then using the above lemma and Lemma 3.1, we have for any $n \geq 2$,

$$\begin{aligned}
\mathbb{P}(A_1 = n) &= 2\mathbb{P}(\max_{1 \leq i \leq n-1} V_i < U_1, V_n > U_1) \\
&= 2\mathbb{P}(V_1 < U_1) \left(\prod_{j=1}^{n-2} \mathbb{P}(V_{j+1} < U_1 \mid \max_{1 \leq i \leq j} V_i < U_1) \right) \mathbb{P}(V_n > U_1 \mid \max_{1 \leq i \leq n-1} V_i < U_1) \\
&= 2 \times \frac{1}{2} \times \prod_{j=1}^{n-2} \frac{j+1}{j+2} \times \frac{1}{n+1} = \frac{2}{n(n+1)}.
\end{aligned}$$

Then statement (2) is proved. □

Proof of Theorem 2.1 - (3): We will show the probability mass functions for the two laws to deduce the tail probabilities. More precisely, we shall first prove the following:

- For any $i \geq 1$ we have $R_i + 1 \leq R_{i+1} \leq A_i$ and for any $1 \leq x \leq A_i - R_i$,

$$\begin{aligned}\mathbb{P}(R_{i+1} = R_i + x \,|\, R_i, A_i) &= \frac{2R_i + 2x}{A_i + R_i + x}\left(\prod_{k=1}^{x-1}\frac{A_i - R_i - k}{A_i + R_i + k}\right) \\ &= \frac{2R_i + 2x}{A_i + R_i + 1}\binom{2A_i}{A_i - R_i - x}\Big/\binom{2A_i}{A_i - R_i - 1},\end{aligned} \tag{3.1}$$

- and for any $y \geq 1$,

$$\begin{aligned}\mathbb{P}(A_{i+1} = A_i + y \,|\, R_i, A_i, R_{i+1}) &= \frac{1}{A_i + R_{i+1} + y}\prod_{k=1}^{y-1}\left(\frac{A_i + R_{i+1} + k - 1}{A_i + R_{i+1} + k}\right) \\ &= \frac{A_i + R_{i+1}}{(A_i + R_{i+1} + y - 1)(A_i + R_{i+1} + y)}.\end{aligned} \tag{3.2}$$

We will use the identification procedure in Section 3.4. Under the assumptions in Step 2, $(0, 1)$ is divided into $R_i + A_i + 1$ subintervals. Among them, there are three categories of subintervals:

(1) there are R_i pairs of subintervals such that each pair share the same U_k as a common end for some $1 \leq k \leq R_i$;
(2) there are two subintervals such that either of them has one end being 0 or 1 (cannot have both $0, 1$ as ends);
(3) the remaining $A_i - R_i - 1$ subintervals can only have ends from $\{V_j : 1 \leq j \leq A_i\}$.

Then following Step 2, we plant remaining sticks starting from $R_i + 1$ successively until the first stick $R_i + x$ that is not neighbour to some V_k and V_j for $1 \leq k \neq j \leq A_i$. Note that every stick j for $R_i + 1 \leq j < R_i + x$ enters a subinterval of category 3, and thus killing one subinterval of category 3 and adding a pair of subintervals of category 1. The stick $R_i + x$ will enter a subinterval of category 1 or 2. In other words,

$$\begin{aligned}\mathbb{P}(R_{i+1} = R_i + x \,|\, R_i, A_i) = \mathbb{P}(&\text{stick } j \text{ enters a subinterval of category 3 for } R_i + 1 \leq j < R_i + x, \\ &\text{and stick } R_i + x \text{ enters a subinterval of category 1 or 2} \,|\, R_i, A_i)\end{aligned}$$

Then clearly we have $R_i + 1 \leq R_{i+1} \leq A_i$ since the number of subintervals of category 3 is $A_i - R_i - 1$. Using Lemma 3.8 and conditional probability formula, the above display yields the first equality in (3.1). The second equality is a direct simplification.

The next step is to plant remaining individuals until stick $R_i + x$ is again neighbour to two planted individuals. We omit the proof of (3.2), which is very similar to that of (3.1).

