An Introduction to Mathematical Population Genetics and Coalescent Processes

Part I: Classical Models

by Jason Schweinsberg

University of California at San Diego

Overview

Goal of population genetics: understand the factors, such as mutation and natural selection, that cause genetic variability.

Mathematical population genetics:

1. Develop mathematical models for how populations evolve.
2. Observe DNA of individuals at the present time.
3. By comparing observations to predictions of the model, draw inferences about evolutionary history of the population.

Coalescent theory: take a sample from the current population, trace ancestral lines backwards in time.

Probabilistic tools: Branching processes, Exchangeability, Urn problems, Random walks, Poisson processes, Stable processes, Brownian motion.

Opportunities for collaborations with Biologists.

References

Richard Durrett, *Probability Models for DNA Sequence Evolution*

Warren J. Ewens, *Mathematical Population Genetics*

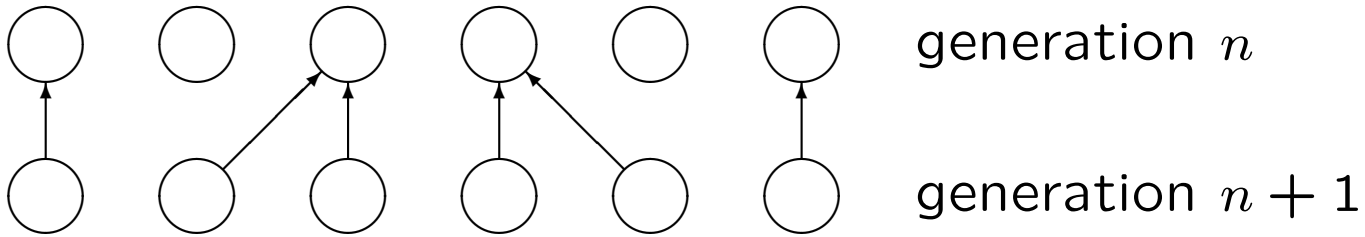
Nathanaël Berestycki, *Recent Progress in Coalescent Theory*

Jim Pitman, *Combinatorial Stochastic Processes*

The Wright-Fisher Model

One of the earliest models in population genetics, goes back to Fisher (1921) and Wright (1930).

- The population has fixed size $2N$.
- Generations do not overlap.
- Each member of the population has one parent, chosen at random from the individuals in the previous generation.



Members of population can be viewed as individuals in a haploid population or as chromosomes in a diploid population.

Partitions

A *partition* of a set S is a collection of disjoint subsets B_i of S such that

$$\bigcup_i B_i = S.$$

The sets B_i are called *blocks* of the partition. Blocks of a partition of size 1 are called *singletons*.

If π is a partition, write $i \sim_\pi j$ if i and j are in the same block. Denote by $\#\pi$ the number of blocks of π.

$\mathcal{P}_\infty =$ set of partitions of $\mathbb{N}$.

$\mathcal{P}_n =$ set of partitions of $\{1, \ldots, n\}$.

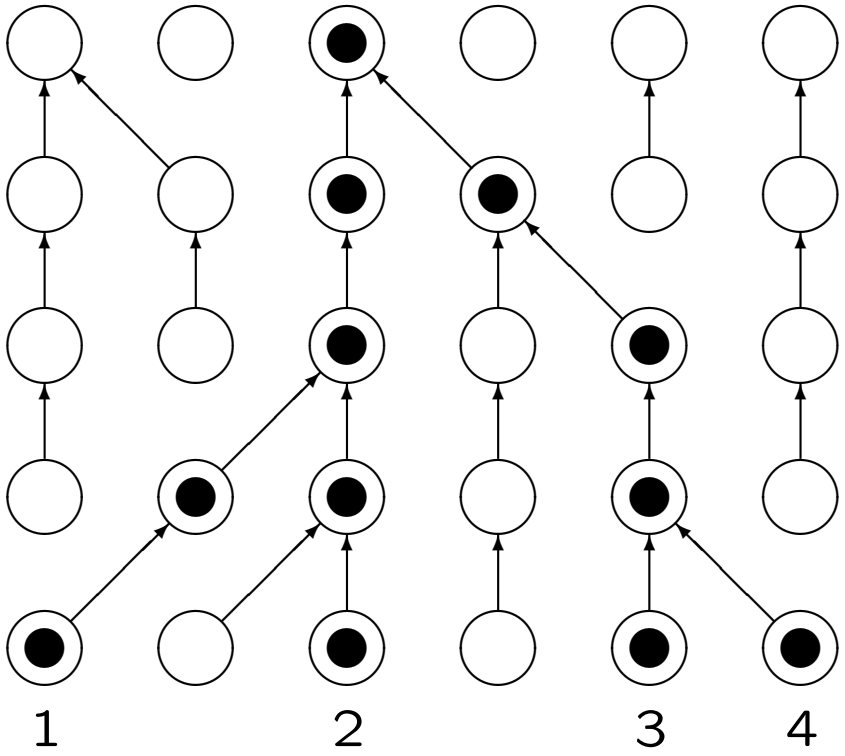
If $\pi \in \mathcal{P}_\infty$, or $\pi \in \mathcal{P}_m$ with $m > n$, then $R_n\pi \in \mathcal{P}_n$ is the restriction of π to $\{1, \ldots, n\}$, which means $i \sim_{R_n\pi} j$ if and only if $i \sim_\pi j$.

Example: $\pi = \{\{1,3,4,7,8\},\{2,5,9\},\{6\}\}$
$R_5\pi = \{\{1,3,4\},\{2,5\}\}$.

Ancestral Processes

Sample n individuals at the present time (generation 0).

Let $\Psi_N(k)$ be the partition of $\{1, \ldots, n\}$ such that $i \sim_{\Psi_N(k)} j$ if and only if the ith and jth sampled individuals have the same ancestor in generation $-k$.



$\Psi_N(4) = \{1, 2, 3, 4\}$

$\Psi_N(3) = \{\{1, 2\}, \{3, 4\}\}$

$\Psi_N(2) = \{\{1, 2\}, \{3, 4\}\}$

$\Psi_N(1) = \{\{1\}, \{2\}, \{3, 4\}\}$

$\Psi_N(0) = \{\{1\}, \{2\}, \{3\}, \{4\}\}$

Consider two individuals in generation 0. The probability that they have the same parent is $1/2N$.

Let T be the number of generations we have to go back before they have the same ancestor. Then

$$P(T > k) = \left(1 - \frac{1}{2N}\right)^k.$$

In particular,

$$P(T > 2Nx) = \left(1 - \frac{1}{2N}\right)^{\lfloor 2Nx \rfloor} \approx e^{-x}.$$

$T/2N$ has approximately an exponential distribution with rate 1.

The probability that three individuals in some generation all have the same parent is $1/(2N)^2$, so it is unlikely that three or more ancestral lines will merge simultaneously.

Kingman's n-Coalescent (Kingman, 1982)

Continuous-time Markov chain $(\Pi_n(t), t \geq 0)$ taking values in $\mathcal{P}_n$.

$\Pi_n(0)$ consists of n singletons.

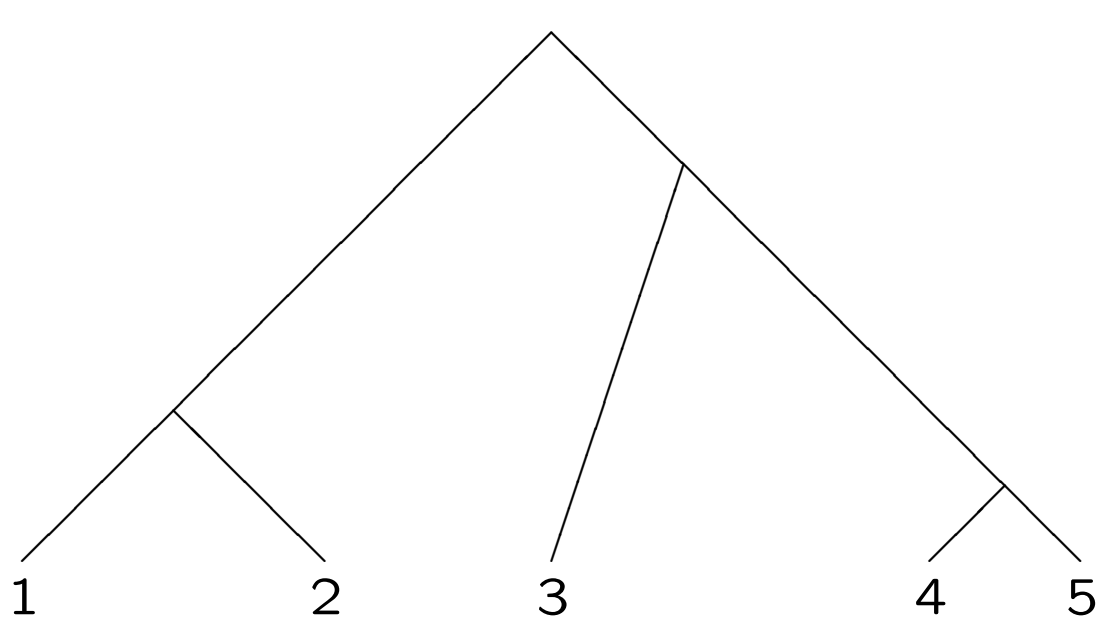
A transition that involves merging two blocks of the partition into one happens at rate 1. No other transitions are possible.

For $\xi, \eta \in \mathcal{P}_n$, write $\xi \prec \eta$ if η is obtained by merging two blocks of ξ. The transition rates are:

$$q(\xi, \eta) = \begin{cases} 1 & \text{if } \xi \prec \eta \\ 0 & \text{otherwise} \end{cases}$$

When there are k blocks, the distribution of the time until the next merger is exponential with rate $k(k-1)/2$. Then two randomly chosen blocks merge.

One time unit in Kingman's n-coalescent corresponds to $2N$ generations in the Wright-Fisher model.



Consistency: if $m > n$, then $(R_n \Pi_m(t), t \geq 0)$ and $(\Pi_n(t), t \geq 0)$ have the same law.

By Kolmogorov's Extension Theorem, there is a continuous-time Markov process $(\Pi_\infty(t), t \geq 0)$ with state space $\mathcal{P}_\infty$ such that $(R_n \Pi_\infty(t), t \geq 0)$ has the same law as $(\Pi_n(t), t \geq 0)$ for all n.

The process $(\Pi_\infty(t), t \geq 0)$ is called *Kingman's coalescent*.

A Limit Theorem

Theorem (Kingman, 1982): Suppose a population evolves according to the Wright-Fisher model with population size $2N$. Sample n individuals at random from the population in generation zero. Let $\Psi_N(k)$ be the partition of $\{1, \ldots, n\}$ such that $i \sim_{\Psi_N(k)} j$ if and only if the ith and jth sampled individuals have the same ancestor in generation $-k$. Let $(\Pi_n(t), t \geq 0)$ be Kingman's n-coalescent. Then, as $N \to \infty$,

$$(\Psi_N(\lfloor 2Nt \rfloor), t \geq 0) \Rightarrow (\Pi_n(t), t \geq 0).$$

Here $\Rightarrow$ denotes weak convergence of stochastic processes with respect to the Skorohod topology.

The Moran Model

Continuous-time model introduced by Moran (1958).

- The population has fixed size $2N$.
- Each individual independently lives for an Exponential(1) time, then is replaced by a new individual.
- If a new individual is born at time t, its parent is chosen uniformly at random from the individuals alive at time $t-$.
- Population can be defined for all $t \in \mathbb{R}$.

Suppose we sample n individuals at random from the population at time 0. Let $\Psi_N(t)$ be the partition of $\{1, \ldots, n\}$ such that $i \sim_{\Psi_N(t)} j$ if and only if the ith and jth individuals in the sample have the same ancestor at time $-t$. Then

$$(\Psi_N(Nt), t \geq 0)$$

is Kingman's n-coalescent.

Merger rate of two lineages is $2 \cdot \dfrac{1}{2N} = \dfrac{1}{N}$.

Exchangeable Sequences

A sequence of random variables $X_1, X_2, \ldots$ is called *exchangeable* if for every n and every permutation σ of $\{1, \ldots, n\}$,

$$(X_1, \ldots, X_n) =_d (X_{\sigma(1)}, \ldots, X_{\sigma(n)}).$$

de Finetti's Theorem: Suppose $X_1, X_2, \ldots$ is an exchangeable sequence of $\{0, 1\}$-valued random variables. Then there is a probability measure μ on $[0, 1]$ such that for any sequence $(a_i)_{i=1}^n$ consisting of k ones and $n - k$ zeros,

$$P(X_1 = a_1, \ldots, X_n = a_n) = \int_0^1 x^k (1 - x)^{n-k} \mu(dx).$$

Two-stage procedure: first pick a number x according to the distribution μ. Conditional on x, choose $X_1, X_2, \ldots$ to be i.i.d., taking the value 1 with probability x and 0 with probability $1 - x$. By the conditional SLLN,

$$\lim_{n \to \infty} \frac{X_1 + \cdots + X_n}{n}$$

exists and has distribution μ.

Polya Urns

Begin with a red balls and b blue balls. Repeatedly draw a ball at random from the urn and return it, along with another ball of the same color.

$$P(RRBBR) = \frac{a}{a+b} \cdot \frac{a+1}{a+b+1} \cdot \frac{b}{a+b+2} \cdot \frac{b+1}{a+b+3} \cdot \frac{a+2}{a+b+4}$$

The order of the 3 red and 2 blue balls does not matter. That is, the sequence is exchangeable.

The probability of getting a particular sequence of k red balls and $n-k$ blue balls is

$$\frac{(a+k-1)!}{(a-1)!} \cdot \frac{(b+n-k-1)!}{(b-1)!} \cdot \frac{(a+b-1)!}{(a+b+n-1)!} = \int_0^1 x^k (1-x)^{n-k} \, \mu(dx),$$

where μ has the Beta(a, b) distribution with density

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \, x^{a-1}(1-x)^{b-1}, \qquad 0 < x < 1.$$

Note: it's not important that a and b are integers.

Yule Processes

Consider a population started with one individual in which there are no deaths, and each individual gives birth at rate one. Let $X(t)$ be the population size at time t. Then $(X(t), t \geq 0)$ is called a *Yule process*.

A Yule process is a continuous-time Markov chain with state space $\mathbb{N}$, transition rates $q(j, j+1) = j$ and $q(j, k) = 0$ if $k \neq j+1$.

We have $E[X(t)] = e^t$, and $X(t)$ has a Geometric distribution with parameter e^{-t}:

$$P(X(t) = k) = e^{-t}(1 - e^{-t})^{k-1}, \qquad k = 1, 2, \ldots$$

We have

$$\lim_{t \to \infty} e^{-t} X(t) = W \text{ a.s.},$$

where W has an Exponential(1) distribution.

Yule Processes and Polya Urns

Consider a Yule process started with k individuals. Let $X_i(t)$ be the number of individuals at time t descended from ith individual at time zero.

Then $e^{-t}X_i(t) \to W_i$, where $W_i \sim$ Exponential(1), and

$$\frac{X_i(t)}{X_1(t) + \cdots + X_k(t)} \to \frac{W_i}{W_1 + \cdots + W_k} \sim \text{Beta}(1, k-1).$$

If descendants of the ith individual are colored red and all other individuals are colored blue, this process is exactly a Polya urn started with 1 red ball and $k-1$ blue balls.

Exchangeable Random Partitions

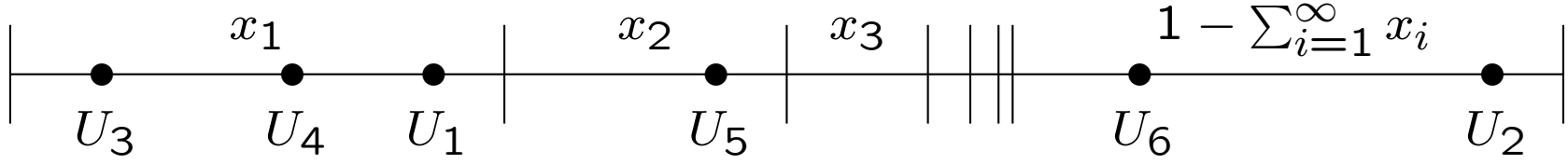
If $\pi \in \mathcal{P}_\infty$ and σ is a permutation of $\mathbb{N}$, define $\sigma\pi \in \mathcal{P}_\infty$ such that $\sigma(i) \sim_{\sigma\pi} \sigma(j)$ if and only if $i \sim_\pi j$.

If Π is a random partition of $\mathbb{N}$, we say Π is *exchangeable* if $\sigma\Pi =_d \Pi$ for all permutations σ of $\mathbb{N}$.

Let $\Delta = \left\{(x_1, x_2, \ldots) : x_1 \geq x_2 \geq \ldots \geq 0, \sum_{i=1}^{\infty} x_i \leq 1\right\}$.

Paintbox (stick-breaking) construction: Let $x = (x_1, x_2, \ldots) \in \Delta$. Divide $[0, 1]$ into subintervals of lengths $x_1, x_2, \ldots$ and $1 - \sum_{i=1}^{\infty} x_i$. Let $U_1, U_2, \ldots$ be i.i.d. Uniform(0,1).

Define Π such that $i \sim_\Pi j$ if and only if U_i and U_j fall in the same subinterval, other than the last interval of length $1 - \sum_{i=1}^{\infty} x_i$.



$$R_6\Pi = \{\{1, 3, 4\}, \{2\}, \{5\}, \{6\}\}.$$

Given $x \in \Delta$, let P^x denote the distribution of the associated paintbox partition.

Theorem (Kingman, 1982): Suppose Π is an exchangeable random partition of $\mathbb{N}$. Then there exists a probability measure μ on Δ such that

$$P(\Pi \in A) = \int_{\Delta} P^x(A)\, \mu(dx)$$

for all measurable subsets A of $\mathcal{P}_\infty$.

We call Π a μ-paintbox partition.

Suppose B is a block of Π. Then

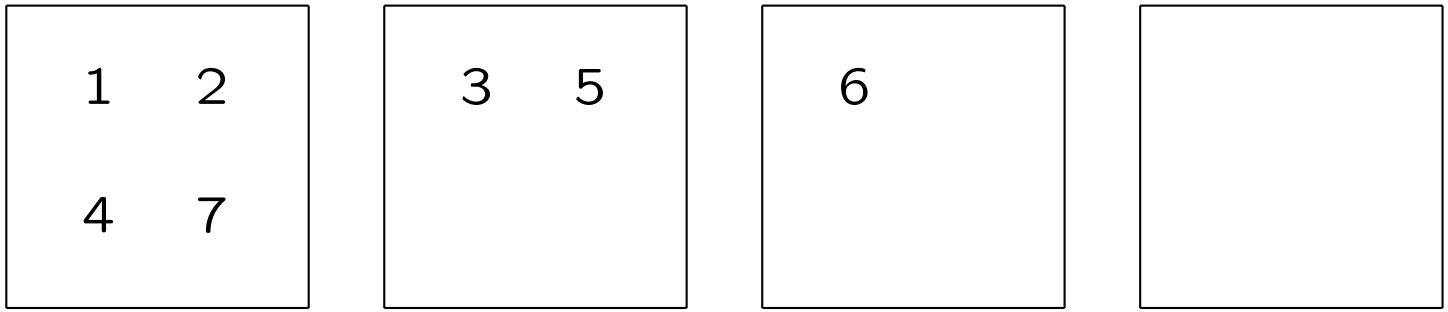
$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} 1_{\{i\in B\}}$$

exists and is called the *asymptotic frequency* of B. The sequence of ranked asymptotic frequencies of blocks has distribution μ.

