

Splitting trees: New tools for branching populations

Amaury Lambert

Unit of Mathematical Evolutionary Biology
Lab of Ecology and Evolution
École Normale Supérieure and Université Pierre et Marie Curie (Paris 6)

Cornell University
June 30th, 2008

LAMBERT (2007). The contour of splitting trees is a Lévy process. Preprint arXiv.
LAMBERT (2008). The allelic partition for coalescent point processes. Preprint arXiv.

Definition

Recursive construction of the splitting tree (Geiger & Kersting 1997)

- $\Lambda =$ **lifespan measure**, such that $\int_0^\infty (1 \wedge r)\Lambda(dr) < \infty$
- A population with at most countably many individuals u with a **date of birth** $\alpha(u)$ and a **date of death** $\omega(u)$
- $\zeta(u) := \omega(u) - \alpha(u)$ is the **lifetime**
- A single ancestor born at time 0, dead at time χ
- the **dates of birth** and the **lifetimes** of the offspring of any individual with lifespan (α, ω) are the atoms of a **Poisson measure** with intensity $dt\Lambda(dr)$ on $(\alpha, \omega) \times (0, \infty)$
- Tree of individuals \mathcal{T} (countable)
- **Splitting tree** $\mathbb{T} = \cup_i(\alpha_i, \omega_i]$ (uncountable)
- \mathbb{T} is a (locally compact) \mathbb{R} -tree, endowed with a Lebesgue measure and a partial order (ancestry).

Definition

Recursive construction of the splitting tree (Geiger & Kersting 1997)

- $\Lambda =$ **lifespan measure**, such that $\int_0^\infty (1 \wedge r)\Lambda(dr) < \infty$
- A population with at most countably many individuals u with a **date of birth** $\alpha(u)$ and a **date of death** $\omega(u)$
- $\zeta(u) := \omega(u) - \alpha(u)$ is the **lifetime**
- A single ancestor born at time 0, dead at time χ
- the **dates of birth** and the **lifetimes** of the offspring of any individual with lifespan (α, ω) are the atoms of a **Poisson measure** with intensity $dt\Lambda(dr)$ on $(\alpha, \omega) \times (0, \infty)$
- Tree of individuals \mathcal{T} (countable)
- **Splitting tree** $\mathbb{T} = \cup_i(\alpha_i, \omega_i]$ (uncountable)
- \mathbb{T} is a (locally compact) \mathbb{R} -tree, endowed with a Lebesgue measure and a partial order (ancestry).

Definition

Recursive construction of the splitting tree (Geiger & Kersting 1997)

- $\Lambda =$ **lifespan measure**, such that $\int_0^\infty (1 \wedge r)\Lambda(dr) < \infty$
- A population with at most countably many individuals u with a **date of birth** $\alpha(u)$ and a **date of death** $\omega(u)$
- $\zeta(u) := \omega(u) - \alpha(u)$ is the **lifetime**
- A single ancestor born at time 0, dead at time χ
- the **dates of birth** and the **lifetimes** of the offspring of any individual with lifespan (α, ω) are the atoms of a **Poisson measure** with intensity $dt\Lambda(dr)$ on $(\alpha, \omega) \times (0, \infty)$
- Tree of individuals \mathcal{T} (countable)
- **Splitting tree** $\mathbb{T} = \cup_i(\alpha_i, \omega_i]$ (uncountable)
- \mathbb{T} is a (locally compact) \mathbb{R} -tree, endowed with a Lebesgue measure and a partial order (ancestry).

Definition

Recursive construction of the splitting tree (Geiger & Kersting 1997)

- $\Lambda =$ **lifespan measure**, such that $\int_0^\infty (1 \wedge r)\Lambda(dr) < \infty$
- A population with at most countably many individuals u with a **date of birth** $\alpha(u)$ and a **date of death** $\omega(u)$
- $\zeta(u) := \omega(u) - \alpha(u)$ is the **lifetime**
- A single ancestor born at time 0, dead at time χ
- the **dates of birth** and the **lifetimes** of the offspring of any individual with lifespan (α, ω) are the atoms of a **Poisson measure** with intensity $dt\Lambda(dr)$ on $(\alpha, \omega) \times (0, \infty)$
- Tree of individuals \mathcal{T} (countable)
- **Splitting tree** $\mathbb{T} = \cup_i(\alpha_i, \omega_i]$ (uncountable)
- \mathbb{T} is a (locally compact) \mathbb{R} -tree, endowed with a Lebesgue measure and a partial order (ancestry).