Next we deduce the tail probabilities. We will only show (2.3) as (2.4) is straightforward. If $x = A_i - R_i$, then we obtain

$$\mathbb{P}(R_{i+1} \geq A_i \,|\, R_i, A_i) = \frac{1}{\binom{2A_i - 1}{A_i - R_i - 1}} = \mathbb{P}(R_{i+1} = A_i \,|\, R_i, A_i).$$

For $1 \le x < A_i - R_i$, we show that (2.3) implies (3.1):

$$\begin{aligned}
&\mathbb{P}(R_{i+1} \ge R_i + x \,|\, R_i, A_i) - \mathbb{P}(R_{i+1} \ge R_i + x + 1 \,|\, R_i, A_i)\\
=&\frac{\binom{2A_i-1}{A_i-R_i-x}}{\binom{2A_i-1}{A_i-R_i-1}} - \frac{\binom{2A_i-1}{A_i-R_i-x-1}}{\binom{2A_i-1}{A_i-R_i-1}}\\
=&\frac{A_i+R_i+x}{A_i+R_i+1}\frac{\binom{2A_i}{A_i-R_i-x}}{\binom{2A_i}{A_i-R_i-1}} - \frac{A_i-R_i-x}{A_i+R_i+1}\frac{\binom{2A_i}{A_i-R_i-x}}{\binom{2A_i}{A_i-R_i-1}}\\
=&\frac{2R_i+2x}{A_i+R_i+1}\frac{\binom{2A_i}{A_i-R_i-x}}{\binom{2A_i}{A_i-R_i-1}}
\end{aligned}$$

which is exactly the probability $\mathbb{P}(R_{i+1} = R_i + x \,|\, R_i, A_i)$ given by (3.1). Then we conclude that (2.3) holds true. □

Finally, all statements in Theorem 2.1 are proved, and the proof is complete.

4. Proof of Theorem 2.2

4.1. *Preliminaries.* In this section, we prove two lemmas for preparation. We use $\xrightarrow{w}$ to denote the weak convergence of probability measures; $\xrightarrow{d}$ for the convergence in distribution for random variables and $\overset{d}{=}$ for being equal in distribution. We use $\mathcal{L}(\cdot)$ to denote the law of a random object.

Lemma 4.1. *Let $n \ge 1$. Consider a random variable $W := W_n$ such that*

$$\mathbb{P}(W = k) = ck\binom{2n}{n-k}, \quad 0 \le k \le n,$$

where $c > 0$ is the normalising constant. Let $t > 0$. Then uniformly in $s \in [0, t]$, we have

$$\sqrt{n}\mathbb{P}(W = \lfloor s\sqrt{n} \rfloor) \longrightarrow 2se^{-s^2}, \text{ as } n \to \infty. \tag{4.1}$$

As a consequence,

$$\mathcal{L}\left(\frac{W^2}{n}\right) \xrightarrow[n\to\infty]{w} Exp(1).$$

Proof: It suffices to prove (4.1). We first find the asymptotic equivalent of c. Let $\alpha \sim B(2n, 1/2)$, a binomial random variable with parameters $2n$ and $1/2$. Note that

$$\frac{1}{c} = \sum_{k=0}^{n} k\binom{2n}{n-k} = \sum_{k=0}^{n} n\binom{2n}{n-k} - \sum_{k=0}^{n}(n-k)\binom{2n}{n-k} =: I_1 - I_2.$$

For the first term, we have

$$2^{-2n}I_1 = n\mathbb{P}(\alpha \le n) = \frac{n}{2} + \frac{n}{2}\mathbb{P}(\alpha = n) = \frac{n}{2} + \frac{n}{2}2^{-2n}\binom{2n}{n} = \frac{n}{2} + \frac{\sqrt{n}}{2\sqrt{\pi}}(1 + o(1)),$$

where the last equality is due to Stirling formula. For the second term, we have

$$\begin{aligned}
2^{-2n}I_2 &= \sum_{k=0}^{n} k\binom{2n}{k}2^{-2n} = 2n\sum_{k=1}^{n}\binom{2n-1}{k-1}2^{-2n} = 2n\sum_{k=0}^{n-1}\binom{2n-1}{k}2^{-2n}\\
&= 2n\frac{\sum_{k=0}^{2n-1}\binom{2n-1}{k}}{2}2^{-2n} = 2n \times \frac{2^{2n-1}}{2}2^{-2n} = \frac{n}{2}.
\end{aligned}$$

Then we obtain that

$$\frac{1}{c} = \frac{\sqrt{n}}{2\sqrt{\pi}}2^{2n}(1 + o(1)). \tag{4.2}$$

Therefore, uniformly for $s \in [0,t]$, as $n \to \infty$,

$$\mathbb{P}(W = \lfloor s\sqrt{n} \rfloor) = \frac{2\sqrt{\pi}}{\sqrt{n}} 2^{-2n} \lfloor s\sqrt{n} \rfloor \binom{2n}{n - \lfloor s\sqrt{n} \rfloor} (1+o(1)) = e^{-s^2} \frac{2s}{\sqrt{n}}(1+o(1)),$$

where the first $o(1)$ comes from (4.2) and the second equality follows from Stirling formula, and both $o(1)$'s converge to 0 uniformly in $s \in [0,t]$ as $n \to \infty$. Then the proof is finished. □