Chinese Restaurant Process (Dubins and Pitman)

Let $\theta > 0$. Consider a restaurant with infinitely many tables.

- The first customer sits at the first table.
- For $n \geq 1$, the $(n+1)$st customer sits at a new table with probability $\theta/(n+\theta)$ otherwise picks one of the previous n customers at random and sits at that person's table.



Define a random partition Π of $\mathbb{N}$ such that $i \sim_\Pi j$ if and only if the ith and jth customers sit at the same table.

Example: $R_7\Pi = \{1,2,4,7\},\{3,5\},\{6\}$

Yule Process with Immigration

Consider a population which evolves as follows:

- Immigrants arrive at times of a rate θ Poisson process.
- Each individual gives birth at rate 1.
- There are no deaths.

Let Π be the random partition of $\mathbb{N}$ such that $i \sim_\Pi j$ if and only if the ith and jth individuals to appear are descended from the same immigrant.

When there are n individuals, the $(n+1)$st individual is a new immigrant with probability $\theta/(n+\theta)$, otherwise is equally likely to be born to any of the n existing individuals.

The distribution of Π is the same as the distribution of the partition obtained from the Chinese restaurant process.

Consider restriction of Π to $\{1, \ldots, n\}$. Example with $n = 7$:

$$\begin{aligned} P(R_7\Pi = \{\{1,2,4,7\},\{3,5\},\{6\}\}) & \\ &= \frac{\theta}{\theta} \cdot \frac{1}{1+\theta} \cdot \frac{\theta}{2+\theta} \cdot \frac{2}{3+\theta} \cdot \frac{1}{4+\theta} \cdot \frac{\theta}{5+\theta} \cdot \frac{3}{6+\theta} \\ &= \frac{\theta^3(4-1)!(2-1)!(1-1)!}{\theta(1+\theta)\ldots(n-1+\theta)}. \end{aligned}$$

Probability of getting a particular partition with a_j blocks of size j for all j:

$$\frac{1}{\theta(1+\theta)\ldots(n-1+\theta)} \prod_{j=1}^{n} \theta^{a_j}[(j-1)!]^{a_j}.$$

This depends only on block sizes, so Π is exchangeable.

Number of partitions with a_j blocks of size j for all j:

$$\frac{n!}{\prod_{j=1}^{n}(j!)^{a_j}a_j!}.$$

Probability of getting a_j blocks of size j for all j:

$$\frac{n!}{\theta(1+\theta)\ldots(n+\theta)} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!}.$$

Beta stick-breaking

Let $k \geq 1$. Consider customers not seated at first $k-1$ tables. Call those at kth table *red* and those at later tables *blue*.

When there are y red customers and z blue customers, the next customer is red with probability $y/(y+z+\theta)$ and blue with probability $(z+\theta)/(y+z+\theta)$.

This is Polya urn started with 1 red ball and θ blue balls, so the long-run fraction of red customers has a Beta$(1,\theta)$ distribution.

Let $W_1, W_2, \ldots$ be i.i.d. with a Beta$(1,\theta)$ distribution.
Let $Y_1 = W_1$, $Y_2 = (1-Y_1)W_2$, $Y_3 = (1-Y_1-Y_2)W_3, \ldots$.
Define $\tilde{Y}_1, \tilde{Y}_2, \ldots$ by ranking $Y_1, Y_2, \ldots$.

The distribution of $(\tilde{Y}_1, \tilde{Y}_2, \ldots)$ is Poisson-Dirichlet $(0,\theta)$.

The partition Π arising from the Chinese restaurant process is a Poisson-Dirichlet $(0,\theta)$-paintbox partition.

A two-parameter generalization (Pitman and Yor, 1997)

Suppose $0 \leq \alpha < 1$ and $\theta > -\alpha$.

Suppose, after n customers are seated, there are k occupied tables with $n_1, \ldots, n_k$ customers. Then the $(n+1)$st customer:

- sits at the ith table with probability $(n_i - \alpha)/(n + \theta)$.
- sits at a new table with probability $(\theta + k\alpha)/(n + \theta)$.

Define a random partition Π of $\mathbb{N}$ such that $i \sim_\Pi j$ if and only if the ith and jth customers sit at the same table.

Let $W_1, W_2, \ldots$ be independent, $W_k \sim \text{Beta}(1 - \alpha, k\alpha + \theta)$.
Let $Y_1 = W_1$, $Y_2 = (1 - Y_1)W_2$, $Y_3 = (1 - Y_1 - Y_2)W_3, \ldots$.
Define $\tilde{Y}_1, \tilde{Y}_2, \ldots$ by ranking $Y_1, Y_2, \ldots$.

The distribution of $(\tilde{Y}_1, \tilde{Y}_2, \ldots)$ is Poisson-Dirichlet (α, θ).

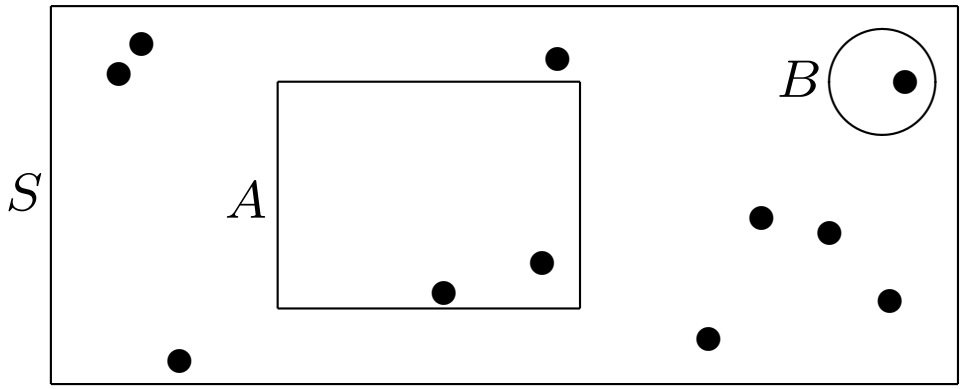
Π is a Poisson-Dirichlet (α, θ)-paintbox partition.

Poisson Point Processes

Let $(S, \mathcal{S}, \mu)$ be a σ-finite measure space.

A Poisson point process on S with *intensity measure* μ is a random measure Θ on $(S, \mathcal{S})$ such that

- If $A_1, \ldots, A_n$ are disjoint subsets of S, then $\Theta(A_1), \ldots, \Theta(A_n)$ are independent.
- If $A \in \mathcal{S}$, then $\Theta(A)$ has Poisson distribution with mean $\mu(A)$.



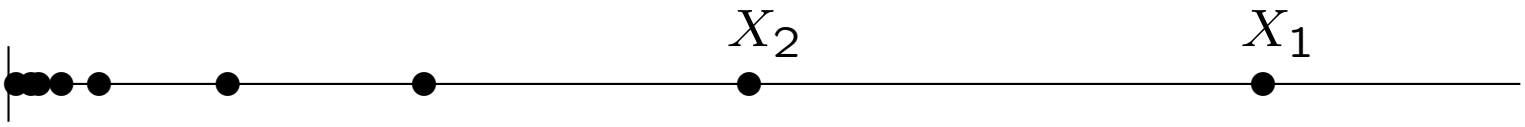
$$\Theta(A) = 2, \quad \Theta(B) = 1$$

Poisson process representation of Poisson-Dirichlet

Theorem (Ferguson, 1973): Let $\theta > 0$. Let $X_1 \geq X_2 \geq \dots$ be the points of a Poisson point process on $\mathbb{R}^+$ with intensity

$$\theta x^{-1} e^{-x}\, dx.$$

Let $X = \sum_{i=1}^{\infty} X_i$. Then the distribution of $(X_1/X, X_2/X, \dots)$ is Poisson-Dirichlet $(0, \theta)$.



Theorem (Perman, Pitman, and Yor (1992)): Let $0 < \alpha < 1$. Let $Y_1 \geq Y_2 \geq \dots$ be points of a Poisson point process on $\mathbb{R}^+$ with intensity

$$C x^{-1-\alpha}\, dx.$$

Let $Y = \sum_{i=1}^{\infty} Y_i$. Then the distribution of $(Y_1/Y, Y_2/Y, \dots)$ is Poisson-Dirichlet $(\alpha, 0)$.

Proof when $\alpha = 0$, $\theta > 0$

Consider a Yule process with immigration, up to a large time t.

Immigrants arrive at times of a rate θ Poisson process. An immigrant who arrives at time s has approximately We^{t-s} descendants alive at time t, where $W \sim$ Exponential(1).

Expected number of families of size greater than xe^t is

$$\theta \int_0^t P(We^{t-s} > xe^t)\, ds = \theta \int_0^t P(W > xe^s)\, ds = \theta \int_0^t e^{-xe^s}\, ds.$$

Change variables $y = xe^s$, this converges as $t \to \infty$ to

$$\theta \int_x^\infty y^{-1} e^{-y}\, dy.$$

Differentiating with respect to x gives the intensity

$$\theta x^{-1} e^{-x}\, dx.$$

Kingman's Coalescent: time back to the MRCA

If we start with a sample of size n, how far back in time do we have to go to find the most recent common ancestor (MRCA) of the sampled individuals?

Let T_k be the time to go from k lineages to $k-1$. The time back to the MRCA is

$$T = T_n + T_{n-1} + \cdots + T_2.$$

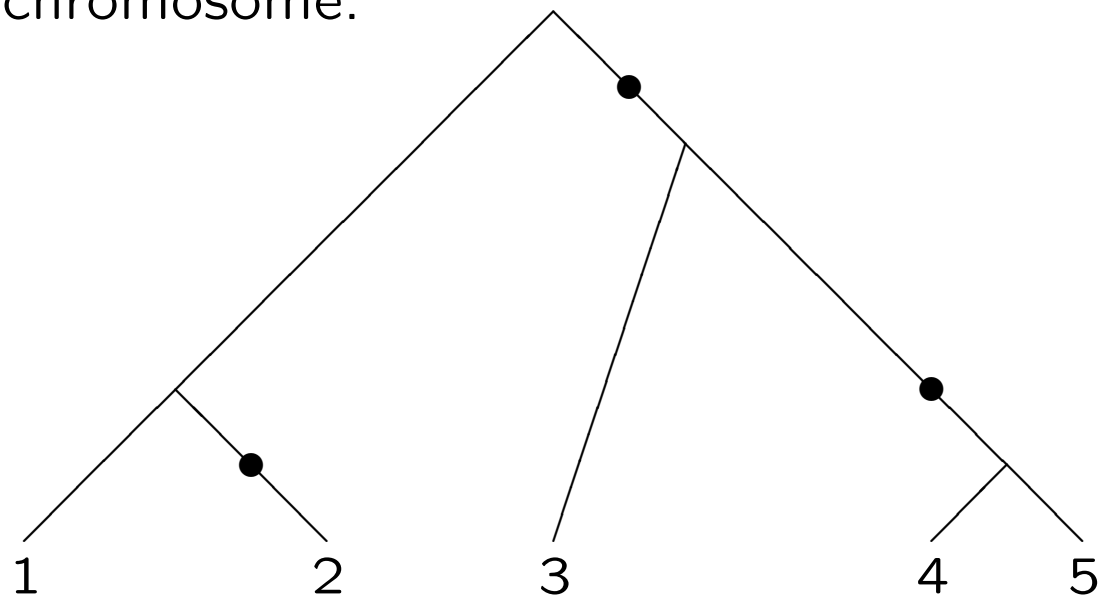
$$E[T] = E\left[\sum_{k=2}^{n} T_k\right] = \sum_{k=2}^{n} E[T_k] = \sum_{k=2}^{n} \frac{2}{k(k-1)} = \sum_{k=2}^{n} \left(\frac{2}{k-1} - \frac{2}{k}\right)$$
$$= \left(2 - \frac{2}{2}\right) + \left(\frac{2}{2} - \frac{2}{3}\right) + \left(\frac{2}{3} - \frac{2}{4}\right) + \cdots + \left(\frac{2}{n-1} - \frac{2}{n}\right) = 2 - \frac{2}{n}.$$

The average time back to the MRCA of 1,000,000 lineages is less than twice the time back of the MRCA of 2 lineages.

Adding Mutations to the Model

Assume mutations happen on each lineage at times of a rate $\theta/2$ Poisson process (mutation probability of $\theta/4N$ per generation).

Infinite sites model: assume each mutation happens at a different site on the chromosome.



1: AC**G**CTAA**T**AGC**A**
2: AC**G**CTAA**T**AGC**T**
3: AC**C**CTAA**T**AGC**A**
4: AC**C**CTAA**C**AGC**A**
5: AC**C**CTAA**C**AGC**A**

Pairwise Differences

1: AC**G**CTAA**T**AGC**A**
2: AC**G**CTAA**T**AGC**T**
3: AC**C**CTAA**T**AGC**A**
4: AC**C**CTAA**C**AGC**A**
5: AC**C**CTAA**C**AGC**A**

For $1 \leq i < j \leq n$, let $\Delta_{i,j}$ be the number of sites at which the ith and jth sequences differ.

Example: $\Delta_{1,2} = 1$, $\Delta_{2,3} = 2$, $\Delta_{4,5} = 0$, etc.

Consider the average number of pairwise differences

$$\Delta_n = \binom{n}{2}^{-1} \sum_{i<j} \Delta_{i,j}.$$

Because mutations contributing to $\Delta_{i,j}$ can occur on either of two lineages before they merge,

$$E[\Delta_n] = E\left[\frac{\theta}{2} \cdot 2 \cdot T_2\right] = \theta E[T_2] = \theta.$$

Segregating Sites

The number of segregating sites S_n is the number of sites on the DNA at which the n sampled individuals do not all agree.

Infinite sites model: S_n is the number of mutations in the tree.

Let L_n be the total branch length. Then

$$E[L_n] = E\left[\sum_{k=2}^{n} kT_k\right] = \sum_{k=2}^{n} k \cdot \frac{2}{k(k-1)} = \sum_{k=2}^{n} \frac{2}{k-1} \approx 2\log n$$

$$\mathrm{Var}(L_n) = \sum_{k=2}^{n} k^2 \mathrm{Var}(T_k) = \sum_{k=2}^{n} k^2 \cdot \frac{4}{k^2(k-1)^2} \to 4 \cdot \frac{\pi^2}{6}$$

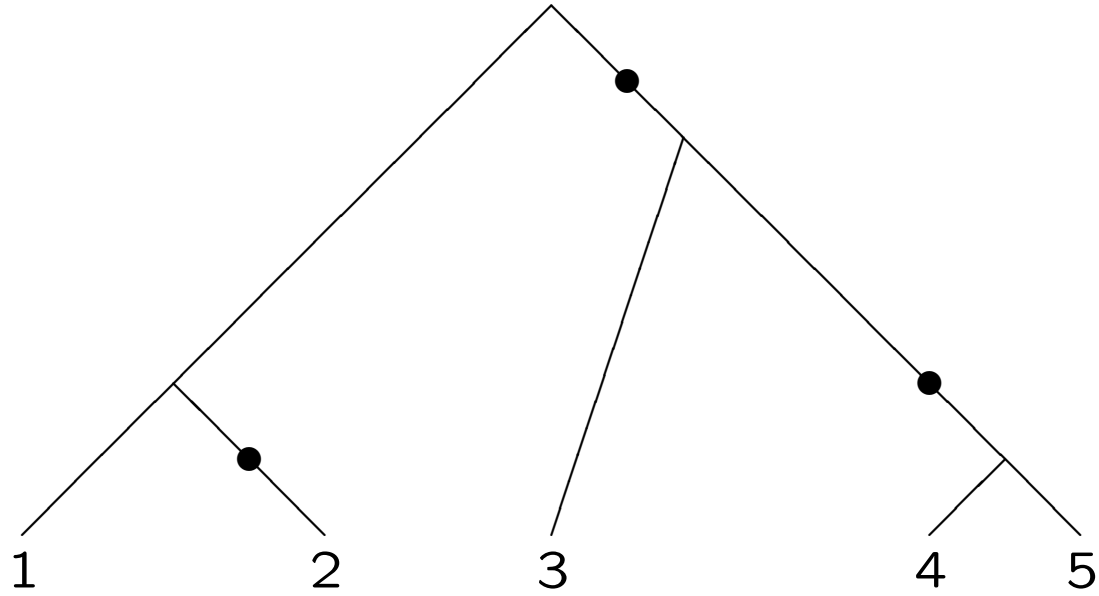
The conditional distribution of S_n given L_n is Poisson$(\theta L_n/2)$.

$$E[S_n] = \frac{\theta}{2} E[L_n] \approx \theta \log n.$$

$$\begin{aligned} \mathrm{Var}(S_n) &= E[\mathrm{Var}(S_n|L_n)] + \mathrm{Var}(E[S_n|L_n]) \\ &= E[(\theta/2)L_n] + \mathrm{Var}((\theta/2)L_n) \approx \theta \log n. \end{aligned}$$

Theorem: $(S_n - E[S_n])/\sqrt{\mathrm{Var}(S_n)} \Rightarrow N(0,1)$.

Other quantities of interest



Allelic partition: blocks represent groups of individuals that got the same mutations. Example: $\Pi_n = \{\{1\}, \{2\}, \{3\}, \{4, 5\}\}$.