Definition

Recursive construction of the splitting tree (Geiger & Kersting 1997)

- $\Lambda =$ **lifespan measure**, such that $\int_0^\infty (1 \wedge r)\Lambda(dr) < \infty$
- A population with at most countably many individuals u with a **date of birth** $\alpha(u)$ and a **date of death** $\omega(u)$
- $\zeta(u) := \omega(u) - \alpha(u)$ is the **lifetime**
- A single ancestor born at time 0, dead at time χ
- the **dates of birth** and the **lifetimes** of the offspring of any individual with lifespan (α, ω) are the atoms of a **Poisson measure** with intensity $dt\Lambda(dr)$ on $(\alpha, \omega) \times (0, \infty)$
- Tree of individuals \mathcal{T} (countable)
- **Splitting tree** $\mathbb{T} = \cup_i(\alpha_i, \omega_i]$ (uncountable)
- \mathbb{T} is a (locally compact) \mathbb{R} -tree, endowed with a Lebesgue measure and a partial order (ancestry).

Recursive construction of the splitting tree (Geiger & Kersting 1997)

- $\Lambda =$ lifespan measure, such that $\int_0^\infty (1 \wedge r)\Lambda(dr) < \infty$
- A population with at most countably many individuals u with a date of birth $\alpha(u)$ and a date of death $\omega(u)$
- $\zeta(u) := \omega(u) - \alpha(u)$ is the lifetime
- A single ancestor born at time 0, dead at time χ
- the dates of birth and the lifetimes of the offspring of any individual with lifespan (α, ω) are the atoms of a Poisson measure with intensity $dt \Lambda(dr)$ on $(\alpha, \omega) \times (0, \infty)$
- Tree of individuals \mathcal{T} (countable)
- Splitting tree $\mathbb{T} = \cup_i (\alpha_i, \omega_i]$ (uncountable)
- \mathbb{T} is a (locally compact) \mathbb{R} -tree, endowed with a Lebesgue measure and a partial order (ancestry).

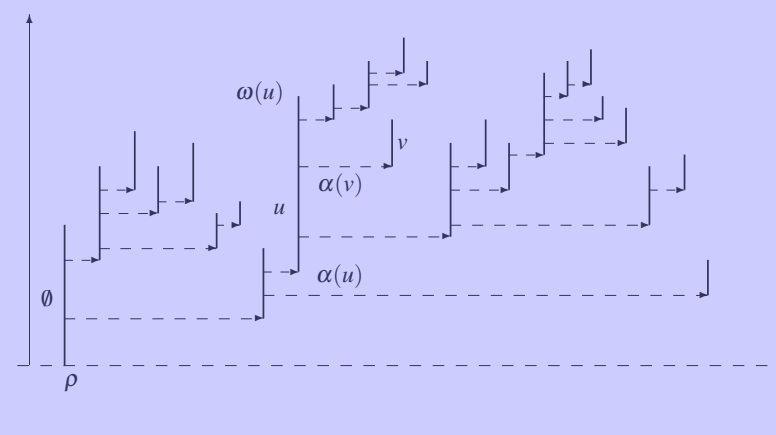
Recursive construction of the splitting tree (Geiger & Kersting 1997)

- $\Lambda =$ lifespan measure, such that $\int_0^\infty (1 \wedge r)\Lambda(dr) < \infty$
- A population with at most countably many individuals u with a date of birth $\alpha(u)$ and a date of death $\omega(u)$
- $\zeta(u) := \omega(u) - \alpha(u)$ is the lifetime
- A single ancestor born at time 0, dead at time χ
- the dates of birth and the lifetimes of the offspring of any individual with lifespan (α, ω) are the atoms of a Poisson measure with intensity $dt \Lambda(dr)$ on $(\alpha, \omega) \times (0, \infty)$
- Tree of individuals \mathcal{T} (countable)
- Splitting tree $\mathbb{T} = \cup_i (\alpha_i, \omega_i]$ (uncountable)
- \mathbb{T} is a (locally compact) \mathbb{R} -tree, endowed with a Lebesgue measure and a partial order (ancestry).

