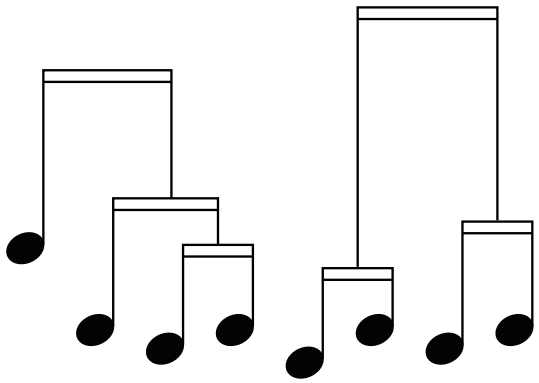
Yun S. Song

University of California, Berkeley

Lecture Notes on Computational and Mathematical Population Genetics



January 11, 2021

Contents

Part III Demography

Part IV Recombination

Preface

These lecture notes are from a graduate-level statistics course I taught at the University of California, Berkeley in 2008, 2011, and 2015. The first six chapters are in decent shape, but the later chapters are somewhat unpolished and have incomplete sections. I have taken a long hiatus from writing and it is unclear when I will be able to get back to it. Hence, I have decided to release this draft version in the hope that some students may find it useful.

There already exist several excellent books (e.g., Durrett (2008); Ewens (2004); Hein et al (2004); Tavaré (2004); Wakeley (2008)) on mathematical population genetics, and I myself have learned from studying them. The reader is strongly encouraged to look into these other resources to get a more complete view of the topic. You will notice that many important topics and references (apologies for not citing your work) are left out from this monograph, as it is NOT meant to be a comprehensive exposition of population genetics. Rather, the primary goal of these notes is to introduce mathematically-inclined students to the basic concepts underpinning the subject, so that they can get started on population genetics research.

The word "computational" is in the title of this work because efficient algorithms are indispensable in empirical population genetics and I try to highlight this aspect when I can. The special topics covered in these notes are mainly from my own research, simply because they reflect my own interest and expertise. Apologies for the personal bias.

Special thanks goes to Paul Jenkins for writing parts of Chapters 8 and 12 while he was a postdoc in my lab. I am also grateful to many students who scribed notes for my lectures and to my lab members for their contributions to research.

Berkeley, CA *Yun S. Song*

References

Durrett R (2008) Probability Models for DNA Sequence Evolution, 2nd edn. Springer, New York

Ewens WJ (2004) Mathematical Population Genetics: I. Theoretical Introduction. Springer Science+Business Media, Inc., New York

Hein J, Schierup M, Wiuf C (2004) Gene Genealogies, Variation and Evolution: A primer in coalescent theory. Oxford University Press, UK

Tavaré S (2004) Ancestral inference in population genetics. In: Tavaré S, Zeitouni O (eds) Lectures on Probability Theory and Statistics. Ecole d'Etes de Probabilite de Saint-Flour XXXI–2001, vol 1837, Springer Berlin/Heidelberg, pp 1–188

Wakeley J (2008) Coalescent Theory: An Introduction. Roberts & Company Publishers

Part I
Genealogical Trees

Chapter 1
Basic properties of the genealogy of a sample

At any given position in the genome, the genealogy of a sample of chromosomes is described by a tree. In this chapter, we will discuss some basic properties of the genealogy of a sample randomly drawn from a very large population. In particular, we will study the distribution of the number of ancestors to the sample as a function of time. We will also characterize other useful genealogical quantities such as the time to the most recent common ancestor and the total tree length.

1.1 A high-level description of the coalescent model

Before we go into the mathematical details, we here provide a high-level description of the underlying model. Consider a population of N chromosomes evolving forward in time. Illustrated in Figure 1.1a is an evolving population of size $N = 5$ with non-overlapping generations, where all individuals are equally likely to reproduce for each birth event. Take a random sample of size n from the present population and follow their ancestry backwards in time; $n = 4$ in Figure 1.1a. Now, as shown in Figure 1.1b, double the population size and rescale time by a factor of two so that one unit of time contains twice as many generations as in the previous case. Again take a sample of size n (the same n as before) and trace their ancestry backwards in time. Double the population size again (Figure 1.1c), rescale time by a factor of two, take a sample of size n, and trace their ancestry backwards in time. As $N \to \infty$ while time is rescaled as described, the distribution of the genealogy of a sample of size n converges to that of a stochastic process called Kingman's n-coalescent (Figure 1.1d), in which at most two lineages may merge at any given time. This kind of weak convergence result holds for a large class of random mating models, and necessary and sufficient conditions for convergence are known (discussed in Chapter 2). In general, how time should be rescaled as $N \to \infty$ depends on the underlying random mating model.

We now say a few words about the limiting model, the n-coalescent. Let $[n]$ denote the n-set $\{1, 2, \ldots, n\}$ and $\mathcal{P}_{[n]}$ the set of all partitions of $[n]$. As detailed in Chapter 2, the n-coalescent $\{C_n(t), t \geq 0\}$ (Kingman, 1982a,b,c) is a $\mathcal{P}_{[n]}$-valued continuous-time Markov process satisfying certain properties. As discussed above, this stochastic process is useful for evolutionary biology because it describes the law of the genealogy of a set of chromosomes randomly drawn from a population. More precisely, associated with each sample path in the n-coalescent is a unique n-leaved tree that is edge-weighted, rooted, binary, and ultrametric,

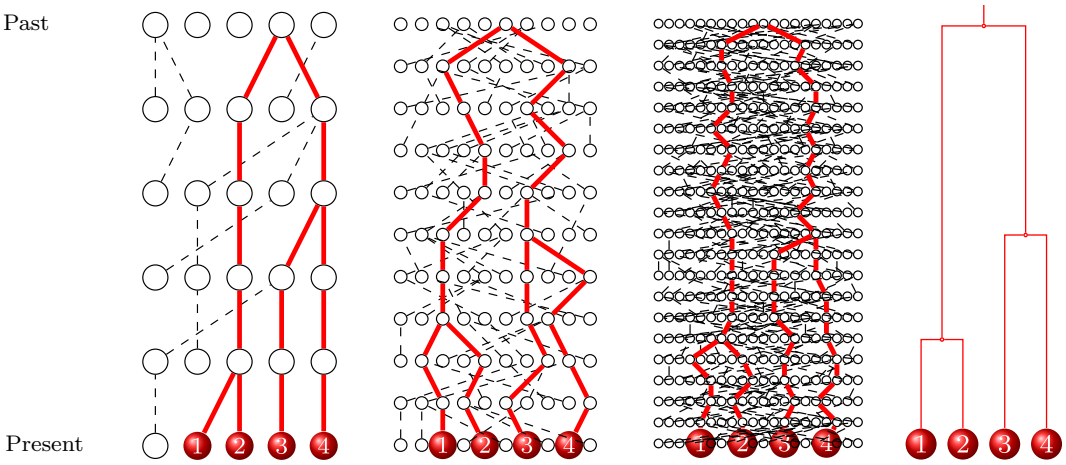


Fig. 1.1: A schematic depiction of convergence to the n-coalescent. Time runs vertically, with the present at the bottom and the past at the top. As $N \to \infty$ while time is rescaled suitably, the distribution of the genealogy of a sample of size n converges to that of a stochastic process called Kingman's n-coalescent. (a) $N = 5, n = 4$. (b) $N = 10, n = 4$ (c). $N = 20, n = 4$. (d) A tree with 4 leaves in the coalescent limit.

and the leaves are bijectively labeled by $[n]$. (Given two leaves a, b of a tree $\mathcal{T}$, let $d(a, b)$ denote the path length between a and b. Then, $\mathcal{T}$ is said to be an *ultrametric* tree if $d(a, b) \leq \max\{d(a, c), d(b, c)\}$ for all distinct leaves a, b, c.) A realization of the n-coalescent for $n = 5$ is illustrated in Figure 1.2. (Note that horizontal edges do not carry any meaning in this representation.) Over the next couple of chapters, we will study some key properties of such trees. Later in the course we will encounter more complicated graphical structures when we incorporate recombination into the coalescent framework.

In Chapter 2, we will derive an explicit expression for $\mathbb{P}(C_n(t) = \alpha)$, where $\alpha \in \mathcal{P}_{[n]}$. In this chapter, we focus on finding $\mathbb{P}(|C_n(t)| = j)$, where $|C_n(t)|$ denotes the number of blocks (non-empty subsets) in $C_n(t)$ and $j \in [n]$. This work will be prove useful when we derive $\mathbb{P}(C_n(t) = \alpha)$.

1.2 Discrete-time ancestral process

Many exchangeable random mating models converge to the n-coalescent. Here, we will consider a well-known random mating model and consider its limiting behavior. The so-called Wright-Fisher model has the following properties:

1. Discrete time (measured in generations) with population size N_g at generation g.
2. Non-overlapping generations. Every individual survives for exactly one generation.
3. All individuals are equally likely to reproduce for each birth event.

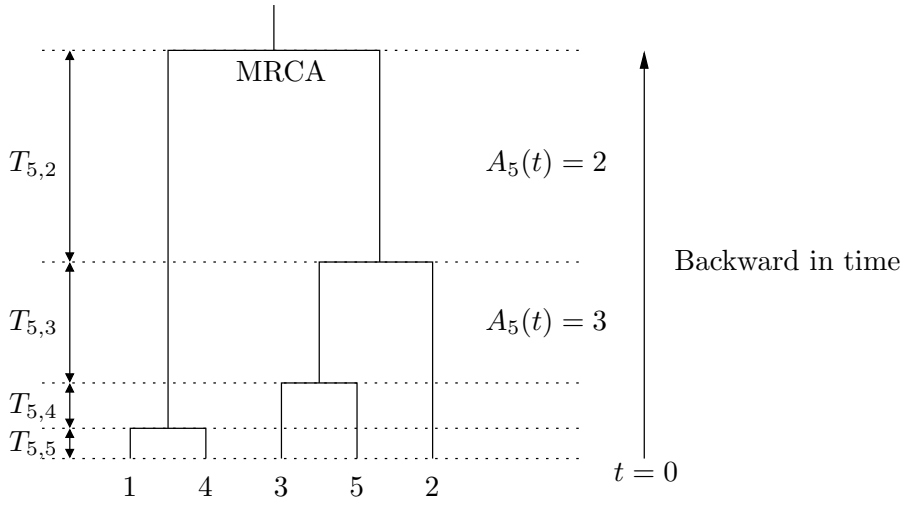


Fig. 1.2: A coalescent tree for a sample of size $n = 5$. The sample is taken at time 0, and t runs backwards in time. Random variable $T_{n,k}$ denotes the time spent while there are k lineages, while $A_n(t)$ denotes the number of ancestral lineages at time t. In the case of a constant population size, $T_{n,n}, \ldots, T_{n,2}$ are mutually independent and $T_{n,k}$ is exponentially distributed with rate $\binom{k}{2}$, for $k = 2, \ldots, n$.

Going backwards in time, each child chooses a parent uniformly at random from the population in the previous generation. Define $p_{kj}^{(g)}$ as

$$p_{kj}^{(g)} := \mathbb{P}(k \text{ particular labeled individuals in generation } g \text{ have } j \text{ distinct parents}). \quad (1.1)$$

We use $g \in \{0, 1, \ldots\}$ to denote the number of generations *back in time*, with $g = 0$ corresponding to the generation in which a sample is taken. Then, for $j = k$,

$$\begin{aligned} p_{22}^{(g)} &= \frac{N_{g+1} - 1}{N_{g+1}}, \\ p_{33}^{(g)} &= \frac{(N_{g+1} - 1)(N_{g+1} - 2)}{(N_{g+1})^2}, \\ p_{kk}^{(g)} &= \frac{(N_{g+1} - 1)(N_{g+1} - 2) \cdots (N_{g+1} - k + 1)}{(N_{g+1})^{k-1}}. \end{aligned}$$

Before we consider general values of j, we first define some notation which will be used time and again:

Definition 1.1 (The jth falling and rising factorials).

- The jth falling factorial of x: $(x)_{j\downarrow} = x(x-1)\cdots(x-j+1)$, with $(x)_{0\downarrow} = 1$.
- The jth rising factorial of x: $(x)_{j\uparrow} = x(x+1)\cdots(x+j-1)$, with $(x)_{0\uparrow} = 1$.

In the literature there are several different notations for falling and rising factorials. The above notation follows Pitman (2006), which is an interesting read. Employing this notation, it is simple to show that

$$p_{kj}^{(g)} = \frac{(N_{g+1})_{j\downarrow}}{(N_{g+1})^k} S(k,j), \tag{1.2}$$

where $S(k,j) = \frac{1}{j!}\sum_{l=0}^{j}(-1)^{j-l}\binom{j}{l}l^k$ are Stirling numbers of the second kind (the number of partitions of $[k] = \{1,\ldots,k\}$ into j non-empty subsets). Note that $\sum_{j=1}^{k} p_{kj}^{(g)} = 1$, for all g, follows from the combinatorial identity

$$x^k = \sum_{j=1}^{k} S(k,j)(x)_{j\downarrow}.$$

Directly evaluating (1.2) warrants caution, since $S(k,j), (N_{g+1})_{j\downarrow}$, and $(N_{g+1})^k$ can each be very large. A more efficient and numerically stable way to compute $p_{kj}^{(g)}$ is to use the following result (Bhaskar et al, 2014):

Proposition 1.2. *For $1 \le j \le k$, $p_{kj}^{(g)}$ satisfies the recursion*

$$p_{kj}^{(g)} = \left(\frac{j}{N_{g+1}}\right) p_{k-1,j}^{(g)} + \left(\frac{N_{g+1}-j+1}{N_{g+1}}\right) p_{k-1,j-1}^{(g)}, \tag{1.3}$$

with boundary conditions $p_{1,1}^{(g)} = 1$ and $p_{k,j}^{(g)} = 0$ if $k < j$.

Proof. This result follows from (1.2) and the well-known recursion

$$S(k,j) = j \cdot S(k-1,j) + S(k-1,j-1)$$

satisfied by $S(k,j)$ for $0 < j \le k$. □

Generalizing the above discussion, we define the following stochastic process:

Definition 1.3 (Discrete-time ancestral process). Suppose a sample of size n is taken at $g = 0$. Let $A_n^{\mathrm{D}}(g)$ denote the number of distinct ancestors of the sample at time g. The process $\{A_n^{\mathrm{D}}(g), g = 0,1,2,...\}$, called the discrete-time ancestral process, has the following properties:

1. $A_n^{\mathrm{D}}(0) = n$.
2. It is a discrete-time Markov chain with state space $[n]$. Specifically, for all $g = 0,1,2,\ldots,$ and $j,k \in [n]$,
$$\mathbb{P}(A_n^{\mathrm{D}}(g+1) = j \mid A_n^{\mathrm{D}}(g) = k) = p_{kj}^{(g)},$$
where $p_{kj}^{(g)}$ is defined in (1.2).
3. It is a pure death process.

Note that $\{A_n^{\mathrm{D}}(g), g \ge 0\}$ is a homogeneous Markov process if and only if the population size is constant; i.e., $N_g = N$ for all g. How can one compute $\mathbb{P}(A_n^{\mathrm{D}}(g) = j)$? Define $\mathbf{P}^{(g)} := (p_{kj}^{(g)})_{k,j\in[n]}$. Then,

$$\mathbb{P}(A_n^{\mathrm{D}}(g) = j) = [\mathbf{P}^{(0)} \cdots \mathbf{P}^{(g-1)}]_{nj}, \tag{1.4}$$

which takes $O(n^3 g)$-time to evaluate using naive matrix multiplication, if $p_{kj}^{(g)}$ are known. A more efficient method is to use the following recursion (Bhaskar et al, 2014), which is left as a simple exercise for the reader to prove:

Proposition 1.4. *For $g > 0$, $\mathbb{P}(A_n^{\mathrm{D}}(g) = j)$ satisfies the recursion*

$$\mathbb{P}(A_n^{\mathrm{D}}(g) = j) = \sum_{k=j}^{n} p_{kj}^{(g-1)} \mathbb{P}(A_n^{\mathrm{D}}(g-1) = k), \tag{1.5}$$

with $\mathbb{P}(A_n^{\mathrm{D}}(0) = j) = \delta_{nj}$.

This recursion takes $O(n^2 g)$-time to evaluate, if $p_{kj}^{(g)}$ are known. Note that Proposition 1.4 also holds for other random mating models, with appropriate one-generation transition probability $p_{kj}^{(g)}$.

The discrete-time process $\{A_n^{\mathrm{D}}(g), g \geq 0\}$ is rather cumbersome to work with, and neither (1.4) nor (1.5) provides much intuition about the dynamics of the ancestral process; e.g., how $\mathbb{P}(A_n^{\mathrm{D}}(g) = j)$ changes over time. We hence turn to a continuous-time approximation to the discrete-time process.

1.3 A large-N limit

In what follows, we assume that $N_g = N > 0$ for all g and use $\{A_{N,n}^{\mathrm{D}}(g), g = 0, 1, 2, \ldots\}$ to denote the corresponding discrete-time ancestral process, with p_{kj} as the one-generation transition probability. Using the fact that $S(k,k) = 1$ and $S(k, k-1) = \binom{k}{2}$, it is simple to show that the transition probability p_{kj} takes the following form:

$$p_{kj} = \frac{(N)_{j\downarrow}}{(N)^k} S(k,j) = \begin{cases} 1 - \binom{k}{2}\frac{1}{N} + O\left(\frac{1}{N^2}\right), & \text{if } j = k, \\ \binom{k}{2}\frac{1}{N} + O\left(\frac{1}{N^2}\right), & \text{if } j = k-1, \\ O\left(\frac{1}{N^{k-j}}\right), & \text{if } j < k-1. \end{cases}$$

Define an n-by-n matrix of transition probabilities as follows:

$$\mathbf{P}_N = (p_{kj}) = I_{n\times n} + Q\frac{1}{N} + O\left(\frac{1}{N^2}\right), \tag{1.6}$$

where

$$Q = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 & 0 \\ -\lambda_2 & \lambda_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -\lambda_n & \lambda_n \end{pmatrix}, \tag{1.7}$$

with $\lambda_k = -\binom{k}{2}$. (Note that $\lambda_1 = 0$.) Now, for a given sample size n and a given real number $t \in \mathbb{R}_{\geq 0}$,

$$\lim_{N\to\infty} (\mathbf{P}_N)^{\lfloor Nt \rfloor} = e^{Qt},$$

where $\lfloor \cdot \rfloor$ denotes the floor function. This shows the following weak convergence:

Theorem 1.5. *Suppose $N_g = N$ for all g. Then, for every fixed sample size n and fixed positive number t, $A^{\mathrm{D}}_{N,n}(\lfloor Nt \rfloor)$ converges weakly to $A_n(t)$ as $N \to \infty$, where $\{A_n(t), t \in \mathbb{R}_{\geq 0}\}$ is a continuous-time homogeneous Markov process with state space $[n]$ and infinitesimal generator Q.*

Remark 1.6. Note that $A_n(0) = n$, and, instantaneously, only jumps of size one are allowed in $\{A_n(t), t \in \mathbb{R}_{\geq 0}\}$.

Using the above result, one can approximate $A^{\mathrm{D}}_n(\lfloor Nt \rfloor)$ by $A_n(t)$, where 1 unit of time in $\{A_n(t), t \in \mathbb{R}_{\geq 0}\}$ roughly corresponds to N generations in $\{A^{\mathrm{D}}_n(g), g = 0, 1, 2, \ldots\}$. When is this a good approximation? To answer this question, let us examine the higher order terms that were ignored in (1.6). In particular,

$$p_{m,m-2} = \Theta\left(\frac{m^4}{N^2}\right) + O\left(\frac{1}{N^3}\right).$$

So, if $m = \Theta(\sqrt{N})$, then $p_{m,m-2} = \Theta(1)$, which is not negligible. Hence, $A_n(t)$ provides a good approximation to the discrete-time ancestral process only if the population size N is sufficiently large compared to the sample size n. This is an important point to note, since the sample size is rapidly increasing in modern population genetics. See Chapter 11 and Bhaskar et al (2014) for a more detailed comparison of the discrete- and continuous-time ancestral processes.

The continuous-time ancestral process is useful because it facilitates computation, as illustrated in the following theorem:

Theorem 1.7 (Tavaré 1984). *For the continuous-time ancestral process $\{A_n(t), t \geq 0\}$,*

$$\mathbb{P}(A_n(t) = j) = \sum_{k=j}^{n} e^{-\binom{k}{2}t} \left[\frac{(-1)^{k-j}(2k-1)(j)_{(k-1)\uparrow}(n)_{k\downarrow}}{j!(k-j)!(n)_{k\uparrow}} \right], \tag{1.8}$$

which corresponds to the probability of there being j ancestors at time t of a sample of size n taken at time 0.

Proof. Recall that $\mathbb{P}(A_n(t) = j) = [e^{Qt}]_{nj}$, where the infinitesimal generator Q is shown in (1.7). The characteristic equation for Q is

$$\det(Q - \lambda I) = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda) = 0,$$

which implies that Q has n distinct eigenvalues (namely, $\lambda_1, \ldots, \lambda_n$), and so Q is diagonalizable. Decompose Q as

$$Q = S \Lambda S^{-1},$$

where S denotes the right eigenvector matrix. The matrices S and S^{-1} are both lower triangular. The right eigenvector with eigenvalue λ_k is $\boldsymbol{r}_k = (r_{k,1}, \ldots, r_{k,n})$, where

$$r_{k,m} = \begin{cases} 0, & m < k, \\ \binom{m}{k} \dfrac{(k)_{k\uparrow}}{(m)_{k\uparrow}}, & m \geq k. \end{cases}$$

The left eigenvector with eigenvalue λ_k is $\boldsymbol{\ell}_k = (\ell_{k,1}, \ldots, \ell_{k,n})$, where

$$\ell_{k,m} = \begin{cases} (-1)^{k-m}\binom{k}{m}\dfrac{(m)_{(k-1)\uparrow}}{(k)_{(k-1)\uparrow}}, & m \le k, \\ 0, & m > k. \end{cases}$$

The kth column of S is $\boldsymbol{r}_k$, while the kth row of S^{-1} is $\boldsymbol{\ell}_k$. (You may convince yourself that $\sum_{m=1}^n r_{i,m}\ell_{j,m} = \delta_{ij}$.) Therefore,

$$[e^{Qt}]_{nj} = \sum_{k=1}^n e^{-\binom{k}{2}t} r_{k,n}\ell_{k,j} = \sum_{k=j}^n e^{-\binom{k}{2}t} r_{k,n}\ell_{k,j},$$

and (1.8) follows after some algebra. □

The closed-form formula (1.8) is useful for gaining intuition about the process. For example, it shows that the transition probability $\mathbb{P}(A_n(t) = j)$ for $j > 1$ decays exponentially with time. However, (1.8) is numerically unstable even for moderate sample sizes (say, $n \ge 50$), because it involves an alternating sum suffering from catastrophic cancellation. A more numerically stable way of evaluating $\mathbb{P}(A_n(t) = j)$ is via exponentiating the rate matrix Q. See Moler and Van Loan (1978) for a useful review of matrix exponentiation.

It turns out that the moments of $A_n(t)$ admits a closed-form formula which is numerically stable to evaluate:

Theorem 1.8 (Tavaré 1984). *The jth factorial moment of $A_n(t)$ is given by*

$$\mathbb{E}[(A_n(t))_{j\downarrow}] = \sum_{k=j}^n e^{-\binom{k}{2}t}\left[(2k-1)\binom{k-1}{j-1}(k)_{(j-1)\uparrow}\frac{(n)_{k\downarrow}}{(n)_{k\uparrow}}\right]. \tag{1.9}$$

Using the above formula, one can easily compute the mean and the variance of $A_n(t)$:

$$\mathbb{E}[A_n(t)] = \sum_{k=1}^n e^{-\binom{k}{2}t}(2k-1)\frac{(n)_{k\downarrow}}{(n)_{k\uparrow}},$$

$$\operatorname{Var}(A_n(t)) = \sum_{k=1}^n e^{-\binom{k}{2}t}(2k-1)(k^2-k+1)\frac{(n)_{k\downarrow}}{(n)_{k\uparrow}} - \left[\sum_{k=1}^n e^{-\binom{k}{2}t}(2k-1)\frac{(n)_{k\downarrow}}{(n)_{k\uparrow}}\right]^2.$$

Since each summand is positive, there is no numerical problem in evaluating the sum. Furthermore, note that the exponential factor $e^{-\binom{k}{2}t}$ decays rapidly for large k, so, when the sample size n is very large, one can obtain an accurate approximation of (1.9) by truncating the summation appropriately (the number of terms required would depend to the value of t).

1.4 Waiting time while there are k ancestors

For $k = n, n-1, \ldots, 2$, let $T_{n,k}$ denote the time spent in $\{A_n(t), t \geq 0\}$ while there are k ancestral lineages. It follows from the infinitesimal generator (1.7) that $T_{n,k} \sim \text{Exp}[\binom{k}{2}]$; i.e., the probability density function of $T_{n,k}$ is

$$f_{T_{n,k}}(t) = \binom{k}{2} \exp\left[-\binom{k}{2} t\right].$$

Another way to see this is to note that $T_{n,k} \stackrel{d}{=} T_{k,k}$ and $\mathbb{P}(T_{k,k} \geq t) = \mathbb{P}(A_k(t) = k) = e^{-\binom{k}{2}t}$, where the last equality follows from Theorem 1.7.

For the constant population size case considered above, $T_{n,n}, T_{n,n-1}, \ldots, T_{n,2}$ are independent random variables, which is a very useful property. For example, noting that $\mathbb{P}(A_n(t) \leq j) = \mathbb{P}(T_{n,n} + \cdots + T_{n,j+1} \leq t)$ and that $T_{n,n}, \ldots, T_{n,j+1}$ are independent random variables, Griffiths (1984) applied the Lyapunov central limit theorem (Billingsley, 2008, Chapter 27) to show that $A_n(t)$ is asymptotically normally distributed in the limit $t \to 0$ and $n \to \infty$ such that $nt \to a > 0$. This is a useful result since one can encounter numerical problems in computing $\mathbb{P}(A_n(t) = j)$ for small values of t.

Later we will study the continuous-time ancestral process for a population with variable size, in which case, the waiting times $T_{n,n}, T_{n,n-1}, \ldots, T_{n,2}$ are no longer independent and the distribution of $T_{n,k}$ is in general different from that of $T_{k,k}$.

1.5 Tree height

Under constant population size, since $T_{n,k} \stackrel{d}{=} T_{k,k}$, people often omit the dependence on the sample size and simply write T_k to denote $T_{n,k}$. Henceforth we will employ this convention.

Definition 1.9 (Time to the MRCA). We use W_n to denote the waiting time to the most recent common ancestor (MRCA):

$$W_n = \inf\{t \geq 0 \mid A_n(t) = 1\}.$$

Note that

$$W_n = T_{n,n} + T_{n,n-1} + \cdots + T_{n,2}. \tag{1.10}$$

Using the results discussed in previous sections, one can show the following:

Theorem 1.10. *Under a constant population size, the probability density function f_{W_n} of W_n is*

$$\begin{aligned} f_{W_n}(t) &= \sum_{k=2}^{n} \binom{k}{2} e^{-\binom{k}{2}t} \left[\frac{(-1)^k (2k-1)(n)_{k\downarrow}}{(n)_{k\uparrow}}\right] \\ &= \sum_{k=2}^{n} \binom{k}{2} e^{-\binom{k}{2}t} \prod_{j=2: j\neq k}^{n} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{k}{2}} \end{aligned} \tag{1.11}$$

Proof. The first equality in (1.11) follows from $\mathbb{P}(W_n \leq t) = \mathbb{P}(A_n(t) = 1)$, which implies

$$f_{W_n}(t) = \frac{d}{dt}\mathbb{P}(A_n(t) = 1),$$

where $\mathbb{P}(A_n(t) = 1)$ is defined in (1.8). The second equality in (1.11) can be shown using (1.10). First, for independent $X_1 \sim \text{Exp}(\lambda_1)$ and $X_2 \sim \text{Exp}(\lambda_2)$, where $\lambda_1 \neq \lambda_2$, note that

$$f_{X_1+X_2}(t) = \int_0^t f_{X_1}(y) f_{X_2}(t-y)dy = \lambda_1 e^{-\lambda_1 t}\frac{\lambda_2}{\lambda_2 - \lambda_1} + \lambda_2 e^{-\lambda_2 t}\frac{\lambda_1}{\lambda_1 - \lambda_2}.$$

More generally, if $X_i \sim \text{Exp}(\lambda_i)$ are independent exponential random variables and λ_i are all distinct for $i = 1, \ldots, m$, then one can derive the following convolution formula:

$$f_{X_1+\cdots+X_m}(t) = \sum_{k=1}^{m}\left[\lambda_k e^{-\lambda_k t}\prod_{j=1, j\neq k}^{m}\frac{\lambda_j}{\lambda_j - \lambda_k}\right]. \tag{1.12}$$

This convolution formula and (1.10) together imply the desired result. □

The expected waiting time $\mathbb{E}(W_n)$ to the MRCA can be computed using (1.11):

$$\mathbb{E}(W_n) = \int_0^\infty t f_{W_n}(t)dt.$$

There is a much simpler way to compute the expectation, however. Using (1.10), we immediately obtain

$$\mathbb{E}(W_n) = \mathbb{E}(T_2 + T_3 + \cdots + T_n) = \sum_{i=2}^{n}\mathbb{E}(T_i) = \sum_{k=2}^{n}\frac{1}{\binom{k}{2}} = 2\left(1 - \frac{1}{n}\right). \tag{1.13}$$

Note that (1.13) is bounded from above by 2, which corresponds to $4N$ discrete generations. Furthermore, $\mathbb{E}[T_2] = 1$, suggesting that the dominant contribution to the expected waiting time to the MRCA comes from T_2 (the time spent while there are 2 lineages). The rate of coalescence is very fast when there are many lineages, but it slows down as the number of lineages decreases.

Similarly, using the independence of $T_2, \ldots, T_n$, we can compute $\text{Var}(W_n)$ as

$$\text{Var}(W_n) = \sum_{k=2}^{n}\text{Var}\left(T_k\right) = \sum_{k=2}^{n}\left[\frac{1}{\binom{k}{2}}\right]^2 = 8\sum_{k=1}^{n-1}\frac{1}{k^2} - 4\left(1 - \frac{1}{n}\right)\left(3 + \frac{1}{n}\right),$$

which implies

$$\lim_{n\to\infty}\text{Var}(W_n) = 8\sum_{i=1}^{\infty}\frac{1}{i^2} - 12 = 8\frac{\pi^2}{6} - 12 \approx 1.16.$$

(The identity $\sum_{i=1}^{\infty}\frac{1}{i^2} = \frac{\pi^2}{6}$ was proved by Euler in the early 18th century, until when providing a proof had remained an intriguing open problem for about 90 years.) Finally, since

$$1 = \text{Var}(T_2) \leq \text{Var}(W_n) \leq 1.16,$$

we see that T_2 is the most variable part of W_n.

1.6 Tree length

Definition 1.11 (Tree length). The length of a coalescent tree is defined as

$$L_n = 2T_{n,2} + 3T_{n,3} + \cdots + nT_{n,n}.$$

Theorem 1.12. *Under a constant population size, the probability density function f_{L_n} of L_n is given by*

$$f_{L_n}(t) = \sum_{k=2}^{n} (-1)^k \binom{n-1}{k-1} \frac{k-1}{2} e^{-\frac{k-1}{2}t} = \frac{n-1}{2} e^{-\frac{t}{2}} \left(1 - e^{-\frac{t}{2}}\right)^{n-2}. \tag{1.14}$$

Proof. Using the convolution formula (1.12), one obtains

$$f_{L_n}(t) = \sum_{k=2}^{n} \frac{k-1}{2} e^{-\frac{k-1}{2}t} \prod_{j=2, j\neq k}^{n} \frac{j-1}{j-k}.$$

The product of fractions can be written as

$$\begin{aligned}
\prod_{j=2, j\neq k}^{n} \frac{j-1}{j-k} &= \left(\frac{1}{2-k}\right)\left(\frac{2}{3-k}\right)\cdots\left(\frac{k-2}{k-1-k}\right)\left(\frac{k}{k+1-k}\right)\cdots\left(\frac{n-1}{n-k}\right) \\
&= \frac{(k-2)!(-1)^{k-2}}{(k-2)!} \times \frac{(n-1)!}{(k-1)!(n-k)!} \\
&= (-1)^k \binom{n-1}{k-1},
\end{aligned}$$

from which the first equality of (1.14) follows. Showing the second equality of (1.14) is left as an exercise. □

The expected tree length is

$$\mathbb{E}(L_n) = \sum_{k=2}^{n} \mathbb{E}(kT_k) = \sum_{k=2}^{n} k\mathbb{E}(T_k) = \sum_{k=2}^{n} \frac{2}{k-1}.$$

For large n, this is approximately equal to $2[\log(n-1) + \gamma_E]$, where γ_E is the Euler-Mascheroni constant. Again, using the independence of $T_2, \ldots, T_n$, the variance of L_n can be computed as

$$\mathrm{Var}(L_n) = \sum_{k=2}^{n} \mathrm{Var}(kT_k) = \sum_{k=2}^{n} k^2 \,\mathrm{Var}(T_k) = \sum_{k=2}^{n} \frac{4}{(k-1)^2} \leq \frac{2\pi^2}{3}.$$

Hence, although $\mathbb{E}(L_n)$ blows up as $n \to \infty$, $\mathrm{Var}(L_n)$ stays finite.

1.7 The ancestral lineage of a particular leaf

Consider a sample of size n and a particular leaf $\ell \in [n]$. Follow the ancestral lineage of ℓ until it is involved in a coalescence event, denoted E_ℓ. Let X_n denote the total number of surviving lineages (including the ancestral lineage of ℓ) when the event E_ℓ occurs, and let Ω_n denote the waiting time until event E_ℓ occurs.

Proposition 1.13. *For $k = 2, 3, \ldots, n$,*

$$\mathbb{P}(X_n = k) = \frac{k-1}{\binom{n}{2}}. \tag{1.15}$$

Proof. While there are j lineages, the probability that a particular lineage is involved in a coalescence event is

$$\frac{j-1}{\binom{j}{2}} = \frac{2}{j},$$

which follows from exchangeability. Hence,

$$\mathbb{P}(X_n = k) = \left[\prod_{j=n}^{k+1}\left(1 - \frac{2}{j}\right)\right]\frac{2}{k} = \left[\frac{k(k-1)}{n(n-1)}\right]\frac{2}{k} = \frac{k-1}{\binom{n}{2}},$$

as desired. □

Proposition 1.14. *The expected waiting time until leaf ℓ coalesces is given by*

$$\mathbb{E}(\Omega_n) = \frac{2}{n}.$$

Proof. Method 1: First decompose $\mathbb{E}(\Omega_n)$ as

$$\mathbb{E}(\Omega_n) = \sum_{k=2}^{n} \mathbb{E}(\Omega_n | X_n = k)\mathbb{P}(X_n = k),$$

and then note that $\mathbb{E}(\Omega_n | X_n = k) = \mathbb{E}[S_{n,k}]$, where

$$S_{n,k} = T_{n,n} + T_{n,n-1} + \cdots + T_{n,k}.$$

Plugging this in, we get

$$\mathbb{E}(\Omega_n) = \sum_{k=2}^{n} \mathbb{E}(S_{n,k})\frac{k-1}{\binom{n}{2}} = \sum_{k=2}^{n}\sum_{j=n}^{k} \mathbb{E}(T_{n,j})\frac{k-1}{\binom{n}{2}}.$$

If the population size is constant, this becomes

$$\mathbb{E}(\Omega_n) = \sum_{k=2}^{n}\sum_{j=n}^{k} \frac{1}{\binom{j}{2}}\frac{k-1}{\binom{n}{2}} = \frac{2}{n}, \tag{1.16}$$

which is the desired result.

Method 2: An alternative approach is to use a recursion (for constant population size). Let F denote the event that the first coalescence events involves ℓ. We know from Proposition 1.13 that $\mathbb{P}(F) = 2/n$. So,

$$\begin{aligned}\mathbb{E}(\Omega_n) &= \mathbb{E}(\Omega_n|F)\mathbb{P}(F) + \mathbb{E}(\Omega_n|F^c)\mathbb{P}(F^c) \\ &= \mathbb{E}(T_{n,n}) \cdot \frac{2}{n} + \mathbb{E}(T_{n,n} + \Omega_{n-1}) \left(1 - \frac{2}{n}\right) \\ &= \mathbb{E}(T_{n,n}) + \left(1 - \frac{2}{n}\right) \mathbb{E}(\Omega_{n-1}) \\ &= \frac{1}{\binom{n}{2}} + \left(1 - \frac{2}{n}\right) \mathbb{E}(\Omega_{n-1}) \end{aligned} \tag{1.17}$$

where $\mathbb{E}(\Omega_2) = \mathbb{E}(T_{n,2}) = 1$. Now we have a recursion for $\mathbb{E}(\Omega_n)$ in terms of $\mathbb{E}(\Omega_{n-1})$, and solving this recursion gives $\mathbb{E}(\Omega_n) = \frac{2}{n}$. □

References

Bhaskar A, Clark AC, Song Y (2014) Distortion of genealogical properties when the sample is very large. Proc Nat Acad Sci 111(6):2385–2390, (PMC3926037)

Billingsley P (2008) Probability and Measure. John Wiley & Sons

Griffiths RC (1984) Asymptotic line-of-descent distributions. Journal of Mathematical Biology 21(1):67–75

Kingman JFC (1982a) The coalescent. Stoch Process Appl 13:235–248

Kingman JFC (1982b) Exchangeability and the evolution of large populations. In: Koch G, Spizzichino F (eds) Exchangeability in Probability and Statistics, North-Holland Publishing Company, pp 97–112

Kingman JFC (1982c) On the genealogy of large populations. J Appl Prob 19A:27–43

Moler C, Van Loan C (1978) Nineteen dubious ways to compute the exponential of a matrix. SIAM Review 20(4):801–836

Pitman J (2006) Combinatorial Stochastic Processes. Springer-Verlag

Tavaré S (1984) Line-of-descent and genealogical processes, and their applications in population genetics models. Theoretical Population Biology 26:119–164

Chapter 2
Kingman's coalescent

In the last chapter we focused on the number of ancestral lineages, but did not consider who is related to whom. Here we study a more refined stochastic process which keeps track of that information.

2.1 The n-coalescent

In this section we introduce Kingman's coalescent and study its probability distribution. First, we define some notation. Recall that $[n]$ denotes the n-set $\{1, 2, \ldots, n\}$ and $\mathcal{P}_{[n]}$ the set of all partitions of $[n]$. Given a partition $\alpha = \{B_1, B_2, \ldots, B_k\} \in \mathcal{P}_{[n]}$ — i.e., $B_i \subset [n]$, $B_i \neq \varnothing$ for all $i = 1, \ldots, k$, and $B_1 \cup \cdots \cup B_k = [n]$ — we employ the following notation:

1. $|\alpha|$:= number of blocks (non-empty subsets of $[n]$) in α. ($|\alpha| = k$ in the above example.)
2. $|B_i|$:= number of elements in B_i. We define $b_i := |B_i|$ in what follows.

The following partial order on $\mathcal{P}_{[n]}$ will be of particular interest:

Definition 2.1. Given $\alpha, \beta \in \mathcal{P}_{[n]}$, we write $\alpha \prec \beta$ if β is obtained from α by merging exactly 2 blocks in α into a single block. For example, if $\alpha = \{\{1, 3\}, \{2, 5\}, \{4\}\}$ and $\beta = \{\{1, 3, 4\}, \{2, 5\}\}$, then $\alpha \prec \beta$, and we say that "α *precedes* β".

We can now provide a formal definition of Kingman's n-coalescent:

Definition 2.2 (Kingman's n-coalescent). The n-coalescent (Kingman, 1982a,b,c), denoted $\{C_n(t), t \geq 0\}$, is a continuous-time Markov process on $\mathcal{P}_{[n]}$ with the following properties:

1. $C_n(0) = \{\{1\}, \{2\}, \ldots, \{n\}\}$.
2. For $\alpha, \beta \in \mathcal{P}_{[n]}$, the infinitesimal generator is given by

$$q_{\alpha\beta} = \begin{cases} -\binom{|\alpha|}{2}, & \text{if } \alpha = \beta, \\ 1, & \text{if } \alpha \prec \beta, \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

3. $\lim_{t\to\infty} C_n(t) = \{\{1, 2, \ldots, n\}\}$ with probability 1.

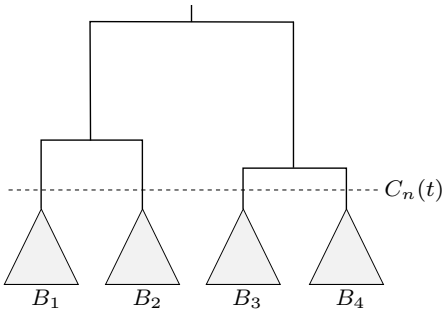


Fig. 2.1: A random partition of the leaf-set $[n]$ induced by cutting a coalescent tree at time t (dashed line). In this example, there are four ancestral lineages at time t, and the leaves they subtend are partitioned into four blocks, $\{B_1, \ldots, B_4\} \in \mathcal{P}_{[n]}$.

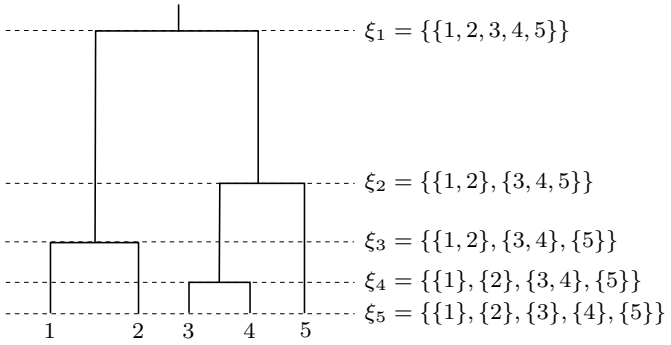


Fig. 2.2: Illustration of the jump chain embedded in the n-coalescent.

As mentioned in Chapter 1 (cf., Figure 1.2), a sample path of the n-coalescent can be represented by an n-leaved rooted binary ultrametric tree with its leaves bijectively labeled by $[n] = \{1, \ldots, n\}$. Suppose the tree is cut horizontally at a given time t when there are k ancestral lineages, leading to a collection of k subtrees underneath t. As illustrated in Figure 2.1, this induces a partition $\alpha = \{B_1, \ldots, B_k\}$ of $[n]$ into k blocks, with the leaf-set of each subtree corresponding to a unique block of α. Our goal is to obtain the probability distribution of such a partition structure under the n-coalescent.

Definition 2.3 (Jump chain). The jump chain $\{\xi_{n,k}, k = n, n-1, \ldots, 2, 1\}$ embedded in the n-coalescent is the discrete-time Markov process on $\mathcal{P}_{[n]}$ obtained by restricting to the times when the state changes in $\{C_n(t), t \geq 0\}$. See Figure 2.2 for an example with $n = 5$.

Several things can be said about this jump chain:

1. $|\xi_{n,k}| = k$.
2. $\xi_{n,n} = C_n(0) = \{\{1\}, \{2\}, \ldots, \{n\}\}$.
3. The infinitesimal generator $q_{\alpha\beta}$ implies

$$\mathbb{P}(\xi_{n,k-1} = \beta \mid \xi_{n,k} = \alpha) = \begin{cases} \frac{1}{\binom{k}{2}}, & \text{if } \alpha \prec \beta \text{ and } |\alpha| = k, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

4. $C_n(t)$ moves through $\xi_{n,n} \succ \xi_{n,n-1} \succ \cdots \succ \xi_{n,1}$.
5. Time spent in state $\xi_{n,k}$ is $T_{n,k} \sim \text{Exp}[\binom{k}{2}]$, where $T_{n,k}$ = time spent while there are k lineages.
6. Let $A_n(t) = |C_n(t)|$. The ancestral process $\{A_n(t), t \geq 0\}$ and the jump chain $\{\xi_{n,k}, k = n, \ldots, 1\}$ are independent processes.
7. $C_n(t) = \xi_{n,A_n(t)}$, for all $t \geq 0$.

Using the above facts, the probability distribution of $C_n(t)$ can be obtained as

$$\begin{aligned}\mathbb{P}\big(C_n(t) = \alpha\big) &= \sum_{j=1}^{n} \mathbb{P}(C_n(t) = \alpha | A_n(t) = j)\mathbb{P}(A_n(t) = j) \\ &= \sum_{j=1}^{n} \mathbb{P}(\xi_{n,j} = \alpha)\mathbb{P}(A_n(t) = j) \\ &= \mathbb{P}(\xi_{n,|\alpha|} = \alpha)\mathbb{P}(A_n(t) = |\alpha|),\end{aligned}$$

where $\alpha \in \mathcal{P}_{[n]}$ and the second term $\mathbb{P}(A_n(t) = |\alpha|)$ is given by (1.8). The following theorem provides a closed-form expression for the first term, which describes how the leaf labels $\{1, 2, \ldots, n\}$ are partitioned into $|\alpha|$ blocks:

Theorem 2.4. *Suppose $\alpha = \{B_1, B_2, \ldots, B_k\} \in \mathcal{P}_{[n]}$ and $b_i = |B_i|$ for $i = 1, \ldots, k$. Then,*

$$\mathbb{P}(\xi_{n,k} = \alpha) = \frac{k!}{\binom{n}{b_1, b_2, \ldots, b_k}} \frac{1}{\binom{n-1}{k-1}}, \tag{2.3}$$

where $\binom{n}{b_1, b_2, \ldots, b_k}$ denotes the multinomial coefficient.

Proof. This result can be proved using backward induction on k, which we sketch here. For the base case of $k = n$, we have $\alpha = \{\{1\}, \{2\}, \ldots \{n\}\}$ and $\mathbb{P}(\xi_{n,n} = \alpha) = 1$, which agrees with the right hand side of (2.3). Now suppose (2.3) holds for $k = n, \ldots, j$. Then, for $\beta \in \mathcal{P}_{[n]}$ with $|\beta| = j - 1$,

$$\begin{aligned}\mathbb{P}(\xi_{n,j-1} = \beta) &= \sum_{\alpha \in \mathcal{P}_{[n]}} \mathbb{P}(\xi_{n,j-1} = \beta \mid \xi_{n,j} = \alpha)\, \mathbb{P}(\xi_{n,j} = \alpha) \\ &= \sum_{\alpha : \alpha \prec \beta} \mathbb{P}(\xi_{n,j-1} = \beta \mid \xi_{n,j} = \alpha)\, \mathbb{P}(\xi_{n,j} = \alpha) \\ &= \sum_{\alpha : \alpha \prec \beta} \frac{1}{\binom{j}{2}} \mathbb{P}(\xi_{n,j} = \alpha).\end{aligned}$$

To complete the proof, we need to use the inductive hypothesis and plug in the expression for $\mathbb{P}(\xi_{n,j} = \alpha)$ in the last equation. That the result is equivalent to (2.3) involves some algebra, which we leave to the reader as an exercise. □

Note that the expression for $\mathbb{P}(\xi_{n,k} = \alpha)$ in (2.3) depends on the sizes of the blocks, but not what they contain; i.e., the partition structure is exchangeable. Specifically, it is proportional to $b_1! \cdots b_k!$. Hence, as illustrated in Table 2.1, uneven partitions are more likely than evenly balanced partitions.

Table 2.1: Comparison of the combinatorial factor in (2.3) that depends on block sizes. This example is for $n = 12$ and $k = 3$. The probability distribution (2.3) is proportional to $b_1! \cdots b_k!$, which implies that uneven partitions are more likely than evenly balanced partitions.

(b_1, b_2, b_3)	$b_1!b_2!b_3!$
(4, 4, 4)	13824
(5, 3, 4)	17280
(9, 2, 1)	725760
(10, 1, 1)	3628800

2.2 Subtree leaf-set sizes

The intuitive interpretation of (2.3) is this: to sample $\xi_{n,k}$, we first sample the sizes of the k subsets uniformly from vectors in $\mathbb{Z}_+^k$ that sum to n, and then conditional on these sizes choose a valid assignment of the sample to these subsets uniformly. More specifically, the following result holds:

Theorem 2.5. *Consider a time t when there are k lineages, label these edges as $e_1, e_2, \ldots, e_k$, and define Z_i to be the number of descendant leaves of e_i, as illustrated in Figure 2.3. Then,*

$$\mathbb{P}(Z_1 = z_1, \ldots, Z_k = z_k \mid A_n(t) = k) = \frac{1}{\binom{n-1}{k-1}} \tag{2.4}$$

when the $z_1, \ldots, z_k$ are positive integers that sum to n.

In words, $Z = (Z_1, \ldots, Z_k)$ is sampled uniformly from all compositions of n into k parts (i.e., all k-dimensional vectors of positive integers that sum to n). Below we present an inductive proof. In Chapter 2.7, we will provide a more direct proof of (2.4) using an urn model.

Proof. We use induction on n. For $n = 2$ and $k = 2$, the only possibility is $z_1 = z_2 = 1$, so (2.4) is certainly true. Assume that (2.4) is true up to $n = m - 1$ and $2 \le k \le m - 1$. Consider $n = m$. If $k = m$, then $z_i = 1$ for all $i = 1, \ldots, k$, so (2.4) holds. If $k < m$, then consider the bottommost coalescent vertex v (i.e., the interior vertex closest to the leaves). The probability that v appears as descendant of e_j is $(z_j - 1)/(m - 1)$. So,

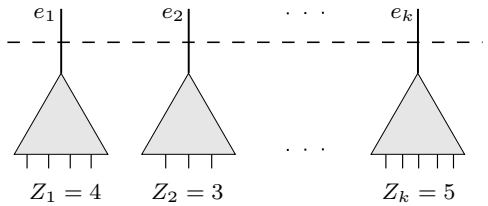


Fig. 2.3: Subtree leaf-set sizes induced by cutting a coalescent tree at a certain time (dotted line) when there are k lineages. The lineages are labeled $e_1, \ldots, e_k$ and the number of leaves subtended by edge e_i is denoted by Z_i.

$$\begin{aligned}\mathbb{P}(z_1,\ldots,z_k) &= \sum_{j=1}^{k} \frac{z_j-1}{m-1}\,\mathbb{P}(z_1,\ldots,z_{j-1},z_j-1,z_{j+1},\ldots,z_k)\\ &= \sum_{j=1}^{k} \frac{z_j-1}{m-1}\,\frac{1}{\binom{m-2}{k-1}}\\ &= \frac{m-k}{m-1}\,\frac{1}{\binom{m-2}{k-1}} \;=\; \frac{1}{\binom{m-1}{k-1}},\end{aligned}$$

where the second line follows from the induction hypothesis. □

Theorem 2.5 can be used to find the marginal probability of $Z_i = m$. If e_i subtends m leaves, there are $n-m$ leaves remaining, and $k-1$ other edges. There are $\binom{n-m-1}{k-2}$ compositions of $n-m$ into $k-1$ positive integers, so we have the following result:

Corollary 2.6. *Let Z_i be defined as in Theorem 2.5. Then, for all $i = 1,\ldots,k$,*

$$\mathbb{P}(Z_i = m \mid A_n(t) = k) = \frac{\binom{n-m-1}{k-2}}{\binom{n-1}{k-1}}. \tag{2.5}$$

2.3 Some properties of a subsample

Here we present a classical result concerning a subsample that can be easily proved using what we have discussed in the previous section. The following theorem follows from exchangeability and applies to an arbitrary variable population size model. In Chapter 2.5, we provide an alternate proof based on counting tree topologies.

Theorem 2.7 (Saunders et al 1984). *Given an n-leaved random coalescent tree from the n-coalescent, take a subsample of size $m < n$ from $[n]$. Let $s_{n,m}$ denote the probability that the subsample has the same most recent common ancestor (MRCA) as the entire sample $[n]$. Then,*

$$s_{n,m} = \left(\frac{m-1}{m+1}\right)\left(\frac{n+1}{n-1}\right).$$

Proof. Here we provide a proof that utilizes Corollary 2.6. In Chapter 2.5, we provide an alternate proof based on counting coalescent tree topologies.

Let $(Z, n-Z)$ denote the number of leaves subtended by the two edges just before the MRCA of the entire sample. Let E denote the event that the subsample of size m has the same MRCA as the entire sample. Then,

$$\mathbb{P}(E \mid Z = z) \;=\; 1 - \frac{\binom{z}{m} + \binom{n-z}{m}}{\binom{n}{m}},$$

where $\binom{k}{j} \equiv 0$ if $k < j$. Note that

$$s_{n,m} = \sum_{z=1}^{n-1} \mathbb{P}(E \mid Z = z)\,\mathbb{P}(Z = z).$$

Therefore, using

$$\begin{aligned}\mathbb{P}(Z = z) &= \frac{1}{n-1},\\ \sum_{z=1}^{n-1} \binom{z}{m} &= \sum_{z=m}^{n-1} \binom{z}{m} = \binom{n}{m+1},\\ \sum_{z=1}^{n-1} \binom{n-z}{m} &= \sum_{z=1}^{n-m} \binom{n-z}{m} = \binom{n}{m+1},\end{aligned}$$

where the first identity follows from (2.5) with $k = 2$, we obtain

$$s_{n,m} = 1 - 2\frac{\binom{n}{m+1}}{\binom{n}{m}} \times \frac{1}{n-1} = 1 - \frac{2(n-m)}{(m+1)(n-1)} = \left(\frac{m-1}{m+1}\right)\left(\frac{n+1}{n-1}\right),$$

which is the desired result. □

Intuitively, $s_{n,m}$ is the probability that the subsample of size m contains at least one individual from each of the two subtrees branching from the MRCA of the entire sample $[n]$. Let Z_L and Z_R respectively denote the number of leaves in the left and the right subtrees. For all $1 \le k \le n-1$, (2.4) implies that $\mathbb{P}(Z_L = k, Z_R = n-k) = 1/(n-1)$, i.e., the distribution is uniform over all compositions of n into two positive parts. So, in the limit $n \to \infty$, the proportion X of leaves in the left subtree is distributed as a $U(0,1)$ random variable, for which the probability density is identically equal to 1. Hence, noting that the subsample is contained in the left (respectively, right) subtree is X^m (respectively, $(1-X)^m$), we obtain

$$s_{\infty,m} = 1 - \int_0^1 [x^m + (1-x)^m]\, dx = \frac{m-1}{m+1}.$$

For the case when $m = 2$ and n is large, $s_{n,m}$ is approximately $1/3$. Combined with recombination, this implies that the genealogy of just a pair of genomes will have parts that go far back into the past with high probability. This provides an intuitive explanation for why a recently developed method called the pairwise sequentially Markovian coalescent (PSMC) Li and Durbin (2011) is able to produce fairly accurate estimates of the effective population size (see Definition 2.21) in the distant past using only a pair of genomes.

We end this section with another interesting result concerning subsamples.

Theorem 2.8 (Forming a subtree). *Consider a subsample $\subset [n]$ of size $m < n$, as in the setting of Theorem 2.7. Let $E_{n,m}$ denote the event that an n-leaved random coalescent tree from the n-coalescent contains an edge that subtends only the subsample; i.e., the subsample forms a subtree. Then,*

$$\mathbb{P}[E_{n,m}] = \frac{2}{(m+1)\binom{n-1}{m-1}}. \tag{2.6}$$

Furthermore, for $m \ge 2$, conditioned on the event $E_{n,m}$, the probability that the first coalescence event back in time involves a pair of leaves in the subsample is

$$\frac{m+1}{n}.$$

Proof. In Chapter 2.5, we provide a proof based on counting coalescent tree topologies. Here, we present an alternate view on the problem. Let $f(n,m) = \mathbb{P}[E_{n,m}]$ and note that it satisfies the recursion

$$f(n,m) = \frac{\binom{m}{2}}{\binom{n}{2}} f(n-1,m-1) + \frac{\binom{n-m}{2}}{\binom{n}{2}} f(n-1,m), \tag{2.7}$$

with boundary conditions $f(n,1) = 1$ for all $n > 1$. We define $f(n,m) = 0$ if $m \geq n$, since this is not a valid sample-subsample pair. Note that the recursion for $f(n,m)$ is strict, since n decreases by 1 on the right hand side. Hence, given the boundary conditions, the recursion has a unique solution. By this uniqueness property, one just needs to show that the expression (2.6) satisfies the recursion (2.7) and the required boundary conditions. It is straightforward to show that both conditions are satisfied.

Now, let C be the event that the first coalescence back in time is between two leaves in the subsample. Then, from Bayes' rule, we have

$$\mathbb{P}[C \mid E_{n,m}] = \frac{\mathbb{P}[E_{n,m} \mid C]\ \mathbb{P}[C]}{\mathbb{P}[E_{n,m}]}.$$

We know $\mathbb{P}[E_{n,m}]$ from the above discussion and $\mathbb{P}[C]$ is simply given by

$$\mathbb{P}[C] = \frac{\binom{m}{2}}{\binom{n}{2}}.$$

Furthermore, $\mathbb{P}[E_{n,m} \mid C]$ is the probability that the subsample after the first coalescence event (which happens between two leaves in the subsample) forms a subtree in the coalescent tree of the remaining sample; i.e., the probability that a subsample of size $m-1$ forms a subtree in the coalescent tree of a sample of size $n-1$. Therefore, $\mathbb{P}[E_{n,m} \mid C] = \mathbb{P}[E_{n-1,m-1}]$. Putting all these things together yields $\mathbb{P}[C \mid E_{n,m}] = \frac{m+1}{n}$. □

Again, an alternate proof of the above result is provided in Chapter 2.5 by counting coalescent tree topologies.

2.4 Forward-in-time jump chain

We now turn our attention to the process forward in time (as opposed to backwards in time like the previous results). The steps below outline a way to run the coalescent process forward in time:

Theorem 2.9. *Given $\xi_{n,k} = \{B_1, B_2, \ldots, B_k\} \in \mathcal{P}_{[n]}$, $\xi_{n,k+1}$ in the jump chain of the n-coalescent can be sampled as follows.*

1. *Choose a block B_i with probability $\frac{b_i - 1}{n-k}$. (We want the numerator to be $b_i - 1$ since we do not want to split a block with only one lineage in it.)*
2. *Choose s uniformly at random from $\{1, 2, \ldots, b_i - 1\}$.*
3. *Choose a bipartition of B_i into blocks of size s and $b_i - s$ uniformly at random over all such partitions.*

Proof. We want to show that this process generates the same distribution over partition structures as the backward in time coalescent process. Let $\alpha, \beta \in \mathcal{P}_{[n]}$ and suppose $\alpha \prec \beta$. Then by Bayes' rule, we know

$$\mathbb{P}\left(\xi_{n,k+1} = \alpha \mid \xi_{n,k} = \beta\right) = \frac{\mathbb{P}(\xi_{n,k} = \beta \mid \xi_{n,k+1} = \alpha)\,\mathbb{P}(\xi_{n,k+1} = \alpha)}{\mathbb{P}(\xi_{n,k} = \beta)}. \tag{2.8}$$

By following the procedure in Theorem 2.9 and splitting block $B_i \in \beta$ into two blocks of size s and $b_i - s$, we obtain a formula for the left hand side:

$$\mathbb{P}\left(\xi_{n,k+1} = \alpha \mid \xi_{n,k} = \beta\right) = \frac{b_i - 1}{n-k} \cdot \frac{1}{b_i - 1} \cdot \frac{2}{\binom{b_i}{s}} = \frac{1}{n-k}\frac{2}{\binom{b_i}{s}},$$

where the factor of 2 in the numerator comes from the fact that the blocks are unordered. Finally, we can show that the right hand side of (2.8) is equal to this expression using (2.2) and (2.3). □

2.5 Tree topologies

Consider a sample path of the embedded jump chain $\{\xi_{n,k}, k = n, \ldots, 1\}$:

$$\{\{1\}, \ldots, \{n\}\} = \alpha_n \prec \alpha_{n-1} \prec \cdots \prec \alpha_1 = \{\{1, \ldots, n\}\}.$$

Since $\{\xi_{n,k}, k = n, \ldots, 1\}$ is a Markov chain, the probability of this sample path is

$$\mathbb{P}(\xi_{n,n} = \alpha_n, \ldots, \xi_{n,1} = \alpha_1) = \prod_{k=2}^{n} \mathbb{P}(\xi_{n,k-1} = \alpha_{k-1} \mid \xi_{n,k} = \alpha_k) = \prod_{k=2}^{n} \frac{1}{\binom{k}{2}} = \frac{2^{n-1}}{n!(n-1)!}. \tag{2.9}$$

So the jump chain has a uniform distribution over all its valid sample paths. We can represent these sample paths with a labeled binary tree, where each interior vertex is labeled by the jump time (where time is run backwards from $n, n-1, \ldots, 1$). We call a binary tree labeled in this way a *coalescent topology:*

Definition 2.10 (Coalescent tree topologies). A coalescent topology is a binary tree with unweighted edges and with a strict linear ordering on the heights of the interior vertices. By convention, we label the highest vertex (the root) as 1, the second highest as 2, and so forth. See Figure 2.4 for an illustration. We denote the set of all coalescent topologies with leaves labeled by $[n]$ as $\mathscr{T}_n^c$.

We now provide alternate proofs of Theorems 2.7 and 2.8 based on counting coalescent tree topologies.

Proof (Theorem 2.7). Consider the partition induced by cutting the tree when there are two lineages. The subsample has a different MRCA from the entire sample if and only if it lies within one of the two blocks. So we need to compute the probability of the subsample lying entirely in one block.

Since Kingman's coalescent generates coalescent topologies following the uniform distribution, we only need to count the number of coalescent topologies where the subsample lies

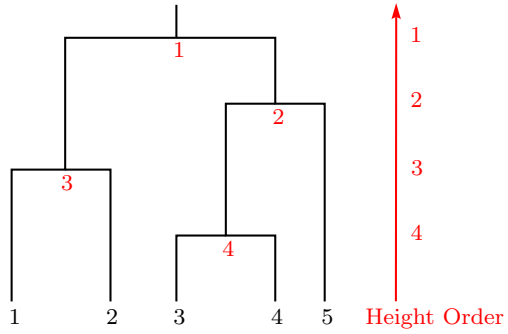


Fig. 2.4: An example of a coalescent topology.

in a single block. To do this, we can partition the sample into two blocks, one of which contains the subsample, put a coalescent topology on each block, and then combine them into a single topology by ordering how coalescent events hit the two blocks.

So fix a partition of two blocks, one of which has size k and contains the subsample. There are $\binom{n-m}{k-m}$ ways to do this, since we must choose $k-m$ elements outside the subsample to be in its block. Next fix a coalescent topology on each block; there are $\frac{k!(k-1)!}{2^{k-1}}\frac{(n-k)!(n-k-1)!}{2^{n-k-1}}$ ways to do this. Finally, note there are $\binom{n-2}{k-1}$ ways to put these two topologies into a single topology, for there are $n-2$ coalescent events before there are two lineages, $k-1$ of which occur on the block with the subsample. Summing over k, we get that the probability of getting a different MRCA for the subsample is

$$\begin{aligned}&\sum_{k=m}^{n-1}\frac{2^{n-1}}{n!(n-1)!}\frac{k!(k-1)!(n-k)!(n-k-1)!}{2^{n-2}}\frac{(n-m)!}{(k-m)!(n-k)!}\frac{(n-2)!}{(k-1)!(n-k-1)!}\\ =&\frac{2}{n-1}\binom{n}{m}^{-1}\sum_{k=m}^{n-1}\binom{k}{m}=\frac{2}{n-1}\binom{n}{m}^{-1}\binom{n}{m+1}=\frac{2(n-m)}{(n-1)(m+1)}\end{aligned}$$

which follows from repeated applications of the identity $\binom{m}{k}=\binom{m-1}{k}+\binom{m-1}{k-1}$. Hence the probability that the MRCA of the subsample and the full sample are the same is

$$1-\frac{2(n-m)}{(n-1)(m+1)}=\frac{(m-1)(n+1)}{(m+1)(n-1)},$$

which proves the claim. □

Proof (Theorem 2.8). We count the coalescent topologies in $E_{n,m}$ by first fixing two topologies, one on the subsample and one on its complement, and then counting the ways to combine them into a single topology. To do this, we choose times for the coalescent events to hit the subsample, we choose a time for the subsample to coalesce with the full sample, and we choose a lineage for the subsample to coalesce with.