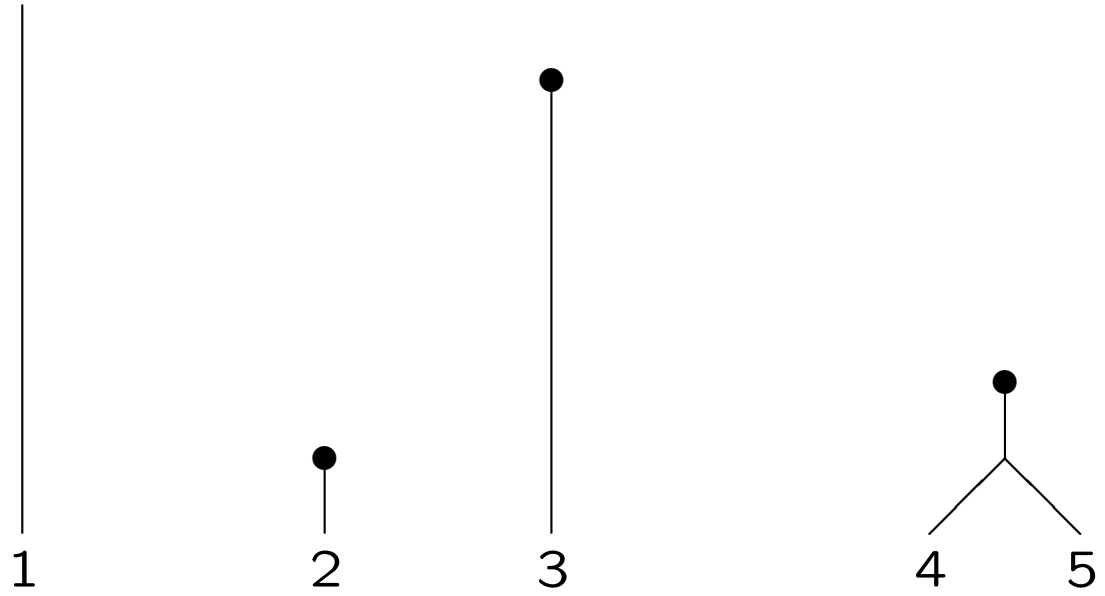
K_n = number of blocks of Π_n (haplotypes). Example: $K_n = 4$.

Allele frequency spectrum: $N_{k,n}$ = number of blocks of size k in allelic partition. Example: $N_{1,5} = 3$, $N_{2,5} = 1$.

Site frequency spectrum: $M_{k,n}$ = number of mutations affecting k individuals. Example: $M_{1,5} = 1$, $M_{2,5} = 1$, $M_{3,5} = 1$.

Allelic Partition

Modify picture by truncating branches at times of mutations.



Backward in time (going from $k+1$ to k lineages): coalescence happens at rate $k(k+1)/2$, with mutations at rate $(k+1)\theta/2$. Probability that mutation happens first is $\theta/(k+\theta)$.

Forward in time (going from k to $k+1$ lineages): start a new lineage with probability $\theta/(k+\theta)$. Otherwise, pick a random lineage to branch into two. This is Chinese Restaurant Process.

Ewens Sampling Formula (Ewens, 1972): We have

$$P(N_{1,n} = a_1, \ldots, N_{n,n} = a_n) = \frac{n!}{\theta(1+\theta)\ldots(n+\theta)} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!}.$$

Using the Chinese restaurant process, we compute

$$E[K_n] = \sum_{k=0}^{n-1} \frac{\theta}{\theta+k} \approx \theta \log n.$$

$$\text{Var}(K_n) = \sum_{k=0}^{n-1} \frac{\theta}{\theta+k}\left(1 - \frac{\theta}{\theta+k}\right) \approx \theta \log n.$$

Theorem (Watterson, 1975): $(K_n - E[K_n])/\sqrt{\text{Var}(K_n)} \Rightarrow N(0,1)$.

We always have $K_n \leq S_n + 1$. Typically, $S_n \approx K_n$ because most mutations occur near bottom of tree.

Theorem (Arratia, Barbour, and Tavaré (1992)): As $n \to \infty$,

$$(N_{1,n}, N_{2,n}, \ldots) \Rightarrow (Y_1, Y_2, \ldots),$$

where $Y_1, Y_2, \ldots$ are independent and $Y_k \sim \text{Poisson}(\theta/k)$. Also, $E[N_{k,n}] \to \theta/k$ as $n \to \infty$.

Site Frequency Spectrum

Theorem: If $1 \leq k \leq n-1$, then $E[M_{k,n}] = \theta/k$.

Proof Sketch: The expected number of mutations while there are j lineages is

$$j \cdot \frac{\theta}{2} \cdot E[T_j] = j \cdot \frac{\theta}{2} \cdot \frac{2}{j(j-1)} = \frac{\theta}{j-1}.$$

The probability that k of the n sampled individuals inherit such a mutation is the probability that, after we add $n-j$ balls to a Polya urn started with one red ball and $j-1$ blue balls, the total number of red balls will be k, which is

$$\binom{n-j}{k-1} \cdot \frac{(k-1)!(n-k-1)!(j-1)!}{(j-2)!(n-1)!} = \frac{(n-j)!(n-k-1)!(j-1)}{(n-j-k+1)!(n-1)!}$$

A combinatorial calculation gives

$$E[M_{k,n}] = \sum_{j=2}^{n} \frac{\theta}{j-1} \cdot \frac{(n-j)!(n-k-1)!(j-1)}{(n-j-k+1)!(n-1)!} = \frac{\theta}{k}.$$

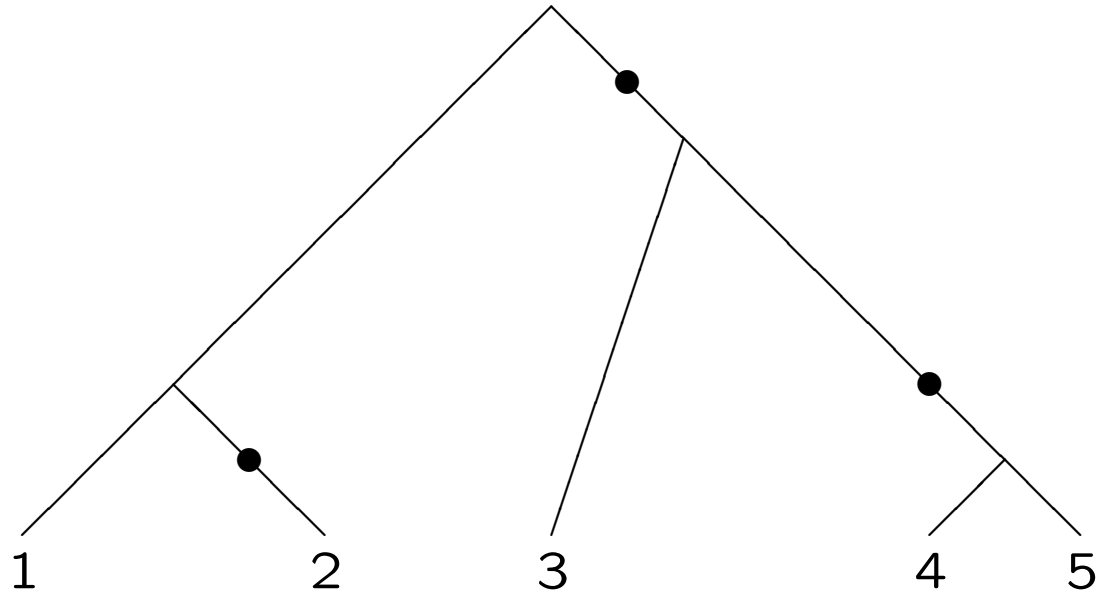
An Introduction to Mathematical Population Genetics and Coalescent Processes

Part II: Large family sizes and Coalescents with multiple mergers

by Jason Schweinsberg

University of California at San Diego

Review of Notation



Segregating sites: $S_n = 3$. Pairwise differences: $\Delta_n = 1.6$.

Allelic partition: blocks represent groups of individuals that got the same mutations. Example: $\Pi_n = \{\{1\}, \{2\}, \{3\}, \{4, 5\}\}$.

$K_n =$ number of blocks of Π_n (haplotypes). Example: $K_n = 4$.

Allele frequency spectrum: $N_{k,n} =$ number of blocks of size k in allelic partition. Example: $N_{1,5} = 3$, $N_{2,5} = 1$.

Site frequency spectrum: $M_{k,n} =$ number of mutations affecting k individuals. Example: $M_{1,5} = 1$, $M_{2,5} = 1$, $M_{3,5} = 1$.

Comparing predictions to data

From examining DNA sequences of n individuals, we can compute the quantities S_n, K_n, $M_{k,n}$, $N_{k,n}$.

Because

$$E[S_n] = \theta \sum_{k=1}^{n-1} \frac{1}{k},$$

we can estimate θ by

$$\hat{\theta} = S_n \Big/ \sum_{k=1}^{n-1} \frac{1}{k}.$$

Because $E[M_{k,n}] = \theta/k$ and $E[N_{k,n}] \approx \theta/k$ for large n, we can compare the observed values $M_{k,n}$ and $N_{k,n}$ to the "expected" value $\hat{\theta}/k$.

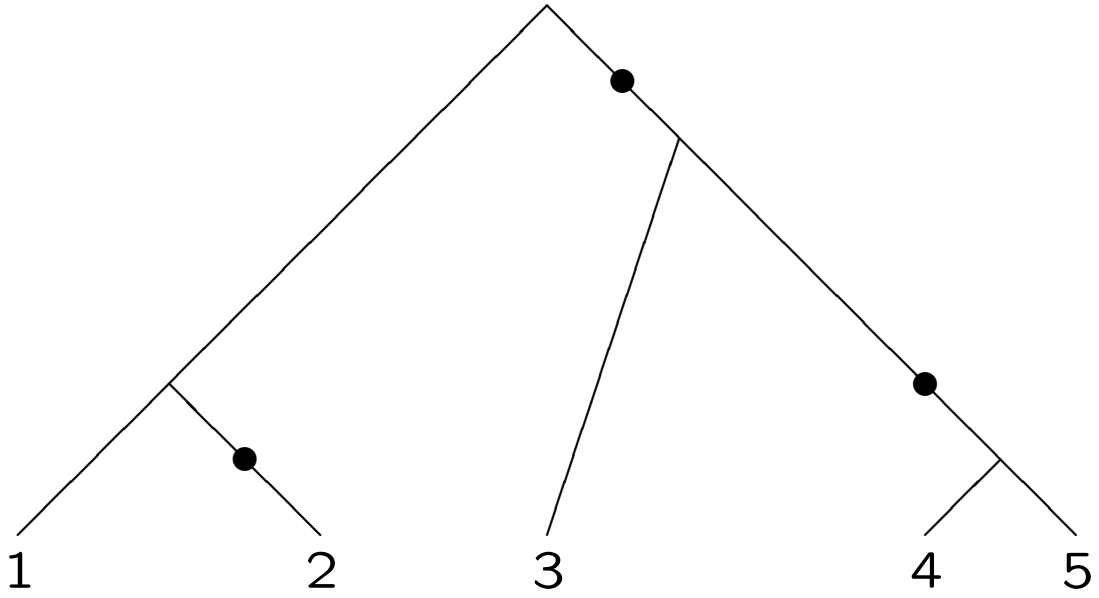
Violations of the infinite sites model

1. Three different nucleotides may appear at one site.

2. Under the infinite sites model, if S_i denotes the set of sequences that acquire the ith mutation, we must have

$$S_i \cap S_j = \emptyset, \quad S_i \subset S_j, \quad \text{or} \quad S_j \subset S_i.$$

   This is not always the case in real data sets.



$S_1 = \{2\}$

$S_2 = \{3, 4, 5\}$

$S_3 = \{4, 5\}$

The folded site frequency spectrum

Unless the genome of a more distantly related individual is available for comparison, we will not know which of the two nucleotides is the mutant and which is the ancestral type.

Consider instead the folded site frequency spectrum

$$\tilde{M}_{k,n} = \begin{cases} M_{k,n} + M_{n-k,n} & \text{if } 1 \leq k < n/2 \\ M_{k,n} & \text{if } k = n/2 \end{cases}$$

Then

$$E[\tilde{M}_{k,n}] = \begin{cases} \dfrac{\theta}{k} + \dfrac{\theta}{n-k} & \text{if } 1 \leq k < n/2 \\ \dfrac{\theta}{k} & \text{if } k = n/2 \end{cases}$$

Example: Mitochondrial DNA from American Indian tribe

Ward, Frazier, Dew-Jager, Pääbo (1991)

63 individuals from Nuu-Chah-Nulth tribe in Pacific Northwest.
Segment 360 nucleotides long from mitochondrial DNA.
There were 26 segregating sites.

Site Frequency Spectrum

k	Observed	Expected
1	5	5.5
2	2	2.8
3	4	1.8
4	1	1.4
5	1	1.1
6	3	0.9
7	1	0.8
8	0	0.7
9+	9	11.0

$S_n = 26$ and $K_n = 28$, violating $K_n \leq S_n + 1$. At least 5 sites have received two or more mutations.

Tajima's D-Statistic

Let $h_n = \sum_{k=1}^{n-1} \frac{1}{k}$.

Recall that $E[S_n/h_n] = \theta$ and $E[\Delta_n] = \theta$.

Thus, $\widehat{\theta}_W = S_n/h_n$ and $\widehat{\theta}_D = \Delta_n$ are unbiased estimates of θ.

If the genealogy of the population follows Kingman's coalescent, then $\widehat{\theta}_W$ and $\widehat{\theta}_D$ should be close.

Tajima's D-Statistic (Tajima, 1989):

$$D = \frac{\Delta_n - S_n/h_n}{\sqrt{a_n S_n + b_n S_n^2}},$$

where a_n and b_n are constants chosen to make Var$(D) \approx 1$.

Confidence intervals for different values of n in Tajima (1989).

Other Statistical Tests

Recall that $E[M_{k,n}] = \theta/k$, so in particular $E[M_{1,n}] = \theta$.

Fu and Li's D-statistic (Fu and Li, 1993):

$$D = \frac{S_n/h_n - M_{1,n}}{\sqrt{c_n S_n + d_n S_n^2}},$$

where c_n and d_n are constants chosen to make $\text{Var}(D) \approx 1$.

Tests based on Fu and Li's D-statistic are powerful when the number of mutations affecting just one individual is unusually high or low.

Other tests use the full site frequency spectrum:

- Fay and Wu (2000): H-statistic
- Zeng, Fu, Shi, and Wu (2006)

Example: *Bacillus anthracis* (Zwick et. al. (2011))

Data from 39 strains of *Bacillus anthracis*.
Sequenced region 303,000 nucleotides long. Some missing data.
There were 240 segregating sites.

Folded Site Frequency Spectrum

k	Observed	Expected
1	141	58.3
2	17	29.9
3	16	20.5
4	13	15.8
5	23	13.0
6	25	11.2
7	0	9.9
8	4	8.9
9+	1	72.5

Tajima's D-statistic: -1.76, p-value = 0.029.
Fu and Li's D-statistic: -3.01, p-value < 0.02.

Possible violations of assumptions

Violations of the assumptions in the Wright-Fisher model could cause the genealogy of to differ from Kingman's coalescent:

1. Non-constant population size
2. Spatial structure
   - Stepping stone model: Kimura (1953); Cox and Durrett (2002); Zähle, Cox, and Durrett (2005)
   - Continuous space: Barton, Etheridge, and Veber (2010)
3. Large family sizes ("sweepstakes reproduction")
   - This may affect the genealogies of some marine species: Hedgecock (1994), Eldon and Wakeley (2006, 2009).
4. Natural selection

Zwick et al. (2011): "Possible explanations for the pattern we observed are rapid demographic expansion of *B. anthracis*, or purifying selection acting to remove deleterious alleles"

Non-constant population size

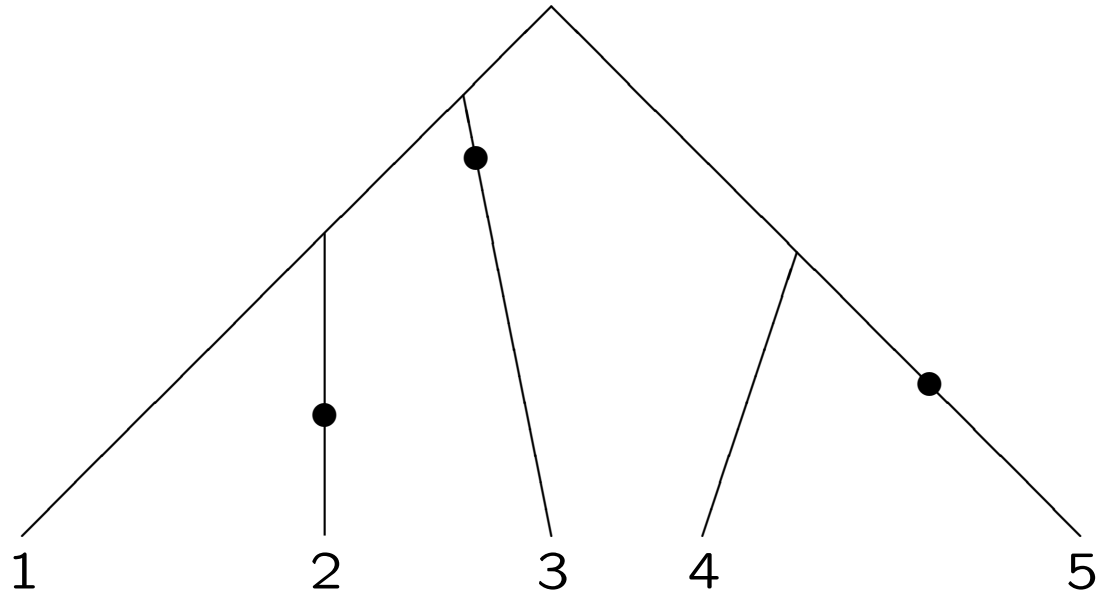
Consider the following modification of the Wright-Fisher model:

- For $t \in \mathbb{Z}$, the population size at time t is $N(t)$.
- Generations do not overlap.
- Each member of the population has one parent, chosen at random from individuals in the previous generation.

The probability that two individuals in generation $t+1$ have the same parent in generation t is $1/N(t)$.

The genealogy can be described by a time-change of Kingman's coalescent, in which the rate of coalescence is inversely proportional to the population size.

If the population size is increasing, then the coalescence rate gets faster as we go further back in time.

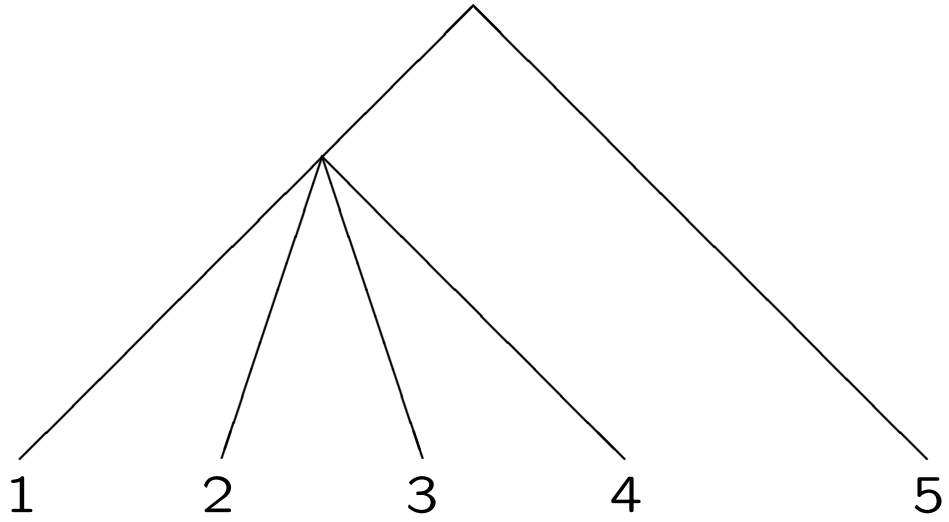


If the population size increases over time, there will be an excess of rare mutations.