Recursive construction of the splitting tree (Geiger & Kersting 1997)

- $\Lambda =$ lifespan measure, such that $\int_0^\infty (1 \wedge r)\Lambda(dr) < \infty$
- A population with at most countably many individuals u with a date of birth $\alpha(u)$ and a date of death $\omega(u)$
- $\zeta(u) := \omega(u) - \alpha(u)$ is the lifetime
- A single ancestor born at time 0, dead at time χ
- the dates of birth and the lifetimes of the offspring of any individual with lifespan (α, ω) are the atoms of a Poisson measure with intensity $dt \Lambda(dr)$ on $(\alpha, \omega) \times (0, \infty)$
- Tree of individuals \mathcal{T} (countable)
- Splitting tree $\mathbb{T} = \cup_i (\alpha_i, \omega_i]$ (uncountable)
- \mathbb{T} is a (locally compact) \mathbb{R} -tree, endowed with a Lebesgue measure and a partial order (ancestry).

Planar representation of a splitting tree



First branching process

The number of individuals alive at τ

$$\Xi_\tau = \text{Card}\{v \in \mathcal{T} : \alpha(v) < \tau \leq \omega(v)\},$$

is a homogeneous binary Crump–Mode–Jagers process. This process is not Markovian, EXCEPT IF

First branching process

The number of individuals alive at τ

$$\Xi_\tau = \text{Card}\{v \in \mathcal{T} : \alpha(v) < \tau \leq \omega(v)\},$$

is a homogeneous binary Crump–Mode–Jagers process. This process is not Markovian, EXCEPT IF

- 1 $\Lambda(dx) = bde^{-dx}$. Then Ξ is a (linear) birth–death process with transition rates

$$\begin{cases} n \rightarrow n+1 & \text{at rate } bn, \\ n \rightarrow n-1 & \text{at rate } dn. \end{cases}$$

- 2 $\Lambda(dx) = b\delta_x(dx)$. Then Ξ is a (Yule) pure birth process with transition rates

$$\begin{cases} n \rightarrow n+1 & \text{at rate } bn. \end{cases}$$

First branching process

The number of individuals alive at τ

$$\Xi_\tau = \text{Card}\{v \in \mathcal{T} : \alpha(v) < \tau \leq \omega(v)\},$$

is a **homogeneous binary Crump–Mode–Jagers process**.

This process is **not** Markovian, EXCEPT IF

- 1 $\Lambda(dx) = bde^{-dx}dx$. Then Ξ is a (linear) **birth–death process** with transition rates

$$\begin{cases} n \rightarrow n+1 & \text{at rate } \mathbf{bn}, \\ n \rightarrow n-1 & \text{at rate } \mathbf{dn}. \end{cases}$$

- 2 $\Lambda(dx) = b\delta_\infty(dx)$. Then Ξ is a (Yule) **pure birth process** with transition rates

$$\{ n \rightarrow n+1 \text{ at rate } \mathbf{bn}. \}$$

Second branching process

The number of individuals of generation n

$$\mathcal{X}_n = \text{Card}\{v \in \mathcal{T} : |v| = n\}.$$

is a **Bienaymé–Galton–Watson (BGW) process** started at 1, with mean number of offspring per individual equal to

$$m := \int_{(0,\infty)} z\Lambda(dz).$$

Third branching process

The sum of lifetimes of all individuals of generation n

$$Z_n = \sum_{|u|=n} \zeta(u)$$

is a **Jirina process**, i.e. a branching process in discrete time and continuous state-space.

Indeed, there are i.i.d. **compound Poisson processes** (S_i) with no drift and Lévy measure Λ such that

$$Z_{n+1} = S_{n+1}(Z_n).$$

See *Bertoin & Le Gall* (2000), for genealogies defined from flows of subordinators.

Third branching process

The sum of lifetimes of all individuals of generation n

$$Z_n = \sum_{|u|=n} \zeta(u)$$

is a **Jirina process**, i.e. a branching process in discrete time and continuous state-space.

Indeed, there are i.i.d. **compound Poisson processes** (S_i) with no drift and Lévy measure Λ such that

$$Z_{n+1} = S_{n+1}(Z_n).$$

See *Bertoin & Le Gall* (2000), for genealogies defined from **flows of subordinators**.

