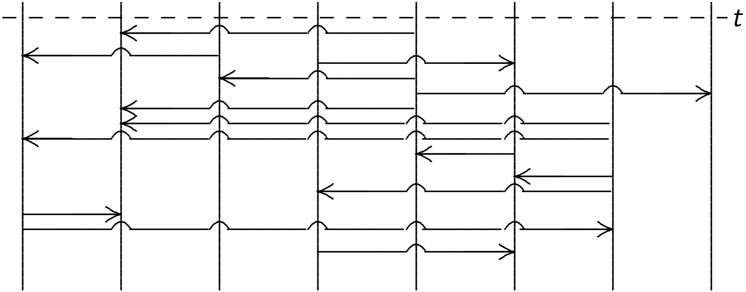
The tree-length of an evolving coalescent

Peter Pfaffelhuber
(joint with Anton Wakolbinger, Heinz Weisshaupt)

Berlin, July 2009

The Moran model

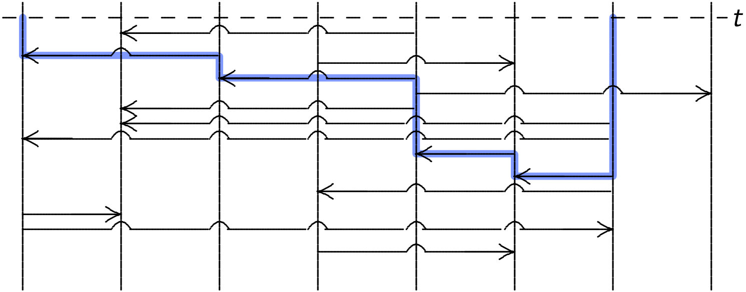
- A population consists of N individuals
- Each pair of individuals **resamples** at rate 1
- Resampling means: one individual dies, the other reproduces



The Moran model

- A population consists of N individuals
- Each pair of individuals **resamples** at rate 1
- Resampling means: one individual dies, the other reproduces

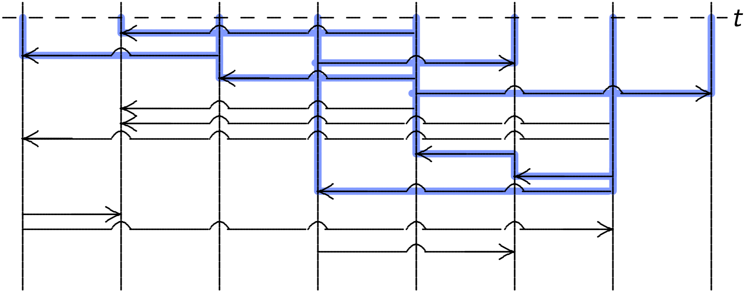
Ancestral lineages coalesce



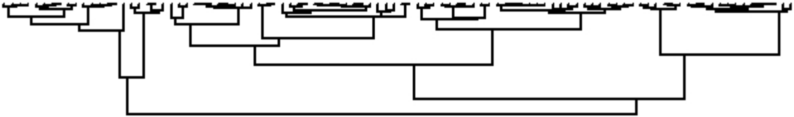
The Moran model

- A population consists of N individuals
- Each pair of individuals **resamples** at rate 1
- Resampling means: one individual dies, the other reproduces

Ancestral lineages coalesce



Kingman's coalescent



- **Genealogies** are given by **Kingman's $\mathbb{N}$-coalescent**
- Coalescence rate is $\binom{k}{2}$.
- What are properties of the **tree length**?