$M_{1,n}$ will be larger than predicted by Kingman's coalescent, as observed in the data from *Bacillus anthracis*.

Coalescents with multiple mergers (Λ-coalescents)

Introduced by Pitman (1999) and Sagitov (1999).
More than two ancestral lines can merge at a time.



Applications of coalescents with multiple mergers:

- Large family sizes (many lineages trace back to individual with large number of offspring).
- Natural selection (many lineages trace back to individual who got a beneficial mutation).

Coalescents with Multiple Mergers

Definition (Pitman, 1999): A *coalescent with multiple mergers* is a $\mathcal{P}_\infty$-valued process $\Pi_\infty = (\Pi_\infty(t), t \geq 0)$ such that:

- $\Pi_\infty(0)$ is the partition of $\mathbb{N}$ into singletons.
- For all $n \in \mathbb{N}$, the process $(R_n\Pi_\infty(t), t \geq 0)$ is a continuous-time $\mathcal{P}_n$-valued Markov chain with the property that when $R_n\Pi_\infty(t)$ has b blocks, each k-tuple of blocks is merging to form a single block at some fixed rate $\lambda_{b,k}$, and no other transitions are possible.

$$\begin{aligned} \{1\},\{2\},\{3\},\{4\} &\to \{1,2,3\},\{4\} && \text{rate } \lambda_{4,3} \\ \{1,2\},\{3,4,5\},\{6\},\{7,8\} &\to \{1,2,3,4,5,6\},\{7,8\} && \text{rate } \lambda_{4,3} \end{aligned}$$

Law of process is determined by $\{\lambda_{b,k}, 2 \leq k \leq b\}$.

Not all collections of rates $\{\lambda_{b,k}, 2 \leq k \leq b\}$ are possible.
Example: We can't have both $\lambda_{3,3} = 1$ and $\lambda_{2,2} = 0$.

Consistency Condition

Consistency condition: $\lambda_{b,k} = \lambda_{b+1,k} + \lambda_{b+1,k+1}$ for $2 \le k \le b$.

$$\begin{array}{rll} \{1\},\ldots,\{b\} \to \{1,\ldots,k\},\{k+1\},\ldots,\{b\} & & \text{rate } \lambda_{b,k} \\ \{1\},\ldots,\{b+1\} \to \{1,\ldots,k,b+1\},\{k+1\},\ldots,\{b\} & & \text{rate } \lambda_{b+1,k+1} \\ \{1\},\ldots,\{b+1\} \to \{1,\ldots,k\},\{k+1\},\ldots,\{b+1\} & & \text{rate } \lambda_{b+1,k} \end{array}$$

Theorem (Pitman, 1999): An array $\{\lambda_{b,k}, 2 \le k \le b\}$ is consistent if and only if

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\,\Lambda(dx)$$

for some finite measure Λ on $[0,1]$.

Definition: We call a process with these rates a Λ-*coalescent*.

Examples:

1. $\Lambda = \delta_0$: Kingman's coalescent ($\lambda_{b,2} = 1, \lambda_{b,k} = 0$ for $k > 2$).
2. $\Lambda = \delta_1$: all blocks merge after exponential(1) hold.

Proof of Pitman's Theorem

Let $(\Pi_\infty(t), t \geq 0)$ be a coalescent with multiple mergers.
Let T be the time when $\{1\}$ and $\{2\}$ merge.

Let $B_1, B_2, \ldots$ be the blocks of $\Pi_\infty(T-)$, ordered by their smallest elements. Assume for now $\#\Pi_\infty(T-) = \infty$.

Let $\xi_i = 1$ if B_i merges with $\{1\}$ and $\{2\}$ at time T, and $\xi_i = 0$ otherwise. Then $(\xi_i)_{i=3}^\infty$ is exchangeable. Thus, by de Finetti's Theorem, there exists a probability measure $\tilde{\Lambda}$ such that

$$P(\xi_3 = \cdots = \xi_k = 1, \xi_{k+1} = \cdots = \xi_b = 0) = \int_0^1 x^{k-2}(1-x)^{b-k}\, \tilde{\Lambda}(dx).$$

We have $P(\xi_3 = \cdots = \xi_k = 1, \xi_{k+1} = \cdots = \xi_b = 0) = \lambda_{b,k}/\lambda_{2,2}$.

Let $\Lambda = \lambda_{2,2}\tilde{\Lambda}$. Then

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\, \Lambda(dx).$$

If $\#\Pi_\infty(T-) < \infty$, then condition $\#\Pi_\infty(T-) \geq k$ and apply Kolmogorov's Extension Theorem.

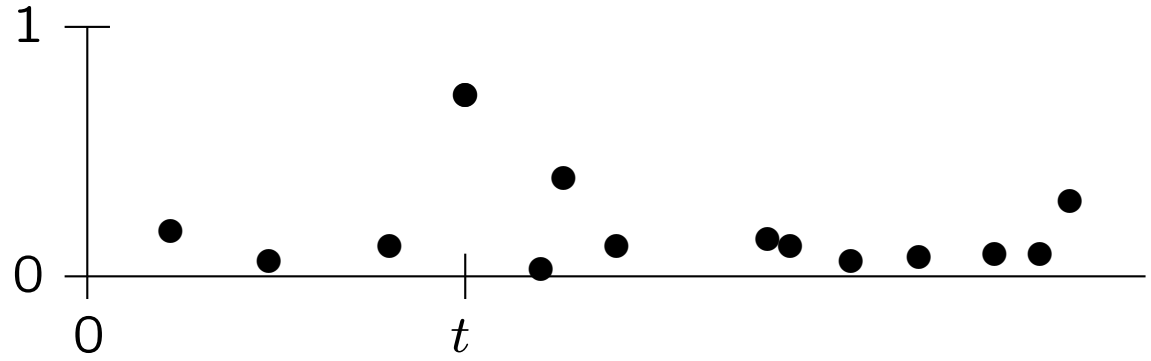
Poisson process construction

Let π be a partition of $\mathbb{N}$ into blocks $B_1, B_2, \ldots$. Let $p \in (0, 1]$. A p-merger of π is obtained as follows:

- Let $\xi_1, \xi_2, \ldots$ be i.i.d. with $P(\xi_i = 1) = p$, $P(\xi_i = 0) = 1 - p$.
- Merge the blocks B_i such that $\xi_i = 1$.

Write $\Lambda = a\delta_0 + \Lambda_0$, where $\Lambda_0(\{0\}) = 0$. Transitions:

- Each pair of blocks merges at rate a.
- Construct a Poisson point process on $[0, \infty) \times (0, 1]$ with intensity $dt \times p^{-2}\Lambda_0(dp)$. If (t, p) is a point of this Poisson process, then an p-merger occurs at time t.



When there are b blocks, $\lambda_{b,k} = \int_0^1 p^{k-2}(1-p)^{b-k}\,\Lambda(dp)$.

A more precise construction

Assume $\Lambda(\{0\}) = 0$. For all $\varepsilon > 0$, we have $\int_\varepsilon^1 p^{-2}\,\Lambda(dp) < \infty$, so p-mergers with $p \geq \varepsilon$ occur at a finite rate. However, the total merger rate $\int_0^1 p^{-2}\,\Lambda(dp)$ could be infinite.

Let Q_p denote the distribution of a sequence of i.i.d. random variables $\xi = (\xi_1, \xi_2, \dots)$ with $P(\xi_i = 1) = p$, $P(\xi_i = 0) = 1 - p$. Construct Poisson point process on $[0, \infty) \times \{0, 1\}^{\mathbb{N}}$ with intensity $dt \times L(d\xi)$, where

$$L(d\xi) = \int_0^1 Q_p(d\xi) \cdot p^{-2}\,\Lambda_0(dp).$$

First, we construct the restriction of Λ-coalescent to $\{1, \dots, n\}$. The rate of points (t, ξ) such that $\sum_{i=1}^n \xi_i \geq 2$ is at most

$$\int_0^1 \binom{n}{2} p^2 \cdot p^{-2}\,\Lambda_0(dp) < \infty.$$

If $B_1, B_2, \dots$ are the blocks at time $t-$, ordered by their smallest elements, at time t we merge blocks B_i such that $\xi_i = 1$.

This construction is consistent for different values of n, so the Λ-coalescent is well-defined.

Basic properties of Λ-coalescents

Suppose $(\Pi_\infty(t), t \geq 0)$ is a Λ-coalescent. Then:

1. Jump-hold property: let $T = \inf\{t : \Pi_\infty(t) \neq \Pi_\infty(0)\}$. If
$$\int_0^1 x^{-2}\, \Lambda(dx) < \infty,$$
then $P(T > 0) = 1$. Otherwise, $P(T = 0) = 1$.

2. Let $X_1(t) \geq X_2(t) \geq \ldots$ be the asymptotic frequencies of the blocks of the exchangeable random partition $\Pi_\infty(t)$. The coalescent has *proper frequencies* if $P(\sum_{k=1}^\infty X_k(t) = 1)$ for all $t > 0$. This is equivalent to:
$$P(\{1\} \text{ is a block of } \Pi_\infty(t)) = 0 \quad \text{for all } t > 0.$$
Thus, the Λ-coalescent has proper frequencies if and only if
$$\int_0^1 x^{-1}\, \Lambda(dx) = \infty.$$

3. If $c > 0$, then $(\Pi_\infty(ct), t \geq 0)$ is a $c\Lambda$-coalescent.

Coming Down from Infinity

Definition: Suppose Π_∞ is a Λ-coalescent. If $\#\Pi_\infty(t) = \infty$ for all $t > 0$, then we say the process *stays infinite*. If $\#\Pi_\infty(t) < \infty$ for all $t > 0$, then we say the process *comes down from infinity*.

Theorem (Pitman, 1999): If $\Lambda(\{1\}) = 0$, then the Λ-coalescent either comes down from infinity almost surely or stays infinite almost surely.

Let T_n be the first time that $1, \ldots, n$ are in the same block. Then $0 < T_2 \leq T_3 \leq \ldots \uparrow T_\infty$. If $T_\infty < \infty$, then all positive integers are in the same block after time T_∞.

For Kingman's coalescent, recall that

$$E[T_n] = \sum_{b=2}^{n} \binom{b}{2}^{-1} = 2 - \frac{2}{n},$$

which implies that $E[T_\infty] = 2$ and $T_\infty < \infty$ a.s.

Thus, Kingman's coalescent comes down from infinity.

Let

$$\lambda_b = \sum_{k=2}^{b} \binom{b}{k} \lambda_{b,k}$$

be the total rate of all mergers when the coalescent has b blocks.

Question: Does the Λ-coalescent come down from infinity if and only if $\sum_{b=2}^{\infty} \lambda_b^{-1} < \infty$?

Answer: No, because $\sum_{b=2}^{n} \lambda_b^{-1}$ overestimates $E[T_n]$.

Let γ_b be the rate at which the number of blocks is decreasing:

$$\gamma_b = \sum_{k=2}^{b} (k-1) \binom{b}{k} \lambda_{b,k}.$$

Theorem (Schweinsberg, 2000): Suppose $\Lambda(\{1\}) = 0$. Then the Λ-coalescent comes down from infinity if and only if

$$\sum_{b=2}^{\infty} \gamma_b^{-1} < \infty.$$

Example: the beta coalescent

Suppose Λ is the beta distribution with density

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1},$$

where $a > 0$ and $b > 0$. Then the process $(\Pi_\infty(t), t \geq 0)$ is called a beta coalescent.

If $a = b = 1$, then Λ is the uniform distribution on $[0, 1]$, and the process is called the Bolthausen-Sznitman coalescent, which was studied by Bolthausen and Sznitman (1998).

The beta coalescent:

- comes down from infinity if and only if $a < 1$.
- has proper frequencies if and only if $a \leq 1$.
- has the jump-hold property if and only if $a > 2$.

Because, for all $\varepsilon > 0$, the rate of p-mergers with $p > \varepsilon$ is finite, only the behavior of Λ near zero affects these properties.

Exchangeable Coalescent Processes

(Schweinsberg (2000), Möhle and Sagitov (2001), Bertoin and Le Gall (2003))

Suppose $\pi, \pi' \in \mathcal{P}_\infty$. Let $B_1, B_2, \ldots$ and $B'_1, B'_2, \ldots$ be the blocks of π and π', ordered by their smallest elements. Let $\text{Coag}(\pi, \pi')$ be the partition of $\mathbb{N}$ with blocks $\bigcup_{i \in B'_j} B_i$ for $j = 1, 2, \ldots$.

Example: $R_7\pi = \{1,2\},\{3\},\{4,5,6\},\{7\}$

$R_4\pi' = \{1,3\},\{2,4\}$

$R_7\text{Coag}(\pi, \pi') = \{1,2,4,5,6\},\{3,7\}$

Definition: An *exchangeable coalescent process* is a $\mathcal{P}_\infty$-valued Markov process $(\Pi(t), t \geq 0)$ such that for all $s, t \geq 0$, the conditional distribution of $\Pi(s+t)$ given $\Pi(s) = \pi$ is the distribution of $\text{Coag}(\pi, \pi_t)$, where π_t is an exchangeable random partition whose distribution depends only on t.

Also called *coalescents with simultaneous multiple mergers*: many mergers, each involving many blocks, may occur simultaneously.

Poisson Process Construction

There is a one-to-one correspondence between exchangeable coalescent processes (started from the partition of $\mathbb{N}$ into singletons) and finite measures on Δ.

Given a finite measure Ξ on Δ, the associated exchangeable coalescent process is called the Ξ-coalescent.

Write $\Xi = a\delta_{(0,0,\dots)} + \Xi_0$, where $\Xi_0(\{(0,0,\dots)\}) = 0$. Then transitions in the Ξ-coalescent $(\Pi(t), t \geq 0)$ are as follows:

- Each pair of blocks merges at rate a.
- Construct a Poisson point process on $[0,\infty)\times\Delta$ with intensity

$$dt \times \left(\sum_{j=1}^{\infty} x_j^2 \right)^{-1} \Xi_0(dx).$$

  If (t,x) is a point of this process, then $\Pi(t) = \mathrm{Coag}(\Pi(t-), \pi)$, where π has the law P^x of a paintbox partition.

Cannings models

Class of population models introduced by Cannings (1974).

- The population has fixed size N.
- Generations do not overlap.
- Let $\nu_{1,N}^{(r)}, \ldots, \nu_{N,N}^{(r)}$ denote the numbers of offspring of the N individuals in generation r. The distribution of $(\nu_{1,N}^{(r)}, \ldots, \nu_{N,N}^{(r)})$ is the same for each r and is exchangeable.
- Family sizes in different generations are independent.

We always have $\nu_{1,N} + \cdots + \nu_{N,N} = N$.

Wright-Fisher model is the special case in which the distribution of $(\nu_{1,N}, \ldots, \nu_{N,N})$ is multinomial$(N; 1/N, \ldots, 1/N)$.

Ancestral process: sample n individuals from generation 0. Let $\Psi_N(k)$ be the partition of $\{1, \ldots, n\}$ such that $i \sim_{\Psi_N(k)} j$ if and only if the ith and jth sampled individuals have the same ancestor in generation $-k$.

A Robustness Result

Notation: write $(x)_k = (x)(x-1)(x-2)\dots(x-k+1)$.

The probability that two individuals have the same parent is

$$c_N = NE\left[\frac{\nu_{1,N}}{N}\cdot\frac{\nu_{1,N}-1}{N-1}\right] = \frac{E[(\nu_{1,N})_2]}{N-1} = \frac{\mathrm{Var}(\nu_{1,N})}{N-1}.$$

The probability that three individuals all have the same parent is

$$\frac{E[(\nu_{1,N})_3]}{(N-1)(N-2)}.$$

Theorem (Möhle, 2000): Consider a Cannings model in which

$$\lim_{N\to\infty}\frac{E[(\nu_{1,N})_3]}{N^2 c_N} = 0.$$

Then, as $N\to\infty$,

$$\Psi_N(\lfloor t/c_N\rfloor, t\geq 0) \Rightarrow (\Pi_n(t), t\geq 0),$$

where $(\Pi_n(t), t\geq 0)$ is Kingman's n-coalescent and $\Rightarrow$ denotes weak convergence with respect to the Skorohod topology.

Models with Large Family Sizes

Theorem (Möhle and Sagitov, 2001): Suppose

$$\lim_{N\to\infty} \frac{E[(\nu_{1,N})_{k_1}\dots(\nu_{r,N})_{k_r}]}{N^{k_1+\dots+k_r-r}c_N}$$

exists for all integers $r \geq 1$ and $k_1,\dots,k_r \geq 2$ and

$$\lim_{N\to\infty} c_N = 0.$$

Then, as $N \to \infty$,

$$(\Psi_N(\lfloor t/c_N \rfloor), t \geq 0) \Rightarrow (\Pi(t), t \geq 0),$$

where $(\Pi(t), t \geq 0)$ is the restriction to $\{1,\dots,n\}$ of an exchangeable coalescent process.

Theorem (Sagitov, 1999): If

$$\lim_{N\to\infty} \frac{E[(\nu_{1,N})_2(\nu_{2,N})_2]}{N^2 c_N} = 0,$$

then $(\Pi(t), t \geq 0)$ is a coalescent with multiple mergers.

Every Ξ-coalescent with $\Xi(\Delta) = 1$ can arise as a limit.

Heavy-tailed offspring distributions

Consider the following population model:

- Population size N in each generation.
- Numbers of offspring $\xi_1, \ldots, \xi_N$ of the N individuals are i.i.d. with $P(\xi_i \geq k) \sim Ck^{-\alpha}$, where $\alpha > 0$, and $E[\xi_i] > 1$.
- Obtain the next generation by sampling N offspring without replacement.

Note: it is possible to have $\xi_1 + \cdots + \xi_N < N$, but the probability of this event decays exponentially in N.

Theorem (Schweinsberg, 2003):

- If $\alpha \geq 2$, the processes $(\Psi_N(\lfloor t/c_N \rfloor), t \geq 0)$ converge to Kingman's coalescent. When $\alpha > 2$, we have $c_N \sim \sigma^2/N$, where σ^2 is the variance of the number of surviving offspring.

- If $1 < \alpha < 2$, the processes $(\Psi_N(\lfloor AN^{\alpha-1}t \rfloor), t \geq 0)$ converge, for some constant A, to the Λ-coalescent, where Λ is the Beta$(2-\alpha, \alpha)$ distribution.

- If $\alpha = 1$, the processes $(\Psi_N(\lfloor (\log N)t \rfloor), t \geq 0)$ converge to the Bolthausen-Sznitman coalescent.

- If $0 < \alpha < 1$, the processes $(\Psi_N(m))_{m=0}^{\infty}$ converge to a discrete-time coalescent with simultaneous multiple mergers, and

$$\Xi(dx) = \left(\sum_{j=1}^{\infty} x_j^2 \right) \Theta(dx),$$

where Θ is the Poisson-Dirichlet $(\alpha, 0)$ distribution.