Lemma 4.2. *The process* $\left(\frac{R_i^2}{A_i}\right)_{i \geq 1}$ *is tight.*

Proof: Tightness means that for any $\epsilon > 0$, there exists a compact set $E \subset [0,\infty)$ such that for any $i \geq 1$, $\mathbb{P}\left(\frac{R_i^2}{A_i} \notin E\right) \leq \epsilon$. To prove this, it suffices to show that there exists $c > 0$ such that the following holds true

$$\mathbb{E}\left[\frac{R_i}{\sqrt{A_i}}\right] \leq c, \quad \forall i \geq 1. \tag{4.3}$$

Indeed, applying Markov inequality, we obtain that $\mathbb{P}\left(\frac{R_i^2}{A_i} > N\right) = \mathbb{P}\left(\frac{R_i}{\sqrt{A_i}} > \sqrt{N}\right) \leq \frac{c}{\sqrt{N}} \leq \epsilon$ if we take N large. Then the compact set can be set as $E = [0,N]$ and the tightness is obtained.

Denote $I := \frac{\sum_{x=1}^{A_i-R_i-1} \binom{2A_i-1}{A_i-R_i-x-1}}{\binom{2A_i-1}{A_i-R_i-1}}$. Note that using (2.3), we have

$$\begin{aligned}
&\mathbb{E}[A_i - R_{i+1} \,|\, R_i, A_i] \\
&= \sum_{x=1}^{A_i-R_i-1} (A_i - R_i - x) \left(\frac{\binom{2A_i-1}{A_i-R_i-x}}{\binom{2A_i-1}{A_i-R_i-1}} - \frac{\binom{2A_i-1}{A_i-R_i-x-1}}{\binom{2A_i-1}{A_i-R_i-1}} \right) \\
&= \sum_{x=1}^{A_i-R_i-1} (A_i - R_i - x) \frac{\binom{2A_i-1}{A_i-R_i-x}}{\binom{2A_i-1}{A_i-R_i-1}} - \sum_{x=1}^{A_i-R_i-1} (A_i - R_i - x - 1 + 1) \frac{\binom{2A_i-1}{A_i-R_i-x-1}}{\binom{2A_i-1}{A_i-R_i-1}} \\
&= \sum_{x=1}^{A_i-R_i-1} (2A_i - 1) \frac{\binom{2A_i-2}{A_i-R_i-x-1}}{\binom{2A_i-1}{A_i-R_i-1}} - \sum_{x=1}^{A_i-R_i-2} (2A_i - 1) \frac{\binom{2A_i-2}{A_i-R_i-x-2}}{\binom{2A_i-1}{A_i-R_i-1}} - I \\
&= (2A_i - 1) \sum_{x=1}^{A_i-R_i-2} \frac{\binom{2A_i-2}{A_i-R_i-x-1} - \binom{2A_i-2}{A_i-R_i-x-2}}{\binom{2A_i-1}{A_i-R_i-1}} + \frac{2A_i - 1}{\binom{2A_i-1}{A_i-R_i-1}} - I \\
&= (2A_i - 1) \frac{\binom{2A_i-2}{A_i-R_i-2}}{\binom{2A_i-1}{A_i-R_i-1}} - I \\
&= A_i - R_i - 1 - I.
\end{aligned}$$

Then

$$\mathbb{E}[R_{i+1} \,|\, R_i, A_i] = R_i + 1 + I. \tag{4.4}$$

Note that

$$\begin{aligned}
I &\leq \frac{\sum_{x=1}^{A_i-R_i} \binom{2A_i}{A_i-R_i-x}}{\binom{2A_i}{A_i-R_i}} && (4.5) \\
&\leq \frac{1}{2R_i+2} \frac{\sum_{x=1}^{A_i-R_i} 2(R_i+x)\binom{2A_i}{A_i-R_i-x}}{\binom{2A_i}{A_i-R_i}} = \frac{A_i+R_i+1}{2R_i+2} \frac{\binom{2A_i}{A_i-R_i-1}}{\binom{2A_i}{A_i-R_i}} = \frac{A_i-R_i}{2R_i+2} \\
&\leq \frac{A_i}{R_i}, && (4.6)
\end{aligned}$$

where the first equality is due to (3.1). Conditional on A_i and R_i, let $\alpha \sim B(2A_i, 1/2)$. If $R_i \le \sqrt{A_i}$, then using (4.5), we have

$$\begin{aligned} I &\le 2^{2A_i} \frac{\sum_{x=1}^{A_i-R_i} \binom{2A_i}{A_i-R_i-x} 2^{-2A_i}}{\binom{2A_i}{A_i-\lfloor\sqrt{A_i}\rfloor}} = 2^{2A_i} \frac{\mathbb{P}(0 \le \alpha < A_i - R_i \,|\, A_i, R_i)}{\binom{2A_i}{A_i-\lfloor\sqrt{A_i}\rfloor}} \\ &\le \frac{2^{2A_i}}{\binom{2A_i}{A_i-\lfloor\sqrt{A_i}\rfloor}} \le c_0\sqrt{A_i}, \quad \text{if } i \text{ (or equivalently } A_i \text{, see (2.2)) is large enough,} \end{aligned} \tag{4.7}$$

