Notes 16 : Kingman's coalescent

MATH 833 - Fall 2012 *Lecturer: Sebastien Roch*

References: [Dur08, Chapter 1.2, 4.4].

1 Coalescent

In the previous lecture, we saw that the infinite population size limit of the first coalescence time of k samples is exponential with mean $\binom{k}{2}^{-1}$. This justifies the introduction of Kingman's n-coalescent which can be thought of as a process on partitions of $\{1, \ldots, n\}$—defined *backwards in time*.

- Let $\Pi_n^n = \{\{1\}, \{2\}, \ldots, \{n\}\}$, $T_n = 0$ and $k := n$.
- Repeat until $k = 1$:
  - Let t_k be exponential with mean $\binom{k}{2}^{-1}$ and set $T_{k-1} = T_k + t_k$.
  - Merge two uniformly random sets in Π_k^n to obtain Π_{k-1}^n and set $k := k - 1$.

Here, the T_k's are the successive coalescent times and the Π_k^n are the states at those times. See Figure 1.5 in [Dur08] for an illustration.

Note that the height T_1 has mean

$$\mathbb{E}[T_1] = \sum_{k=2}^{n} \mathbb{E}[t_k] = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n}\left(\frac{1}{k-1} - \frac{1}{k}\right) = 2\left(1 - \frac{1}{n}\right).$$

2 Effective population size

As we discussed in the previous lecture, the limiting procedure above is quite robust and often does not depend on the details of the model. We give a brief example—without proof—in the case of two-sex models.

Consider a population with N_m male diploids and N_f female diploids where N_m and N_f are constant in time and of the same order as $N = N_m + N_f$. Number the two gene copies in each individual 1 and 2. Imagine that, at each generation, the two copies of each individual decide uniformly at random which one inherits its state from a male. Then the corresponding chromosome picks a father uniformly at random from the male population and a uniformly chosen gene copy from that father. Similarly for the female chromosome.

Consider two gene copies from two different individuals in the population. Irrespective of the sex of the corresponding individuals, the probability of coalescence at the previous generation is

$$\left(\frac{1}{4}\frac{1}{N_m} + \frac{1}{4}\frac{1}{N_f}\right)\frac{1}{2} \equiv \frac{1}{2N_e},$$

independently of the other generations, where the $1/4$ comes from the choice of the same sex, the $1/N_m$ comes from the choice of the same parent and the $1/2$ comes from the choice of the same gene copy. The quantity N_e is called the effective population size. Rescaling time by $2N_e$ and taking a limit $N \to \infty$ leads to an exponential distribution for the coalescence time of 2 copies.

There is one caveat however. If the two chromosomes come from the same individual, then they *cannot* coalesce at the previous generation, as one of them chooses a male parent and the other a female parent. However, the probability of being in that "almost coalesced" state is of order $O(1/N)$ and one immediately leaves that state at the previous generation. Therefore, the amount spent on that problematic state is negligible and the limit is not affected.

One can show that the limit is Kingman's coalescent.

3 Shape of the coalescent

We begin our study of the coalescent with some simple calculations.

THM 16.1 (Distribution) *We have*

$$\mathbb{P}[\Pi_i^n = \Pi] = \frac{i!(n-i)!(i-1)!}{n!(n-1)!}\prod_{k=1}^{i}\lambda_k!,$$

where Π has $|\Pi| = i$ sets, their sizes being $\lambda_1, \ldots, \lambda_i$. We denote the first factor by $c_{n,i}$ and the second one by $w(\Pi)$.

Proof: We proceed by induction, the case $i = n$ being trivial. We write $\Pi' \prec \Pi$ if Π is obtained from Π' by merging two sets. To use induction, we condition on the state at time i,

$$\mathbb{P}[\Pi^n_{i-1} = \Pi \,|\, \Pi^n_i = \Pi'] = \frac{1}{\binom{i}{2}}.$$

Then

$$\mathbb{P}[\Pi^n_{i-1} = \Pi] = \frac{2}{i(i-1)} \sum_{\Pi' \prec \Pi} \mathbb{P}[\Pi^n_i = \Pi'].$$

If Π has sets of size $\lambda_1, \ldots, \lambda_{i-1}$, then $\Pi' \prec \Pi$ has sets of size

$$\lambda_1, \ldots, \lambda_{\ell-1}, \nu, \lambda_\ell - \nu, \ldots, \lambda_{i-1},$$

for some $1 \leq \ell \leq i-1$ and $1 \leq \nu < \lambda_\ell$. Hence,

$$\begin{aligned}
\mathbb{P}[\Pi^n_{i-1} = \Pi] &= \frac{2}{i(i-1)} \sum_{\ell=1}^{i-1} \sum_{\nu=1}^{\lambda_\ell - 1} c_{n,i} w_{\ell,\nu} \binom{\lambda_\ell}{\nu} \frac{1}{2} \\
&= w(\Pi) \frac{c_{n,i}}{i(i-1)} \sum_{\ell=1}^{i-1} \sum_{\nu=1}^{\lambda_\ell - 1} 1 \\
&= w(\Pi) \frac{c_{n,i}}{i(i-1)} (n - (i-1)),
\end{aligned}$$

where

$$w_{\ell,\nu} = \lambda_1! \cdots \lambda_{\ell-1}! \nu! (\lambda_\ell - \nu)! \cdots \lambda_{i-1}!.$$

Note that the factor $\binom{\lambda_\ell}{\nu}\frac{1}{2}$ gives the number of ways of choosing two subsets from the merged set where the two subsets are themselves unordered.

A simple calculation concludes the proof. ■

We immediately obtain the following result on the sizes of the sets which will be useful later.

THM 16.2 *Assume that Π^n_i has sets of size $\lambda_1, \ldots, \lambda_i$ where the sets are ordered such that the first one contains 1, the second contains the smallest remaining element, etc. Let π be a permutation on $\{1, \ldots, i\}$ and define $\mu_\ell = \lambda_{\pi(\ell)}$, for $\ell = 1, \ldots, i$. Then the vector $(\mu_1, \ldots, \mu_i)$ is distributed uniformly over vectors summing to n.*

Proof: By the previous theorem, permuting the sets in Π^n_i uniformly, each arrangement has probability

$$c_{n,i} \lambda_1! \cdots \lambda_i! \frac{1}{i!}.$$

Picking the elements in each set randomly, we obtain the probability of their sizes as

$$c_{n,i}\lambda_1!\cdots\lambda_i!\frac{1}{i!}\frac{n!}{\lambda_1!\cdots\lambda_i!}$$

which is independent of the sizes. ■

Further reading

The material in this section was taken from Sections 1.2 and 4.4 of the excellent monograph [Dur08]. For more details on the robustness of the coalescent, see [Wak08].

References

[Dur08] Richard Durrett. *Probability models for DNA sequence evolution*. Probability and its Applications (New York). Springer, New York, second edition, 2008.

[Wak08] J. Wakeley. *Coalescent Theory: An Introduction*. Roberts and Company Publishers, Greenwich Village, CO, 2008.