There are $\frac{m!(m-1)!}{2^{m-1}}\frac{(n-m)!(n-m-1)!}{2^{n-m-1}}$ ways to choose topologies on the subsample and its complement. Fixing the time i when subsample coalesces with its complement, there are $\binom{i-1}{m-1}$ ways to choose times for coalescences to hit the subsample, and $n-i$ lineages for the subsample to coalesce with. Hence the number of topologies in $E_{n,m}$ is

$$\begin{aligned}
&\frac{m!(m-1)!(n-m)!(n-m-1)!}{2^{n-2}} \sum_{i=m}^{n-1} \frac{(i-1)!}{(m-1)!(i-m)!}(n-i) \\
&= \frac{m!(m-1)!(n-m)!(n-m-1)!}{2^{n-2}} \left[n\binom{n-1}{m} - m\binom{n}{m+1} \right] \\
&= \frac{m!(m-1)!(n-m)!(n-m-1)!}{2^{n-2}} \frac{n!}{m!(n-1-m)!} \left[1 - \frac{m}{m+1} \right] \\
&= \frac{(m-1)!(n-m)!n!}{2^{n-2}(m+1)}.
\end{aligned}$$

Multiplying each topology by $\frac{2^{n-1}}{n!(n-1)!}$ yields a probability of

$$\frac{2(m-1)!(n-m)!}{(n-1)!(m+1)} = \frac{2}{m+1} \frac{1}{\binom{n-1}{m-1}}.$$

To show the second part of the theorem, we count coalescent topologies in $E_{n,m}$ that start with a coalescence on the subsample. There are $\binom{m}{2}$ ways to pick a pair from the subsample to coalesce. After doing this, there are

$$\frac{(m-2)!(n-m)!(n-1)!}{2^{n-3}m}$$

ways to draw a coalescent topology on the configuration after the first coalescence, such that the lineages of the subsample form a subtree. Hence the conditional probability is

$$\binom{m}{2} \frac{(m-2)!(n-m)!(n-1)!}{2^{n-3}m} \left[\frac{(m-1)!(n-m)!n!}{2^{n-2}(m+1)} \right]^{-1} = \frac{m+1}{n},$$

which completes the proof. □

For some problems, we may not care about the ordering of the interior vertices apart from ancestral relationships. So in addition to coalescent topologies, we will be interested in another type of tree topology, obtained by erasing the labels of the interior vertices:

Definition 2.11 (Rooted binary tree topologies). A rooted topology is a binary tree with unweighted edges and a root vertex, but with no ordering on two interior vertices when neither is an ancestor of the other. We denote the set of all rooted topologies with leaves labeled by $[n]$ as $\mathscr{T}_n^r$.

The number of coalescent and rooted topologies are given by the following formulas:

$$\begin{aligned}
|\mathscr{T}_n^c| &= \frac{n!(n-1)!}{2^{n-1}} \\
|\mathscr{T}_n^r| &= (2n-3)!! = (2n-3)(2n-5)\cdots 5 \cdot 3 \cdot 1.
\end{aligned}$$

So for $n \geq 4$, we have the strict inequality $|\mathscr{T}_n^r| < |\mathscr{T}_n^c|$. In fact, it is not hard to compute the number of coalescent topologies that are consistent with a given rooted topology:

Theorem 2.12. *Call a coalescent topology $\tau^c \in \mathscr{T}_n^c$ consistent with a rooted topology $\tau^r \in \mathscr{T}_n^r$ if τ^r is obtained from τ^c by ignoring the order of the interior vertices when there is no*

ancestral relationship between them. For any given τ^r, the number of coalescent topologies consistent with τ^r is

$$(n-1)! \prod_{v \in \mathring{V}(\tau^r)} \frac{1}{l(v)-1},$$

where $\mathring{V}(\tau^r)$ denotes the set of interior vertices of τ^r, and $l(v)$ denotes the number of descendant leaves of $v \in \mathring{V}(\tau^r)$.

Proof. Let $\tau^r \in \mathscr{T}_n^r$ denote a rooted topology with n leaves. Then, τ^r has $(n-1)$ interior vertices. Given $n-1$ objects, there are $(n-1)!$ ways to linearly order them. In coalescent topologies, a linear ordering of interior vertices is constrained as follows: For $v \in V^\circ(\tau^r)$, let $D(v)$ denote the set of interior vertices (not including v) that are descendants of v. Since $|D(v)| = \ell(v) - 2$, for a fixed linear ordering of the elements in $D(v)$, there are $\ell(v) - 1$ ways to extend the linear ordering to $D(v) \cup \{v\}$, but exactly one of those extensions is realizable in coalescent topologies. Hence, the number of coalescent topologies that have the same shape and leaf label assignment as τ^r is

$$(n-1)! \prod_{v \in \mathring{V}(\tau^r)} \frac{1}{l(v)-1}.$$

Alternatively, one can also prove the result using induction. Consider the left subtree τ_L and the right subtree τ_R attached to the root of τ^r, and invoke induction hypothesis on those subtrees. Then, given a linear ordering of the interior vertices in each subtree, consider the number of ways to linearly order the interior vertices in τ_L relative to those in τ_R. Multiplying the above factors gives the desired result. □

This leads to the following corollary:

Corollary 2.13. *Under the n-coalescent, the probability of generating a particular rooted topology $\tau \in \mathscr{T}_n^r$ is given by*

$$\frac{2^{n-1}}{n!} \prod_{v \in \mathring{V}(\tau^r)} \frac{1}{l(v)-1}.$$

Proof. Follows from (2.9) and Theorem 2.12. □

2.6 The Yule-Harding process

The Yule-Harding process is a forward-in-time process that generates coalescent topologies in $\mathscr{T}_n^c$ with the same law as the n-coalescent. As an application, we will use the Yule-Harding process to give a simple proof of Theorem 2.5 via an urn model.

Coalescent topologies have a clearer connection to the urn model than rooted topologies, hence we will show that the Yule-Harding process induces the correct law on $\mathscr{T}_n^c$.

Definition 2.14 (Yule-Harding process). The Yule-Harding process is a forward-in-time Markov chain with index set $\{1, \ldots, n\}$ and state space consisting of trees in which leaves are labeled by subsets of $[n]$. Denoting by $\mathscr{Y}_k$ the state of the chain at time k, we generate a realization of the Yule-Harding process as follows:

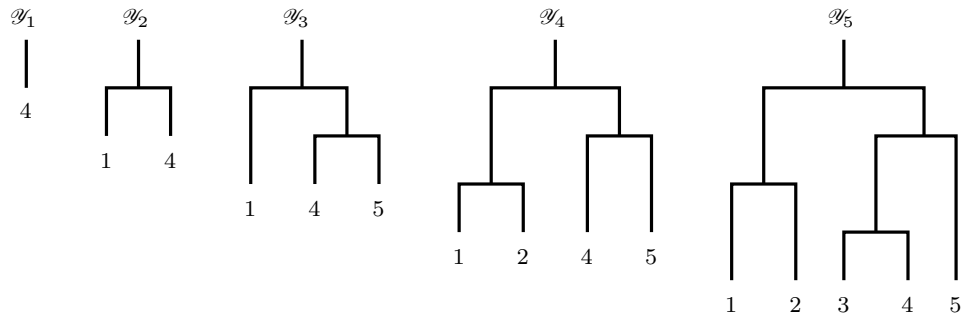


Fig. 2.5: A sample path in the Yule-Harding model, when $n = 5$. If we keep track of when the lineages branch, the resulting coalescent topology is the same as in Figure 2.4.

1. $\mathscr{Y}_1$ consists of a single lineage, labeled uniformly at random by an element of the sample $[n]$.
2. To generate $\mathscr{Y}_{k+1}$ from $\mathscr{Y}_k$, select a leaf lineage uniformly at random from $\mathscr{Y}_k$ and split it into two. Label the new lineage uniformly at random with an unused element of $[n]$.

To make this definition clear, we illustrate a realization of the Yule-Harding process in Figure 2.5. The usefulness of the Yule-Harding model is a result of the following theorem:

Theorem 2.15. *Generate a coalescent topology from the Yule-Harding process by keeping track of the time indices when interior vertices split but ignoring the labels of interior lineages. Then the Yule-Harding process induces the same law on $\mathscr{T}_n^c$ as the n-coalescent.*

Proof. The single lineage in $\mathscr{Y}_1$ is labeled by a particular element of $[n]$ with probability $1/n$. At time $k+1$, we select a lineage to split with probability $1/k$, then select an unused lineage to add with probability $1/(n-k)$. Hence, each sample path in the Yule-Harding process has probability

$$\frac{1}{n}\prod_{k=1}^{n-1}\frac{1}{k(n-k)} = \frac{1}{n!(n-1)!}.$$

We can represent a sample path in the Yule-Harding process by a binary tree, where the interior vertices are ordered by the time in which they split, and where each edge is labeled by the lineage in $[n]$ that it corresponds to (note that we place an edge above the root vertex as well). See Figure 2.6 for an illustration. Observe that the parent edge of a leaf must have the same label as the leaf, while the parent edge of an interior vertex must have the same label as one of its child edges.

We obtain a coalescent topology from a sample path by ignoring the labels of the edges. So we count how many sample paths correspond to a coalescent topology. Working up from the bottom of the tree, there are two choices when labeling the parent edge of an interior vertex. Since there are $n-1$ interior vertices, there are 2^{n-1} sample paths that correspond to a single coalescent topology.

Hence the Yule-Harding process generates each coalescent topology with probability $\frac{2^{n-1}}{n!(n-1)!}$, i.e. it is uniform over $\mathscr{T}_n^c$. □

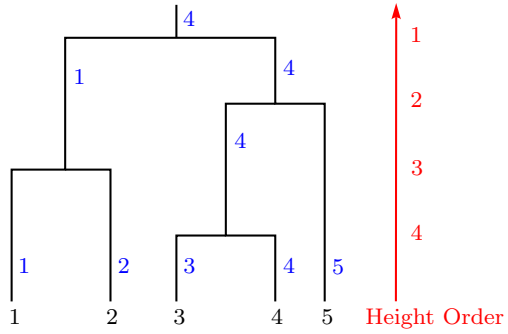


Fig. 2.6: The sample path in Figure 2.5, but represented as a single binary tree with labeled edges and an ordering on the interior vertices by branching time (height).

2.7 Urn models with stochastic replacement

At each time step in the Yule-Harding process, we select a lineage uniformly at random and connect it to an unused leaf lineage. We can model this as a balls-and-urn process, where we repeatedly sample a ball from an urn and return it with another ball. Here, each ball in the urn corresponds to a lineage that has been attached to the tree, whereas the ball we add corresponds to the unused leaf from $[n]$ that we connect to the tree. This perspective will give us an easy proof of (2.4) in Theorem 2.5, i.e., that cutting the n-coalescent at level k generates a composition with k parts uniformly at random.

We begin by stating some facts about urn models, and then elaborate on their connection with the Yule-Harding process and compositions. While urn models are traditionally attributed to Eggenberger and Polya (1923), all the results we will be using were obtained by Markov (1917).

Theorem 2.16. *Consider an urn with c_j balls of color j, where $j \in \{1, \ldots, k\}$. Sample a ball from the urn, and return it with s copies of the same color. Repeat this m times, and let X_j be the number of times color j was drawn. Then,*

$$\mathbb{P}(X_1 = i_1, \ldots, X_k = i_k) = \binom{m}{i_1, \ldots, i_k} \frac{\prod_{j=1}^{k} (\frac{c_j}{s})_{i_j\uparrow}}{(\frac{c_1+\cdots+c_k}{s})_{m\uparrow}}$$

when $i_1, \ldots, i_k$ are non-negative integers that sum to m.

Proof. Let Y_i be the color of the ith ball drawn from the urn. We have that

$$\mathbb{P}(Y_1 = y_1, \ldots, Y_m = y_m) = \mathbb{P}(Y_1 = y_1) \prod_{i=2}^{m} \mathbb{P}(Y_i = y_i \mid Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}). \quad (2.10)$$

Now suppose that the color y_i appears $(l-1)$ times in $(y_1, \ldots, y_{i-1})$. Then

$$\mathbb{P}(Y_i = y_i \mid Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}) = \frac{c_{y_i} + (l-1)s}{\sum_{j=1}^{k} c_j + (i-1)s},$$

since there are $\sum_j c_j + (i-1)s$ balls in the urn at time i, of which $c_{y_i} + (l-1)s$ have color y_i. Hence the ith draw contributes

$$\sum_{j=1}^{k} c_j + (i-1)s = \left[\frac{\sum_{j=1}^{k} c_j}{s} + (i-1) \right] s$$

to the denominator of the product (2.10), while the lth draw of color j contributes

$$c_j + (l-1)s = \left[\frac{c_j}{s} + (l-1) \right] s$$

to the numerator of that product. Hence, we get

$$\mathbb{P}(Y_1 = y_1, \ldots, Y_m = y_m) = \frac{\prod_{j=1}^{k} (\frac{c_j}{s})_{i_j\uparrow}}{(\frac{c_1+\cdots+c_k}{s})_{m\uparrow}},$$

where $i_j = \sum_{l=1}^{m} \mathbb{I}(y_l = j)$ denotes the total number of times that color j appears in $(y_1, \ldots, y_m)$. Since there are $\binom{m}{i_1,\ldots,i_k}$ sequences where color j is selected exactly i_j times, we get the desired result. □

We now use the above urn model to provide an alternate proof of Theorem 2.5.

Proof (Theorem 2.5). Consider the time when there are k lineages in the Yule-Harding process. Give each lineage a unique color, then run the Yule-Harding process as normal; however, whenever we add a new lineage to the tree, give it the color of the lineage we are attaching it to. Then the total number of leaves of color j will follow the urn model above, with $s = 1$, $c_j = 1$ for all $j \in \{1, \ldots, k\}$, and $m = n - k$. Therefore, letting Z_j denote the number of descendant leaves of the jth lineage, Theorem 2.16 implies

$$\begin{aligned}
\mathbb{P}(Z_1 = i_1, \ldots, Z_k = i_k) &= \binom{n-k}{i_1, \ldots, i_k} \frac{\prod_{j=1}^{k} i_j!}{(k)_{(n-k)\uparrow}} \\
&= \frac{(n-k)!}{(k)_{(n-k)\uparrow}} \\
&= \frac{(n-k)!(k-1)!}{(n-1)!} = \frac{1}{\binom{n-1}{k-1}},
\end{aligned}$$

which agrees with (2.4). So the law of $(Z_1, \ldots, Z_k)$ is uniform over all compositions of size n with k parts. □

2.8 Sufficient conditions for weak convergence to the n-coalescent

In Chapter 1.2, we considered a popular discrete-time random mating model, namely the Wright-Fisher model, and saw that probability computation under the model can be rather cumbersome. We therefore sought a continuous-time model that facilitates computation while providing an accurate approximation to the Wright-Fisher model when the population size is large. In this section, we extend this concept to a more general class of random

mating models called *Cannings exchangeable models*. For this class of models, we will prove the celebrated result of Kingman which establishes a set of sufficient conditions for the genealogical process associated with an exchangeable random mating model to converge in distribution to the n-coalescent defined in Chapter 2.1. We begin with some definitions needed to state the main theorem.

Definition 2.17 (Exchangeability). Let $\mathscr{S}_k$ denote the symmetric group on $[k]$. A k-tuple $(X_1, X_2, \ldots, X_k)$ of random variables is said to be *exchangeable* if for all permutations $\pi \in \mathscr{S}_k$, $(X_1, X_2, \ldots, X_k)$ has the same distribution as $(X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(k)})$.

Definition 2.18 (Cannings exchangeable models). Denote by $\nu_k(\tau)$ the number of offspring born to individual $k \in [N]$ at time τ. Cannings exchangeable models (Cannings, 1974) are a class of random mating models with the following properties:

1. Every individual survives for exactly one generation. Time, denoted by τ, is measured in generations.
2. For all generations, population size remains constant at N, i.e., $\nu_1(\tau) + \nu_2(\tau) + \cdots + \nu_N(\tau) = N$ for all τ. Individuals are labeled $1, 2, \ldots, N$.
3. For each generation τ, the N-tuple $(\nu_1(\tau), \nu_2(\tau), \ldots, \nu_N(\tau))$ is exchangeable.
4. $\{(\nu_1(\tau), \nu_2(\tau), \ldots, \nu_N(\tau))\}_{\tau \in \mathbb{Z}_{\geq 0}}$ are independently and identically distributed random vectors. (For a fixed time τ, the $\nu_i(\tau)$ are not independent of one another, since they sum to N.)

Remark 2.19. We note the following simple observations:

1. $\mathbb{P}(\nu_i(\tau) = m) = \mathbb{P}(\nu_j(\tau) = m)$ for all $i, j \in [N]$ and for all $m \in \{0, \ldots, N\}$. This follows directly from exchangeability.
2. Exchangeability and $\nu_1(\tau) + \cdots + \nu_N(\tau) = N$ together imply $\mathbb{E}(\nu_i(\tau)) = 1$ for all $i \in [N]$.
3. For all generations τ and index $i \in [N]$, $\mathrm{Var}(\nu_i(\tau)) = \sigma_N^2$, which may depend on N but not on i.

In what follows, we drop the dependence on τ when writing offspring numbers, which is allowed since $\{(\nu_1(\tau), \nu_2(\tau), \ldots, \nu_N(\tau))\}_{\tau \in \mathbb{Z}_{\geq 0}}$ are i.i.d. random vectors. A well-known example of Cannings exchangeable models is the Wright-Fisher model, in which the offspring numbers $(\nu_1, \nu_2, \ldots, \nu_N)$ have a multinomial distribution in each generation, i.e.,

$$\mathbb{P}(\nu_1 = m_1, \ldots, \nu_N = m_N) = \binom{N}{m_1, \ldots, m_N} \frac{1}{N^N}.$$

Noting that ν_i are marginally binomial with success probability $1/N$, one can easily check that $\mathbb{E}(\nu_i) = 1$ and $\mathrm{Var}(\nu_i) = 1 - \frac{1}{N}$, which do not depend on i.

We now present the aforementioned weak convergence result due to Kingman.

Theorem 2.20 (Kingman 1982c). *In a Cannings exchangeable model with population size N, let $\{D_n^N(\tau), \tau = 0, 1, \ldots\}$ denote the $\mathcal{P}_{[n]}$-valued backward-in-time Markov process describing the genealogy of a sample of size n. (Take a sample of size n from the population at time 0 and trace its genealogy backwards in time.) For all $t \in \mathbb{R}_+$, $D_n^N\left(\lfloor tN/\sigma^2 \rfloor\right)$ converges weakly to $C_n(t)$ (the n-coalescent at time t) as $N \to \infty$, if the following conditions hold:*

1. $\mathrm{Var}(\nu_1) = \sigma_N^2 \to \sigma^2 \in (0, \infty)$ *as $N \to \infty$. (Note that σ^2 is strictly positive.)*

2. $\sup_N \mathbb{E}(\nu_1^k) < \infty$*, for all* $k = 3, 4, \ldots$

Proof. Suppose $D_n^N(\tau) = \alpha$ and $D_n^N(\tau+1) = \beta$, where $\alpha = (A_1, \ldots, A_k) \in \mathcal{P}_{[n]}$ and $\beta = (B_1, \ldots, B_l) \in \mathcal{P}_{[n]}$. Let $p_{\alpha\beta}$ denote the one-step transition probability of $D_n^N(\cdot)$ and define $P_N = (p_{\alpha\beta})_{\alpha,\beta\in\mathcal{P}_{[n]}}$. Our goal is to show that

$$P_N = I + \frac{\sigma^2}{N} Q + o\left(\frac{1}{N}\right), \tag{2.11}$$

where $Q = (q_{\alpha\beta})_{\alpha,\beta\in\mathcal{P}_{[n]}}$ is the infinitesimal generator of the n-coalescent (2.1). If this holds, then $\lim_{N\to\infty} P_N^{\lfloor Nt/\sigma^2 \rfloor} = e^{Qt}$, for all $t \in \mathbb{R}_+$, thus implying the desired weak convergence result.

For $1 \le i \le l$, suppose block B_i contains $k_i > 0$ blocks in α, such that $k = \sum_{i=1}^l k_i$. (If this is not the case, then $p_{\alpha\beta} = 0$.) Let $j_1, \ldots, j_l \in [N]$ be l distinct individuals at time $\tau+1$, and $\nu_{j_1}, \ldots, \nu_{j_l}$ their corresponding offspring numbers. Then, $p_{\alpha\beta}$ is given by

$$p_{\alpha\beta} = \mathbb{E}\left[\sum_{j_1,\ldots,j_l\in[N],\text{all distinct}} \frac{(\nu_{j_1})_{k_1\downarrow}\cdots(\nu_{j_l})_{k_l\downarrow}}{(N)_{k\downarrow}}\right], \tag{2.12}$$

where the expectation is taken over the distribution of $\nu_{j_1}, \ldots, \nu_{j_l}$.

We want to show that if it requires more than a single pairwise merger to get from α to β, then $p_{\alpha\beta}$ decays faster than $1/N$. Indeed, for $k \ge 3$ and $l < k-1$, we have

$$\begin{aligned}
p_{\alpha\beta} &\le \mathbb{E}\left[\sum_{j_1,\ldots,j_l\in[N],\text{all distinct}} \frac{\nu_{j_1}^{k_1}\cdots\nu_{j_l}^{k_l}}{(N)_{k\downarrow}}\right] \\
&\le \sum_{j_1,\ldots,j_l\in[N],\text{all distinct}} \frac{1}{(N)_{k\downarrow}} [\mathbb{E}(\nu_{j_1}^k)]^{\frac{k_1}{k}} \cdots [\mathbb{E}(\nu_{j_l}^k)]^{\frac{k_l}{k}} \\
&= \sum_{j_1,\ldots,j_l\in[N],\text{all distinct}} \frac{1}{(N)_{k\downarrow}} \mathbb{E}(\nu_1^k) \\
&= \frac{(N)_{l\downarrow}}{(N)_{k\downarrow}} \mathbb{E}(\nu_1^k) \\
&= o\left(\frac{1}{N}\right),
\end{aligned}$$

where the second inequality follows from Hölder's inequality, the third line follows from exchangeability, and the last line follows from $l < k-1$ and the second condition of the theorem.

Now suppose $\alpha \prec \beta$; i.e., transitioning from α to β involves exactly a single pairwise merger. In this case, $l = k-1$. Without loss of generality, assume that two blocks in α find individual j_1 in generation $\tau+1$ as a common parent. Then, setting $(k_1, k_2, \ldots, k_l) = (2, 1, \ldots, 1)$ and $l = k-1$ in (2.12), we get

$$p_{\alpha\beta} = \mathbb{E}\left[\sum_{j_1,\ldots,j_{k-1}\in[N],\text{all distinct}} \frac{\nu_{j_1}(\nu_{j_1}-1)\nu_{j_2}\cdots\nu_{j_{k-1}}}{(N)_{k\downarrow}}\right].$$

By removing the restriction that $j_1,\ldots,j_{k-1}$ be distinct, we can upper bound the above expression as

$$p_{\alpha\beta} \le \mathbb{E}\left[\sum_{j=1}^{N} \frac{\nu_j(\nu_j-1)N^{k-2}}{(N)_{k\downarrow}}\right] = \frac{N^{k-2}}{(N)_{k\downarrow}} N\,\mathbb{E}(\nu_1(\nu_1-1)) = \frac{N^{k-2}}{(N)_{k\downarrow}} N\sigma_N^2 = \frac{\sigma^2}{N} + o\left(\frac{1}{N}\right),$$

where the first equality follows from exchangeability, and the difference between σ^2 and σ_N^2 has been absorbed into $o(1/N)$ in the last equality.

To put a lower bound on $p_{\alpha\beta}$ when $\alpha \prec \beta$, first note

$$p_{\alpha\beta} \ge \mathbb{E}\left\{\frac{1}{(N)_{k\downarrow}}\sum_{j=1}^{N}\nu_j(\nu_j-1)\left[(N-\nu_j)^{k-2} - \binom{k-2}{2}\sum_{i\neq j}\nu_i^2(N-\nu_j)^{k-4}\right]\right\}. \quad (2.13)$$

To see why this is a lower bound, think of assigning the k blocks of α to the individuals at time τ. Recall $\alpha \prec \beta$, and suppose A_1 and A_2 are the two blocks of α that merge. Then, note that $\sum_{j_1,\ldots,j_{k-1}\text{ distinct}}\nu_{j_1}(\nu_{j_1}-1)\nu_{j_2}\cdots\nu_{j_{k-1}}$ corresponds to the number of ways of assigning A_1 and A_2 to two distinct descendants of some individual j in generation $\tau+1$, and assigning the remaining $k-2$ blocks of α to non-descendants of j such that no two of those blocks have the same parent. Now note that $(N-\nu_j)^{k-2}$ corresponds to the number of ways to map $A_3,\ldots,A_k$ to the non-descendants of j. From this, we want to subtract the number of maps that would lead to some of $A_3,\ldots,A_k$ being siblings; $\binom{k-2}{2}\sum_{i\neq j}\nu_i^2(N-\nu_j)^{k-4}$ is an upper bound on that number.

From (2.13), we obtain

$$\begin{aligned}
p_{\alpha\beta} &\ge \frac{1}{(N)_{k\downarrow}}\mathbb{E}\left\{\sum_{j=1}^{N}\nu_j(\nu_j-1)\left[(N-\nu_j)^{k-2} - \binom{k-2}{2}\sum_{i\neq j}\nu_i^2 N^{k-4}\right]\right\} \\
&\ge \frac{1}{N^k}\mathbb{E}\left\{\left[\sum_{j=1}^{N}\nu_j(\nu_j-1)N^{k-2}\right] - \left[(k-2)\sum_{j=1}^{N}\nu_j^3 N^{k-3}\right] - \binom{k-2}{2}\sum_{i\neq j}\nu_i^2\nu_j^2 N^{k-4}\right\} \\
&\ge \frac{1}{N}\mathbb{E}[\nu_1(\nu_1-1)] - \frac{(k-2)}{N^2}\mathbb{E}[\nu_1^3] - \binom{k-2}{2}\frac{1}{N^2}\mathbb{E}[\nu_1^2\nu_2^2] \\
&= \frac{\sigma^2}{N} + o\left(\frac{1}{N}\right),
\end{aligned}$$

where the second line follows from $\frac{1}{(N)_{k\downarrow}} > \frac{1}{N^k}$ and $(N-\nu_j)^{k-2} \ge N^{k-2} - (k-2)N^{k-3}\nu_j$ for N sufficiently large, and the last line follows from the Cauchy-Schwarz inequality $\mathbb{E}[\nu_1^2\nu_2^2] \le [\mathbb{E}(\nu_1^4)\mathbb{E}(\nu_2^4)]^{1/2}$ and the second condition of the theorem. Hence, for $\alpha \prec \beta$, since $p_{\alpha\beta}$ is bounded from both above and below by $\frac{\sigma^2}{N} + o\left(\frac{1}{N}\right)$, we conclude that

$$p_{\alpha\beta} = \frac{\sigma^2}{N} + o\left(\frac{1}{N}\right),$$

and the one-step transition matrix P_N has the desired form (2.11). □

Note that the conditions in Theorem 2.20 are sufficient, but not necessary. Necessary and sufficient conditions will be discussed in Chapter 2.10.

Consider a general Canning exchangeable model with population size N satisfying the conditions in Theorem 2.20. The convergence result discussed above implies that the genealogical process at time τ for such a model follows approximately the same law as the genealogical process at time τ for the Wright-Fisher model with population size N/σ^2, provided that N is sufficiently large. This motivates the following definition:

Definition 2.21 (Coalescent effective population size). Let σ^2 be defined as in Theorem 2.20. Then, the *coalescent effective population size* of the Canning exchangeable model is defined as

$$N_e = \frac{N}{\sigma^2}.$$

One unit of time in the n-coalescent corresponds to roughly N_e generations in the discrete-time model.

2.9 Moran models

Moran models are another kind of random mating model widely used in population genetics. Unlike the Wright-Fisher model, Moran models allow for overlapping generations. At any given point in time, at most one birth-death event occurs. This simplification often allows one to carry out exact computation. Furthermore, if one chooses a suitable birth-death rate in the model, the coalescent process can be obtained without the need of taking $N \to \infty$. There are three different ways to define a Moran model with population size N:

1. (Continuous-time) Each individual dies at rate $\frac{N}{2}$, and is replaced by a duplicate of another individual chosen uniformly at random from the current population. This model requires no rescaling of time for the genealogical process of a sample to converge to the coalescent.
2. (Continuous-time) Each individual dies at rate 1, and is replaced by a duplicate as in the above case. Weak convergence to the coalescent requires rescaling time by $\frac{N}{2}$.
3. (Discrete-time) In each generation, exactly one individual dies and is replaced by a duplicate of another individual chosen uniformly at random from the current generation. Every generation has exactly one birth-death event. The offspring number vector $(\nu_1, \ldots, \nu_N)$ is defined to have exactly one 0 (representing death) and one 2 (representing duplication); all other entries are 1s. Note that $\mathbb{E}[\nu_i] = 2 \cdot \frac{1}{N} + 1 \cdot \frac{N-2}{N} = 1$ and $\mathrm{Var}(\nu_i) = \mathbb{E}[\nu_i^2] - 1 = \frac{4}{N} + \frac{N-2}{N} - 1 = \frac{2}{N}$. In this model, one would need to rescale time by $\frac{N^2}{2}$ to converge to the coalescent.

See Figure 2.7 for an illustration of a continuous-time Moran model with population size $N = 6$. There are N parallel lines, each corresponding to an individual. The horizontal axis corresponds to time. An arrow from i to j indicates that i reproduces and j dies, with the newborn individual replacing j. The genealogical tree for a sample is obtained as follows:

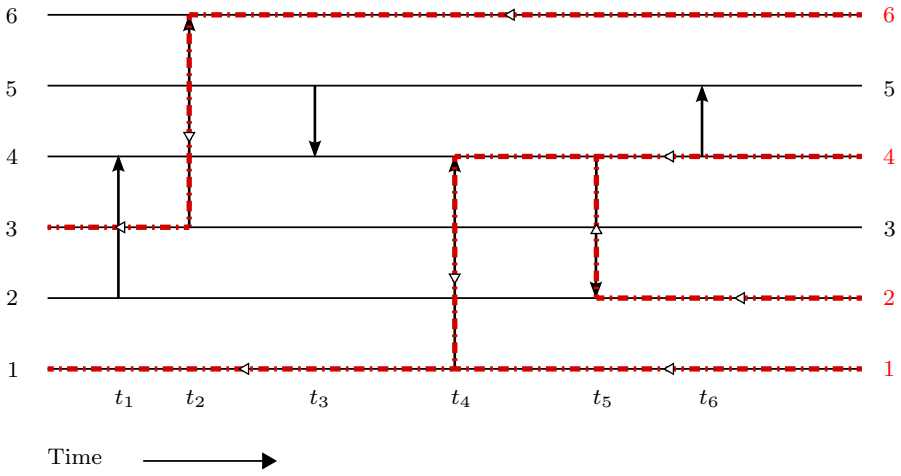


Fig. 2.7: Illustration of a continuous-time Moran model with $N = 6$.

1. For each individual i in the sample, start from the far right end and trace back the lineage labeled i.
2. When the head of an arrow is encountered (which means the individual is replaced at this time and its parent is specified by the tail of the arrow), follow the arrow and then continue on the line specified by the arrow's tail.

Figure 2.7 illustrates the genealogy for the sample $\{1, 2, 4, 6\}$. The partition structure at time t_3 is $\{\{6\}, \{1, 2, 4\}\}$. In the case of the first Moran model mentioned above, the rate of pairwise merger is $\binom{k}{2}$ while there are k ancestral lineages.

2.10 Necessary and sufficient conditions for weak convergence

We now present another milestone in the n-coalescent literature. Möhle and Sagitov (2001, 2003) established necessary and sufficient conditions for a process to converge in distribution to the n-coalescent. Their results are phrased in terms of two quantities c_N and d_N, which are defined below.

Definition 2.22 (2-merger probability). Let c_N denote the probability that two individuals chosen at random without replacement from the same generation have the same parent one generation back in time. For an exchangeable discrete-time random mating model with a constant population size N,

$$c_N = \sum_{i=1}^{N} \frac{\mathbb{E}(\nu_i(\nu_i - 1))}{N(N-1)} = \frac{\mathbb{E}(\nu_1(\nu_1 - 1))}{N-1}.$$

Definition 2.23 (3-merger probability). Let d_N denote the probability that three individuals chosen at random without replacement from the same generation have the same

parent one generation back in time. For an exchangeable discrete-time random mating model with a constant population size N,

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

Given a sample of size two from a population of size N, let X denote the waiting time until their most recent common ancestor is found. Then, $X \sim \text{Geometric}(c_N)$, which implies that c_N determines the right time scaling to obtain convergence to Kingman's coalescent. Further, for $t \in \mathbb{R}_+$, note that $\mathbb{P}(X > \lfloor t/c_N \rfloor) = (1-c_N)^{\lfloor t/c_N \rfloor} \to e^{-t}$ if $c_N \to 0$. Hence, we need $c_N \to 0$ as $N \to \infty$ for the waiting time to converge to an exponentially distributed random variable with rate 1, as in Kingman's coalescent. Another key property of Kingman's coalescent is that only pairwise mergers are allowed at any given time, so we want d_N to be negligibly small compared to c_N for large N. It turns out that the above two conditions are in fact necessary and sufficient for weak convergence:

Theorem 2.24 (Möhle and Sagitov 2001, 2003). *The discrete-time $\mathcal{P}_{[n]}$-valued process $\left\{D_n^N\left(\lfloor t/c_N \rfloor\right), t \geq 0\right\}$ converges weakly to the n-coalescent $\left\{C_n\left(t\right), t \geq 0\right\}$ as $N \to \infty$ if and only if the following conditions hold:*

1. $\lim_{N\to\infty} c_N = 0.$
2. $\lim_{N\to\infty} \frac{d_N}{c_N} = 0.$

Remark 2.25. We conclude with a few remarks on the above result:

1. The coalescent effective population for this general setting is $1/c_N$.
2. In the discrete-time Moran model, $\text{Var}(\nu_1) = 2/N \to 0$ as $N \to \infty$, so Theorem 2.20 cannot be applied to show weak convergence. However, it satisfies the conditions of Theorem 2.24. Specifically, $c_N = \text{Var}(\nu_1)/(N-1) = 2/[N(N-1)] \to 0$ as $N \to \infty$, while $d_N = 0$ for all N.

2.11 Coming down from infinity

The n-coalescent satisfies the following *consistency property*: Suppose $m < n$. The restriction of $\{C_n(t), t \geq 0\}$ to $[m]$ has the same law as the process $\{C_m(t), t \geq 0\}$. Moreover, there exists a unique $\mathcal{P}_{\mathbb{N}}$-valued Markov process $\{C_\infty(t), t \geq 0\}$, called Kingman's coalescent, such that for every $n \in \mathbb{N}$, the process $\{C_\infty(t), t \geq 0\}$ restricted to $[n]$ has the same law as $\{C_n(t), t \geq 0\}$. See Section 2.1 of Berestycki (2009) for further details.

A surprising property satisfied by Kingman's coalescent is that it "comes down from infinity." That is, although the number of blocks in $C_\infty(0)$ is infinite, after any finite amount of time $t > 0$, the number of blocks remaining in $C_\infty(t)$ is finite almost surely:

Theorem 2.26 (Coming down from infinity). *Let $A_\infty(t) = |C_\infty(t)|$ denote the ancestral process for $C_\infty(t)$. Then, $\mathbb{P}(A_\infty(t) < \infty \text{ for all } t > 0) = 1$.*

Proof. It suffices to show that for every $\epsilon > 0$, there exists a positive integer M, such that $\mathbb{P}(A_\infty(t) > M) \leq \epsilon$. Consider the restriction $C_n(t)$ of $C_\infty(t)$ to $[n]$ and let $A_n(t) = |C_n(t)|$. Then,

$$\begin{aligned}
\mathbb{P}(A_n(t) > M) &= \mathbb{P}\left(\sum_{k=M+1}^{n} T_{n,k} > t\right) \\
&\leq \frac{1}{t}\mathbb{E}\left(\sum_{k=M+1}^{n} T_{n,k}\right) \\
&= \frac{1}{t}\sum_{k=M+1}^{n} \frac{1}{\binom{k}{2}} \\
&\leq \frac{2}{tM},
\end{aligned}$$

where the second line follows from Markov's inequality and the third line follows from the fact that $T_{n,k} \sim \text{Exp}(\binom{k}{2})$ for a constant population size model. Now, setting $M \geq \frac{2}{t\epsilon}$, we conclude $\limsup_{n\to\infty} \mathbb{P}(A_n(t) > M) \leq \epsilon$, which implies the desired result. □

References

Berestycki N (2009) Recent progress in coalescent theory. Ensaios Matematicos 16(1):1–193

Cannings C (1974) The latent roots of certain markov chains arising in genetics: a new approach, I. haploid models. Advances in Applied Probability 6:260–290

Eggenberger F, Polya G (1923) Uber die statistik verketter vorgange. Zeitschrift fur Angewandte Mathematik und Mechanik 3.4:279–289

Kingman JFC (1982a) The coalescent. Stoch Process Appl 13:235–248

Kingman JFC (1982b) Exchangeability and the evolution of large populations. In: Koch G, Spizzichino F (eds) Exchangeability in Probability and Statistics, North-Holland Publishing Company, pp 97–112

Kingman JFC (1982c) On the genealogy of large populations. J Appl Prob 19A:27–43

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475:493–496

Markov A (1917) On limiting formulas for computing probabilities. Izvestiah Imperatovsky Akademy Nauk 6:177–186

Möhle M, Sagitov S (2001) A classification of coalescent processes for haploid exchangeable population models. The Annals of Probability 29(4):1547–1562

Möhle M, Sagitov S (2003) Coalescent patterns in diploid exchangeable population models. Journal of Mathematical Biology 47(4):337–352

Saunders IW, Tavaré S, Watterson G (1984) On the genealogy of nested subsamples from a haploid population. Advances in Applied probability pp 471–491

Part II
Neutral Mutations on Trees at Equilibrium

Chapter 3
Number of mutations

In this chapter we consider mutations in the coalescent process. In Figure 3.1, mutations $m_1, \ldots, m_4$ are marked by "×"s on the coalescent tree. In most biological applications, we do not know the true genealogical histories. Rather, we only observe genetic variation data at the leaves, and the objective is to make ancestral inference using the observed sequence data. Mutation is a major evolutionary mechanism responsible for generating genetic variation in a population.

Mutations change the allele type in a lineage, and exactly what kind of changes are introduced depends on the assumed model of mutation. In this chapter, we derive several probabilities that are independent of the details of the assumed mutation model. We first describe mutation in the discrete-time Wright-Fisher model and then generalize it to continuous time. We assume selective neutrality; that is, genetic types do not influence reproductive success.

Throughout this chapter, we assume that the population size remains constant over time.

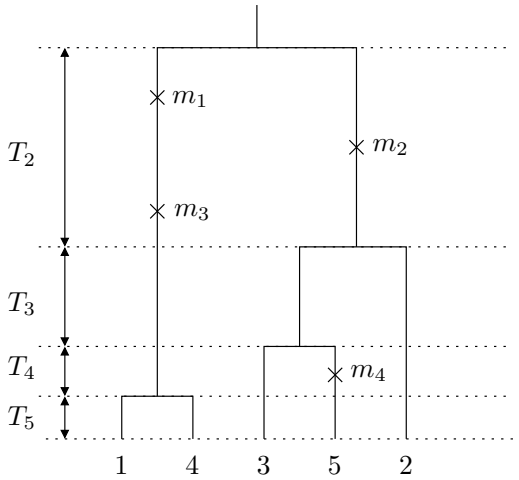


Fig. 3.1: A coalescent tree with mutations marked by "×"s and labeled $m_1, \ldots, m_4$. Mutations on each edge arise according to a Poisson point process with rate $\frac{\theta}{2}$, independently of all other edges

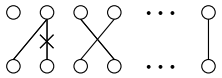


Fig. 3.2: Discrete-time Wright-Fisher model with mutations.

3.1 Mutations in a single lineage

In the discrete-time Canning exchangeable model, let T_{Mut} denote the number of generations until a mutation is encountered when the lineage of a single individual is traced back in time. Consider Figure 3.2, which illustrates two generations. The symbol "$\times$" represents a mutation that occurs when the second individual in the top row reproduces and creates the second individual in the bottom row. If the probability u of mutation per locus per generation per individual is constant over time, then the distribution of T_{Mut} satisfies

$$\mathbb{P}\big(T_{\text{Mut}} > k\,\big) = (1-u)^k.$$

Hence, for $t \in \mathbb{R}_{\geq 0}$,

$$\mathbb{P}\big(T_{\text{Mut}} > \lfloor t/c_N \rfloor\big) = (1-u)^{\lfloor t/c_N \rfloor} \to e^{-\frac{\theta}{2}t},$$

as $c_N \to 0$ and $u \to 0$ such that $\frac{u}{c_N} \to \frac{\theta}{2}$. (The factor of 2 in the denominator is conventionally introduced to simplify formulas in many quantities of interest.) Incidentally, u is typically very small ($u \approx 1.2 \times 10^{-8}$ for humans) while the population N is large, so assuming this limit is fine for most organisms. In the above limit, the waiting time to encountering a mutation is distributed as an exponential random variable with parameter $\theta/2$. This is equivalent to saying that mutations arrive according to a Poisson point process with intensity $\theta/2$, so

$$\mathbb{P}(m \text{ mutations on an edge of length } t) = \frac{1}{m!}\left(\frac{\theta}{2}t\right)^m e^{-\frac{\theta}{2}t}.$$

3.2 Number of mutations in a coalescent tree with n leaves

Generalizing the above result to the whole coalescent tree, we conclude that mutations on each edge arise according to a Poisson point process intensity rate $\frac{\theta}{2}$, independently of all other edges.

Definition 3.1. *($M_{n,k}$ and M_n).* Let $M_{n,k}$ denote the number of mutations while there are k lineages in a coalescent tree with n leaves. Let $M_n = M_{n,2} + \cdots + M_{n,n}$, the total number of mutations in the tree.

Proposition 3.2. *The probability that there are m mutations in a coalescent tree with n leaves is*

$$\mathbb{P}(M_n = m) = \sum_{k=2}^{n} (-1)^k \binom{n-1}{k-1} \frac{k-1}{\theta+k-1} \left(\frac{\theta}{\theta+k-1}\right)^m. \tag{3.1}$$

Proof. We will prove this result using two different methods. The first method will involve a direct computation of the probability, while the second approach will use the method of generating functions.

METHOD 1: Let $X_1 \sim \text{Poi}(\lambda_1), X_2 \sim \text{Poi}(\lambda_2)$ be independent Poisson random variables. Recall that the sum of two independent Poisson variables is itself a Poisson variable. More exactly,

$$X_1 + X_2 \sim \text{Poi}(\lambda_1 + \lambda_2).$$

This property implies that $M_n|T_2, \ldots, T_n \sim \text{Poi}\left(\frac{\theta}{2}L_n\right)$, where $L_n = \sum_{k=2}^{n} kT_k$ denotes the tree length. Thus

$$\mathbb{P}(M_n = m \mid L_n = t) = \frac{1}{m!}\left(\frac{\theta}{2}t\right)^m e^{-\frac{\theta}{2}t}.$$

The marginal distribution of M_n is then

$$\mathbb{P}(M_n = m) = \int_0^\infty \mathbb{P}(M_n = m \mid L_n = t) f_{L_n}(t) dt, \tag{3.2}$$

where $f_{L_n}(t)$ is the probability density function of L_n discussed in Lecture 2:

$$f_{L_n}(t) = \sum_{k=2} (-1)^k \binom{n-1}{k-1} \frac{k-1}{2} e^{-\frac{k-1}{2}t}.$$

METHOD 2: The second method uses the probability generating function $G_n(z)$ defined as follows.

$$G_n(z) = \sum_{m=0}^{\infty} \mathbb{P}(M_n = m) z^m = \sum_{m=0}^{\infty} \mathbb{E}\left[\frac{1}{m!}\left(\frac{\theta}{2}L_n\right)^m e^{-\frac{\theta}{2}L_n}\right] z^m.$$

For $z \geq 0$, the monotone convergence theorems imply that we can bring the summation inside the expectation, and we obtain

$$\begin{aligned}
G_n(z) &= \mathbb{E}\left[\sum_{m=0}^{\infty} \frac{1}{m!}\left(\frac{\theta}{2}L_n\right)^m e^{-\frac{\theta}{2}L_n} z^m\right] \\
&= \mathbb{E}\left[e^{(z-1)\frac{\theta}{2}L_n}\right] \\
&= \prod_{k=2}^{m} \mathbb{E}\left[e^{(z-1)\frac{\theta}{2}kT_k}\right] \\
&= \prod_{k=2}^{m} \int_0^\infty e^{(z-1)\frac{\theta}{2}t} \frac{k-1}{2} e^{-\frac{k-1}{2}t} dt \\
&= \prod_{k=2}^{m} \frac{1}{1 - \left(\frac{z-1}{k-1}\right)\theta}.
\end{aligned}$$

Lastly, we can obtain $\mathbb{P}(M_n = m)$ by finding the coefficient of z^m in $G_n(z)$. □

Remark 3.3. Although (3.1) is a nice closed-form result, numerical problems often arise when evaluating it. A numerically stable way of computing $\mathbb{P}(M_n = m)$ is to evaluate the integral in (3.2) numerically using the following form of the probability density of L_n:

$$f_{L_n}(t) = \frac{n-1}{2} e^{-\frac{t}{2}} \left(1 - e^{-\frac{t}{2}}\right)^{n-2}.$$

Proposition 3.4. *The probability mass function of $M_{n,k}$ is*

$$\mathbb{P}(M_{n,k} = m) = \frac{k-1}{\theta + k - 1} \left(\frac{\theta}{\theta + k - 1}\right)^m. \tag{3.3}$$

Proof. There are several ways to prove this result. We will mention three of them here. The first method will involve a direct computation, the second a generating function, and the third a coalescent argument.

Method 1: A direct computation of $\mathbb{P}(M_{n,k} = m)$ is

$$\begin{aligned}\mathbb{P}(M_{n,k} = m) &= \int_0^\infty \mathbb{P}(M_{n,k} = m \mid kT_k = t) f_{kT_k}(t)\, dt \\ &= \int_0^\infty \frac{1}{m!} \left(\frac{\theta}{2} t\right)^m e^{-\frac{\theta}{2}t} \left(\frac{k-1}{2}\right) e^{-\frac{k-1}{2}t}\, dt.\end{aligned}$$

Method 2: We can define a generating function as follows.

$$G_{n,k}(z) = \sum_{m=0}^\infty \mathbb{P}(M_{n,k} = m) z^m = \mathbb{E}\left[e^{(z-1)\frac{\theta}{2}kT_k}\right] = \frac{1}{1 - \left(\frac{z-1}{k-1}\right)\theta}.$$

We can then obtain $\mathbb{P}(M_{n,k} = m)$ by finding the coefficient of z^m in $G_{n,k}(z)$.

Method 3: Let $X_1 \sim \text{Exp}(\lambda_1)$, $X_2 \sim \text{Exp}(\lambda_2)$ be independent exponential random variables. Then $\mathbb{P}(X_1 < X_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2}$. While there are k lineages, the rate of mutation events is $k\frac{\theta}{2}$ and the rate of coalescence is $\binom{k}{2}$. Hence,

$$\begin{aligned}\mathbb{P}(\text{mutation} \mid \text{there are } k \text{ lineages}) &= \frac{\theta}{\theta + k - 1}, \\ \mathbb{P}(\text{coalescence} \mid \text{there are } k \text{ lineages}) &= \frac{k-1}{\theta + k - 1},\end{aligned}$$

and (3.3) immediately follows from these two probabilities. □

Note that the generating function discussed in Method 2 allows us to obtain the moments of $M_{n,k}$ straightforwardly:

$$\begin{aligned}\mathbb{E}(M_{n,k}) &= \frac{d}{dz} G_{n,k}(z)\Big|_{z=1} = \frac{\theta}{k-1}, \\ \text{Var}(M_{n,k}) &= \left[\frac{d}{dz} z \frac{d}{dz} G_{n,k}(z) - \left(\frac{d}{dz} G_{n,k}(z)\right)^2\right]_{z=1} = \frac{\theta}{k-1} + \left(\frac{\theta}{k-1}\right)^2.\end{aligned}$$

Further, from $\mathbb{E}(M_{n,k})$ and $\text{Var}(M_{n,k})$, we can compute $\mathbb{E}(M_n)$ and $\text{Var}(M_n)$ as follows:

$$\mathbb{E}(M_n) = \theta \sum_{j=1}^{n-1} \frac{1}{j},$$
$$\mathrm{Var}(M_n) = \sum_{k=2}^{n} \mathrm{Var}(M_{n,k}) = \sum_{j=1}^{n-1} \left[\frac{\theta}{j} + \left(\frac{\theta}{j} \right)^2 \right],$$

where the first equality in the second line follows from the independence of $M_{n,i}$ and $M_{n,j}$ for $i \neq j$. As $n \to \infty$, both $\mathbb{E}(M_n) \to \theta \log(n)$ and $\mathrm{Var}(M_n) \to \theta \log(n)$, up to some additive constants. Under the infinite-sites model of mutation which will be discussed later, $\mathbb{E}(M_n)$ is equal to the expected number of *segregating* (or polymorphic) sites.

3.3 Waiting times conditioned on the number of mutations

The waiting time T_k while k lineages is distributed as $T_k \sim \mathrm{Exp}\left[\binom{k}{2}\right]$. The following result establishes the conditional probability density of T_k given that there are m mutations while k lineages:

Proposition 3.5. *The conditional probability density function of T_k given $M_{n,k} = m$ is*

$$f_{T_k}(t|M_{n,k} = m) = \frac{1}{m!} \left[\frac{k}{2}(\theta + k - 1) \right]^{m+1} t^m e^{-\frac{k}{2}(\theta+k-1)t}. \tag{3.4}$$

Proof. Applying Bayes' rule, the conditional probability density function of T_k given $M_{n,k} = m$ can be written as

$$f_{T_k}(t|M_{n,k} = m) = \frac{\mathbb{P}(M_{n,k} = m \mid T_k = t)}{\mathbb{P}(M_{n,k} = m)} f_{T_k}(t).$$

Because mutations for each lineage occur according to a Poisson point process with rate $\theta/2$,

$$\mathbb{P}(M_{n,k} = m \mid T_k = t) = \frac{1}{m!} \left(\frac{\theta}{2} kt \right)^m e^{-\frac{\theta}{2}kt}.$$

We now have all the parts to compute the conditional probability density of T_k given $M_{n,k} = m$, and (3.4) follows after some algebra. □

In fact, Proposition 3.5 can be obtained in a much simpler way. While there are k lineages, the total rate of coalescence and mutation is $\frac{1}{2}k(\theta + k - 1)$. So, if there are m mutation events before coalescence, the total waiting time is a sum of $m+1$ i.i.d. exponential random variables each with rate $\frac{1}{2}k(\theta+k-1)$, which gives the gamma distribution with shape $m+1$ and rate $\frac{1}{2}k(\theta + k - 1)$. So,

$$T_k \mid (M_{n,k} = m) \sim \mathrm{Gamma}\Big(m + 1, \frac{1}{2}k(\theta + k - 1)\Big),$$

and indeed (3.4) is the density function of this gamma distribution.

Using (3.4), one can show that the conditional expectation of T_k given $M_{n,k} = m$ is

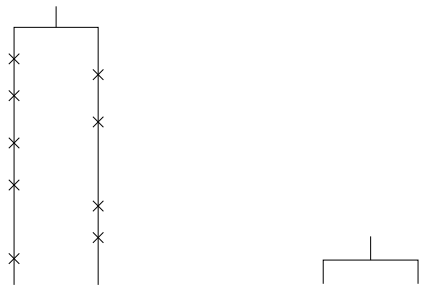


Fig. 3.3: Waiting time T_2 conditional of $M_2 = m$. (a) θ small and m large. (b) θ large and m small.

$$\mathbb{E}(T_k \mid M_{n,k} = m) = \frac{2(m+1)}{k(\theta + k - 1)}. \tag{3.5}$$

Compare this with the unconditional expectation of T_k:

$$\mathbb{E}(T_k) = \frac{2}{k(k-1)}.$$

Example 3.6. Consider the case of $n = k = 2$. Using $k = 2$ in (3.5) we obtain $\mathbb{E}(T_2 \mid M_2 = m) = \frac{m+1}{\theta+1}$, while $\mathbb{E}(T_2) = 1$. Suppose m is large, while θ is small. Then, $\mathbb{E}(T_2 \mid M_2 = m) > \mathbb{E}(T_2)$. This makes sense since, as illustrated in Figure 3.3a, a long coalescence time is expected for there to be many mutations when the mutation rate is low. On the other hand, suppose m is small, while θ is large. Then, this implies a short coalescence time, as illustrated in Figure 3.3b. For there to be few mutations when the mutation rate is high, a short coalescence is expected.

Example 3.7 (T_{MRCA} of human Y chromosomes). Dorit et al (1995) examined a region in the upstream of the Zinc-Finger (ZFY) locus on the Y-chromosome. No polymorphism was observed in the sample, which consisted of 729 base-pairs in 38 individuals.

Let W_n denote the waiting time until the most recent common ancestor. Recall that $W_n = T_n + \cdots + T_2$. In estimating the conditional expectation of W_n given $M_n = 0$, Dorit et al used an incorrect formula for $\mathbb{P}(M_n = 0 \mid W_n = t)$ and also incorrectly assumed that $f_{W_n}(t) = 1$. The correct answer can be computed as follows. Using (3.5), we obtain

$$\mathbb{E}(W_n \mid M_n = 0) = \sum_{k=2}^{n} \mathbb{E}(T_k \mid M_{n,k} = 0) = \sum_{k=2}^{n} \frac{2}{k(\theta + k - 1)}.$$

Suppose the long-term *effective* population size N_e^Y for Y-chromosomes is around $5,000$ (Harding et al, 1997; Harpending et al, 1998). To translate the coalescent time to years, the estimate from the above expression needs to be multiplied by N_e^Y (the number of generations per unit of coalescent time) and the average number G of years per generation. For the 729 bp region, $\theta = 2N_e^Y \times 729 \times 1.25 \times 10^{-8}$. Table 3.1 shows the expected time to the MRCA for different values of N_e and $G = 29$. For $N_e^Y = 5,000$, the expected W_n is about 252,884 years. Incidentally, using $u = 2.7 \times 10^{-8}$ and $G = 20$, Dorit et al (1995) estimated W_n to be

Table 3.1: T_{MRCA} estimates for different effective population sizes, assuming the $u = 1.25 \times 10^{-8}$ per individual per generation per bp and $G = 29$ years per generation.

N_e^Y	$N_e^Y \times G \times \mathbb{E}(W_n \mid M_n = 0)$, for $n = 38$
2,500	133,285 years
5,000	252,884 years
10,000	460,498 years

around 270,000 years, with a 95% confidence interval of (0, 800000). (There is not enough information in 729 bp, and hence the large confidence interval.)

References

Dorit R, Akashi H, Gilbert W (1995) Absence of polymorphism at the zfy locus on the human y chromosome. Science 268:1183–1185

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet 60:772–789

Harpending HC, Batze MA, Gurven M, Jorde LB, Rogers AR, Sherry ST (1998) Genetic traces of ancient demography. Proc Nat Acad Sci 95:1961–1967

Chapter 4
Infinite-alleles model and random combinatorial structures

In this chapter, we will discuss a classical mutation model called the *infinite-alleles* model (Kimura and Crow, 1964), in which each mutation gives rise to a new allele (genetic type) that has never been seen before. In addition to having had numerous applications in population genetics, this model turns out to have fascinating connections with random combinatorial structures. We start with a specific example of this connection. See Arratia et al (2003) for a more in-depth discussion of this topic, and Crane (2016) for a recent survey.

4.1 θ-biased random permutations

Consider the permutation group $\mathscr{S}_n$ on $[n]$.

Definition 4.1 (Cycle type). For a given $\sigma \in \mathscr{S}_n$, let c_i represent the number of cycles of length i in the cycle decomposition of σ. Then the n-tuple $\boldsymbol{c} = (c_1, \ldots, c_n)$ is called the *cycle type* of σ.

For example, if

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 6 & 8 & 3 & 5 & 7 & 1 & 4 & 2 \end{pmatrix} = (3)(16)(28)(457),$$

then $\boldsymbol{c}(\sigma) = (1, 2, 1, 0, 0, 0, 0, 0)$. Given an n-tuple $\boldsymbol{a} = (a_1, \ldots, a_n)$ of non-negative integers, we say that $\boldsymbol{a}$ partitions n, and write $\boldsymbol{a} \vdash n$, if $\sum_{i=1}^{n} i a_i = n$.

Theorem 4.2 (Cauchy). *For a given $\boldsymbol{a} \vdash n$, define $N(n, \boldsymbol{a}) \stackrel{\text{def}}{=} |\{\sigma \in \mathscr{S}_n : \boldsymbol{c}(\sigma) = \boldsymbol{a}\}|$, the number of permutations in $\mathscr{S}_n$ with cycle type $\boldsymbol{a}$. Then,*

$$N(n, \boldsymbol{a}) = n! \prod_{i=1}^{n} \left(\frac{1}{i}\right)^{a_i} \frac{1}{a_i!}. \tag{4.1}$$

Proof. This result can be derived as follows. Given a cycle type $\boldsymbol{a} \vdash n$, there are $n!$ ways to assign the elements of $[n]$ to distinct positions of the following cycle decomposition:

$$\overbrace{(\cdot), \ldots, (\cdot)}^{a_1}, \overbrace{(\cdot,\cdot), \ldots, (\cdot,\cdot)}^{a_2}, \overbrace{(\cdot,\cdot,\cdot), \ldots, (\cdot,\cdot,\cdot)}^{a_3}, \ldots.$$

Some of the $n!$ assignments lead to equivalent permutations. More precisely, for a given cycle, cyclic permutations of its entries do not change the cycle. For example, $(x, y, z), (y, z, x)$ and (z, x, y) are the same cycle. Hence, for each cycle of length i, a factor of $\frac{1}{i}$ comes from the fact that there are i equivalent ways to write down the cycle. Further, a factor of $\frac{1}{a_i!}$ comes from the $a_i!$ possible arrangements of the cycles of length i. □

Now suppose a permutation is randomly drawn from some distribution and let for $\boldsymbol{a} \vdash n$. If $\sigma \sim \text{Unif}\{\mathscr{S}_n\}$, then

$$\mathbb{P}(\boldsymbol{c}(\sigma) = \boldsymbol{a}) = \frac{N(n, \boldsymbol{a})}{n!} = \prod_{i=1}^{n} \left(\frac{1}{i}\right)^{a_i} \frac{1}{a_i!}.$$

In the θ-biased model of random permutation, for $\theta > 0$, the probability of selecting a permutation is proportional to $\theta^{\|\boldsymbol{a}\|_1}$, where $\|\boldsymbol{a}\|_1 \stackrel{\text{def}}{=} \sum_{i=1}^{n} a_i$ corresponds to the total number of cycles. Under this model,

$$\mathbb{P}_\theta(\boldsymbol{c}(\sigma) = \boldsymbol{a}) = \frac{(\theta)^{\|\boldsymbol{a}\|_1}}{(\theta)_{n\uparrow}} N(n, \boldsymbol{a}) = \frac{n!(\theta)^{\|\boldsymbol{a}\|_1}}{(\theta)_{n\uparrow}} \prod_{i=1}^{n} \left(\frac{1}{i}\right)^{a_i} \frac{1}{a_i!}. \tag{4.2}$$

When θ is large, permutations with many small cycles are more likely. Conversely, when θ is small, permutations with fewer, large cycles are more likely. Note that $a_1 \in \{0, 1, \ldots, n\}$, while $a_n \in \{0, 1\}$.

As we will see presently, a remarkable fact is that (4.2) is precisely equal to the sampling probability under the infinite-alleles model.

4.2 The infinite-alleles model and the Ewens sampling formula

As mentioned earlier, in the *infinite-alleles* model of mutation, each mutation gives rise to a new allele that has never been seen before. This can be modeled mathematically as follows. Every time a mutation occurs, we label it by a new random number drawn from $\text{Unif}[0, 1]$. We label each leaf by the label of the most recent mutation encountered when the ancestral lineage of the leaf is followed backward in time. See Figure 4.1 for an illustration. The leaf labels represent observed allelic types. Given two alleles, we can tell whether they are the same or not, but not how distantly they are related.

The configuration of a sample of size n is specified by an n-tuple $\boldsymbol{a} = (a_1, \ldots, a_n) \vdash n$, where a_i denotes the number of allelic types that appear exactly i times in the sample. For the example illustrated in Figure 4.1, $\boldsymbol{a} = (4, 1, 0, 0, 0, 0)$ since four alleles appear once and one appears twice.

Definition 4.3. Let $\boldsymbol{a} = (a_1, \ldots, a_n) \vdash n$ denote a sample configuration.

1. The sample size is denoted $|\boldsymbol{a}| = \sum_{i=1}^{n} i a_i = n$.
2. The number of distinct allele types observed is obtained by taking the 1-norm of $\boldsymbol{a}$:

$$\|\boldsymbol{a}\|_1 = \sum_{i=1}^{n} |a_i| = \sum_{i=1}^{n} a_i,$$

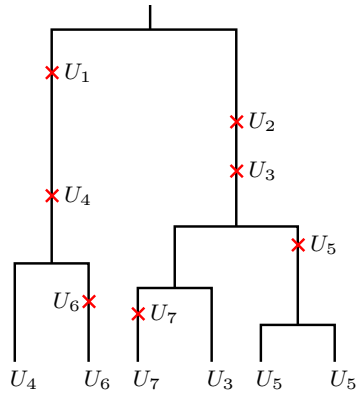


Fig. 4.1: A coalescent tree for a sample of size 6 under the infinite-alleles model of mutation. Each "×" indicates a mutation that gives rise to a new allele. Each leaf is assigned the label of the most recent mutation encountered when the ancestral lineage of the leaf is followed backward in time.

where the second equality follows from the fact that the a_i are non-negative.

3. For a sample of size n, the probability of observing a sample configuration $\boldsymbol{a}$ under the infinite-alleles model is denoted by p_n^θ. For all given $\theta > 0$ and $n \in \mathbb{N}$, the sampling probability distribution is normalized as

$$\sum_{\boldsymbol{a}:\boldsymbol{a}\vdash n} p_n^\theta(\boldsymbol{a}) = 1.$$

4. The unit n-vector consisting of a 1 in the ith position and a 0 elsewhere is denoted

$$\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0).$$

We aim to show that the formula (4.2) for the θ-biased random permutation model coincides with the sampling formula for the infinite-alleles model. First, we prove that the sampling formula p_n^θ satisfies the following recursion relation:

Theorem 4.4 (Probability recursion for an unordered sample). *Let $\boldsymbol{a} \vdash n$ be a sample configuration with sample size n. Then, under the infinite-alleles model, the sampling probability at stationary satisfies the following recursion:*

$$\begin{aligned} p_n^\theta(\boldsymbol{a}) = \frac{\theta}{n-1+\theta}\left[\frac{a_1}{n}p_n^\theta(\boldsymbol{a}) + \sum_{j=2}^{n}\frac{j(a_j+1)}{n}p_n^\theta(\boldsymbol{a}-\boldsymbol{e}_1-\boldsymbol{e}_{j-1}+\boldsymbol{e}_j)\right] \\ + \frac{n-1}{n-1+\theta}\sum_{j=1}^{n-1}\frac{j(a_j+1)}{n-1}p_{n-1}^\theta(\boldsymbol{a}-\boldsymbol{e}_{j+1}+\boldsymbol{e}_j), \end{aligned} \tag{4.3}$$

with boundary conditions $p_1^\theta(\boldsymbol{e}_1) = 1$ and $p_n^\theta(\boldsymbol{b}) = 0$ if $b_i < 0$ for any i.

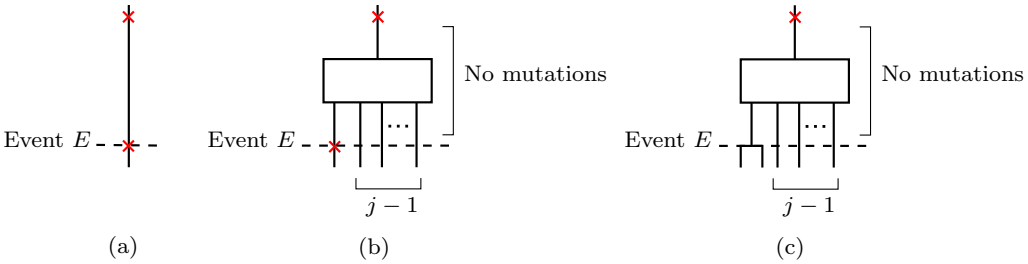


Fig. 4.2: Illustration of possible first events. (a) A mutation occurred in an ancestor that belonged to a 1-class. (b) A a mutation occurred in an ancestor that belonged to a j-class. (c) An ancestor in a j-class reproduced.