where $c_0 > 0$ does not depend on anything and the last inequality is by Stirling formula. In conclusion, for i large enough,

$$\mathbb{E}[R_{i+1} \,|\, R_i, A_i] \le \begin{cases} R_i + 1 + \frac{A_i}{R_i}, & \text{if } R_i > \sqrt{A_i}; \\ R_i + 1 + c_0\sqrt{A_i}, & \text{if } R_i \le \sqrt{A_i}. \end{cases} \tag{4.8}$$

Next we note that there exists $0 < c_1 < 1$ such that for $i \ge 1$

$$\begin{aligned} \mathbb{E}\left[\frac{1}{\sqrt{A_{i+1}}} \,|\, R_i, A_i, R_{i+1}\right] &\le \mathbb{P}(A_{i+1} \ge 2A_i \,|\, R_i, A_i, R_{i+1}) \frac{1}{\sqrt{2A_i}} + \mathbb{P}(A_{i+1} < 2A_i \,|\, R_i, A_i, R_{i+1}) \frac{1}{\sqrt{A_i}} \\ &\le \left(\frac{1}{2\sqrt{2}} + \frac{1}{2}\right) \frac{1}{\sqrt{A_i}} \\ &\le \frac{c_1}{\sqrt{A_i}}, \end{aligned} \tag{4.9}$$

where the first inequality uses $A_{i+1} > A_i$, see (2.2), and the second inequality uses $\mathbb{P}(A_{i+1} \ge 2A_i \,|\, R_i, A_i, R_{i+1}) \ge \frac{1}{2}$ which is a consequence of (2.4). Then following (4.4), (4.8) and (4.9), we obtain for i large enough,

$$\begin{aligned} \mathbb{E}\left[\frac{R_{i+1}}{\sqrt{A_{i+1}}} \,|\, R_i, A_i\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{R_{i+1}}{\sqrt{A_{i+1}}} \,|\, R_i, A_i, R_{i+1}\right] \,|\, R_i, A_i\right] \\ &\le c_1 \mathbb{E}\left[\frac{R_{i+1}}{\sqrt{A_i}} \,|\, R_i, A_i\right] \\ &\le \begin{cases} c_1 \frac{R_i}{\sqrt{A_i}} + c_1 \frac{1}{\sqrt{A_i}} + c_1 \frac{\sqrt{A_i}}{R_i}, & \text{if } R_i > \sqrt{A_i} \\ c_1 \frac{R_i}{\sqrt{A_i}} + c_1 \frac{1}{\sqrt{A_i}} + c_0 c_1, & \text{if } R_i \le \sqrt{A_i} \end{cases} \\ &\le \begin{cases} c_1 \frac{R_i}{\sqrt{A_i}} + 2c_1, & \text{if } R_i > \sqrt{A_i} \\ c_1 \frac{R_i}{\sqrt{A_i}} + c_1(1 + c_0), & \text{if } R_i \le \sqrt{A_i} \end{cases} \\ &\le c_1 \frac{R_i}{\sqrt{A_i}} + c_2 \end{aligned}$$

where $c_2 = \max\{2c_1, c_1(1 + c_0)\}$, and for the third inequality we used (2.2). Therefore, we obtain the following

$$\mathbb{E}\left[\frac{R_{i+1}}{\sqrt{A_{i+1}}}\right] \le c_1 \mathbb{E}\left[\frac{R_i}{\sqrt{A_i}}\right] + c_2, \quad \text{if } i \text{ is large enough.} \tag{4.10}$$

Since $0 < c_1 < 1$, the above display yields (4.3). Then the proof is finished. □

4.2. *Proof of Theorem 2.2*. We start with two propositions. Either of them proves a partial result of Theorem 2.2 whose proof is provided at the end of this section. The first proposition is about the ergodicity of the Markov chain $(\xi_i)_{i \ge 0}$, introduced in (2.5). Let P be the transition kernel of $(\xi_i)_{i \ge 0}$. Introduce the following weighted supremum norm for a function $f : [0, \infty) \mapsto \mathbb{R}$

$$\|f\| = \sup_{x \ge 0} \frac{|f(x)|}{x + 1}.$$

Let C_b be the set of bounded continuous functions from $[0,\infty)$ to $\mathbb{R}$. For $t \geq 0$, let $\mathcal{E}_t$ be the law as follows:

$$\mathcal{E}_t := \mathcal{L}(X \mid X \geq t) = \mathcal{L}(t+X), \qquad \text{for } X \sim \text{Exp}(1), t \geq 0, \tag{4.11}$$

In particular, $\mathcal{E}_0 = \text{Exp}(1)$.