Kingman's coalescent

- $\mathscr{L}_t^{\mathrm{N}}$: tree length of N-coalescent at time t

$$\mathbb{E}[\mathscr{L}_t^{\mathrm{N}}] = \sum_{k=2}^{\mathrm{N}} k \frac{1}{\binom{k}{2}} \overset{\mathrm{N}\to\infty}{\approx} 2\log(\mathrm{N})$$

$$\mathbb{V}[\mathscr{L}_t^{\mathrm{N}}] = \sum_{k=2}^{\mathrm{N}} k^2 \frac{1}{\binom{k}{2}^2} \overset{\mathrm{N}\to\infty}{\approx} 4\frac{\pi^2}{6}$$

Kingman's coalescent

- $\mathscr{L}_t^{\mathrm{N}}$: tree length of N-coalescent at time t
- $\mathcal{E}(\cdot)$: independent exponential distributions

$$\frac{1}{2}\mathscr{L}_t^{\mathrm{N}} \stackrel{d}{=} \frac{1}{2}\sum_{k=2}^{\mathrm{N}} k \cdot \mathcal{E}\left(\binom{k}{2}\right) \stackrel{d}{=} \sum_{k=1}^{\mathrm{N}-1} \mathcal{E}(k) \stackrel{d}{=} \max_{1\le k\le \mathrm{N}-1} \mathcal{E}(1)$$

$$\mathbb{P}[\tfrac{1}{2}(\mathscr{L}_t^{\mathrm{N}} - 2\log(\mathrm{N})) \le t] = (1 - e^{-\log(\mathrm{N})-t})^{\mathrm{N}-1} \approx e^{-e^{t}}$$

- $\Rightarrow \frac{1}{2}(\mathscr{L}_t^{\mathrm{N}} - 2\log(\mathrm{N})) \xrightarrow{\mathrm{N}\to\infty}$ Gumbel

The Gumbel variable in the coalescent

- Are there **stronger versions** of tree length convergence on a coalescent?

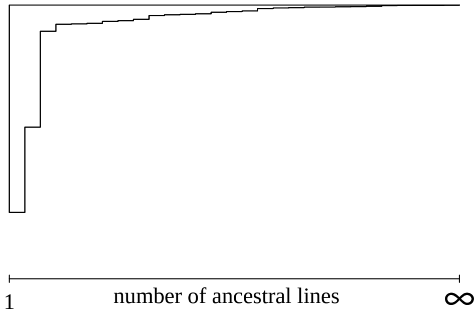
- Consider subtrees with N leaves of a full Kingman coalescent
- $X_i \overset{d}{=} \mathcal{E}\binom{i}{2}$: time full coalescent stays with i lines

$$L^1_{\mathrm{N}} = \sum_{i=2}^{\mathrm{N}} iX_i \qquad \textbf{Temporal coupling}$$

- K_i^{N}: # lines in N-tree while full tree has i lines

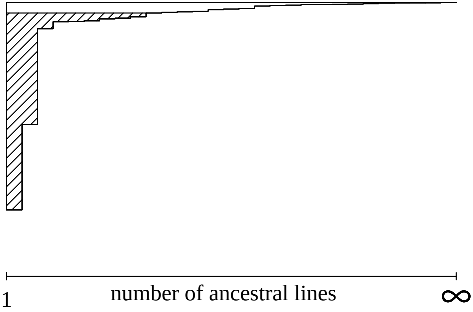
$$L^2_{\mathrm{N}} = \sum_{i=2}^{\infty} K_i X_i \qquad \textbf{Natural coupling}$$