Note that the right hand side contains terms corresponding to mutation events (the first line) and coalescence events (the second line). The sample size stays the same in the former case, whereas the sample size decreases by one in the latter case.

Proof. The basic idea underlying the proof is to condition on the first event back in time. Let $\mathcal{F}$ denote the set of all possible types of the first event. Then,

$$p_n^\theta(\boldsymbol{a}) = \sum_{E\in\mathcal{F}} \mathbb{P}(\boldsymbol{a}|E)\ \mathbb{P}(E).$$

Given the type of the first event, we then find the conditional probability of obtaining the configuration $\boldsymbol{a}$. For this reason, the method is sometimes referred to as the "backward-forward" argument.

Let $\boldsymbol{b}$ denote the sample configuration immediately after (going backwards in time) the event E. Then,

$$p_n^\theta(\boldsymbol{a}) = \sum_{E\in\mathcal{F}} \sum_{\boldsymbol{b}} \mathbb{P}(\boldsymbol{a}|\boldsymbol{b},E)\ \mathbb{P}(\boldsymbol{b}|E)\ \mathbb{P}(E).$$

As further detailed below, only certain configurations $\boldsymbol{b}$ will have non-zero forward probabilities $p_n^\theta(\boldsymbol{a}|\boldsymbol{b},E)$. First, we note that $\mathbb{P}(E)$ is simple to compute:

1. If the first event E back in time is a mutation event, then $\mathbb{P}(E) = \frac{\theta}{\theta+n-1}$.
2. If the first event E back in time is a coalescence event, then $\mathbb{P}(E) = \frac{n-1}{\theta+n-1}$.

We expand upon these two cases below.

Case 1 (E is a mutation event): Going forward in time, suppose a mutation occurred in an ancestor that belonged to a j-class (i.e., an allele type seen j times), where $1 \le j \le n$. If $j = 1$ (as illustrated in Figure 4.2a), then $\boldsymbol{b} = \boldsymbol{a}$ and

$$\mathbb{P}(\boldsymbol{a} \mid \boldsymbol{b}, E) = \frac{b_1}{n} = \frac{a_1}{n}.$$

If $j > 1$ (as illustrated in Figure 4.2b), then $\boldsymbol{b} = \boldsymbol{a} - \boldsymbol{e}_1 - \boldsymbol{e}_{j-1} + \boldsymbol{e}_j$ and

$$\mathbb{P}(\boldsymbol{a} \mid \boldsymbol{b}, E) = \frac{jb_j}{n} = \frac{j(a_j+1)}{n}.$$

At stationarity, $\mathbb{P}(\boldsymbol{b}|E) = p_n^\theta(\boldsymbol{b})$. Combining the above factors, we obtain the first line of (4.3).

Case 2 (E is a coalescence event): Suppose an ancestor in a j-class of $\boldsymbol{b}$ reproduced, where $1 \le j \le n-1$ (equivalently, two members of a $(j+1)$-class of $\boldsymbol{a}$ coalesced), as illustrated in Figure 4.2c. Then $\boldsymbol{b} = \boldsymbol{a} - \boldsymbol{e}_{j+1} + \boldsymbol{e}_j$ and $|\boldsymbol{b}| = n-1$. Since all ancestors are equally likely to reproduce, we have

$$\mathbb{P}(\boldsymbol{a} \mid \boldsymbol{b}, E) = \frac{jb_j}{n-1} = \frac{j(a_j+1)}{n-1}.$$

At stationarity, $\mathbb{P}(\boldsymbol{b}|E) = p_{n-1}^\theta(\boldsymbol{b})$. Summing over j gives the second line of (4.3). □

The following closed-form sampling formula for the infinite-alleles model was first obtained by Ewens (1972) and later proved by Karlin and McGregor (1972):

Theorem 4.5 (Ewens Sampling Formula (ESF)). *Suppose $\boldsymbol{a} \vdash n$ is a sample configuration with sample size $|\boldsymbol{a}| = n$. Then, at stationarity,*

$$\mathrm{ESF}^\theta(\boldsymbol{a}) \stackrel{\text{def}}{=} p_n^\theta(\boldsymbol{a}) = \frac{n!\theta^{\|\boldsymbol{a}\|_1}}{(\theta)_{n\uparrow}} \prod_{i=1}^{n} \left(\frac{1}{i}\right)^{a_i} \frac{1}{a_i!}. \tag{4.4}$$

Proof. This result can be proved by double induction on $|\boldsymbol{a}|$ and $\|\boldsymbol{a}\|_1$ using the recursion (4.3) in Theorem 4.4. The details are left as an exercise. □

As mentioned earlier, this sampling formula for the infinite-allele model is the same as the sampling formula $\mathbb{P}_\theta(\boldsymbol{c}(\sigma) = \boldsymbol{a})$ in (4.2) for the θ-based random permutation model.

4.3 The coalescent with killing

Consider a sequential sampling scheme in which we draw alleles one at a time from the population. Label the first allele by A_1. In the ith draw, if the observed allele is the same as one of the first $i-1$ alleles, label it by that allele label. Otherwise, give it a new label; A_k denotes the kth distinct allele type observed. Repeat until the sample size is n and define the following notation:

- Let K denote the number of distinct alleles observed in the sample.
- Let n_i denote the number of times that allele type A_i is observed, and define $\boldsymbol{n} = (n_1, \ldots, n_K)$.

This sequential sampling induces a unique partition $\{B_1, \ldots, B_K\}$ of $[n]$, where $i \in B_k$ if and only if the ith allele drawn is type A_k.

Consider the following example with $K = 4$ and $\boldsymbol{n} = (3, 1, 2, 1)$:

i :	1	2	3	4	5	6	7
ith allele :	A_1	A_2	A_1	A_3	A_1	A_3	A_4

The partition corresponding this ordered sample is

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

Under the infinite-alleles model, what is the probability of observing a particular partition of $[n]$? To answer this question, we introduce the *ordered* sampling probability $q_n^\theta(\boldsymbol{n})$, which denotes the probability of an ordered sample with configuration $\boldsymbol{n}$. It it left as a simple exercise to show the following lemma:

Lemma 4.6. *Given an ordered sample with configuration $\boldsymbol{n} = (n_1, \ldots, n_K)$, let $\boldsymbol{a} \vdash n$ with $\|\boldsymbol{a}\|_1 = K$ be the corresponding unordered sample configuration. (Note that $\boldsymbol{a}$ is completely determined by $\boldsymbol{n}$.) Then,*

$$p_n^\theta(\boldsymbol{a}) = \left[\prod_{j=1}^{n} \frac{1}{a_j!}\right] \binom{n}{n_1, \ldots, n_K} q_n^\theta(\boldsymbol{n}). \tag{4.5}$$

Genealogically, an ordered sample corresponds to choosing a particular assignment of the alleles in a sample to the leaf labels $1, \ldots, n$, while the probability of an unordered sample sums over all inequivalent such assignments. In general it is more convenient to work with an ordered sample than an unordered one. For the infinite-alleles model, this aspect is clearly illustrated by the following recursion for $q_n^\theta(\boldsymbol{n})$, which is much simpler than the recursion (4.3) satisfied by $p_n^\theta(\boldsymbol{a})$:

Theorem 4.7 (Probability recursion for an ordered sample). *The ordered sampling probability q_n^θ satisfies the following recursion:*

$$q_n^\theta(\boldsymbol{n}) = \frac{\theta}{n-1+\theta} \sum_{i=1}^{K} \frac{\delta_{n_i,1}}{n} q_{n-1}^\theta(\boldsymbol{n} - \boldsymbol{e}_i) + \frac{n-1}{n-1+\theta} \sum_{i=1}^{K} \frac{n_i(n_i-1)}{n(n-1)} q_{n-1}^\theta(\boldsymbol{n} - \boldsymbol{e}_i), \tag{4.6}$$

with boundary conditions $q_1(\boldsymbol{e}_i) = 1$ for all $i \in [K]$.

Proof. This recursion can be proved using Theorem 4.4 and Lemma 4.6. A simpler strategy is to construct a recursion directly for $q_n^\theta(\boldsymbol{n})$, by following a similar line of argument as in the proof of Theorem 4.4, adapted for an ordered sample. For example, suppose the sample is ordered such that allele i appears before allele j if $i < j$. In the second term of (4.6), the factor $\frac{n_i-1}{n-1}$ corresponds to the probability that type i branches given that there is a branching, while the factor $\frac{n_i}{n}$ corresponds to the probability of inserting the newborn (which is of allele type i) into the ordered sample of size $n-1$ with $n_i - 1$ copies of type i, such that the ordering convention still holds after the insertion (i.e., allele i appears before allele j if $i < j$). □

Using Theorem 4.5 and Lemma 4.6, one can obtain the following formula for the ordered sampling probability:

Theorem 4.8 (Ordered ESF). *For an ordered sample with configuration $\boldsymbol{n} = (n_1, \ldots, n_K)$ and sample size $n = \sum_i n_i$,*

$$q_n^\theta(\boldsymbol{n}) = \frac{\theta^K}{(\theta)_{n\uparrow}} \prod_{i=1}^{K} (n_i - 1)!. \tag{4.7}$$

The recursion (4.6) for the ordered sampling probability is strictly recursive in the sample size since we only have q_{n-1}^θ terms appearing on the right hand side. This suggests that we

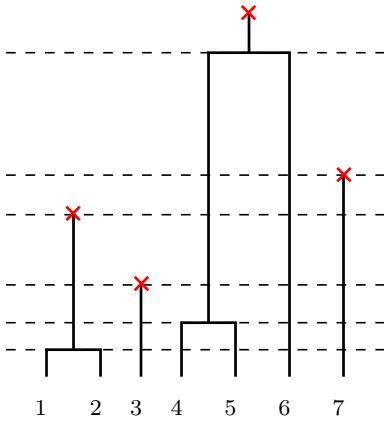


Fig. 4.3: Illustration of a sample path from the coalescent with killing. The partition of [7] induced by this realization is $\{\{1,2\},\{3\},\{4,5,6\},\{7\}\}$. There are two 1-leaved subtrees, one 2-leaved subtree, and one 3-leaved subtree, so $\boldsymbol{a} = (2,1,1,0,0,0,0)$.

can define a simpler stochastic process in which a lineage is lost whenever either a mutation or a coalescence occurs. More precisely, (4.6) implies that such a stochastic process should follow the dynamics described below:

Definition 4.9 (The Coalescent with Killing). The coalescent with killing for a sample of size n starts with n leaves. Then, for $k = n, n-1, \ldots, 2, 1$,

1. Draw a waiting time $T_k \sim \text{Exp}\left[\binom{k}{2} + k\frac{\theta}{2}\right]$.
2. After time T_k since the last event, the next event is chosen as follows.
   a. With probability $\frac{\theta}{k-1+\theta}$, kill a lineage chosen uniformly at random due to a mutation.
   b. With probability $\frac{k-1}{k-1+\theta}$, merge two lineages chosen uniformly at random.

See Figure 4.3 for an illustration of the process. The coalescent with killing gives rise to a forest of subtrees, where each mutation event leads to a subtree. Further, the forest defines a partition of $[n]$, with the leaves of each subtree defining a block of the partition. It follows from Theorem 4.8 that the probability distribution of such a random partition is given by:

Theorem 4.10. *Let Π_n denote a random partition of $[n]$ induced by the coalescent with killing for n leaves. For $\pi = \{B_1, \ldots, B_K\} \in \mathcal{P}_{[n]}$,*

$$\mathbb{P}^{\theta}_{\text{CK}}(\Pi_n = \pi) = \frac{\theta^K}{(\theta)_{n\uparrow}} \prod_{i=1}^{K} (|B_i| - 1)!. \tag{4.8}$$

Further, let $\Lambda_{n,i}$ denote the number of blocks of size i in Π_n, and define $\boldsymbol{\Lambda}_n = (\Lambda_{n,1}, \Lambda_{n,2}, \ldots, \Lambda_{n,n})$. Then, given $\boldsymbol{a} \vdash n$,

$$\mathbb{P}^{\theta}_{\text{CK}}(\boldsymbol{\Lambda}_n = \boldsymbol{a}) = \frac{n!\theta^{\|\boldsymbol{a}\|_1}}{(\theta)_{n\uparrow}} \prod_{i=1}^{n} \left(\frac{1}{i}\right)^{a_i} \frac{1}{a_i!} = \text{ESF}^{\theta}(\boldsymbol{a}). \tag{4.9}$$

4.4 Ancestral process under the coalescent with killing

Consider the ancestral process $\{A_n^\theta(t), t \geq 0\}$ for the coalescent with killing, where $A_n^\theta(t) \in \{0, 1, \ldots, n\}$ denotes the number of surviving lineages at time t. Because of the killing due to mutation, $\lim_{t\to\infty} A_n^\theta(t) = 0$ almost surely. By finding a spectral decomposition of the generator for this process, Tavaré (1984) obtained the following result:

Theorem 4.11 (Tavaré 1984). *For $\theta > 0$, the ancestral process for the coalescent with killing has the following distribution:*

$$\mathbb{P}_{\mathrm{CK}}^\theta(A_n^\theta(t) = j) = \begin{cases} \displaystyle\sum_{k=1}^{n} e^{-[\binom{k}{2}+\frac{k\theta}{2}]t} \frac{(-1)^{k-j}(2k+\theta+1)(j+\theta)_{(k-1)\uparrow}(n)_{k\downarrow}}{j!(k-j)!(n+\theta)_{k\uparrow}}, & \text{if } 1 \leq j \leq n, \\ \displaystyle 1 + \sum_{k=1}^{n} e^{-[\binom{k}{2}+\frac{k\theta}{2}]t} \frac{(-1)^{k}(2k+\theta+1)(\theta)_{(k-1)\uparrow}(n)_{k\downarrow}}{k!(n+\theta)_{k\uparrow}}, & \text{if } j = 0. \end{cases}$$

The above result reduces to (1.8) as $\theta \to 0$. Similarly, Theorem 1.8 can be generalized as follows:

Theorem 4.12. *The jth factorial moment of $A_n^\theta(t)$ is given by*

$$\mathbb{E}\left[(A_n^\theta(t))_{j\downarrow}\right] = \sum_{k=j}^{n} e^{-[\binom{k}{2}+\frac{k\theta}{2}]t}(2k+\theta-1)\binom{k-1}{j-1}\frac{(\theta+k)_{(j-1)\uparrow}(n)_{k\downarrow}}{(n+\theta)_{k\uparrow}}.$$

4.5 Hoppe's urn model

In the coalescent with killing, a lineage is lost whenever either a mutation or a coalescence occurs. We can think of this as the sample size decreasing by one. Let $\boldsymbol{\Lambda}_n \to \boldsymbol{\Lambda}_{n-1} \to \cdots \to \boldsymbol{\Lambda}_1$ denote a random sequence of unordered sample configurations encountered by going backwards in time under the coalescent with killing. We will be interested in a sequence of this kind when we discuss importance sampling later on: Given $\boldsymbol{\Lambda}_n = \boldsymbol{a}$, where $\boldsymbol{a} \vdash n$ denotes observed sample configuration, we want to sample a configuration $\boldsymbol{\Lambda}_{n-1}$ one event back in time from the posterior distribution, and iterate this procedure until arriving at $\boldsymbol{\Lambda}_1 = \boldsymbol{e}_1$.

The above sequence is a Markov chain, but its transition probability is not immediately obvious. However, by reversing the time direction of the coalescent with killing, we obtain a forward-in-time Markov chain for which the transition probability is simple to write down. This process can be formulated as a Pólya-like urn model (sampling with replacement), due to Hoppe (1984), in which the urn contains balls of mass either θ or 1. The urn starts with a single black ball of mass θ, and one draws and replaces balls in the urn as follows.

1. Draw a ball X from the urn with probability proportional to the mass of the ball.
   a. If X is black, return it together with a ball of a new color with mass 1.
   b. If X is not black, return it together with a ball of the same color as X, also with mass 1.

2. Repeat sampling n times.

Note that the number of distinct colors (aside from black) in the urn after the urn process has been finished is the number of times that the black ball of mass θ is drawn. Let $\Gamma_{n,i}$ denote the number of colors, discounting the black θ-ball, represented i times in the urn upon n draws, and let $\boldsymbol{\Gamma}_n = (\Gamma_{n,1}, \Gamma_{n,2}, \ldots, \Gamma_{n,n})$. Then, one can easily show the following result on transition probability:

Lemma 4.13. *For $\boldsymbol{b} \vdash (k+1)$ and $\boldsymbol{a} \vdash k$, the transition probability under Hoppe's urn model is given by*

$$\mathbb{P}^{\theta}_{\mathrm{H}}(\boldsymbol{\Gamma}_{k+1} = \boldsymbol{b} \mid \boldsymbol{\Gamma}_k = \boldsymbol{a}) = \begin{cases} \frac{\theta}{k+\theta}, & \text{if } \boldsymbol{b} = (\boldsymbol{a} + \boldsymbol{e}_1, 0), \\ \frac{ia_i}{k+\theta}, & \text{if } \boldsymbol{b} = (\boldsymbol{a} - \boldsymbol{e}_i + \boldsymbol{e}_{i+1}, 0), \text{ for } i = 1, \ldots, k-1, \\ \frac{k}{k+\theta}, & \text{if } \boldsymbol{b} = \boldsymbol{e}_{k+1} \text{ and } \boldsymbol{a} = \boldsymbol{e}_k, \\ 0, & \text{otherwise.} \end{cases}.$$

(Here, $(\boldsymbol{a} + \boldsymbol{e}_1, 0)$ means concatenating a 0 to the k-tuple $\boldsymbol{a} + \boldsymbol{e}_1$ to obtain a $(k+1)$-tuple; $(\boldsymbol{a} - \boldsymbol{e}_i + \boldsymbol{e}_{i+1}, 0)$ is similarly defined.)

Hoppe (1984) proved the following result:

Theorem 4.14 (Hoppe 1984). *For a given partition $\boldsymbol{a} \vdash n$, the probability distribution $\mathbb{P}^{\theta}_{\mathrm{H}}$ under the above urn model satisfies*

$$\mathbb{P}^{\theta}_{\mathrm{H}}(\boldsymbol{\Gamma}_n = \boldsymbol{a}) = \mathrm{ESF}^{\theta}(\boldsymbol{a}),$$

where $\mathrm{ESF}^{\theta}(\boldsymbol{a})$ is defined in (4.4).

By construction, the forward-in-time transition probability under the coalescent with killing is the same as the transition probability in Hoppe's urn model:

Proposition 4.15. *The forward-direction process $\boldsymbol{\Lambda}_1 \to \boldsymbol{\Lambda}_2 \to \cdots \to \boldsymbol{\Lambda}_n$ under the coalescent with killing is a Markov chain. Furthermore, the transition probability for this process is equivalent to that of Hoppe's urn model; i.e.,*

$$\mathbb{P}^{\theta}_{\mathrm{CK}}(\boldsymbol{\Lambda}_{k+1} = \boldsymbol{b} \mid \boldsymbol{\Lambda}_k = \boldsymbol{a}) = \mathbb{P}^{\theta}_{\mathrm{H}}(\boldsymbol{\Gamma}_{k+1} = \boldsymbol{b} \mid \boldsymbol{\Gamma}_k = \boldsymbol{a}).$$

The reverse transition probability $\mathbb{P}^{\theta}_{\mathrm{CK}}(\boldsymbol{\Lambda}_k = \boldsymbol{a} \mid \boldsymbol{\Lambda}_{k+1} = \boldsymbol{b})$ for the coalescent with killing can then be computed using Bayes' rule, together with Proposition 4.15 and equation (4.9).

$$\mathbb{P}^{\theta}_{\mathrm{CK}}(\boldsymbol{\Lambda}_k = \boldsymbol{a} \mid \boldsymbol{\Lambda}_{k+1} = \boldsymbol{b}) = \begin{cases} \frac{b_1}{k+1}, & \text{if } \boldsymbol{b} = (\boldsymbol{a} + \boldsymbol{e}_1, 0), \\ \frac{ib_i}{k+1}, & \text{if } \boldsymbol{b} = (\boldsymbol{a} - \boldsymbol{e}_{i-1} + \boldsymbol{e}_i, 0), \text{ for } i = 2, \ldots, k, \\ 1, & \text{if } \boldsymbol{b} = \boldsymbol{e}_{k+1} \text{ and } \boldsymbol{a} = \boldsymbol{e}_k, \\ 0, & \text{otherwise.} \end{cases} \tag{4.10}$$

Note that this reverse transition probability does not depend on the mutation rate. Furthermore, it implies that, given $\boldsymbol{\Lambda}_{k+1} = \boldsymbol{b}$, a configuration $\boldsymbol{\Lambda}_k$ one event back in time can be sampled from the correct posterior distribution by picking an allele in $\boldsymbol{b}$ uniformly at

random and removing it. (If the chosen allele is a singleton, then it gets killed by a mutation event. If it has copy number greater than one, then it coalesces with another allele of the same type.

4.6 Chinese Restaurant Process

If you were asked to implement an algorithm to sample a random permutation from $\mathscr{S}_n$, how would you do it? Enumerating all possible permutations is out of question for even a moderate n, since $|\mathscr{S}_n| = n!$ would be a very large number (e.g., for $n = 20$, $n! > 2.4 \times 10^{18}$). As we will see below, the Chinese Restaurant Process (CRP) can be used to generate a θ-biased random permutation in $O(n)$ time. It was first considered by Aldous (1985) and then later in population genetics by Joyce and Tavaré (1987). The latter rephrased CRP as Hoppe's Urn model plus some extra bookkeeping. Imagine a restaurant with infinitely many tables and n customers who are lined up outside the door in order, with 1 first in line. The customers enter the restaurant one at a time and seat themselves as follows:

1. The kth customer chooses to do the following:
   a. with probability $\frac{\theta}{k-1+\theta}$, she starts a new table, or
   b. with probability $\frac{1}{k-1+\theta}$, she sits to the left of a particular person already seated.
2. After all the customers have been seated, each non-empty table defines a cycle and the collection of all non-empty tables defines a random permutation of $[n]$.

Example 4.16. Suppose there are six customers who sit down as follows:

- 1 sits at table 1.
- 2 sits at table 2.
- 3 sits to the left of 1.
- 4 sits to the left of 2.
- 5 sites to the left of 2.
- 6 sits at table 3.

Then the resulting permutation is (13)(254)(6).

Below we relate the CRP with the infinite alleles model.

Theorem 4.17. *In the CRP, the probability of generating a permutation $\sigma \in \mathscr{S}_n$ is*

$$\mathbb{P}^{\theta}_{\mathrm{CRP}}(\sigma) = \frac{\theta^k}{(\theta)_{n\uparrow}},$$

where k is the number of cycles in σ.

Proof. Each new cycle requires that someone starts a new table, so k tables are started. Therefore,

$$\mathbb{P}^{\theta}_{\mathrm{CRP}}(\sigma) = \theta^k \prod_{i=1}^{n} \frac{1}{i-1+\theta} = \frac{\theta^k}{(\theta)_{n\uparrow}},$$

which is the desired result. □

Theorem 4.18. *Let $\boldsymbol{\Omega}_n = (\Omega_{n,1}, \Omega_{n,2}, \ldots, \Omega_{n,n})$, where $\Omega_{n,i}$ denotes the number of cycles of length i in a CRP-generated random permutation of $[n]$. Let $\mathbb{P}^{\theta}_{\mathrm{CRP}}$ denote the probability distribution under the CRP. Then, for every partition $\boldsymbol{a} \vdash n$,*

$$\mathbb{P}^{\theta}_{\mathrm{CRP}}(\boldsymbol{\Omega}_n = \boldsymbol{a}) = \mathbb{P}^{\theta}_{\mathrm{CK}}(\boldsymbol{\Lambda}_n = \boldsymbol{a}).$$

Proof. We know from earlier that $\mathbb{P}^{\theta}_{\mathrm{CK}}(\boldsymbol{\Lambda}_n = \boldsymbol{a}) = \mathrm{ESF}^{\theta}(\boldsymbol{a})$. For CRP,

$$\begin{aligned}
\mathbb{P}^{\theta}_{\mathrm{CRP}}(\boldsymbol{\Omega}_n = \boldsymbol{a}) &= \sum_{\sigma : \boldsymbol{c}(\sigma) = \boldsymbol{a}} \mathbb{P}^{\theta}_{\mathrm{CRP}}(\sigma) \\
&= |\{\sigma \in \mathscr{S}_n \mid \boldsymbol{\Omega}_n(\sigma) = \boldsymbol{a}\}| \cdot \frac{\theta^k}{(\theta)_{n\uparrow}} \\
&= \mathrm{ESF}^{\theta}(\boldsymbol{a}).
\end{aligned}$$

In summary, $\mathbb{P}^{\theta}_{\mathrm{CRP}}(\boldsymbol{\Omega}_n = \boldsymbol{a}) = \mathrm{ESF}^{\theta}(\boldsymbol{a}) = \mathbb{P}^{\theta}_{\mathrm{CK}}(\boldsymbol{\Lambda}_n = \boldsymbol{a})$. □

We end this section with the following result by Goncharov on the distribution of $w_{n,j}$:

Theorem 4.19 (Goncharov 1944). *Consider the CRP with $\theta = 1$ for sampling a permutation in $\mathscr{S}_n$ under the uniform distribution. Then, for $j \in [n]$,*

$$\mathbb{P}^{\theta}_{\mathrm{CRP}}(\Omega_{n,j} = w) = \frac{1}{j^w w!} \sum_{l=0}^{\lfloor n/j \rfloor - w} (-1)^l \frac{1}{j^l l!}.$$

4.7 The number of distinct allele types

Given a permutation $\sigma \in \mathscr{S}_n$, recall that $\boldsymbol{c}(\sigma) = (c_1, \ldots, c_n)$ denotes the cycle type of σ, with c_j being the number of cycles of length j in σ. Let $K_n(\sigma) = \|\boldsymbol{c}(\sigma)\|_1$ denote the total number of cycles in σ. We will prove the following convergence result for the function K_n:

Theorem 4.20. *For all θ-biased random permutation models with $\theta > 0$,*

$$\frac{K_n - \mathbb{E}(K_n)}{\sqrt{\mathrm{Var}(K_n)}} \xrightarrow{d} N(0,1) \qquad \text{as } n \to \infty.$$

So far we learned that the coalescent under the infinite-alleles model of mutation is equivalent to the coalescent with killing. Furthermore, we learned that the coalescent with killing is related to the Chinese Restaurant Process (CRP), which is a model for generating θ-biased permutations. We will prove Theorem 4.20 by utilizing this connection between the θ-biased random permutation model and the infinite-alleles model with mutation rate $\theta/2$. We will also obtain closed-form formulas for $\mathbb{E}(K_n)$ and $\mathrm{Var}(K_n)$.

The distribution of K_n can be found in closed form:

Theorem 4.21. *Suppose $\boldsymbol{a} \vdash n$. For all $k \in [n]$,*

$$\mathbb{P}(K_n = k \mid \theta) = \frac{\theta^k}{(\theta)_{n\uparrow}} |s(n,k)|,$$

where $|s(n,k)| = |\{\sigma \in \mathscr{S}_n : \sigma \text{ has } k \text{ cycles}\}|$ are unsigned Stirling numbers of the first kind.

Proof. There are at least two ways to prove this result. The first one is far easier, but it is hard to use when trying to calculate $\mathbb{E}(K_n)$. The second one invokes intuition that will pay off in later calculations.

Method 1: Calculate $\mathbb{P}(K_n = k|\theta)$ directly by marginalizing over the ESF.

$$\mathbb{P}(K_n = k \mid \theta) = \sum_{\boldsymbol{a}:\boldsymbol{a}\vdash n, \|\boldsymbol{a}\|_1 = k} \mathrm{ESF}^\theta(\boldsymbol{a}) = \frac{\theta^k}{(\theta)_{n\uparrow}} \sum_{\boldsymbol{a}:\boldsymbol{a}\vdash n, \|\boldsymbol{a}\|_1 = k} N(n,\boldsymbol{a}) = \frac{\theta^k}{(\theta)_{n\uparrow}} |s(n,k)|.$$

The last equality follows from the definition of $|s(n,k)|$.

Method 2: In the coalescent with killing, let I_j be the indicator variable defined as

$$I_j = \begin{cases} 1, & \text{if the } j\text{th event is a killing by mutation,} \\ 0, & \text{if the } j\text{th event is a coalescence.} \end{cases}$$

Then, $K_n = I_1 + I_2 + \cdots + I_n$. These indicator variables are independent (though not identically distributed) random variables, so the probability generating function $G_{K_n}(z)$ of K_n can be obtained by multiplying the probability generating functions of I_j, which are given by

$$G_{I_j}(z) = \mathbb{P}(I_j = 0) + \mathbb{P}(I_j = 1)z.$$

Clearly $\mathbb{P}(I_n = 1) = 1$, since the last lineage is always killed by mutation (i.e., there is no other lineage to coalesce with). For $1 \le j \le n-1$, we have $\mathbb{P}(I_j = 1) = \frac{\theta}{n-j+\theta}$, which gives

$$G_{I_j}(z) = \frac{n-j}{n-j+\theta} + \frac{\theta}{n-j+\theta} z = \frac{n-j+\theta z}{n-j+\theta}.$$

Therefore,

$$G_{K_n}(z) = \prod_{j=1}^{n} G_{I_j}(z) = \frac{\prod_{j=1}^{n}(n-j+\theta z)}{\prod_{j=1}^{n}(n-j+\theta)} = \frac{(\theta z)_{n\uparrow}}{(\theta)_{n\uparrow}}.$$

Now, letting $x = \theta z$ in the combinatorial identity

$$(x)_{n\uparrow} = \sum_{k=1}^{n} (-1)^{n-k} s(n,k) x^k = \sum_{k=1}^{n} |s(n,k)| x^k$$

and recalling the definition $G_{K_n}(z) = \sum_{k=1}^{n} \mathbb{P}(K_n = k|\theta) z^k$, we can match coefficients to obtain $\mathbb{P}(K_n = k|\theta) = \frac{\theta^k}{(\theta)_{n\uparrow}} |s(n,k)|$. □

In the above proof, Method 2 is harder than Method 1 for deriving the distribution of K_n, but it is much more powerful. For example, by the linearity of expectation, we immediately obtain,

$$\mathbb{E}(K_n) = \sum_{j=1}^{n} \mathbb{E}(I_j) = 1 + \sum_{j=1}^{n-1} \frac{\theta}{j+\theta}. \tag{4.11}$$

Further, since $I_1, \ldots, I_n$ are independent, we obtain

$$\mathrm{Var}(K_n) = \sum_{j=1}^{n} \mathrm{Var}(I_j) = \sum_{j=1}^{n} \left[\frac{\theta}{n-j+\theta} - \left(\frac{\theta}{n-j+\theta} \right)^2 \right] = \theta \sum_{j=1}^{n-1} \frac{j}{(j+\theta)^2}.$$

Lastly, Theorem 4.20 can be proved by applying the Lyapunov Central Limit Theorem (Billingsley, 2008, Chapter 27) to $I_1 + \cdots + I_n$ as $n \to \infty$.

4.8 A sufficient statistic for $\boldsymbol{\theta}$

In the coalescent with killing, the conditional distribution of block structure given the number of distinct alleles does not depend on the mutation rate:

$$\mathbb{P}^{\theta}_{\mathrm{CK}}(\boldsymbol{\Lambda}_n = \boldsymbol{a} \mid K_n = k) = \frac{n!}{|s(n,k)|} \prod_{j=1}^{n} \frac{1}{j^{a_j} a_j!}.$$

Therefore, K_n is a sufficient statistic for θ and the maximum likelihood estimate $\hat{\theta}$ of θ can be found from $L(\theta) = \mathbb{P}(K_n = k \mid \theta)$:

$$\frac{\partial}{\partial \theta} \log L(\theta) = 0 \implies k = 1 + \sum_{j=1}^{n-1} \frac{\hat{\theta}}{j + \hat{\theta}}.$$

Comparing this with (4.11), we see that the maximum likelihood estimate (MLE) of θ is equal to the moment estimate.

4.9 Population-wide distribution of allele frequencies

In the remainder of this chapter, we will discuss the population-wide distribution associated with the Ewens sampling formula. Consider a population of N individuals evolving under some discrete-time model, such as the Wright-Fisher model, with some mutation matrix on K alleles. Assuming equilibrium has been reached, take a random sample of size n from the population. Then for several such discrete-models, it has been shown that as the population size N and the number of alleles K go to ∞ while keeping the population-scaled mutation rate θ fixed, the probability distribution of the allelic configuration in the sample converges to that described by the Ewens sampling formula.

Suppose the alleles in a population of size N were labeled from 1 to K, with frequencies $X_1, \ldots, X_K$, where $0 < X_k < 1$ and

$$\sum_{k=1}^{K} X_k = 1.$$

Kingman was interested in characterizing the limiting distribution of these population frequencies given that the Ewens sampling formula holds in the limit. In particular, Kingman

(1977) showed that the joint distribution of the decreasing order statistics of the allele frequencies, $\{X_{(k)}\}$, converges to a *Poisson-Dirichlet* distribution parameterized by θ if and only if the allele frequencies generate the Ewens sampling formula in the limit of N and K going to ∞.

In what follows, we will describe some of these connections between the Ewens sampling formula and the Poisson-Dirichlet distribution, and in particular, to prove a special case of Kingman's result. We begin with some definitions.

4.9.1 Size-biased representation, stick breaking process, and the GEM distribution

Consider an infinite population with countably infinite number of alleles labeled by $\mathbb{N}$, where X_k is the frequency of allele k in the population, $0 < X_k < 1$ for all k and

$$\sum_{k=1}^{\infty} X_k = 1.$$

Sequentially sample n alleles according to the frequencies $\{X_k\}$, and let α_i be the ith allele drawn. Conditional on the population frequencies $\{X_k\}$, the random variables $\alpha_1, \ldots, \alpha_n$ are i.i.d. where

$$\mathbb{P}(\alpha_i = j \mid \{X_k\}) = X_j.$$

Let K_n be the number of distinct alleles in the sequence of alleles $\alpha_1, \ldots, \alpha_n$, and let $d_1, \ldots, d_{K_n}$ be the distinct alleles in the order in which they first appear in the sequence $\alpha_1, \ldots, \alpha_n$. Note that alleles with higher frequencies are more likely to be sampled first.

Example 4.22. Let $n = 10$ and $\alpha_1, \ldots, \alpha_n = 2, 4, 1, 2, 2, 5, 10, 4, 1, 2$. Then, $K_n = 5$ and $d_1, \ldots, d_5 = 2, 4, 1, 5, 10$.

Definition 4.23 (Size-biased representation). Taking the sample size $n \to \infty$, the infinite sequence $\{X_{d_i}\}$ is called the size-biased representation of the infinite sequence $\{X_k\}$. We can also call the permutation over $\mathbb{N}$ defined by the map $i \mapsto d_i$ as the size-biased permutation generated by the sequence $\{X_k\}$.

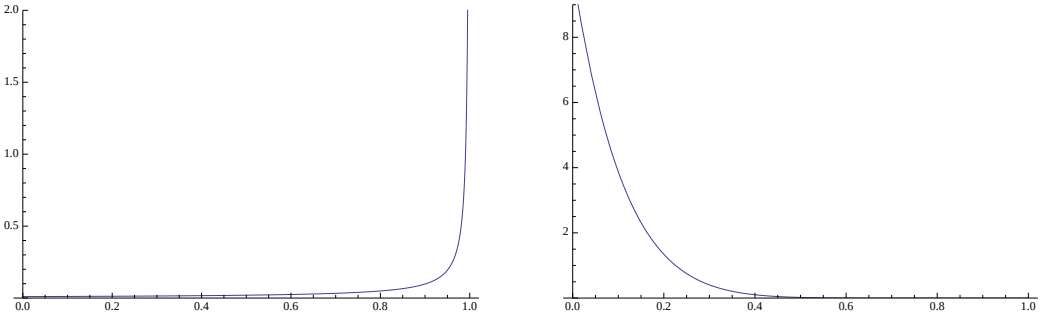
We now describe the so-called *stick breaking process* for generating a random measure over $\mathbb{N}$. Consider the infinite sequence of independent and identically distributed random variables $\{R_k\}$, where $R_k \sim \text{Beta}(1, \theta)$. These random variables correspond to a sequence of "residual fractions" in the stick breaking process. The probability density function of R_k is given by

$$f_{R_k}(x) = \theta(1-x)^{\theta-1},$$

where $0 < x < 1$ and $\theta > 0$. Define the sequence of random variables $\{X_k\}$ inductively as:

$$\begin{aligned} X_1 &= R_1, \\ X_k &= (1 - X_1 - \cdots - X_{k-1})R_k, \qquad \text{for } k > 1. \end{aligned}$$

Pictorially, you can think of a stick of length 1, which is broken up into infinitely many pieces with lengths given by $\{X_k\}$. The first piece has length distributed as $\text{Beta}(1, \theta)$ of the



(a) $\theta = 0.01$ (b) $\theta = 10$

Fig. 4.4: Probability density function of $\text{Beta}(1, \theta)$

stick length, the second piece has length distributed as a $\text{Beta}(1, \theta)$ fraction of the remaining length of the stick, and so on, ad infinitum.

For $\theta \ll 1$, most of the mass of the $\text{Beta}(1, \theta)$ distribution is concentrated near 1 (Figure 4.4a), and we expect to see one long piece with rest being very short. For $\theta \gg 1$, $\text{Beta}(1, \theta)$ is concentrated near 0, and we expect to see several pieces of roughly equal length (Figure 4.4b) taking up most of the stick length.

Definition 4.24 (GEM distribution). The joint distribution of the sequence of random variables $\{X_k\}$ generated by the stick breaking process with parameter θ is called $\text{GEM}(\theta)$.

The GEM distribution is named after Griffiths (1979), Engen (1978) and McCloskey (1965), who made connections between this distribution and population genetics and ecology. In fact, the following theorem establishes how it is related to the infinite alleles model:

Theorem 4.25. *Take a sample of size n from a population with infinitely many alleles of which frequencies are given by an infinite sequence $\{X_k\}$ distributed as the GEM(θ) distribution. Let $C_{n,i}$ be the number of alleles represented i times in the sample, and $\boldsymbol{C_n} = (C_{n,1}, C_{n,2}, \ldots, C_{n,n})$. Then,*

$$\mathbb{P}(\boldsymbol{C_n} = \boldsymbol{a}) = \text{ESF}^{\theta}(\boldsymbol{a}).$$

Remark 4.26. Here are some useful facts about the GEM distribution.

1. Since the residual fractions $\{R_k\}$ are independent in the stick breaking construction, it is easy to see that
$$\mathbb{E}[X_j] = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^{j-1},$$
which is a decreasing function of j for all $\theta > 0$.
2. The GEM distribution is invariant under size-biased permutations. More precisely, if $\{X_k\}$ is drawn from $\text{GEM}(\theta)$, then $\{X_{d_k}\}$ is also distributed as $\text{GEM}(\theta)$, where $\{d_k\}$ is a size-biased random permutation generated by $\{X_k\}$.
3. Let $X_{(1)} > X_{(2)} > \cdots$ denote the decreasing order statistics of $X_1, X_2, \ldots$; i.e., $X_{(j)}$ is the jth largest element of $\{X_k\}$. Then, the sequence $\{X_{(k)}\}$ follows the so-called Poisson-Dirichlet distribution with parameter θ, defined below.

4.9.2 Poisson-Dirichlet point process

Definition 4.27 (Poisson-Dirichlet point process and distribution). Let $\{Y_k\}$ be a non-homogeneous Poisson point process on $\mathbb{R}_+$ with intensity measure density $\frac{\theta e^{-y}}{y}$ for $y > 0$ and $\theta > 0$. This is called the Poisson-Dirichlet point process with parameter θ. Let $\{Y_{(k)}\}$ be the decreasing order statistics of $\{Y_k\}$ (i.e., $Y_{(1)} > Y_{(2)} > \cdots$) and let $Y = \sum_{k=1}^{\infty} Y_k$. Define $X_{(k)} = Y_{(k)}/Y$. The law of the sequence of random variables $\{X_{(k)}\}$ is called the Poisson-Dirichlet distribution with parameter θ, denoted by $\mathrm{PD}(\theta)$.

Note that the intensity measure density $\frac{\theta e^{-y}}{y}$ blows up as $y \to 0$, so there is an accumulation of points near 0. However, the density vanishes as $y \to \infty$, and the decreasing order statistics $\{Y_{(k)}\}$ in Definition 4.27 exist because the intensity measure of any interval (a, b) for $0 < a < b$ is finite. In particular, $\int_a^{\infty} \frac{\theta e^{-y}}{y} dy < \infty$.

We will see below that $Y \sim \mathrm{Gamma}(\theta, 1)$, i.e., the probability density function of Y is given by

$$f_Y(y) = \frac{y^{\theta-1} e^{-y}}{\Gamma(\theta)},$$

and that Y is independent of $\{X_{(k)}\}$.

4.9.3 Probability generating functional

Here, we take a slight detour to discuss a useful result regarding Poisson point processes, which we will employ in the next section to prove Kingman's result mentioned in Chapter 4.9. Recall that the probability generating function of a random variable Y is given by

$$G_Y(z) = \mathbb{E}[z^Y].$$

For example, for $Y \sim \mathrm{Poisson}(\lambda)$,

$$G_Y(z) = e^{\lambda(z-1)}. \tag{4.12}$$

This definition can be extended to d random variables $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ as

$$G_{\boldsymbol{Y}}(\boldsymbol{z}) = G_{\boldsymbol{Y}}(z_1, \ldots, z_d) = \mathbb{E}\left[\prod_{i=1}^{d} z_i^{Y_i}\right].$$

However, when working with a point process such as a Poisson point process on $\mathbb{R}^+$, we have one random variable per point, and the number of these random variables itself is random. In this case, we can define a generalization called the *probability generating functional*. For a Poisson point process $\{N(y), y > 0\}$, where $N(y)$ denotes the total number of points in $(0, y]$, it is defined as

$$G[f] = \mathbb{E}\left[\exp\left(\int_{\mathbb{R}^+} \log f(y) dN(y)\right)\right], \tag{4.13}$$

where f is assumed to be some smooth function of y which is different from 1 only on some bounded subset of $\mathbb{R}^+$. Let $\{Y_i\}$ denote the coordinates of the random points in the Poisson point process. Then, since $N(y)$ is a step function, the integral in (4.13) can be replaced with $\sum_i \log f(Y_i)$, thus yielding

$$G[f] = \mathbb{E}\left[\prod_i f(Y_i)\right]. \tag{4.14}$$

The aforementioned restriction on f ensures that (4.14) converges. If the intensity measure density of the Poisson point process is $\rho(y)$, then by conditioning on the number of points and using the tower rule, one can show that

$$G[f] = \exp\left(\int_{\mathbb{R}^+} \rho(y)[f(y) - 1]dy\right). \tag{4.15}$$

Note the similarity with (4.12). A derivation of (4.15) from (4.13) can be found in Section 2.7 of Cox and Isham (1980). The probability generating functional uniquely determines a Poisson point process, and we will use this fact in the next section.

4.9.4 Limit of a symmetric mutation model with K alleles

We will now prove a special case of Kingman's result described in Chapter 4.9. We can take a symmetric K-alleles mutation model and let K go to infinity in order to create an infinite alleles model, whose sampling probability is given by Ewens sampling formula. We will show that the decreasing order statistics of the allele frequencies in the K-alleles model with mutation rate $\theta/2$ converges to the Poisson-Dirichlet distribution PD(θ) as $K \to \infty$.

Consider the K-allele parent-independent mutation model, with mutation rate $\theta/2$ and mutation matrix entries $P_{ij} = \pi_j$. Then, as we will discuss in Chapter 6, at stationarity the population-wide allele frequencies $X_1, \ldots, X_K$ are distributed as

$$(X_1, \ldots, X_K) \sim \text{Dirichlet}(\theta\pi_1, \ldots, \theta\pi_K).$$

Now set $\pi_i = 1/K$ for all i, so that

$$(X_1, \ldots, X_K) \sim \text{Dirichlet}\left(\frac{\theta}{K}, \ldots, \frac{\theta}{K}\right).$$

We would like to take $K \to \infty$ and show that the joint distribution of the decreasing order statistics of $X_1, \ldots, X_K$ converges to PD(θ). However $X_1, \ldots, X_K$ are not independent random variables (in particular, $X_1 + \cdots + X_K = 1$) and this might be hard to do. To circumvent this, we use the following theorem about the connection between the Gamma and Dirichlet distributions.

Theorem 4.28. *Let $U_i \sim \text{Gamma}(r_i, \lambda)$, for $i = 1, \ldots, K$, be independent random variables, where r_i and λ are shape and scale parameters, respectively. Then, if we define $U = U_1 + \cdots + U_K$ and $V_i = U_i/U$, it follows that*

$$U \sim \text{Gamma}(r_1 + \cdots + r_k, \lambda),$$
$$(V_1, \ldots, V_K) \sim \text{Dirichlet}(r_1, \ldots, r_k),$$

and U is independent of $(V_1, \ldots, V_K)$.

Let $Y_1, \ldots, Y_K$ be i.i.d. random variables distributed as Gamma$(\frac{\theta}{K}, 1)$. Then, from Theorem 4.28, it follows that $Y = Y_1 + \cdots + Y_K$ is distributed as Gamma$(\theta, 1)$ and is independent of $\frac{Y_1}{Y}, \ldots, \frac{Y_K}{Y}$. Moreover,

$$\left(\frac{Y_1}{Y}, \ldots, \frac{Y_K}{Y}\right) \stackrel{d}{=} (X_1, \ldots, X_K).$$

Hence, to show that the distribution of the decreasing order statistics of $(X_1, \ldots, X_K)$ converges to PD(θ), it suffices to show that as $K \to \infty$, $Y_1, \ldots, Y_K$ look like points drawn from the Poisson-Dirichlet process.

If we treat the random variables $Y_1, \ldots, Y_K$ as points realized from a point process, then using (4.14), we have

$$\begin{aligned} G_K[f] &= \mathbb{E}\left[\prod_{i=1}^{K} f(Y_i)\right] \\ &= \left(\int_0^\infty f(y) \frac{e^{-y} y^{\frac{\theta}{K}-1}}{\Gamma(\frac{\theta}{K})} dy\right)^K \\ &= \left(1 + \frac{1}{K}\int_0^\infty [f(y)-1] \frac{\theta e^{-y} y^{\frac{\theta}{K}-1}}{\Gamma(\frac{\theta}{K}+1)} dy\right)^K \\ &\xrightarrow[K\to\infty]{} \exp\left(\int_0^\infty [f(y)-1)] \frac{\theta e^{-y}}{y} dy\right). \end{aligned} \tag{4.16}$$

For the Poisson-Dirichlet process with parameter θ, from Definition 4.27 and (4.15), we know that its probability generating functional is given by

$$G[f] = \exp\left(\int_0^\infty (f(y)-1) \frac{\theta e^{-y}}{y} dy\right),$$

which is identical to (4.16). Since the probability generating functional completely characterizes a Poisson point process, we get that in the limit as $K \to \infty$, the sequence $\{Y_i\}$ converges to a Poisson-Dirichlet process with parameter θ, and hence the distribution of the decreasing order statistics of $(X_1, \ldots, X_K)$ approaches PD(θ).

Remark 4.29. Let $\{X_k\}$ be a sequence drawn from GEM(θ), and $\{X_{(k)}\}$ be the decreasing order statistics of $\{X_k\}$. Then $\{X_{(k)}\}$ is distributed as PD(θ). This can be directly seen from Kingman's characterization of the Ewens sampling formula (Kingman, 1977), since by Theorem 4.25, we know that drawing population frequencies according to GEM(θ) generates Ewens sampling formula.

References

Aldous DJ (1985) Exchangeability and related topics. In: Hennequin P (ed) École d'Été de Probabilités de Saint-Flour XIII — 1983, Lecture Notes in Mathematics, vol 1117, Springer Berlin Heidelberg, pp 1–198

Arratia A, Barbour AD, Tavaré S (2003) Logarithmic Combinatorial Structures: A Probabilistic Approach. European Mathematical Society Publishing House, Switzerland

Billingsley P (2008) Probability and Measure. John Wiley & Sons

Cox D, Isham V (1980) Point processes. Monographs on applied probability and statistics, Chapman and Hall

Crane H (2016) The ubiquitous Ewens sampling formula. Statistical Science 31(1):1–19

Engen S (1978) Stochastic abundance models, with emphasis on biological communities and species diversity. Monographs on applied probability and statistics, Chapman and Hall, London

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theoretical Population Biology 3:87–112

Goncharov VL (1944) Some facts from combinatorics. Izvestia Akademii Nauk SSSR, Ser Mat 8:3–48

Griffiths RC (1979) Exact sampling distributions from the infinite neutral alleles model. Advances in Applied Probability 11(2):326–354

Hoppe F (1984) Pólya-like urns and the Ewens' sampling formula. J Math Biol 20:91–94

Joyce P, Tavaré S (1987) Cycles, permutations and the structure of the yule process with immigration. Stochastic processes and their applications 25:309–314

Karlin S, McGregor J (1972) Addendum to a paper of w. ewens. Theoretical Population Biology 3(1):113–116

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49(4):725–738

Kingman JFC (1977) The population structure associated with the Ewens sampling formula. Theoretical Population Biology 11(2):274 – 283

McCloskey JW (1965) A model for the distribution of individuals by species in an environment. PhD thesis, Michigan State University, East Lansing, Michigan

Tavaré S (1984) Line-of-descent and genealogical processes, and their applications in population genetics models. Theoretical Population Biology 26:119–164

Chapter 5
Infinite-sites model of mutation

As we saw in the previous chapter, the infinite-alleles model has several interesting connections with other probabilistic models and various closed-form results can be derived. However, one notable drawback is that there is no notion of closeness between different alleles in the infinite-alleles model. That is, given a pair of alleles, we are able to tell whether they are the same or not, but we cannot say how closely related they are. Hence, this model is too coarse for describing the pattern of shared mutations in a collection of DNA sequences. As we will see presently, however, what we learned about the infinite-alleles model in the previous chapter is still very useful for studying some important properties of more realistic mutation models. In this chapter, we consider a widely used model called the infinite-sites model of mutation, in which each genome is assumed to be infinitely long and every mutation occurs at a unique genomic position that has never mutated previously.

5.1 Model description

As usual, mutations arrive according to a Poisson point process with intensity $\theta/2$ on each edge independently of all other edges. Similar to the infinite-alleles case discussed in Chapter 4.2, every time a mutation occurs, we label it by a new random number drawn from Unif$[0, 1]$. In the infinite-alleles model, each leaf was assigned the label of the most recent mutation encountered when the ancestral lineage of the leaf is followed backward in time. In contrast, in the infinite-sites model, each leaf is assigned the unit interval $[0, 1]$ marked by the mutations encountered in the path from the leaf to the root of the tree. Such a marked interval assigned to a leaf is referred to as a *haplotype*. See Figure 5.1 for an illustration. Since the genomic location of each mutation is drawn independently from Unif$([0, 1])$, note that every mutation arises at a unique location almost surely.

Definition 5.1 (Segregating site). A position in the genome is said to be *segregating* (or polymorphic) if not every individual in the sample has the same allele. The total number of segregating sites in a sample of size n is denoted by S_n.

In the example shown in Figure 5.1, $n = 6$ and $S_n = 7$. Note that $S_n \mid L_n \sim \text{Poisson}(\frac{\theta}{2} L_n)$, where L_n is the tree length defined in Definition 1.11. Therefore, $\mathbb{E}(S_n) = \mathbb{E}[\mathbb{E}(S_n \mid L_n)] = \mathbb{E}(\frac{\theta}{2} L_n)$, which leads to the following result:

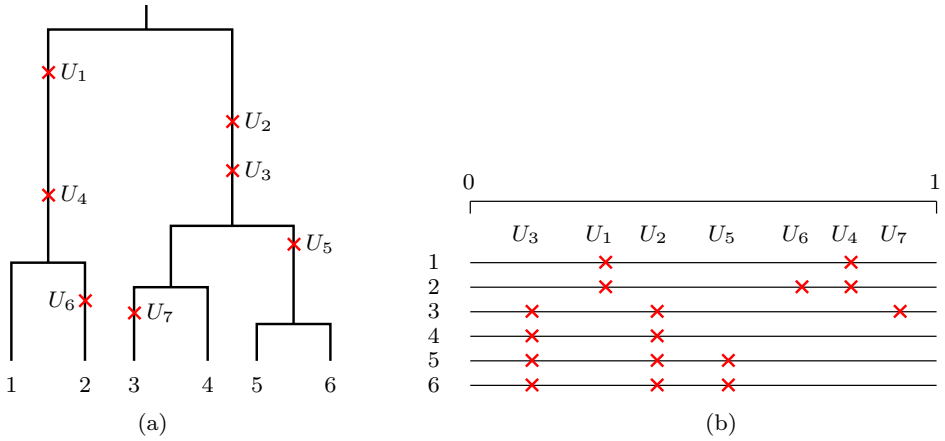


Fig. 5.1: Illustration of generating a sample under the infinite-sites model of mutation. (a) At each mutation event, a random number U_i is drawn from Unif$[0,1]$. (b) Each leaf is assigned the unit interval $[0,1]$ marked by the mutations encountered in the path from the leaf to the root of the tree.

Proposition 5.2 (Expected number of segregating sites). *For a sample of size n, the mean number of segregating sites is given by*

$$\mathbb{E}(S_n) = \frac{\theta}{2} \sum_{k=2}^{n} k\, \mathbb{E}[T_{n,k}], \tag{5.1}$$

where $T_{n,k}$ is the waiting time while there are k lineages.

Under a constant population size, $\mathbb{E}(T_{n,k}) = \frac{1}{\binom{k}{2}}$ and (5.1) simplifies to

$$\mathbb{E}(S_n) = \theta \sum_{j=1}^{n-1} \frac{1}{j}.$$

As we will see below, it will be useful to consider the following quantity:

Definition 5.3 (Unnormalized site frequency spectrum). Let $\zeta_{n,b}$ denote the number of segregating sites each with b derived alleles in a sample of n sequences. The vector $(\zeta_{n,1}, \ldots, \zeta_{n,n-1})$ is called the unnormalized *site frequency spectrum* (SFS).

Note that $S_n = \sum_{b=1}^{n-1} \zeta_{n,b}$. For the example shown in Figure 5.1, $\zeta_{6,1} = 2, \zeta_{6,2} = 3, \zeta_{6,4} = 2$, and $\zeta_{6,b} = 0$ for all other values of b. When there is ambiguity as to which allelic type is ancestral and which is derived, a "folded" version is often used:

Definition 5.4 (Folded site frequency spectrum). For $i \in \{1, 2, \ldots, \lfloor \frac{n}{2} \rfloor\}$, we define

$$\eta_{n,i} = \frac{\zeta_{n,i} + \zeta_{n,n-i}}{1 + \delta_{i,n-i}}.$$

Note that the vector $(\eta_{n,1}, \ldots, \eta_{n,\lfloor \frac{n}{2} \rfloor})$ is obtained by folding $(\zeta_{n,1}, \ldots, \zeta_{n,n-1})$ in half.

By conditioning on polymorphism, we obtain a normalized version of the SFS:

Definition 5.5 (Normalized Site Frequency Spectrum). Let $q_{n,b}$ denote the conditional probability that, at a particular site, b alleles in a sample of size n are derived alleles, conditional on the site segregating. The vector $(q_{n,1}, \dots, q_{n,n-1})$ is called the normalized site frequency spectrum.

5.2 Connections with the infinite-alleles model

In this section, we apply what we know about the infinite-alleles model to obtain some useful results for the infinite-sites model.

Proposition 5.6 (Expected number of distinct haplotypes). *Let H_n denote the number of distinct haplotypes in a sample of size n. Then,*

$$\mathbb{E}(H_n) = 1 + \sum_{j=1}^{n-1} \frac{\theta}{j+\theta}. \tag{5.2}$$

Proof. Note that H_n is equal to K_n in the infinite-alleles model. For example, compare Figure 5.1 with Figure 4.1. Hence, (5.2) follows from (4.11). □

Since generating a new haplotype requires a mutation, while not all mutations increase the number of distinct haplotypes, we have $S_n \geq H_n - 1$. Furthermore, note that

$$\mathbb{E}[S_n - (H_n - 1)] = \theta^2 \sum_{j=1}^{n-1} \frac{1}{j(j+\theta)}, \tag{5.3}$$

which is small if θ is small.

Proposition 5.7 (Expected number of uniquely represented haplotypes). *Let $H_{n,1}$ denote the number of distinct haplotypes represented exactly once in a sample of size n. Then,*

$$\mathbb{E}(H_{n,1}) = \frac{n\theta}{n-1+\theta}. \tag{5.4}$$

Proof. The easiest way to prove this result is by using the Chinese Restaurant Process and exchangeability. The probability that the first customer is sitting by themself after all n customers have seated is

$$\begin{aligned}
&\left(1 - \frac{1}{1+\theta}\right)\left(1 - \frac{1}{2+\theta}\right)\left(1 - \frac{1}{3+\theta}\right)\cdots\left(1 - \frac{1}{n-1+\theta}\right) \\
&= \left(\frac{\theta}{1+\theta}\right)\left(\frac{1+\theta}{2+\theta}\right)\left(\frac{2+\theta}{3+\theta}\right)\cdots\left(\frac{n-2+\theta}{n-1+\theta}\right) \\
&= \left(\frac{\theta}{n-1+\theta}\right),
\end{aligned}$$

where the last line follows from telescoping. By exchangeability, the probability that any customer $j \in [n]$ is sitting by themself after all n customers have seated is also $\left(\frac{\theta}{n-1+\theta}\right)$. Then (5.4) follows from $H_{n,1} = \sum_{j=1}^{n} \mathbb{I}\{j \text{ is sitting by themself}\}$. □

In the coalescent with killing, call the last one to be killed the ancestral allele. In Figure 4.3, leaves labeled $4, 5, 6$ have the ancestral allele by this definition. Non-ancestral alleles are called *derived* (or mutant) alleles.

Theorem 5.8 (Polarized sampling probability). *Consider an infinite-alleles model with mutation rate $\theta/2$, and assume that the ancestral allelic type is known. Suppose a sample of size n consists of n_d derived alleles and $n - n_d$ ancestral alleles. Then, the probability $\mathbb{P}_n^\theta(n_d = b)$ of observing such a sample is*

$$\mathbb{P}_n^\theta(n_d = b) = \frac{(n-1)!}{b!} \frac{(\theta)_{b\uparrow}}{(1+\theta)_{(n-1)\uparrow}}.$$

Proof. We use the duality between Hoppe's urn model and the infinite-alleles model. The first draw is the black ball with probability 1. The new ball returned to the urn together with the black ball is designated as the ancestral type. At step j of the urn model, $j-1$ non-black balls have already been placed in the urn, so the total mass in the urn is $j-1+\theta$. Suppose the urn has $n_a^{(j-1)}$ ancestral balls and $n_d^{(j-1)}$ mutant balls, where $n_a^{(j-1)} + n_d^{(j-1)} = j-1$. Then, the probability of picking the ancestral type is $\frac{n_a^{(j-1)}}{j-1+\theta}$, while the probability of picking a mutant (non-ancestral) type or the black ball is $\frac{n_d^{(j-1)}+\theta}{j-1+\theta}$. Hence, a particular sequence of draws leading to n_a ancestral types and $n_d = n - n_a$ mutation types in the urn has probability

$$\frac{(\prod_{i=1}^{n_a-1} i)[\prod_{j=0}^{n_d-1}(j+\theta)]}{\prod_{k=1}^{n-1}(k+\theta)}. \tag{5.5}$$

Since there are $n-1$ draws after the ancestral type is first introduced to the urn and we need $n_a - 1$ of those steps to be drawing an ancestral ball, we obtain $\binom{n-1}{n_a-1}$ distinct sequences of draws. Multiplying this combinatorial factor with (5.5) gives the desired result. □

If the mutation rate is small, we may want to consider a first-order approximation to the above sampling formula. Doing a Taylor expansion of (5.5) about $\theta = 0$, the probability of observing $b \geq 1$ derived alleles is $\frac{(n-1)!}{b!}\frac{(b-1)!}{(n-1)!}\theta + O(\theta^2)$, which simplifies to

$$\mathbb{P}_n^\theta(n_d = b) = \frac{\theta}{b} + O(\theta^2). \tag{5.6}$$

While this is a nice formula, perhaps there are models of mutation that better describe the biological scenario of long sequences with low per-base mutation rates. One such a model is the infinite-sites model of mutation.

Consider now a sample of n sequences, each with L sites at which mutations arise. Assume that the total mutation rate for the entire region is $\frac{\theta}{2}$. The n sequences are related by a single coalescent tree and mutations are placed on that tree as a Poisson point process with rate $\frac{\theta}{2}\frac{1}{L}$ for each of the L sites according to the infinite alleles model. Note that this is equivalent to placing mutations on the genealogy at rate $\frac{\theta}{2}$ and then having each mutation hit one of the L sites uniformly at random. If we take the limit of $L \to \infty$, we arrive at the infinite-sites model of mutation.

We want to compute $\mathbb{E}(\zeta_{n,b})$, which will illustrate the computational convenience of taking L to ∞. Let n_d^ℓ denote the number of sequences with a derived allele at site ℓ. Then, we can see, using (5.6) and summing over all L sites,

$$\mathbb{E}(\zeta_{n,b}) = \sum_{\ell=1}^{L} \mathbb{P}_n^\theta(n_d^\ell = b) = \sum_{\ell=1}^{L} \left[\frac{\theta}{L}\frac{1}{b} + O\left(\frac{\theta^2}{L^2}\right)\right] = \frac{\theta}{b} + O\left(\frac{\theta^2}{L}\right). \tag{5.7}$$

In the limit as $L \to \infty$, the normalized SFS can be obtained as

$$q_{n,b} = \lim_{L\to\infty} \mathbb{P}_n^\theta(n_d = b \mid n_d > 0) = \lim_{L\to\infty} \frac{\mathbb{P}_n^\theta(n_d = b)}{\mathbb{P}_n^\theta(n_d > 0)} = \frac{\frac{1}{b}}{\sum_{j=1}^{n-1} \frac{1}{j}}. \tag{5.8}$$

5.3 Site frequency spectrum (SFS) under the infinite-sites model

In this section, we obtain more general formulas for the unnormalized and normalized site frequency spectra under the infinite-sites model. These formulas will hold for arbitrary population size functions.

Definition 5.9. Let $\tau_{n,b}$ denote the sum of the lengths of all edges each subtending exactly b leaves. See Figure 5.2 for an example.