Proposition 4.3 (Geometric ergodicity)**.** *The Markov chain $(\xi_i)_{i\geq 0}$ admits a unique invariant measure $\mu = Exp(1)$. Moreover, there exists $C > 0$ and $\rho \in (0,1)$ such that*

$$\|P^n f - \mu(f)\| \leq C\rho^n \|f - \mu(f)\|, \qquad \textit{for any } f \in C_b, \tag{4.12}$$

where $\mu(f) = \int_0^\infty f(x)\mu(\mathrm{d}x)$.

Proof: First of all, we verify that Exp(1) is an invariant measure. Thanks to the construction (2.5), it suffices to show that, if $\xi_0 \sim \text{Exp}(1)$, then $\xi_1 \sim \text{Exp}(1)$. Recall $\xi_1 = (\xi_0 + X_1)\eta_1$ where ξ_0, X_1, η_1 are independent and $X_1 \sim \text{Exp}(1), \eta_1 \sim \text{Uni(0,1)}$ (see (2.5)). If we assume $\xi_0 \sim \text{Exp}(1)$, then for any integer $k \geq 0$, we have

$$\begin{aligned}
\mathbb{E}[\xi_1^k] &= \mathbb{E}[(\xi_0 + X_1)^k]\mathbb{E}[\eta_1^k] \\
&= \frac{1}{k+1}\mathbb{E}[(\xi_0 + X_1)^k] \\
&= \frac{1}{k+1}\sum_{i=0}^{k}\binom{k}{i}\mathbb{E}[\xi_0^i]\mathbb{E}[X_1^{k-i}] = k! = \mathbb{E}[\xi_0^k].
\end{aligned}$$

Since the above display is true for any k, we conclude that $\xi_1 \overset{d}{=} \xi_0 \sim \text{Exp}(1)$. Here we used the method of moments, see Billingsley (1995, Theorem 30.1). Thus Exp(1) is indeed an invariant measure of $(\xi_i)_{i\geq 0}$.

Note that (4.12) implies that $\mu = \text{Exp}(1)$ is the unique invariant measure. For the rest, we will prove (4.12) by applying Theorem 3.6 in Hairer (2021) in our context. It suffices to verify the following condition which combines Assumption 3.1 and Remark 3.5 in Hairer (2021).

Condition: (1) Let V be the identity function: $V(x) = x, \forall x \geq 0$. Then $PV(x) = \frac{V(x)+1}{2}$ for any $x \geq 0$. (2) Let $R > 0$. Then for any f with $\sup_{x\geq 0}|f(x)| \leq 1$, we have

$$|Pf(x) - Pf(y)| \leq 2(1 - e^{-R}), \qquad \forall\, 0 \leq x \leq y \leq R.$$

The verification of the above condition is as follows. Note that

$$PV(x) = \mathbb{E}[V(\xi_1) \mid \xi_0 = x] = \mathbb{E}[\xi_0 + X_1]\mathbb{E}[\eta_1] = \frac{x+1}{2} = \frac{V(x)+1}{2}.$$

Then the first statement is proved. For the second statement, recall that ξ_0, X_1, η_1 are independent. Then we observe that

$$\begin{aligned}
Pf(x) &= \mathbb{E}[f(\xi_1 \mid \xi_0 = x)] \\
&= \mathbb{E}[f((\xi_0 + X_1)\eta_1) \mid \xi_0 = x] \\
&= \mathbb{E}[f((\xi_0 + X_1)\eta_1)\mathbf{1}_{X_1 \geq y-x} \mid \xi_0 = x] + \mathbb{E}[f((\xi_0 + X_1)\eta_1)\mathbf{1}_{X_1 < y-x} \mid \xi_0 = x] \\
&= \mathbb{P}(X_1 \geq y - x \mid \xi_0 = x)\mathbb{E}[f((\xi_0 + X_1)\eta_1) \mid X_1 \geq y - x, \xi_0 = x] \\
&\qquad + \mathbb{E}[f((\xi_0 + X_1)\eta_1)\mathbf{1}_{X_1 < y-x} \mid \xi_0 = x] \\
&= \mathbb{P}(X_1 \geq y - x)Pf(y) + \mathbb{E}[f((\xi_0 + X_1)\eta_1)\mathbf{1}_{X_1 < y-x} \mid \xi_0 = x].
\end{aligned}$$

Here for the last equality we used that $\mathcal{L}(X_1 \mid X_1 \geq y - x) = \mathcal{L}(X_1 + y - x)$ since $X_1 \sim \text{Exp}(1)$, see (4.11). Then since $|f(\cdot)| \leq 1$ and $0 \leq x \leq y \leq R$, we obtain

$$|Pf(x) - Pf(y)| \leq 2\mathbb{P}(X_1 < y - x) \leq 2(1 - e^{-R}).$$

Thus the condition holds true and the proof is finished. □

The second proposition is to prove the one dimensional convergence of $\left(\frac{R_i^2}{A_i}\right)$. For preparation, the following lemma is needed.