The Gumbel variable in the coalescent



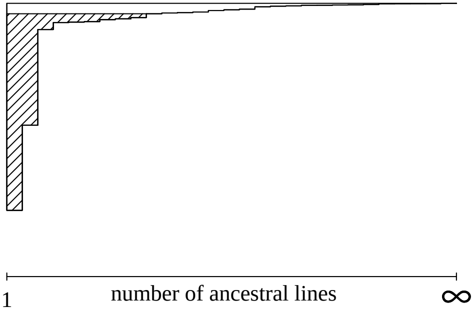
- The **full** Kingman coalescent

The Gumbel variable in the coalescent



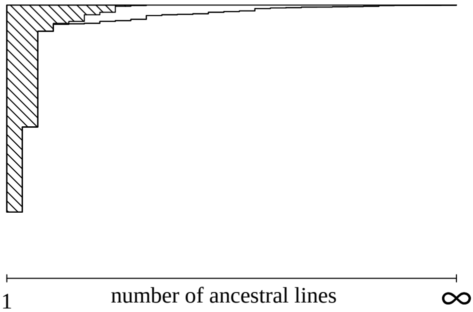
- $\mathscr{L}_t^{N,1} = \sum_{i=2}^{N} iX_i$
- X_i: time full coalescent stays with i lines

The Gumbel variable in the coalescent



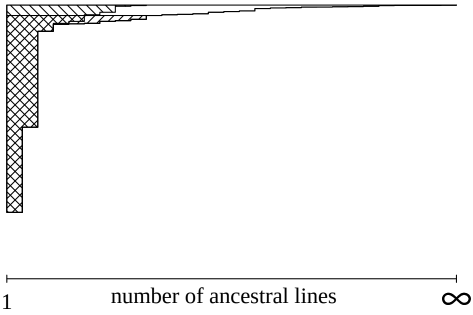
- $\frac{1}{2}(\mathscr{L}_t^{N,1} - 2\log N) \xrightarrow{N\to\infty} \mathscr{L}_t$ **almost surely and in L^2**
- $\mathscr{L}_t$: Gumbel distributed

The Gumbel variable in the coalescent



- K_i^{N}: # lines in N-tree while full tree has i lines
- $\mathscr{L}_t^{\mathrm{N},2} = \sum_{i=2}^{\infty} K_i X_i$

The Gumbel variable in the coalescent



- $\mathscr{L}_{\mathbf{t}}^{\mathbf{N,1}} - \mathscr{L}_{\mathbf{t}}^{\mathbf{N,2}} \xrightarrow{\mathbf{N \to \infty}} \mathbf{0}$ **in** L^2

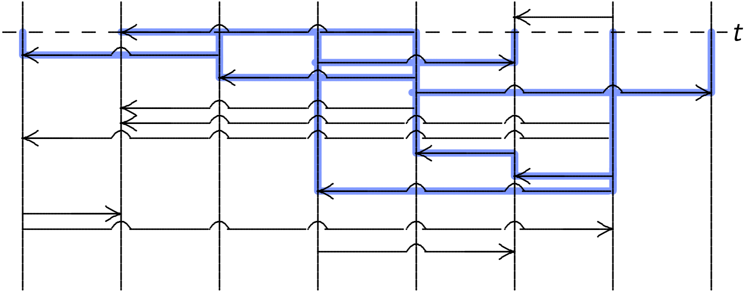
Sample path

- Genealogies evolve together with the population
- Show movie
- Rest of the talk:

  What does the evolution of tree lengths $\mathscr{L}_t^{\mathrm{N}}$ look like?

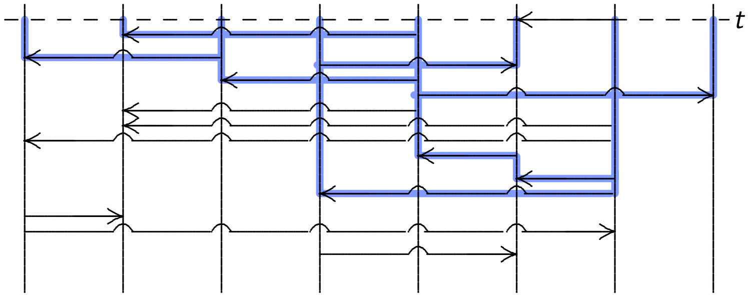
The Moran model

- Genealogies evolve as time proceeds
- Tree growth at speed $\mathrm{N}dt$ between resampling events