Idea of the proof $(1 < \alpha < 2)$

Let $\mu = E[\xi_i]$ be the mean of the offspring distribution.

We get a p-merger with $p \geq x$ if

$$\frac{\xi}{\xi + N\mu} \geq x \qquad \Longleftrightarrow \qquad \xi \geq \frac{x}{1-x} \cdot N\mu$$

The probability of such a family in a given generation is

$$NP\left(\xi \geq \frac{x}{1-x} \cdot N\mu\right) \sim NC\left(\frac{x}{1-x} \cdot N\mu\right)^{-\alpha}.$$

The rate of such mergers in the Beta$(2-\alpha, \alpha)$-coalescent is

$$\frac{1}{\Gamma(\alpha)\Gamma(2-\alpha)} \int_x^1 p^{-1-\alpha}(1-p)^{\alpha-1}\, dp = \frac{1}{\alpha\Gamma(\alpha)\Gamma(2-\alpha)}\left(\frac{x}{1-x}\right)^{-\alpha}.$$

An Introduction to Mathematical Population Genetics and Coalescent Processes

Part III: Continuous-state branching process and Beta coalescents

by Jason Schweinsberg

University of California at San Diego

Galton-Watson Processes

Definition: Let $(p_k)_{k=0}^{\infty}$ be a sequence of nonnegative numbers such that $\sum_{k=0}^{\infty} p_k = 1$. Consider a population with the following properties:

- There is one individual in generation zero.
- An individual has k offspring with probability p_k.
- The numbers of offspring of different individuals are independent.

Let Z_n be the population size in generation n. Then $(Z_n)_{n=0}^{\infty}$ is a *Galton-Watson process* with offspring distribution $(p_k)_{k=0}^{\infty}$.

Let L be a random variable such that

$$P(L = k) = p_k, \qquad k = 0, 1, 2, \ldots.$$

Let $m = E[L]$. We say the Galton-Watson process is *subcritical* if $m < 1$, *critical* if $m = 1$, and *supercritical* if $m > 1$.

Theorem: Let $q = P(Z_n = 0 \text{ for some } n)$ be the extinction probability. Then $q < 1$ if and only if $m > 1$ or $p_1 = 1$.

Theorem: Suppose $m = 1$ and $\text{Var}(L) = \sigma^2 < \infty$. Then

- (Kolmogorov, 1938): As $n \to \infty$, we have
$$P(Z_n > 0) \sim \frac{2}{n\sigma^2}.$$
- (Yaglom, 1947): As $n \to \infty$, the conditional distribution of Z_n/n given $Z_n > 0$ converges to the Exponential distribution with rate $2/\sigma^2$.

Theorem (Kesten and Stigum, 1966): Suppose $1 < m < \infty$. Then
$$\lim_{n\to\infty} Z_n/m^n = W \quad \text{a.s.}$$
If $E[L \log^+ L] < \infty$, then $P(W = 0) = q$ and $E[W] = 1$. Otherwise, $P(W = 0) = 1$.

Theorem (Seneta, 1968; Heyde, 1970): If $1 < m < \infty$, then there exist constants $(c_n)_{n=0}^{\infty}$ such that
$$\lim_{n\to\infty} Z_n/c_n$$
exists almost surely, and the limit is in $(0, \infty)$ almost surely on the event of nonextinction.

Lévy Processes

Definition: $(X(t), t \geq 0)$ is called a *Lévy process* if:

- If $0 = t_0 < t_1 < \cdots < t_n$, then the increments $X(t_1) - X(t_0), X(t_2) - X(t_1), \ldots, X(t_n) - X(t_{n-1})$ are independent.
- If $s, t \geq 0$, then $X(t+s) - X(t)$ has same distribution as $X(s)$.
- Almost surely $t \to X(t)$ is right continuous.

Examples:

1. Brownian motion with variance parameter σ^2. Then we have $E[e^{iuX(t)}] = \exp(-\frac{1}{2}\sigma^2 u^2 t)$.
2. Deterministic drift at rate d. Then $E[e^{iuX(t)}] = \exp(idut)$.
3. Jumps of size x at times of a rate λ Poisson process. Then $E[e^{iuX(t)}] = \exp(\lambda t(e^{iux} - 1))$.

Lévy-Khintchine Formula: Suppose $(X(t), t \geq 0)$ is a Lévy process. Then there is a function Φ called the *characteristic exponent* of the Lévy process such that $E[e^{iuX(t)}] = e^{t\Phi(u)}$. We have

$$\Phi(u) = idu - \frac{\sigma^2 u^2}{2} + \int_{-\infty}^{\infty} (e^{iux} - 1 - iux\mathbf{1}_{\{|x|\leq 1\}})\, \nu(dx),$$

where $d \in \mathbb{R}$, $\sigma^2 \geq 0$, and ν is a *Lévy measure* with $\nu(\{0\}) = 0$ and $\int_{-\infty}^{\infty}(1 \wedge x^2)\, \nu(dx) < \infty$.

If the Lévy process has no negative jumps, then for $\lambda \geq 0$, we have $E[e^{-\lambda X(t)}] = e^{t\Psi(\lambda)}$, where

$$\Psi(\lambda) = -d\lambda + \frac{\sigma^2 \lambda^2}{2} + \int_0^{\infty} (e^{-\lambda x} - 1 + \lambda x\mathbf{1}_{\{x\leq 1\}})\, \nu(dx).$$

The function $-\Psi$ is called the *Laplace exponent*.

A nondecreasing Lévy process is called a *subordinator*. Then

$$\Psi(\lambda) = -d\lambda + \int_0^{\infty} (e^{-\lambda x} - 1)\, \nu(dx),$$

where $d \geq 0$ and $\int_0^{\infty}(1 \wedge x)\, \nu(dx) < \infty$.

Examples of Subordinators

The *gamma subordinator* is the subordinator with

$$\Psi(\lambda) = \log\left(\frac{1}{1+\lambda}\right) = \int_0^\infty (e^{-\lambda x} - 1)\, x^{-1} e^{-x}\, dx$$

The density of $X(t)$ is

$$f(x) = \frac{1}{\Gamma(t)} x^{t-1} e^{-x}\, dx, \qquad x > 0.$$

For $0 < \alpha < 1$, the *stable subordinator with index* α has

$$\Psi(\lambda) = -\lambda^\alpha = \frac{\alpha}{\Gamma(1-\alpha)} \int_0^\infty (e^{-\lambda x} - 1)\, x^{-1-\alpha}\, dx.$$

If $J_1 \geq J_2 \geq \ldots$ are the jump sizes before time θ of the gamma subordinator, then the distribution of $(J_1/X(\theta), J_2/X(\theta), \ldots)$ is Poisson-Dirichlet $(0, \theta)$.

If $J_1 \geq J_2 \geq \ldots$ are the jump sizes before time t of the stable subordinator with index α, the distribution of $(J_1/X(t), J_2/X(t), \ldots)$ is Poisson-Dirichlet $(\alpha, 0)$.

Stable Distributions

Definition: The distribution of X is called *stable* if for all n there are constants a_n and b_n such that if $X_1, \ldots, X_n$ are i.i.d. and have the same distribution as X, then

$$\frac{X_1 + \cdots + X_n - b_n}{a_n} =_d X.$$

We have $a_n = n^{1/\alpha}$ with $0 < \alpha \leq 2$, in which case we say X has a stable law of index α.

The stable laws of index 2 are the normal distributions.

If $0 < \alpha < 2$, then $P(|X| > x) \sim Cx^{-\alpha}$ and

$$E[e^{iuX}] = idu + \int_{-\infty}^{\infty} (e^{iux} - 1 - iux\mathbf{1}_{\{|x| \leq 1\}})\, \nu(dx),$$

where

$$\nu(dx) = \begin{cases} c_1 x^{-1-\alpha}\, dx & \text{if } x > 0 \\ c_2 |x|^{-1-\alpha}\, dx & \text{if } x < 0 \end{cases}$$

Continuous-State Branching Processes

Definition: A continuous-state branching process (CSBP) is a $[0,\infty]$-valued Markov process $(X(t), t \geq 0)$ whose transition functions satisfy

$$p_t(a+b,\cdot) = p_t(a,\cdot) * p_t(b,\cdot).$$

Theorem (Lamperti, 1967): CSBPs are the processes that can be obtained as limits of processes $(X_n(t), t \geq 0)$ with

$$X_n(t) = \frac{Z_n(\lfloor nt \rfloor)}{a_n},$$

where each process $(Z_n(m))_{m=0}^{\infty}$ is a Galton-Watson process and $Z_n(0) \to \infty$ as $n \to \infty$.

Theorem (Lamperti, 1967): Let $(Y(s), s \geq 0)$ be a Lévy process with no negative jumps with $Y(0) > 0$, stopped when it hits zero. Let

$$S(t) = \inf\{u : \int_0^u Y(s)^{-1}\, ds > t\}.$$

Let $X(t) = Y(S(t))$. Then $(X(t), t \geq 0)$ is a CSBP. Every CSBP with no instantaneous jump to ∞ can be obtained this way.

Suppose $(X(t), t \geq 0)$ is a CSBP obtained from the Lévy process $(Y(t), t \geq 0)$. If $Y(0) = a$, then $E[e^{-\lambda Y(t)}] = e^{-\lambda a t} e^{t\Psi(\lambda)}$, where

$$\Psi(\lambda) = -d\lambda + \frac{\sigma^2\lambda^2}{2} + \int_0^\infty (e^{-\lambda x} - 1 + \lambda x \mathbf{1}_{\{x \leq 1\}})\, \nu(dx).$$

The function Ψ is called the *branching mechanism* of the CSBP.

Let $m = -\Psi'(0)$. Then $E[X(t)] = e^{mt}$. We call the process *subcritical* if $m < 0$, *critical* if $m = 0$, and *supercritical* if $m > 0$.

Theorem (Grey, 1974): Let $q = P(X(t) = 0 \text{ for some } t)$ be the extinction probability. Suppose $m \leq 0$. Then $q = 1$ if

$$\int_1^\infty \frac{1}{\Psi(\lambda)}\, d\lambda < \infty.$$

Otherwise, $q = 0$.

Theorem (Grey, 1974): The process $(X(t), t \geq 0)$ is conservative, meaning $P(X(t) < \infty \text{ for all } t) = 1$, if and only if for all $\delta > 0$,

$$\int_0^\delta \frac{1}{|\Psi(\lambda)|}\, d\lambda = \infty.$$

Examples of CSBPs

1. Feller's branching diffusion: $\Psi(\lambda) = \frac{1}{2}\lambda^2$. Then $(X(t), t \geq 0)$ satisfies the SDE

$$dX(t) = \sqrt{X(t)}\, dB(t),$$

where $(B(t), t \geq 0)$ is standard Brownian motion.

2. α-stable CSBP with $1 < \alpha < 2$:

$$\Psi(\lambda) = \lambda^\alpha, \qquad \nu(dx) = \frac{\alpha(\alpha-1)}{\Gamma(2-\alpha)}\, x^{-1-\alpha}\, dx.$$

3. Neveu's CSBP (Neveu, 1992):

$$\Psi(\lambda) = \lambda \log \lambda, \qquad \nu(dx) = x^{-2}\, dx$$

4. α-stable CSBP with $0 < \alpha < 1$:

$$\Psi(\lambda) = -\lambda^\alpha, \qquad \nu(dx) = \frac{\alpha}{\Gamma(1-\alpha)}\, x^{-1-\alpha}\, dx.$$

Note: we get extinction when $\alpha > 1$, explosion when $\alpha < 1$.

The genealogy of a CSBP

Three approaches to describing the genealogy of a CSBP:

- (Bertoin and Le Gall, 2000): Construct a flow of subordinators $(S^{(s,t)}(a), 0 \leq s \leq t, a \geq 0)$, such that $S^{(s,t)}(a)$ is the number of individuals at time t descended from the first a individuals at time s. If $0 < y < X(t)$, then $\inf\{b : S^{(s,t)}(b) \geq y\}$ is the ancestor of y at time s.

- (Le Gall and Le Jean, 1998; Duquesne and Le Gall, 2002): For CSBPs that will go extinct, can represent genealogy using the height process, construct random tree.

- (Donnelly and Kurtz, 1999): Represent the population by a countable system of particles.

The Lookdown Construction

(Donnelly and Kurtz, 1999)

Let $(X(t), t \geq 0)$ be a CSBP with $X(0) = a$.

For all $t \geq 0$, there is a particle at each *level* $j \in \mathbb{N}$, and the particle at level j has a *type* in E denoted by $\xi_j(t)$. Define

$$R(t) = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \delta_{\xi_j(t)}.$$

Let $M_t = X(t)R(t)$, so $(M_t, t \geq 0)$ is a measure-valued process.

In some models, the type of a particle could represent the spatial location or genetic type of an individual, could change over time.

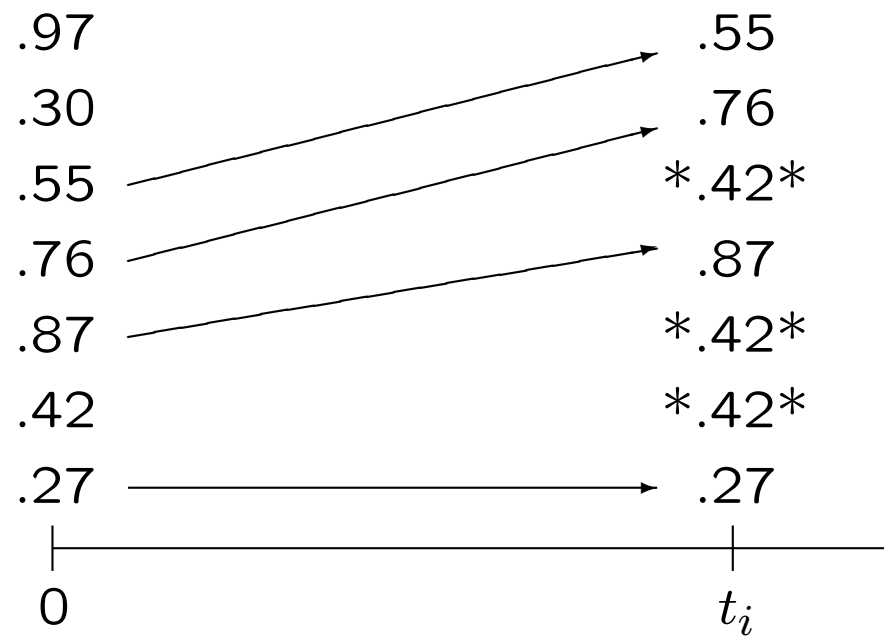
We will assign the types at time 0 to be i.i.d. Uniform$(0, a)$ random variables. Types change only at birth/death events.

Here $M_t(A)$ is number of individuals at time t descended from individuals with types in A. If $b < a$, then $(M_t([0, b]), t \geq 0)$ has the same law as the CSBP started at b.

Assume $\sigma^2 = 0$ (no Brownian component).

Consider the points (t_i, y_i), where the t_i are the jump times and $y_i = (X(t_i) - X(t_i-))/X(t_i)$.

At time t_i, toss a coin for each level with probability y_i of heads. All the levels whose coins come up heads adopt the type of the smallest such level. Other types are shifted upward.



If $0 < s < t$, one can identify the particle at time s from which a particular particle at time t inherited its type.

The genealogy of Neveu's CSBP

Theorem (Bertoin and Le Gall, 2000): Let $(X(t), t \geq 0)$ be a CSBP with $\Psi(\lambda) = \lambda \log \lambda$. Fix $t > 0$. Let $\Pi(s)$ be the partition of $\mathbb{N}$ such that $i \sim_{\Pi(s)} j$ if and only if the particles at levels i and j at time t have the same ancestor at time $t - s$. Then $(\Pi(s), 0 \leq s \leq t)$ is the Bolthausen-Sznitman coalescent.

Bertoin and Le Gall's proof uses the flow of subordinators and depends on comparing the finite-dimensional distributions of the two processes:

- The process $(M_t([0, a]), a \geq 0)$ is a stable subordinator of index $\alpha = e^{-t}$.
- (Bolthausen and Sznitman, 1998) The distribution of asymptotic block frequencies of $\Pi(t)$ is Poisson-Dirichlet $(e^{-t}, 0)$.

Finite-dimensional distributions for other Λ-coalescents unknown.

Obtaining the genealogy from the jump rates

- Assume $\nu(dx) = g(x)\,dx$. Let A be current population size.
- A jump in the population of size x happens at rate $Ag(x)\,dx$.
- After the jump, a fraction $p = x/(A+x)$ of the population was born at the time of the jump. Tracing ancestral lines backwards in time, a p-merger occurs at this time.
- We have $x = Ap/(1-p)$ and $dx/dp = A/(1-p)^2$, so the rate of p-mergers is

$$\eta(dp) = Ag\left(\frac{Ap}{1-p}\right) \cdot \frac{A}{(1-p)^2}\,dp.$$

Example: Suppose $g(x) = cx^{-1-\alpha}$ for $0 < \alpha < 2$. Then

$$\eta(dp) = cA^{1-\alpha}p^{-1-\alpha}(1-p)^{\alpha-1}\,dp.$$

After speeding up time by $A^{\alpha-1}$, we get a Λ-coalescent with

$$\Lambda(dp) = A^{\alpha-1}p^2\eta(dp) = cp^{1-\alpha}(1-p)^{\alpha-1},$$

which is Beta$(2-\alpha, \alpha)$ up to a constant.

A More General Result

Theorem (Birkner, Blath, Capaldo, Etheridge, Möhle, Schweinsberg, and Wakolbinger (2005)): Let $(X(t), t \geq 0)$ be a CSBP with $\sigma^2 = 0$ and $\nu(dx) = cx^{-1-\alpha}\, dx$, where $0 < \alpha < 2$. Let

$$T(s) = \inf\left\{u : \int_0^u X(t)^{1-\alpha}\, dt > s\right\}.$$

Fix $t > 0$, and define the $\mathcal{P}_\infty$-valued process $(\Pi(s), 0 \leq s \leq t)$ such that $i \sim_{\Pi(s)} j$ if and only if the particles at levels i and j at time $T(t)$ inherited their types from the same ancestor at time $T(t-s)$. Then $(\Pi(s), 0 \leq s \leq t)$ is the Beta$(2-\alpha, \alpha)$-coalescent.

Remarks:

- When $\alpha = 1$, we recover result of Bertoin and Le Gall (2000).
- When $1 < \alpha < 2$, $X(t) = 0$ for large t, but $X(T(t)) > 0$ for all t because $\lim_{t\to\infty} T(t) = \tau_0 = \inf\{t : X(t) = 0\}$.
- When $0 < \alpha < 1$, $X(t) = \infty$ for large t, but $X(T(t)) < \infty$ for all t because $\lim_{t\to\infty} T(t) = \tau_\infty = \inf\{t : X(t) = \infty\}$.