Theorem 5.10. *For $b \in \{1, \dots, n-1\}$,*

$$\mathbb{E}(\tau_{n,b}) = \sum_{k=2}^{n} \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}} k\mathbb{E}(T_{n,k}). \tag{5.9}$$

Proof. Note that $\tau_{n,b}$ be can be decomposed into epochs while there are k lineages:

$$\tau_{n,b} = \sum_{k=2}^{n} \sum_{e\in E_k} \mathbb{I}\{\text{edge } e \text{ subtends exactly } b \text{ leaves}\} T_{n,k},$$

where E_k denotes the set of edges in the epoch while there are k lineages and, as usual, $T_{n,k}$ denotes the waiting time while there are k lineages. See Figure 5.2b for an illustration. Now, note that the random variables $\mathbb{I}\{\text{edge } e \text{ subtends exactly } b \text{ leaves}\}$ and $T_{n,k}$ are independent, so $\mathbb{E}(\mathbb{I}\{\text{edge } e \text{ subtends exactly } b \text{ leaves}\}T_{n,k})$ can be factorized as $\mathbb{E}(\mathbb{I}\{\text{edge } e \text{ subtends exactly } b \text{ leaves}\}\mathbb{E}(T_{n,k})$. Furthermore, from Corollary 2.6, we have

$$\mathbb{P}\{\text{edge } e \text{ subtends exactly } b \text{ leaves}\} = \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}}.$$

Since E_k contains k edges, (5.9) then follows. □

For $b \in \{1, \dots, n-1\}$, since $\zeta_{n,b}|\tau_{n,b} \sim \text{Poisson}(\frac{\theta}{2}\tau_{n,b})$, the unnormalized SFS entries are given by

$$\mathbb{E}(\zeta_{n,b}) = \mathbb{E}(\mathbb{E}(\zeta_{n,b} \mid \tau_{n,b})) = \frac{\theta}{2}\mathbb{E}(\tau_{n,b}).$$

For a constant population size, $\mathbb{E}(T_{n,k}) = 1/\binom{k}{2}$, and plugging this into (5.9) and simplifying gives $\mathbb{E}(\tau_{n,b}) = \frac{2}{b}$, which implies $\mathbb{E}(\zeta_{n,b}) = \frac{\theta}{b}$. This result agrees with the leading order term in (5.7). Entries of the normalized SFS, for $b \in \{1, \dots, n-1\}$, are given by

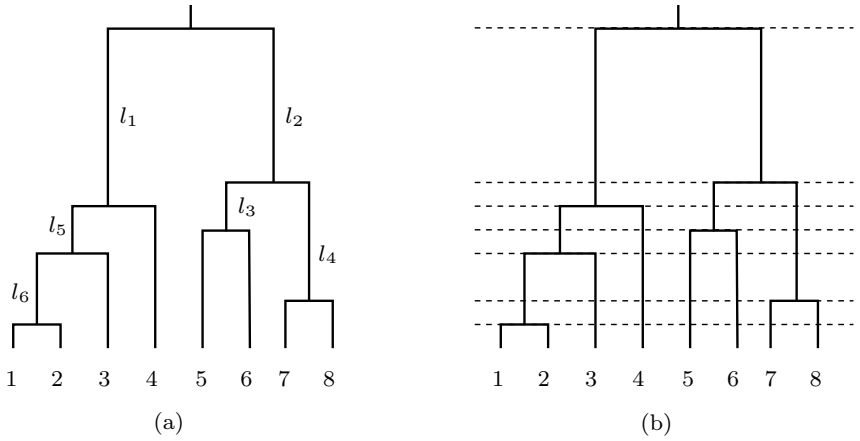


Fig. 5.2: A coalescent tree with interior edges labeled with their lengths. (a) For this particular tree, $\tau_{8,5} = \tau_{8,6} = \tau_{8,7} = 0$, while $\tau_{8,2} = l_3 + l_4 + l_6$, $\tau_{8,3} = l_5$, $\tau_{8,4} = l_1 + l_2$, and $\tau_{8,1} =$ sum of the lengths of all pendant edges. (b) The same tree sliced by epochs while there are k lineages, for $k = n, n-1, \ldots, 2$.

$$q_{n,b} = \frac{\mathbb{E}(\zeta_{n,b})}{\sum_{j=1}^{n-1} \mathbb{E}(\zeta_{n,j})} = \frac{\mathbb{E}(\zeta_{n,b})}{\mathbb{E}(S_n)}, \tag{5.10}$$

which agrees with (5.8) in the case of a constant population size.

5.4 A warning on conditioning on the number of segregating sites

Sometimes people simulate sequence data under the coalescent while conditioning on the number of segregating sites. One should note that the SFS simulated in this way has a distribution different from $q_{n,b}$. To illustrate this point, suppose you simulate a random 3-leaved coalescent tree and drop a single mutation on the tree uniformly at random. This will lead to either a *singleton* site with one derived allele or a *doubleton* site with two derived alleles. Suppose you repeat this experiment many times to estimate the frequency of observing a doubleton site. What is the result you expect to obtain? In fact, this example is simple enough to find the exact answer analytically. Given the inter-coalescence times $T_{3,3}$ and $T_{3,2}$ of a sampled tree, the conditional probability that the mutation leads to a doubleton site is $T_{3,2}/(3T_{3,3} + 2T_{3,2})$, where the denominator corresponds to the total length of the tree. Furthermore, since $T_{3,3}$ and $T_{3,2}$ are independent when the population size remains constant over time, $\mathbb{E}[T_{3,2}/(3T_{3,3} + 2T_{3,2})]$ can be easily evaluated as

$$\mathbb{E}\left[\frac{T_{3,2}}{3T_{3,3} + 2T_{3,2}}\right] = \int_0^\infty \int_0^\infty \frac{t_2}{3t_3 + 2t_2} 3e^{-3t_3} e^{-t_2} dt_3 dt_2 = 1 - \log 2 \approx 0.307. \tag{5.11}$$

In contrast, we know from (5.8) that $q_{3,2} = 1/3 \approx 0.333$. The key difference is that, as can be seen in (5.10), $q_{n,b}$ is a ratio of the expectations $\mathbb{E}[T_{3,2}]$ and $\mathbb{E}[3T_{3,3} + 2T_{3,2}]$, while (5.11) is an expectation of the ratio $\frac{T_{3,2}}{3T_{3,3}+2T_{3,2}}$.

5.5 The age of a mutation

We present here a general result concerning the age of a mutation under the infinite-sites model, first obtained by Griffiths and Tavaré (1998).

Theorem 5.11 (The age of a mutation). *Assume the infinite-sites model of mutation, and let $\mathcal{M}_{n,b}$ denote the event that a site is segregating with b mutant alleles and $n-b$ ancestral alleles. Let A denote the age of the corresponding mutation (i.e., how long back in time it arose). Then,*

$$\mathbb{E}(A \mid \mathcal{M}_{n,b}) = \frac{\sum_{k=2}^{n} k p_{n,k}(b) \mathbb{E}[T_{n,k}(\frac{1}{2}T_{n,k} + T_{n,k+1} + \cdots + T_{n,n})]}{\sum_{j=2}^{n} j p_{n,j}(b) \mathbb{E}(T_{n,j})}, \tag{5.12}$$

where $p_{n,k}(b)$ is defined as

$$p_{n,k}(b) = \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}}.$$

Proof. Assume that our chromosome is $[0,1]$ and that mutations are located in this interval according to the uniform distribution. Let $M(x, \delta x, b)$ be the event that there is a segregating site with b mutant alleles in $(x, x+\delta x)$, and let $M_k(x, \delta x, b)$ be the event of there being such a segregating site, and that it arose when there were k lineages. Define $\boldsymbol{T} = (T_{n,2}, T_{n,3}, \ldots, T_{n,n})$ and let $A(x, \delta x)$ be the age of the first mutation in $(x, x+\delta x)$.

Conditional on $\boldsymbol{T}$, the number of mutations in $(x, x+\delta x)$ is Poisson with rate proportional to δx, so the probability of more than one mutation is $O((\delta x)^2)$. Therefore,

$$\mathbb{P}[M(x, \delta x, b) \mid \boldsymbol{T}] = \sum_{k=2}^{n} \mathbb{P}[M_k(x, \delta x, b) \mid \boldsymbol{T}] + o(\delta x) = \sum_{k=2}^{n} \frac{\theta \delta x}{2} k T_{n,k} p_{n,k}(b) + o(\delta x).$$

This follows from the fact that the number of mutations when there are k lineages is Poisson with rate $\frac{\theta \delta x}{2} k T_{n,k}$, so the probability of getting such a mutation is $\frac{\theta \delta x}{2} k T_{n,k} + o(\delta x)$. Once we have a mutation, the probability that it has b descendants is given by $p_{n,k}(b)$.

By the same argument,

$$\begin{aligned}
\mathbb{E}[A(x, \delta x) \cdot \mathbb{I}_{M(x,\delta x,b)} \mid \boldsymbol{T}] &= \sum_{k=2}^{n} \mathbb{E}[A(x, \delta x) \cdot \mathbb{I}_{M_k(x,\delta x,b)} \mid \boldsymbol{T}] + o(\delta x) \\
&= \sum_{k=2}^{n} \mathbb{E}[A(x, \delta x) \mid M_k(x, \delta x, b), \boldsymbol{T}] \; \mathbb{P}[M_k(x, \delta x, b) \mid \boldsymbol{T}] + o(\delta x) \\
&= \sum_{k=2}^{n} \left(T_{n,n} + \cdots + T_{n,k+1} + \frac{1}{2} T_{n,k} \right) \frac{\theta \delta x}{2} k T_{n,k} p_{n,k}(b) + o(\delta x).
\end{aligned}$$

Since mutations arise according to a Poisson point process, given that a mutation occurs in the time interval $T_{n,k}$, its position is uniformly distributed in the interval. This explains the factor of $1/2$ in front of $T_{n,k}$.

Finally, using the tower property to find $\mathbb{E}[A(x, \delta x) \mathbb{I}_{M(x,\delta x,b)}]$, we obtain

$$\begin{aligned}\mathbb{E}[A(x,\delta x) \mid M(x,\delta x,b)] &= \frac{\mathbb{E}[A(x,\delta x)\mathbb{I}_{M(x,\delta x,b)}]}{\mathbb{P}[M(x,\delta x,b)]} \\ &= \frac{\sum_k kp_{n,k}(b)\mathbb{E}[(T_{n,n}+\cdots+T_{n,k+1}+\frac{1}{2}T_{n,k})T_{n,k}]+o(\delta x)}{\sum_k kp_{n,k}(b)\mathbb{E}[T_{n,k}]+o(\delta x)},\end{aligned}$$

and letting $\delta x \to 0$ yields the desired result. □

Griffiths and Tavaré (1998) actually obtained stronger results; they obtained the moment generating function as well as the probability density of the mutation age. The expression in (5.12) simplifies considerably for the case of a constant population size.

Corollary 5.12. *In the case of a constant population size,* (5.12) *simplies to*

$$\mathbb{E}(A \mid \mathcal{M}_{n,b}) = \frac{2b}{n-b}\sum_{k=b+1}^{n}\frac{1}{k}. \tag{5.13}$$

Proof. In the case of a constant population size, the denominator of (5.12) is given by $\sum_{k=2}^{n} kp_{n,k}(b)\mathbb{E}(T_{n,k}) = \frac{2}{b}$. The numerator of (5.12) can be computed as follows. First,

$$\begin{aligned}\mathbb{E}\left[T_{n,k}\left(\frac{1}{2}T_{n,k}+T_{n,k+1}+\cdots+T_{n,n}\right)\right] &= \frac{1}{2}\frac{2}{\binom{k}{2}^2}+\frac{1}{\binom{k}{2}}\sum_{j=k+1}^{n}\frac{1}{\binom{j}{2}} = \frac{1}{\binom{k}{2}}\sum_{j=k}^{n}\frac{1}{\binom{j}{2}} \\ &= \frac{1}{\binom{k}{2}}\sum_{j=k}^{n}2\left[\frac{1}{j-1}-\frac{1}{j}\right] = \frac{2}{\binom{k}{2}}\left[\frac{1}{k-1}-\frac{1}{n}\right].\end{aligned}$$

Now, since

$$kp_{n,k}(b) = k\frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}} = \frac{k(k-1)}{b}\frac{\binom{n-k}{b-1}}{\binom{n-1}{b}},$$

we obtain

$$\begin{aligned}\mathbb{E}(A \mid \mathcal{M}_{n,b}) &= \frac{b}{2}\sum_{k=2}^{n}\frac{k(k-1)}{b}\frac{\binom{n-k}{b-1}}{\binom{n-1}{b}}\times\frac{2}{\binom{k}{2}}\frac{n-k+1}{n(k-1)} \\ &= \frac{2}{\binom{n-1}{b}}\sum_{k=2}^{n}\frac{n-k+1}{n(k-1)}\binom{n-k}{b-1} \\ &= \frac{2b}{n-b}\sum_{k=2}^{n-b+1}\frac{\binom{n-b}{k-1}}{\binom{n}{k-1}}\frac{1}{k-1} = \frac{2b}{n-b}\sum_{k=1}^{n-b}\frac{\binom{n-b}{k}}{\binom{n}{k}}\frac{1}{k}.\end{aligned}$$

To simplify the above expression, consider a sequence of numbers $c_{0,0}, c_{0,1}, \ldots, c_{0,m}$. Let $c_{k-1,j+1} = c_{k-1,j} + c_{k,j}$ for $k = 1,\ldots,m$ and $j = 0,\ldots,m-k$. Then,

$$\sum_{k=0}^{m}\binom{m}{k}c_{k,0} = c_{0,m}. \tag{5.14}$$

If $c_{0,0} = 0$ and $c_{0,j} = \sum_{k=0}^{j-1}\frac{1}{n-j}$ for $j = 1,\ldots,n-b$, one can show that $c_{k,0} = \frac{1}{k\binom{n}{k}}$. Hence, with $m = n-b$ in (5.14), the left hand side is

$$\sum_{k=0}^{n-b} \binom{n-b}{k} c_{k,0} = \sum_{k=0}^{n-b} \binom{n-b}{k} \frac{1}{\binom{n}{k} k},$$

while the right hand side is

$$c_{0,n-b} = \sum_{k=0}^{n-b-1} \frac{1}{n-j} = \sum_{k=b+1}^{n} \frac{1}{k}.$$

We therefore conclude that

$$\mathbb{E}(A \mid \mathcal{M}_{n,b}) = \frac{2b}{n-b} \sum_{k=b+1}^{n} \frac{1}{k},$$

which proves the claim. □

We now use the above result to reproduce a classical result in population genetics. From (5.13), we have

$$\begin{aligned}
\mathbb{E}(A \mid \mathcal{M}_{n,b}) &= \frac{2b/n}{1-b/n} \left[\left(\sum_{k=1}^{n} \frac{1}{k} \right) - \left(\sum_{j=1}^{n(b/n)} \frac{1}{j} \right) \right] \\
&= \frac{2b/n}{1-b/n} \left[\left(\sum_{k=1}^{n} \frac{1}{k} - \log(n) \right) - \left(\sum_{j=1}^{n(b/n)} \frac{1}{j} - \log(n(b/n)) \right) + \log(n) - \log(b) \right].
\end{aligned}$$

Therefore, as $n \to \infty, b \to \infty$, and $b/n \to x$,

$$\mathbb{E}(A \mid \mathcal{M}_{n,b}) \to \frac{2x}{1-x}[\gamma_E - \gamma_E - \log(x)] = -\frac{2x}{1-x} \log(x),$$

where γ_E denotes the Euler constant. This corresponds to the population-wide expectation of the age of a neutral mutation, a result obtained by Kimura and Ohta (1973) using the diffusion theory.

5.6 Unbiased moment estimators of θ

In Chapter 4.7, we saw that the number K_n of distinct alleles in a sample of size n is a sufficient statistic for θ under the infinite-alleles model. Further, we saw that the moment estimator based on $\mathbb{E}(K_n)$ is equal to the maximum likelihood estimate. In the infinite-sites model, the number of distinct haplotypes is *not* a sufficient statistic for θ; in fact, no simple sufficient statistic for θ is known. In this section, we discuss several classical unbiased moment estimators of the mutation rate.

Definition 5.13 (Watterson's estimator of θ). Watterson (1975) suggested the following estimator of θ:

$$\hat{\theta}_W = \frac{2S_n}{\sum_{k=2}^{n} k\mathbb{E}[T_{n,k}]}, \tag{5.15}$$

where S_n denotes the observed number of segregating sites in a sample of size n.

Under a constant population size model, for which $\mathbb{E}[T_{n,k}] = 1/\binom{k}{2}$, Watterson's estimator becomes $\hat{\theta}_W = S_n / \sum_{k=1}^{n-1} \frac{1}{k}$. Note that (5.1) immediately implies that (5.15) satisfies $\mathbb{E}(\hat{\theta}_W) = \theta$, so this estimator is unbiased. Furthermore, it turns out that $\hat{\theta}_W$ is a weakly consistent estimator; e.g., under a constant population size model,

$$\mathrm{Var}(\hat{\theta}_W) = \frac{\sum_{j=1}^{n-1} \left[\frac{\theta}{j} + \left(\frac{\theta}{j} \right)^2 \right]}{\left[\sum_{j=1}^{n-1} \frac{1}{j} \right]^2} \to 0 \quad \text{as} \quad n \to \infty,$$

which implies $\hat{\theta}_W \xrightarrow{p} \theta$ as $n \to \infty$, which can be shown using Chebyshev's inequality.

Definition 5.14 (Tajima's estimator of θ). Tajima (1983) proposed the following estimator of θ:

$$\hat{\theta}_T = \frac{1}{\binom{n}{2} \mathbb{E}[T_{2,2}]} \sum_{i<j} \Pi_{i,j},$$

where $\Pi_{i,j}$ denote the Hamming distance between haplotypes i and j; i.e., the number of differences between them.

Note that

$$\mathbb{E}(\hat{\theta}_T) = \frac{1}{\binom{n}{2} \mathbb{E}[T_{2,2}]} \sum_{i<j} \mathbb{E}(\Pi_{i,j}) = \frac{\mathbb{E}(\Pi_{1,2})}{\mathbb{E}[T_{2,2}]} = \frac{\mathbb{E}(S_2)}{\mathbb{E}[T_{2,2}]} = \theta,$$

so Tajima's estimator is also unbiased.

Definition 5.15 (Fu and Li's estimator of θ). For $n > 2$, Fu and Li (1993) proposed the following estimator of θ:

$$\hat{\theta}_{FL} = \frac{2\eta_{n,1}}{\mathbb{E}(\tau_{n,1}) + \mathbb{E}(\tau_{n,n-1})},$$

where $\eta_{n,1}$ denotes the number of folded singleton sites (cf., Definition 5.4), and $\tau_{n,b}$ denotes the total length of all edges each subtending exactly b leaves in a coalescent tree with n leaves (cf., Definition 5.9).

For a constant population size model, we have $\mathbb{E}(\tau_{n,k}) = \frac{2}{k}$, so $\hat{\theta}_{FL} = \frac{n-1}{n}\eta_{n,1}$. Note that $\mathbb{E}(\hat{\theta}_{FL}) = \theta$ for $n > 2$. Neither $\hat{\theta}_T$ nor $\hat{\theta}_{FL}$ converges to θ as $n \to \infty$. Moreover, $\hat{\theta}_{FL}$ is not very reliable, since $\eta_{n,1}$ is rather sensitive to sequencing errors.

It turns out that many estimators of θ can be written as a linear combination of the entries $\zeta_{n,k}$ of the unnormalized SFS defined in Definition 5.3: i.e., $\hat{\theta} = \sum_{k=1}^{n-1} a_{n,k}\zeta_{n,k}$, where $a_{n,k}$ are constants. For example, the estimators described above can be written as follows:

$$\hat{\theta}_W = \frac{2}{\sum_{j=2}^{n} j\mathbb{E}[T_{n,j}]} \sum_{k=1}^{n-1} \zeta_{n,k},$$
$$\hat{\theta}_T = \frac{1}{\binom{n}{2}\mathbb{E}[T_{2,2}]} \sum_{k=1}^{n-1} k(n-k)\zeta_{n,k},$$
$$\hat{\theta}_{FL} = \frac{2}{\mathbb{E}(\tau_{n,1}) + \mathbb{E}(\tau_{n,n-1})}(\zeta_{n,1} + \zeta_{n,n-1}).$$

A natural question to ask is, which linear estimator is better? Fu (1994) showed that Watterson's estimator $\hat{\theta}_W$ is approximately the best linear unbiased estimator (BLUE) for small θ. However, Futschik and Gach (2008) showed that $\hat{\theta}_W$ is inadmissible and obtained a uniformly better, biased estimator using shrinkage under the mean squared error loss function.

The variance of an estimator of the form $\hat{\theta} = \sum_{k=1}^{n-1} a_{n,k}\zeta_{n,k}$ can be computed using the following result:

Theorem 5.16. *Suppose $\hat{\theta} = \sum_{k=1}^{n-1} a_{n,k}\zeta_{n,k}$. Then,*

$$\operatorname{Var}(\hat{\theta}) = \frac{\theta}{2} \sum_{k=1}^{n-1} a_{n,k}^2 \mathbb{E}(\tau_{n,k}) + \frac{\theta^2}{4} \sum_{j=1}^{n-1} \sum_{k=1}^{n-1} a_{n,j} a_{n,k} \operatorname{Cov}(\tau_{n,j}, \tau_{n,k}), \tag{5.16}$$

where $\tau_{n,b}$ is defined in Definition 5.9.

Proof. Consider the following decomposition of the variance:

$$\operatorname{Var}(\hat{\theta}) = \mathbb{E}(\operatorname{Var}(\hat{\theta}|\boldsymbol{\tau}_n)) + \operatorname{Var}(\mathbb{E}(\hat{\theta}|\boldsymbol{\tau}_n)),$$

where $\boldsymbol{\tau}_n = (\tau_{n,1}, \ldots, \tau_{n,n-1})$. Here, the first factor captures the "mutational part" (i.e., variance due to the placement of mutations on the tree) of the variance, while the second factor captures the "genealogical part" (i.e., variance due to the randomness in the genealogy). Now, noting that $\zeta_{n,1}, \ldots, \zeta_{n,n-1}$ are conditionally independent given $\boldsymbol{\tau}_n$ and that $\zeta_{n,k} \mid \tau_{n,k} \sim \text{Poisson}(\frac{\theta}{2}\tau_{n,k})$, we get

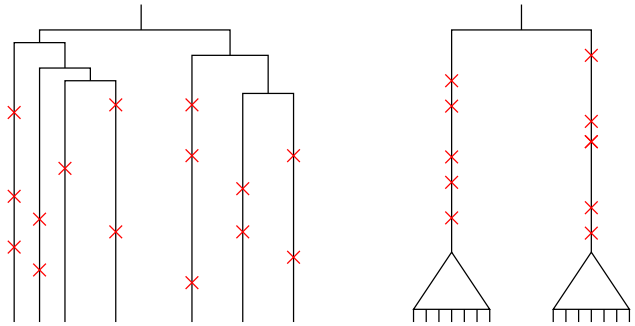
$$\operatorname{Var}(\hat{\theta}|\boldsymbol{\tau}_n) = \sum_{k=1}^{n-1} a_{n,k}^2 \operatorname{Var}(\zeta_{n,k}|\boldsymbol{\tau}_n) = \sum_{k=1}^{n-1} a_{n,k}^2 \operatorname{Var}(\zeta_{n,k}|\tau_{n,k}) = \sum_{k=1}^{n-1} a_{n,k}^2 \frac{\theta}{2}\tau_{n,k},$$

which explains the first term of (5.16). Using $\mathbb{E}(\sum_{k=1}^{n-1} a_{n,k}\zeta_{n,k}|\boldsymbol{\tau}_n) = \sum_{k=1}^{n-1} a_{n,k}\frac{\theta}{2}\tau_{n,k}$ and taking the variance leads to the second term of (5.16). □

Computing $\operatorname{Cov}(\tau_{n,j}, \tau_{n,k})$ involves second-order moments $\mathbb{E}(T_{n,j}T_{n,k})$ of inter-coalescence times. In the case of a constant population size,

$$\mathbb{E}(T_{n,j}T_{n,k}) = \frac{1 + \delta_{jk}}{\binom{j}{2}\binom{k}{2}},$$

and Fu (1995) used this result to derive a closed-form formula for $\operatorname{Cov}(\tau_{n,j}, \tau_{n,k})$. The work of Polanski et al (2003) can be used to obtain a general formula for $\mathbb{E}(T_{n,j}T_{n,k})$ under an arbitrary variable population size function, and Živković and Wiehe (2008) independently obtained a formula for the special case of piecewise-constant population size functions, but neither result is numerically stable for large values of n.



(a) (b)

Fig. 5.3: Two extreme examples. (a) Large $\tau_{n,1}$. (b) Long $T_{n,2}$, with about $n/2$ leaves on each subtree adjacent to the root.

5.7 Tests of selective neutrality

Deviations from the assumed model can have significantly different effects on different estimators of θ. Motivated by this observation, Tajima (1989) considered a test statistic defined as

$$D = \frac{\hat{\theta}_T - \hat{\theta}_W}{\sqrt{\mathrm{Var}(\hat{\theta}_T - \hat{\theta}_W)}}.$$

Under a constant population size model, $\mathbb{E}(D) = 0$, and Tajima used simulations to show that D approximately follows a rescaled Beta distribution and can be used to test for selective neutrality. He provided critical values for rejecting the null model (neutrality) at a significance level of α. For $\alpha = 5\%$, one rejects neutrality if $|D| > 2$. Consider the following two extreme examples to see what the statistic D is trying to capture:

1. (**Strong positive selection**) Consider the tree shape with long terminal branches ($\tau_{n,1}$), as shown in Figure 5.3a. In this case $\zeta_{n,1} \approx S_n$, so $\hat{\theta}_T \approx \frac{1}{\binom{n}{2}}(n-1)S_n = \frac{2}{n}S_n$. Since $\hat{\theta}_W \approx \frac{S_n}{\log(n-1)}$, we would expect $D < 0$.
2. (**Balancing selection**) Consider the tree shape with long internal branches ($T_{n,2}$) adjacent to the root, as shown in Figure 5.3b, with approximately $n/2$ leaves in each of left and right subtrees. In this case, $\hat{\theta}_T \approx \frac{1}{\binom{n}{2}}(\frac{n}{2})^2 S_n = \frac{n}{2(n-1)}S_n$. Since $\hat{\theta}_W \approx \frac{S_n}{\log(n-1)}$, we would expect $D > 0$.

Remark 5.17. Selectively neutral models with non-trivial population demography can also lead to tree shapes similar to those shown in Figure 5.3. For example, the first kind of tree shape can result from a severe bottleneck followed by a rapid expansion, while the second kind can result from population substructure.

Using their moment estimator θ_{FL}, Fu and Li (1993) considered the following statistics for testing neutrality:

$$D^* = \frac{\hat{\theta}_W - \hat{\theta}_{FL}}{\sqrt{\mathrm{Var}(\hat{\theta}_W - \hat{\theta}_{FL})}}, \qquad F^* = \frac{\hat{\theta}_T - \hat{\theta}_{FL}}{\sqrt{\mathrm{Var}(\hat{\theta}_T - \hat{\theta}_{FL})}}.$$

See their paper for further details on applications of these statistics. More recently, Achaz (2009) considered finding estimatores $\hat{\theta}_1 = \sum_{k=1}^{n-1} a_{n,k}^{(1)} \zeta_{n,k}$ and $\hat{\theta}_2 = \sum_{k=1}^{n-1} a_{n,k}^{(2)} \zeta_{n,k}$, such that the statistic $(\hat{\theta}_1 - \hat{\theta}_2)/\sqrt{\mathrm{Var}(\hat{\theta}_1 - \hat{\theta}_2)}$ provides more powerful tests of deviations from selective neutrality.

5.8 A direct method of computing the full likelihood

In Chapter 5.6, we discussed several unbiased moment estimators of the mutation parameter θ. Recall that they are all based on summary statistics of data (namely, the unnormalized frequency spectra $\zeta_{n,i}$). In what follows, we will discuss a full-likelihood method. The main advantage of such a method is that it improves the statistical efficiency of the estimate by utilizing as much of the information in the data as possible.

Given a data set X with n haplotypes generated under the infinite-sites model without recombination, our goal is to compute the likelihood $L(\theta) = p_n^\theta(X)$, also referred to as the sampling probability. As usual, let n denote the sample size, which is equal to the number of rows in X. We first consider a direct approach to computing $p_n^\theta(X)$. To be concrete, assume that the root sequence is the all-0 sequence and let 1 denote the derived allele at each segregating site. Consider the following simple example, where rows correspond to haplotypes and columns segregating sites:

$$X = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}. \tag{5.17}$$

The basic idea we employ is to compute the conditional probability of observing X given a coalescent tree topology, and then to sum over all possible topologies. For $n = 3$, there are three possible coalescent tree topologies. They are illustrated in Figure 5.4, where the leaf labels correspond to the labels of three distinct individuals in a population. The probability of each coalescent tree topology is $\frac{1}{3}$.

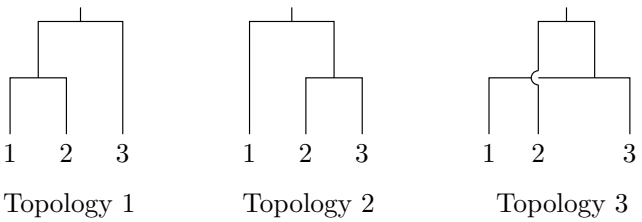


Fig. 5.4: The three possible coalescent tree topologies for X in (5.17). Each topology occurs with probability $\frac{1}{3}$.

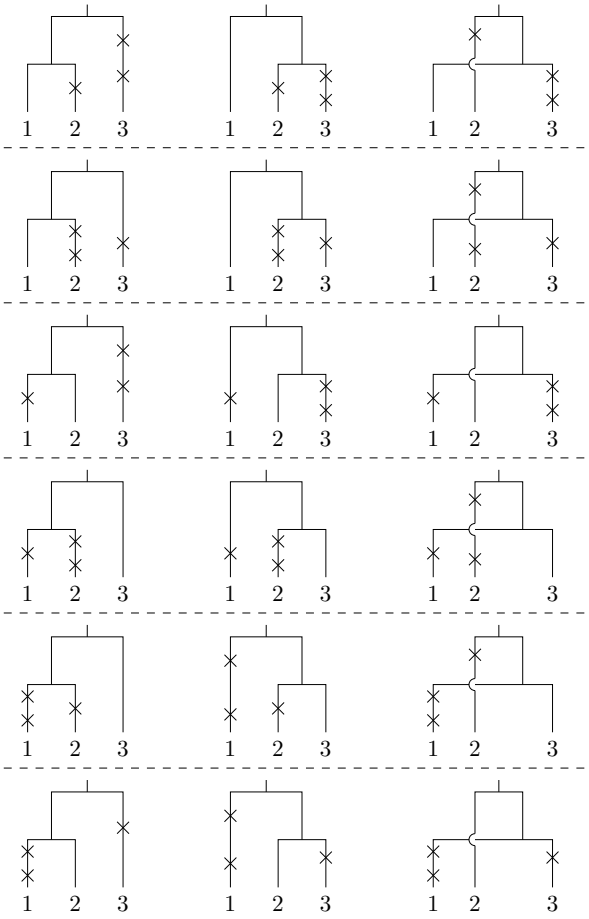


Fig. 5.5: Possible coalescent histories (topology + mutations) consistent with X in (5.17).

For each coalescent tree topology, there are six inequivalent ways to generate the data in (5.17), depending on which haplotype is associated with which individual. A complete list of coalescent histories (topology + mutations) for X is shown in Figure 5.5. Let h_i denote the ith row of X. In the coalescent histories shown in the top row of Figure 5.5, haplotype h_i is associated with leaf i. We use $q_n^\theta(X)$ to denote the probability of this "ordered sample." Note that each of the three coalescent histories in the top row of Figure 5.5 appears exactly 6 times in the full list, with permuted assignments of h_1, h_2, h_3 to leaves $1, 2, 3$. Therefore, $p_n^\theta(X) = 6 \times q_n^\theta(X)$.

Henceforward, assume a constant population size model. Using the fact that mutations occur according to a Poisson point process with rate $\theta/2$ for each lineage, independently of all other lineages, the conditional probability $q_n^\theta(X \mid \text{topology 1})$ can be computed as

$$\begin{aligned} q_n^\theta(X \mid \text{topology 1}) &= \mathbb{E}\left[e^{-\frac{\theta}{2}T_3}\left(\frac{\theta}{2}T_3\right)e^{-\frac{\theta}{2}T_3}\frac{1}{2!}\left[\frac{\theta}{2}(T_2+T_3)\right]^2 e^{-\frac{\theta}{2}(T_2+T_3)}e^{-\frac{\theta}{2}T_2}\right] \\ &= \mathbb{E}\left[\frac{\theta^3}{16}T_3(T_2+T_3)^2 e^{-\theta(T_2+\frac{3}{2}T_3)}\right] \\ &= \int_0^\infty\int_0^\infty\left[\frac{\theta^3}{16}t_3(t_2+t_3)^2 e^{-\theta(t_2+\frac{3}{2}t_3)}\right]e^{-t_2}3e^{-3t_3}\mathrm{d}t_2\,\mathrm{d}t_3 \\ &= \frac{\theta^3}{16}\frac{(24+32\theta+11\theta^2)}{(1+\theta)^3(2+\theta)^4}, \end{aligned}$$

where T_k denotes the waiting time while there are k lineages. The conditional probabilities $q_n^\theta(X \mid \text{topology 2})$ and $q_n^\theta(X \mid \text{topology 3})$ can be computed in a similar fashion:

$$q_n^\theta(X \mid \text{topology 2}) = \frac{\theta^3}{16}\mathbb{E}[T_3^3 e^{-\theta(T_2+\frac{3}{2}T_3)}] = \frac{\theta^3}{16}\frac{32}{9}\frac{1}{(1+\theta)(2+\theta)^4} \tag{5.18}$$

$$q_n^\theta(X \mid \text{topology 3}) = \frac{\theta^3}{16}\mathbb{E}[(T_2+T_3)T_3^2 e^{-\theta(T_2+\frac{3}{2}T_3)}] = \frac{\theta^3}{16}\frac{16}{9}\frac{4+3\theta}{(1+\theta)^2(2+\theta)^4}, \tag{5.19}$$

yielding

$$q_n^\theta(X) = \sum_{i=1}^3 q_n^\theta(X \mid \text{topology } i)\times\frac{1}{3} = \frac{\theta^3}{18}\frac{(12+18\theta+7\theta^2)}{(1+\theta)^3(2+\theta)^4}. \tag{5.20}$$

Although, this direct approach is straightforward, there are at least two problems:

1. As discussed in Chapter 2.5, the number of inequivalent coalescent tree topologies with n leaves is $n!(n-1)!/2^{n-1}$, which increases super exponentially with the sample size n
2. The integral we need to perform for each coalescent tree topology is $(n-1)$-dimensional.

So, the computational complexity of this direct approach grows quickly with n, and we need a more efficient approach that avoids having to enumerate all possible histories explicitly.

5.9 Perfect phylogeny

It turns out that there is an efficient test to check whether a given data set is compatible with the infinite-sites model without recombination. To illustrate, consider the following data set:

$$X = \begin{pmatrix} 0&1&0&1&0\\ 1&1&1&0&1\\ 0&1&1&1&0\\ 1&1&1&1&0\\ 0&0&0&0&1 \end{pmatrix}.$$

In this data set, the haplotypes (rows) restricted to the first and the fourth columns, corresponding to the first and the fourth segregating sites, contain all four possible configurations $00, 01, 10$, and 11. It is easy to see that, in the absence of recombination, at least three mutations are needed to generate all four configurations. This violates the infinite-

sites assumption, which implies that at most one mutation occurs per site. In summary, a sufficient condition for $p_n^\theta(X) = 0$ is that X is not compatible with the model assumption. Equivalently, under the infinite-sites model without recombination, $\mathbb{P}(X \mid \theta) = 0$ if X does not admit a perfect phylogeny, defined as follows:

Definition 5.18. A perfect phylogeny for a binary matrix X is a tree which satisfies the following conditions:

1. There is a 1-1 correspondence between the leaves of the tree and the rows of X.
2. Mutations occur on the edges of the tree and every interior edge has at least one mutation on it.
3. There is at most one mutation per column of X.
4. For any two leaves i and j of the tree, the path between them contains a mutation for column k if and only if $X_{i,k} \neq X_{j,k}$.

Note that a perfect phylogeny need not be a binary tree. Given an n-by-m binary matrix X, there exists an $O(nm)$-time algorithm (Gusfield, 1991) to test whether X admits a perfect phylogeny and to construct one if it exists. Also, one can show the following results [proofs can be found in (Gusfield, 1997)]:

Theorem 5.19 (Root known). *Suppose that the root sequence is the all-0 sequence. Then, there exists a rooted perfect phylogeny for X if and only if no two columns contain all of* $01, 10$ *and* 11 *configurations.*

Theorem 5.20 (Root unknown). *There exists an unrooted perfect phylogeny for X if and only if no two columns contain all of* $00, 01, 10$ *and* 11 *configurations.*

Given an input binary matrix X with m columns, the *majority-allele sequence* is a length-m binary string in which the ith character is the majority allele at column i of X; if there is a tie, then arbitrarily choose 0 or 1.

Theorem 5.21. *There exists an unrooted perfect phylogeny for X if and only if there exists a rooted perfect phylogeny for X with the majority-allele sequence as the root sequence.*

5.10 Probability recursion for gene trees

In this section, we develop a systematic approach to computing $p_n^\theta(X)$. Assume that the root sequence is the all-0 sequence. Then, given an n-by-m binary matrix X, first test for the existence of a rooted perfect phylogeny and construct one if it exists. Then, for each distinct haplotype, record its multiplicity and follow the path to the root to record the mutations encountered. For example, consider

$$X = \begin{pmatrix} 1\,1\,0\,0\,0 \\ 1\,1\,0\,0\,0 \\ 0\,0\,1\,0\,0 \\ 0\,0\,0\,1\,0 \\ 0\,0\,0\,1\,1 \end{pmatrix}. \tag{5.21}$$

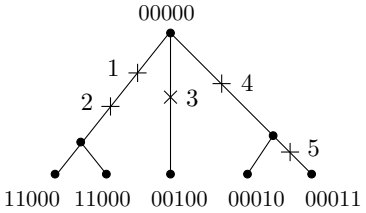


Fig. 5.6: A rooted perfect phylogeny for the data X in (5.21). We use the column numbers to label mutations, marked by crosses.

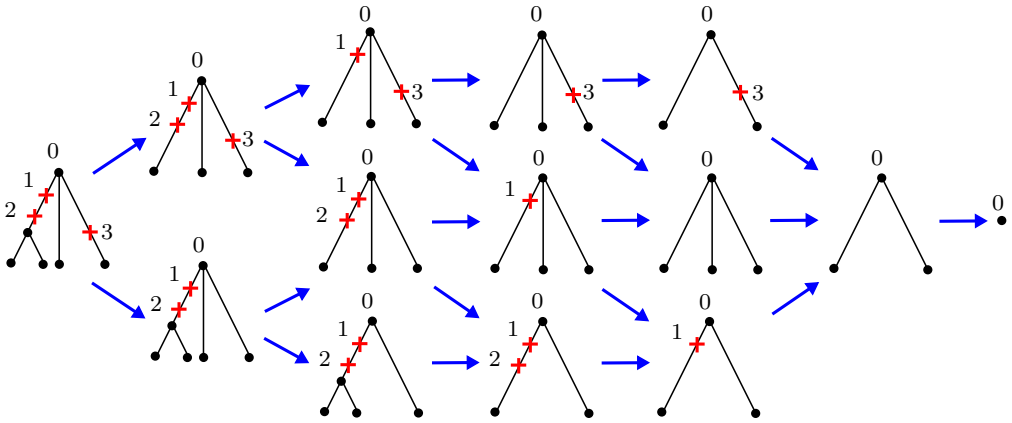


Fig. 5.7: A graphical illustration of the probability recursion for the infinite-sites model. Transitions are denoted by blue arrows.

This data set admits a rooted perfect phylogeny (see Figure 5.6) with the root being the all-0 sequence. The ordering of the columns of X is not important in the absence of recombination, and likewise the ordering of mutations within each edge is irrelevant. Two trees related by permutations of mutation labels within each edge are considered equivalent. For the perfect phylogeny in Figure 5.6, the corresponding distinct mutation lists and their multiplicities are

$$\begin{aligned} \boldsymbol{x}_1 &= (2,1,0),\ n_1 = 2, \\ \boldsymbol{x}_2 &= (3,0),\quad n_2 = 1, \\ \boldsymbol{x}_3 &= (4,0),\quad n_3 = 1, \\ \boldsymbol{x}_4 &= (5,4,0),\ n_4 = 1, \end{aligned}$$

where we appended a 0 to each $\boldsymbol{x}_i$ to denote the root. In general, if there are d distinct haplotypes, we encode the perfect phylogeny as $(\boldsymbol{T}, \boldsymbol{n})$, where $\boldsymbol{T} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$ is a d-tuple of distinct mutation lists and $\boldsymbol{n} = (n_1, \ldots, n_d)$ is the associated multiplicity vector. The pair $(\boldsymbol{T}, \boldsymbol{n})$ is referred to as a *gene tree*.

The goal is to construct a recursion satisfied by $(\boldsymbol{T}, \boldsymbol{n})$ by summing over all possible immediately preceding events. The basic idea is illustrated in Figure 5.7 for a simpler example with 4 haplotypes and 3 segregating sites.

To write down the recursion mathematically, we introduce the following operators that modify $\boldsymbol{T}$ or $\boldsymbol{n}$:

Definition 5.22. Shift operators $\mathcal{S}$ and $\mathcal{S}_k$, and the removal operator $\mathcal{R}_k$ are defined as follows.

1. $\mathcal{S}\boldsymbol{x}_i$: Delete the first entry of $\boldsymbol{x}_i$.
2. $\mathcal{S}_k\boldsymbol{T}$: Delete the first entry of $\boldsymbol{x}_k$ in $\boldsymbol{T} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$.
3. $\mathcal{R}_k\boldsymbol{T}$: Remove $\boldsymbol{x}_k$ from $\boldsymbol{T} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$
4. $\mathcal{R}_k\boldsymbol{n}$: Remove n_k from $\boldsymbol{n} = (n_1, \ldots, n_d)$.

Ethier and Griffiths (1987) and Griffiths (1989) established the following result for the case of a constant population size:

Theorem 5.23 (Probability recursion for an unordered sample). *Let $p_n^\theta(\boldsymbol{T}, \boldsymbol{n})$ denote the stationary probability of a labeled unordered sample with configuration $(\boldsymbol{T}, \boldsymbol{n})$ and sample size $n = \sum_{i=1}^d n_i$. For $\theta > 0$, $p_n^\theta(\boldsymbol{T}, \boldsymbol{n})$ satisfies the following recursion equation:*

$$p_n^\theta(\boldsymbol{T}, \boldsymbol{n}) = \frac{n-1}{n-1+\theta}\left[\sum_{k:n_k\geq 2}\frac{n_k-1}{n-1}p_n^\theta(\boldsymbol{T}, \boldsymbol{n}-\boldsymbol{e}_k)\right] + \frac{\theta}{n-1+\theta}\times$$
$$\left[\sum_{\substack{k\,:\,n_k=1,\\ x_{k,1}\text{unique in }\boldsymbol{T},\\ \mathcal{S}\boldsymbol{x}_k\neq\boldsymbol{x}_j,\forall j}}\frac{1}{n}p_n^\theta(\mathcal{S}_k\boldsymbol{T}, \boldsymbol{n}) + \sum_{\substack{k\,:\,n_k=1,\\ x_{k,1}\text{unique in }\boldsymbol{T}}}\sum_{j:\mathcal{S}\boldsymbol{x}_k=\boldsymbol{x}_j}\frac{n_j+1}{n}p_n^\theta(\mathcal{R}_k\boldsymbol{T}, \mathcal{R}_k(\boldsymbol{n}+\boldsymbol{e}_j))\right],$$

with the boundary conditions $p_n^\theta(\boldsymbol{T}_0, (1)) = 1$, where $\boldsymbol{T}_0 = ((0))$.

Let $q_n^\theta(\boldsymbol{T}, \boldsymbol{n})$ denote the stationary probability of a labeled ordered sample with configuration $(\boldsymbol{T}, \boldsymbol{n})$ and sample size $n = \sum_{i=1}^d n_i$. Note that

$$p_n^\theta(\boldsymbol{T}, \boldsymbol{n}) = \binom{n}{n_1, \ldots, n_d} q_n^\theta(\boldsymbol{T}, \boldsymbol{n}).$$

Theorem 5.24 (Probability recursion for an ordered sample). *For $\theta > 0$, $q_n^\theta(\boldsymbol{T}, \boldsymbol{n})$ satisfies the following recursion equation:*

$$q_n^\theta(\boldsymbol{T}, \boldsymbol{n}) = \frac{n-1}{n-1+\theta}\left[\sum_{k:n_k\geq 2}\frac{n_k(n_k-1)}{n(n-1)}q_n^\theta(\boldsymbol{T}, \boldsymbol{n}-\boldsymbol{e}_k)\right] + \frac{\theta}{n-1+\theta}\times$$
$$\left[\sum_{\substack{k\,:\,n_k=1,\\ x_{k,1}\text{unique in }\boldsymbol{T},\\ \mathcal{S}\boldsymbol{x}_k\neq\boldsymbol{x}_j,\forall j}}\frac{1}{n}q_n^\theta(\mathcal{S}_k\boldsymbol{T}, \boldsymbol{n}) + \sum_{\substack{k\,:\,n_k=1,\\ x_{k,1}\text{unique in }\boldsymbol{T}}}\sum_{j:\mathcal{S}\boldsymbol{x}_k=\boldsymbol{x}_j}\frac{1}{n}q_n^\theta(\mathcal{R}_k\boldsymbol{T}, \mathcal{R}_k(\boldsymbol{n}+\boldsymbol{e}_j))\right],$$

with the boundary condition $q_n^\theta(\boldsymbol{T}_0, (1)) = 1$, where $\boldsymbol{T}_0 = ((0))$.

Remark 5.25.

1. The above results can be proved using the backward-forward argument, similar to that used in Chapter 4.2.
2. The condition "$x_{k,1}$unique in $\boldsymbol{T}$" is needed so that the root does not get removed.

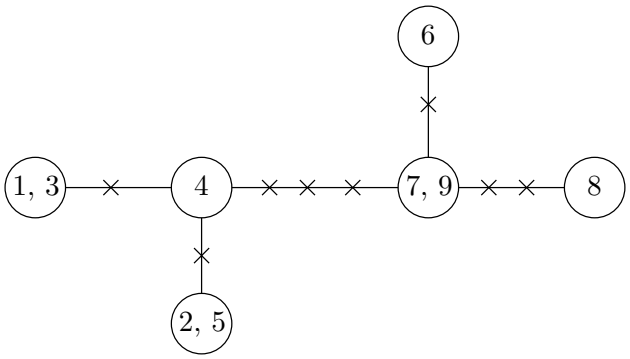


Fig. 5.8: A minimal unrooted tree for $n = 9$.

3. The recursions are purely algebraic. No messy integration appears in the recursions.
4. The recursions can be solved numerically using dynamic programming or Monte Carlo methods.
5. Wu (2010) has developed a dynamic programming algorithm to solve the above recursions numerically for moderate-size data sets.

5.11 Root unknown case

In the case the root is unknown, construct the unrooted perfect phylogeny, if there exists one, and contract all edges with no mutations. Denote the resulting unrooted tree by τ and call it a minimal unrooted tree. An example is provided in Figure 5.8.

Theorem 5.26. *Let m denote the number of segregating sites (or columns) in the input matrix X and suppose that X admits an unrooted perfect phylogeny. Then, up to permutations of mutation labels, the number of inequivalent rooted perfect phylogenies for X is $1 + m$.*

Proof. For a given minimal unrooted tree τ, we can place the root at any vertex or between two consecutive mutations on an edge. Thus, up to permutations of mutation labels, the total number of inequivalent perfect phylogenies is

$$|V(\tau)| + \sum_{e \in E(\tau)} (m_e - 1) = |V(\tau)| - |E(\tau)| + \sum_{e \in E(\tau)} m_e,$$

where $V(\tau)$ and $E(\tau)$ are the vertex and edge sets of τ, respectively, and m_e denotes the number of mutations on edge e. Now, $|V(\tau)| - |E(\tau)| = 1$, while $\sum_{e \in E(\tau)} m_e = m$, and therefore we have the desired result.

If the root is unknown and X admits an unrooted perfect phylogeny, one way to compute the probability of X is to sum $p_n^\theta(X)$ over all $1 + m$ rooted perfect phylogenies. A more efficient method is to use a probability recursion for the unrooted case. See Tavaré (2004) for details.

References

Achaz G (2009) Frequency spectrum neutrality tests: one for all and all for one. Genetics 183(1):249–258

Ethier SN, Griffiths RC (1987) The infinitely-many-sites model as a measure valued diffusion. Ann Probab 15:515–545

Fu YX (1994) Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of dna sequences. Genetics 138(4):1375–1386

Fu YX (1995) Statistical properties of segregating sites. Theoretical Population Biology 48:172–197

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133(3):693–709

Futschik A, Gach F (2008) On the inadmissibility of watterson's estimator. Theoretical population biology 73(2):212–221

Griffiths R, Tavaré S (1998) The age of a mutation in a general coalescent tree. Communications in Statistics Stochastic Models 14(1-2):273–295

Griffiths RC (1989) Genealogical-tree probabilities in the infinitely-many-site mode. J Math Biol 27:667–680

Gusfield D (1991) Efficient algorithms for inferring evolutionary trees. Networks 21:19–28

Gusfield D (1997) Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology. Cambridge university press

Kimura M, Ohta T (1973) The age of a neutral mutant persisting in a finite population. Genetics 75(1):199–212

Polanski A, Bobrowski A, Kimmel M (2003) A note on distributions of times to coalescence, under time-dependent population size. Theoretical Population Biology 63(1):33–40

Tajima F (1983) Evolutionary relationship of dna sequences in finite populations. Genetics 105(2):437–460

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by dna polymorphism. Genetics 123(3):585–595

Tavaré S (2004) Ancestral inference in population genetics. In: Tavaré S, Zeitouni O (eds) Lectures on Probability Theory and Statistics. Ecole d'Eté de Probabilités de Saint-Flour XXXI–2001, Springer Berlin/Heidelberg, Lecture Notes in Mathematics, vol 1837, pp 1–188

Watterson G (1975) On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7(2):256–276

Wu Y (2010) Exact computation of coalescent likelihood for panmictic and subdivided populations under the infinite sites model. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 7(4):611–618

Živković D, Wiehe T (2008) Second-order moments of segregating sites under variable population size. Genetics 180(1):341–357

Chapter 6
Finite-alleles model of mutation

In this chapter, we will study a more realistic model of mutation. Specifically, we will consider a finite collection of loci each with a finite number of possible alleles. When mutation occurs at a particular locus, the allelic type at that locus changes according to a Markov process. Furthermore, unlike in the infinite-sites model, the same locus can mutate multiple times in this model. Unfortunately, except for a rather special Markov model called the *parent-independent mutation* model, no exact sampling formula is known even for the case of a single locus. We will therefore discuss Monte Carlo algorithms for approximating the sampling probability.

6.1 Sampling probability

We consider a finite number of loci and assume that the allelic type space E at each locus is finite. A particularly interesting example is $E = \{A, C, G, T\}$, corresponding to the four DNA nucleotides at a particular site in the genome. More generally, we consider

$$E = \{1, 2, 3, \ldots, K\},$$

for some $K \in \mathbb{N}$. This case could correspond to the situation where we have a locus consisting of L sites, each of which could be one of $\{A, C, G, T\}$, giving $|E| = 4^L$ unique alleles at the locus. We make three key assumptions in this model:

1. Mutations arrive according a Poisson point process with intensity $\theta/2$,
2. Given that a mutation has occurred, allelic type change is governed by a Markov chain with transition matrix $\mathbf{P} = (P_{\alpha\beta})_{\alpha,\beta\in E}$, where $P_{\alpha\beta}$ gives the probability of transitioning from allele α to allele β (going forward in time).
3. There exists a unique stationary distribution $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_K)$ of the Markov chain.

Definition 6.1 (Sample configuration). A sample configuration is denoted by $\mathbf{n} = (n_1, n_2, \ldots, n_K)$, where $n_\alpha \in \mathbb{N}_0$ corresponds to the number of times that allele $\alpha \in E$ is represented in the sample. The size of the sample $\mathbf{n}$ is denoted $|\mathbf{n}| = \sum_{\alpha\in E} n_\alpha$.

We are interested in the stationary probability $p_n(\mathbf{n} \mid \theta, \mathbf{P})$ of an unordered sample of size n having configuration $\mathbf{n}$, where $|\mathbf{n}| = n$. For ease of notation, we drop the dependence

on θ and $\mathbf{P}$ when writing the sampling probability. By conditioning on the first event back in time and using a similar "backward-forward" argument as in the proof of Theorem 4.4, one can show the following result:

Theorem 6.2 (Probability recursion for an unordered sample). *Suppose a sample of n alleles is drawn from an infinite population at stationary. Then, the probability $p_n(\mathbf{n})$ of observing a particular unordered sample with configuration $\mathbf{n} = (n_1, \ldots, n_K) \in \mathbb{N}_0^K$, where $|\mathbf{n}| = n$, satisfies the recursion*

$$p_n(\mathbf{n}) = \frac{\theta}{n-1+\theta} \sum_{\beta \in E} \sum_{\alpha \in E : n_\alpha \geq 1} \left(\frac{n_\beta + 1 - \delta_{\alpha\beta}}{n} \right) P_{\beta\alpha}\, p_n(\mathbf{n} - \mathbf{e}_\alpha + \mathbf{e}_\beta) + \frac{n-1}{n-1+\theta} \sum_{\alpha \in E : n_\alpha \geq 2} \frac{n_\alpha - 1}{n-1} p_{n-1}(\mathbf{n} - \mathbf{e}_\alpha), \tag{6.1}$$

where $\delta_{\alpha\beta}$ denotes the Kronecker delta. Boundary conditions are $p_1(\mathbf{e}_\alpha) = \varphi_\alpha$ for all $\alpha \in E$.

Note the condition $n_\alpha \geq 1$ in the first line of the above recursion. Unlike in the recursion for the infinite-sites model (cf., Theorem 5.23), where at most one mutation occurs per site, in (6.1) alleles of type α can undergo mutation (backward in time) before coalescence events reduce the copy number to one.

Now, suppose n alleles are sequentially sampled one by one from the population and the order in which they appear is recorded. Then, the probability $q_n(\mathbf{n})$ of observing a particular ordered sample with configuration $\mathbf{n}$, where $|\mathbf{n}| = n$, is related to the unordered sampling probability $p_n(\mathbf{n})$ as

$$p_n(\mathbf{n}) = \binom{n}{n_1, \ldots, n_K} q_n(\mathbf{n}). \tag{6.2}$$

The recursion for $q_n(\mathbf{n})$ takes on a slightly simpler form than (6.1):

Theorem 6.3 (Probability recursion for an ordered sample). *Suppose a sample of n alleles is sequentially drawn from an infinite population at stationarity. Then, the probability $q_n(\mathbf{n})$ of observing a particularly ordered sample with configuration $\mathbf{n} = (n_1, \ldots, n_K) \in \mathbb{N}_0^K$, where $|\mathbf{n}| = n$, satisfies the recursion*

$$q_n(\mathbf{n}) = \frac{\theta}{n-1+\theta} \sum_{\beta \in E} \sum_{\alpha \in E : n_\alpha \geq 1} \frac{n_\alpha}{n} P_{\beta\alpha}\, q_n(\mathbf{n} - \mathbf{e}_\alpha + \mathbf{e}_\beta) + \frac{n-1}{n-1+\theta} \sum_{\alpha \in E : n_\alpha \geq 2} \frac{n_\alpha(n_\alpha - 1)}{n(n-1)} q_{n-1}(\mathbf{n} - \mathbf{e}_\alpha), \tag{6.3}$$

with boundary conditions are $q_1(\mathbf{e}_\alpha) = \varphi_\alpha$ for all $\alpha \in E$.

Note that neither (6.1) nor (6.3) is strictly recursive. For a general transition matrix $\mathbf{P}$, an exact closed-form solution to (6.1) or (6.3) is unknown. However, sampling probabilities can be computed numerically by solving systems of coupled linear equations sequentially for sample sizes 2 through n. Graphically, assign a vertex to each distinct sample configuration and draw an edge from vertex $v_{\mathbf{m}}$ to vertex $v_{\mathbf{m}'}$ if $\mathbf{m}$ appears on the left hand side of (6.1) while $\mathbf{m}'$ appears on the right hand side. Topologically sort this directed graph, find the highest strongly connected component (SCC), and solve the system of couple linear

equations corresponding to that SCC. Upon recording that solution, move down in the topological order, solve the system of couple linear equations corresponding to that new SCC, and iterate the procedure until the SCC for sample size n has been solved. The desired probability $p_n(\mathbf{n})$ appears as one of the variables in the last system of linear equations. Clearly this method will not scale to large sample sizes, since the SCC at level m has $O(m^K)$ vertices. In this chapter, we will discuss Monte Carlo methods for solving the recursion (6.1).

6.2 Parent-independent mutation

The solution to (6.1) remains unknown in general, but with additional assumptions and conditions, we can obtain a more tractable model for which an exact solution exists. Such a more tractable model is *parent-independent mutation* (PIM), in which $P_{\beta\alpha} = \varphi_\alpha$, i.e., the probability of transitioning to allele α (forward in time) is independent of the current allelic state β. We discuss below what this buys us.

Let X_α denote the population-wide frequency of allele $\alpha \in E$. Then, the sampling probability $p_n(\mathbf{n})$ can be written as

$$p_n(\mathbf{n}) = \binom{n}{n_1, \ldots, n_K} \mathbb{E}\left[\prod_{\alpha \in E} X_\alpha^{n_\alpha}\right]$$

Using this representation and the fact $\sum_{\alpha \in E} X_\alpha = 1$, one can show that for all sample configurations $\mathbf{n}$ with $|\mathbf{n}| = n \geq 1$,

$$\sum_{\beta \in E} \frac{n_\beta + 1}{n+1} p_{n+1}(\mathbf{n} + \mathbf{e}_\beta) = p_n(\mathbf{n}), \tag{6.4}$$

which implies

$$\sum_{\beta \in E} \frac{n_\beta + 1 - \delta_{\alpha\beta}}{n} p_n(\mathbf{n} - \mathbf{e}_\alpha + \mathbf{e}_\beta) = p_{n-1}(\mathbf{n} - \mathbf{e}_\alpha). \tag{6.5}$$

Hence, if $P_{\beta\alpha} = \varphi_\alpha$, then (6.5) can be used to simplify (6.1) as

$$p_n(\mathbf{n}) = \frac{\theta}{n-1+\theta} \sum_{\alpha \in E : n_\alpha \geq 1} \varphi_\alpha \, p_{n-1}(\mathbf{n} - \mathbf{e}_\alpha) + \frac{n-1}{n-1+\theta} \sum_{\alpha \in E : n_\alpha \geq 2} \frac{n_\alpha - 1}{n-1} p_{n-1}(\mathbf{n} - \mathbf{e}_\alpha), \tag{6.6}$$

which is strictly recursive in n. The unique solution to (6.6) with boundary conditions $p_1(\mathbf{e}_\alpha) = \varphi_\alpha$, for all $\alpha \in E$, is given by

$$p_n(\mathbf{n}) = \binom{n}{n_1, \ldots, n_K} \frac{\prod_{\alpha \in E} (\theta \varphi_\alpha)_{n_\alpha \uparrow}}{(\theta)_{n\uparrow}}. \tag{6.7}$$

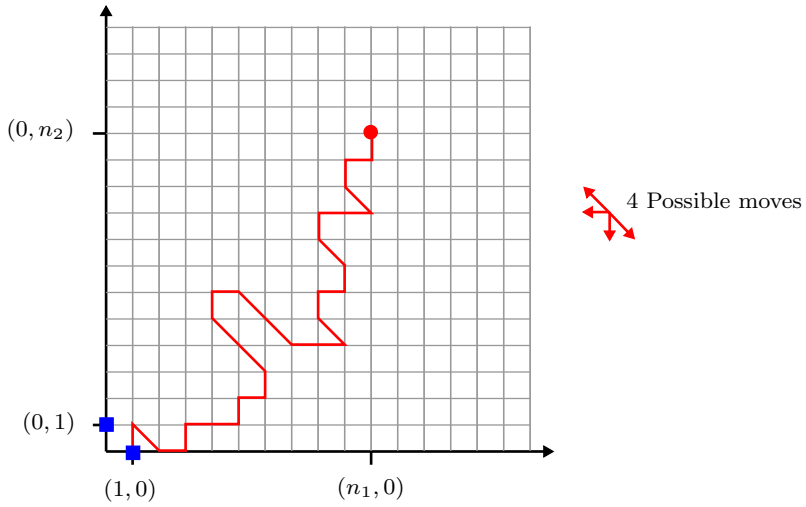


Fig. 6.1: Illustration of a random walk in the case of $K = 2$. The probability $p_n(\mathbf{n})$ can be approximated by averaging a certain function over random sample paths each of which starts at $\mathbf{n} = (n_1, n_2)$ (denoted by a red circle) and ends at one of the absorbing states $\{\mathbf{e}_\alpha\}$ (denoted by blue squares). There are four allowed moves in the interior of the lattice, while only two moves are allowed on the x- and y-axes.

6.3 A simple Monte Carlo method for approximating the likelihood

Here, we discuss a simple Monte Carlo algorithm for finding an approximate solution to (6.1). This approach was originally proposed by Griffiths and Tavaré (1994a,b,c). Figure 6.1 illustrates the method in the case of $K = 2$. The input sample configuration $\mathbf{n} = (n_1, n_2)$ is identified with a point in the lattice $\mathbb{N}_0^2$, and starting from that point one takes a random walk through $\mathbb{N}_0^2$ using allowed moves, until hitting one of the absorbing states $\mathbf{e}_1, \ldots, \mathbf{e}_K$.

The transition probability $P_{\alpha\alpha}$ may be non-zero for some $\alpha \in E$, which means $p_n(\mathbf{n})$ will appear on both sides of (6.1). To make the Monte Carlo algorithm more efficient, we manipulate the original recursion to eliminate terms corresponding to self-loops:

$$\left[1 - \frac{\theta}{n-1+\theta} \sum_{\alpha \in E} P_{\alpha\alpha} \frac{n_\alpha}{n}\right] p_n(\mathbf{n}) = \frac{\theta}{n-1+\theta} \sum_{\alpha,\beta \in E: \alpha \neq \beta, n_\alpha \geq 1} P_{\beta\alpha} \frac{n_\beta + 1}{n} p_n(\mathbf{n} - \mathbf{e}_\alpha + \mathbf{e}_\beta) + \frac{n-1}{n-1+\theta} \sum_{\alpha \in E: n_\alpha \geq 2} \frac{n_\alpha - 1}{n-1} p_{n-1}(\mathbf{n} - \mathbf{e}_\alpha) \tag{6.8}$$

Let $b(\mathbf{n})$ denote the coefficient of $p_n(\mathbf{n})$ on the left hand side of (6.8), and define

$$a(\mathbf{n}) = \frac{\theta}{n-1+\theta} \sum_{\alpha,\beta \in E: \alpha \neq \beta, n_\alpha \geq 1} P_{\beta\alpha} \frac{n_\beta + 1}{n} + \frac{1}{n-1+\theta} \sum_{\alpha \in E: n_\alpha \geq 2} (n_\alpha - 1).$$

This gives us

$$p_n(\mathbf{n}) = \frac{a(\mathbf{n})}{b(\mathbf{n})}\Bigg[\sum_{\alpha,\beta\in E:\alpha\neq\beta,n_\alpha\geq 1}\overbrace{\frac{\theta}{a(\mathbf{n})(n-1+\theta)}\frac{P_{\beta\alpha}(n_\beta+1)}{n}}^{\lambda_{\beta\alpha}(\mathbf{n})}p_n(\mathbf{n}-\mathbf{e}_\alpha+\mathbf{e}_\beta) + \sum_{\alpha\in E:n_\alpha\geq 2}\underbrace{\frac{n_\alpha-1}{a(\mathbf{n})(n-1+\theta)}}_{\mu_\alpha(\mathbf{n})}p_{n-1}(\mathbf{n}-\mathbf{e}_\alpha)\Bigg]. \tag{6.9}$$

The coefficients $\mu_\alpha(\mathbf{n})$ and $\lambda_{\beta\alpha}(\mathbf{n})$ are defined as shown above, and they satisfy

$$\sum_{\alpha,\beta\in E:\alpha\neq\beta,n_\alpha\geq 1}\lambda_{\beta\alpha}(\mathbf{n}) + \sum_{\alpha,\beta\in E:\alpha\neq\beta,n_\alpha\geq 1}\mu_\alpha(\mathbf{n}) = 1.$$

Hence, the form of (6.9) suggests constructing a discrete-time Markov chain $\{\mathbf{Z}_j, j\geq 0\}$ with state space $\mathbb{N}_0^K$ as follows:

1. $\mathbf{Z}_0 = \mathbf{n}$.
2. For $n\geq 2$, possible transitions are
   a. $\mathbf{n}\to\mathbf{n}-\mathbf{e}_\alpha+\mathbf{e}_\beta$ with probability $\lambda_{\beta\alpha}(\mathbf{n})$, for $\alpha,\beta\in E, \alpha\neq\beta, n_\alpha\geq 1$,
   b. $\mathbf{n}\to\mathbf{n}-\mathbf{e}_\alpha$ with probability $\mu_\alpha(\mathbf{n})$, for $\alpha\in E, n_\alpha\geq 2$.
3. Absorbing states are $\mathcal{A} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K\}$.
4. The random hitting time is $\tau = \inf\{j\geq 0 \mid \mathbf{Z}_j\in\mathcal{A}\}$.

In terms of this Markov chain, (6.9) can be rephrased as follows.