The Moran model

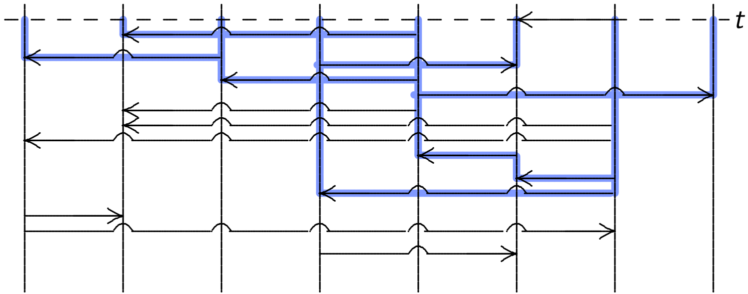
- Genealogies evolve as time proceeds
- Tree growth at speed $\mathrm{N}dt$ between resampling events



The Moran model

- Genealogies evolve as time proceeds
- Tree growth at speed $\mathrm{N}dt$ between resampling events
- At resampling times the tree length changes

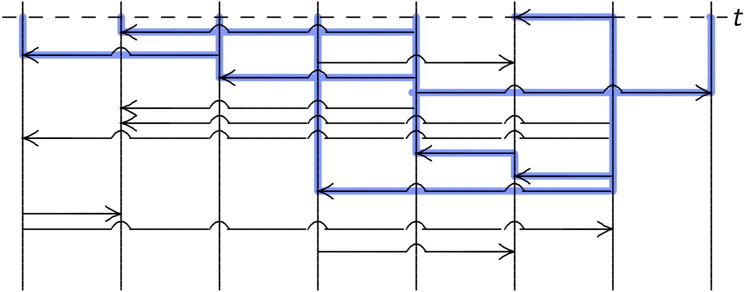
External branches break off



The Moran model

- Genealogies evolve as time proceeds
- Tree growth at speed $\mathrm{N}dt$ between resampling events
- At resampling times the tree length changes

External branches break off



Typical jumps

- F: jump time of N-coalescent
- J^{N}: length of a random external branch

$$\mathscr{L}_F^{\mathrm{N}} - \mathscr{L}_{F-}^{\mathrm{N}} \stackrel{d}{=} J^{\mathrm{N}}$$

- Fu, Li (1993); Durrett (2002); Caliebe, Neiniger, Krawczak, Rösler (2007)

$$\mathrm{N} \cdot J^{\mathrm{N}} \xrightarrow[]{\mathrm{N}\to\infty} J, \qquad \mathbb{E}[J] = 2, \quad \mathbb{V}[J] = \infty$$

- $\Rightarrow \binom{\mathrm{N}}{2}$ jumps of size $2/\mathrm{N}$ per time unit.

Main result

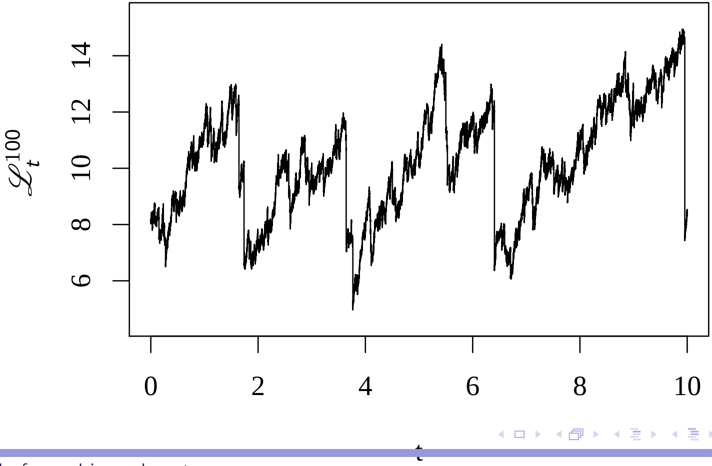
- **There is a** process $\mathscr{L} = (\mathscr{L}_t)_{t\in\mathbb{R}}$ with càdlàg paths such that

$$\mathscr{L}^{\mathrm{N}} - 2\log(\mathrm{N}) \Longrightarrow \mathscr{L} \text{ as } \mathrm{N} \to \infty.$$