Comparison with discrete case

- Feller's branching diffusion arises as a limit of critical Galton-Watson processes for which the variance of the offspring distributions is σ^2. Kingman's coalescent is the genealogy in the discrete and continuous cases.

- When $1 < \alpha < 2$, the α-stable CSBP arises as a limit of critical Galton-Watson processes whose offspring distribution satisfies $P(X_1 \geq k) \sim Ck^{-\alpha}$. We get the same coalescent as in the discrete model. Time change is the same as in the discrete case, where $c_N \sim 1/AN^{\alpha-1}$.

- Correspondence between discrete and continuous cases fails when $0 < \alpha < 1$. No simultaneous mergers in continuous case.

A small-time approximation

Berestycki, Berestycki, and Limic (2012): For small times, the Λ-coalescent is well-approximated by the genealogy of a CSBP with $X(0) = 1$ and branching mechanism

$$\Psi(\lambda) = \int_0^1 (e^{-\lambda x} - 1 + \lambda x) x^{-2} \, \Lambda(dx).$$

Theorem (Bertoin and Le Gall, 2006): The Λ-coalescent comes down from infinity if and only if

$$\int_1^\infty \frac{1}{\Psi(\lambda)} \, d\lambda < \infty.$$

Theorem (Berestycki, Berestycki, and Limic (2010)): Let $N(t)$ be the number of blocks in the Λ-coalescent after time t. Define $v(t)$ so that

$$\int_{v(t)}^\infty \frac{1}{\Psi(\lambda)} \, d\lambda = t.$$

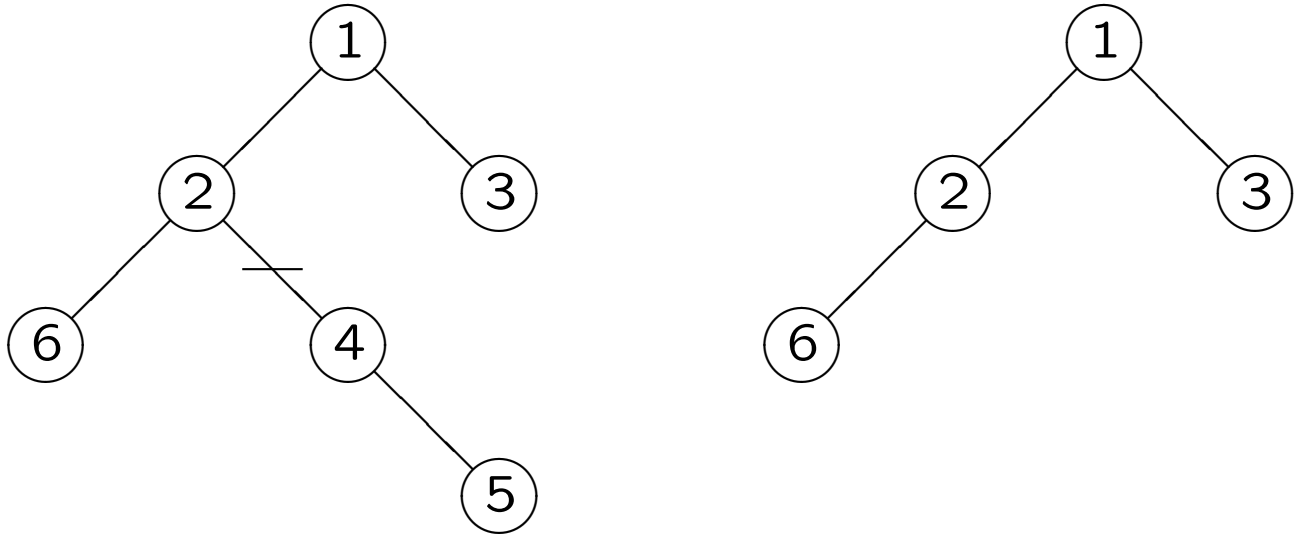
Then

$$\lim_{t \to 0} \frac{N(t)}{v(t)} = 1 \quad \text{a.s.}$$

Random recursive trees

Definition: A tree on n vertices labeled $1, \ldots, n$ is called a *recursive tree* if the root is labeled 1 and, for $2 \leq k \leq n$, the labels on the path from the root to k are increasing.

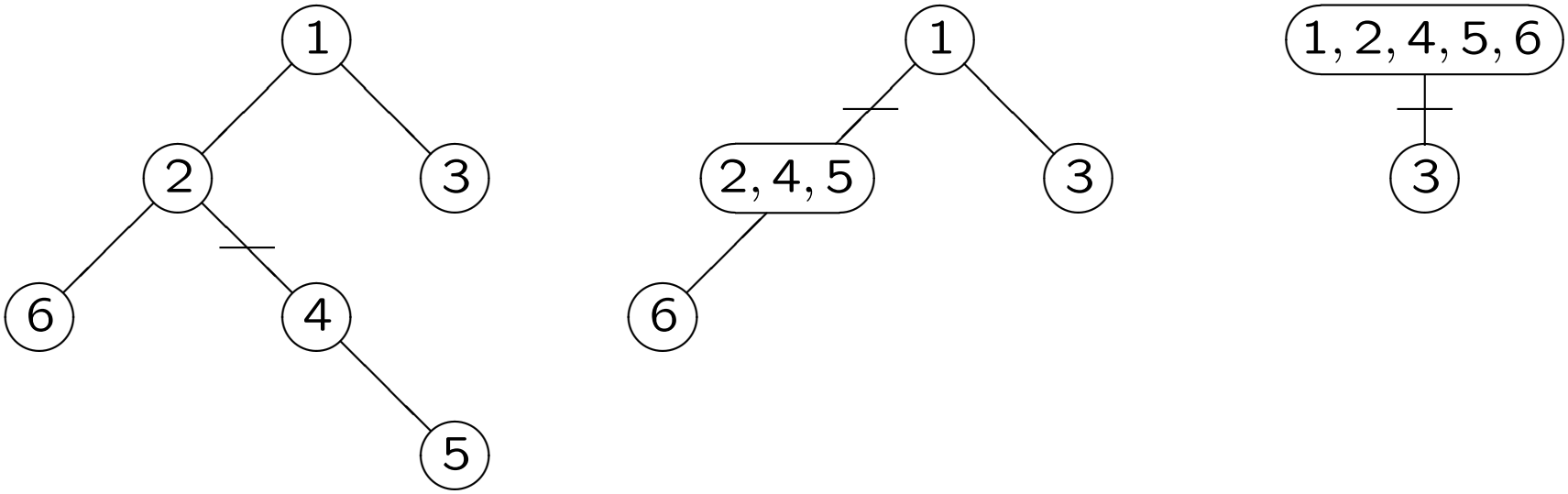
There are $(n-1)!$ recursive trees. To construct a random recursive tree, attach k to one of the previous $k-1$ vertices uniformly at random.



Cutting procedure (Meir and Moon, 1974): Pick an edge at random, and delete it along with the subtree below it. What remains is a random recursive tree on the new label set.

Connection with Bolthausen-Sznitman coalescent

Theorem (Goldschmidt and Martin, 2005): Cut each edge at the time of an exponential(1) random variable, and add the labels below the cut to the vertex above. The labels form a partition of $\{1, \ldots, n\}$ which evolves as a Bolthausen-Sznitman coalescent.



Proof idea: Given $\ell_1 < \cdots < \ell_k$, there are $(k-2)!$ recursive trees involving $\ell_2, \ldots, \ell_k$ and $(n-k)!$ recursive trees on the remaining vertices. The probability that $\ell_1, \ldots, \ell_k$ could merge is

$$\frac{(k-2)!(n-k)!}{(n-1)!} = \int_0^1 x^{k-2}(1-x)^{n-k}\, dx = \lambda_{n,k}.$$

Time back to MRCA

Theorem (Goldschmidt and Martin, 2005): Let T_n be the time back to the MRCA for the Bolthausen-Sznitman coalescent. For all $x \in \mathbb{R}$,

$$\lim_{n\to\infty} P\Big(T_n - \log\log n \leq x\Big) = e^{-e^{-x}}.$$

Proof idea: The last cut must involve one of the edges attached to the root. Because there are approximately

$$\sum_{k=2}^{n} \frac{1}{k-1} \approx \log n$$

such edges, the time back to the MRCA behaves like the maximum of $\log n$ exponential(1) random variables. By extreme value theory, the mean is approximately $\log\log n$, and the asymptotic distribution is the Gumbel distribution.

Remark: The Bolthausen-Sznitman just barely stays infinite.

The one-dimensional distributions

Build a random recursive tree. Mark edges whose random variable is less than t, which has probability $1 - e^{-t}$. Marked edges are cut before time t.

An integer n is the smallest integer in its block at time t if and only if there are no marked edges on the path from root to n.

Suppose $1, \ldots, n$ are in k blocks of sizes $n_1, \ldots, n_k$. Then:

- The integer $n + 1$ starts a new block if it attaches to one of the k integers that is the smallest in its block, and if the edge is unmarked. The probability is ke^{-t}/n.
- The integer $n + 1$ joins the block of size n_i if it attaches to one of the $n_i - 1$ integers other than the smallest in the block, or if it attaches to the smallest integer in the block and the edge is marked. The probability is $(n_i - e^{-t})/n$.

Chinese restaurant process with $\alpha = e^{-t}$ and $\theta = 0$. Distribution of the asymptotic block frequencies is Poisson-Dirichlet $(e^{-t}, 0)$.

Number of Segregating Sites

Theorem (Drmota, Iksanov, Möhle, and Rösler, 2007): For the Bolthausen-Sznitman coalescent with mutations at rate θ, as $n \to \infty$,

$$\frac{(\log n)^2}{\theta n}\left(S_n - \frac{\theta n}{\log n} - \frac{\theta n \log\log n}{(\log n)^2}\right) \Rightarrow X,$$

where X has a stable law of index 1. Using γ to denote Euler's constant,

$$\begin{aligned} E[e^{iuX}] &= \exp\left(-\frac{\pi}{2}|u| + iu\log|u|\right) \\ &= \exp\left(iu(1-\gamma) - \int_{-\infty}^{0}\left(1 - e^{iux} + iux\mathbf{1}_{\{|x|\le 1\}}\right)x^{-2}\,dx\right). \end{aligned}$$

Proof Idea: Relate total tree length L_n to the number of mergers before all integers are in one block, which is the number of cuts needed to reduce a random recursive tree down to one vertex. Was studied by Panholzer (2004).

Allele Frequency Spectrum

Theorem (Basdevant and Goldschmidt, 2008): For the Bolthausen-Sznitman coalescent, as $n \to \infty$,

$$\frac{\log n}{n} N_{1,n} \to_p \theta$$

and for $k \geq 2$,

$$\frac{(\log n)^2}{n} N_{k,n} \to_p \frac{\theta}{k(k-1)}.$$

Remarks:

- The fraction of blocks in the allelic partition that have size one tends to one as $n \to \infty$.
- It follows that

$$\frac{\log n}{n} K_n \to_p \theta.$$

- The same results should hold for the site frequency spectrum.

Segregating sites $(1 < \alpha < 2)$

Theorem (Berestycki, Berestycki, and Schweinsberg (2008)): For the Beta$(2-\alpha, \alpha)$-coalescent with $1 < \alpha < 2$ and mutations at rate θ,

$$\frac{S_n}{n^{2-\alpha}} \to_p \frac{\theta\alpha(\alpha-1)\Gamma(\alpha)}{2-\alpha}.$$

Limiting distribution of S_n obtained by Kersting (2011):

stable law of index α	if $1 < \alpha < \sqrt{2}$
mixture	if $\alpha = \sqrt{2}$
normal	if $\sqrt{2} < \alpha < 2$

The phase transition at $\alpha = \sqrt{2}$ was conjectured by Delmas, Dhersin, and Siri-Jegousse (2008).

A generalization

Theorem (Berestycki, Berestycki, and Limic (2012)): Suppose the Λ-coalescent comes down from infinity. Let

$$\Psi(\lambda) = \int_0^1 (e^{-\lambda x} - 1 + \lambda x) x^{-2} \, \Lambda(dx).$$

Then

$$\frac{S_n}{\int_1^n q \Psi(q)^{-1} \, dq} \to_p \theta,$$

and same result holds with K_n in place of S_n. If $\Lambda(dx) = f(x) \, dx$ with $f(x) \sim C x^{-1-\alpha}$ as $x \to 0$, then the convergence holds a.s.

For the Beta$(2-\alpha, \alpha)$ coalescent with $1 < \alpha < 2$, we recover the result of Berestycki, Berestycki, and Schweinsberg (2008), but with almost sure convergence both for S_n and for K_n.

Proof idea

Let $\lambda_b =$ total merger rate when b lineages.
Let $G_n(b) = P$(there are exactly b lineages at some time).

$$E[S_n] = \theta \sum_{b=2}^{n} b\lambda_b^{-1} G_n(b).$$

Total merger rate when there are b lineages is

$$\lambda_b = \sum_{k=2}^{b} \binom{b}{k} \lambda_{b,k} \sim \frac{1}{\alpha\Gamma(\alpha)} b^{\alpha}.$$

When there are b lineages, the expected number of lineages that are lost after the next merger converges to $1/(\alpha - 1)$ as $b \to \infty$ (Bertoin and Le Gall, 2005).

A renewal argument gives $G_n(b) \approx \alpha - 1$ for large n and b.

$$E[S_n] \approx \theta\alpha(\alpha - 1)\Gamma(\alpha) \int_0^n x^{1-\alpha}\, dx = \theta \frac{\alpha(\alpha - 1)\Gamma(\alpha)}{2 - \alpha} n^{2-\alpha}.$$

Site and allele frequency spectrum ($1 < \alpha < 2$)

Theorem (Berestycki, Berestycki, and Schweinsberg (2007)): For the Beta$(2-\alpha, \alpha)$-coalescent with $1 < \alpha < 2$, we have

$$\frac{M_{k,n}}{S_n} \to_p \frac{(2-\alpha)\Gamma(k+\alpha-2)}{\Gamma(\alpha-1)k!} = a_k$$

and $N_{k,n}/K_n \to_p a_k$.

Remarks:

- $a_1 = 2 - \alpha$, so can estimate α by $2 -$ fraction of singletons.
- $a_k \sim Ck^{\alpha-3}$, so smaller α means more low frequency mutants.
- Kingman's coalescent: $E[M_{k,n}] = \theta/k$ corresponds to $\alpha = 2$. Bolthausen-Sznitman: $N_{k,n} \approx C\theta/[k(k-1)]$ matches $\alpha = 1$.
- Berestycki, Berestycki, and Limic (2012) proved almost sure convergence.
- Original proof used connections with CSBP.

The case $\alpha < 1$

Theorem (Möhle, 2006): Suppose

$$\int_0^1 x^{-1}\Lambda(dx) < \infty,$$

so the Λ-coalescent does not have proper frequencies. Then if mutations occur at rate θ,

$$\frac{S_n}{\theta n} \Rightarrow S,$$

where

$$S =_d \int_0^\infty e^{-X(t)}\, dt$$

and $(X(t), t \geq 0)$ is a subordinator with zero drift whose Lévy measure is the image of the measure $x^{-2}\Lambda(dx)$ under the map $x \mapsto -\log(1-x)$.

This result includes the Beta$(2-\alpha, \alpha)$-coalescent for $0 < \alpha < 1$.

Abraham and Delmas (2012) gave a combinatorial construction of Beta$(3/2, 1/2)$-coalescent by pruning a random binary tree.

Example: Pacific Oyster

Data on 141 Pacific Oysters from British Columbia.
Data from Boom, Boulding, and Beckenbach (1994).
Analyzed by Sargsyan and Wakeley (2008).

There were 48 segregating sites
$M_{1,n} = 29$, $M_{2,n} = 12$, $M_{3,n} = 4$, $M_{6,n} = 2$, and $M_{67,n} = 1$.

Predictions with Kingman's coalescent: estimate θ by

$$\widehat{\theta} = 48 \Big/ \sum_{j=1}^{140} \frac{1}{j} \approx 8.7.$$

Then predict $M_{k,n}$ to be $\widehat{\theta}/k$.

Predictions with beta coalescent: predict $M_{k,n} = 48a_k$. Choose the α that gives the best fit to the data.

Comparision of predictions from Kingman's coalescent and from the beta coalescent with $\alpha = 1.35$.

Site Frequency Spectrum

k	Observed	Kingman	beta
1	29	8.7	31.2
2	12	4.3	5.5
3	4	2.9	2.5
4	0	2.2	1.4
5	0	1.7	1.0
6	2	1.4	0.7
7+	1	26.7	5.7

Neither fit is good. The fit from the beta coalescent is better.

Example: Atlantic Cod

Data on 1278 Atlantic Cod, segment 250 base pairs long.
Data from Arnason (2004).
Analyzed by Birkner and Blath (2007).
There were 59 haplotypes.
Estimate $\alpha = 1.43$ for the beta coalescent.

Allele Frequency Spectrum

k	Observed	Kingman	beta
1	32	7.6	33.6
2	7	3.8	7.2
3	6	2.5	3.4
4	2	1.9	2.1
5	3	1.5	1.4
6	1	1.3	1.0
7	1	1.1	0.8
8+	7	39.2	9.3

Statistical analysis in Birkner and Blath (2007) allows one to reject the Kingman's coalescent hypothesis. However, it is not possible to estimate α precisely.

Limitations to this analysis

1. Violations of assumptions. For example, the Atlantic Cod data had only 39 segregating sites, but 59 haplotypes.

2. (Durrett, Huerta-Sanchez): It seems that

$$\frac{M_{k,n}}{S_n} = a_k + O\left(\frac{1}{\log n}\right),$$

   so the a_k are not precise for finite values of n.

3. Different coalescent processes can lead to similar values for the site frequency spectrum and allele frequency spectrum. It is difficult to distinguish the effects of large family sizes from the effects of changing population size.

Block sizes of exchangeable random partitions

Let Π be an exchangeable random partition of $\mathbb{N}$.

Let $\Pi_n = R_n\Pi$ be the restriction of Π to $\{1, \ldots, n\}$.

Let K_n be the number of blocks of Π_n, and let $N_{k,n}$ be the number of blocks of size k.

Theorem (Karlin, 1967; Gnedin, Hansen, and Pitman, 2007): Suppose $1 < \alpha < 2$. If $K_n/n^{2-\alpha} \to c > 0$ a.s., then

$$\frac{N_{k,n}}{K_n} \to a_k = \frac{(2-\alpha)\Gamma(k+\alpha-2)}{\Gamma(\alpha-1)k!} \text{ a.s.}$$

Remarks:

- (Schweinsberg, 2010) If instead $K_n/n^{2-\alpha} \to_p c > 0$, then we can conclude $N_{k,n}/K_n \to_p a_k$.
- This result implies our asymptotic result for the allele frequency spectrum of the beta coalescent with $1 < \alpha < 2$.

Increasing Population Size (Schweinsberg, 2010)

Consider the following population model:

- There are N individuals in generation zero, and for $t \in \mathbb{N}$, there are $\lceil Nt^{-\gamma} \rceil$ individuals in generation $-t$, where $\gamma > 0$.
- Each member of the population has one parent, chosen at random from the individuals in the previous generation.