Lemma 4.4. *For any $c \geq 0$, we have $\mathcal{L}\left(\frac{A_i}{A_{i+1}} \,|\, R_i, R_{i+1} = \lfloor cA_i \rfloor\right) = \mathcal{L}\left(\frac{A_i}{A_{i+1}} \,|\, R_{i+1} = \lfloor cA_i \rfloor\right)$. Moreover, for any $C > 0, f \in C_b$,*

$$\sup_{0\leq c\leq C}\left|\mathbb{E}\left[f\left(\frac{A_i}{A_{i+1}}\right) \,|\, R_{i+1} = \lfloor\sqrt{cA_i}\rfloor\right] - \mathbb{E}[f(\eta)]\right| \xrightarrow{i\to\infty} 0,$$

where $\eta \sim Uni(0,1)$.

Proof: Note that here we allow $c = 0$ which means $R_i = 0$. For Kingman's coalescent, R_i can never be 0, but the transition probabilities (3.1) and (3.2) do allow the more general case with $R_i = 0$.

The lemma is a direct consequence of (2.4) or (3.2) and the fact that $A_i \to \infty$ as $i \to \infty$. We omit details of the proof. □

Proposition 4.5. *The law of $\frac{R_i^2}{A_i}$ converges weakly to Exp(1) as $i \to \infty$.*

Proof: We claim that it suffices to prove that for any $t \geq 0$ and any $f \in C_b$,

$$\sup_{0\leq s\leq t}\left|\mathbb{E}\left[f\left(\frac{R_{i+1}^2}{A_{i+1}}\right) \,|\, R_i = \lfloor\sqrt{sA_i}\rfloor\right] - Pf(s)\right| \xrightarrow{i\to\infty} 0. \tag{4.13}$$

Indeed, if (4.13) is true, then we have that for any $\epsilon > 0$ there exists $T > 0$ such that

$$\inf_{0\leq s\leq t}\mathbb{P}\left(0 \leq \frac{R_{i+1}^2}{A_{i+1}} \leq T \,|\, R_i = \lfloor\sqrt{sA_i}\rfloor\right) \geq 1-\epsilon, \quad i \text{ large enough.} \tag{4.14}$$

Then as $i \to \infty$, uniformly for $s \in [0,t]$, we have

$$\begin{aligned}\mathbb{E}\left[f\left(\frac{R_{i+2}^2}{A_{i+2}}\right) \,|\, R_i = \lfloor\sqrt{sA_i}\rfloor\right] &= \mathbb{E}\left[\mathbb{E}\left[f\left(\frac{R_{i+2}^2}{A_{i+2}}\right) \,|\, \frac{R_{i+1}^2}{A_{i+1}}\right] \,|\, R_i = \lfloor\sqrt{sA_i}\rfloor\right]\\ &= \mathbb{E}\left[Pf\left(\frac{R_{i+1}^2}{A_{i+1}}\right) \,|\, R_i = \lfloor\sqrt{sA_i}\rfloor\right] + o(1)\\ &= P^2f(s) + o(1),\end{aligned}$$

where the first equivalence is due to (4.14) and (4.13) and $f \in C_b, Pf \in C_b$; the second equivalence is due to (4.13) and $Pf \in C_b$. Therefore we obtain

$$\sup_{0\leq s\leq t}\left|\mathbb{E}\left[f\left(\frac{R_{i+2}^2}{A_{i+2}}\right) \,|\, R_i = \lfloor\sqrt{sA_i}\rfloor\right] - P^2f(s)\right| \xrightarrow{i\to\infty} 0.$$

Iterating this procedure, it follows that for any $n \geq 1$,

$$\sup_{0\leq s\leq t}\left|\mathbb{E}\left[f\left(\frac{R_{i+n}^2}{A_{i+n}}\right) \,|\, R_i = \lfloor\sqrt{sA_i}\rfloor\right] - P^nf(s)\right| \xrightarrow{i\to\infty} 0,$$