Theorem 6.4 (Griffiths and Tavaré 1994c). *Define $f(\mathbf{e}_\alpha) = \varphi_\alpha$ for $\alpha\in E$ and $f(\mathbf{m}) = a(\mathbf{m})/b(\mathbf{m})$ for $|\mathbf{m}|\geq 2$. Then, the sampling probability $p_n(\mathbf{n})$ is given by*

$$p_n(\mathbf{n}) = \mathbb{E}_{\mathbf{n}}\left[\prod_{j=0}^{\tau} f(\mathbf{Z}_j)\right], \tag{6.10}$$

where the expectation is taken over the random path $\mathbf{Z}_0,\ldots,\mathbf{Z}_\tau$ conditioned on $\mathbf{Z}_0 = \mathbf{n}$.

This result implies that the probability $p_n(\mathbf{n})$ can be approximated by

$$p_n(\mathbf{n}) \approx \frac{1}{M}\sum_{s=1}^{M} f(\mathbf{Z}_0^{(s)})f(\mathbf{Z}_1^{(s)})\cdots f(\mathbf{Z}_{\tau_s}^{(s)}),$$

where $\mathbf{Z}_0^{(s)}, \mathbf{Z}_1^{(s)}, \ldots, \mathbf{Z}_{\tau_s}^{(s)}$ are independent sample paths drawn from the above Markov chain.

Remark 6.5.

1. A potential problem with this approach is that some sample paths might start out with low probability but become more probable later, compared to other sample paths. Such paths are unlikely to be sampled from the above Markov chain, thus reducing efficiency.
2. The above technique has been extended to more general models with additional features, including variable population size (Griffiths and Tavaré, 1994b), recombination (Larribe et al, 2002), natural selection (Coop and Griffiths, 2004), and subdivided population structure (Bahlo and Griffiths, 2000).

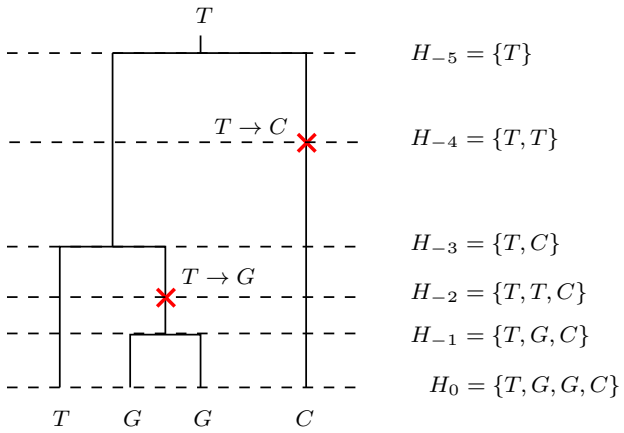


Fig. 6.2: An example history $\mathbf{H} = (H_0, H_{-1}, \ldots, H_{-\tau})$, where H_{-j} denotes the configuration after j events. This history has three coalescence and two mutation events until the root in reached, so $\tau = 5$. Mutation events are denoted by the $\times$ symbol.

6.4 Sequential importance sampling (SIS)

In this section, we discuss a more efficient Monte Carlo method for approximating the likelihood. First, we define the concept of a *coalescent history*.

Definition 6.6 (Coalescent history). A *coalescent history* records the sample configuration after each coalescence or mutation event. It is defined as $\mathbf{H} = (H_0, H_{-1}, \ldots, H_{-\tau})$, where H_{-j} denotes the configuration after j events and τ denotes the random number of events until the configuration size has decreased to 1. An example history is illustrated in Figure 6.2.

Note that there is a one-to-one correspondence between the set of sample paths in the Markov chain discussed in the previous section and the set of coalescent histories.

Let $\mathcal{H}_n$ denote the set of all coalescent histories $\mathbf{H}$ with $|H_0| = n$, and let $\mathbb{P}_n$ denote the coalescent distribution over $\mathcal{H}_n$. Then,

$$p_n(\mathbf{n}) = \sum_{\mathbf{H} \in \mathcal{H}_n} \mathbb{P}_n(\mathbf{n} \mid \mathbf{H})\, \mathbb{P}_n(\mathbf{H}), \tag{6.11}$$

where

$$\mathbb{P}_n(\mathbf{n} \mid \mathbf{H}) = \begin{cases} 1, & \text{if } \mathrm{config}(H_0) = \mathbf{n}, \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathrm{config}(H_0)$ denotes the allele configuration of H_0. (For ease of notation, we sometimes use H_t and $\mathrm{config}(H_t)$ interchangeably in what follows.) It would be hopeless to approximate (6.11) using

$$p_n(\mathbf{n}) \approx \frac{1}{M} \sum_{s=1}^{M} \mathbb{P}_n(\mathbf{n} \mid \mathbf{H}^{(s)})$$

where $\mathbf{H}^{(s)}$ are independent samples of coalescent histories from the coalescent prior $\mathbb{P}_n(\cdot)$, since most such histories would have $\mathbb{P}_n(\mathbf{n} \mid \mathbf{H}^{(s)}) = 0$. A much better scheme is to sample coalescent histories $\mathbf{H}^{(s)}$ conditioned on $\text{config}(H_0^{(s)}) = \mathbf{n}$. Importance sampling will enable us to achieve this goal. The basic idea is as follows. Suppose Q is a probability measure such that the support of $\mathbb{P}_n$ is contained in the support of Q. Then, we can write

$$\begin{aligned} p_n(\mathbf{n}) &= \sum_{\mathbf{H} \in \mathcal{H}_n} \mathbb{P}_n(\mathbf{n} \mid \mathbf{H}) \frac{\mathbb{P}_n(\mathbf{H})}{Q(\mathbf{H})} Q(\mathbf{H}) \\ &\approx \frac{1}{M} \sum_{s=1}^{M} \mathbb{P}_n(\mathbf{n} \mid \mathbf{H}^{(s)}) \frac{\mathbb{P}_n(\mathbf{H}^{(s)})}{Q(\mathbf{H}^{(s)})}, \end{aligned} \tag{6.12}$$

where $\mathbf{H}^{(s)}$ are independent samples from Q. The ratio $w^{(s)} = \mathbb{P}(\mathbf{H}^{(s)})/Q(\mathbf{H}^{(s)})$ is called the *importance weight*. The key to the success of this algorithm depends critically on the choice of the proposal distribution Q.

Proposition 6.7 (Optimal proposal distribution). *The optimal proposal distribution $Q^*(\mathbf{H})$ is given by the posterior distribution $\mathbb{P}_n(\mathbf{H} \mid H_0 = \mathbf{n})$. This implies that we can compute the true probability $p_n(\mathbf{n})$ using a single draw and that the importance weight $w^{(s)}$ has zero variance.*

Proof. Suppose $\mathbf{H}$ is sampled from the posterior distribution $\mathbb{P}_n(\cdot|\mathbf{n})$. Then, $\mathbb{P}_n(\mathbf{n} \mid \mathbf{H}) = 1$ and using Bayes' rule we see that the importance weight is given by

$$\frac{\mathbb{P}_n(\mathbf{H})}{\mathbb{P}_n(\mathbf{H} \mid \mathbf{n})} = \frac{\mathbb{P}_n(\mathbf{H}) p_n(\mathbf{n})}{\mathbb{P}_n(\mathbf{n} \mid \mathbf{H}) \mathbb{P}_n(\mathbf{H})} = p_n(\mathbf{n}),$$

so a single draw from $\mathbb{P}_n(\cdot|\mathbf{n})$ is sufficient to find the true likelihood. □

Unfortunately, sampling from the optimal proposal distribution Q^* is as hard as computing $p_n(\mathbf{n})$. A major advance on this problem was made when Stephens and Donnelly (2000) showed that it is possible to characterize the optimal proposal distribution Q^* in terms of a simpler sampling distribution. In the remainder of this section, we provide an alternate derivation of their results.

6.4.1 The coalescent prior distribution of histories

A history $\mathbf{H}$ can be sampled from the coalescent prior $\mathbb{P}_n(\cdot)$ as follows. Starting from n lineages, sample a sequence of coalescence and mutation events backwards in time until only one lineage remains. Then, sample the ancestral allele type from the stationary distribution $\boldsymbol{\varphi}$, and propagate the information forward in time conditioned on the sequence of coalescence and mutation events sampled in the first step. A graphical model describing the conditional independence structure of this generative model is shown in Figure 6.3, where A_{-t} denotes the number of lineages remaining after t events.

To compute the numerator $\mathbb{P}_n(\mathbf{H})$ of the importance weight, first note that

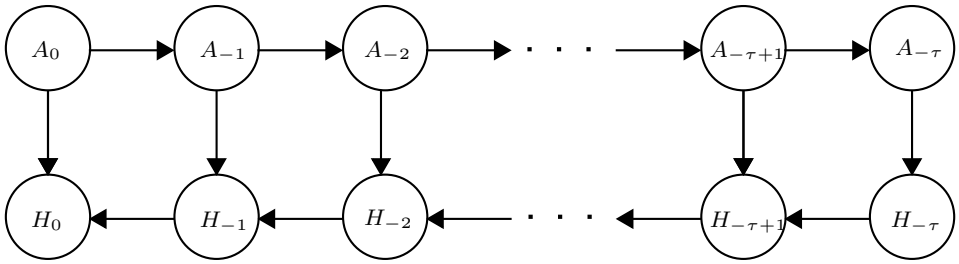


Fig. 6.3: Graphical model depicting the conditional independence structure of the process for generating a coalescent history. The random variable A_{-t} denotes the number of lineages remaining after t events backwards in time. The sequence $A_0, A_{-1}, \ldots, A_{-\tau}$ is a Markov chain, with $A_0 = n$.

$$\begin{aligned}\mathbb{P}_n(\mathbf{H}) &= \sum_{a_1,\ldots,a_\tau} \mathbb{P}_n(\mathbf{H} \mid A_{-1} = a_1, \ldots, A_{-\tau} = a_\tau)\mathbb{P}_n(A_{-1} = a_1, \ldots, A_{-\tau} = a_\tau) \\ &= \mathbb{P}_n(\mathbf{H} \mid A_{-1} = |H_{-1}|, \ldots, A_{-\tau} = |H_{-\tau}|)\mathbb{P}_n(A_{-1} = |H_{-1}|, \ldots, A_{-\tau} = |H_{-\tau}|).\end{aligned}$$

For ease of notation, we define $n_t = |H_t|$. Then, the conditional independence structure of the model (Figure 6.3) implies that the above equation factorizes as

$$\mathbb{P}_n(\mathbf{H}) = \mathbb{P}_n(H_{-\tau} \mid A_{-\tau} = n_{-\tau}) \prod_{t=-\tau+1}^{0} \mathbb{P}(A_{t-1} = n_{t-1} \mid A_t = n_t)\mathbb{P}(H_t \mid H_{t-1}, A_t = n_t)$$

The definition of τ implies $A_{-\tau} = 1$ and $H_{-\tau} = \{\alpha\}$ for some $\alpha \in E$, so the first term in the above factorization is $\mathbb{P}(H_{-\tau} = \{\alpha\}|A_{-\tau} = 1) = p_1(\mathbf{e}_\alpha) = \varphi_\alpha$. The backward transition probability $\mathbb{P}(A_{t-1} = j \mid A_t = k)$ in the number of lineages is given by

$$\mathbb{P}(A_{t-1} = j \mid A_t = k) = \begin{cases} \dfrac{k-1}{k-1+\theta}, & \text{if } j = k-1, \\ \dfrac{\theta}{k-1+\theta}, & \text{if } j = k, \\ 0, & \text{otherwise,} \end{cases} \tag{6.13}$$

where the first case corresponds to a coalescence event, and the second case a mutation event. Lastly, for $\text{config}(H_{t-1}) = (n_{t-1,\alpha})_{\alpha\in E}$ where $\sum_{\alpha\in E} n_{t-1,\alpha} = n_{t-1}$ and $\text{config}(H_t) = (n_{t,\alpha})_{\alpha\in E}$ where $\sum_{\alpha\in E} n_{t,\alpha} = n_t$, the forward transition probability $\mathbb{P}(H_t|H_{t-1}, A_t = n_t)$ is given by

$$\mathbb{P}(H_t|H_{t-1}, A_t = n_t) = \begin{cases} \dfrac{n_{t-1,\alpha}}{n_{t-1}} = \dfrac{n_{t,\alpha}-1}{n_t-1}, & \text{if } H_t = H_{t-1} + \alpha, \\ \dfrac{n_{t-1,\beta}}{n_{t-1}} P_{\beta\alpha} = \dfrac{n_{t,\beta}+1-\delta_{\alpha\beta}}{n_t} P_{\beta\alpha}, & \text{if } H_t = H_{t-1} - \beta + \alpha, \\ 0, & \text{otherwise.} \end{cases} \tag{6.14}$$

The first case is the probability of a lineage of type α splitting into two, given that a branching occurs. The second case is the probability of a lineage of type β mutating to α, given that a mutation occurs.

6.4.2 Reverse transition probability

Noting that $H_0, H_{-1}, \ldots, H_{-\tau}$ is a Markov chain, the optimal proposal distribution $Q^*(\mathbf{H}) = \mathbb{P}_n(\mathbf{H} \mid H_0 = \mathbf{n})$ can be written in terms of the reverse transition probabilities:

$$\mathbb{P}_n(\mathbf{H} \mid H_0 = \mathbf{n}) = \mathbb{P}(H_{-1}|H_0 = \mathbf{n})\mathbb{P}(H_{-2}|H_{-1}) \cdots \mathbb{P}(H_{-\tau}|H_{-\tau+1}).$$

Noting that $\mathbb{P}(H_{t-1}|H_t) = \mathbb{P}(H_{t-1}|H_t, A_t = n_t)$, where $n_t = |H_t|$, we can utilize the conditional independence structure illustrated in Figure 6.3 to obtain

$$\begin{aligned}
&\mathbb{P}(H_{t-1}|H_t, A_t = n_t) \\
&\quad = \mathbb{P}(H_t \mid A_t = n_t, H_{t-1}) \frac{\mathbb{P}(H_{t-1} \mid A_t = n_t)}{\mathbb{P}(H_t \mid A_t = n_t)} \\
&\quad = \mathbb{P}(H_t \mid A_t = n_t, H_{t-1}) \frac{\mathbb{P}(H_{t-1}, A_{t-1} = n_{t-1} \mid A_t = n_t)}{\mathbb{P}(H_t \mid A_t = n_t)} \\
&\quad = \mathbb{P}(H_t \mid A_t = n_t, H_{t-1}) \mathbb{P}(A_{t-1} = n_{t-1} \mid A_t = n_t) \frac{\mathbb{P}(H_{t-1} \mid A_{t-1} = n_{t-1})}{\mathbb{P}(H_t \mid A_t = n_t)}.
\end{aligned}$$

The first two terms in the last line are known; the first term is given by (6.14) while the second term is given by (6.13). Unfortunately, the last term is challenging to evaluate, since $\mathbb{P}(H_{t-1} \mid A_{t-1} = n_{t-1}) = p_{n_{t-1}}(H_{t-1})$ and $\mathbb{P}(H_t \mid A_t = n_t) = p_{n_t}(H_t)$, which are unknown.

Let $\pi(\alpha \mid H)$ denote the conditional probability that an additionally sampled allele is of type α, given that a sample H has been observed already. This conditional sampling probability can be written as a simple ratio of ordered probabilities. Specifically, if $|H| = h$, then

$$\pi(\alpha \mid H) = \frac{q_{h+1}(H + \alpha)}{q_h(H)}. \tag{6.15}$$

Then, using (6.15) and the relation (6.2) between unordered and ordered sampling probabilities, we obtain

$$\frac{\mathbb{P}(H_{t-1}|A_{t-1} = n_{t-1})}{\mathbb{P}(H_t|A_t = n_t)} = \begin{cases} \dfrac{n_{t,\alpha}}{n_t} \dfrac{1}{\pi(\alpha|H_t - \alpha)}, & \text{if } H_t = H_{t-1} + \alpha, \\ \left(\dfrac{n_{t,\alpha}}{n_{t,\beta} + 1 - \delta_{\alpha\beta}} \right) \dfrac{\pi(\beta|H_t - \alpha)}{\pi(\alpha|H_t - \alpha)}, & \text{if } H_t = H_{t-1} - \beta + \alpha, \\ 0, & \text{otherwise.} \end{cases}$$

Putting all these terms together, we obtain the following result:

Theorem 6.8 (Reverse transition probability). *Suppose* $\text{config}(H_t) = (n_{t,\alpha})_{\alpha \in E}$ *and* $n_t = \sum_{\alpha \in E} n_{t,\alpha}$. *Then, the transition probability of the Markov chain* $H_0, H_{-1}, \ldots, H_{-\tau}$ *is given by (Stephens and Donnelly, 2000)*

$$\mathbb{P}(H_{t-1}|H_t) = \begin{cases} \dfrac{2}{n_t(n_t-1+\theta)}\binom{n_{t,\alpha}}{2}\dfrac{1}{\pi(\alpha|H_t-\alpha)}, & \text{if } H_{t-1} = H_t - \alpha, \\ \dfrac{2}{n_t(n_t-1+\theta)}\dfrac{\theta}{2}P_{\beta\alpha}n_{t,\alpha}\dfrac{\pi(\beta|H_t-\alpha)}{\pi(\alpha|H_t-\alpha)}, & \text{if } H_{t-1} = H_t - \alpha + \beta, \\ 0, & \text{otherwise.} \end{cases} \tag{6.16}$$

The above theorem relates the multi-dimensional distribution $\mathbb{P}(H_{t-1}|H_t)$ to a one-dimension conditional sampling distribution (CSD) π, which is much simpler to work with. In particular, this connection allows us to construct a useful approximation of Q^* by approximating π. We turn to this topic in the next section.

6.5 Approximate conditional sampling distribution (CSD)

The exact conditional sampling probability $\pi(\alpha \mid H)$ is not known for a general mutation model with an arbitrary transition matrix $\mathbf{P}$. In what follows, we consider a useful approximation proposed by Stephens and Donnelly (2000).

6.5.1 A single site

Stephens and Donnelly (2000) suggested an approximation of $\pi(\alpha \mid H)$ based on the following urn model: Suppose $\text{config}(H) = \mathbf{n} = (n_\beta)_{\beta\in E}$ with $n = \sum_{\beta\in E} n_\beta$. Then, consider an urn containing n_β copies of allele β for each $\beta \in E$, sample an allele uniformly at random from the urn, and then mutate it a random number M of times according to the transition matrix $\mathbf{P}$. Specifically, M is assumed to follow the geometric distribution with success probability $\frac{n}{n+\theta}$, so that

$$\mathbb{P}(M = m) = \left(\frac{\theta}{n+\theta}\right)^m \frac{n}{n+\theta}.$$

Under this model, the corresponding approximate conditional sampling probability is easy to compute in closed form:

$$\hat{\pi}_{\text{SD}}(\alpha \mid \boldsymbol{n}) = \sum_{\beta\in E}\frac{n_\beta}{n}\sum_{m=0}^{\infty}\left(\frac{\theta}{\theta+n}\right)^m \frac{n}{\theta+n}[\mathbf{P}^m]_{\beta\alpha}. \tag{6.17}$$

The above expression has a genealogical interpretation. First, note that

$$\left(\frac{\theta}{\theta+n}\right)^m \frac{n}{\theta+n} = \int_0^\infty ne^{-nt}\left[\frac{1}{m!}(\theta t)^m e^{-\theta t}\right]dt,$$

which is the probability in a Poisson point process with intensity θ that there are m points in a random interval $[0, X]$ where $X \sim \text{Exp}(n)$. The random variable X can be interpreted as the waiting time until the lineage corresponding to the additionally sampled allele α coalesces with one of the n lineages corresponding to $\mathbf{n}$ (the n lineages are assumed to extend to infinity without mutating or coalescing among them). The probability that the

additional lineage coalesces with a lineage corresponding to type β is n_β/n. Conditioned on $X = t$, the number of mutations hitting the additional lineage is distributed as a Poisson random variable with mean θt. The mutation rate in this interpretation is θ instead of $\theta/2$, to make up for the fact that the lineages corresponding to $\mathbf{n}$ are assumed to be static (in particular, they do not mutate). In summary, we can rewrite (6.17) into a genealogically interpretable form as

$$\hat{\pi}_{\mathrm{SD}}(\alpha \mid \boldsymbol{n}) = \sum_{\beta \in E} \frac{n_\beta}{n} \int_0^\infty n e^{-nt} \left[e^{\theta t (\mathbf{P} - \mathbf{I})} \right]_{\beta\alpha} dt,$$

where we have used the identity

$$\sum_{m=0}^{\infty} \frac{1}{m!} (\theta t)^m e^{-\theta t} [\mathbf{P}^m]_{\beta\alpha} = \left[e^{\theta t (\mathbf{P} - \mathbf{I})} \right]_{\beta\alpha}. \tag{6.18}$$

As detailed below, the approximation $\hat{\pi}_{\mathrm{SD}}$ has several nice properties. First, it is actually equivalent to the true conditional sampling distribution π under the PIM model, in which $(P_{\beta\alpha}) = (\varphi_\alpha)$ for all $\alpha, \beta \in E$, and (6.7) implies

$$\pi(\alpha \mid \mathbf{n}) = \frac{n_\alpha + \theta \varphi_\alpha}{n + \theta}. \tag{6.19}$$

Proposition 6.9. *Under a PIM model, $\hat{\pi}_{SD}(\alpha \mid \mathbf{n}) = \pi(\alpha \mid \mathbf{n})$ for all sample configuration $\mathbf{n}$ and allele $\alpha \in E$.*

Proof. Under a PIM model, the transition matrix $\mathbf{P}$ takes the form

$$\mathbf{P} = \begin{pmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_K \\ \varphi_1 & \varphi_2 & \cdots & \varphi_K \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_1 & \varphi_2 & \cdots & \varphi_K \end{pmatrix}.$$

So, since $\sum_{\alpha \in E} \varphi_\alpha = 1$, $\mathbf{P}^m = \mathbf{P}$ for all positive integers $m \in \mathbb{N}$, while $[\mathbf{P}^0]_{\beta\alpha} = \delta_{\beta\alpha}$. Therefore,

$$\begin{aligned} \hat{\pi}_{\mathrm{SD}}(\alpha \mid \mathbf{n}) &= \sum_{\beta \in E} \frac{n_\beta}{n} \left[\frac{n}{n+\theta} \delta_{\beta\alpha} + \varphi_\alpha \sum_{m=1}^{\infty} \left(\frac{\theta}{\theta + n} \right)^m \frac{n}{n+\theta} \right] \\ &= \frac{n_\alpha}{n+\theta} + \varphi_\alpha \sum_{m=1}^{\infty} \left(\frac{\theta}{\theta + n} \right)^m \frac{n}{n+\theta} \\ &= \frac{n_\alpha}{n+\theta} + \frac{n}{n+\theta} \varphi_\alpha \left(\frac{1}{1 - \frac{\theta}{n+\theta}} - 1 \right) \\ &= \frac{n_\alpha + \theta \varphi_\alpha}{n+\theta}, \end{aligned}$$

which is equal to (6.19). □

Proposition 6.10. *Suppose $|\mathbf{n}| = 1$; i.e., $\mathbf{n} = \mathbf{e}_\alpha$ for some $\alpha \in E$. If $\boldsymbol{P}$ is reversible, then $\hat{\pi}_{\boldsymbol{SD}}(\alpha \mid \mathbf{e}_\beta) = \pi(\alpha \mid \mathbf{e}_\beta)$ for all $\alpha, \beta \in E$.*

Proof. First, note that

$$\pi(\alpha \mid \mathbf{e}_\beta) = \frac{q_2(\mathbf{e}_\beta + \mathbf{e}_\alpha)}{q_1(\mathbf{e}_\beta)} = \left(\frac{\delta_{\alpha\beta} + 1}{2}\right) \frac{p_2(\mathbf{e}_\beta + \mathbf{e}_\alpha)}{p_1(\mathbf{e}_\beta)}. \tag{6.20}$$

The ordered case is simpler, but the unordered case helps to understand why the combinatorial coefficient in (4.15) is necessary, so we will work with the latter case. Let $(\varphi_\beta)_{\beta\in E}$ denote the unique stationary distribution of $\mathbf{P}$. Then, $p_1(\mathbf{e}_\beta) = \varphi_\beta$. To compute $p_2(\mathbf{e}_\beta + \mathbf{e}_\alpha)$, consider a coalescent tree for sample size 2, and let m_i denote the number of mutations from leaf i to the root. Then, given m_1 and m_2, the conditional probability of observing alleles $\alpha, \beta \in E$ at the leaves is

$$\begin{aligned} p_2(\mathbf{e}_\beta + \mathbf{e}_\alpha \mid m_1, m_2) &= (2 - \delta_{\alpha\beta}) \sum_{\gamma\in E} \varphi_\gamma [\mathbf{P}^{m_1}]_{\gamma\beta} [\mathbf{P}^{m_2}]_{\gamma\alpha} \\ &= (2 - \delta_{\alpha\beta}) \sum_{\gamma\in E} \varphi_\beta [\mathbf{P}^{m_1}]_{\beta\gamma} [\mathbf{P}^{m_2}]_{\gamma\alpha} \\ &= (2 - \delta_{\alpha\beta}) \varphi_\beta [\mathbf{P}^{m_1+m_2}]_{\beta\alpha}, \end{aligned}$$

where the symmetry factor $(2 - \delta_{\alpha\beta})$ comes from assigning α and β to the two leaves (since we are dealing with an unordered sampling distribution), and the second line follows from the reversibility of $\mathbf{P}$, which implies $\varphi_\gamma P_{\gamma\beta} = \varphi_\beta P_{\beta\gamma}$.

According to (3.1), the probability that there are m mutations in a coalescent tree with 2 leaves is

$$\left(\frac{\theta}{1+\theta}\right)^m \frac{1}{1+\theta}.$$

Hence, (6.20) can be written as

$$\begin{aligned} \pi(\alpha \mid \mathbf{e}_\beta) &= \left(\frac{\delta_{\alpha\beta} + 1}{2}\right) \frac{1}{\varphi_\beta} \sum_{m=0}^{\infty} \left(\frac{\theta}{1+\theta}\right)^m \frac{1}{1+\theta} (2 - \delta_{\alpha\beta}) \varphi_\beta [\mathbf{P}^m]_{\beta\alpha} \\ &= \sum_{m=0}^{\infty} \left(\frac{\theta}{1+\theta}\right)^m \frac{1}{1+\theta} [\mathbf{P}^m]_{\beta\alpha}, \end{aligned}$$

which is equal to $\hat{\pi}_{\mathrm{SD}}(\alpha \mid \mathbf{e}_\beta)$. □

Proposition 6.11. *The conditional sampling distribution $\hat{\pi}_{SD}$ satisfies*

$$\hat{\pi}_{SD}(\alpha \mid \mathbf{n}) = \sum_{\beta\in E} \hat{\pi}_{SD}(\beta \mid \mathbf{n}) \frac{n_\alpha + \theta P_{\beta\alpha}}{n + \theta}. \tag{6.21}$$

Proving this result is left as an exercise. This property implies that $\hat{\pi}_{\mathrm{SD}}$ is a stationary distribution of the Markov chain on E with transition matrix

$$T_{\beta\alpha} = \frac{n_\alpha + \theta P_{\beta\alpha}}{n + \theta}.$$

Under the PIM model, note that $T_{\beta\alpha} = \pi(\alpha \mid \mathbf{n})$. De Iorio and Griffiths (2004) proposed an alternate approach to constructing an approximation to the CSD π by using the generator

of the Wright-Fisher diffusion, which is dual to the coalescent process. They showed that the approximation of Stephens and Donnelly that $\hat{\pi}_{\mathrm{SD}}$ satisfies the stationary condition (6.21) is equivalent to the diffusion-generator approximation that De Iorio and Griffiths developed.

Stephens and Donnelly (2000) showed that if their approximation $\hat{\pi}_{\mathrm{SD}}$ is used in (6.16), then the probability that allele type α is involved in the transition $H_t \to H_{t-1}$ is n_α/n. This suggests a more efficient sampling scheme that avoids having to compute $\mathbb{P}(H_{t-1} \mid H_t)$ for all possible H_{t-1}:

1. Choose an allele uniformly at random from H_t. Denote this allele by α.
2. Then, set $H_{t-1} = H_t - \alpha + \beta$ with probability proportional to $\theta P_{\beta\alpha}\hat{\pi}_{\mathrm{SD}}(\beta \mid H_t - \alpha)$, or set $H_{t-1} = H_t - \alpha$ with probability proportional to $n_\alpha - 1$.

We have seen that $\hat{\pi}_{\mathrm{SD}}$ has several nice properties, but it also has an undesirable property, namely that it violates exchangeability in general. Specifically, the probability of a sample configuration should not depend on the order in which the elements are sampled; formally,

$$\pi(\alpha \mid \mathbf{n})\pi(\beta \mid \mathbf{n} + \mathbf{e}_\alpha) = \pi(\beta \mid \mathbf{n})\pi(\alpha \mid \mathbf{n} + \mathbf{e}_\beta).$$

However, for a general transition matrix $\mathbf{P}$, $\hat{\pi}_{\mathrm{SD}}$ does not satisfy the above condition. Despite this problem, $\hat{\pi}_{\mathrm{SD}}$ has proved useful in approximating the optimal proposal distribution Q^* in sequential importance sampling.

6.5.2 Generalization to multiple sites

Consider a non-recombining locus with L sites each with a finite type space E and per-site mutation rate $\theta/2$. The allele type of the entire locus is a haplotype $\boldsymbol{x} \in E^L$, and we can generalize $\hat{\pi}_{\mathrm{SD}}$ to approximate the conditional probability $\pi(\boldsymbol{x} \mid \mathbf{n})$ that an additionally sampled haplotype is of type $\boldsymbol{x} \in E^L$ given that a sample with configuration $\mathbf{n} = (n_{\boldsymbol{x}})_{\boldsymbol{x}\in E^L}$ has been already observed. We define $n = \sum_{\boldsymbol{x}\in E} n_{\boldsymbol{x}}$.

Similar to the single-site case, one considers an urn containing $n_{\boldsymbol{x}}$ copies of haplotype $\boldsymbol{x} \in E^L$, sample a haplotype uniformly at random from the urn, and then mutate it a random number M of time according to the transition matrix $\mathbf{P}$. Here, M is assumed to follow the geometric distribution with success probability $\frac{n}{n+L\theta}$:

$$\mathbb{P}(M = m) = \left(\frac{L\theta}{n + L\theta}\right)^m \frac{n}{n + L\theta}.$$

In the L-site model, these M mutations are distributed uniformly into the L sites; that is, letting M_ℓ denote the number of times that site l is mutated, the joint distribution of $(M_1, \ldots, M_L)$ is given by

$$\mathbb{P}((M_1, \ldots, M_L) = (m_1, \ldots, m_L)) = \mathbb{I}\{m_1 + \cdots + m_L = m\}\frac{1}{L^m}\binom{m}{m_1, \ldots, m_L}.$$

Combining everything, the approximate conditional sampling distribution proposed by Stephens and Donnelly (2000) is

$$\hat{\pi}_{\mathrm{SD}}(\boldsymbol{x} \mid \mathbf{n}) = \sum_{\boldsymbol{x}' \in E^L} \frac{n_{\boldsymbol{x}'}}{n} \sum_{m_1=0}^{\infty} \cdots \sum_{m_L=0}^{\infty} \binom{m}{m_1, \ldots, m_L} \frac{1}{L^m} \frac{n}{n + L\theta} \prod_{\ell=1}^{L} \left(\frac{L\theta}{L\theta + n} \right)^{m_\ell} [\mathbf{P}^{m_\ell}]_{x'_\ell x_\ell}, \tag{6.22}$$

where $m = m_1 + \cdots + m_\ell$, $\boldsymbol{x} = (x_1, \ldots, x_L)$, and $\boldsymbol{x}' = (x'_1, \ldots, x'_L)$. Now, using the identity

$$\begin{aligned} &\binom{m}{m_1, \ldots, m_L} \frac{1}{L^m} \left(\frac{n}{n + L\theta} \right) \prod_{\ell=1}^{L} \left(\frac{L\theta}{L\theta + n} \right)^{m_\ell} [\mathbf{P}^{m_\ell}]_{x'_\ell x_\ell} \\ &\qquad = \int_0^\infty n e^{-nt} \left[\prod_{\ell=1}^{L} e^{-\theta t} \frac{1}{m_\ell!} (\theta t)^{m_\ell} [\mathbf{P}^{m_\ell}]_{x'_\ell x_\ell} \right] dt \end{aligned}$$

and (6.18), we can rewrite (6.22) as

$$\hat{\pi}_{\mathrm{SD}}(\boldsymbol{x} \mid \mathbf{n}) = \sum_{\boldsymbol{x}' \in E^L} \frac{n_{\boldsymbol{x}'}}{n} \int_0^\infty n e^{-nt} \prod_{\ell=1}^{L} \left[e^{\theta t (\mathbf{P} - \mathbf{I})} \right]_{x'_\ell x_\ell} dt.$$

6.6 The infinite-sites model revisited

Before we conclude this chapter, we briefly revisit the infinite-sites model. Sequential importance sampling works differently under the latter model. We mentioned above that if the approximation $\hat{\pi}_{\mathrm{SD}}$ is used in (6.16), then the probability that allele type α is involved in the transition $H_t \to H_{t-1}$ is n_α/n. Recall that the same probability also applies to the infinite-alleles model; see (4.10) and the discussion therein. Under that model, one event back in time can be sampled from the correct posterior distribution by picking an allele uniformly at random from the current configuration, since the event is completely determined by the allele type chosen: If the chosen allele is a singleton, then it gets killed by a mutation event. If it has copy number greater than one, then it coalesces with another allele of the same type.

Motivated by these results, the following importance sampling scheme has been proposed for the infinite-sites model: Suppose the root sequence is unknown and consider the unrooted perfect phylogeny (cf., Figure 5.8) corresponding to an input data set. Consider the set of all haplotypes h_α with (a) multiplicity $n_\alpha \geq 2$, or (b) $n_\alpha = 1$ and the degree of its corresponding vertex in the unrooted perfect phylogeny is 1. Then, choose a haplotype uniformly at random from this set. This uniquely determines the event to be sampled: If the multiplicity of the chosen haplotype is > 1, then it is involved in a coalescence event. Otherwise, a singleton mutation is removed from the haplotype.

See Hobolth et al (2008) for an improved version of SIS for the infinite-sites model.

6.7 Posterior probability of the first event back in time

Assume a constant population size, and consider the coalescent for a K-allelic model with mutation rate $\theta/2$ and transition matrix $\mathbf{P} = (P_{ij})_{i,j \in [K]}$. Let $\boldsymbol{n} = (n_1, \ldots, n_K)$ denote the

configuration of a sample of size $n = n_1 + \cdots + n_K$. Using what has been covered so far in this chapter, we would like to address the following two questions:

1. Let T denote the time to the first event back in time. What is the posterior distribution of T given $\boldsymbol{n}$?
2. Assume that the mutation model is PIM. Let C_α denote the event that "the first event back in time is a coalescence between two lineages of type α". Further, let M_α denote the event that "the first event back in time is a mutation in a lineage of type α". Find $\mathbb{P}(C_\alpha \mid \boldsymbol{n})$ and $\mathbb{P}(M_\alpha \mid \boldsymbol{n})$.

The reader is strongly encouraged to think about these problems before taking a look at the solutions below.

Answer to Question 1: This is a trick question. It so happens that T is independent of $\boldsymbol{n}$, and therefore $\mathbb{P}(T \in dt \mid \boldsymbol{n}) = \lambda e^{-\lambda t} dt$, where $\lambda = \binom{n}{2} + \frac{n\theta}{2}$ and $t > 0$.

We check the independence. Let $\boldsymbol{u}$ be the full ordered sample, and let $\boldsymbol{v}$ be the full ordered sample after the first event. Let X be the first event back in time (so it is a coalescence for a specific pair, or a mutation on a specific individual).

We can generate $\boldsymbol{u}$ as follows: wait for exponential time T with rate λ. When the time hits, we select what event X occurs independently of T, with probability proportional to the rate at which that event occurs (for example, each pairwise coalescence event occurs with rate 1). We can do this because superpositions of Poisson processes are still Poisson processes; that is, the above is equivalent to "racing" the waiting times of each event.

Once we have selected the time T and the event X, we generate $\boldsymbol{v}$ by sampling the rest of the coalescent tree and dropping mutations on it. Finally, to generate $\boldsymbol{u}$, there are two cases: if X is a coalescence then $\boldsymbol{u}$ is determined from $\boldsymbol{v}$, otherwise we generate $\boldsymbol{u}$ from $\boldsymbol{v}$ according to $\mathbf{P}$. Summarizing, we have

$$\mathbb{P}(T, X, \boldsymbol{v}, \boldsymbol{u}) = \mathbb{P}(T)\mathbb{P}(X)\mathbb{P}(\boldsymbol{v} \mid X)\mathbb{P}(\boldsymbol{u} \mid X, \boldsymbol{v}) = f(T)g(X, \boldsymbol{v}, \boldsymbol{u}).$$

So T is independent of $\boldsymbol{u}$, and hence also independent of $\boldsymbol{n} = \boldsymbol{n}(\boldsymbol{u})$.

Answer to Question 2: Let C denote the event that the first event back in time is a coalescence. Then, the recursion shown in (6.1) implies

$$\mathbb{P}(C \mid \boldsymbol{n}) = \frac{\mathbb{P}(\boldsymbol{n} \mid C)\mathbb{P}(C)}{p_n(\boldsymbol{n})} = \frac{n-1}{n-1+\theta} \sum_{\alpha=1}^{K} \frac{n_\alpha - 1}{n-1} \cdot \frac{p_{n-1}(\boldsymbol{n} - \boldsymbol{e}_\alpha)}{p_n(\boldsymbol{n})}, \tag{6.23}$$

which can be written as

$$\mathbb{P}(C \mid \boldsymbol{n}) = \frac{1}{n(n-1+\theta)} \sum_{\alpha=1}^{K} \frac{n_\alpha(n_\alpha - 1)}{\pi(\alpha \mid \boldsymbol{n} - \boldsymbol{e}_\alpha)},$$

where

$$\pi(\alpha \mid \boldsymbol{n} - \boldsymbol{e}_\alpha) = \frac{n_\alpha}{n} \frac{p_n(\boldsymbol{n})}{p_{n-1}(\boldsymbol{n} - \boldsymbol{e}_\alpha)}. \tag{6.24}$$

Similarly, the posterior probability of the event C_α is

$$\mathbb{P}(C_\alpha \mid \boldsymbol{n}) = \frac{1}{n(n-1+\theta)} \frac{n_\alpha(n_\alpha - 1)}{\pi(\alpha \mid \boldsymbol{n} - \boldsymbol{e}_\alpha)}.$$

Using (6.24) and

$$\pi(\beta \mid \boldsymbol{n} - \boldsymbol{e}_\alpha) = \frac{n_\beta + 1 - \delta_{\alpha\beta}}{n} \frac{p_n(\boldsymbol{n} - \boldsymbol{e}_\alpha + \boldsymbol{e}_\beta)}{p_{n-1}(\boldsymbol{n} - \boldsymbol{e}_\alpha)}, \tag{6.25}$$

one can show that the posterior probability of M_α is given by

$$\mathbb{P}(M_\alpha \mid \boldsymbol{n}) = \frac{\theta}{n(n-1+\theta)} \sum_\beta P_{\beta\alpha} n_\alpha \frac{\pi(\beta \mid \boldsymbol{n} - \boldsymbol{e}_\alpha)}{\pi(\alpha \mid \boldsymbol{n} - \boldsymbol{e}_\alpha)}.$$

Under the PIM model where $P_{\beta\alpha} = \varphi_\alpha$,

$$\pi(\alpha \mid \boldsymbol{n} - \boldsymbol{e}_\alpha) = \frac{n_\alpha - 1 + \theta\varphi_\alpha}{n - 1 + \theta},$$

so

$$\mathbb{P}(C \mid \boldsymbol{n}) = \frac{1}{n} \sum_{\alpha=1}^{K} \frac{n_\alpha(n_\alpha - 1)}{n_\alpha - 1 + \theta\varphi_\alpha},$$

while

$$\mathbb{P}(C_\alpha \mid \boldsymbol{n}) = \frac{n_\alpha(n_\alpha - 1)}{n(n_\alpha - 1 + \theta\varphi_\alpha)},$$
$$\mathbb{P}(M_\alpha \mid \boldsymbol{n}) = \frac{n_\alpha \theta\varphi_\alpha}{n(n_\alpha - 1 + \theta\varphi_\alpha)}.$$

6.8 Closed-form asymptotic sampling formulae for small θ

As mentioned above, finding an exact, closed-form sampling formula for non-PIM models has remained a challenging open problem. However, it is possible to derive approximate, closed-form sampling formulae that are very accurate when θ is small, by considering the Taylor expansion of the sampling probability $p_n(\boldsymbol{n} \mid \theta, \mathbf{P})$ about $\theta = 0$ (Bhaskar et al, 2012; Jenkins and Song, 2011). If $\mathbf{P}$ is irreducible when restricted to the set $\mathcal{O}_\mathbf{n}$ of distinct observed alleles in the sample $\mathbf{n}$, then the leading order term in the expansion is proportional to $\theta^{|\mathcal{O}_\mathbf{n}|-1}$. Hence,

$$p_n(\mathbf{n} \mid \theta, \mathbf{P}) = \theta^{|\mathcal{O}_\mathbf{n}|-1} f(\mathbf{n} \mid \mathbf{P}) + O(\theta^{|\mathcal{O}_\mathbf{n}|}), \tag{6.26}$$

where $f(\mathbf{n} \mid \mathbf{P})$ is the leading order coefficient that depends on the mutation transition matrix $\mathbf{P}$ but not on the mutation rate θ.

By decomposing the set of possible mutation events in the coalescent genealogy into different classes, Jenkins and Song (2011) obtained closed-form formulae for $f(\mathbf{n} \mid \mathbf{P})$ for a general transition matrix $\mathbf{P}$ when $|\mathcal{O}_\mathbf{n}| \leq 3$. Bhaskar et al (2012) later provided alternate proofs using martingale arguments and an urn construction related to the coalescent, which led to a recursion that could be solved in closed-form using combinatorial techniques. Furthermore, Bhaskar et al (2012) extended the results to obtain a closed-form formula for $f(\mathbf{n} \mid \mathbf{P})$ when $|\mathcal{O}_\mathbf{n}| = 4$ and the transition matrix $\mathbf{P}$ is reversible restricted to $\mathcal{O}_\mathbf{n}$. We summarize these results below, where we use h_k to denote the kth harmonic number:

$$h_k = \sum_{j=1}^{k} \frac{1}{j}.$$

Theorem 6.12 (Bi-allelic observation). *Consider a sample configuration* $\mathbf{n} = (n_\alpha)_{\alpha\in E}$ *with* $|\mathbf{n}| = n$ *and* $|\mathcal{O}_\mathbf{n}| = 2$*. For an arbitrary mutation transition matrix* $\mathbf{P}$ *that is irreducible when restricted to* $\mathcal{O}_\mathbf{n}$*, the sampling probability* $p_n(\mathbf{n})$ *is given by (Bhaskar et al, 2012; Jenkins and Song, 2011)*

$$p_n(\mathbf{n}) = \theta \sum_{a,b\in\mathcal{O}_\mathbf{n}: a\neq b} \frac{\pi_a P_{ab}}{n_b} + O(\theta^2).$$

Theorem 6.13 (Tri-allelic observation). *Consider a sample configuration* $\mathbf{n} = (n_\alpha)_{\alpha\in E}$ *with* $|\mathbf{n}| = n$ *and* $|\mathcal{O}_\mathbf{n}| = 3$*. For an arbitrary mutation transition matrix* $\mathbf{P}$ *that is irreducible when restricted to* $\mathcal{O}_\mathbf{n}$*, then* $p_n(\mathbf{n})$ *is given by (Bhaskar et al, 2012; Jenkins and Song, 2011)*

$$p_n(\mathbf{n}) = \theta^2 \sum_{\text{distinct } a,b,c\in\mathcal{O}_\mathbf{n}} \left\{ \varphi_a P_{ab} P_{ac} \left[\frac{1}{n_c(n_b+n_c)} - d(n_a,n_b,n_c) \right] + \varphi_a P_{ab} P_{bc} d(n_a,n_b,n_c) \right\} + O(\theta^3)$$

where

$$d(n_a,n_b,n_c) = \frac{1}{(n_a+n_b)(n_a+n_b-1)} \left[1 + \frac{n}{n_c} - \frac{2n(h_n - h_{n_c-1})}{n_a+n_b+1} \right].$$

Corollary 6.14 (Tri-allelic observation under a reversible model). *Suppose* $|\mathcal{O}_\mathbf{n}| = 3$ *with sample configuration* $\mathbf{n} = n_a\mathbf{e}_a + n_b\mathbf{e}_b + n_c\mathbf{e}_c$*, where* a, b, c *are distinct alleles in* E*. If the mutation transition matrix* $\mathbf{P}$ *is reversible and irreducible when restricted to the observed alleles* $\mathcal{O}_\mathbf{n}$*, then* $p_n(\mathbf{n})$ *is given by (Bhaskar et al, 2012)*

$$p_n(\mathbf{n}) = \theta^2 \left(\frac{\pi_a P_{ab} P_{ac}}{n_b n_c} + \frac{\pi_b P_{ba} P_{bc}}{n_a n_c} + \frac{\pi_c P_{ca} P_{cb}}{n_a n_b} \right) + O(\theta^3).$$

Theorem 6.15 (Quadra-allelic observation under a reversible model). *Consider a sample configuration* $\mathbf{n} = (n_\alpha)_{\alpha\in E}$ *with* $|\mathbf{n}| = n$ *and* $|\mathcal{O}_\mathbf{n}| = 4$*. If the mutation transition matrix* $\mathbf{P}$ *is reversible and irreducible when restricted to the observed alleles* $\mathcal{O}_\mathbf{n}$*, then* $p_n(\mathbf{n})$ *is given by (Bhaskar et al, 2012)*

$$p_n(\mathbf{n}) = \theta^3 \sum_{\text{distinct } a,b,c,d\in\mathcal{O}_\mathbf{n}} [\pi_a P_{ab} P_{ac} P_{ad} \gamma(\boldsymbol{n},a,b,c,d) + \pi_a P_{ab} P_{ac} P_{bd} \delta(\boldsymbol{n},a,b,c,d)] + O(\theta^4),$$

where

$$\gamma(\boldsymbol{n},a,b,c,d) = \frac{1}{n_b n_c n_d} \left\{ \left[\frac{n_a - 1}{2(n_a+n_b+n_c-1)} - \frac{2n_b n_d}{(n_a+n_b+n_c)_{2\downarrow}} \right] + \frac{n_d}{2(n_b+n_c+n_d)} - \left[\frac{n_d(n_a-1)}{(n_c+n_d)(n_a+n_b-1)} - \frac{2n_b n_d}{(n_a+n_b)_{2\downarrow}} \right] \right\} + \frac{2n}{n_c(n_a+n_b+n_c+1)_{3\downarrow}} (h_n - h_{n_d-1}) - \frac{2n}{n_c(n_a+n_b+1)_{3\downarrow}} (h_n - h_{n_c+n_d-1}),$$

and

$$\delta(\boldsymbol{n}, a, b, c, d) = \frac{1}{n_b n_c n_d}\left\{\left[\frac{n_b}{n_a+n_b+n_c-1} + \frac{2n_b n_d}{(n_a+n_b+n_c)_{2\downarrow}}\right] - \left[\frac{n_b n_d}{(n_c+n_d)(n_a+n_b-1)} + \frac{2n_b n_d}{(n_a+n_b)_{2\downarrow}}\right]\right\} - \frac{2n}{n_c(n_a+n_b+n_c+1)_{3\downarrow}}(h_n - h_{n_d-1}) + \frac{2}{n_c(n_a+n_b+1)_{3\downarrow}}(h_n - h_{n_c+n_d-1}).$$

6.9 How many triallic sites do we expect to see in a sample of n genomes?

Here, we utilize asymptotic sampling formulae to obtain a rough approximation of the expected number of triallic sites in a sample of n genomes. The results will be expressed in terms of harmonic numbers, for which we use the following notation:

$$H_n = \sum_{j=1}^{n} \frac{1}{j}, \qquad \text{and} \qquad H_n^{(2)} = \sum_{j=1}^{n} \frac{1}{j^2}.$$

Further, let $c_n^{(s)}$ denote the sth order generalized harmonic number (Roman, 1993), defined for $s \geq 0$ and $n \geq 1$ by

$$c_n^{(s)} = \begin{cases} 1, & \text{if } s = 0, \\ \sum_{j=1}^{n} \frac{c_j^{(s-1)}}{j}, & \text{if } s > 0. \end{cases}$$

In particular,

$$c_n^{(1)} = H_n \;\text{ and }\; c_n^{(2)} = \sum_{j=1}^{n} \frac{H_j}{j} = \frac{1}{2}\left[(H_n)^2 + H_n^{(2)}\right]. \tag{6.27}$$

In what follows, we assume the coalescent under a constant population size.

Theorem 6.16 (Jenkins and Song 2011). *Let $M_{n,s}$ denote the event that there are exactly s mutation events in the history of a sample of size n. Then,*

$$\mathbb{P}(M_{n,s}) = \sum_{j=0}^{\infty} \theta^{s+j} c_{n-1}^{(s+j)} (-1)^j \binom{s+j}{j}. \tag{6.28}$$

The series converges for $\theta < 1$.

It follows from (6.27) and (6.28) that

$$\mathbb{P}(M_{n,0}) = 1 - \theta H_{n-1} + O(\theta^2),$$

and that the probability of the event $\mathit{\Pi}_n$ that a particular site is polymorphic in a sample of size n genomes is

$$\mathbb{P}(\Pi_n) = \theta H_{n-1} + O(\theta^2). \tag{6.29}$$

Suppose the pattern of mutation is governed by an ergodic, irreducible Markov chain on $\Gamma = \{A, C, G, T\}$, with transition matrix $\mathbf{P} = (P_{ji})$, where P_{ji} denotes the probability of allele j mutating to allele i forward in time given that a mutation occurs. The stationary distribution of $\mathbf{P}$ is denoted by $(\pi_A, \pi_C, \pi_G, \pi_T)$.

Theorem 6.17 (Bhaskar et al 2012; Jenkins and Song 2011). *Given a sample of size n, denote by α the allele of the most recent common ancestor to the sample. Let $O_{n,3;a}$ denote the event of a triallelic polymorphism at a given site, with one of the observed alleles being $a \in \Gamma$. Then,*

$$\mathbb{P}(O_{n,3;a}, M_{n,2} \mid \alpha = a) = \theta^2 \Bigg\{ [1 - (\mathbf{P}^2)_{aa}] \left(H_n + \frac{1}{n} - 2 \right) + [1 - (\mathbf{P}\mathbf{P}^T)_{aa}] \left[\frac{(H_{n-1})^2}{2} - \frac{H_{n-1}^{(2)}}{2} - H_n - \frac{1}{n} + 2 \right] \Bigg\} + O(\theta^3).$$

Corollary 6.18. *Let $O_{n,3}$ denote the event of a triallelic polymorphism at a given site for a sample of n genomes. Then, Theorem 6.17 implies*

$$\begin{aligned} \mathbb{P}(O_{n,3}) &= \sum_{a \in \Gamma} \mathbb{P}(O_{n,3;a}, M_{n,2} \mid \alpha = a) \cdot \pi_a + O(\theta^3) \\ &= \sum_{a \in \Gamma} \theta^2 \Bigg\{ [1 - (\mathbf{P}^2)_{aa}] \left(H_n + \frac{1}{n} - 2 \right) \\ &\qquad + [1 - (\mathbf{P}\mathbf{P}^T)_{aa}] \left[\frac{(H_{n-1})^2}{2} - \frac{H_{n-1}^{(2)}}{2} - H_n - \frac{1}{n} + 2 \right] \Bigg\} + O(\theta^3). \end{aligned} \tag{6.30}$$

Consider the mutation transition matrix

$$\mathbf{P} = \begin{pmatrix} 0.503 & 0.082 & 0.315 & 0.100 \\ 0.186 & 0.002 & 0.158 & 0.655 \\ 0.654 & 0.158 & 0.000 & 0.189 \\ 0.097 & 0.303 & 0.085 & 0.515 \end{pmatrix},$$

which was estimated from human data, and suppose $\theta = 0.0014$. For these parameters, (6.29) implies that the probability of a given site in a sample of size $n = 2000$ genomes being polymorphic is

$$\mathbb{P}(\Pi_n) = 0.011,$$

while (6.30) implies $\mathbb{P}(O_{n,3}) = 3.8 \times 10^{-5}$. Combining the two results gives $\mathbb{P}(O_{n,3} \mid \Pi_n) = 0.003$. For genomes of length 3×10^9, we conclude that the expected number of triallelic sites in a sample of size $n = 2000$ is about

$$3 \times 10^9 \times \mathbb{P}(O_{n,3}) \approx 1.1 \times 10^5.$$

References

Bahlo M, Griffiths RC (2000) Inference from gene trees in a subdivided population. Theoretical Population Biology 57:79–95

Bhaskar A, Kamm JA, Song YS (2012) Approximate sampling formulae for general finite-alleles models of mutation. Advances in Applied Probability 44:408–428, (PMC3953561)

Coop G, Griffiths RC (2004) Ancestral inference on gene trees under selection. Theoretical Population Biology 66(3):219–232

De Iorio M, Griffiths RC (2004) Importance sampling on coalescent histories. I. Adv Appl Prob 36:417–433

Griffiths RC, Tavaré S (1994a) Ancestral inference in population genetics. Stat Sci 9:307–319

Griffiths RC, Tavaré S (1994b) Sampling theory for neutral alleles in a varying environment. Proc R Soc London B 344:403–410

Griffiths RC, Tavaré S (1994c) Simulating probability distributions in the coalescent. Theoretical Population Biology 46:131–159

Hobolth A, Uyenoyama M, Wiuf C (2008) Importance sampling for the infinite sites model. Statistical Applications in Genetics and Molecular Biology 7(1):32

Jenkins PA, Song YS (2011) The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. Theoretical Population Biology 80(2):158–173, (PMC3143209)

Larribe F, Lessard S, Schork NJ (2002) Gene mapping via the ancestral recombination graph. Theoretical Population Biology 62(2):215–229

Roman S (1993) The harmonic logarithms and the binomial formula. J Comb Theory A 63:143–163

Stephens M, Donnelly P (2000) Inference in molecular population genetics. JR Stat Soc Ser B 62:605–655

Part III
Demography

Chapter 7
Variable population size

In this chapter, we relax the assumption that the population size remains constant over time. This is a necessary extension for modeling any real population. When the population size is variable, inter-coalescence times of the n-coalescent are not independent of each other, thus introducing significant complication to analysis. However, we will see that the expected site frequency spectrum (SFS) can still be computed efficiently and numerically stably. We will also discuss the theoretical question of whether and under what conditions the expected SFS uniquely determines the population size function.

7.1 Discrete-time model

Consider the Wright-Fisher model and let $\mathcal{N}(\tau)$ denote the population size in generation $\tau = 0, 1, 2, \ldots$, with $\tau = 0$ corresponding to the present, $\tau = 1$ one generation back in time, and so on.

Definition 7.1 (Relative population size). For $t \in \mathbb{R}_{\geq 0}$, the relative population size $\eta_N(\mathrm{t})$ is defined as

$$\eta_N(t) = \frac{\mathcal{N}(\lfloor tN \rfloor)}{N}, \tag{7.1}$$

where N is some reference population size.

To take a large-N limit, we make the following assumptions:

1. $\mathcal{N}$ is a strictly positive deterministic function.
2. $\mathcal{N}(\tau)$ is sufficiently large for all τ.
3. $\lim_{N\to\infty} \eta_N(t) = \eta(t)$ exists for all $t \geq 0$.
4. $\eta(t) > 0$ for all $t \geq 0$. (Note that the function η is strictly greater than zero.)

Example 7.2. Suppose $\mathcal{N}(\tau) = \left\lfloor (1-\frac{\alpha}{N})^{\tau} N \right\rfloor$, so that the population size decays exponentially backwards in time. Then, $\lim_{N\to\infty} \eta_N(t) = e^{-\alpha t}$, where α is a population-scaled decay rate.

The n-coalescent $\{C_n(t), t \geq 0\}$ obtained in the $N \to \infty$ limit is a time-inhomogeneous $\mathcal{P}_{[n]}$-valued Markov process where the instantaneous rate of coalescence between a pair of lineages at time t is given by $\frac{1}{\eta(t)}$. Intuitively, in smaller populations it is easier for lineages to

find a common ancestor and so the rate of coalescence is higher, while in larger populations the opposite is true. The embedded jump chain $\{\xi_{n,k}, k = n, n-1, \dots, 1\}$ is the same as in the constant population size case.

7.2 Inter-coalescence times and the ancestral process

As usual, we use $T_{n,k}$ to denote the waiting time while there are k ancestral lineages for a sample of size n taken at time 0. In the constant population size case with $\eta \equiv 1$, $T_{n,k}$ are independent exponential random variables with rate $\binom{k}{2}$. For non-constant η, the inter-coalescence times $T_{n,k}$ are no longer independent: depending on the previous inter-coalescence times, the rate of coalescence (determined by η) may be different.

We define the following notation, which will play an important role in the subsequent discussion:

Definition 7.3 (Time rescaling). Given a population size function η, we define

$$R_\eta(t) := \int_0^t \frac{1}{\eta(s)}\, ds, \tag{7.2}$$

which corresponds to the total intensity of pairwise coalescences up to time t

Note that this function is monotonically increasing and continuous, and hence invertible. This function will allow us to reparameterize time, thereby relating the time-inhomogeneous coalescent process with a time-homogeneous one.

We now discuss the distribution of inter-coalescence times, starting with $T_{2,2}$.

Proposition 7.4. *The probability density function of $T_{2,2}$ is given by*

$$f_{T_{2,2}}(t) = \frac{1}{\eta(t)} e^{-R_\eta(t)}. \tag{7.3}$$

Proof. We will show that $\mathbb{P}(T_{2,2} > t) = e^{-R_\eta(t)}$, differentiating which implies the desired result. In the discrete-time Wright-Fisher model,

$$\mathbb{P}(T^{\mathrm{WF}}_{2,2} > \lfloor Nt \rfloor) = \prod_{j=1}^{\lfloor Nt \rfloor} \left[1 - \frac{1}{\mathcal{N}(j)}\right], \tag{7.4}$$

where $1 - \frac{1}{\mathcal{N}(j)}$ is the probability that two individuals will find different parents one generation back. Taking the log of (7.4) and noting $x \le -\log(1-x) \le \frac{x}{1-x}$ for all $x \in (0,1)$, we obtain

$$\sum_{j=1}^{\lfloor Nt \rfloor} \frac{1}{\mathcal{N}(j)} \le -\sum_{j=1}^{\lfloor Nt \rfloor} \log\left[1 - \frac{1}{\mathcal{N}(j)}\right] \le \sum_{j=1}^{\lfloor Nt \rfloor} \frac{1}{\mathcal{N}(j) - 1} \tag{7.5}$$

Then, in the limit as $N \to \infty$, the left and right hand sides of (7.5) both converge to $R_\eta(t)$, so by the squeeze theorem we obtain

$$-\lim_{N\to\infty}\sum_{j=1}^{\lfloor Nt\rfloor}\log\left[1-\frac{1}{\mathcal{N}(j)}\right]=R_\eta(t). \tag{7.6}$$

Therefore, we conclude $\mathbb{P}(T_{2,2}>t)=e^{-R_\eta(t)}$. □

Remark 7.5. We want $\lim_{t\to\infty}\mathbb{P}(T_{2,2}\le t)=1$, so we require η to satisfy $\lim_{t\to\infty}R_\eta(t)=\infty$.

Proposition 7.6. *Define $\sigma_{n+1}=0$ and $\sigma_k=t_n+t_{n-1}+\cdots+t_k$ for $k=2,\dots,n$. The joint probability density function of $T_{n,n},\dots,T_{n,2}$ is given by*

$$f_{T_{n,n},\dots,T_{n,2}}(t_n,\dots,t_2)=\prod_{k=2}^{n}\binom{k}{2}\frac{1}{\eta(\sigma_k)}e^{-\binom{k}{2}[R_\eta(\sigma_k)-R_\eta(\sigma_{k+1})]}. \tag{7.7}$$

Proof. Noting that $T_{n,n},\dots,T_{n,2}$ is a Markov chain, we can decompose the joint distribution of $T_{n,n},\dots,T_{n,2}$ as a product of the following conditional distributions:

$$\begin{aligned}
&T_{n,n-1}\mid T_{n,n}=t_n\\
&T_{n,n-2}\mid (T_{n,n},T_{n,n-1})=(t_n,t_{n-1})\\
&\qquad\vdots\\
&T_{n,2}\mid (T_{n,n},\dots,T_{n,3})=(t_n,\dots,t_3)
\end{aligned}$$

Now note that the condition distribution of $T_{n,k}$ given $(T_{n,n},\dots,T_{n,k+1})=(t_n,\dots,t_{k+1})$ depends on $T_{n,n},\dots,T_{n,k+1}$ only through the sum $T_{n,n}+\cdots+T_{n,k+1}=t_n+\cdots+t_{k+1}=\sigma_{k+1}$. Further, this conditional distribution is equal to the distribution of the time to the first coalescence event in the coalescent process starting at time σ_{k+1} with k lineages. Therefore,

$$\begin{aligned}
f_{T_{n,k}\mid(T_{n,n},\dots,T_{n,k+1})=(t_n,\dots,t_{k+1})}(t_k)&=\binom{k}{2}\frac{1}{\eta(\sigma_k)}\exp\left[-\binom{k}{2}\int_{\sigma_{k+1}}^{\sigma_k}\frac{1}{\eta(s)}ds\right]\\
&=\binom{k}{2}\frac{1}{\eta(\sigma_k)}\exp\left\{-\binom{k}{2}[R(\sigma_k)-R(\sigma_{k+1})]\right\},
\end{aligned}$$

and combining these conditional densities yields the desired result. □

Let $\{A_n^{(\eta)}(t),t\ge 0\}$ denote the ancestral process for population size η, where $A_n^{(\eta)}(t)$ denotes the number of ancestral lineages at time t for a sample of size n taken at time 0. For non-constant η, $\{A_n^{(\eta)}(t),t\ge 0\}$ is a time-inhomogeneous Markov chain on $[n]$. For $0<h\ll 1$, note that

$$\begin{aligned}
\mathbb{P}(A_2^{(\eta)}(t+h)=1\mid A_2^{(\eta)}(t)=2)&=\mathbb{P}(T_{2,2}\le t+h\mid T_{2,2}>t)\\
&=\frac{\mathbb{P}(t<T_{2,2}\le t+h)}{\mathbb{P}(t>T_{2,2})}\\
&=\frac{e^{-R_\eta(t)}-e^{-R_\eta(t+h)}}{e^{-R_\eta(t)}}\\
&=\frac{1}{\eta(t)}h+o(h).
\end{aligned}$$

In general, for $n \geq 2$ and $0 < h \ll 1$,

$$\mathbb{P}(A_n^{(\eta)}(t+h) = j \mid A_n^{(\eta)}(t) = k) = \begin{cases} \binom{k}{2}\frac{1}{\eta(t)}h + o(h), & j = k-1, \\ 1 - \binom{k}{2}\frac{1}{\eta(t)}h + o(h), & j = k, \\ o(h), & \text{otherwise.} \end{cases} \tag{7.8}$$

What this implies is that we can relate $\{A_n^{(\eta)}(t), t \geq 0\}$ to the ancestral process for the constant population size model with $\eta \equiv 1$. More precisely, we have

$$A_n^{(\eta)}(t) \stackrel{d}{=} A_n^{(1)}(R_\eta(t)).$$

So, $\mathbb{P}(A_n^{(\eta)}(t) = j)$ is obtained from (1.8) by simply replacing $e^{-\binom{k}{2}t}$ with $e^{-\binom{k}{2}R_\eta(t)}$.

This suggests an algorithm for sampling inter-coalescence times $T_{n,n}^{(\eta)}, \ldots, T_{n,2}^{(\eta)}$ under a variable population size function η:

1. Simulate $T_n, \ldots, T_2$, where $T_k \sim \text{Exp}[\binom{k}{2}]$, and let $s_n, \ldots, s_2$ denote the sampled times.
2. Solve for t_n using $R_\eta(t_n) = s_n$ and then set $\sigma_n = t_n$.
3. For $k = n-1, \ldots, 2$, solve for t_k using $s_k = R_\eta(\sigma_{k+1} + t_k) - R_\eta(\sigma_{k+1})$ and then set $\sigma_k = \sigma_{k+1} + t_k$.