Sample n individuals in generation zero. Define the ancestral process Ψ_N as before. Then

$$(\Psi_N(\lfloor N^{1/(1+\gamma)} t \rfloor), t \geq 0) \Rightarrow (\Pi(t), t \geq 0),$$

where $(\Pi(t), t \geq 0)$ is a time-changed Kingman's coalescent in which at time t, each pair of lineages is merging at rate t^γ.

Let $\alpha = (2+\gamma)/(1+\gamma) \in (1,2)$, then

$$\frac{K_n}{n^{2-\alpha}} \to_p \frac{\theta 2^{\alpha-1}(\alpha-1)^{2-\alpha}\pi}{\sin(\pi(2-\alpha))}.$$

Thus, $N_{n,k}/K_n \to_p a_k$.

An Introduction to Mathematical Population Genetics and Coalescent Processes

Part IV: Natural Selection

by Jason Schweinsberg

University of California at San Diego

A model of selection

Consider the following modification of the Moran model:

- The population has fixed size $2N$.
- At time zero, there is a beneficial mutation on one chromosome, so $2N-1$ chromosomes have the b allele and one has the advantageous B allele.
- Each individual independently lives for an Exponential(1) time, then is replaced by a new individual chosen uniformly at random from the population.
- A replacement of a B by a b is rejected with probability s, in which case there is no change to the population.

We assume that $s > 0$ is a fixed constant (strong selection), though one could also allow s to tend to zero as $N \to \infty$.

A birth and death process

Let $X(t)$ be the number of individuals with the B allele at time t. Then $(X(t), t \geq 0)$ is a continuous-time birth and death process.

When $X(t) = k$,

- Birth rate: $b_k = (2N - k)\left(\frac{k}{2N}\right) = \frac{k(2N-k)}{2N}$.
- Death rate: $d_k = (k)\left(\frac{2N-k}{2N}\right)(1-s) = \frac{k(2N-k)(1-s)}{2N}$.

We have $X(0) = 1$. Let $\tau = \inf\{t : X(t) = 0 \text{ or } X(t) = 2N\}$.

If $X(\tau) = 0$, then the B allele disappears.

If $X(\tau) = 2N$, then eventually all $2N$ chromosomes have the B allele. This is called a *selective sweep*.

The probability of a selective sweep

Let $h(k)$ be the probability that a selective sweep occurs when k individuals have the B allele.

$$b_k = \frac{k(2N-k)}{2N}, \qquad d_k = \frac{k(2N-k)(1-s)}{2N}.$$

When $X(t) = k$, the next event is a birth with probability $1/(2-s)$ and a death with probability $(1-s)/(2-s)$, so

$$h(k) = \frac{1}{2-s}\, h(k+1) + \frac{1-s}{2-s}\, h(k-1)$$

with $h(0) = 0$ and $h(2N) = 1$. Solving the recursion gives

$$h(k) = \frac{1-(1-s)^k}{1-(1-s)^{2N}}.$$

Therefore, the probability of a selective sweep is

$$h(1) = \frac{s}{1-(1-s)^{2N}} \approx s.$$

If $s = 0$, the probability of fixation is $1/N$.

The path to fixation

Let $Y(t) = X(t)/2N$ be the fraction of the population with B.

When $X(t) = k$, the rate at which $Y(t)$ is increasing is:

$$\frac{b_k - d_k}{2N} = \frac{sk(2N-k)}{(2N)^2} = sY(t)(1-Y(t)).$$

We can approximate $(Y(t), t \geq 0)$ by the solution to the logistic differential equation

$$\frac{d}{dt}Y(t) = sY(t)(1-Y(t)), \qquad Y(0) = \frac{1}{2N}.$$

We get

$$Y(t) \approx \frac{1}{1 + (2N-1)e^{-st}}.$$

The duration of a selective sweep

Theorem (Kimura and Ohta, 1969): As $N \to \infty$, we have

$$E[\tau | X(\tau) = 2N] \sim \frac{2}{s} \log N.$$

Proof Idea: One can use the approximation

$$Y(t) \approx \frac{1}{1 + (2N-1)e^{-st}}$$

to see that $Y(t) = 1 - 1/2N$ when

$$t = \frac{2}{s} \log(2N-1) \sim \frac{2}{s} \log N.$$

Alternatively, one can use random walk calculations to compute, for $1 \leq k \leq 2N-1$, the expected duration of time for which $X(t) = k$, then sum over k.

With the ordinary Moran model, the time for two lineages to coalesce is $O(N)$. Lineages that coalesce during a selective sweep do so almost instantaneously on the $O(N)$ time scale.

Recombination

Individuals may inherit pieces of each of a parent's two chromosomes. Consider a site on the chromosome nearby where a beneficial mutation occurs.

Suppose one site has a B or b allele, B advantageous.

The site of interest has an A or a allele, neither is advantageous.

Initially there is just one B.

When a new individual is born:

- The B or b comes from a randomly chosen parent. (Change from B to b is rejected with probability s.)
- With probability $1-r$, the A or a comes from the same parent.
- With probability r, the A or a allele comes from a parent chosen independently at random.

The genealogy at the A/a site

Sample n individuals at the time τ when a selective sweep ends.

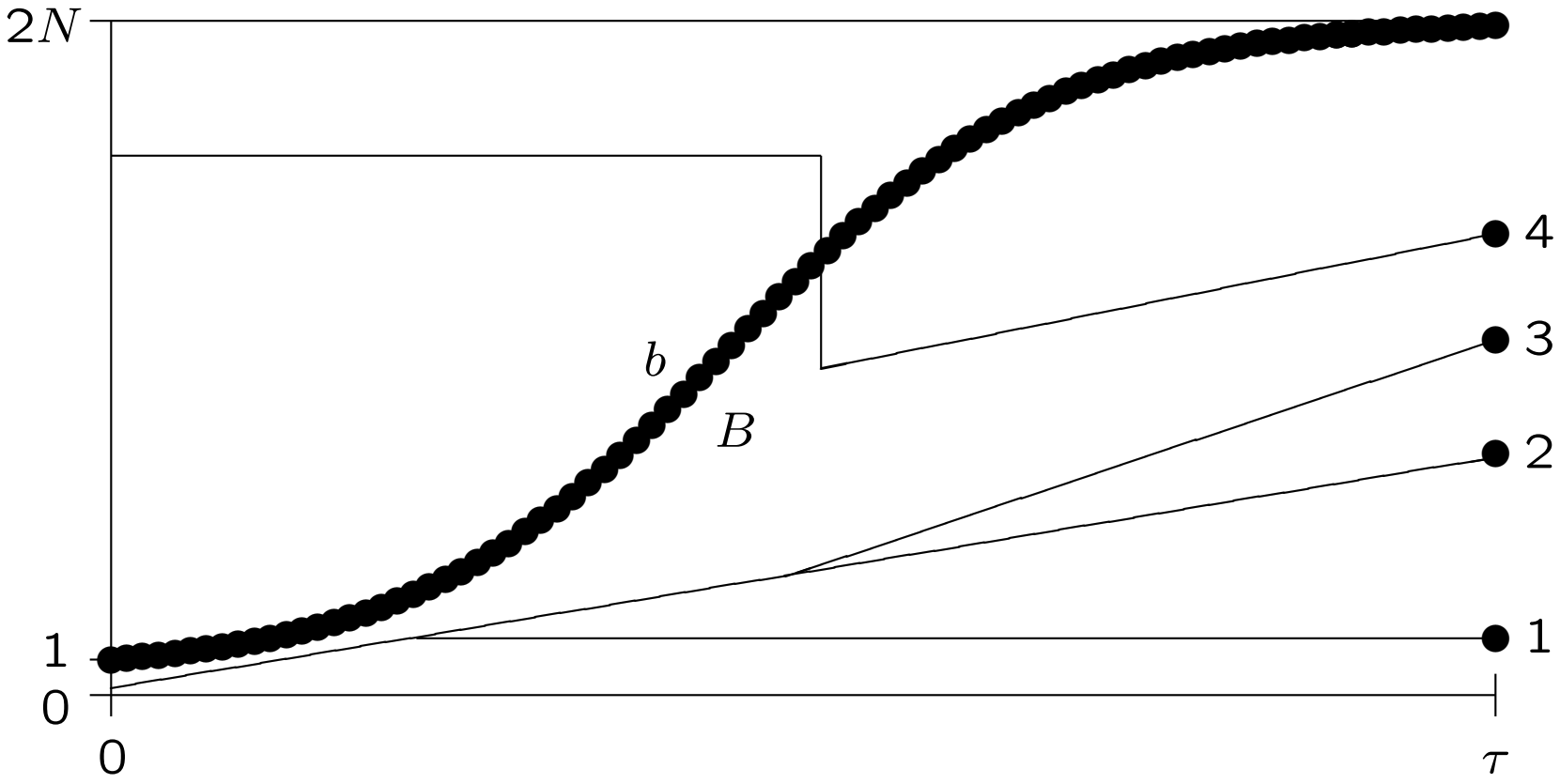
All n individuals in the sample inherited their B allele from the same individual at time 0.

Let Θ be a random partition of $\{1, \ldots, n\}$ such that $i \sim_\Theta j$ if and only if the ith and jth sampled individuals inherited their A/a allele from the same individual at time zero.

Goal: to describe the distribution of the random partition Θ.

Previous work: Maynard Smith and Haigh (1974); Kaplan, Hudson, and Langley (1989); Stephan, Wiehe, and Lenz (1992); Barton (1998, 2000).

Illustration of a selective sweep



$$\Theta = \{\{1, 2, 3\}, \{4\}\}.$$

If the A/a allele of one individual comes from an individual that had the b allele at time zero, we say the lineage *escapes* the selective sweep.

Estimating the probability p of failing to escape

There is a small probability that a given lineage is affected by recombination each time there is a change in the population.

The probability that a lineage escapes the sweep at a time when the number of B individuals goes from k to $k+1$ is:

$$\frac{1}{k+1} \cdot r \cdot \frac{2N-k}{2N}.$$

Make similar calculations for when the number of B individuals goes from k to $k-1$ or k, calculate expected number of such changes to get

$$p \approx \exp\left(-\frac{r}{s} \log(2N) \right).$$

Probability of two recombinations is $O(1/(\log N)^2)$.
Probability of coalescence, then recombination is $O(1/(\log N))$.
Probability of recombination, then coalescence is $O((\log N)^2/N)$.

If $A_1, \ldots, A_n$ are the events that n lineages escape the sweep, then $A_1, \ldots, A_n$ are approximately independent for large N.

A simple approximation

Define a random partition Θ_p of $\{1, \ldots, n\}$ as follows:

- Flip n independent coins with probability p of heads.
- One block of Θ_p is $\{i : \text{the } i\text{th coin is heads}\}$.
- The other blocks are singletons.

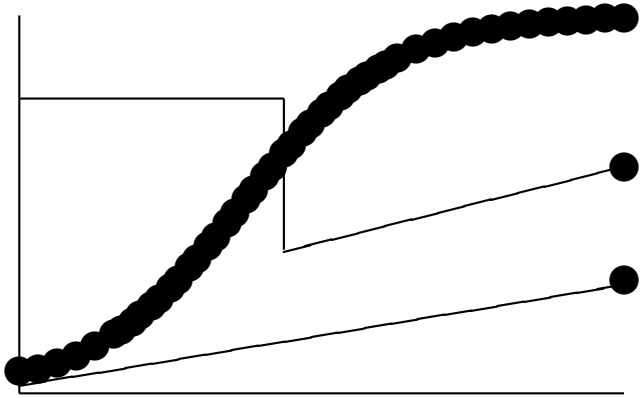
Theorem (Schweinsberg and Durrett, 2005): Suppose s is constant and $r \sim c/(\log N)$ for some constant c. Let $a = r \log(2N)/s$. Let $p = e^{-a}$. Then there exists a positive constant C such that

$$|P(\Theta = \pi) - P(\Theta_p = \pi)| \leq \frac{C}{\log N}$$

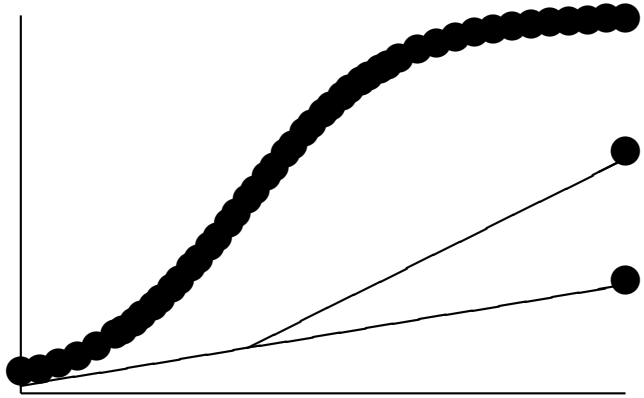
for all N and all partitions π of $\{1, \ldots, n\}$.

Simulations

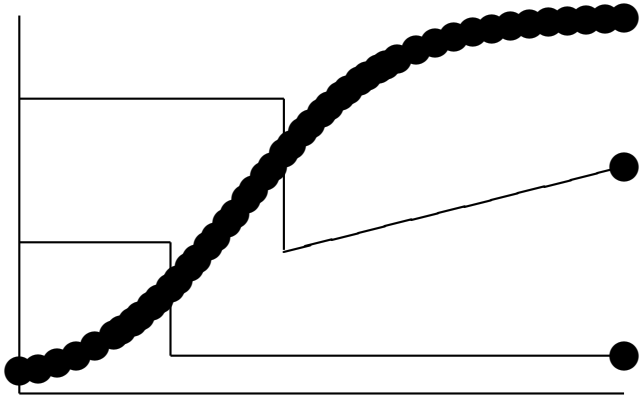
Keep track of the fraction of lineages that escape the sweep.
Also, we have the following possibilities for two lineages:



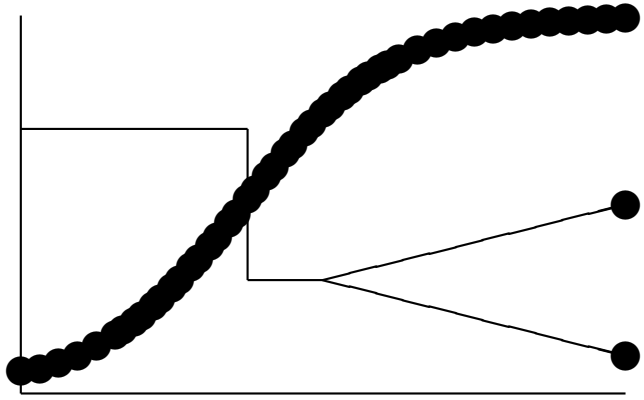
$B - b$



BB



$b - b$



bb

Simulation results

Choose r so that $1 - e^{-a} = 0.4$, where $a = r \log(2N)/s$.

N = 10,000; s = 0.03	b	B–b	BB	bb	b–b
simulations	.295	.303	.553	.067	.077
approximation	.400	.480	.360	.000	.160

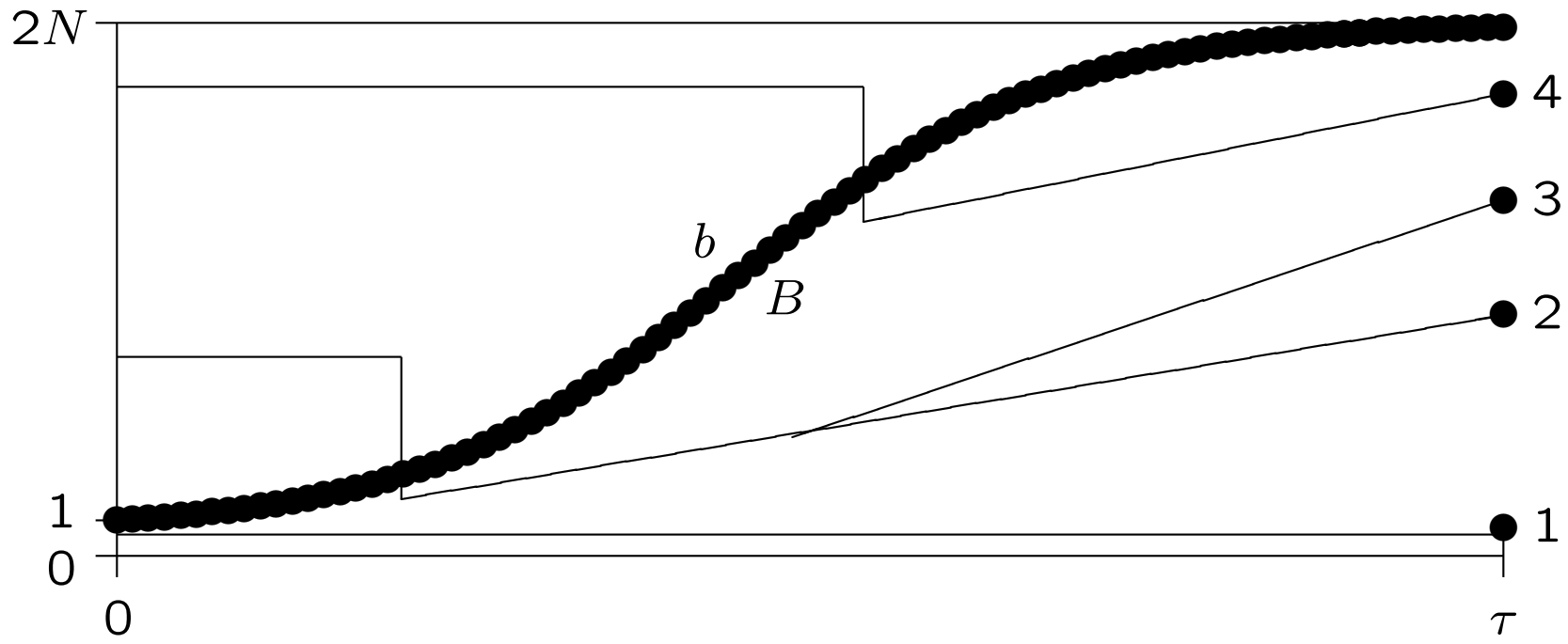
N = 100,000; s = 0.03	b	B–b	BB	bb	b–b
simulations	.318	.352	.505	.046	.096
approximation	.400	.480	.360	.000	.160

N = 1,000,000; s = 0.01	b	B–b	BB	bb	b–b
simulations	.308	.355	.515	.039	.091
approximation	.400	.480	.360	.000	.160

Approximation is poor, error $O(1/\log N)$.

Dominant source of error

A recombination soon after the beneficial mutation may cause several lineages that have already coalesced to be descended from the same individual in the b population (Barton, 1998).



Then Θ has more than one large block:

- One corresponds to lineages that fail to escape
- Others correspond to groups of lineages that coalesce and escape near the beginning.

The beginning of a selective sweep

The recombinations that cause additional large blocks in $\ominus$ are those that occur when the number of B's is small.