which implies that $\mathcal{L}(\frac{R_{i+n}^2}{A_{i+n}} \,|\, R_i = \lfloor\sqrt{sA_i}\rfloor)$ converges weakly to $\mathcal{L}(\xi_n \,|\, \xi_0 = s)$ for any $s \geq 0$ as $i \to \infty$. By Proposition 4.3, $\mathcal{L}(\xi_n \,|\, \xi_0 = s)$ converges weakly to Exp(1) as $n \to \infty$. Finally we apply Lemma 4.2 to conclude that this proposition holds true. This reasoning also leads to $\left(\frac{R_i^2}{A_i}, \frac{R_{i+1}^2}{A_{i+1}} \cdots, \frac{R_{i+k}^2}{A_{i+k}}\right) \xrightarrow[i\to\infty]{d} (\xi_0, \xi_1, \cdots, \xi_k)$ for any positive integer k if $\xi_0 \sim$ Exp(1). But we are content with the one dimensional convergence for this proposition and the multidimensional convergence will be proved in the proof of Theorem 2.2 more straightforwardly. Another way to prove this proposition using (4.13) is to apply Karr (1975, Theorem (1)). The only problem is that $\left(\frac{R_i^2}{A_i}\right)$ is not a Markov chain. It suffices to enlarge it into $\left(\frac{R_i^2}{A_i}, \frac{1}{A_i}\right)$. We omit the detailed steps.

To prove (4.13), we first recall the random variable $W = W_n$ in Lemma 4.1. For $0 \le k \le n$, (3.1) yields

$$\mathcal{L}(W \,|\, W > k) = \mathcal{L}(R_{i+1} \,|\, R_i = k, A_i = n).$$

Then using (4.1), we obtain

$$\sup_{0\le s\le t}\left|\mathbb{E}\left[f\left(\frac{R_{i+1}^2}{A_i}\right) \,|\, R_i = \lfloor\sqrt{sA_i}\rfloor\right] - \mathbb{E}[f(Z_s)]\right| \xrightarrow{i\to\infty} 0, \tag{4.15}$$

where $Z_s \sim \mathcal{E}_s$ (see (4.11)). As a consequence, similar to (4.14), for any $\epsilon > 0$, there exists $C > 0$ such that

$$\inf_{0\le s\le t}\mathbb{P}(0 \le R_{i+1}^2 \le CA_i \,|\, R_i = \lfloor\sqrt{sA_i}\rfloor) \ge 1-\epsilon, \quad i \text{ large enough.} \tag{4.16}$$

Then using Lemma 4.4, we have

$$\sup_{0\le s\le t, s\le c\le C}\left|\mathbb{E}\left[f\left(\frac{A_i}{A_{i+1}}\right) \,|\, R_i = \lfloor\sqrt{sA_i}\rfloor, R_{i+1} = \lfloor\sqrt{cA_i}\rfloor\right] - \mathbb{E}[f(\eta)]\right| \xrightarrow{i\to\infty} 0. \tag{4.17}$$

where $\eta \sim \text{Uni}(0,1)$. Note that $Pf(s) = \mathbb{E}[f(Z_s\eta)]$ if we assume η is independent of Z_s. Moreover, $\frac{R_{i+1}^2}{A_{i+1}} = \frac{R_{i+1}^2}{A_i}\frac{A_i}{A_{i+1}}$. Then using the above three displays, we conclude that (4.13) holds true and thus the proof for the proposition is finished. □

Proof of Theorem 2.2: By Proposition 4.3, $(\xi_i)_{i\ge 0}$ is a stationary Markov chain if $\xi_0 \sim \text{Exp}(1)$. Using the definition (2.5), we conclude that $\mathcal{W} = (\xi_i, \eta_i)_{i\ge 1}$ is a stationary Markov chain. Next we show (2.6). Using (2.5), we have

$$\mathbb{P}(\xi_1 \le s, \eta_1 \le t) = \mathbb{P}((\xi_0 + X_1)\eta_1 \le s, \eta_1 \le t) = \int_0^t \gamma\left(2, \frac{s}{u}\right) \mathrm{d}u,$$

where $\gamma(\cdot,\cdot)$ is the lower incomplete gamma function. Here we used that ξ_0, X_1, η_1 are independent and $\xi_0 \overset{d}{=} X_1 \sim \text{Exp}(1)$ and $\eta_1 \sim \text{Uni}(0,1)$, and such that $\xi_0 + X_1$ follows the Gamma distribution with shape parameter 2 and scale parameter 1. Taking the partial derivative with respect to s and t yields (2.6). It remains to prove $\mathcal{W}^{(n)} \Longrightarrow \mathcal{W}$ as $n \to \infty$.

Let $\xi_0 \sim \text{Exp}(1)$. Using Proposition 4.5 and (4.15), we have

$$\left(\frac{R_n^2}{A_n}, \frac{R_{n+1}^2}{A_n}\right) \xrightarrow[n\to\infty]{d} (\xi_0, \xi_0 + X_1). \tag{4.18}$$

Together with Lemma 4.4 and the fact that $\left(\frac{R_i^2}{A_i}\right)_{i\ge 1}$ is tight, we obtain further that

$$\left(\frac{R_n^2}{A_n}, \frac{R_{n+1}^2}{A_n}, \frac{A_n}{A_{n+1}}, \frac{A_{n+1}}{A_{n+2}}, \cdots\right) \overset{n\to\infty}{\Longrightarrow} (\xi_0, \xi_0 + X_1, \eta_1, \eta_2, \cdots). \tag{4.19}$$