In general it is not possible to solve for t_k in closed form. An exception is when $\eta(t) = e^{-\alpha t}$, in which case $R_\eta(t) = \frac{1}{\alpha}(e^{\alpha t} - 1)$ and $t_k = \frac{1}{\alpha}\log(1 + \alpha s_k e^{-\alpha\sigma_{k+1}})$.

7.3 The expected SFS under variable population size

Recall that we use $\tau_{n,b}$ to denote the sum of the lengths of all edges each subtending exactly b leaves; see Definition 5.9 and Figure 5.2. The expected site frequency spectrum (SFS) can be easily found if we know $\mathbb{E}[\tau_{n,b}]$. Good news is that the formula for $\mathbb{E}[\tau_{n,b}]$ shown in (5.9) applies to an arbitrary population size function. However, computing the expectation $\mathbb{E}[T_{n,k}]$ is generally challenging, except in the case of a constant population size. For one thing, although the joint density of $T_{n,n}, \ldots, T_{n,2}$ is straightforward to write down (see (7.7)), finding the marginal distribution of $T_{n,k}$ is not easy. Fortunately, there is a very nice solution to this problem, which we detail below.

7.3.1 Inter-coalescence times in terms of first-coalescence times

Polanski and Kimmel (2003) came up with a beautiful solution to computing the expected SFS under variable population size. Building on their earlier work (Polanski et al, 2003), they found an efficient, numerically stable way of computing $\mathbb{E}[\tau_{n,b}]$ without having to compute $\mathbb{E}[T_{n,k}]$ explicitly. Here, we provide an alternate proof of their result. First we establish a useful lemma by using exchangeability and the consistency property of the coalescent (cf., Chapter 2.11), in a similar vein as Kamm et al (2017).

Lemma 7.7. *For $k+1 \leq n$,*

$$\mathbb{E}[T_{n-1,k}] = \frac{k(k+1)}{n(n-1)}\mathbb{E}[T_{n,k+1}] + \left[1 - \frac{k(k-1)}{n(n-1)}\right]\mathbb{E}[T_{n,k}]. \tag{7.9}$$

Proof. Let $\{\xi_{n,k}, k = n, \ldots, 2\}$ denote the jump chain embedded in the n-coalescent. Then,

$$T_{n-1,k} = \mathbb{I}(\{n\} \in \xi_{n,k+1})T_{n,k+1} + \mathbb{I}(\{n\} \notin \xi_{n,k})T_{n,k}, \tag{7.10}$$

which can be seen as follows. In a coalescent tree with n leaves labeled by $1, \ldots, n$, consider the subtree corresponding to the subsample $1, 2, \ldots, n-1$. As illustrated in Figure 7.1, let $T_{n-1,n-1}, \ldots, T_{n-1,2}$ denote the inter-coalescence times of that subtree, while $T_{n,n}, \ldots, T_{n,2}$ denotes the inter-coalescence times of the full tree. Suppose that when the nth lineage in the full tree coalesces with another lineage, there are j lineages in the subtree. Then, for $k > j$, we have $T_{n-1,k} = T_{n,k+1}$, while for $k < j$, we have $T_{n-1,k} = T_{n,k}$. Furthermore, $T_{n-1,j} = T_{n,j+1} + T_{n,j}$. This establishes (7.10).

Now, noting that $\mathbb{I}(\{n\} \in \xi_{n,k+1})$ and $T_{n,k+1}$ are independent random variables, and likewise for $\mathbb{I}(\{n\} \notin \xi_{n,k})$ and $T_{n,k}$, we obtain

$$\mathbb{E}[T_{n-1,k}] = \mathbb{P}(\{n\} \in \xi_{n,k+1})\mathbb{E}[T_{n,k+1}] + \mathbb{P}(\{n\} \notin \xi_{n,k})\mathbb{E}[T_{n,k}].$$

To complete the proof, note that

$$\mathbb{P}(\{n\} \in \xi_{n,i}) = \prod_{l=i+1}^{n}\left[1 - \frac{l-1}{\binom{l}{2}}\right] = \frac{i(i-1)}{n(n-1)},$$

which follows from the fact that the probability that a particular lineage is involved in a coalescence event while there are l lineages is $\frac{l-1}{\binom{l}{2}}$. □

Using (7.9), we can rewrite $\mathbb{E}[T_{n,k}]$ as

$$\mathbb{E}[T_{n,k}] = \left[1 - \frac{k(k-1)}{n(n-1)}\right]^{-1}\left(\mathbb{E}[T_{n-1,k}] - \frac{k(k+1)}{n(n-1)}\mathbb{E}[T_{n,k+1}]\right).$$

By iteratively applying this recursion, we can represent $\mathbb{E}[T_{n,k}]$ as a linear combination of $\mathbb{E}[T_{m,m}]$:

$$\mathbb{E}[T_{n,k}] = \sum_{m=2}^{n} A_{n,k,m}\mathbb{E}[T_{m,m}], \tag{7.11}$$

where $A_{n,k,m}$ are combinatorial factors that do not depend on η. (Note that $A_{n,k,m} = 0$ for $m < k$.) This is a useful result, since $\mathbb{E}[T_{m,m}]$ can be computed easily:

$$\mathbb{E}[T_{m,m}] = \int_0^\infty \mathbb{P}(T_{m,m} > t)dt = \int_0^\infty e^{-\binom{m}{2}R_\eta(t)}dt. \tag{7.12}$$

Unfortunately, the coefficients $A_{n,k,m}$ are extremely large and have alternating signs, so evaluating (7.11) is numerically unstable due to catastrophic cancellation. However, plugging (7.11) into (5.9), we obtain the following result:

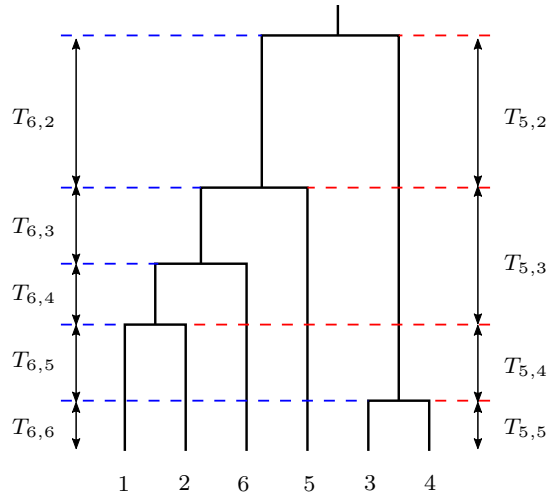


Fig. 7.1: Inter-coalescence times for the full sample $\{1,\dots,n\}$ (shown on the left) and that for the subsample $\{1,\dots,n-1\}$ (shown on the right), where $n=6$. Note that $T_{5,5} = T_{6,6}, T_{5,4} = T_{6,5}$ and $T_{5,2} = T_{6,2}$, while $T_{5,3} = T_{6,4} + T_{6,3}$.

Theorem 7.8 (Polanski and Kimmel 2003). *For all $n > 1$, there exist universal constants $W_{n,b,m}$ independent of η, where $b = 1,\dots,n-1$ and $m = 2,\dots,n$, such that*

$$\mathbb{E}[\tau_{n,b}] = \sum_{m=2}^{n} W_{n,b,m}\mathbb{E}[T_{m,m}]. \tag{7.13}$$

Proof. Using (7.11) in (5.9), we obtain

$$\mathbb{E}(\tau_{n,b}) = \sum_{k=2}^{n} p_{n,k}(b)\cdot k \sum_{m=2}^{n} A_{n,k,m}\mathbb{E}[T_{m,m}] = \sum_{m=2}^{n}\left(\sum_{k=2}^{n} k\, p_{n,k}(b)A_{n,k,m}\right)\mathbb{E}[T_{m,m}],$$

where

$$p_{n,k}(b) = \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}}. \tag{7.14}$$

So, setting $W_{n,b,m} = \sum_{k=2}^{n} k\, p_{n,k}(b)A_{n,k,m}$ completes the proof. □

Empirically, the constants $W_{n,b,m}$ grow much less quickly than do $A_{n,k,m}$, so (7.13) is much more numerically stable than first computing $\mathbb{E}[T_{n,k}]$ using (7.11) and plugging them into (5.9). Furthermore, Polanski and Kimmel (2003) showed that $W_{n,b,m}$ can be computed recursively:

$$W_{n,b,2} = \frac{6}{n+1},$$
$$W_{n,b,3} = 30\frac{(n-2b)}{(n+1)(n+2)},$$
$$W_{n,b,m+2} = -\frac{(1+m)(3+2m)(n-m)}{m(2m-1)(n+m+1)}W_{n,b,m} + \frac{(3+2m)(n-2b)}{m(n+m+1)}W_{n,b,m+1}.$$

7.3.2 Monotonicity and convexity

Sargsyan and Wakeley (2008) proved interesting properties of $\mathbb{E}[\tau_{n,b}]$ that hold for an arbitrary population size function η. Recall the combinatorial coefficient $p_{n,k}(b)$ defined in (7.14). For $k = 2$, we have $p_{n,k}(b) = 1/(n-1)$ for all b. For $k > 2$ and $1 \le b \le n-k$, Sargsyan and Wakeley noted that $p_{n,k}(b)$ satisfies

$$p_{n,k}(b) - p_{n,k}(b+1) = p_{n,k}(b)\frac{k-2}{n-b-1} > 0.$$

Also, since $p_{n,k}(b) = 0$ if $b > n-k+1$, we have $p_{n,k}(b) - p_{n,k}(b+1) \ge 0$ for $b \ge n-k+1$. Hence, for fixed values of n and k, where $2 \le k \le n$, we conclude that $p_{n,k}(b)$ is a non-increasing function of b. Then, since $\mathbb{E}[\tau_{n,b}]$ is a linear combination of $p_{n,2}(b), \ldots, p_{n,n}(b)$ with positive coefficients (namely, $k\mathbb{E}[T_{n,k}]$) and each $p_{n,k}(b)$ is a non-increasing function of b, we conclude that $\mathbb{E}[\tau_{n,b}]$ is a non-increasing function of b.

In addition, Sargsyan and Wakeley showed that $p_{n,k}(b)$ is convex in b, which implies that $\mathbb{E}[\tau_{n,b}]$ is also convex in b.

7.4 SFS-based likelihoods

Consider a locus with m sites and let $\theta/2$ denote the population-scaled mutation rate for the whole locus. If the per-site per-generation mutation rate is μ and N_{ref} denotes a reference population size, then

$$\theta = 4N_{\text{ref}}m\mu.$$

In a sample of size n, suppose we observe $\boldsymbol{s} = (s_1, \ldots, s_{n-1})$ as the frequency spectrum at the locus, where s_b denotes the number of sites each with b derived alleles. What is the probability of this event under the infinite-sites model of mutation and a given demographic model Φ? Below we closely follow the exposition of Bhaskar et al (2015) to present two extreme cases, namely, completely linked and completely unlinked loci.

7.4.1 Completely linked case

If the locus under consideration is completely linked, then the n haplotypes in the sample are related by the same coalescent tree T at all sites of the locus, and we have

$$\mathbb{P}(\boldsymbol{s} \mid T, \Phi, \theta) = \prod_{b=1}^{n-1} \exp\left[-\frac{\theta}{2}\tau_{n,b}(T)\right] \frac{\left[\frac{\theta}{2}\tau_{n,b}(T)\right]^{s_b}}{s_b!}, \tag{7.15}$$

where $\tau_{n,b}(T)$ is the sum of the lengths of all branches in T that subtend b descendant leaves. To compute the probability of observing $\boldsymbol{s}$, we need to integrate (7.15) over the distribution $f(T \mid \Phi)$ of T under the demography Φ. Let $\mathcal{T}_n$ denote the space of coalescent trees with n leaves (capturing both tree topologies and branch lengths). Then, abusing notation, the probability $\mathbb{P}(\boldsymbol{s} \mid \Phi, \theta)$ can be written as

$$\begin{aligned}
\mathbb{P}(\boldsymbol{s} \mid \Phi, \theta) &= \int_{\mathcal{T}_n} \mathbb{P}(\boldsymbol{s} \mid T, \Phi, \theta) f(T \mid \Phi)\, dT \\
&= \int_{\mathcal{T}_n} \left\{ \prod_{b=1}^{n-1} \frac{\left[\frac{\theta}{2}\tau_{n,b}(T)\right]^{s_b}}{s_b!} \right\} \exp\left[-\frac{\theta}{2}\tau_n(T)\right] f(T \mid \Phi)\, dT \\
&= \int_{\mathcal{T}_n} \binom{s}{s_1, \ldots, s_{n-1}} \left\{ \prod_{b=1}^{n-1} \left[\frac{\tau_{n,b}(T)}{\tau_n(T)}\right]^{s_b} \right\} \exp\left[-\frac{\theta}{2}\tau_n(T)\right] \frac{\left[\frac{\theta}{2}\tau_n(T)\right]^{s}}{s!} f(T \mid \Phi)\, dT,
\end{aligned} \tag{7.16}$$

where $s = \sum_{b=1}^{n-1} s_b$ and $\tau_n(T) = \sum_{b=1}^{n-1} \tau_{n,b}(T)$, the total branch length of T. Lohse et al (2011, 2016) recently developed an interesting technique based on generating functions for computing likelihoods of this kind, but unfortunately the approach does not scale well with sample size. When the sample size is moderate to large, it is unknown how to efficiently and exactly compute (7.16), even for a constant population size demographic model. So, recent works (Keinan and Clark, 2012; Nelson et al, 2012) approximated the integral in (7.16) by sampling coalescent trees under the demographic model Φ, and tried to find the MLE for θ by repeating this Monte-Carlo integration for each value of θ in some grid.

7.4.2 Completely unlinked case: Poisson Random Field

The opposite extreme is the Poisson Random Field (PRF) approximation of Sawyer and Hartl (1992), which assumes that all the sites in a given locus are completely unlinked; i.e., the underlying coalescent tree at each site is independent of the trees at other sites. Under this assumption, the probability of observing the frequency spectrum $\boldsymbol{s}$ is given by

$$\begin{aligned}
\mathbb{P}(\boldsymbol{s} \mid \Phi, \theta) &= \left[\prod_{b=1}^{n-1} \frac{\left(\frac{\theta}{2}\mathbb{E}_{\Phi}[\tau_{n,b}]\right)^{s_b}}{s_b!} \right] \exp\left(-\frac{\theta}{2}\mathbb{E}_{\Phi}[\tau_n]\right) \\
&= C \left[\prod_{b=1}^{n-1} \left(\frac{\theta}{2}\mathbb{E}_{\Phi}[\tau_{n,b}]\right)^{s_b} \right] \exp\left(-\frac{\theta}{2}\mathbb{E}_{\Phi}[\tau_n]\right),
\end{aligned} \tag{7.17}$$

where $C = \prod_{b=1}^{n-1} \frac{1}{s_b!}$ is a data-dependent constant that can be ignored for maximum likelihood estimation, and the expectations $\mathbb{E}_{\Phi}[\cdot]$ are taken over the distribution on coalescent trees with n leaves under the demographic model Φ. Therefore, under the PRF approximation, the problem of computing the likelihood in (7.17) reduces to that of computing the expectations $\mathbb{E}_{\Phi}[\tau_{n,b}]$ and $\mathbb{E}_{\Phi}[\tau_n]$. Using the results discussed in Chapter 7.3, this can be done *numerically stably* and *exactly* for a wide class of population size functions. Taking logarithms on both sides of (7.17), we obtain the following log-likelihood for the demographic model Φ and mutation rate θ:

$$\mathcal{L}(\Phi, \theta) = \log \mathbb{P}(\boldsymbol{s} \mid \Phi, \theta) = \sum_{b=1}^{n-1} s_b \left(\log \mathbb{E}_{\Phi}[\tau_{n,b}] + \log \theta\right) - \frac{\theta}{2} \mathbb{E}_{\Phi}[\tau_n] + \text{constant}(\boldsymbol{s}), \tag{7.18}$$

where $\text{constant}(\boldsymbol{s})$ depends on $\boldsymbol{s}$ but not on Φ or θ.

Now suppose there are L independent loci with observed frequency spectrum $\boldsymbol{s}^{(l)} = (s_1^{(l)}, \ldots, s_{n-1}^{(l)})$ and mutation rate $\theta^{(l)}/2$ for the lth locus. Assuming that each locus is completely unlinked, the one-locus log-likelihood in (7.18) can be summed across all loci $l = 1, \ldots, L$, yielding

$$\begin{aligned}\mathcal{L}(\Phi, \{\theta^{(l)}\}_{l=1}^L) &= \log \mathbb{P}(\{\boldsymbol{s}^{(l)}\}_{l=1}^L \mid \Phi, \{\theta^{(l)}\}_{l=1}^L) \\ &= \sum_{l=1}^{L} \left[\sum_{b=1}^{n-1} s_b^{(l)} \left(\log \mathbb{E}_{\Phi}[\tau_{n,b}] + \log \theta^{(l)}\right) - \frac{\theta^{(l)}}{2} \mathbb{E}_{\Phi}[\tau_n]\right] + \text{constant}(\{\boldsymbol{s}^{(l)}\}_{l=1}^L).\end{aligned} \tag{7.19}$$

It is easy to see that $\mathcal{L}$ is a concave function of $\theta^{(l)}$, since the Hessian of $\mathcal{L}$ with respect to $\boldsymbol{\theta} = (\theta^{(1)}, \ldots, \theta^{(L)})$ is negative definite for all $\boldsymbol{\theta} \succ \mathbf{0}$, as can be seen from

$$\frac{\partial^2 \mathcal{L}}{\partial \theta^{(l)} \partial \theta^{(l')}} = -\delta_{l,l'} \frac{1}{[\theta^{(l)}]^2} \sum_{b=1}^{n-1} s_b^{(l)}. \tag{7.20}$$

Hence, the mutation rates that maximize $\mathcal{L}$ are the solutions of

$$0 = \frac{\partial \mathcal{L}}{\partial \theta^{(l)}} = \frac{1}{\theta^{(l)}} \sum_{b=1}^{n-1} s_b^{(l)} - \frac{1}{2} \mathbb{E}_{\Phi}[\tau_n], \tag{7.21}$$

yielding the following MLE for $\theta^{(l)}$ given the demographic model Φ:

$$\widehat{\theta}^{(l)} = \frac{2 \sum_{b=1}^{n-1} s_b^{(l)}}{\mathbb{E}_{\Phi}[\tau_n]}, \tag{7.22}$$

which generalizes Watterson's (1975) estimator to the case of variable population size. Substituting this MLE for $\theta^{(l)}$ into (7.19), we obtain the following log-likelihood for Φ with the optimal mutation rates:

$$\mathcal{L}(\Phi) = \sum_{b=1}^{n-1} \left[\left(\sum_{l=1}^{L} s_b^{(l)}\right) \log \left(\frac{\mathbb{E}_{\Phi}[\tau_{n,b}]}{\mathbb{E}_{\Phi}[\tau_n]}\right)\right] + \text{constant}(\{\boldsymbol{s}^{(l)}\}_{l=1}^L). \tag{7.23}$$

Now, define the discrete probability distribution $\boldsymbol{p}_n = \{p_{n,b}\}_{b=1}^{n-1}$, where

$$p_{n,b} = \frac{\sum_{l=1}^{L} s_b^{(l)}}{\sum_{k=1}^{n-1}\sum_{l=1}^{L} s_k^{(l)}},$$

and let $\boldsymbol{q}_n(\Phi) = \{q_{n,b}(\Phi)\}_{b=1}^{n-1}$, where

$$q_{n,b}(\Phi) = \frac{\mathbb{E}_\Phi[\tau_{n,b}]}{\mathbb{E}_\Phi[\tau_n]}.$$

Then, the MLE of the likelihood function $\mathcal{L}(\Phi)$ in (7.23) is given by (Bhaskar et al, 2015)

$$\widehat{\Phi} = \arg\max_{\Phi} \mathcal{L}(\Phi) = \arg\min_{\Phi} \mathrm{KL}(\boldsymbol{p}_n \| \boldsymbol{q}_n(\Phi)), \tag{7.24}$$

where $\mathrm{KL}(P\|Q)$ denotes the Kullback-Liebler divergence of distribution Q from P.

In summary, for a given demographic model Φ, one can compute the log-likelihood using (7.23), and infer the optimal mutation rate at each locus using (7.22). The gradient of $\mathcal{L}(\Phi)$ with respect to Φ can be computed using automatic differentiation (Griewank and Corliss, 1991), which allows one to search over the space of demographic models more efficiently using standard gradient-based optimization algorithms.

7.5 A recursion for efficiently computing $\mathbb{P}(A_m(t) = k)$

For a fixed positive integer n, Kamm et al (2017) showed that $\mathbb{P}(A_m(t) = k)$ for all values of k, m satisfying $1 \le k \le m \le n$ can be computed efficiently in $O(n^2)$ time. The idea is to use exchangeability and the consistency property of the coalescent, as in the proof of Lemma 7.7. Let $\{C_m(t), t \ge 0\}$ denote the m-coalescent. Then, note that

$$\begin{aligned}
\mathbb{P}(A_{m-1}(t) = k) &= \mathbb{P}(A_m(t) = k+1, \{m\} \in C_m(t)) + \mathbb{P}(A_m(t) = k, \{m\} \notin C_m(t)) \\
&= \mathbb{P}(\{m\} \in \xi_{m,k+1})\mathbb{P}(A_m(t) = k+1) + \mathbb{P}(\{m\} \notin \xi_{m,k})\mathbb{P}(A_m(t) = k) \\
&= \frac{k(k+1)}{m(m-1)}\mathbb{P}(A_m(t) = k+1) + \left[1 - \frac{k(k-1)}{m(m-1)}\right]\mathbb{P}(A_m(t) = k).
\end{aligned}$$

Upon rearranging the above equation, we obtain the recursion

$$\mathbb{P}(A_m(t) = k) = \left[1 - \frac{k(k-1)}{m(m-1)}\right]^{-1}\left[\mathbb{P}(A_{m-1}(t) = k) - \frac{k(k+1)}{m(m-1)}\mathbb{P}(A_m(t) = k+1)\right]$$

with boundary conditions

$$\mathbb{P}(A_m(t) = m) = e^{-\binom{m}{2}R_\eta(t)}.$$

Hence, after computing $R_\eta(t)$, we can use the above recursion and memoization to solve for all of the $O(n^2)$ terms $\mathbb{P}(A_m(t) = k)$, where $1 \le k \le m \le n$, in $O(n^2)$ time.

7.6 Identifiability of population size histories from the SFS

Numerous empirical studies in population genetics have been based on the SFS, which describes the distribution of the number of mutant alleles at a polymorphic site in a sample of DNA sequences. This widely-used summary statistic provides a highly efficient dimensional reduction of large-scale population genomic variation data, summarizing the information in n sequences of arbitrary length in just $n-1$ numbers.

Suppose the mutation rate is low so that the infinite-sites model is applicable. Then, given a sample of n sequences, its empirical SFS is $\hat{\boldsymbol{f}}_n = (\hat{f}_{n,1}, \ldots, \hat{f}_{n,n-1})$, where $\hat{f}_{n,b}$, for $b \in \{1, \ldots, n-1\}$, is defined as

$$\hat{f}_{n,b} = \frac{\#\text{ sites with } b \text{ mutant alleles and } n-b \text{ ancestral alleles}}{\#\text{ segregating sites}}.$$

Example 7.9. Suppose 1 denotes the mutant type and 0 the ancestral type, and consider the following four sequences:

Sequence 1 =	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0
Sequence 2 =	1	0	0	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
Sequence 3 =	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Sequence 4 =	1	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
Count of 1s:	1		3			1						1			1				2			1		

There are six segregating sites and the count of 1s at those sites are shown above. The empirical SFS corresponding to the above toy data are

$$\hat{f}_{4,1} = \frac{4}{6}, \qquad \hat{f}_{4,2} = \frac{1}{6}, \qquad \hat{f}_{4,3} = \frac{1}{6},$$

which sum to 1.

Under a given model Θ of population size history, the *expected* SFS for a sample of n sequences is defined as $\boldsymbol{f}_n^{(\Theta)} = (f_{n,1}^{(\Theta)}, \ldots, f_{n,n-1}^{(\Theta)})$, where $f_{n,b}^{(\Theta)}$ is the conditional probability that a site has b mutant alleles and $n-b$ ancestral alleles given that the site is segregating. As discussed in Chapter 7.3, the mathematical dependence of the expected SFS $\boldsymbol{f}_n^{(\Theta)}$ on the underlying population demography Θ is well understood, and we can compute $\boldsymbol{f}_n^{(\Theta)}$ efficiently.

By the law of large numbers, as the number of segregating sites tends to infinity, the empirical SFS converges almost surely to the expected SFS corresponding to the true demographic model. The question of identifiability is whether the expected SFS uniquely determines the population size history. In other words, how much information about the underlying demography is captured by the expected SFS? Which population size functions are identifiable and what is the sample size n needed to guarantee unique recovery of the true population size function from the expected SFS?

7.6.1 An analogy: Can you hear the shape of a drum?

Before we proceed with the SFS question, it would help to draw an analogy at this point. A famous question in analysis (Kac, 1966) is whether one can uniquely determine the shape

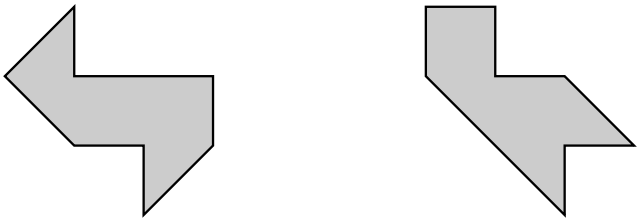


Fig. 7.2: A counterexample to the "can you hear the shape of a drum" problem constructed by Gordon et al (1992). The two drumheads have different shapes, but they have exactly the same set of Dirichlet eigenvalues.

of a two-dimensional drumhead from the sound it makes, more precisely by knowing the complete list of overtones that the drumhead is capable of producing. Mathematically this question can be translated to whether two plane regions with different boundaries can have exactly the same set of eigenvalues of the Laplacian, with the Dirichlet boundary condition that the eigenfunction vanishes at the boundary.

Gordon et al (1992) proved that the answer to this question is negative if arbitrarily shaped drumheads are allowed. They actually constructed an explicit counterexample, illustrated in Figure 7.2. Note that these shapes have sharp corners and do not exhibit symmetries. Interestingly, Zelditch (2000) showed that if drumhead shapes are constrained to have analytical boundaries and certain reflection symmetries, then the answer to the question is positive.

7.6.2 Non-identifiability and an explicit counterexample

Suppose we have perfect information about the expected SFS $\boldsymbol{f}_n$ for n genomes randomly sampled from a population. Can the data uniquely determine $\eta(t)$? If $\eta(t)$ is allowed to be arbitrary, the answer to this question is negative. Myers et al (2008) proved that different population size functions can generate the same expected SFS $\boldsymbol{f}_n = (f_{n,1}, \ldots, f_{n,n-1})$ for all sample sizes n. Using Müntz-Szasz theory, they showed that there exist smooth functions F such that for every population size function η, both η and $\eta + F$ generate the same expected SFS for all n. Furthermore, they constructed an explicit example of such a function $F(t)$:

$$F(t) = \int_0^t g_0(t-u) g_1(u) du,$$

where $g_1(t) = \frac{\cos(\pi^2/t)\exp(-t/8)}{\sqrt{t}}$ and $g_0(t) = \exp(-1/t^2)$. Because of the $\cos(\pi^2/t)$ term, this $F(t)$ oscillates at an increasingly higher frequency as time approaches the present.

The population size functions involved in the counterexample are arguably unrealistic for biological populations, since birth and death rates in a real population are bounded. Then, one may ask,

1. What kind of population size models are identifiable?
2. What sample sizes are required for identifiability?

Bhaskar and Song (2014) provided answers to these questions. As in the drum analogy discussed in the previous section, although the population size function η is in general not identifiable from the SFS if η is allowed to be arbitrary, we can achieve identifiability if we impose suitable constraints on η.

Definition 7.10 ($\mathcal{F}$, family of piecewise continuous population size functions). A family $\mathcal{F}$ of piecewise continuous population size functions is a set of positive piecewise continuous functions $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_+$ of a particular type parameterized by a collection of variables.

7.6.3 Rule of signs

Definition 7.11 (A zero of multiplicity m). Let g be a smooth function on $[a, b] \subset \mathbb{R}$. We say that g has a zero (or root) of multiplicity $m \in \mathbb{N}$ at the point $x_0 \in [a, b]$ if

$$g(x_0) = g'(x_0) = \cdots = g^{(m-1)}(x_0) = 0 \quad \text{and} \quad g^{(m)}(x_0) \neq 0.$$

Let $Z(g, I)$ denote the number of zeros of g in interval I, counted with their multiplicity.

Theorem 7.12 (Rolle's Theorem). *Let $g : \mathbb{R} \to \mathbb{R}$ be a differentiable function on $[a, b]$ that satisfies $g(a) = g(b)$. Then, there exists at least one point x_0 in the open interval (a, b) such that $g'(x_0) = 0$.*

Lemma 7.13. *For all smooth function g and interval I, $Z(g, I) \leq Z(g', I) + 1$.*

Proof. Suppose g has n zeros $x_1, \ldots, x_n$ in I with multiplicity $m_1, \ldots, m_n$, respectively. Then, for all $r = 1, \ldots n$, the point x_r is a zero of g' with multiplicity $m_r - 1$. Furthermore, by Theorem 7.12, g' has at least one zero between each consecutive points of $x_1, \ldots, x_n$. So,

$$Z(g', I) \geq n - 1 + \sum_{r=1}^{n} (m_r - 1) = n - 1 + Z(g, I) - n = Z(g, I) - 1.$$

□

Suppose $a_n, \ldots, a_0 \in \mathbb{R}$ and $p_n, \ldots, p_1 \in \mathbb{R}$ where $p_n > p_{n-1} > \cdots > p_1$.

- Ordinary polynomials:

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0, \text{ where } x \in \mathbb{R}.$$

- Generalized polynomials:

$$Q(t) = a_n t^{p_n} + a_{n-1} t^{p_{n-1}} + \cdots + a_1 t^{p_1}, \text{ where } t \in \mathbb{R}_+.$$

- Generalized Dirichlet polynomials:

$$D(x) = a_n e^{p_n x} + a_{n-1} e^{p_{n-1} x} + \cdots + a_1 e^{p_1 x}, \text{ where } x \in \mathbb{R}.$$

Theorem 7.14 (Descartes' rule of signs). *Counted with multiplicity, let $Z_+(P)$ denote the number of positive zeros of P and let $Z(Q), Z(D)$ denote the number of zeros of Q, D over their respective domains. Let $S[(a_j)]$ denote the number of sign changes in the sequence $a_n, a_{n-1}, \ldots, a_0$. Then,*

$$Z_+(P) \leq S[(a_j)], \qquad Z(Q) \leq S[(a_j)], \qquad Z(D) \leq S[(a_j)]$$

Proof. We prove by induction on the number of sign changes. Let $D(x) = \sum_{j=1}^{n} a_j e^{p_j x}$, where $p_1 < p_2 < \cdots < p_n$. Suppose $S[(a_j)] = 0$. Then, either $D(x) > 0$ for all x or $D(x) < 0$ for all $x \in \mathbb{R}$, so $Z(D) = 0$. Assume $Z(D) \leq S[(a_j)]$ holds when there are $\leq s$ sign changes. Suppose $S[(a_j)] = s + 1$ and suppose that a sign change in $a_1, \ldots, a_n$ occurs at index k. Choose p such that $p_{k-1} < p < p_k$ and define

$$F(x) = e^{-px} D(x) = \sum_{j=1}^{n} a_j e^{(p_j - p)x}.$$

Note that D and F have the same zeros. Further

$$F'(x) = \sum_{j=1}^{n} (p_j - p) a_j e^{(p_j - p)x}.$$

Define $b_j = (p_j - p)a_j$. Then, the sequence $b_1, \ldots, b_n$ does not change sign at index k, so $S[(b_j)] = s$ and the induction hypothesis implies $Z(F') \leq s$. Finally, $Z(D) = Z(F) \leq Z(F') + 1 \leq s + 1$. □

Example 7.15. How many positive roots does the following polynomial have?

$$P(x) = x^{100} - 7x^{77} + 13x^{14} + 19x^3 + 8.$$

Since the sequence of coefficients has two sign changes, Descartes' rule of signs implies $Z_+(P) \leq 2$. In fact, $P(x)$ has exactly two positive roots. How many negative roots does it have? Since

$$P(-x) = x^{100} + 7x^{77} + 13x^{14} - 19x^3 + 8$$

has two sign changes, applying Descartes' rule of signs implies to this polynomial implies that P has at most 2 negative roots. In fact, it turns out that P has exactly two negative roots.

We use $\sigma(g)$ to denote the number of sign changes of a function g; it counts the number of times the function g changes value from positive to negative (and vice versa) while ignoring intervals where it is identically zero. See Bhaskar and Song (2014) for a more precise definition.

Theorem 7.16 (Generalized Descartes' rule of signs). *Let $g : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be a piecewise-continuous function which is not identically zero and with a finite number $\sigma(g)$ of sign changes. Then, the function $G(x)$ defined by*

$$G(x) = \int_0^\infty g(t) e^{-tx} \, dt$$

has at most $\sigma(g)$ roots in $\mathbb{R}$ (counted with multiplicity).

Proof. By induction on the number of sign changes of g. If g has zero sign changes, then without loss of generality, $g(t) \geq 0$ for $t \in (0, \infty)$ and $g(t) > 0$ for some interval $(a, b) \subseteq (0, \infty)$. Hence, $G(x) > 0$ for all x, and the base case holds. Assume the statement holds up to m and suppose g has $m + 1$ sign change points $t_0, \ldots, t_m$, where $m \geq 0$.

Note that $G(x)$ and $F(x) = e^{t_0 x} G(x)$ have the same real-valued roots (with multiplicity) since $e^{t_0 x} > 0$ for all $x \in \mathbb{R}$. Further, $F'(x)$ is given by

$$F'(x) = \frac{d}{dx}\left(\int_0^\infty g(t) e^{-(t-t_0)x}\, dt\right) = \int_0^\infty (t_0 - t) g(t) e^{-(t-t_0)x}\, dt,$$

where the interchange of the differential and integral operators in the second equality is justified by the Leibniz integral rule.

Note that the set of sign change points of $(t_0 - t)g(t)$ is $\{t_1, \ldots, t_m\}$. Hence $(t_0 - t)g(t)$ has only m sign changes.

By the induction hypothesis, F' has at most m real-valued roots. By Theorem 7.12, the number of real-valued roots of F is at most one more than the number of real-valued roots of F'. Hence, F has at most $m + 1$ real-valued roots, implying that G has at most $m + 1$ real-valued roots. □

7.6.4 Identifiability

Recall the time-rescaling function R_η defined in (7.2). Using this function, we define the time-rescaled population size function $\widetilde{\eta}$ as

$$\widetilde{\eta}(\tau) = \eta(R_\eta^{-1}(\tau)).$$

Then, $\mathbb{E}[T_{m,m}]$ shown in (7.12), the expected time to the first coalescence for a sample of size m, can be written as

$$\mathbb{E}[T_{m,m}] = \int_0^\infty \widetilde{\eta}(\tau)\, e^{-\binom{m}{2}\tau} d\tau, \tag{7.25}$$

which is the Laplace transform of $\widetilde{\eta}$ evaluated at $-\binom{m}{2}$. By Theorem 7.8, we know that $\mathbb{E}[\tau_{n,b}]$ is determined by the collection $\mathbb{E}[T_{2,2}], \ldots, \mathbb{E}[T_{n,n}]$ of expected first coalescence times for sample sizes 2 to n. Furthermore, one can prove that the $(n-1)$-by-$(n-1)$ matrix $\boldsymbol{W}_n = (W_{n,b,m})$ appearing in Theorem 7.8 is invertible. Therefore, there is a one-to-one correspondence between $\{\mathbb{E}[\tau_{n,b}]\}_{b=1,\ldots,n-1}$ and $\{\mathbb{E}[T_{m,m}]\}_{m=2,\ldots,n}$. Hence, the question of identifiability is equivalent to asking under what conditions the collection $\{\mathbb{E}[T_{m,m}]\}_{m=2,\ldots,n}$ uniquely determines the population size function. Recently Bhaskar and Song (2014) found sufficient conditions for identifiability and their results are given in terms of the following quantity:

Definition 7.17 (Sign change complexity). We define the sign change complexity $\mathscr{S}(\mathcal{F})$ of a function family $\mathcal{F}$ as

$$\mathscr{S}(\mathcal{F}) = \sup_{\eta_1, \eta_2 \in \mathcal{F}} \{\sigma(\widetilde{\eta}_1 - \widetilde{\eta}_2)\}.$$

Lemma 7.18 (Bhaskar and Song 2014). *If $\mathscr{S}(\mathcal{F}) < \infty$ and $n \geq \mathscr{S}(\mathcal{F}) + 2$, then the collection $\{\mathbb{E}[T_{m,m}]\}_{m=2,\ldots,n}$ of expected first-coalescence times uniquely determines the population size function in $\mathcal{F}$. In other words, for $n \geq \mathscr{S}(\mathcal{F})+2$, the map $(\mathbb{E}[T_{2,2}], \ldots, \mathbb{E}[T_{n,n}]) : \mathcal{F} \to \mathbb{R}_+^{n-1}$ is injective.*

Proof. Suppose there exist two distinct models $\eta_1, \eta_2 \in \mathcal{F}$ that produce exactly the same $\mathbb{E}[T_{m,m}]$ for all $2 \leq m \leq n$. Then, by (7.25),

$$\int_0^\infty [\widetilde{\eta}_1(\tau) - \widetilde{\eta}_2(\tau)] e^{-\binom{m}{2}\tau} d\tau = 0$$

for all $2 \leq m \leq n$. Define the function $G(x)$ as

$$G(x) = \int_0^\infty [\widetilde{\eta}_1(\tau) - \widetilde{\eta}_2(\tau)] e^{x\tau} d\tau.$$

Since $-\binom{m}{2}$ is a root of $G(x)$ for $2 \leq m \leq n$, we conclude $Z(G) \geq n - 1$. However, by generalized Descartes' rule of signs (cf., Theorem 7.16),

$$Z(G) \leq \sigma(\widetilde{\eta}_1 - \widetilde{\eta}_2) \leq \mathscr{S}(\mathcal{F}),$$

where the second inequality the definition of $\mathscr{S}(\mathcal{F})$ (cf., Definition 7.17). Hence, we get a contradiction if $n - 1 > \mathscr{S}(\mathcal{F})$, which is equivalent to the condition $n \geq \mathscr{S}(\mathcal{F}) + 2$. □

Putting together the above discussion and recalling that the unnormalized SFS is given by

$$\mathbb{E}[\zeta_{n,b}] = \frac{\theta}{2} \mathbb{E}[\tau_{n,b}],$$

we obtain the following result:

Theorem 7.19 (Sufficient conditions for identifiability). *If $\mathscr{S}(\mathcal{F}) < \infty$ and $n \geq \mathscr{S}(\mathcal{F}) + 2$, then the expected unnormalized SFS $(\mathbb{E}[\zeta_{n,1}], \ldots, \mathbb{E}[\zeta_{n,n-1}])$ uniquely determines the population size function in $\mathcal{F}$.*

Example 7.20. The following conditions are sufficient for guaranteeing identifiability of population size functions from the expected unnormalized SFS:

1. $n \geq 2K$ for $\mathcal{F} =$ the set of piecewise-constant functions with K pieces.
2. $n \geq 4K - 1$ for $\mathcal{F} =$ the set of piecewise-exponential functions with K pieces.
3. $n \geq 6K - 2$ for $\mathcal{F} =$ the set of piecewise-generalized-exponential functions with K pieces.

Consider a piecewise-defined population size function $\eta \in \mathcal{F}$ that produces $\mathbb{E}[T_{m,m}] = c_m$ for $2 \leq m \leq n$. Suppose $\mathscr{S}(\mathcal{F}) < \infty$ and $n \geq \mathscr{S}(\mathcal{F}) + 2$. Then, for every fixed $s \in \mathbb{R}_+$, there exists a unique population size function $\eta_s \in \mathcal{F}$ with $\mathbb{E}[T_{m,m}] = s \cdot c_m$ for $2 \leq m \leq n$. Furthermore, this population size function η_s is given by $\eta_s(t) = s \cdot \eta(t/s)$.

Definition 7.21 (Equivalent piecewise population size functions). Given two models $\eta_1, \eta_2 \in \mathcal{F}$, we say that η_1 and η_2 are equivalent, and write $\eta_1 \sim \eta_2$, if they are related by rescaling time and population sizes as described above.

Since $f_{n,b} = \mathbb{E}[\zeta_{n,b}] / \sum_{k=1}^{n-1} \mathbb{E}[\zeta_{n,k}]$, we note that η and η_s produce the same $f_{n,b}$. So, there is a one-parameter family of population size functions in $\mathcal{F}$ that produce exactly the same expected normalized SFS, as formalized by the following theorem (Bhaskar and Song, 2014):

Theorem 7.22 (Identifiability from the expected SFS). *If $\mathscr{S}(\mathcal{F}) < \infty$ and $n \geq \mathscr{S}(\mathcal{F}) + 2$, then the expected normalized SFS $\boldsymbol{f}_n = (f_{n,1}, \ldots, f_{n,n-1})$ uniquely determines a unique equivalence class $[\eta]$ of models in $\mathcal{F}/\!\sim$.*

See Bhaskar and Song (2014) for results on the identifiability of population size functions from the folded SFS. Since the folded SFS has roughly half the number of entries as the unfolded SFS (cf., Definition 5.4), we expect to require roughly twice as large of a sample to achieve identifiability, and this intuition turns out to be correct. More precisely, one gains a factor of 2 in front of the sign change complexity $\mathscr{S}(\mathcal{F})$ term in the sample size lower bound.

7.7 Minimax error for population size estimation based on the SFS

As detailed in the previous section, the expected SFS uniquely determines the population size function $\eta(t)$ provided that it has a finite number of oscillations and the sample size n is sufficiently large. We discussed a general bound on n sufficient to guarantee identifiability and saw that the bound depends on the type of population size function. When studying identifiability, we assume that an *infinite* amount of data is available; that is, we assume that the number s of segregating sites is unlimited so that we can obtain perfect estimates of the expected SFS. In practice, s is finite and only a perturbed version of the expected frequency spectrum, say $\hat{\boldsymbol{f}}_n$, is observed. From a practical standpoint, it is important to understand how these perturbations ultimately affect the parameter estimate $\hat{\eta}(t)$.

This question was addressed by Terhorst and Song (2015), who showed that using the SFS to estimate the size history of a population has a minimax error of at least $O(1/\log s)$, where s is the number of independent segregating sites used in the analysis. This is an information-theoretic result in nature and applies to *any* estimator that operates solely on the SFS. This lower bound of $O(1/\log s)$ on the minimax error is rather surprising in a couple of ways:

1. The minimax error for many classical estimation problems in statistics (for example, nonparametric regression or density estimation) decays inverse polynomially in the amount of data. Compared with these problems, exponentially more data would be required to estimate a population size history function to within a similar magnitude of error.
2. The bound does not depend on n, which means that, for a fixed number s of segregating sites considered, using more individuals does not help to reduce the minimax error bound.

This minimax error result pertains to populations that have experienced a bottleneck, and it can be expected to apply to many populations in nature.

7.8 Geometry of the SFS

Rosen et al (2018)

References

Bhaskar A, Song YS (2014) Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. Annals of Statistics 42(6):2469–2493

Bhaskar A, Wang YXR, Song YS (2015) Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. Genome Research 25(2):268–279

Gordon C, Webb D, Wolpert S (1992) Isospectral plane domains and surfaces via Riemannian orbifolds. Inventiones mathematicae 110(1):1, DOI 10.1007/BF01231320, URL http://dx.doi.org/10.1007/BF01231320

Griewank A, Corliss GF (1991) Automatic differentiation of algorithms: theory, implementation, and application. Society for industrial and Applied Mathematics Philadelphia, PA

Kac M (1966) Can one hear the shape of a drum? The American Mathematical Monthly 73(4):1–23

Kamm JA, Terhorst J, Song YS (2017) Efficient computation of the joint sample frequency spectra for multiple populations. Journal of Computational and Graphical Statistics 26(1):182–194

Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336(6082):740–743

Lohse K, Harrison R, Barton NH (2011) A general method for calculating likelihoods under the coalescent process. Genetics 189(3):977–987

Lohse K, Chmelik M, Martin SH, Barton NH (2016) Efficient strategies for calculating blockwise likelihoods under the coalescent. Genetics 202(2):775–786

Myers S, Fefferman C, Patterson N (2008) Can one learn history from the allelic spectrum? Theoretical Population Biology 73(3):342–348

Nelson MR, Wegmann D, Ehm MG, Kessner D, Jean PS, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D, et al (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337(6090):100–104

Polanski A, Kimmel M (2003) New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics 165(1):427–436

Polanski A, Bobrowski A, Kimmel M (2003) A note on distributions of times to coalescence, under time-dependent population size. Theoretical Population Biology 63(1):33–40

Rosen Z, Bhaskar A, Roch S, Song YS (2018) Geometry of the sample frequency spectrum and the perils of demographic inference. Genetics 210(2):665–682

Sargsyan O, Wakeley J (2008) A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. Theoretical Population Biology 74(1):104–114

Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132(4):1161–76

Terhorst J, Song YS (2015) Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. Proc Nat Acad Sci 112(25):7677–7682

Watterson G (1975) On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7(2):256–276

Zelditch S (2000) Spectral determination of analytic bi-axisymmetric plane domains. Geometric and Functional Analysis 10(3):628–677

Chapter 8
Multiple populations

In this chapter, we consider the *structured coalescent* (see also Wakeley (2008, Chapter 5) for an introduction). This is the first example of a coalescent process in which exchangeability is violated. The structured coalescent is appropriate when there is non-random mating, for example by geographic isolation of individuals into distinct demes. However, a number of other modeling assumptions have the structured coalescent as their limiting behavior.

8.1 The structured coalescent

We assume a finite number d of subpopulations, or *demes*, labeled $1, \ldots, d$, with random mating (and hence exchangeability) within each deme, but no mating between demes. Each deme $\alpha \in \{1, \ldots, d\}$ is large and—for convenience—haploid, and we denote its size by N_α. Per-generation rates of movement between demes are assumed to be low, so that migration events occur on the same timescale as coalescence events. (The so-called *strong migration limit* (Nagylaki, 1980; Notohara, 1993), in which migration events occur on a much faster timescale than coalescence events, leads to another set of interesting dynamics which we do not go into here.) Lineages can be thought of as carrying labels which identify the deme in which they reside and influence their rate of coalescence. The structured coalescent will evolve on a single timescale N, and we define the relative size of each deme by

$$r_\alpha = \frac{N_\alpha}{N}.$$

The assumption that each deme is large corresponds to holding each r_α fixed as $N_\alpha \to \infty$ and $N \to \infty$. Note that there is more than one choice for the timescale N; two natural choices are

1. $N = \sum_{\alpha=1}^{d} N_\alpha$, the total population size, and
2. $N = N_\alpha$, the size of a reference deme α.

Results given in coalescent units are with respect to the chosen timescale; different authors sometimes quote different results due to the choice of timescale. While there exist a_α lineages in deme α, the rate of coalescence of these lineages is

$$\binom{a_\alpha}{2}\frac{1}{r_\alpha}. \tag{8.1}$$

This agrees with our intuition; smaller demes should exhibit *faster* rates of coalescence.

Finally, we allow migration to occur between demes at certain rates. In the discrete model of population reproduction, let $c_{\alpha\beta}$ denote the probability that the parent of an individual from deme α is from deme β one generation back in time. As with other coalescent parameters, we hold $m_{\alpha\beta} = 2Nc_{\alpha\beta}$ fixed while letting $N \to \infty$ and $c_{\alpha\beta} \to 0$. Note that these parameters are defined going *backwards* in time.

Our first task is to find the ancestral process. The ancestral process for the unstructured coalescent was a death process on the natural numbers. The ancestral process for the structured coalescent, $\{\boldsymbol{A}_n(t) : t \geq 0\}$, applies to vectors of the form $\boldsymbol{A}_n(t) = (a_1, \ldots, a_d)$ where a_α denotes the number of lineages in deme α at time t. We must have $a_\alpha \geq 0$ for each α and each $t \geq 0$, and also $\sum_{\alpha=1}^d a_\alpha = n$ when $t = 0$.

Transition rates can be obtained by similar arguments to the unstructured case. Mimicking the argument given in Chapter 1, we write down the leading-order contributions to the one-step transition probability $\mathbb{P}[\boldsymbol{A}_n^{\mathrm{WF}}(\tau+1) = \boldsymbol{b} \mid \boldsymbol{A}_n^{\mathrm{WF}}(\tau) = \boldsymbol{a}]$ in the discrete-time Wright-Fisher model:

$$\begin{aligned} g_{\boldsymbol{ab}} &= \mathbb{P}[\boldsymbol{A}_n^{\mathrm{WF}}(\tau+1) = \boldsymbol{b} \mid \boldsymbol{A}_n^{\mathrm{WF}}(\tau) = \boldsymbol{a}] \\ &= \begin{cases} \frac{1}{N} a_\alpha \frac{m_{\alpha\beta}}{2} + O\left(\frac{1}{N^2}\right), & \text{if } \boldsymbol{b} = \boldsymbol{a} - \boldsymbol{e}_\alpha + \boldsymbol{e}_\beta, \\ \frac{1}{N}\binom{a_\alpha}{2}\frac{1}{r_\alpha} + O\left(\frac{1}{N^2}\right), & \text{if } \boldsymbol{b} = \boldsymbol{a} - \boldsymbol{e}_\alpha, \\ 1 - \frac{1}{N}\sum_{\alpha=1}^d \left[a_\alpha \frac{m_\alpha}{2} + \binom{a_\alpha}{2}\frac{1}{r_\alpha}\right] + O\left(\frac{1}{N^2}\right), & \text{if } \boldsymbol{b} = \boldsymbol{a}, \\ O\left(\frac{1}{N^2}\right), & \text{otherwise,} \end{cases} \end{aligned}$$

where $m_\alpha = \sum_{\beta:\beta\neq\alpha} m_{\alpha\beta}$, and $\boldsymbol{e}_\alpha$ denotes a unit vector with the αth entry equal 1 and the rest zero. As in the Kingman case, we write $G = (g_{\boldsymbol{ab}}) = I + Q\frac{1}{N} + O(\frac{1}{N^2})$, so that, for all $t \geq 0$,

$$G^{\lfloor Nt \rfloor} = \left[I + Q\frac{1}{N} + O\left(\frac{1}{N^2}\right)\right]^{\lfloor Nt \rfloor} \to e^{Qt},$$

as $N \to \infty$. The ancestral process for the structured coalescent converges to a continuous-time Markov process with generator $Q = (q_{\boldsymbol{ab}})$, where

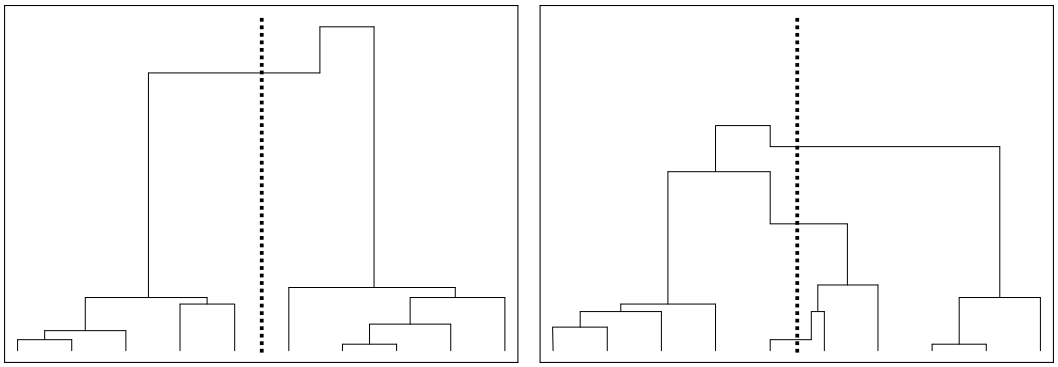


Fig. 8.1: Two realizations of the structured coalescent for $n = 10$ sequences, five from each of two demes (distinguished by the dashed line).

$$q_{\boldsymbol{ab}} = \begin{cases} a_\alpha \dfrac{m_{\alpha\beta}}{2}, & \text{if } \boldsymbol{b} = \boldsymbol{a} - \boldsymbol{e}_\alpha + \boldsymbol{e}_\beta, \\ \dbinom{a_\alpha}{2} \dfrac{1}{r_\alpha}, & \text{if } \boldsymbol{b} = \boldsymbol{a} - \boldsymbol{e}_\alpha, \\ -\displaystyle\sum_{\alpha=1}^{d} \left[a_\alpha \frac{m_\alpha}{2} + \binom{a_\alpha}{2} \frac{1}{r_\alpha} \right], & \text{if } \boldsymbol{b} = \boldsymbol{a}, \\ 0, & \text{otherwise.} \end{cases} \tag{8.2}$$

It is worth emphasizing that coalescence occurs only *within* demes. Unlike coalescence events, the total number of migration events is random and unbounded. At least $d-1$ are required when initially we have $a_\alpha > 0$ for each α. In general, e^{Qt} does not admit simple expressions, so, unlike in the un-structured coalescent, $\mathbb{P}(\boldsymbol{A}_n(t) = \boldsymbol{a})$ is not known in closed form.

Figure 8.1 illustrates two realizations of the structured coalescent process when $d = 2$. Unlike variable population size, exchangeability is violated and so both branch lengths and topologies differ from the usual unstructured (Kingman) coalescent. The left-hand genealogy in Figure 8.1 contains only one migration event, and so might be more typical of very low migration rates. The branches while there exist only two ancestors can be very long while we wait for the necessary migration event. Mutations on these branches leave a very pronounced signature in the site frequency spectrum for the whole population; we therefore observe an excess of variants at intermediate frequency.

8.2 Coalescence time for a pair of lineages

Here, we focus on results for a sample of size 2. Denote by $T_{\mathrm{w}}^{(\alpha)}$ the time to coalescence when both sequences are sampled from the same deme α, and denote by $T_{\mathrm{b}}^{(\alpha\beta)}$ the time to coalescence when one sequence is sampled from deme α and one is sampled from deme β.

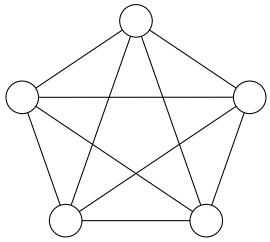


Fig. 8.2: The symmetric island model. Subpopulation sizes are all equal, as are migration rates between any two islands.

The subscripts "w" and "b" respectively indicate "within" and "between". By utilizing the Markov nature of the process and considering the most recent event back in time, we obtain the following system of equations:

$$\mathbb{E}[T_{\mathrm{w}}^{(\alpha)}] = \frac{1}{m_\alpha + \frac{1}{r_\alpha}} + \sum_{\beta \neq \alpha} \frac{m_{\alpha\beta}}{m_\alpha + \frac{1}{r_\alpha}} \mathbb{E}[T_{\mathrm{b}}^{(\alpha\beta)}], \tag{8.3}$$

$$\begin{aligned}\mathbb{E}[T_{\mathrm{b}}^{(\beta)}] = {} & \frac{1}{\frac{m_\alpha}{2} + \frac{m_\beta}{2}} + \frac{m_{\alpha\beta}}{m_\alpha + m_\beta} \mathbb{E}[T_{\mathrm{w}}^{(\beta)}] + \frac{m_{\beta\alpha}}{m_\alpha + m_\beta} \mathbb{E}[T_{\mathrm{w}}^{(\alpha)}] \\ & + \sum_{\gamma \neq \alpha, \beta} \left[\frac{m_{\alpha\gamma}}{m_\alpha + m_\beta} \mathbb{E}[T_{\mathrm{b}}^{(\gamma\beta)}] + \frac{m_{\beta\gamma}}{m_\alpha + m_\beta} \mathbb{E}[T_{\mathrm{b}}^{(\alpha\gamma)}] \right].\end{aligned} \tag{8.4}$$

This system does not exhibit a simple closed-form in general, and so we look at a special case.

8.2.1 Symmetric Island Model

Suppose each island has the same size and each pair has the same migration rates (Figure 8.2), so that $N = N_\alpha$ and $m_\alpha = m$, for all α (note the different meaning of N in this model). The migration parameter between any two demes is $m_{\alpha\beta} = \frac{m}{d-1}$.

The system (8.3)–(8.4) becomes

$$\begin{aligned}\mathbb{E}[T_{\mathrm{w}}] &= \frac{1}{m+1} + \frac{m}{m+1} \mathbb{E}[T_{\mathrm{b}}], \\ \mathbb{E}[T_{\mathrm{b}}] &= \frac{1}{m} + \frac{1}{d-1} \mathbb{E}[T_{\mathrm{w}}] + \frac{d-2}{d-1} \mathbb{E}[T_{\mathrm{b}}],\end{aligned}$$

which can be solved to obtain

$$\mathbb{E}[T_{\mathrm{w}}] = d, \qquad \mathbb{E}[T_{\mathrm{b}}] = d + \frac{d-1}{m}. \tag{8.5}$$

This result deserves several remarks.

1. Time to coalescence depends on the number of demes (and is proportional to the total population size dN).

2. $\mathbb{E}[T_{\mathrm{w}}] \leq \mathbb{E}[T_{\mathrm{b}}]$, with equality as $m \to \infty$.
3. $\mathbb{E}[T_{\mathrm{w}}]$ is independent of the migration parameter m. This at-first-sight surprising *invariance result* is sometimes called Strobeck (1987)'s theorem. An intuition is as follows. The time to coalescence is shorter when there is less migration to take the two lineages apart, and there is less migration with decreasing m. On the other hand, if there is a migration event the time to coalescence is subsequently shorter when there is *more* migration—so that the two lineages can be brought back into the same deme. It turns out that these effects on $\mathbb{E}[T_{\mathrm{w}}]$ completely cancel out. More generally, this invariance result holds whenever the migration model is *isotropic*, that is, we have the same pattern of migration for all demes. Formally, we may define population structure to be *isotropic* if for all α, β, there exists a permutation $\sigma \in \mathscr{S}_d$ such that $\sigma(\alpha) = \beta$ and $m_{\gamma\delta} = m_{\sigma(\gamma)\sigma(\delta)}$, for each γ, δ (Strobeck, 1987). Isotropy is a special case of *conservative migration* (see below).
4. Higher order moments *do* depend on m. For example,

$$\mathrm{Var}[T_{\mathrm{w}}] = d^2 + 2\frac{(d-1)^2}{m},$$
$$\mathrm{Var}[T_{\mathrm{b}}] = d^2 + 2\frac{(d-1)^2}{m} + \frac{(d-1)^2}{m^2}.$$

The mean and variance of T_{w} and T_{b} as a function of m is plotted in Figure 8.3 in the case $d = 2$. Note the unusual behavior as $m \to 0$. $\mathbb{E}[T_{\mathrm{w}}]$ remains bounded while $\mathrm{Var}[T_{\mathrm{w}}]$ increases, so that the p.d.f. $f_{T_{\mathrm{w}}}(t)$ becomes skewed as $m \to 0$. In this limit we might expect, but fail, to observe convergence of the moments $\mathbb{E}[T_{\mathrm{w}}]$ and $\mathrm{Var}[T_{\mathrm{w}}]$ to those of the unstructured coalescent for the deme considered in isolation: We know that for $m = 0$ we have $\mathbb{E}[T_{\mathrm{w}}] = 1$ and $\mathrm{Var}[T_{\mathrm{w}}] = 1$. There is therefore a discontinuity in these moments as $m \to 0$. In fact, we do have the following convergence result (which is nonetheless not strong enough to ensure convergence of these moments).

Theorem 8.1 (Nath and Griffiths 1993). *For $d = 2$,*

$$f_{T_w}(t) \to e^{-t} \text{ and } f_{T_b}(t) \to 0$$

as $m \to 0$, where convergence is pointwise.

The p.d.f.s $f_{T_{\mathrm{w}}}(t)$ and $f_{T_{\mathrm{b}}}(t)$ are plotted in Figure 8.4 for the example case $d = 2$.

8.2.2 Identity-by-descent (IBD) in the symmetric island model

Definition 8.2. A sample of size two is said to be *identical-by-descent* (IBD) if the two samples coalesce with each other with no intervening mutations. Under the infinite-alleles model a sample is IBD iff the samples are identical. Under more general mutation models the samples could be identical despite having experienced (recurrent, parallel, ...) mutations; that is, they could be *identical-by-state*.

In a similar manner to coalescence times, we can calculate IBD probabilities in the restricted setting of the symmetric island model. We recall the mutation parameter $\theta = 2Nu$, and denote by $p_{\mathrm{w}}(\theta)$ the probability that two sequences sampled from the same deme share IBD.

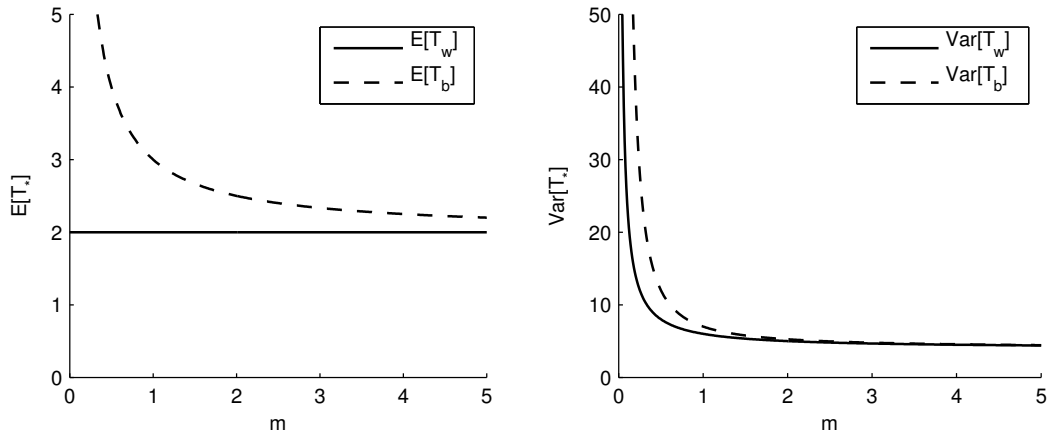


Fig. 8.3: The mean and variance in the time to coalescence for two sequences drawn within the same deme (w) or between demes (b), for a symmetric, two-island model.

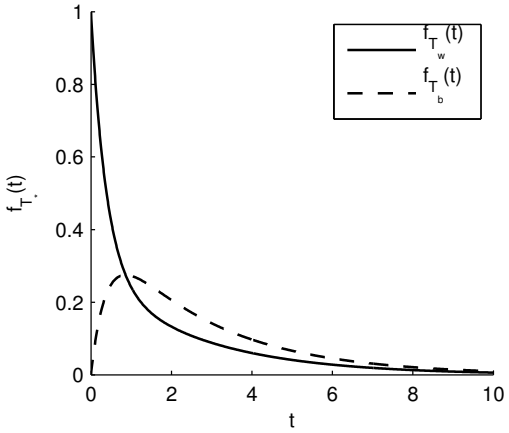


Fig. 8.4: The p.d.f.s for the time to coalescence for two sequences drawn within the same deme (w) or between demes (b), for a symmetric, two-island model. Here, $m = 1$.

Similarly, denote by $p_{\mathrm{b}}(\theta)$ the probability that two sequences sampled from the different demes share IBD.

Proposition 8.3. *Let* $\bar{m} = \dfrac{m}{d-1}$ *and* $D = \theta^2 + \theta(1 + d\bar{m}) + \bar{m}$. *Then,*

$$p_w(\theta) = \frac{\theta + \bar{m}}{D} \qquad \textit{and}\; p_b(\theta) = \frac{\bar{m}}{D}.$$

Before proving this result, we make a couple of observations:

1. $p_{\mathrm{b}}(\theta) \leq p_{\mathrm{w}}(\theta)$, with equality only if $\theta = 0$.
2. As $\bar{m} \to \infty$, $p_{\mathrm{w}}(\theta) + p_{\mathrm{b}}(\theta) \to \dfrac{1}{d\theta + 1}$.

Each of these observations should agree with our intuition. (Why?)