When the B-population is small, it is approximately a continuous-time branching process in which each individual dies at rate $1-s$ and gives birth at rate 1.

The number of lineages with an infinite line of descent is a branching process with no deaths and births at rate s.

Define $0 = \tau_1 < \tau_2 < \ldots$ such that τ_k is the first time at which there are k individuals with an infinite line of descent.

If there is recombination along a lineage with an infinite line of descent between times τ_k and τ_{k+1}, descendants of that lineage will have a different ancestor at the beginning of the sweep than descendants of the other $k-1$ lineages.

What fraction of the population is descended from this lineage?

A connection with Polya urns

Color the B lineage that gets the recombination red, and the other $k-1$ lineages blue.

When a lineage splits into two, give the new lineage the same color as the parent.

When there are x red lineages and y blue lineages,

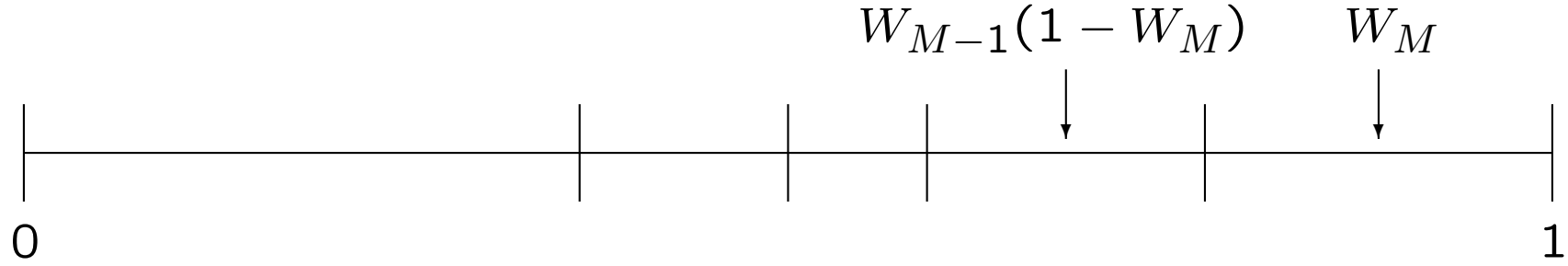
$$P(\text{next individual is red}) = \frac{x}{x+y}.$$

These are the same dynamics as the Polya urn, started with 1 red ball and $k-1$ blue balls.

The long-run fraction of red individuals (descended from the lineage with the recombination) has a Beta$(1, k-1)$ distribution.

Stick-breaking construction

Stick-breaking (paintbox) construction:



Let $M = \lfloor 2Ns \rfloor$. For $k = M, M-1, M-2, \ldots, 3, 2$, we break off a fraction W_k of the interval that is left.

W_k corresponds to the fraction of lineages that escape the sweep between times τ_k and τ_{k+1}.

Expected number of recombinations between τ_k and τ_{k+1} is r/s. Assume the number is 0 or 1.

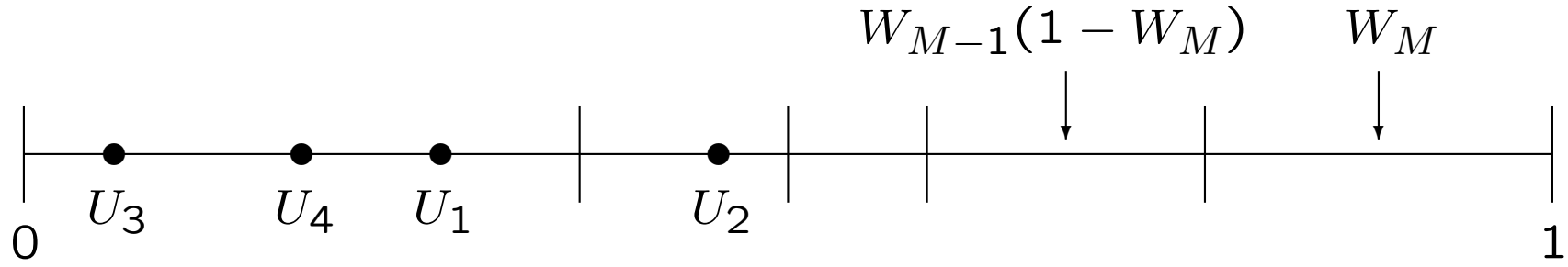
With probability r/s, W_k has the Beta$(1, k-1)$ distribution.

With probability $1 - r/s$, $W_k = 0$.

A second approximation

Let $U_1, U_2, \ldots, U_n$ be i.i.d. with the uniform distribution on $[0, 1]$.

Let Π be the random partition of $\{1, \ldots, n\}$ such that $i \sim_\Pi j$ if and only if U_i and U_j are in the same subinterval.



Example: $\Pi = \{\{1, 3, 4\}, \{2\}\}$.

Theorem (Schweinsberg and Durrett, 2005): If s is constant and $r \sim c/\log N$, then there exists a constant C such that for all N and all partitions π of $\{1, \ldots, n\}$, we have

$$|P(\Theta = \pi) - P(\Pi = \pi)| \leq \frac{C}{(\log N)^2}.$$

Simulation results

Choose r so that $1 - e^{-a} = 0.4$, where $a = r \log(2N)/s$.

N = 10,000; s = 0.03	b	B–b	BB	bb	b–b
simulations	.295	.303	.553	.067	.077
approximation	.301	.318	.540	.059	.082

N = 100,000; s = 0.03	b	B–b	BB	bb	b–b
simulations	.318	.352	.505	.046	.096
approximation	.321	.358	.501	.044	.098

N = 1,000,000; s = 0.01	b	B–b	BB	bb	b–b
simulations	.308	.355	.515	.039	.091
approximation	.308	.358	.513	.038	.091

The stick-breaking approximation works much better than the coin tossing approximation.

Other Time Scales

1. The previous theorems hold for "strong selection" when the selective advantage s is $O(1)$.

2. One can also consider "weak selection" when s is $O(1/N)$. There is diffusion limit, studied by Krone and Neuhauser (1997); Donnelly and Kurtz (1999); Barton, Etheridge, and Sturm (2004).

3. Etheridge, Pfaffelhuber, and Wakolbinger (2006) show that same approximations work in the diffusion limit, if we set $s = \alpha/N$ and then let $\alpha \to \infty$.

Recurrent selective sweeps

Gillespie (2000) proposed that selective sweeps happen at times of a Poisson process.

If selective sweeps happen at rate $O(N^{-1})$, then the ancestral processes converge to a coalescent with multiple mergers.

A better approximation for finite N can be obtained using a coalescent with simultaneous multiple mergers.

Limiting coalescents (Durrett and Schweinsberg, 2005):

- No selection: $\Lambda = \delta_0$ (Kingman's coalescent).
- If the mutations all occur at the same site, then we get $\Lambda = \delta_0 + \alpha p^2 \delta_p$.
- If mutations and recombinations occur uniformly along the chromosome, then $\Lambda(dx) = \delta_0 + \beta x \, dx$.
- Other Λ could arise under different assumptions.

Another population model with selection

Brunet, Derrida, Mueller, Munier (2006, 2007)

- Population has fixed size N.
- Each individual has $k \geq 2$ offspring.
- The fitness of each offspring is the parent's fitness plus an independent random variable with distribution μ.
- Of the kN offspring, the N with the highest fitness survive to form the next generation.

Durrett and Mayberry (2009) studied related model in context of predator-prey systems.

Related work: Bérard and Gouéré (2010); Durrett and Remenik (2011).

Three conjectures

Brunet, Derrida, Mueller, and Munier (2006, 2007)

1. (Brunet and Derrida, 1997) Let L_m be the maximum of the fitnesses of the individuals in generation m. Then $L_m/m \to v_N$ a.s. Let $v_\infty = \lim_{N\to\infty} v_N$. Then
$$v_\infty - v_N \sim \frac{C}{(\log N)^2}.$$
2. If two individuals are chosen from some generation, then the number of generations back to their most recent common ancestor is $O((\log N)^3)$.
3. If n individuals are sampled from some generation, then the coalescence of the ancestral lineages is governed by the Bolthausen-Sznitman coalescent.

Theorem (Bérard and Gouéré, 2010): Conjecture 1 holds in the case $k = 2$ under suitable regularity conditions on μ.

Goal: prove rigorous versions of Conjectures 2 and 3.

Branching Brownian motion with absorption

- Begin with some configuration of particles in $(0, \infty)$.
- Each particle independently moves according to standard one-dimensional Brownian motion with drift $-\mu$.
- Each particle splits into two at rate 1.
- Particles are absorbed if they reach the origin.

Interpretation: particles represent individuals in a population
branching events represent births
positions of particles represent fitnesses
absorption models deaths of unfit individuals

Theorem (Kesten, 1978): Starting with one particle at $x > 0$, this process dies out almost surely if $\mu \geq \sqrt{2}$. If $\mu < \sqrt{2}$, the number of particles grows exponentially with positive probability.

We take $\mu = \mu_N = \sqrt{2 - \dfrac{2\pi^2}{(\log N + 3 \log\log N)^2}}$.

The $O((\log N)^{-2})$ correction is related to Conjecture 1.

Notation

Let $M_N(t)$ be the number of particles at time t.

Let $X_{1,N}(t) \geq X_{2,N}(t) \geq \cdots \geq X_{M_N(t),N}(t)$ be the positions of the particles at time t.

Let $L = \frac{1}{\sqrt{2}}\left(\log N + 3 \log \log N \right)$.

Let $Y_N(t) = \sum_{i=1}^{M_N(t)} e^{\mu X_{i,N}(t)}$.

Let $Z_N(t) = \sum_{i=1}^{M_N(t)} e^{\mu X_{i,N}(t)} \sin\left(\frac{\pi X_{i,N}(t)}{L}\right) \mathbf{1}_{\{X_{i,N}(t) \leq L\}}$.

$Z_N(t)$ will be a useful measure of the "size" of the process at time t, disregarding particles to the right of L.

Main results

Theorem (Berestycki, Berestycki, and Schweinsberg (2010)): Suppose $Z_N(0)/[N(\log N)^2] \Rightarrow \nu$ and $Y_N(0)/[N(\log N)^3] \Rightarrow 0$. For some $a \in \mathbb{R}$, the finite-dimensional distributions of

$$\left(\frac{1}{2\pi N} M_N((\log N)^3 t), t > 0 \right)$$

converge to those of the CSBP with initial distribution ν and branching mechanism $\Psi(u) = au + 2\pi^2 u \log u$.

Theorem (Berestycki, Berestycki, and Schweinsberg (2010)): Let $t > 0$, and pick n particles at random at time $t(\log N)^3$. Let $\Pi_N(s)$ be the partition of $\{1, \ldots, n\}$ such that $i \sim_{\Pi_N(s)} j$ if and only if the ith and jth sampled particles have the same ancestor at time $(t - s/2\pi)(\log N)^3$. If the above assumptions hold and $\nu(\{0\}) = 0$, then the finite-dimensional distributions of $(\Pi_N(s), 0 \leq s \leq 2\pi t)$ converge to those of the Bolthausen-Sznitman coalescent.

The key heuristic

Brunet, Derrida, Mueller, and Munier (2006, 2007)

Occasionally, a particle gets very far to the right.

This particle has a large number of surviving descendants, as the descendants avoid the barrier at zero.

This leads to sudden jumps in the number of particles, and multiple mergers of ancestral lines.

The proof strategy

Find the level L such that a particle must reach L to give rise to a jump in the number of particles.

Show that the behavior of branching Brownian motion with particles killed at 0 and L is approximately deterministic.
(This is a "Law of Large Numbers" or "fluid limit" result that is proved by calculating first and second moments.)

Separately determine the (random) contribution of the particles that reach L.

Branching Brownian motion in a strip

Consider Brownian motion killed at 0 and L. If there is initially one particle at x, the "density" of the position at time t is:

$$q_t(x,y) = \frac{2}{L}\sum_{n=1}^{\infty} e^{-\pi^2 n^2 t/2L^2} \sin\left(\frac{n\pi x}{L}\right)\sin\left(\frac{n\pi y}{L}\right).$$

Add branching and drift of $-\mu$, "density" becomes:

$$p_t(x,y) = q_t(x,y)\cdot e^{\mu(x-y)-\mu^2 t/2}\cdot e^t,$$

meaning that if $B \subset (0,L)$, the expected number of particles in B at time t is

$$\int_B p_t(x,y)\,dy.$$

For $t \gg L^2$,

$$p_t(x,y) \approx \frac{2}{L} e^{(1-\mu^2/2-\pi^2/2L^2)t} e^{\mu x}\sin\left(\frac{\pi x}{L}\right) e^{-\mu y}\sin\left(\frac{\pi y}{L}\right).$$

Obervations related to density formula

$$p_t(x,y) \approx \frac{2}{L} e^{(1-\mu^2/2-\pi^2/2L^2)t} e^{\mu x} \sin\left(\frac{\pi x}{L}\right) e^{-\mu y} \sin\left(\frac{\pi y}{L}\right).$$

- When $1-\mu^2/2-\pi^2/2L^2 = 0$, the formula does not depend on t. We choose μ to satisfy this equation, to keep the number of particles relatively stable.
- Formula is proportional to $e^{\mu x}\sin(\pi x/L)$. Summing over multiple particles at time zero, this becomes $Z_N(0)$. Thus, $Z_N(0)$ determines how many particles will be in a given set at future times.
- With μ chosen as above, $(Z_N(t), t \geq 0)$ is a martingale.
- Formula is proportional to $e^{-\mu y}\sin(\pi y/L)$. For $t \gg (\log N)^2$, particles settle into a fairly stable limiting configuration.

A continuous-time branching process

Consider branching Brownian motion with drift $-\sqrt{2}$ started with one particle at L.

Let $M(y)$ be the number of particles that reach $L-y$, if particles are killed upon reaching $L-y$.

Conditional on $M(x)$, the distribution of $M(x+y)$ is the distribution of $M(x)$ independent random variables with the same distribution as $M(y)$. Therefore, $(M(y), y \geq 0)$ is a continuous-time branching process.

Maillard (2010) showed that the offspring distribution satisfies

$$\sum_{k=n}^{\infty} p_k \sim \frac{C}{n(\log n)^2}.$$

Offspring distribution has finite mean but is not in $L \log L$ class.

A limit theorem for the branching process

Theorem (Neveu, 1987): There exists a random variable W such that almost surely

$$\lim_{y\to\infty} ye^{-\sqrt{2}y}M(y) = W.$$

Furthermore, for all $u \in \mathbb{R}$,

$$E[e^{-e^{\sqrt{2}u}W}] = \psi(u),$$

where ψ satisfies Kolmogorov's equation $\frac{1}{2}\psi'' - \sqrt{2}\psi' = \psi(1-\psi)$.

Proposition: As $x \to \infty$, we have $P(W > x) \sim \dfrac{1}{x\sqrt{2}}$.

Proof Idea: Use a Tauberian theorem to reduce this to a problem about $E[e^{-\lambda W}]$ for small λ, and thus about asymptotics of $\psi(u)$ as $u \to -\infty$. Then follow Harris (1999) to obtain result from a property of the three-dimensional Bessel process.

Neveu's CSBP and Bolthausen-Sznitman

Waiting time for a particle to hit L is $O((\log N)^3)$.

Rate at which particles reach L is proportional to $Z_N(t)$.

If a particle hits L, its contribution is proportional to the number of descendants that hit $L - y$ for large y, and therefore is approximately proportional to W.

The probability that $Z_N(t)/(N(\log N)^2)$ jumps by at least x is approximately Cx^{-1}.

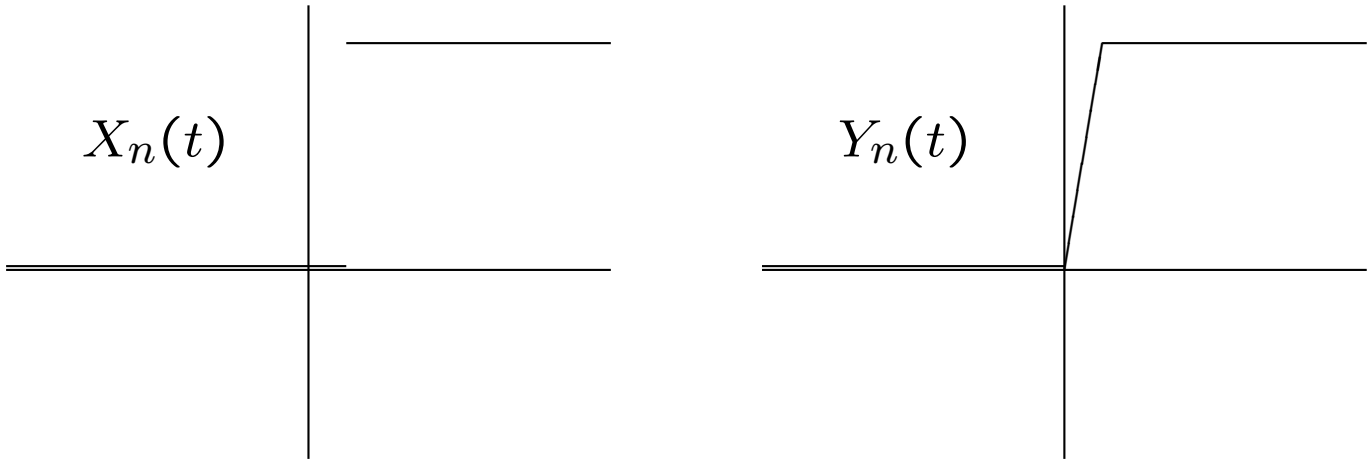
For any CSBP, the rate of jumps of size at least x is proportional to the value of the process.

For Neveu's CSBP, the rate of jumps of size at least x is proportional to $\int_x^\infty y^{-2}\,dy = x^{-1}$.

Types of Convergence

Convergence to Neveu's CSBP and to Bolthausen-Sznitman is convergence of finite-dimensional distributions, not weak convergence with respect to the Skorohod topology.

Let $X_n(t) = \begin{cases} 0 & \text{if } t < 1/n \\ 1 & \text{if } t \geq 1/n \end{cases}$ and $Y_n(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ nt & \text{if } 0 < t < 1/n \\ 1 & \text{if } t \geq 1/n \end{cases}$



Let $X(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 & \text{if } t \geq 0 \end{cases}$. Then $X_n \Rightarrow X$ but $Y_n \not\Rightarrow X$.

Conclusions

- To learn about the history of a population, it is often useful to work backwards in time, study the coalescent process that describes merging of ancestral lines.
- Under standard assumptions, the genealogy of a population can be described by Kingman's coalescent: only two lineages merge at a time.
- In populations with large family sizes or selection, there could be multiple mergers of ancestral lines.
- Beta coalescents describe genealogies of certain populations with large family sizes, Bolthausen-Sznitman coalescent describes genealogy of some populations undergoing selection.
- Large family sizes and selection could explain excess of low-frequency mutations in some data sets, but so could demographic factors.