The process $\mathscr{L}$ has **infinite infinitesimal variance**; in particular

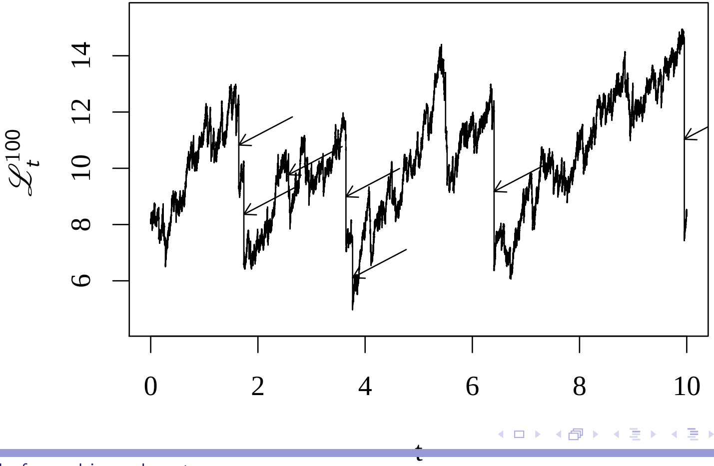
$$\frac{1}{t|\log t|}\mathbb{E}[(\mathscr{L}_t - \mathscr{L}_0)^2] \overset{t\to 0}{\sim} 2.$$

Sample path

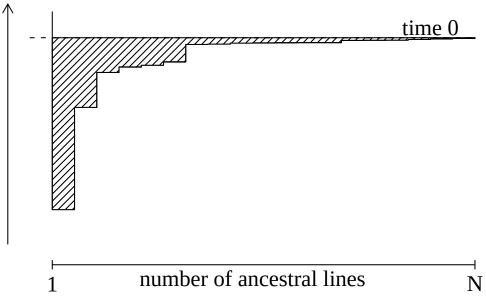


Sample path

When MRCA jumps, tree lengths jump as well

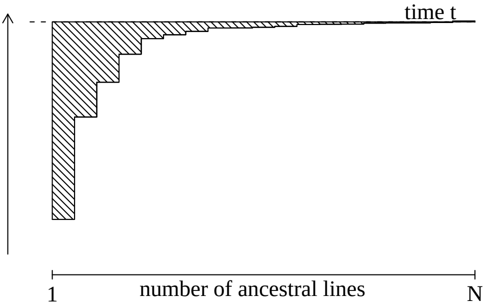


Infinitesimal variance

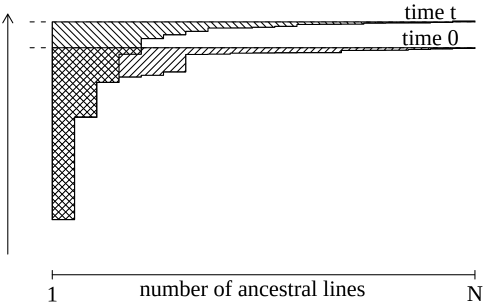


- $\mathbf{S_s^N}$:= # ancestors at time $-s$ (:=0 before MRCA)
- $\mathscr{L}_0^N = \int_0^\infty S_s^N ds$

Infinitesimal variance

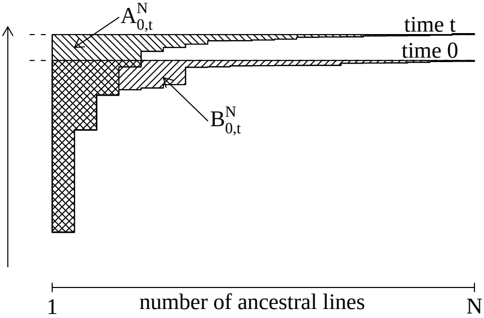


Infinitesimal variance



- Trees at times 0, t **overlap**

Infinitesimal variance



- $\mathscr{L}^N_t - \mathscr{L}^N_0 \stackrel{d}{=} A^N_{0,t} - B^N_{0,t}$
- For inf. variance, compute $\lim_{N\to\infty} \mathbb{V}[A^N_{0,t} - B^N_{0,t}]$ for $t \to 0$.