Proof. By considering the most recent event back in time, we can write down the following system of equations:

$$p_{\mathrm{w}}(\theta) = \frac{1}{1+\theta+m} + \frac{m}{1+\theta+m} p_{\mathrm{b}}(\theta),$$
$$p_{\mathrm{b}}(\theta) = \frac{m/(d-1)}{\theta+m} p_{\mathrm{w}}(\theta) + \frac{m(d-2)/(d-1)}{\theta+m} p_{\mathrm{b}}(\theta),$$

which can easily be solved to obtain the given expressions.

8.2.3 Wright's F_{ST}

We can relate the parameters of the coalescent model to Wright's classical measure of population structure, F_{ST}, using its definition in terms of IBD:

$$F_{ST} = \frac{f_0 - \bar{f}}{1 - \bar{f}},$$

where f_0 is the probability of IBD for a sample of size 2 taken from within one deme, and $\bar{f}$ is the probability of IBD for a sample of size 2 taken at random from the whole population.

The precise meaning of F_{ST} depends on the underlying demographic model. Under the symmetric island model, $f_0 = p_{\mathrm{w}}(\theta)$ and $\bar{f} = \frac{1}{d} p_{\mathrm{w}}(\theta) + \frac{d-1}{d} p_{\mathrm{b}}(\theta)$, so we can use the results of the previous section to get

$$F_{ST} = \frac{1}{1 + \frac{md^2}{(d-1)^2} + \frac{\theta d}{d-1}} \approx \frac{1}{1 + \frac{md^2}{(d-1)^2}}, \text{ when } \theta \ll 1.$$

A direct coalescent interpretation of F_{ST} for the symmetric island model was provided by Slatkin (1991) when θ is small. Substituting for the approximations

$$p_{\mathrm{w}}(\theta) = \mathbb{E}[e^{-\theta T_{\mathrm{w}}}] \approx 1 - \theta \mathbb{E}[T_{\mathrm{w}}],$$
$$p_{\mathrm{b}}(\theta) = \mathbb{E}[e^{-\theta T_{\mathrm{b}}}] \approx 1 - \theta \mathbb{E}[T_{\mathrm{b}}],$$

for the symmetric island model we find

$$F_{ST} \approx \frac{\mathbb{E}[T] - \mathbb{E}[T_{\mathrm{w}}]}{\mathbb{E}[T]},$$

where $T = \frac{1}{d} T_{\mathrm{w}} + \frac{d-1}{d} T_{\mathrm{b}}$ is the coalescence time for a sample of size two drawn at random from the whole population.

Consider the demographic model shown in Figure 8.5, depicting a clean split model. If time is measured in units of N generations, then

$$\mathbb{E}[T_b] = 2 + t,$$
$$\mathbb{E}[T_w] = 1 + e^{-t},$$

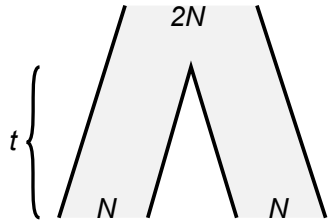


Fig. 8.5: A clean split model with divergence time t (in coalescent unit).

and

$$F_{ST} \approx \frac{\mathbb{E}[T_b] - \mathbb{E}[T_w]}{\mathbb{E}[T_b] + \mathbb{E}[T_w]} = \frac{1 + t - e^{-t}}{3 + t + e^{-t}}.$$

8.3 Conservative migration

It is worth returning to the parameter $c_{\alpha\beta}$ introduced at the beginning. We ought to give it some further justification. We interpret $c_{\alpha\beta}$ as the probability that the parent of an individual in deme α came from deme β one generation ago in our discrete model of population of reproduction, but what if these migrations occur with very high frequency, or asymmetrically? It seems feasible that so much migration might affect our assumption of constant deme sizes. For concreteness assume a Wright-Fisher model in which N_α individuals have N_α offspring in each round of reproduction and then die, with migration rates forward in time denoted by $f_{\alpha\beta}$. If each offspring then migrated independently, we can no longer guarantee that each N_α remains fixed. In effect, we require *conservative migration*: that the total number of non-emigrants plus the number of immigrants is equal to N_α in each generation, for each α:

$$N_\alpha \left(1 - \sum_{\beta:\beta\neq\alpha} f_{\alpha\beta}\right) + \sum_{\beta:\beta\neq\alpha} N_\beta f_{\beta\alpha} = N_\alpha,$$

which simplifies slightly to

$$N_\alpha \sum_{\beta:\beta\neq\alpha} f_{\alpha\beta} = \sum_{\beta:\beta\neq\alpha} N_\beta f_{\beta\alpha}.$$

When this holds, *individuals in the population do not migrate independently.* A useful way to think of this is that a "chunk" of the offspring (of deterministically chosen size $N_\alpha f_{\alpha\beta}$) in each deme α is chosen to move to deme β in each generation. Considering this flux backwards in time, we see that the backwards migration parameters $c_{\alpha\beta}$ are related to $f_{\alpha\beta}$ by

$$c_{\alpha\beta} = \frac{N_\beta f_{\beta\alpha}}{N_\alpha},$$

known as the *backward migration fraction.* For the symmetric island model, $c_{\alpha\beta} = f_{\alpha\beta} = f_{\beta\alpha}$.

(One might wonder what happens if we insist upon independence for the migration of individuals in the population. An alternative formulation of the Wright-Fisher model with

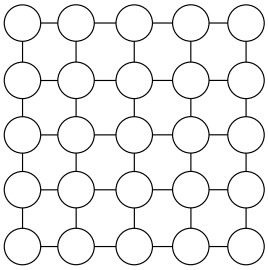


Fig. 8.6: The two-dimensional stepping-stone model. Migration occurs only between adjacent islands.

migration is as follows. Suppose N_α individuals each migrate independently with probability $f_{\alpha\beta}$ to each deme β. After migration, deme α now has size $N_\alpha + X_\alpha$, where X_α is the random number of net migrants. These $N_\alpha + X_\alpha$ individuals then have N_α offspring and die, to form the next generation. Notohara (1990) showed that this converges to the same structured coalescent process as the simpler deterministic formulation given above.)

8.4 Further extensions

We have focused on simple island models to illustrate the structured coalescent process. We briefly mention further extensions to this model which attempt to capture more biologically realistic scenarios. Perhaps the biggest restriction is the lack of a spatial component relating the demes. Many species exhibit *isolation by distance*—a correlation between pairwise genetic differences and pairwise geographic distances. It is implicit in simple models like the island model that the habitat limits an individual's ability to disperse. Isolation by distance suggests that one should also account for the physical ability of an individual to disperse itself. As a compromise, one modification to the simple island model is to consider a one- or two-dimensional array of islands, with non-zero migration occurring only between adjacent islands (Figure 8.6). More ambitiously, one might model the spatial co-ordinate of individuals as diffusion processes in $\mathbb{R}$: when two diffusions collide, the individuals coalesce. Extending this to $\mathbb{R}^2$ is mathematically challenging, since non-trivial diffusion processes do not collide in $\mathbb{R}^2$.

Finally, we can attempt to combine population substructure and migration with variable population size to postulate non-equilibrium models of population history. Two geographically distinct demes exhibit limited migration until some time in the past when they split, and before which all individuals mated randomly. A well-known example is the "out-of-Africa" expansion of modern humans. A model for this history could also include recent population growth, a bottleneck shortly after the split, and so on. Example coalescent models incorporating many of these features simultaneously are discussed in Excoffier et al (2013); Gravel et al (2011); Gutenkunst et al (2009); Tennessen et al (2012).

8.5 Multi-population SFS

Kamm et al (2020, 2017)

References

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. PLoS Genetics 9(10):e1003,905

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD, Altshuler DL, et al (2011) Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences 108(29):11,983–11,988

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics 5(10):e1000,695

Kamm J, Terhorst J, Durbin R, Song YS (2020) Efficiently inferring the demographic history of many populations with allele count data. Journal of the American Statistical Association 115(531):1472–1487

Kamm JA, Terhorst J, Song YS (2017) Efficient computation of the joint sample frequency spectra for multiple populations. Journal of Computational and Graphical Statistics 26(1):182–194

Nagylaki T (1980) The strong-migration limit in geographically structured populations. Journal of Mathematical Biology 9(2):101–114

Nath H, Griffiths R (1993) The coalescent in two colonies with symmetric migration. Journal of Mathematical Biology 31(8):841–851

Notohara M (1993) The strong-migration limit for the genealogical process in geographically structured populations. Journal of Mathematical Biology 31(2):115–122

Slatkin M (1991) Inbreeding coefficients and coalescence times. Genetical Research 58(02):167–175

Strobeck C (1987) Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics 117(1):149–153

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337(6090):64–69

Wakeley J (2008) Coalescent Theory: An Introduction. Roberts & Company Publishers

Part IV
Recombination

Chapter 9
The coalescent with recombination

In this chapter, we consider generalizing the coalescent to incorporate *recombination*, which is an important biological process common to most forms of life. A far-reaching consequence of recombination is that different positions in the genome can have different evolutionary histories. Figure 9.1 illustrates why this is the case. There are two types of outcome of meiotic recombination, namely *crossover* and *gene-conversion*; see Paigen and Petkov (2010); Sasaki et al (2010) for reviews of recombination pathways. Here, we consider only the first type and use the term recombination to mean crossover. The coalescent with crossover recombination has been studied by a number of researchers in the past. Early works on the topic include Griffiths (1981, 1991); Griffiths and Marjoram (1997); Hudson (1983). See Wiuf (2000) for a coalescent approach to gene-conversion.

Throughout this chapter, we assume a constant population size. This assumption will be relaxed in the next chapter, where we discuss sampling distributions.

9.1 Wright-Fisher model with recombination

The genealogical process for N diploid individuals is typically approximated by a model with $2N$ haploid individuals. This approximation is reasonable for describing the dynamics at a sufficiently long evolutionary timescale. Here, we consider the Wright-Fisher model with $2N$ chromosomes and, for ease of description, assume that the per-generation crossover probability is constant across the chromosome. The per-generation crossover probability for the entire region is denoted by r.

Mathematically crossover recombination is modeled as follows: independently of all other chromosomes in the population, with probability $1 - r$ each chromosome chooses only one parent from the previous generation (corresponding to there being no recombination), or with probability r two parents are chosen and a crossover breakpoint is introduced uniformly at random (Figure 9.2). We refer to $\rho = 4Nr$ as the population-scaled recombination rate which is held fixed as $N \to \infty$ and $r \to 0$.

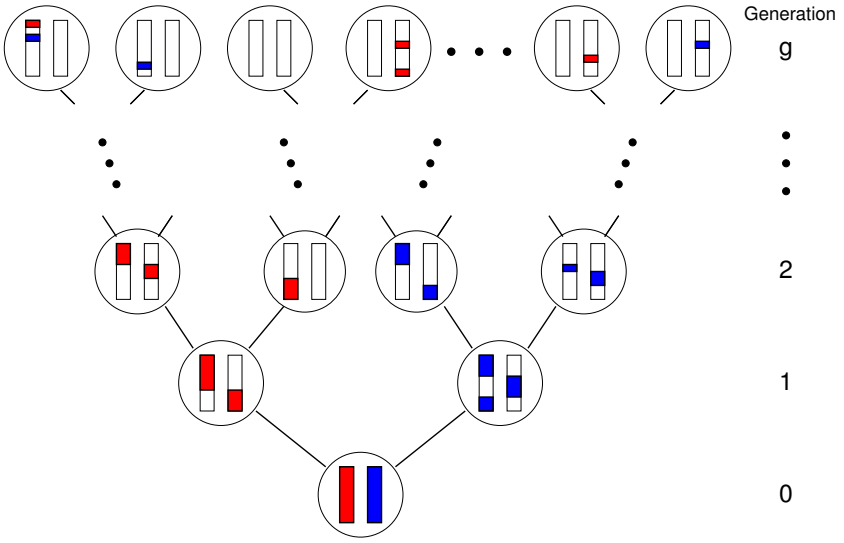


Fig. 9.1: Illustration of genetic material contributions from ancestors. Each circle represents a diploid individual with two *homologous* chromosomes, one from their mother and the other from their father. During meiosis, recombination takes the two *homologous* chromosomes of an individual and creates a mosaic chromosome, which gets transmitted to the individual's child.

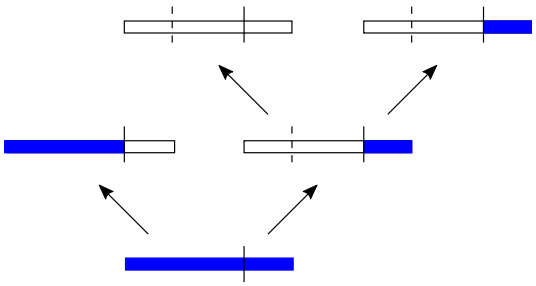


Fig. 9.2: The blue chromosome at the bottom inherited genetic material from two parental chromosomes in the previous generation, which underwent recombination. Recombination breakpoints are marked by vertical bars. One of the grandparental chromosomes (upper left one) is not a genetic ancestor to the blue chromosome since it contributes no genetic material.

9.2 Geneaological ancestral process

A *genealogical* ancestor of an individual i is any ancestor related to i through a sequence of parent-child relationships, whereas a *genetic* ancestor is a genealogical ancestor who actually contributes some genetic material to i. In Figure 9.2, one of the grandparents (upper left one) is a genealogical ancestor but not a genetic ancestor.

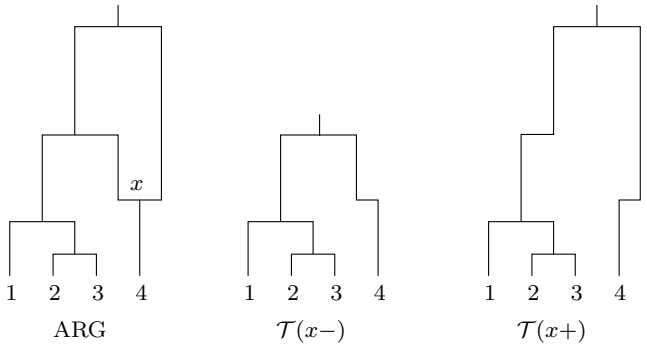


Fig. 9.3: An ancestral recombination graph (ARG) and the embedded marginal trees for a sample of size 4. The ARG has a single recombination event with breakpoint at position x along the chromosome. When there is a recombination event, a lineage splits into two, going backwards in time. The marginal coalescent tree for positions to the left of the breakpoint x is $\mathcal{T}(x-)$ and that to the right of x is $\mathcal{T}(x+)$.

Definition 9.1 (Genealogical ancestral process). In the Wright-Fisher model described in Section 9.1, let $A_n^{(N,\rho)}(\tau)$ denote the number of genealogical ancestors in generation τ for a sample of size n taken in generation 0. As usual, τ runs backwards in time. The genealogical ancestral process $\{A_n^{(N,\rho)}(\tau), \tau = 0, 1, \ldots\}$ is a birth-death process with state space $\mathbb{N}$ and transition probabilities

$$\mathbb{P}(A_n^{(N,\rho)}(\tau+1) = j \mid A_n^{(N,\rho)}(\tau) = k) = \begin{cases} 1 - \frac{1}{2N}\binom{k}{2} - \frac{1}{2N}\frac{k\rho}{2} + O(\frac{1}{N^2}), & \text{if } j = k, \\ \frac{1}{2N}\binom{k}{2} + O(\frac{1}{N^2}), & \text{if } j = k-1, \\ \frac{1}{2N}\frac{k\rho}{2}, & \text{if } j = k+1, \\ O(\frac{1}{N^2}), & \text{otherwise.} \end{cases}$$

Taking the usual coalescent limit by rescaling time and sending $N \to \infty$, the genealogical ancestral process converges to a continuous-time process. Specifically, for all $t \in \mathbb{R}_{\geq 0}$,

$$A_n^{(N,\rho)}(\lfloor 2Nt \rfloor) \xrightarrow{d} A_n^{(\rho)}(t).$$

The state space of $\{A_n^{(\rho)}(t), t \geq 0\}$ is also $\mathbb{N}$ and state 1 is reached with probability 1 since, when there are k lineages, the coalescence rate is quadratic in k while the recombination rate is linear in k. A sample path from the process can be represented a data structure called the *ancestral recombination graph* (ARG), illustrated in Figure 9.3.

9.2.1 Grand MRCA

The time to the grand most recent common ancestor (GMRCA) of a sample is defined as

$$W_n^{(\rho)} = \inf\{t \geq 0 \mid A_n^{(\rho)}(t) = 1\}.$$

Griffiths (1991) obtained the following result on the expectation of $W_n^{(\rho)}$:

Theorem 9.2 (Expected time to the GMRCA). *The expected value of the waiting time until the GMRCA is*

$$\mathbb{E}[W_n^{(\rho)}] = \frac{2}{\rho}\int_0^1 \frac{1-x^{n-1}}{1-x}\left[e^{\rho(1-x)}-1\right]dx. \tag{9.1}$$

Remark 9.3. Before presenting a proof of the above result, we first make a few observations:

1. $\mathbb{E}[W_n^{(\rho)}]$ is a monotonically increasing function of ρ.
2. As $\rho \to 0$, $\mathbb{E}[W_n^{(\rho)}] \to 2\left(1-\frac{1}{n}\right)$, which is the correct answer for $\rho = 0$ (c.f., (1.13)).
3. For $n = 2$,
$$\mathbb{E}[W_2^{(\rho)}] = \frac{2}{\rho^2}[e^\rho - 1 - \rho]. \tag{9.2}$$

The expectation $\mathbb{E}[W_2^{(\rho)}]$ grows very fast with ρ, reaching above 400 for $\rho = 10$.

Proof. By conditioning on the first event back in time, we obtain the recursion

$$\mathbb{E}[W_n^{(\rho)}] = \frac{1}{\binom{n}{2}+\frac{n\rho}{2}} + \frac{n-1}{n-1+\rho}\mathbb{E}[W_{n-1}^{(\rho)}] + \frac{\rho}{n-1+\rho}\mathbb{E}[W_{n+1}^{(\rho)}], \tag{9.3}$$

where the first term corresponds to the expected waiting time until the first event. When $n = 1$, the GMRCA has been reached already, so $\mathbb{E}[W_1^{(\rho)}] = 0$. However, the recursion in (9.3) is not bounded from above (i.e., the index n will keep increasing), so it cannot be solved directly. Griffiths (1991) suggested a clever trick of considering a related process with a reflecting barrier at $b > n$. Let $W_n^{(\rho,b)}$ denote the time to the GMRCA in this new process. Then $\mathbb{E}[W_n^{(\rho,b)}]$ satisfies (9.3) for $2 \le n \le b-1$, while for $n = b$

$$\mathbb{E}[W_b^{(\rho,b)}] = \mathbb{E}[W_{b-1}^{(\rho)}].$$

With this boundary condition, (9.3) can be solved in closed form:

$$\mathbb{E}[W_n^{(\rho,b)}] = 2\sum_{k=2}^{n}(k-2)!\sum_{j=0}^{b-k-1}\frac{\rho^j}{(j+k)!},$$

where $2 \le n \le b-1$. Then, taking the limit as the reflecting barrier $b \to \infty$, we get

$$\lim_{b\to\infty}\mathbb{E}[W_n^{(\rho,b)}] = \mathbb{E}[W_n^{(\rho)}] = 2\sum_{k=2}^{n}(k-2)!\sum_{j=0}^{\infty}\frac{\rho^j}{(j+k)!},$$

which can be shown to be equal to the integral representation shown in (9.1). See Griffiths (1991) for further details. □

Theorem 9.4 (Expected time while j lineages). *Let $W_{n,j}^{(\rho)}$ denote the total waiting time while there are j lineages until the GMRCA is reached. Then,*

$$\mathbb{E}[W_{n,j}^{(\rho)}] = 2 \sum_{k=2}^{\min(j,n)} (k-2)! \frac{\rho^{j-k}}{j!},$$

for $j = 2, 3, \ldots$.

Proof. As noted by Griffiths (1991), this result can be obtained by simply modifying the recursion (9.3) as

$$\mathbb{E}[W_{n,j}^{(\rho)}] = \frac{\delta_{n,j}}{\binom{n}{2} + \frac{n\rho}{2}} + \frac{n-1}{n-1+\rho}\mathbb{E}[W_{n-1,j}^{(\rho)}] + \frac{\rho}{n-1+\rho}\mathbb{E}[W_{n+1,j}^{(\rho)}],$$

and using the same trick of introducing a reflecting barrier. □

9.2.2 The width of a genealogical ARG

In the previous section, we considered the expected height of a genealogical ARG (Figure 9.4). Here, we consider its *width*; more precisely, the maximum number of genealogical ancestors for a sample of size n before the GMRCA is reached. We use $A_{\max}(n, \rho)$ to denote this number.

Theorem 9.5. *For* $m \geq n$,

$$\mathbb{P}(A_{\max}(n,\rho) \leq m) = \sum_{j=n-1}^{m-1} \frac{j!}{\rho^j} \Bigg/ \sum_{k=0}^{m-1} \frac{k!}{\rho^k}. \tag{9.4}$$

Proof. Define $c_n(m) = \mathbb{P}(A_{\max}(n,\rho) \leq m)$. After observing that $c_1(m) = 1$ for all $m \geq 1$ and $c_n(m) = 0$ for all $m < n$, we may construct a recurrence relation for $c_n(m)$ by conditioning on the first event (coalescence or recombination) back in time:

$$c_n(m) = \frac{n-1}{n-1+\rho} c_{n-1}(m) + \frac{\rho}{n-1+\rho} c_{n+1}(m), \tag{9.5}$$

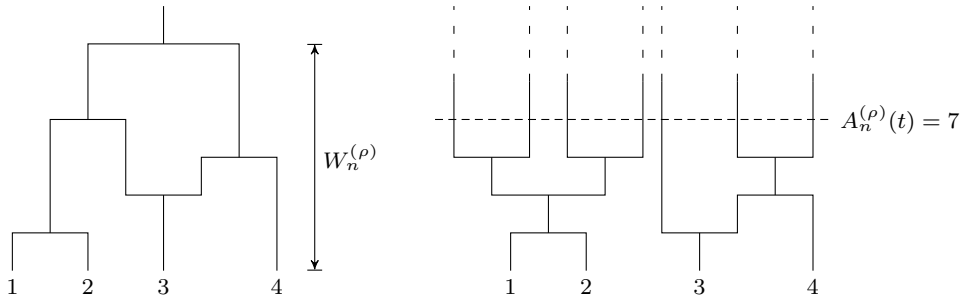


Fig. 9.4: Ancestral recombination graphs illustrating height and width. The figure on the right shows how recombination can cause the graph to get much wider than the sample size.

for $2 \leq n \leq m$. We now detail how to solve this recursion. First, rewrite the left-hand side as

$$\Big(\frac{n-1}{n-1+\rho} + \frac{\rho}{n-1+\rho}\Big)c_n(m) = \frac{n-1}{n-1+\rho}c_{n-1}(m) + \frac{\rho}{n-1+\rho}c_{n+1}(m). \tag{9.6}$$

Then, letting $d_n(m) = c_n(m) - c_{n-1}(m)$, we obtain

$$\begin{aligned} & \frac{n-1}{n-1+\rho}d_n(m) = \frac{\rho}{n-1+\rho}d_{n+1}(m), \\ \implies \quad & d_{n+1}(m) = \frac{n-1}{\rho}d_n(m), \\ \implies \quad & d_{j+1}(m) = \frac{(j-1)!}{\rho^{j-1}}d_2(m), \text{ for } 1 \leq j \leq m. \end{aligned}$$

Now, by telescoping,

$$\begin{aligned} c_{n+1}(m) - c_1(m) &= [c_{n+1}(m) - c_n(m)] + [c_n(m) - c_{n-1}(m)] + \cdots + [c_2(m) - c_1(m)] \\ &= d_{n+1}(m) + d_n(m) + \cdots + d_2(m) \\ &= \sum_{j=2}^{n+1} d_j(m) = d_2(m) \sum_{j=0}^{n-1} \frac{j!}{\rho^j}. \end{aligned} \tag{9.7}$$

To solve for $d_2(m)$, set $n = m$ in (9.7), and plug in $c_{m+1}(m) = 0$ and $c_1(m) = 1$ to obtain

$$d_2(m) = -\left[\sum_{j=0}^{m-1} \frac{j!}{\rho^j}\right]^{-1}.$$

Using this result in (9.7), we get

$$c_n(m) = c_1(m) + d_2(m) \sum_{j=0}^{n-2} \frac{j!}{\rho^j} = 1 - \sum_{j=0}^{n-2} \frac{j!}{\rho^j} \Bigg/ \sum_{j=0}^{m-1} \frac{j!}{\rho^j} = \sum_{j=n-1}^{m-1} \frac{j!}{\rho^j} \Bigg/ \sum_{j=0}^{m-1} \frac{j!}{\rho^j},$$

which is the desired result. □

Finally, it bears mentioning that

$$\frac{A_{\max}(n,\rho)}{n} \xrightarrow{p} 1 \text{ as } n \to \infty. \tag{9.8}$$

Intuitively this is because the rate of coalescence is quadratic in n while the rate of recombination is linear in n.

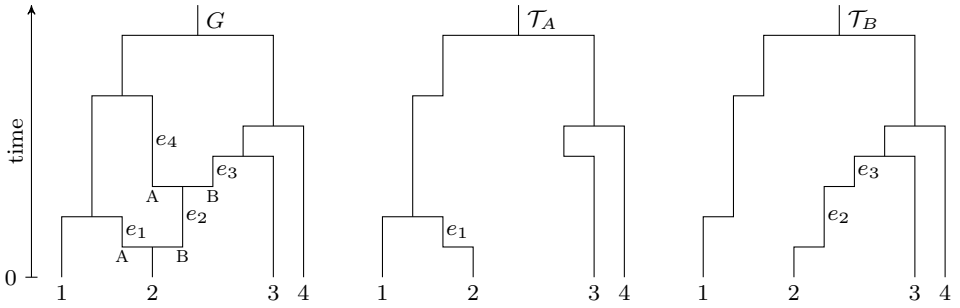


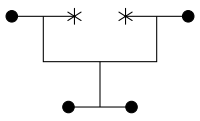
Fig. 9.5: A genealogical ancestral recombination graph G with two marginal coalescent trees.

9.3 Unreduced and reduced ancestral processes for ARGs

Consider the genealogical ancestral recombination graph G shown on the left of Figure 9.5, where four of the edges resulting from recombination are labeled (e_1, e_2, e_3, e_4). For simplicity, we consider just two loci in the subsequent discussion:

Locus A ●——● Locus B

At each of the two recombination events displayed in Figure 9.5, we assume that the parent on the left supplies Locus A, while the parent on the right supplies Locus B. This induces coalescent trees for the two locus, $\mathcal{T}_A$ and $\mathcal{T}_B$ (Figure 9.5, center and right), that are of different topologies. The more recent of the two recombination events can be depicted this way:



A filled ball denotes an ancestral locus — one that contributes genetic material to at least one of the descendants in the sample. A star denotes a non-ancestral locus — one that contributes no genetic material to the sample. The diagram shows that the child has two ancestral loci, whereas each parent has one ancestral and one non-ancestral locus.

Let $E(G)$ denoted the set of vertical edges of G. (Horizontal edges have no biological meaning.) We can partition $E(G)$ into four types, depending on whether each locus is ancestral or non-ancestral.

Type A (●—∗): $\mathcal{A}(G) = \{e \in E(G) \mid e \in E(\mathcal{T}_A) \text{ and } e \notin E(\mathcal{T}_B)\} = \{e_1\}$
Type B (∗—●): $\mathcal{B}(G) = \{e \in E(G) \mid e \notin E(\mathcal{T}_A) \text{ and } e \in E(\mathcal{T}_B)\} = \{e_2, e_3\}$
Type C (●—●): $\mathcal{C}(G) = \{e \in E(G) \mid e \in E(\mathcal{T}_A) \text{ and } e \in E(\mathcal{T}_B)\} = E(G) \setminus \{e_1, e_2, e_3, e_4\}$
Type D (∗—∗): $\mathcal{D}(G) = \{e \in E(G) \mid e \notin E(\mathcal{T}_A) \text{ and } e \notin E(\mathcal{T}_B)\} = \{e_4\}$

We can now define the unreduced Markov chain for an ancestral recombination graph G with two loci.

Definition 9.6 (Unreduced ancestral process). Partition the edge set $E(G)$ into $\mathcal{A}(G), \mathcal{B}(G), \mathcal{C}(G), \mathcal{D}(G)$ as above, and let $N_{\mathcal{A}}(t)$ denote the number of edges in $\mathcal{A}(G)$ at

time t, and let $N_{\mathcal{B}}(t)$, $N_{\mathcal{C}}(t)$, and $N_{\mathcal{D}}(t)$ be defined analogously for $\mathcal{B}(G)$, $\mathcal{C}(G)$, and $\mathcal{D}(G)$. The *unreduced* ancestral process for G is:

$$\big\{U_t = \big(N_A(t), N_B(t), N_C(t), N_D(t)\big), t \geq 0\big\}.$$

Remark 9.7. We first note a few facts:

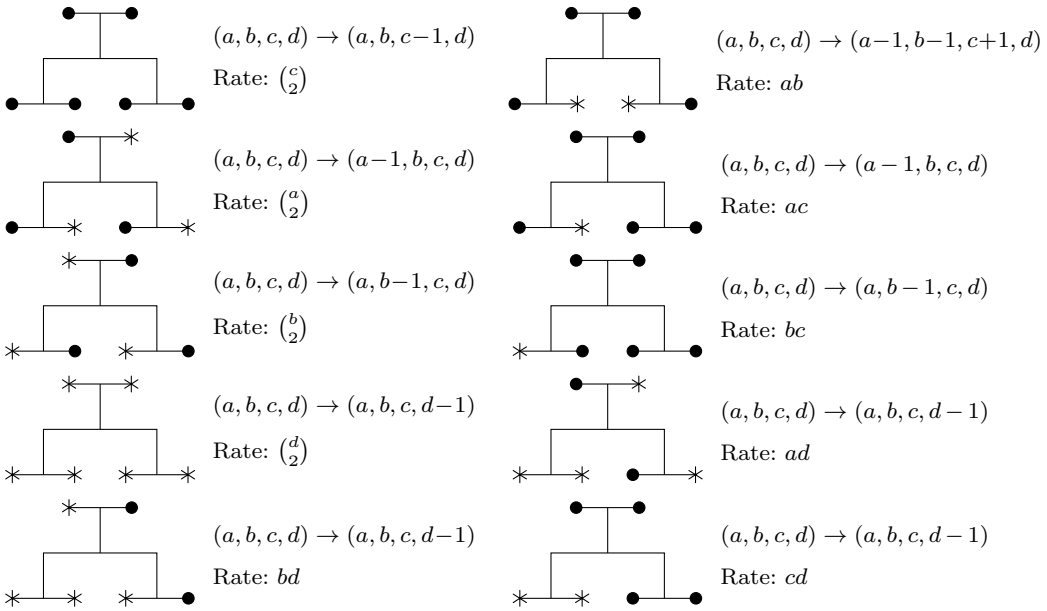
1. $N_{\mathcal{A}}(t) + N_{\mathcal{B}}(t) + N_{\mathcal{C}}(t) + N_{\mathcal{D}}(t) = A_n^{(\rho)}(t)$, the genealogical ancestral process, i.e., the number of genealogical ancestors at time t of a sample of size n.
2. $N_{\mathcal{A}}(t) + N_{\mathcal{C}}(t)$ is the marginal ancestral process of Locus A.
3. $N_{\mathcal{B}}(t) + N_{\mathcal{C}}(t)$ is the marginal ancestral process of Locus B.
4. We have a closed-form formula for $\mathbb{P}(C_n^A(t) = \alpha)$, where $C_n(t)$ is the marginal n-coalescent and $\alpha \in \mathcal{P}_{[n]}$, but we lack a closed-form formula for the joint distribution for two loci $\mathbb{P}(C_n^A(t) = \alpha, C_n^B(t) = \beta)$, where $\alpha, \beta \in \mathcal{P}_{[n]}$, except when $\rho = 0$ or $\rho = \infty$. When $\rho = 0$, the two loci are never separated and

$$\mathbb{P}(C_n^A(t) = \alpha, C_n^B(t) = \beta) = \begin{cases} \mathbb{P}(C_n^A(t) = \alpha), & \text{if } \alpha = \beta, \\ 0, & \text{otherwise.} \end{cases}$$

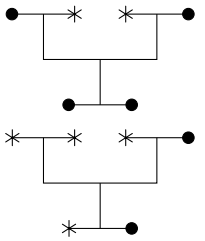
   When $\rho = \infty$, the two loci are completely independent and

$$\mathbb{P}(C_n^A(t) = \alpha, C_n^B(t) = \beta) = \mathbb{P}(C_n^A(t) = \alpha)\mathbb{P}(C_n^B(t) = \beta).$$

We can obtain the transition rates for U_t by enumerating the various kinds of coalescence and recombination events. There are ten kinds of coalescence events:



There are four kinds of recombinations events:

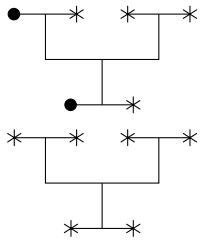


$(a, b, c, d) \to (a+1, b+1, c-1, d)$

Rate: $\frac{c\rho}{2}$

$(a, b, c, d) \to (a, b, c, d+1)$

Rate: $\frac{b\rho}{2}$



$(a, b, c, d) \to (a, b, c, d+1)$

Rate: $\frac{a\rho}{2}$

$(a, b, c, d) \to (a, b, c, d+1)$

Rate: $\frac{d\rho}{2}$

Grouping together transitions of the same type yields

$$(a,b,c,d) \to \begin{cases} (a,b,c-1,d) & \text{at rate } \binom{c}{2}, \\ (a-1,b-1,c+1,d) & \text{at rate } ab, \\ (a-1,b,c,d) & \text{at rate } \binom{a}{2} + ac, \\ (a,b-1,c,d) & \text{at rate } \binom{b}{2} + bc, \\ (a,b,c,d-1) & \text{at rate } (a+b+c)d + \binom{d}{2}, \\ (a+1,b+1,c-1,d) & \text{at rate } \frac{c\rho}{2}, \\ (a,b,c,d+1) & \text{at rate } \frac{(a+b+d)\rho}{2}. \end{cases} \tag{9.9}$$

The total transition rate is $\binom{m}{2} + \frac{m\rho}{2}$, where $m = a + b + c + d$.

For most questions of interest, note that it is unnecessary to keep track of Type D edges ($*\longrightarrow*$), since what happens to them does not affect the sample. If we take the unreduced Markov chain U_t and delete the fourth element, which is the count of Type D edges, we obtain the following reduced Markov chain:

Definition 9.8 (Reduced ancestral process). Let $N_{\mathcal{A}}(t)$, $N_{\mathcal{B}}(t)$, and $N_{\mathcal{C}}(t)$ be defined as for U_t. The reduced ancestral process is:

$$\big\{R_t = \big(N_A(t), N_B(t), N_C(t)\big), t \geq 0\big\}.$$

The transitions for R_t are the same as for U_t, without the two that involve a change in the fourth element of U_t.

$$(a,b,c) \to \begin{cases} (a,b,c-1) & \text{at rate } \binom{c}{2}, \\ (a-1,b-1,c+1) & \text{at rate } ab, \\ (a-1,b,c) & \text{at rate } \binom{a}{2} + ac, \\ (a,b-1,c) & \text{at rate } \binom{b}{2} + bc, \\ (a+1,b+1,c-1) & \text{at rate } \frac{c\rho}{2}. \end{cases} \tag{9.10}$$

The total transition rate for this reduced Markov chain is $\binom{a+b+c}{2} + \frac{c\rho}{2}$. This is a much more efficient representation of the genealogical process compared to U_t.

9.4 Covariance of marginal TMRCAs at a pair of loci.

The TMRCA for loci A and B may be different, as shown in Figure 9.6. The following theorem characterizes how they covary as a function of ρ:

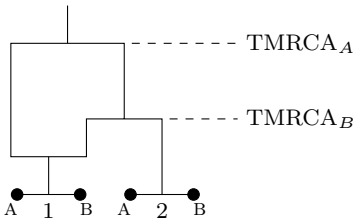


Fig. 9.6: An ancestral recombination graph showing how marginal TMRCAs can differ.

Theorem 9.9 (Griffiths 1981, 1991). *Let R_t be a reduced ancestral process. Suppose $R_0 = (a, b, c)$, where $a+c = 2$ and $b+c = 2$. Let W_A and W_B denote the marginal TMRCAs for loci A and B, respectively. Then,*

$$\begin{aligned}
\mathrm{Cov}(W_A, W_B \mid R_0 = (0,0,2)) &= \frac{\rho + 18}{\rho^2 + 13\rho + 18}, \\
\mathrm{Cov}(W_A, W_B \mid R_0 = (1,1,1)) &= \frac{6}{\rho^2 + 13\rho + 18}, \\
\mathrm{Cov}(W_A, W_B \mid R_0 = (2,2,0)) &= \frac{4}{\rho^2 + 13\rho + 18}.
\end{aligned}$$

Remark 9.10. Some remarks before proving this result:

1. The $*$ symbols in type A and B lineages at time 0 (i.e., in the sample) should be interpreted as missing data.
2. As one might expect, in each case, covariance goes to zero as $\rho \to \infty$.
3. Even when $R_0 = (2, 2, 0)$, it is possible for the TMRCAs to be the same if each Type A taxon coalesces with a Type B taxon before doing anything else.
4. $\mathrm{Cov}(W_A, W_B \mid R_0 = (0, 0, 2))$ is the largest and $\mathrm{Cov}(W_A, W_B \mid R_0 = (2, 2, 0))$ is the smallest, which makes intuitive sense.

Proof. First note that

$$\mathrm{Cov}(W_A, W_B \mid R_0 = R) = \mathbb{E}_R(W_A W_B) - \mathbb{E}_R(W_A)\mathbb{E}_R(W_B) = \mathbb{E}_R(W_A W_B) - 1,$$

where $\mathbb{E}_R$ denotes the conditional expectation given that the process starts in state R. Let $T \sim \mathrm{Exp}(\binom{a+b+c}{2} + \frac{c\rho}{2})$ denote the time of the first jump, and let $R' = (a', b', c')$ be the resulting state. Then,

$$\begin{aligned}
&\mathbb{E}_R(W_A W_B \mid T, R') \\
&= \mathbb{E}_R((W_A - T + T)(W_B - T + T) \mid T, R') \\
&= \mathbb{E}_R((W_A - T)(W_B - T) \mid T, R') + T[\mathbb{E}_R(W_A - T \mid T, R') + \mathbb{E}_R(W_B - T \mid T, R')] + T^2 \\
&= \mathbb{E}_{R'}(W_A W_B) + T[\mathbb{E}_{R'}(W_A) + \mathbb{E}_{R'}(W_B)] + T^2,
\end{aligned}$$

where we have used the Markov property in the last line. Now, note that T and R' are independent, and take the expectation with respect to T and R' to obtain

$$f(R) = \mathbb{E}_R[f(R')] + \mathbb{E}_R(T)\mathbb{E}_R[\mathbb{E}_{R'}(W_A) + \mathbb{E}_{R'}(W_B)] + \mathbb{E}_R(T^2), \tag{9.11}$$

where we have introduced the notation $f(R) := \mathbb{E}_R(W_A W_B)$. The expectations $\mathbb{E}_R(T)$ and $\mathbb{E}_R(T^2)$ are given by

$$\mathbb{E}_R(T) = \frac{1}{\binom{a+b+c}{2} + \frac{c\rho}{2}} \qquad \text{and} \qquad \mathbb{E}_R(T^2) = \frac{2}{[\binom{a+b+c}{2} + \frac{c\rho}{2}]^2},$$

while

$$\mathbb{E}_{R'}(W_A) = \begin{cases} 1, & \text{if } a' + c' = 2, \\ 0, & \text{if } a' + c' = 1 \end{cases} \qquad \text{and} \qquad \mathbb{E}_{R'}(W_B) = \begin{cases} 1, & \text{if } b' + c' = 2, \\ 0, & \text{if } b' + c' = 1 \end{cases}.$$

Using these identities in (9.11), we obtain the following system of coupled linear equations:

$$\begin{aligned} f(0,0,2) &= \frac{\rho}{\rho+1} f(1,1,1) + \frac{2}{\rho+1}, \\ f(1,1,1) &= \frac{2}{\rho+6} f(0,0,2) + \frac{\rho}{\rho+6} f(2,2,0) + \frac{4}{\rho+6}, \\ f(2,2,0) &= \frac{2}{3} f(1,1,1) + \frac{1}{3}, \end{aligned}$$

which can be solved for $f(0,0,2), f(1,1,1), f(2,2,0)$. These results can then be used to determine $\mathrm{Cov}(W_A, W_B \mid R_0 = R) = f(R) - 1$. □

References

Griffiths RC (1981) Neutral two-locus multiple allele models with recombination. Theoretical Population Biology 19:169–186

Griffiths RC (1991) The two-locus ancestral graph. Selected Proceedings of the Sheffield Symposium on Applied Probability IMS Lecture Notes–Monograph Series 18:100–117

Griffiths RC, Marjoram P (1997) An ancestral recombination graph. In: Donnelly P, Tavaré S (eds) Progress in population genetics and human evolution, vol 87, Springer-Verlag, Berlin, pp 257–270

Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology 23:183–201

Paigen K, Petkov P (2010) Mammalian recombination hot spots: properties, control and evolution. Nature Reviews Genetics 11(3):221–233

Sasaki M, Lange J, Keeney S (2010) Genome destabilization by homologous recombination in the germ line. Nature Reviews Molecular Cell Biology 11(3):182–195

Wiuf C (2000) A coalescence approach to gene conversion. Theoretical Population Biology 57:357–367

Chapter 10
Exact and approximate likelihoods under the coalescent with recombination

As discussed in earlier chapters, the probability of observing a sample of DNA sequences plays a fundamental role in various applications, including parameter estimation and ancestral inference. In this chapter, we address the problem of finding such sampling probabilities under the coalescent with recombination. Because it involves integrating out latent variables (genealogical histories) that live in extremely high dimensions, finding the sampling distribution under the full model is a difficult problem. Deriving closed-form formulas is notoriously challenging, while numerical and Monte Carlo approaches are in general computationally intensive and tend to not scale well with data size. To address this challenge, it is important to consider making principled approximations to arrive at a simpler model that accurately captures the essential features of the full model, while facilitating computation. We will discuss some recent work along this line of research.

10.1 Two-locus sampling distributions

For simplicity, we consider a two-locus model with recombination. The two loci are denoted by A and B, and their population-scaled mutation rates are $\frac{\theta_A}{2}$ and $\frac{\theta_B}{2}$, respectively. In the case of a finite-alleles model, we define K and L as the number of possible allele types at loci A and B, respectively, while for an infinite-alleles model, K and L correspond to the number of distinct allele types observed at the two loci. The population-scaled recombination rate is denoted by $\frac{\rho}{2}$.

Going backwards in time, a recombination event breaks up a haplotype into two fragments. Therefore, when we think about a sample's genealogical history and try to obtain a closed system of recursion for the sampling distribution, the two-locus type space must be extended to allow some haplotypes to be specified at only one of the two loci.

Definition 10.1 (Two-locus sample configuration). The two-locus sample configuration is denoted by $\boldsymbol{n} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$, where

- $\boldsymbol{a} = (a_i)_{i\in[K]}$, with a_i being the number of haplotypes with allele i at locus A and unspecified alleles at locus B,
- $\boldsymbol{b} = (b_j)_{j\in[L]}$, with b_j being the number of haplotypes with unspecified alleles at locus A and allele j at locus B,

- $\boldsymbol{c} = (c_{ij})_{i,j\in[K]\times[L]}$, with c_{ij} being the multiplicity of haplotypes with allele i at locus A and allele j at locus B.

Even if we observe both alleles of every haplotype in a sample (in the form $(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{c})$), the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are still needed when describing the sample's ancestry. An ancestral haplotype might transmit genetic material to extant haplotypes in the sample at only one of the two loci, whereupon we use $\boldsymbol{a}$ or $\boldsymbol{b}$ to avoid specifying the allele at the non-ancestral locus. Throughout, we use the following notation:

$$|\boldsymbol{a}| = \sum_{i=1}^{K} a_i, \quad c_{i\cdot} = \sum_{j=1}^{L} c_{ij}, \quad |\boldsymbol{c}| = \sum_{i=1}^{K}\sum_{j=1}^{L} c_{ij},$$
$$|\boldsymbol{b}| = \sum_{j=1}^{L} b_j, \quad c_{\cdot j} = \sum_{i=1}^{K} c_{ij}, \quad |\boldsymbol{n}| = |\boldsymbol{a}| + |\boldsymbol{b}| + |\boldsymbol{c}|.$$

10.1.1 Probability recursion for an arbitrary finite-alleles model

For fixed $a, b, c \in \mathbb{Z}_{\geq 0}$, let $p_{a,b,c}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ denote the probability of observing an *unordered* sample with configuration $\boldsymbol{n} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$, conditioned on $|\boldsymbol{a}| = a, |\boldsymbol{b}| = b$, and $|\boldsymbol{c}| = c$. The sampling distribution $p_{a,b,c}$ is normalized so that

$$\sum_{\boldsymbol{a}:|\boldsymbol{a}|=a} \sum_{\boldsymbol{b}:|\boldsymbol{b}|=b} \sum_{\boldsymbol{c}:|\boldsymbol{c}|=c} p_{a,b,c}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = 1. \tag{10.1}$$

We use $q_{a,b,c}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ to denote the sampling probability of an *ordered* sample with configuration $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$. As in the one-locus case, it is easier to work with ordered samples than unordered samples.

For a *finite-alleles* model, the distribution $q_{a,b,c}$ is related to $p_{a,b,c}$ as

$$p_{a,b,c}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \binom{|\boldsymbol{a}|}{a_1, a_2, \ldots, a_K} \binom{|\boldsymbol{b}|}{b_1, b_2, \ldots, a_L} \binom{|\boldsymbol{c}|}{c_{11}, c_{1,2}, \ldots, c_{KL}} q_{a,b,c}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}). \tag{10.2}$$

Let $\mathbf{P}^A = (P^A_{ij})$ and $\mathbf{P}^B = (P^B_{ij})$ denote the mutation transition matrices at the two loci. Then, $q_{a,b,c}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ satisfies the following recursion at stationarity:

$$\begin{aligned}
&[n(n-1) + \theta_A(a+c) + \theta_B(b+c) + \rho c] q_{a,b,c}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \\
&\sum_{i=1}^{K} a_i(a_i - 1 + 2c_{i\cdot}) q_{a-1,b,c}(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b}, \boldsymbol{c}) + \sum_{j=1}^{L} b_j(b_j - 1 + 2c_{\cdot j}) q_{a,b-1,c}(\boldsymbol{a}, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{c}) \\
&+ \sum_{i=1}^{K}\sum_{j=1}^{L} [c_{ij}(c_{ij} - 1) q_{a,b,c-1}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} - \boldsymbol{e}_{ij}) + 2a_i b_j q_{a-1,b-1,c+1}(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{c} + \boldsymbol{e}_{ij})] \\
&+ \theta_A \sum_{i=1}^{K} \left[\sum_{j=1}^{L} c_{ij} \sum_{t=1}^{K} P^A_{ti} q_{a,b,c}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} - \boldsymbol{e}_{ij} + \boldsymbol{e}_{tj}) + a_i \sum_{t=1}^{K} P^A_{ti} q_{a,b,c}(\boldsymbol{a} - \boldsymbol{e}_i + \boldsymbol{e}_t, \boldsymbol{b}, \boldsymbol{c}) \right]
\end{aligned}$$

$$+\theta_B \sum_{j=1}^{L} \left[\sum_{i=1}^{K} c_{ij} \sum_{t=1}^{L} P^B_{tj} q_{a,b,c}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} - \boldsymbol{e}_{ij} + \boldsymbol{e}_{it}) + b_j \sum_{t=1}^{L} P^B_{tj} q_{a,b,c}(\boldsymbol{a}, \boldsymbol{b} - \boldsymbol{e}_j + \boldsymbol{e}_t, \boldsymbol{c}) \right]$$
$$+\rho \sum_{i=1}^{K} \sum_{j=1}^{L} c_{ij} q_{a+1,b+1,c-1}(\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{b} + \boldsymbol{e}_j, \boldsymbol{c} - \boldsymbol{e}_{ij}), \tag{10.3}$$

with boundary conditions $q_{1,0,0}(\boldsymbol{e}_i, \boldsymbol{0}, \boldsymbol{0}) = \pi^A_i$ for all $i \in [K]$, $q_{0,1,0}(\boldsymbol{0}, \boldsymbol{e}_j, \boldsymbol{0}) = \pi^B_j$ for all $j \in [L]\}$, and $q_{0,0,1}(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{e}_{ij}) = \pi^A_i \pi^B_j$ for all $(i, j) \in [K] \times [L]$, where $(\pi^A_i)_{i\in[K]}$ and $(\pi^B_j)_{j\in[L]}$ are stationary distributions corresponding to $\mathbf{P}^A$ and $\mathbf{P}^B$, respectively. As we have seen for one-locus sampling distributions, one can derive the above recursion by considering the probabilities of the first event back in time in the coalescent with recombination. Alternatively, the recursion can be obtained from the Wright-Fisher diffusion process dual to the coalescent (Ethier and Griffiths, 1990; Griffiths and Tavaré, 1994).

10.1.2 Probability recursion for the infinite-alleles model

Suppose that each locus evolves according to the infinite-alleles model. Golding (1984) first considered generalizing the infinite-alleles model to include recombination, and Ethier and Griffiths (1990) later undertook a more mathematical analysis of the model and provided several interesting theoretical results. Given a two-locus sample configuration $\boldsymbol{n} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ with $|\boldsymbol{a}| = a, |\boldsymbol{b}| = b, |\boldsymbol{c}| = c$, the ordered sampling probability $q_{a,b,c}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ under the infinite-alleles model satisfies the following recursion at stationarity (Ethier and Griffiths, 1990; Golding, 1984):

$$[n(n-1) + \theta_A(a+c) + \theta_B(b+c) + \rho c] q_{a,b,c}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) =$$
$$\sum_{i=1}^{K} a_i(a_i - 1 + 2c_{i\cdot}) q_{a-1,b,c}(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b}, \boldsymbol{c}) + \sum_{j=1}^{L} b_j(b_j - 1 + 2c_{\cdot j}) q_{a,b-1,c}(\boldsymbol{a}, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{c})$$
$$+\sum_{i=1}^{K} \sum_{j=1}^{L} [c_{ij}(c_{ij} - 1) q_{a,b,c-1}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} - \boldsymbol{e}_{ij}) + 2a_i b_j q_{a-1,b-1,c+1}(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{c} + \boldsymbol{e}_{ij})]$$
$$+\theta_A \sum_{i=1}^{K} \left[\sum_{j=1}^{L} \delta_{a_i + c_{i\cdot},1} \delta_{c_{ij},1} q_{a,b+1,c-1}(\boldsymbol{a}, \boldsymbol{b} + \boldsymbol{e}_j, \boldsymbol{c} - \boldsymbol{e}_{ij}) + \delta_{a_i,1} \delta_{c_{i\cdot},0} q_{a-1,b,c}(\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b}, \boldsymbol{c}) \right]$$
$$+\theta_B \sum_{j=1}^{L} \left[\sum_{i=1}^{K} \delta_{b_j + c_{\cdot j},1} \delta_{c_{ij},1} q_{a+1,b,c-1}(\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{b}, \boldsymbol{c} - \boldsymbol{e}_{ij}) + \delta_{b_j,1} \delta_{c_{\cdot j},0} q_{a,b-1,c}(\boldsymbol{a}, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{c}) \right]$$
$$+\rho \sum_{i=1}^{K} \sum_{j=1}^{L} c_{ij} q_{a+1,b+1,c-1}(\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{b} + \boldsymbol{e}_j, \boldsymbol{c} - \boldsymbol{e}_{ij}), \tag{10.4}$$

with boundary conditions $q_{1,0,0}(\boldsymbol{e}_i, \boldsymbol{0}, \boldsymbol{0}) = q_{0,1,0}(\boldsymbol{0}, \boldsymbol{e}_j, \boldsymbol{0}) = q_{0,0,1}(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{e}_{ij}) = 1$ for all $i \in [K]$ and $j \in [L]$. We define $q_{a,b,c}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = 0$ whenever any entry in $\boldsymbol{a}$, $\boldsymbol{b}$, or $\boldsymbol{c}$ is negative. For notational convenience, we deviate from Ethier and Griffiths (1990) and allow each summation to range over all allelic types.

Remark 10.2. Note that the mutation terms in (10.4) are simpler than that in (10.3). Specifically, as in the one-locus coalescent with killing (cf., Chapter 4.3), a lineage can be treated as being lost when mutation occurs in the infinite-alleles model. This fact facilitates computation substantially when solving the recursion numerically. The same simplification occurs in (10.3) under a parent-independent mutation model.

The ordered and the unordered sampling distributions under the infinite-alleles model are related as

$$p_{a,b,c}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) = \frac{1}{\sigma(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})}\binom{|\boldsymbol{a}|}{a_1,a_2,\ldots,a_K}\binom{|\boldsymbol{b}|}{b_1,b_2,\ldots,a_L}\binom{|\boldsymbol{c}|}{c_{11},c_{1,2},\ldots,c_{KL}} q_{a,b,c}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}),$$

where $\sigma(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$ denotes the number of pairs of permutations $(\alpha,\beta) \in \mathscr{S}_K \times \mathscr{S}_L$ of the allele labels that leave the sample configuration $(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$ invariant.

10.2 Asymptotic sampling distributions

Unfortunately there are no known closed-form solutions to (10.3) or (10.4). These recursions can be solved numerically, but computation quickly becomes intractable with growing sample size. A more scalable approach is to employ importance sampling (Fearnhead and Donnelly, 2001; Griffiths et al, 2008), generalizing the one-locus case discussed in Chapter 6.4. For large recombination rates, the genealogies sampled by Monte Carlo methods are typically very complicated, containing many recombination events. In contrast to this increased complexity in the coalescent, however, we in fact expect the dynamics to be easier to study for large recombination rates, since the loci under consideration would then be less dependent. That is, there may exist a simpler stochastic process that provides an effective description of the dynamics for large recombination rates.

10.2.1 Two loci

Motivated by the above intuition, Jenkins and Song (2009, 2010, 2012) developed a new approach to computing the two-locus sampling probability when the recombination rate is moderate to large. Specifically, for ρ large, they proposed to find an asymptotic expansion of the form

$$q(\boldsymbol{n}) = Q_0(\boldsymbol{n}) + \frac{1}{\rho}Q_1(\boldsymbol{n}) + \frac{1}{\rho^2}Q_2(\boldsymbol{n}) + O\Big(\frac{1}{\rho^3}\Big), \tag{10.5}$$

where $Q_k(\boldsymbol{n})$ are coefficients independent of ρ. (For ease of notation, we henceforth write q instead of $q_{a,b,c\cdot}$.) Using the above perspective, it turns out that it is possible to obtain useful analytic results. We first need to define some notation before presenting these results.

Definition 10.3 (Marginal sampling configuration). Let $\boldsymbol{c}_A = (c_{i\cdot})$ and $\boldsymbol{c}_B = (c_{\cdot j})$ denote the marginal sample configurations of $\boldsymbol{c}$ restricted to loci A and B, respectively.

Note the distinction between the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, which represent haplotypes with alleles specified at only one of the two loci, and the vectors $\boldsymbol{c}_A$ and $\boldsymbol{c}_B$, which represent the one-locus marginal configurations of haplotypes with both alleles observed.

Now, note that $Q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ is the exact sampling distribution when the two loci are unlinked ($\rho = \infty$):

$$\lim_{\rho \to \infty} q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = Q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = q^A(\boldsymbol{a} + \boldsymbol{c}_A) q^B(\boldsymbol{b} + \boldsymbol{c}_B), \tag{10.6}$$

where q^A and q^B are one-locus sampling distributions for loci A and B, respectively. The functional form of $Q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ in (10.6) is *universal* in the sense that it holds true to all mutation models. Surprising, Q_1 also satisfies such a universality property:

Theorem 10.4 (Jenkins and Song 2009, 2010). *For both the infinite-alleles model and an arbitrary finite-alleles model, the first-order term $Q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ in the asymptotic expansion* (10.5) *is given by*

$$\begin{aligned} Q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = & \binom{c}{2} q^A(\boldsymbol{a} + \boldsymbol{c}_A) q^B(\boldsymbol{b} + \boldsymbol{c}_B) \\ & - q^B(\boldsymbol{b} + \boldsymbol{c}_B) \sum_{i=1}^{K} \binom{c_{i\cdot}}{2} q^A(\boldsymbol{a} + \boldsymbol{c}_A - \boldsymbol{e}_i) \\ & - q^A(\boldsymbol{a} + \boldsymbol{c}_A) \sum_{j=1}^{L} \binom{c_{\cdot j}}{2} q^B(\boldsymbol{b} + \boldsymbol{c}_B - \boldsymbol{e}_j) \\ & + \sum_{i=1}^{K} \sum_{j=1}^{L} q^A(\boldsymbol{a} + \boldsymbol{c}_A - \boldsymbol{e}_i) q^B(\boldsymbol{b} + \boldsymbol{c}_B - \boldsymbol{e}_j), \end{aligned} \tag{10.7}$$

where $c = |\boldsymbol{c}|$ and $\boldsymbol{e}_i$ is a unit vector with a 1 at entry i and 0 otherwise.

Remark 10.5. The expression (10.7) for $Q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ has an interesting application. Using the definition of the asymptotic series, Jenkins and Song (2012) obtained the following result:

Theorem 10.6 (A sufficient condition for the MLE to be finite). *Suppose $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ is used to estimate ρ. If $Q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) > 0$, then the maximum likelihood estimate of ρ is finite.*

The converse is not true in general; Jenkins and Song constructed an explicit counterexample.

We now outline an overall strategy for proving Theorem 10.4; the interested reader is referred to Jenkins and Song (2009, 2010) for further details. As shown in (10.3) and (10.4), the full sampling distribution $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ satisfies a recursion relation. Plugging in the proposed expansion (10.5) into the recursion and matching the coefficients of $\frac{1}{\rho}$, one can show that $Q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ satisfies

$$Q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = f_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) + \sum_{i=1}^{K} \sum_{j=1}^{L} \frac{c_{ij}}{c} Q_1(\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{b} + \boldsymbol{e}_j, \boldsymbol{c} - \boldsymbol{e}_{ij}), \tag{10.8}$$

where $f_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ is a function of the zeroth-order term Q_0, which depends on the marginal one-locus sampling distributions q^A and q^B. Repeatedly applying (10.8), one can show that it admits a probabilistic interpretation:

$$Q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \sum_{m=1}^{c} \mathbb{E}[f_1(\boldsymbol{A}^{(m)}, \boldsymbol{B}^{(m)}, \boldsymbol{C}^{(m)})], \tag{10.9}$$

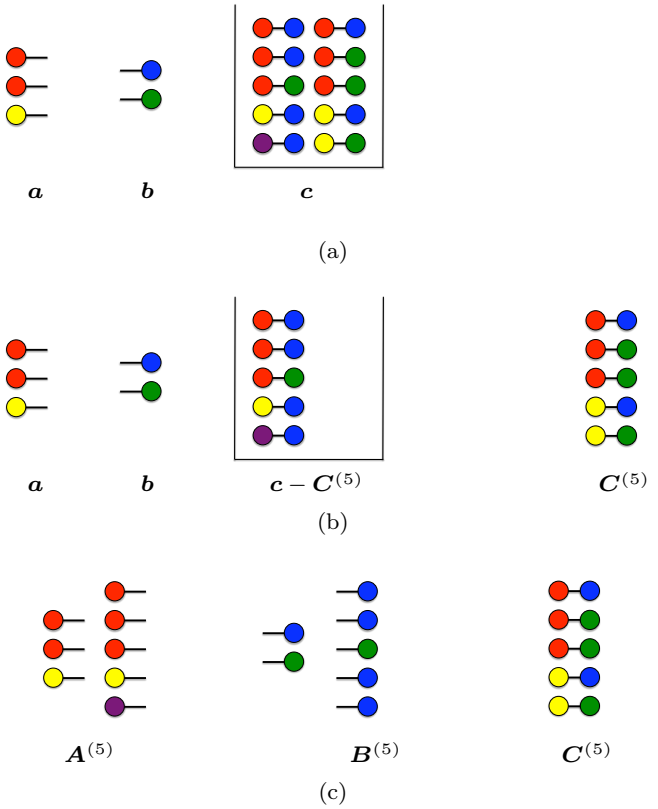


Fig. 10.1: A random sampling procedure leading to random vectors $\boldsymbol{A}^{(k)}$ and $\boldsymbol{B}^{(k)}$ specifying left and right half fragments, respectively, and a random matrix $\boldsymbol{C}^{(k)}$ specifying full haplotypes. (a) A sample $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$, with full haplotypes in an urn. (b) k haplotypes are randomly drawn from the urn and this random sample is denoted by $\boldsymbol{C}^{(k)}$. In this example, $k = 5$. (c) Break apart every haplotype remaining in the urn, and then add the half fragments to $\boldsymbol{a}$ and $\boldsymbol{b}$ appropriately.

where $\boldsymbol{C}^{(m)} = (C_{ij}^{(m)})$ is a multivariate hypergeometric$(|\boldsymbol{c}|, \boldsymbol{c}, m)$ random variable, and $\boldsymbol{A}^{(m)}$ and $\boldsymbol{B}^{(m)}$ depend on $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$, and $\boldsymbol{C}^{(m)}$ as described in Figure 10.1. The expectation in (10.9) is with respective to the random sampling procedure described in the figure, and this expectation can be evaluated analytically, leading to (10.7).

More generally, the Mth order term Q_M can be written as

$$Q_M(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = Q_M(\boldsymbol{a} + \boldsymbol{c}_A, \boldsymbol{b} + \boldsymbol{c}_B, \boldsymbol{0}) + \sum_{k=1}^{|\boldsymbol{c}|} \mathbb{E}[f_M(\boldsymbol{A}^{(k)}, \boldsymbol{B}^{(k)}, \boldsymbol{C}^{(k)})], \tag{10.10}$$

where f_M is a degree-$2M$ polynomial in the entries of the matrix $\boldsymbol{C}^{(k)}$, and the random vectors $\boldsymbol{A}^{(k)}, \boldsymbol{B}^{(k)}$ and the random matrix $\boldsymbol{C}^{(k)}$ are defined as above. Finding f_M and evaluating the expectation is cumbersome. For $M = 2$, Jenkins and Song (2009, 2010) found a closed-form formula for $\sum_{k=1}^{|\boldsymbol{c}|} \mathbb{E}[f_M(\boldsymbol{A}^{(k)}, \boldsymbol{B}^{(k)}, \boldsymbol{C}^{(k)})]$ for both an arbitrary finite-

alleles model and the infinite-alleles model. The first term $Q_M(\boldsymbol{a}+\boldsymbol{c}_A, \boldsymbol{b}+\boldsymbol{c}_B, \boldsymbol{0})$ in (10.10) is either zero or negligibly small. When it is non-zero, it can be computed using dynamic programming, but its burden in computational time increases with sample size.

The above results raise a few follow-up questions.

1. Is it possible to compute the higher-order coefficients $q_M(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ for $M > 2$?
2. For a given finite $\rho > 0$, does the series converge as more terms are added?
3. If not, how should one make use of the expansion in practice?
4. Can we incorporate other important mechanisms of evolution such as natural selection?