Note that

$$\frac{R_{i+1}^2}{A_{i+1}} = \frac{R_{i+1}^2}{A_i}\frac{A_i}{A_{i+1}}, \quad \forall i \ge 1.$$

The above two displays entail

$$\left(\frac{R_n^2}{A_n}, \frac{R_{n+1}^2}{A_n}, \frac{R_{n+1}^2}{A_{n+1}}, \frac{A_n}{A_{n+1}}, \frac{A_{n+1}}{A_{n+2}}, \cdots\right) \overset{n\to\infty}{\Longrightarrow} (\xi_0, \xi_0 + X_1, \xi_1, \eta_1, \eta_2, \cdots).$$

Since the asymptotic behaviour of $\frac{R_{i+1}^2}{A_i}$ only depends on that of $\frac{R_i^2}{A_i}$ (see (4.18) and (4.19)), we have

$$\left(\frac{R_n^2}{A_n}, \frac{R_{n+1}^2}{A_n}, \frac{R_{n+1}^2}{A_{n+1}}, \frac{R_{n+2}^2}{A_{n+1}}, \frac{A_n}{A_{n+1}}, \frac{A_{n+1}}{A_{n+2}}, \cdots\right) \overset{n\to\infty}{\Longrightarrow} (\xi_0, \xi_0 + X_1, \xi_1, \xi_1 + X_2, \eta_1, \eta_2, \cdots).$$

Iterating this procedure, we obtain that for any $k \geq 0$,

$$\left(\frac{R_n^2}{A_n}, \frac{R_{n+1}^2}{A_n}, \cdots, \frac{R_{n+k}^2}{A_{n+k}}, \frac{R_{n+k+1}^2}{A_{n+k}}, \frac{A_n}{A_{n+1}}, \frac{A_{n+1}}{A_{n+2}}, \cdots\right)$$
$$\overset{n\to\infty}{\Longrightarrow} (\xi_0, \xi_0 + X_1, \cdots, \xi_k, \xi_k + X_{k+1}, \eta_1, \eta_2, \cdots).$$

which implies $\mathcal{W}^{(n)} \overset{n\to\infty}{\Longrightarrow} \mathcal{W}$. Then the proof of Theorem 2.2 is finished. □

Acknowledgements

The author thanks Matthias Birkner for suggesting the name *provisional external branch length sequence* for (L_n), and Clément Foucart and Huili Liu for helpful comments on a draft version. The author thanks an anonymous referee for carefully reading the manuscript and for pointing out several mistakes and giving insightful and constructive comments.

References

Aldous, D. J. Deterministic and stochastic models for coalescence (aggregation and coagulation): a review of the mean-field theory for probabilists. *Bernoulli*, **5** (1), 3–48 (1999). MR1673235.

Berestycki, J., Berestycki, N., and Schweinsberg, J. Small-time behavior of beta coalescents. *Ann. Inst. Henri Poincaré Probab. Stat.*, **44** (2), 214–238 (2008). MR2446321.

Berestycki, N. *Recent progress in coalescent theory*, volume 16 of *Ensaios Matemáticos [Mathematical Surveys]*. Sociedade Brasileira de Matemática, Rio de Janeiro (2009). ISBN 978-85-85818-40-1. MR2574323.

Billingsley, P. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edition (1995). ISBN 0-471-00710-2. MR1324786.

Crane, H. The ubiquitous Ewens sampling formula. *Statist. Sci.*, **31** (1), 1–19 (2016). MR3458585.

Devroye, L. *Nonuniform random variate generation*. Springer-Verlag, New York (1986). ISBN 0-387-96305-7. MR836973.

Dhersin, J.-S., Freund, F., Siri-Jégousse, A., and Yuan, L. On the length of an external branch in the beta-coalescent. *Stochastic Process. Appl.*, **123** (5), 1691–1715 (2013). MR3027896.

Favaro, S., Feng, S., and Jenkins, P. A. Bayesian nonparametric analysis of Kingman's coalescent. *Ann. Inst. Henri Poincaré Probab. Stat.*, **55** (2), 1087–1115 (2019). MR3949965.

Hairer, M. *Convergence of Markov processes.* Lecture notes, version of 2021-05-26 (2021). Available at `https://www.hairer.org/notes/Convergence.pdf`.

Karr, A. F. Weak convergence of a sequence of Markov chains. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, **33** (1), 41–48 (1975). MR394882.

Kingman, J. F. C. The coalescent. *Stochastic Process. Appl.*, **13** (3), 235–248 (1982). MR671034.

Yuan, L. On the measure division construction of Λ-coalescents. *Markov Process. Related Fields*, **20** (2), 229–264 (2014). MR3241525.