Infinitesimal variance

- **Gain in tree length**: $A^{\mathrm{N}}_{0,t} \stackrel{d}{=} \int_0^t S^{\mathrm{N}}_s ds$
- Coalescent comes down from $\infty \Rightarrow S^{\mathrm{N}}_s \xRightarrow{\mathrm{N}\to\infty} S_s$
- Aldous (1999):

$$\frac{S_s - 2/s}{\sqrt{2/(3s)}} \xRightarrow{s\to 0} N(0,1)$$

- Extension: $r \leq s \Rightarrow \mathbb{COV}[S_r, S_s] \stackrel{s\to 0}{\approx} \frac{2}{3}\frac{r}{s^2}$

$$\lim_{\mathrm{N}\to\infty} \mathbb{V}\left[A^{\mathrm{N}}_{0,t}\right] = 2\int_0^t \int_0^s \mathbb{COV}[S_r, S_s] drds \stackrel{t\to 0}{\approx} \frac{2}{3}t$$

Infinitesimal variance

- **Loss in tree length**

$$B^{\mathrm{N}}_{0,t} \stackrel{d}{=} \sum_{i=2}^{\mathrm{N}} (i - K_i^{\mathrm{N},S_t^{\mathrm{N}}}) \mathcal{E}\left(\binom{i}{2}\right)$$

- $K_i^{\mathrm{N},\mathrm{K}} := \#$ lines in K-tree while the N-tree has i lines.
- $S_t^{\mathrm{N}} \approx 2/t$, $\quad (K_i^{\mathrm{N},\mathrm{K}})_{i=\mathrm{N},\mathrm{N}-1,\dots}$ is Markov Chain

$$\xRightarrow{\text{some calculations}} \lim_{\mathrm{N}\to\infty} \mathbb{V}[B^{\mathrm{N}}_{0,t}] \stackrel{t\to 0}{\approx} 2t|\log t|$$

Infinitesimal variance

- **Collecting** terms

$$\lim_{N\to\infty} \mathbb{E}[(\mathscr{L}_t^N - \mathscr{L}_0^N)^2] = \lim_{N\to\infty} \mathbb{V}[A_{0,t}^N - B_{0,t}^N]$$
$$\overset{t\to 0}{\approx} \lim_{N\to\infty} \mathbb{V}[B_{0,t}^N]$$
$$\overset{t\to 0}{\approx} 2t|\log(t)|$$

Outlook

- General theory shows: $\mathscr{L}^{\mathrm{N}} \xrightarrow{\mathrm{N}\to\infty} \mathscr{L}$ **in probability** on the Lookdown probability space
  Does **almost sure convergence** hold as well?
- What is the joint evolution of $(\mathcal{D}_t, \mathcal{L}_t)$ of the tree?
  ($\mathcal{D}_t =$ **depth of the tree** at time t)
- Take a Cannings model with **finite offspring variance**. Does

$$\mathscr{L}^{\text{Cannings},\mathrm{N}} \xRightarrow{\mathrm{N}\to\infty} \mathscr{L}?$$

- What are **other limits** of $\mathscr{L}^{\text{Cannings},\mathrm{N}}$ for Cannings models with infinite offspring variance?

Summary

- Tree lengths in Kingman's coalescent are **Gumbel** distributed
- Evolution of tree lengths gives a **càdlàg** process $\mathscr{L}$
- $\mathscr{L}$ has **infinite quadratic variation**