To answer the first question, Jenkins and Song (2012) developed a new approach based on the diffusion process; the key idea is to utilize the diffusion generator to organize computation in a simple, transparent fashion. The same basic procedure, which is purely algebraic, applies to all orders and the computation can be completely automated. It turns out that the asymptotic series (10.5) diverges in general, but Jenkins and Song (2012) came up with a solution to this problem. More precisely, they proposed to approximate $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ by a rational function known as the Padé approximant:

$$q_{\text{Padé}}^{[U/V]}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \frac{A_0 + A_1\big(\frac{1}{\rho}\big) + \cdots + A_U\big(\frac{1}{\rho}\big)^U}{1 + B_1\big(\frac{1}{\rho}\big) + \cdots + B_V\big(\frac{1}{\rho}\big)^V}.$$

The coefficients are chosen so that the first $U + V + 1$ terms in a Maclaurin series of $q_{\text{Padé}}^{[U/V]}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ agree with the first $M + 1$ terms in a Maclaurin series of $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$, which is just our partial sum:

$$q^{(M)}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = Q_0(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) + \frac{Q_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}{\rho} + \cdots + \frac{Q_M(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})}{\rho^M},$$

where $M = U + V$. The Padé approximant is a natural method to employ because of the following result:

Proposition 10.7 (Characterization of the exact sampling distribution, Jenkins and Song 2012). *The exact sampling distribution $q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ is a rational function of $1/\rho$, and the degree of the numerator is equal to the degree of the denominator:*

$$q(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \frac{\alpha_0 + \alpha_1\big(\frac{1}{\rho}\big) + \cdots + \alpha_d\big(\frac{1}{\rho}\big)^d}{1 + \beta_1\big(\frac{1}{\rho}\big) + \cdots + \beta_d\big(\frac{1}{\rho}\big)^d}.$$

Finally, for any given sample configuration $\boldsymbol{n} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$, the following theorem guarantees that the *exact* two-locus sampling distribution can be obtained as an analytic function of ρ via the method of the Padé approximant:

Theorem 10.8 (Convergence to the exact distribution, Jenkins and Song 2012). *For every two-locus sample configuration $\boldsymbol{n} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$, there exists a positive integer $C(\boldsymbol{n})$ such that for all $U \geq C(\boldsymbol{n})$ and $V \geq C(\boldsymbol{n})$, the Padé approximant $q_{\text{Padé}}^{[U/V]}(\boldsymbol{n})$ is exactly equal to $q(\boldsymbol{n})$ for all $\rho \in [0, \infty)$.*

Now suppose the alleles at locus A have different fitnesses. The effect of natural selection on genealogies is complicated (cf., the ancestral selection graph (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997)), while the effect on the diffusion process is relatively simple.

Jenkins and Song (2012) showed that their method extends naturally to incorporate weak selection at one locus. ("Weak" selection here means selection well-modeled by a diffusion process.) In particular, the universality property of Q_0 and Q_1 still holds in the presence of selection. More precisely, Q_0 and Q_1 are given by (10.6) and (10.7), respectively, with q^A corresponding the one-locus sampling distribution under selection.

We now return to the motivating remark made in the beginning of this section, namely that there may exist a simpler stochastic process that describes the important dynamics of the ARG for large recombination rates. Recently, Jenkins et al (2015) actually found such a simpler genealogical process with a closed-form sampling distribution that agrees with the "truth" up to $O(1/\rho^2)$: $Q_0(\boldsymbol{n}) + Q_1(\boldsymbol{n})/\rho$. Furthermore, they were able to make a similar statement about the Wright-Fisher diffusion dual to the coalescent with recombination. Hence, both the ARG and the Wright-Fisher diffusion with recombination exhibit a deep and regular structure when the recombination rate increases, and this structure can be exploited to derive simple approximations to these models.

10.2.2 Multiple loci

Bhaskar and Song (2012) extended some of the analytical results described above to more than two loci. More precisely, they obtained closed-form formulas for the first two terms (analogous to Q_0 and Q_1 in the two-locus case) in an asymptotic expansion of the sampling distribution for an arbitrary number of loci. In the case of L loci, there are $L-1$ possible recombination breakpoints each with its own rate. The number of possible allelic combinations grows exponentially with the number of loci, and the system of equations that needs to be solved is considerably more complex than that in the two-locus case. The authors employed combinatorial techniques (particularly the inclusion-exclusion principle) to make progress on this general case. Their work showed that the universality property of Q_0 and Q_1 previously observed in the two-locus case also applies to the case of an arbitrary number of loci. We refer the reader to Bhaskar and Song (2012) for the details.

10.3 Two-locus likelihoods under variable population size

Kamm et al (2016), Ragsdale and Gutenkunst (2017)

10.4 Application of two-locus likelihoods: fine-scale recombination rate estimation

Myers et al (2005), Chan et al (2012), Spence and Song (2019)

References

Bhaskar A, Song YS (2012) Closed-form asymptotic sampling distributions under the coalescent with recombination for an arbitrary number of loci. Advances in Applied Probability 44:391–407, (PMC3409093)

Chan AH, Jenkins PA, Song YS (2012) Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. PLoS Genetics 8(12):e1003,090

Ethier SN, Griffiths RC (1990) On the two-locus sampling distribution. J Math Biol 29:131–159

Fearnhead P, Donnelly P (2001) Estimating recombination rates from population genetic data. Genetics 159:1299–1318

Golding GB (1984) The sampling distribution of linkage disequilibrium. Genetics 108:257–274

Griffiths RC, Tavaré S (1994) Simulating probability distributions in the coalescent. Theoretical Population Biology 46:131–159

Griffiths RC, Jenkins PA, Song YS (2008) Importance sampling and the two-locus model with subdivided population structure. Advances in Applied Probability 40:473–500

Jenkins PA, Song YS (2009) Closed-form two-locus sampling distributions: accuracy and universality. Genetics 183:1087–1103

Jenkins PA, Song YS (2010) An asymptotic sampling formula for the coalescent with recombination. Annals of Applied Probability 20:1005–1028, (PMC2910927)

Jenkins PA, Song YS (2012) Padé approximants and exact two-locus sampling distributions. Annals of Applied Probability 22:576–607, (PMC3685441)

Jenkins PA, Fearnhead P, Song YS (2015) Tractable diffusion and coalescent processes for weakly correlated loci. Electronic Journal of Probability 20(58):1–26

Kamm JA, Spence JP, Chan J, Song YS (2016) Two-locus likelihoods under variable population size and fine-scale recombination rate estimation. Genetics 203:1381–1399, DOI 10.1534/genetics.115.184820

Krone SM, Neuhauser C (1997) Ancestral processes with selection. Theoretical Population Biology 51(3):210–237

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310(5746):321–324

Neuhauser C, Krone SM (1997) The genealogy of samples in models with selection. Genetics 145:519–34

Ragsdale AP, Gutenkunst RN (2017) Inferring demographic history using two-locus statistics. Genetics 206(2):1037–1048

Spence JP, Song YS (2019) Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. Science Advances 5(10):eaaw9206

Part V
Further Extensions: Multiple and Simultaneous Mergers

Chapter 11
Accuracy of the coalescent when the sample is very large

As discussed in Chapter 1.3, the coalescent provides a good approximation to the genealogical process of a discrete-time random mating model only if the population size N of the discrete-time model is sufficiently large compared to the sample size n. Since study sample sizes in population genomics are growing rapidly, it is important to examine the accuracy of the standard coalescent approximation, which underlies many commonly-used analytical tools. To this end, Bhaskar et al (2014) developed a method to perform exact computation in the discrete-time Wright-Fisher model and compared several key genealogical quantities of interest with the corresponding quantities under the coalescent. In particular, they studied the number of multiple- and simultaneous-merger events under the discrete-time WF model, which are absent in the coalescent by construction, and also examined the resulting distortion in the expected sample frequency spectrum. The goal of this chapter is to present the computational details of this work. In the subsequent chapters, we will discuss extensions of the coalescent process that allow multiple- and simultaneous-merger events.

11.1 Computing the expected number of multiple- and simultaneous-mergers

As in (1.1), let $p_{kj}^{(g)}$ denote the probability that k particular labeled individuals at generation g have j distinct ancestors at generation $g+1$. For an algorithmic reason that will become clear presently, we assume that there is a critical generation g_c such that $N_g = N$ (some constant) for all $g > g_c$. This assumption is not so restrictive since for sufficiently large g, there will be only 1 lineage left with high probability, and the genealogical properties we study will not be affected. For $g > g_c$, we drop the dependence on g in the probabilities $p_{n,m}^{(g)}$, and simply write them as $p_{n,m}$.

Let $X_{n,k}^{(g)}$ denote the number of k-mergers that occur in a random genealogical tree for a sample of n individuals from generation g. The expectation of $X_{n,k}^{(0)}$ can be computed by using the recursion

$$\mathbb{E}[X_{n,k}^{(g)}] = \binom{n}{k} \sum_{m=k+1}^{n} p_{k,1}^{(g)} p_{n-k,m-k}^{(g)} \frac{N_{g+1} - m + k}{N_{g+1}} + \sum_{m=1}^{n} p_{n,m}^{(g)} \mathbb{E}[X_{m,k}^{(g+1)}], \tag{11.1}$$

for $k < n$, with the boundary condition $\mathbb{E}[X_{k,k}^{(g)}] = p_{k,1}^{(g)} + p_{k,k}^{(g)}\mathbb{E}[X_{k,k}^{(g+1)}]$. This recursion follows from conditioning on the mergers that occur between generations g and $g+1$. If the population size remains constant (e.g., for generation $g > g_c$), we can drop the dependence on g in the notation $\mathbb{E}[X_{n,k}^{(g)}]$, and obtain the following recursion for $\mathbb{E}[X_{n,k}]$:

$$\mathbb{E}[X_{n,k}] = \binom{n}{k} \sum_{m=k+1}^{n} \frac{p_{k,1}p_{n-k,m-k}}{1-p_{n,n}} \frac{N-m+k}{N} + \sum_{m=1}^{n-1} \frac{p_{n,m}}{1-p_{n,n}} \mathbb{E}[X_{m,k}], \tag{11.2}$$

for $k < n$, with the boundary condition $\mathbb{E}[X_{k,k}] = \frac{p_{k,1}}{1-p_{k,k}}$.

One can write similar recursions for the expected number of simultaneous mergers by conditioning on the mergers that occur during each generation of reproduction.

11.2 Computing the expected SFS in a discrete-time model

For a given set of n individuals sampled at present, the ancestral process in the discrete-time WF model generates a genealogical tree with n leaves, with the individuals in the sample forming the leaves of the tree. Let $\tau_{n,b}^{\text{WF}}$ denote the total length (in number of generations) of all branches each subtending exactly b leaves. We can use dynamic programming to compute $\mathbb{E}[\tau_{n,b}^{\text{WF}}]$ efficiently as follows. Let $\gamma_{a,b}^{(g)}$ be a random variable denoting the total branch length of a subtree that subtends a particular set of a labeled individuals in a larger set of $a+b$ individuals observed at generation g. Then, the exchangeability of the individuals in the sample implies

$$\mathbb{E}[\tau_{n,b}^{\text{WF}}] = \binom{n}{b} \mathbb{E}[\gamma_{b,n-b}^{(0)}], \tag{11.3}$$

since there are $\binom{n}{b}$ subsamples of b individuals out of the n individuals in the original sample. The expected unnormalized SFS is then given by $(\mu\mathbb{E}[\tau_{n,1}^{\text{WF}}], \ldots, \mu\mathbb{E}[\tau_{n,n-1}^{\text{WF}}])$, where μ corresponds to the mutation probability per generation.

By conditioning on the mergers between lineages that take place between generations g and $g+1$, we obtain the following recursion for $\mathbb{E}[\gamma_{a,b}^{(g)}]$:

$$\mathbb{E}[\gamma_{a,b}^{(g)}] = \begin{cases} \displaystyle\sum_{j=1}^{a}\sum_{k=1}^{b} p_{a,j}^{(g)} p_{b,k}^{(g)} \frac{(N_{g+1}-j)_{k\downarrow}}{(N_{g+1})_{k\downarrow}} \mathbb{E}[\gamma_{j,k}^{(g+1)}], & \text{if } a > 1, \\ \displaystyle 1 + \sum_{m=1}^{b} \frac{N_{g+1}-m}{N_{g+1}} p_{b,m}^{(g)} \mathbb{E}[\gamma_{1,m}^{(g+1)}], & \text{if } a = 1 \text{ and } b > 1, \\ 1 + p_{2,2}^{(g)} \mathbb{E}[\gamma_{1,1}^{(g+1)}], & \text{if } a = b = 1. \end{cases} \tag{11.4}$$

If the population size remains constant (e.g., for generation $g > g_c$), we can drop the dependence on g in the notation $\gamma_{a,b}^{(g)}$, and by conditioning on the previous genealogical event, we can derive the following recursion and boundary condition for $\mathbb{E}[\gamma_{a,b}]$:

$$\mathbb{E}[\gamma_{a,b}] = \begin{cases} \sum_{j=1}^{a}\sum_{k=1}^{b}(1-\delta_{j,a}\delta_{k,b})\dfrac{p_{a,j}p_{b,k}}{1-p_{a+b,a+b}}\dfrac{(N-j)_{k\downarrow}}{(N)_{k\downarrow}}\mathbb{E}[\gamma_{j,k}], & \text{if } a>1, \\ \dfrac{1}{1-p_{b+1,b+1}} + \sum_{m=1}^{b-1}\dfrac{N-m}{N}\dfrac{p_{b,m}}{1-p_{b+1,b+1}}\mathbb{E}[\gamma_{1,m}], & \text{if } a=1 \text{ and } b>1, \\ N, & \text{if } a=b=1. \end{cases} \tag{11.5}$$

From recursions (11.4) and (11.5), the expected unnormalized SFS for a sample of size n can be computed in $O(n^4)$ and $O(n^4 g_c)$ time for the constant and variable population cases, respectively. However, if one truncates the summation range for the indices j and k in (11.4) and (11.5) to only those j,k values where $p_{a,j}^{(g)}$ and $p_{b,k}^{(g)}$ (respectively, $p_{a,j}$ and $p_{b,k}$) are greater than some small tolerance parameter $\varepsilon > 0$, the time complexity of the above dynamic programs can be improved to $\widetilde{O}(n^2)$ and $\widetilde{O}(n^2 g_c)$, where the $\widetilde{O}$ notation signifies the dependence of the quantities on the truncation parameter ε.

Bhaskar et al (2014) used (11.3) along with a truncation parameter of $\varepsilon = 10^{-120}$ to compute the expected SFS.

11.3 Comparison between the discrete-time WF model and the coalescent

11.3.1 Multiple and simultaneous mergers

11.3.2 Ancestral process

11.3.3 Expected SFS

11.4 A two-phase hybrid approach

References

Bhaskar A, Clark AC, Song Y (2014) Distortion of genealogical properties when the sample is very large. Proc Nat Acad Sci 111(6):2385–2390, (PMC3926037)

Chapter 12
Λ-coalescents

Kingman's coalescent models the ancestry of a sample of genes in a randomly-mating, selectively neutral, constant-sized population (Kingman, 1982a,b,c). We have seen that many of the constraints of the model can be relaxed in turn, and here we consider generalizing the coalescent process to allow two or more lineages to merge in a single coalescence event. A motivation for this is, for example, to model an organism in which the variance in offspring numbers is large (Eldon and Wakeley, 2006; Möhle and Sagitov, 2001), a condition we make precise below. Other relevant contexts include models of recurrent selective sweeps (Durrett and Schweinsberg, 2004, 2005) and populations undergoing continuous strong selection (Neher and Hallatschek, 2013; Schweinsberg, 2015). The resulting coalescent process is known as the coalescent with *multiple mergers*, or the *Λ-coalescent*, and was introduced by Pitman (1999) and Sagitov (1999). Recent reviews are in Berestycki (2009) and Birkner and Blath (2009). For now, we will still prohibit *simultaneous* mergers, which result in an even more general class of coalescent processes (Möhle and Sagitov, 2001; Schweinsberg, 2000).

12.1 Characterizing a consistent collection of multiple-merger rates

We first recall some facts about Kingman's coalescent. It is a unique time-homogeneous Markov process $\{C_\infty(t), t \geq 0\}$ on the space of partitions of $\mathbb{N}$ with the following properties, the first two of which are sufficient to ensure uniqueness (Kingman, 1982a):

1. $C_\infty(0) = \{\{1\}, \{2\}, \ldots\}$, the partition consisting of only singletons. (Coalescents with this initial condition are referred to as *standard*).
2. For each n, the restriction $(C_n(t), t \geq 0)$ of $(C_\infty(t), t \geq 0)$ to $[n]$ is a Markov chain with càdlàg paths and transition rates specified by saying that while there are b blocks in the partition, each of the $\binom{b}{2}$ pairs of blocks is merging independently at rate one. The process $(C_n(t), t \geq 0)$ is referred to as an *n-coalescent*.
3. For each $t \geq 0$, $C_\infty(t)$ is an exchangeable random partition of $\mathbb{N}$.
4. $C_\infty(t)$ is consistent in a temporal sense: if $C_n(s) = \{B_1, \ldots, B_m\}$, where $B_i \subseteq [n]$, then the subsequent process on the indices of these blocks is an m-coalescent, for all $s \geq 0$ and all $n \in \mathbb{N}$.
5. $C_\infty(t)$ is consistent in a spatial sense: for all $n \in \mathbb{N}$, the restriction of $\{C_n(t), t \geq 0\}$ to m, where $m < n$, is an m-coalescent.

Notice that the rates of merging do not depend on the sizes of the blocks involved. We seek to generalize these merger rates beyond just binary mergers.

Definition 12.1. While there are $|C_n(t)| = b$ blocks, let $\lambda_{b,k}$ denote the rate at which each k-tuple of blocks of $C_n(t)$ is merging to form a single block.

In Kingman's coalescent, we have $\lambda_{b,k} = \delta_{k,2}$ (the Kronecker delta). How can we define the collection $\{\lambda_{b,k}, 2 \le k \le b \le \infty\}$ in a consistent way? For example, we could not have $\lambda_{3,3} = 1$ and $\lambda_{2,2} = 0$, since the restriction of a process with three blocks to two should still have the possibility of a two-merger. Obviously we need $\lambda_{b,k} \ge 0$, and a moment's thought tells us we also require that

$$\lambda_{b,k} = \lambda_{b+1,k} + \lambda_{b+1,k+1}, \tag{12.1}$$

for each $2 \le k \le b$. We see this by the following argument: We have b blocks, in which a particular k-tuple is merging at rate $\lambda_{b,k}$. Now *reveal* a $(b+1)$th block. Either it was also secretly participating in this merger (rate $\lambda_{b+1,k+1}$) or it was not (rate $\lambda_{b+1,k}$), and so the sum of these two possibilities must equal the original rate. It turns out that the rates $(\lambda_{b,k})$ satisfying (12.1) can be completely characterized (Pitman, 1999). We now detail the argument, which is closely related to de Finetti's theorem:

Theorem 12.2 (De Finetti). *Let $(Z_1, Z_2, \ldots)$ be an infinite sequence of exchangeable Bernoulli random variables. Then there exists a probability distribution F on $[0,1]$ such that*

$$p_{n,k} := \mathbb{P}(Z_1 = 1, \ldots, Z_k = 1, Z_{k+1} = 0, \ldots, Z_n = 0) = \mathbb{E}[X^k(1-X)^{n-k}],$$

where X is distributed according to F. That is, conditionally given $X = x$, the Z_i are i.i.d. Bernoulli(x) *random variables, and we must have*

$$\frac{Z_1 + Z_2 + \cdots + Z_n}{n} \to X \sim F$$

almost surely as $n \to \infty$.

By exchangeability, $p_{n,k}$ is equal to the probability of any sequence $(Z_1, \ldots, Z_n)$ with exactly k ones and $n-k$ zeros. Using a similar consistency argument as that we used for (12.1), we conclude

$$p_{n-1,k} = p_{n,k} + p_{n,k+1}, \tag{12.2}$$

for $k = 0, \ldots, n-1$. To prove de Finetti's theorem, we will utilize the following result:

Theorem 12.3 (Hausdorff moment problem). *The sequence $(c_0, c_1, \ldots)$ is a sequence of moments of a probability distribution on $[0,1]$* (i.e., $c_n = \mathbb{E}[X^n]$ for all $n \in \mathbb{N}_0$), if and only if $c_0 = 1$ and it is *completely monotonic*:

$$(-1)^k \Delta^k c_n \ge 0,$$

for all $n, k \ge 0$, where Δ denotes the difference operator defined as $\Delta c_n := c_{n+1} - c_n$.

We now define $c_0 = 1$ and $c_n = p_{n,n}$, and show that the sequence (c_i) satisfies the necessary and sufficient condition in Theorem 12.3. The trick is to express the $p_{n,k}$ in terms of the (c_i). Setting $k = n-1$ in (12.2) and rearranging terms, we obtain

$$p_{n,n-1} = p_{n-1,n-1} - p_{n,n} = c_{n-1} - c_n = -\Delta c_{n-1}.$$

Similarly, setting $k = n - 2$ in (12.2), we obtain

$$p_{n,n-2} = p_{n-1,n-2} - p_{n,n-1} = -\Delta c_{n-2} + \Delta c_{n-1} = \Delta^2 c_{n-2},$$

and continuing by induction we get

$$p_{n,k} = p_{n-1,k} - p_{n,k+1} = (-1)^{n-k} \Delta^{n-k} c_k \geq 0. \tag{12.3}$$

The right-hand inequality holds because $p_{n,k}$ is a probability. Moreover, it tells us that (by definition) the sequence (c_n) is *completely monotonic*. Hence, by Theorem 12.3, we conclude that there exists a random variable X distributed according to some probability measure on $[0, 1]$ such that $c_n = \mathbb{E}[X^n]$, for all $n \in \mathbb{N}_0$. Furthermore, since

$$\Delta^m c_k = \sum_{j=0}^{m} \binom{m}{j} (-1)^{j+m} c_{k+j},$$

$p_{n,k}$ can be written as $\sum_{j=0}^{n-k} \binom{n-k}{j} (-1)^j c_{k+j}$, from which we conclude

$$p_{n,k} = \mathbb{E}[X^k (1 - X)^{n-k}].$$

We have essentially just proved de Finetti's theorem (this is the approach of Feller 1966, Chapter VII.4), but the reason for going through the details is that we notice the similarity between the relationship (12.2) for $p_{n,k}$ and (12.1) for $\lambda_{b,k}$. Can we apply de Finetti's theorem to $(\lambda_{b,k})$? There are two issues before we can: First, the indices for $\lambda_{b,k}$ start from 2 rather than from 0. Second, the $\lambda_{b,k}$ are rates rather than probabilities. We deal with these by defining $u_{n,j} = \lambda_{n+2,j+2}/\lambda_{2,2}$. Now, applying de Finetti's theorem to the $(u_{n,j})$, we conclude that there exists a probability distribution F such that

$$u_{n,j} = \int_0^1 x^j (1 - x)^{n-j} F(dx),$$

and therefore

$$\lambda_{b,k} = \int_0^1 x^k (1 - x)^{b-k} \frac{\Lambda(dx)}{x^2}, \tag{12.4}$$

where $\Lambda = \lambda_{2,2} F$ is a finite (but not necessarily probability) measure on $[0, 1]$. Since F is a probability measure, note that $\Lambda([0, 1]) = \lambda_{2,2}$. This remarkable result provides a correspondence between consistent collections of coalescence rates and finite measures on $[0, 1]$. The notation explains the nomenclature "Λ-coalescent". Moreover, using Kolmogorov's extension theorem in the same way that it was used to construct Kingman's coalescent, we can get away with speaking of "the" Λ-coalescent, a unique partition-valued Markov process $(C_\infty^\Lambda(t), t \geq 0)$ whose restriction to $[n]$ is an n-Λ-coalescent, for each n.

Example 12.4. We consider a few concrete examples.

1. $\Lambda = \delta_0$, a unit mass at 0, recovers Kingman's coalescent.
2. $\Lambda = \text{Beta}(2-\alpha, \alpha)$, a one-parameter subfamily of beta distributions, where $\alpha \in (0, 2)$. This is known as the *beta* coalescent. It arises as the limit from a finite Cannings model (see below) in which the tails of the offspring distributions decay like $\mathbb{P}(\nu_1 > Nx) \sim Cx^{-\alpha}$, when $\alpha \geq 1$.

3. The special case $\alpha = 1$ gives $\Lambda = \text{Uniform}[0, 1]$. This is known as the *Bolthausen-Sznitman* coalescent.
4. $\Lambda = \delta_1$, a unit mass at 1. This is a curious degenerate case in which $\lambda_{b,b} = 1$ and $\lambda_{b,k}$ is zero for any other k. Thus, the coalescent process waits an exponential(1) amount of time before *all* blocks coalesce. Drawing the resulting phylogeny explains the name: the *star-shaped* coalescent.

12.2 Poisson point process construction

The relationship (12.4) can be well understood by constructing the Λ-coalescent from a Poisson point process on $(0, 1] \times (0, \infty)$ with intensity measure $\Lambda(dx)x^{-2} \otimes dt$ (Pitman, 1999). (We see that we are in trouble if Λ has mass at zero, so for now we exclude this possibility.) It is straightforward to see that we can recover a realization of the Λ-coalescent if, at a point (x, t) of the point process, we independently include each block in the merger by coin toss, with probability x. The term $x^k(1-x)^{b-k}$ in (12.4) is then seen to be the probability of the outcomes of these b Bernoulli trials. The correspondence between this point process and the Λ-coalescent is illustrated in Figure 12.1. In fact, one could even incorporate the coin tosses into the definition of the point process by instead defining it on the space $\{0, 1\}^\infty \times (0, \infty)$ with intensity measure $L(d\boldsymbol{\chi}) \otimes dt$, where $L(d\boldsymbol{\chi}) = \int_0^1 \Lambda(dx)x^{-2}P_x$. Here, P_x determines the outcomes of $\boldsymbol{\chi} = (\chi_1, \chi_2, \ldots) \in \{0, 1\}^\infty$, an infinite sequence of independent Bernoulli trials such that $P_x(\chi_i = 1) = x$ for each i (Pitman, 1999).

For $\Lambda(\{0\}) = 0$, it is useful to think of $\Lambda(dx)/x^2$ as the rate at which a fraction x of all blocks coalesce (Berestycki, 2009). However, there is a slight subtlety here: this is only true if there are an infinite number of blocks. Pitman's solution is to make precise statements

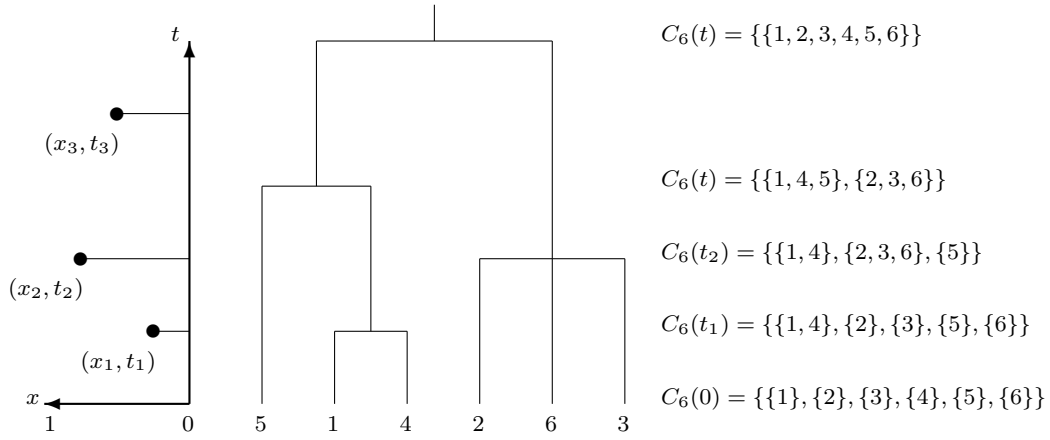


Fig. 12.1: A realization of a Λ-coalescent, together with a realization of a Poisson point process (with an unmarked, independent Kingman component) that gives rise to it. The Kingman component means that binary mergers may (as in $\{1\}$ and $\{4\}$ merging) or may not (as in $\{1, 4\}$ and $\{5\}$ merging) correspond to points of the point process. Also note that a point – e.g., see (x_3, t_3) – in the Poisson Point Process may not lead to any merger event.

like this only for Λ-coalescents which stay infinite (Pitman, 1999). Otherwise, we are tacitly assuming that the blocks reside in an otherwise unobserved infinite population of blocks. This is of course the natural way of thinking of them in a biological setting.

To incorporate mass at zero, we can simply write $\Lambda = \rho\delta_0 + \hat{\Lambda}$ where $\hat{\Lambda}$ has no mass at zero. We then apply the Poisson point process construction to $\hat{\Lambda}$ and say that pairs of blocks also independently merge at rate ρ. A more satisfying construction can be found in Schweinsberg (2000).

12.3 Interpretation of the measure Λ

The Poisson point process construction (for $\Lambda(\{0\}) = 0$) gives an intuitive interpretation for $\Lambda(dx)x^{-2}$. An interpretation of $\Lambda(dx)$ itself is as follows. Let $t_{i,j}$, which are equal in distribution to $t_{1,2}$, be the coalescence time for integers i and j; that is, i and j reside in the same block of $C_\infty(t_{i,j})$ and not of $C_\infty(t_{i,j}-)$. Notice that $t_{i,j} \sim \text{Exp}(\lambda_{2,2}) = \text{Exp}(\Lambda([0,1]))$. At a point (x,t) of the point process, i and j merge with probability $\mathbb{P}(\chi_i = \chi_j = 1) = x^2$. Thus, thinning the point process so that each point is retained independently with probability x^2 results (by the standard theory of Poisson processes) in another Poisson point process, this time with intensity $\Lambda(dx)dt$. Thus, two *particular* integers coalesce at rate $\Lambda([0,1])$, and $\Lambda(dx)/\Lambda([0,1])$ gives the distribution of the fraction of blocks which participate in their merger (again, the merger is understood to occur in an infinite population of blocks).

12.4 When can we apply a Λ-coalescent to biology?

We need to address two issues if the Λ-coalescent is to be an appropriate model for the gene genealogy of some organism—first, does it come down from infinity? And second, when does the (time-rescaled) limit of the finite-population dynamics converge to a Λ-coalescent as the population size $N \to \infty$?

12.4.1 Coming down from infinity

One should expect that the entire population of an organism has a finite time to the most recent common ancestor. In our notation, for a standard coalescent this is defined as

$$T_{\text{MRCA}} := \inf\{t \geq 0 : |C_\infty(t)| = 1\}.$$

A necessary and sufficient condition for $\mathbb{E}[T_{\text{MRCA}}] < \infty$ is that the coalescent *comes down from infinity*, defined as follows.

Definition 12.5. The coalescent is said to *come down from infinity* if for all $t > 0$, $|C_\infty(t)| < \infty$ almost surely. It is said to *stay infinite* if for all $t \geq 0$, $|C_\infty(t)| = \infty$.

Notice that we opt for a stronger definition than simply "there exists" some time for which the number of blocks is finite. This is because, with the sole exception of the star-shaped

coalescent (or indeed any coalescent with a star-shaped component, $\Lambda(\{1\}) > 0$), a coalescent either comes down from infinity almost surely or stays infinite almost surely (Pitman, 1999). The star-shaped coalescent with all its mass at 1 neither comes down from infinity (in this stronger sense) nor does it stay infinite.

As discussed in Section 2.11, Kingman's coalescent comes down from infinity. The Bolthausen-Sznitman coalescent stays infinite, but only just: for the family of Beta$(2-\alpha, \alpha)$-coalescents, $\alpha \in (0, 2)$, of which the Bolthausen-Sznitman coalescent is a member ($\alpha = 1$), the process comes down from infinity if $\alpha < 1$ and stays infinite if $\alpha \geq 1$. There exist a number of conditions on Λ related to whether the coalescent comes down from infinity (Pitman, 1999, and references therein); we content ourselves with stating a concise necessary and sufficient condition:

Theorem 12.6. *The Λ-coalescent comes down from infinity if and only if*

$$\sum_{b=2}^{\infty} \left[\sum_{k=2}^{b} (k-1) \binom{b}{k} \lambda_{b,k} \right]^{-1} < \infty.$$

12.4.2 Convergence to the coalescent

When does the genealogical process of a finite population converge to a Λ-coalescent as the population size $N \to \infty$? We will state necessary and sufficient conditions for convergence for Cannings exchangeable models, defined in Section 2.8. Recall the *offspring vector* $(\nu_1, \ldots, \nu_N)$, where ν_i denotes the number of offspring of individual i. As discussed in Section 2.10, the crucial parameter for determining convergence to a coalescent process is the probability that two individuals shared a common ancestor in the previous generation (c.f., Definition 2.22):

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

where the first equality follows by summing over the possible parent labels for the parent of the two coalescing individuals and taking expectation over $(\nu_1, \ldots, \nu_N)$, and the second equality follows from exchangeability. Sagitov (1999) (see also Möhle and Sagitov 2001) established the following necessary and sufficient conditions for convergence to a Λ-coalescent:

Theorem 12.7 (Sagitov 1999). *Let $q_{\alpha\beta}^{(N)}$ denote the transition rate from partition α to partition β in the genealogical process of a Cannings model with population size N. The rates $q_{\alpha\beta}^{(N)}$ converge to those of a Λ-coalescent on a timescale of c_N if and only if*

1. $c_N \to 0$,
2. $\dfrac{1}{c_N} \dfrac{\mathbb{E}[\nu_1(\nu_1 - 1)\nu_2(\nu_2 - 1)]}{N^2} \to 0$, *and*
3. $\dfrac{N\mathbb{P}(\nu_1 > Nx)}{c_N} \to \displaystyle\int_{(x,1]} \frac{F(dy)}{y^2}$, *at all points x with $F(\{x\}) = 0$, where F is a probability distribution on $[0, 1]$,*

as $N \to \infty$.

Each condition has a natural interpretation. The first tells us that the rate of coalescence must go to zero if we are to get a continuous-time approximation in the limit. The second tells us that the rate of *simultaneous* multiple mergers, compared to that of binary mergers, must go to zero, and the third provides an explicit relationship between the distribution F and the distribution of offspring of an individual when they replace a substantial ($x > 0$) fraction of the population. In the Kingman case, condition 2 is analogous to $\mathbb{E}[\nu_1(\nu_1-1)(\nu_1-2)]/(N^2 c_N) \to 0$, which tells us that the rate of 3-mergers relative to binary mergers goes to zero in the limit.

This convergence result is in fact much more general than the original conditions of Kingman (1982a), who required that $\mathbb{E}[\nu_1(\nu_1-1)]$ tended to a positive constant. This condition does not include, for example, the Moran model, for which $c_N = 2/[N(N-1)] \to 0$.

Example 12.8. Condition 3 of Theorem 12.7 shows us when a Λ-coalescent might be applicable: when the distribution of offspring is highly skewed, such that there is a positive probability that the offspring of a single individual might replace a substantial fraction of the population in a single generation. One such reproductive model that fits this category is those species that undergo broadcast spawning. A nice discussion is in Eldon and Wakeley (2006), who suggest that Λ-coalescents might be suitable for the Pacific oyster, the American lobster, and the Atlantic cod. They test the fit of a model in which $F = a\delta_0 + (1-a)\delta_\psi$, $a \in [0,1]$, a mixture of a Kingman coalescent and rare events in which a single individual replaces a fraction ψ of the population.

References

Berestycki N (2009) Recent progress in coalescent theory. Ensaios Matematicos 16(1):1–193

Birkner M, Blath J (2009) Measure-valued diffusions, general coalescents and population genetic inference. In: Blath J, Mörters P, Scheutzow M (eds) Trends in Stochastic Analysis, Cambridge University Press, pp 329–363

Durrett R, Schweinsberg J (2004) Approximating selective sweeps. Theoretical Population Biology 66:129–138

Durrett R, Schweinsberg J (2005) A coalescent model for the effect of advantageous mutations on the genealogy of a population. Stochastic Processes Appl 115:1628–1657

Eldon B, Wakeley J (2006) Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics 172:2621–2633

Feller W (1966) An introduction to probability theory and its applications, vol 2. John Wiley, New York

Kingman JFC (1982a) The coalescent. Stoch Process Appl 13:235–248

Kingman JFC (1982b) Exchangeability and the evolution of large populations. In: Koch G, Spizzichino F (eds) Exchangeability in Probability and Statistics, North-Holland Publishing Company, pp 97–112

Kingman JFC (1982c) On the genealogy of large populations. J Appl Prob 19A:27–43

Möhle M, Sagitov S (2001) A classification of coalescent processes for haploid exchangeable population models. The Annals of Probability 29(4):1547–1562

Möhle M, Sagitov S (2001) A classification of coalescent processes for haploid exchangeable population models. The Annals of Probability 29(4):1547–1562

Neher RA, Hallatschek O (2013) Genealogies of rapidly adapting populations. Proceedings of the National Academy of Sciences 110(2):437–442

Pitman J (1999) Coalescents with multiple collisions. Annals of Probability 27:1870–1902
Sagitov S (1999) The general coalescent with asynchronous mergers of ancestral lines. Journal of Applied Probability 36(4):1116–1125
Schweinsberg J (2000) Coalescents with simultaneous multiple collisions. Electronic Journal of Probability 5:1–50
Schweinsberg J (2015) Rigorous results for a population model with selection ii: genealogy of the population. arXiv preprint arXiv:150700394

Chapter 13
The site frequency spectrum for general coalescent models

In Chapter 12, we saw that Kingman's coalescent can be viewed as a special case of Λ-coalescents (Pitman, 1999; Sagitov, 1999). Λ-coalescents can in turn be seen as special cases of a broader class of models called Ξ-coalescents (Schweinsberg, 2000), in which more than one multiple merger event can occur simultaneously. Such events can arise in certain models of selection (Huillet, 2014), models of selective sweeps (Durrett and Schweinsberg, 2005), models with repeated strong bottlenecks (Birkner et al, 2009), and certain diploid mating models (Möhle and Sagitov, 2003). In this chapter, we discuss the problem of computing the expected site frequency spectrum (SFS) for general Ξ-coalescents. The exposition here closely follows Spence et al (2016).

13.1 A brief introduction to Ξ-coalescents

Formally, time-homogeneous Ξ-coalescents are governed by a measure $\Xi(d\boldsymbol{x})$ on the set $\{(x_1, x_2, \ldots) : x_1 \geq x_2 \geq \cdots \geq 0 \text{ and } \sum_{i=1}^{\infty} x_i \leq 1\}$. Following Spence et al (2016), we consider a general time-inhomogeneous Ξ-coalescent governed by a measure of the form $\frac{\Xi(d\boldsymbol{x})}{\zeta(t)}$, where $\Xi(d\boldsymbol{x})$ is a time-independent measure and $\zeta : \mathbb{R}_{\geq 0} \to \mathbb{R}_+$ is a strictly positive function of time representing (for historical reasons) the inverse intensity. For example, for Kingman's coalescent, $\Xi(d\boldsymbol{x}) = \delta_{\mathbf{0}}(d\boldsymbol{x})$, the point mass at zero, and $\zeta(t)$ corresponds to the scaled effective population size at time t. For other models, $\zeta(t)$ does not necessarily correspond to the population size, but has an interpretation specific to the model. For example, Neher and Hallatschek (2013) showed empirically that the rate of coalescence in a model of continuous strong selection is a nonlinear function of the population size and the first two moments of the distribution of mutational effects.

We use $\mathbf{Q}$ to denote the rate matrix of the ancestral process (also known as the block-counting process) of the time-homogeneous coalescent corresponding to $\Xi(d\boldsymbol{x})$. More specifically, $\mathbf{Q}$ is a lower triangular matrix where $(\mathbf{Q})_{ij}$ is the instantaneous rate at which i unlabeled lineages merge to form j unlabeled lineages when $\zeta \equiv 1$. For example, for Kingman's coalescent,

$$(\mathbf{Q})_{ij} = \begin{cases} \binom{i}{2}, & j = i - 1, \\ -\binom{i}{2}, & j = i, \\ 0, & \text{otherwise.} \end{cases}$$

For Λ-coalescents,

$$(\mathbf{Q})_{ij} = \binom{i}{i-j+1} \lambda_{i,i-j+1},$$

where $\lambda_{b,k}$ are defined in (12.4). In what follows, we define $q_i := \sum_{k=1}^{i-1} (\mathbf{Q})_{ik} = -(\mathbf{Q})_{ii}$.

See Schweinsberg (2000) for a more detailed description of Ξ-coalescents. For an alternative perspective based on a lookdown construction of particle systems with general reproduction mechanisms, see Donnelly et al (1999) and Birkner et al (2009).

13.2 Previous work on the expected SFS for Ξ-coalescents

As discussion in Chapter 7, the expected SFS for Kingman's coalescent is well understood, and can, in fact, be computed for an arbitrary ζ in $O(n^2)$ time (Bhaskar et al, 2015; Polanski and Kimmel, 2003) for a sample of size n. For the special case of constant ζ, Birkner et al (2013) and Blath et al (2016) developed methods to compute the expected SFS for Λ- and Ξ-coalescents, respectively. Birkner et al's method for Λ-coalescents takes $O(n^4)$ time and Blath et al's method for Ξ-coalescents takes time exponential in n, since it requires performing a sum over partitions of n numbers.

Berestycki et al (2007, 2014) studied the asymptotic behavior of the expected SFS as $n \to \infty$. Specifically, they derived simple formulae for time-homogeneous Λ-coalescents that come down from infinity. These asymptotic formulae can be rather inaccurate for finite n, however. Indeed, Birkner et al (2013) showed that, even for $n = 10,000$, there is a substantial discrepancy between the asymptotic formulae for some Λ-coalescents and the SFS obtained by simulation, illustrating the need for finite-sample computation. Nevertheless, such asymptotic results highlight some interesting properties of Λ-coalescents, as reviewed in Berestycki (2009).

In this chapter, we discuss the work of Spence et al (2016), who devised a method that can compute the expected SFS for time-inhomogeneous Λ- and Ξ-coalescents with arbitrary ζ in $O(n^3)$ time. In the case where ζ is a constant function, their method can compute the expected SFS in $O(n^2)$ time given the rate matrix $\mathbf{Q}$ of the ancestral process. Briefly, recall that the expected unnormalized SFS for a sample of size n is given by $\frac{\theta}{2}(\mathbb{E}[\tau_{n,1}], \ldots, \mathbb{E}[\tau_{n,n-1}])$, where $\frac{\theta}{2}$ denotes the population-scaled mutation rate and $\tau_{n,b}$ denote the sum of the lengths of all edges each subtending exactly b leaves (cf., Definition 5.9). Spence et al (2016) used subsampling arguments to show that the expectations $\mathbb{E}[\tau_{n,1}], \ldots, \mathbb{E}[\tau_{n,n-1}]$ can be computed from $\mathbb{E}T_2^{\mathrm{MRCA}}, \ldots, \mathbb{E}T_n^{\mathrm{MRCA}}$, where $\mathbb{E}T_k^{\mathrm{MRCA}}$ denotes the expected time to the most recent common ancestor for sample size $k \in \{2, \ldots, n\}$. Then, they showed how to compute $\mathbb{E}T_k^{\mathrm{MRCA}}$ using a spectral decomposition of the rate matrix $\mathbf{Q}$. We detail these two steps in the ensuing sections.

13.3 Relating the SFS to the TMRCAs

The expression for $\mathbb{E}[\tau_{n,b}]$ in Theorem 5.10 and the recursion for expected inter-coalescence times $\mathbb{E}[T_{n-1,k}]$ in Lemma 7.7 apply to Kingman's coalescent, but in general they do not hold for other Ξ-coalescent models. However, using exchangeability and a subsampling argument

similar to that in the proof of Lemma 7.7, one can obtain the following recursion for $\mathbb{E}[\tau_{n-1,k}]$, where $k < n-1$, for an arbitrary Ξ-coalescent:

$$\mathbb{E}[\tau_{n-1,k}] = \frac{k+1}{n}\mathbb{E}[\tau_{n,k+1}] + \frac{n-k}{n}\mathbb{E}[\tau_{n,k}], \tag{13.1}$$

which follows from removing a leaf uniformly at random from a sample of size k. Using (13.1), we show in the lemma below how the expectations $\mathbb{E}[\tau_{n,1}], \mathbb{E}[\tau_{n,2}], \ldots, \mathbb{E}[\tau_{n,n-1}]$ for a fixed sample size n are related to the "anti-singleton" expectations $\mathbb{E}[\tau_{2,1}], \mathbb{E}[\tau_{3,2}], \ldots, \mathbb{E}[\tau_{n,n-1}]$ for sample sizes $2, \ldots, n$. (The name follows from the fact that $\frac{\theta}{2}\mathbb{E}[\tau_{2,1}], \frac{\theta}{2}\mathbb{E}[\tau_{3,2}], \ldots, \frac{\theta}{2}\mathbb{E}[\tau_{n,n-1}]$ correspond to the anti-singleton entries — i.e., entries where exactly one haplotype has the ancestral allele and all other haplotypes have the derived allele — of the expected SFS for sample sizes $2, \ldots, n$.)

Lemma 13.1. *For an arbitrary time-inhomogeneous Ξ-coalescent governed by a measure $\frac{\Xi(d\boldsymbol{x})}{\zeta(t)}$, we have*

$$\begin{pmatrix} \mathbb{E}[\tau_{n,1}] \\ \mathbb{E}[\tau_{n,2}] \\ \vdots \\ \mathbb{E}[\tau_{n,n-1}] \end{pmatrix} = \mathbf{B} \begin{pmatrix} \mathbb{E}[\tau_{2,1}] \\ \mathbb{E}[\tau_{3,2}] \\ \vdots \\ \mathbb{E}[\tau_{n,n-1}] \end{pmatrix}, \qquad \text{for all } n \geq 2, \tag{13.2}$$

where $\mathbf{B}$ is an $(n-1)$-by-$(n-1)$ matrix that does not depend on the measure. Specifically, the entries of $\mathbf{B}$ are given by

$$(\mathbf{B})_{ij} = \begin{cases} (-1)^{i-j}\frac{1}{j+1}\binom{n-i-1}{j-i}\binom{n}{i}, & i \leq j, \\ 0, & i > j. \end{cases}$$

Proof. Define the *level* of $\mathbb{E}[\tau_{n,i}]$ as $n-i$. We use induction on the level to show that, for all $n \geq 2$ and $1 \leq i \leq n-1$,

$$\mathbb{E}[\tau_{n,i}] = \binom{n}{i}\sum_{j=i}^{n-1}(-1)^{i-j}\frac{1}{j+1}\binom{n-i-1}{j-i}\mathbb{E}[\tau_{j+1,j}]. \tag{13.3}$$

First, note that (13.3) holds for level 1, i.e., for $i = n-1$. Assume that (13.3) holds for level $n-i-1$. Then,

$$\begin{aligned}
\mathbb{E}[\tau_{n,i}] &= \frac{n}{n-i}\mathbb{E}[\tau_{n-1,i}] - \frac{i+1}{n-i}\mathbb{E}[\tau_{n,i+1}] \\
&= \frac{n}{n-i}\left[\binom{n-1}{i}\sum_{j=i}^{n-2}(-1)^{i-j}\frac{1}{j+1}\binom{n-i-2}{j-i}\mathbb{E}[\tau_{j+1,j}]\right] \\
&\quad - \frac{i+1}{n-i}\left[\binom{n}{i+1}\sum_{j=i+1}^{n-1}(-1)^{i+1-j}\frac{1}{j+1}\binom{n-i-2}{j-i-1}\mathbb{E}[\tau_{j+1,j}]\right] \\
&= \binom{n}{i}\left\{\frac{1}{i+1}\mathbb{E}[\tau_{i+1,i}] + (-1)^{n-1-i}\frac{1}{n}\mathbb{E}[\tau_{n,n-1}]\right.
\end{aligned}$$

$$+ \sum_{j=i+1}^{n-2} (-1)^{i-j} \frac{1}{j+1} \left[\binom{n-i-2}{j-i} + \binom{n-i-2}{j-i-1} \right] \mathbb{E}[\tau_{j+1,j}] \Bigg\}$$

$$= \binom{n}{i} \sum_{j=i}^{n-1} (-1)^{j-i} \frac{1}{j+1} \binom{n-i-1}{j-i} \mathbb{E}[\tau_{j+1,j}],$$

where the first equality holds by the recursion (13.1) and the second equality holds by the inductive hypothesis, by noting that $\mathbb{E}[\tau_{n-1,i}]$ and $\mathbb{E}[\tau_{n,i+1}]$ are both one level below $\mathbb{E}[\tau_{n,i}]$. □

The following lemma shows that the anti-singleton expectation $\mathbb{E}[\tau_{k,k-1}]$ on the right hand side of (13.2) can be written has a linear combination of the expected TMRCAs for samples sizes $2, \ldots, n$, with coefficients independent of the measure of the Ξ-coalescent:

Lemma 13.2. *For an arbitrary time-inhomogeneous Ξ-coalescent governed by a measure $\frac{\Xi(d\boldsymbol{x})}{\zeta(t)}$, we have*

$$\begin{pmatrix} \mathbb{E}[\tau_{2,1}] \\ \mathbb{E}[\tau_{3,2}] \\ \vdots \\ \mathbb{E}[\tau_{n,n-1}] \end{pmatrix} = \mathbf{C} \begin{pmatrix} \mathbb{E}[T_2^{MRCA}] \\ \mathbb{E}[T_3^{MRCA}] \\ \vdots \\ \mathbb{E}[T_n^{MRCA}] \end{pmatrix}, \qquad \textit{for all } n \geq 2,$$

where $\mathbf{C}$ is an $(n-1)$-by-$(n-1)$ bi-diagonal matrix with entries

$$(\mathbf{C})_{ij} = \begin{cases} i+1, & \textit{for } i = j, \\ -(i+1), & \textit{for } i = j+1, \\ 0, & \textit{otherwise.} \end{cases}$$

Proof. As in (13.1), we employ a subsampling argument. Consider a sample of size $k+1$. The only way that a subsample of size k can have a different time to most recent common ancestor is if the removed individual is a singleton after all of the other lineages have coalesced. The probability that we remove that singleton to form our subsample is $\frac{1}{k+1}$. Then, the expected amount of time during which there is one singleton and all of the other individuals have coalesced scaled by the mutation rate is exactly the anti-singleton entry. Thus,

$$\frac{1}{k+1} \mathbb{E}[\tau_{k+1,k}] = \mathbb{E}[T_{k+1}^{\mathrm{MRCA}}] - \mathbb{E}[T_k^{\mathrm{MRCA}}]$$

for $k > 1$. When $k = 1$, there are only 2 lineages, so the total branch length is the anti-singleton entry. Thus, $\tau_{2,1} = 2\mathbb{E}[T_2^{\mathrm{MRCA}}]$. Rewriting this as a matrix equation for $k \in \{1, \ldots, n-1\}$ completes the proof. □

By combining Lemmas 13.1 and 13.2, we obtain the following result:

Theorem 13.3. *For an arbitrary time-inhomogeneous Ξ-coalescent governed by a measure $\frac{\Xi(d\boldsymbol{x})}{\zeta(t)}$, there exists a universal $(n-1)$-by-$(n-1)$ matrix $\mathbf{A}$ that does not depend on the measure such that*

$$\begin{pmatrix} \mathbb{E}[\tau_{n,1}] \\ \mathbb{E}[\tau_{n,2}] \\ \vdots \\ \mathbb{E}[\tau_{n,n-1}] \end{pmatrix} = \mathbf{A} \begin{pmatrix} \mathbb{E}[T_2^{MRCA}] \\ \mathbb{E}[T_3^{MRCA}] \\ \vdots \\ \mathbb{E}[T_n^{MRCA}] \end{pmatrix}.$$

More precisely, $\mathbf{A} = \mathbf{BC}$*, where* $\mathbf{B}$ *and* $\mathbf{C}$ *are defined in Lemmas 13.1 and 13.2, respectively.*

As in the case of Kingman's coalescent (cf., Chapter 7.3), the expected first coalescence times $\mathbb{E}[T_{2,2}], \ldots, \mathbb{E}[T_{n,n}]$ play an important role in computing the SFS for general coalescent models. These expected coalescence times can be computed as (Bhaskar et al, 2015; Polanski and Kimmel, 2003)

$$\begin{aligned} \mathbb{E}[T_{k,k}] &= \int_0^\infty \mathbb{P}\left\{\text{time of first coalescence for } k \text{ individuals} > t\right\} dt \\ &= \int_0^\infty e^{(\mathbf{Q})_{kk} \int_0^t \frac{1}{\zeta(s)} ds} dt. \end{aligned} \tag{13.4}$$

For the general time-inhomogeneous case, we discuss in the next section how to compute $\mathbb{E}[T_2^{\text{MRCA}}], \ldots, \mathbb{E}[T_n^{\text{MRCA}}]$ using $\mathbb{E}[T_{2,2}], \ldots, \mathbb{E}[T_{n,n}]$ in $O(n^3)$ time. When ζ is a constant function, it is possible to do this in $O(n^2)$ time using the following lemma:

Lemma 13.4. *For a* Ξ*-coalescent governed by a measure of the form* $\frac{\Xi(d\boldsymbol{x})}{\zeta(t)}$ *where* ζ *is a constant function,* $\mathbb{E}[T_2^{MRCA}], \ldots, \mathbb{E}[T_n^{MRCA}]$ *can be computed recursively from* $\mathbb{E}[T_{2,2}], \ldots, \mathbb{E}[T_{n,n}]$ *and* $\mathbf{Q}$ *as follows:*

$$\begin{aligned} &\mathbb{E}[T_2^{MRCA}] = \mathbb{E}[T_{2,2}], \\ &\mathbb{E}[T_k^{MRCA}] = \mathbb{E}[T_{k,k}] + \sum_{l=2}^{k-1} \frac{(\mathbf{Q})_{kl}}{q_k} \mathbb{E}[T_l^{MRCA}], \qquad \text{for } k > 2. \end{aligned}$$

Proof. The formulae follow immediately from the homogeneity of the process, recursing on the sample size, and noting that the probability that the first coalescence event for a sample of size k results in k lineages merging down to l lineages is $\frac{(\mathbf{Q})_{kl}}{q_k}$. □

13.4 Relating the TMRCAs to the first coalescence times

In this section, we relate the expected TMRCAs $\mathbb{E}[T_2^{\text{MRCA}}], \ldots, \mathbb{E}[T_n^{\text{MRCA}}]$ to the expected first coalescence times $\mathbb{E}[T_{2,2}], \ldots, \mathbb{E}[T_{n,n}]$ for sample sizes $2, \ldots, n$. First, we establish a useful result on the spectral decomposition of the rate matrix $\mathbf{Q}$; this result was also obtained by Möhle and Pitters (2014, Equation 2.3) for the Bolthausen-Sznitman coalescent.

Lemma 13.5. *Fix an arbitrary* Ξ*-coalescent with* $q_i \neq q_j$ *for* $i \neq j$*, where* $q_i := \sum_{k=1}^{i-1} (\mathbf{Q})_{ik} = -(\mathbf{Q})_{ii}$*. Let* $\mathbf{Q} \in \mathbb{R}^{n \times n}$ *denote the rate matrix of the ancestral process corresponding to* $\Xi(d\boldsymbol{x})$ *(i.e., the process counting the number of extant lineages at time* t*). Then,*

$$\mathbf{Q} = \mathbf{U}\mathbf{E}\mathbf{U}^{-1},$$

where $(\mathbf{E})_{ij} = \delta_{ij}(\mathbf{Q})_{ii}$*, and*

$$(\mathbf{U})_{ij} = \begin{cases} 1, & i = j, \\ \frac{1}{q_i - q_j} \sum_{k=j}^{i-1} (\mathbf{Q})_{ik} (\mathbf{U})_{kj}, & i > j, \\ 0, & \textit{otherwise.} \end{cases}$$

Proof. By the construction of $\mathbf{U}$,

$$(\mathbf{U})_{ij}(\mathbf{Q})_{jj} = \sum_{k=j}^{i} (\mathbf{Q})_{ik}(\mathbf{U})_{kj},$$

which implies that $\mathbf{UE} = \mathbf{QU}$. Then, since $\mathbf{U}$ is triangular and has strictly positive diagonal entries, it is invertible. Therefore, $\mathbf{Q} = \mathbf{UEU}^{-1}$. □

The following result relates $\mathbb{E}[T_2^{\text{MRCA}}], \ldots, \mathbb{E}[T_n^{\text{MRCA}}]$ and $\mathbb{E}[T_{2,2}], \ldots, \mathbb{E}[T_{n,n}]$:

Lemma 13.6. *Fix an arbitrary Ξ measure and a strictly positive function ζ. Now consider a time-inhomogeneous coalescent governed by $\frac{\Xi(d\boldsymbol{x})}{\zeta(t)}$. If $\mathbb{E}[T_{k,k}] < \infty$, for $2 \le k \le n$, then*

$$\begin{pmatrix} \mathbb{E}[T_2^{MRCA}] \\ \mathbb{E}[T_3^{MRCA}] \\ \vdots \\ \mathbb{E}[T_n^{MRCA}] \end{pmatrix} = -(\mathbf{UD})_{2:n,2:n} \begin{pmatrix} \mathbb{E}[T_{2,2}] \\ \mathbb{E}[T_{3,3}] \\ \vdots \\ \mathbb{E}[T_{n,n}] \end{pmatrix},$$

where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is the diagonal matrix $\operatorname{diag}([\mathbf{U}^{-1}]_{\cdot,1})$, with $[\mathbf{U}^{-1}]_{\cdot,1}$ denoting the first column of $\mathbf{U}^{-1}$, and $(\mathbf{UD})_{2:n,2:n}$ denotes the submatrix of $\mathbf{UD}$ in rows and columns 2 through n.

Proof. Note that $\mathbb{E}T_k^{\text{MRCA}} = \int_0^\infty \mathbb{P}\{T_k^{\text{MRCA}} > t\}\, dt$. Therefore,

$$\begin{aligned} \mathbb{E}T_k^{\text{MRCA}} = \int_0^\infty \mathbb{P}\{T_k^{\text{MRCA}} > t\}\, dt &= \int_0^\infty \sum_{l=2}^{k} \left[e^{\mathbf{Q}\int_0^t \frac{1}{\zeta(s)} ds}\right]_{kl} dt \\ &= \int_0^\infty \sum_{l=2}^{n} \left[e^{\mathbf{Q}\int_0^t \frac{1}{\zeta(s)} ds}\right]_{kl} dt \\ &= \int_0^\infty \sum_{l=2}^{n} \left[\mathbf{U} e^{\mathbf{E}\int_0^t \frac{1}{\zeta(s)} ds} \mathbf{U}^{-1}\right]_{kl} dt, \end{aligned}$$

where the third equality follows from the fact that $\mathbf{Q}$ is lower triangular and hence so is its exponential. Now, since $\mathbf{U}$ is lower triangular, its inverse is as well. Therefore, we may ignore the value of $[e^{\mathbf{E}\int_0^t \frac{1}{\zeta(s)} ds}]_{1,1}$. Letting $\mathbf{F}(t) := e^{\mathbf{E}\int_0^t \frac{1}{\zeta(s)} ds}$ but with $\mathbf{F}_{1,1}(t) := 0$, note that $\int_0^\infty \mathbf{F}(t)dt = \operatorname{diag}(0, \mathbb{E}[T_{2,2}], \mathbb{E}[T_{3,3}], \ldots, \mathbb{E}[T_{n,n}])$. Then we have

$$\mathbb{E}T_k^{\text{MRCA}} = \int_0^\infty \sum_{l=2}^{n} [\mathbf{U}\mathbf{F}(t)\mathbf{U}^{-1}]_{kl} dt = \sum_{l=2}^{n} [\mathbf{U} \operatorname{diag}(0, \mathbb{E}[T_{2,2}], \ldots, \mathbb{E}[T_{n,n}])\mathbf{U}^{-1}]_{kl}.$$

Now, note that $(\mathbf{U})_{i,1} = 1$ for all i by Lemma 13.5 and induction. This implies $\sum_{l=1}^{n} [\mathbf{U}^{-1}]_{il} = \delta_{i1}$, or $\sum_{l=2}^{n} [\mathbf{U}^{-1}]_{il} = \delta_{i1} - [\mathbf{U}^{-1}]_{i1}$. Using this identity, we can rewrite the above expression for $\mathbb{E}T_k^{\text{MRCA}}$ as

$$\mathbb{E}T_k^{\mathrm{MRCA}} = -\sum_{j=2}^{n}[(\mathbf{UD})_{2:n,2:n}]_{k-1,j}\mathbb{E}[T_{j,j}],$$

where $\mathbf{D} = \mathrm{diag}([\mathbf{U}^{-1}]_{\cdot,1})$. Collecting these equations over $k \in \{2,\dots,n\}$ in matrix form leads to the desired result. □

Lemma 13.5 provides a recursion to compute $\mathbf{U}$, and $\mathbf{D}$ may be computed by noting that $(\mathbf{U}^{-1})_{11} = 1$ and then since $\mathbf{U}\mathbf{U}^{-1} = \mathbf{I}$ we have

$$\mathbf{U}^{-1}_{i1} = -\sum_{j=1}^{i-1}(\mathbf{U})_{ij}(\mathbf{U}^{-1})_{j1}.$$

Since the matrices $\mathbf{A}$ in Theorem 13.3 and $\mathbf{UD}$ in Lemma 13.6 do not depend on ζ, note that the SFS depends on time and the inhomogeneity of the coalescent process only through the expected first coalescence times.

Combining the results discussed so far, we obtain the following main result:

Theorem 13.7. *For an arbitrary time-inhomogeneous Ξ-coalescent governed by a measure $\frac{\Xi(d\boldsymbol{x})}{\zeta(t)}$, the expected SFS for a sample of size n can be computed in $O(n^3)$ time.*

Proof. Theorem 13.3 and Lemma 13.6 describe how to compute $\mathbb{E}[\tau_{n,1}],\dots,\mathbb{E}[\tau_{n,n-1}]$ from $\mathbb{E}[T_{2,2}],\dots,\mathbb{E}[T_{n,n}]$. For the runtime, note that each of the $O(n^2)$ entries of $\mathbf{U}$ requires $O(n)$ computations, and so computing $\mathbf{U}$ is $O(n^3)$. The matrices composing $\mathbf{A}$ are known in closed form, however, and constructing $\mathbf{D}$ only requires filling $O(n)$ entries, each requiring $O(n)$ computations for a total of $O(n^2)$. To then obtain the SFS from $\mathbb{E}[T_{2,2}],\dots,\mathbb{E}[T_{n,n}]$ simply requires iterated matrix vector products taking $O(n^2)$ time. The overall procedure thus requires $O(n^3)$. □

Remark 13.8. Other than computing $\mathbf{U}$, the algorithm presented above for computing the SFS is $O(n^2)$. Thus, for the Bolthausen-Sznitman Coalescent (Bolthausen and Sznitman, 1998) or Kingman's coalescent, where $\mathbf{U}$ is known in closed form (Möhle and Pitters, 2014, Theorem 1.1 and Appendix), the SFS can be computed in $O(n^2)$ time even for non-constant ζ.

13.5 Identifiability results

References

Berestycki J, Berestycki N, Schweinsberg J (2007) Beta-coalescents and continuous stable random trees. Annals of Probability pp 1835–1887

Berestycki J, Berestycki N, Limic V (2014) Asymptotic sampling formulae for Λ-coalescents. In: Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, Institut Henri Poincaré, vol 50, pp 715–731

Berestycki N (2009) Recent progress in coalescent theory. Ensaios Matematicos 16(1):1–193

Bhaskar A, Wang YXR, Song YS (2015) Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. Genome Research 25(2):268–279

Birkner M, Blath J, Möhle M, Steinrücken M, Tams J (2009) A modified lookdown construction for the xi-fleming-viot process with mutation and populations with recurrent bottlenecks. ALEA 6:25–61

Birkner M, Blath J, Eldon B (2013) Statistical properties of the site-frequency spectrum associated with lambda-coalescents. Genetics pp genetics–113

Blath J, Cronjäger MC, Eldon B, Hammer M (2016) The site-frequency spectrum associated with ξ-coalescents. Theoretical Population Biology 110:36–50

Bolthausen E, Sznitman AS (1998) On Ruelle's probability cascades and an abstract cavity method. Commun Math Phys 197:247–276

Donnelly P, Kurtz TG, et al (1999) Particle representations for measure-valued population models. The Annals of Probability 27(1):166–205

Durrett R, Schweinsberg J (2005) A coalescent model for the effect of advantageous mutations on the genealogy of a population. Stochastic Processes Appl 115:1628–1657

Huillet TE (2014) Pareto genealogies arising from a poisson branching evolution model with selection. Journal of Mathematical Biology 68(3):727–761

Möhle M, Pitters H (2014) A spectral decomposition for the block counting process of the bolthausen-sznitman coalescent. Electron Commun Probab 19(47):1–11

Möhle M, Sagitov S (2003) Coalescent patterns in diploid exchangeable population models. Journal of Mathematical Biology 47(4):337–352

Neher RA, Hallatschek O (2013) Genealogies of rapidly adapting populations. Proceedings of the National Academy of Sciences 110(2):437–442

Pitman J (1999) Coalescents with multiple collisions. Annals of Probability 27:1870–1902

Polanski A, Kimmel M (2003) New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics 165(1):427–436

Sagitov S (1999) The general coalescent with asynchronous mergers of ancestral lines. Journal of Applied Probability 36(4):1116–1125

Schweinsberg J (2000) Coalescents with simultaneous multiple collisions. Electronic Journal of Probability 5:1–50

Spence JP, Kamm JA, Song YS (2016) The site frequency spectrum for general coalescents. Genetics 202(4):1549–1561

Part VI
Diffusion Processes

Index