Three branching processes

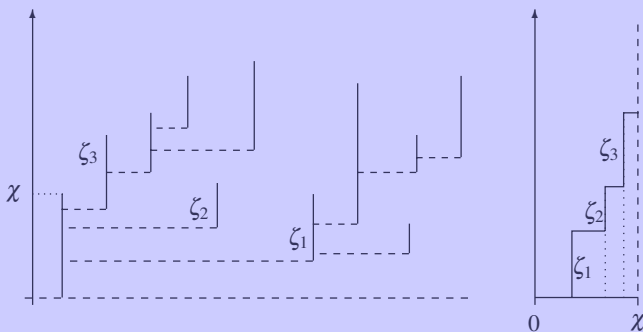


FIG.: A splitting tree and the birth process of the ancestor, a compound Poisson process on $[0, \chi]$.

The « Jumping Chronological Contour Process »

Theorem

Let $\tau > 0$ and X the **jumping (chronological) contour process (JCCP)** of the tree \mathbb{T}_τ obtained from \mathbb{T} by removing all points of \mathbb{T} at distance $> \tau$ from the root.

Then X is a **compound Poisson process** with no negative jumps reflected below τ and killed upon hitting 0. It has Lévy measure Λ and drift coefficient -1 .

The « Jumping Chronological Contour Process »

Theorem

Let $\tau > 0$ and X the jumping (chronological) contour process (JCCP) of the tree \mathbb{T}_τ obtained from \mathbb{T} by removing all points of \mathbb{T} at distance $> \tau$ from the root.

Then X is a compound Poisson process with no negative jumps reflected below τ and killed upon hitting 0. It has Lévy measure Λ and drift coefficient -1 .

The « Jumping Chronological Contour Process »

Theorem

Let $\tau > 0$ and X the jumping (chronological) contour process (JCCP) of the tree \mathbb{T}_τ obtained from \mathbb{T} by removing all points of \mathbb{T} at distance $> \tau$ from the root.

Then X is a compound Poisson process with no negative jumps reflected below τ and killed upon hitting 0. It has Lévy measure Λ and drift coefficient -1 .

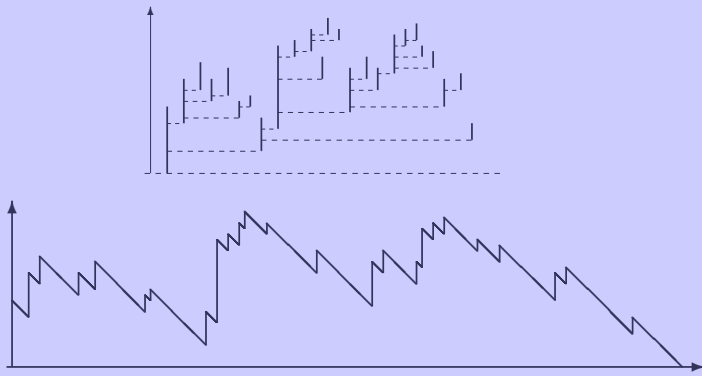


FIG.: A splitting tree and its associated JCCP.

The Lévy process Y (1)

The JCCP has the law of the compensated compound Poisson process Y (reflected below some truncation level τ)

$$t \mapsto Y_t := -t + \sum_{s \leq t} \Delta_s,$$

where $(\Delta_r, t \geq 0)$ is a Poisson point process with intensity measure $\text{Leb} \otimes \Lambda$.

Then Y is a Lévy process with no negative jumps and Laplace exponent ψ

$$\psi(\lambda) := \lambda - \int_0^\infty (1 - \exp(-\lambda r)) \Lambda(dr) \quad \lambda \geq 0,$$

that is,

$$\mathbb{E}_0(\exp(\lambda Y_t)) = \exp(-t\psi(\lambda)).$$

The Lévy process Y (1)

The JCCP has the law of the compensated compound Poisson process Y (reflected below some truncation level τ)

$$t \mapsto Y_t := -t + \sum_{s \leq t} \Delta_s,$$

where $(\Delta_r, t \geq 0)$ is a Poisson point process with intensity measure $\text{Leb} \otimes \Lambda$.

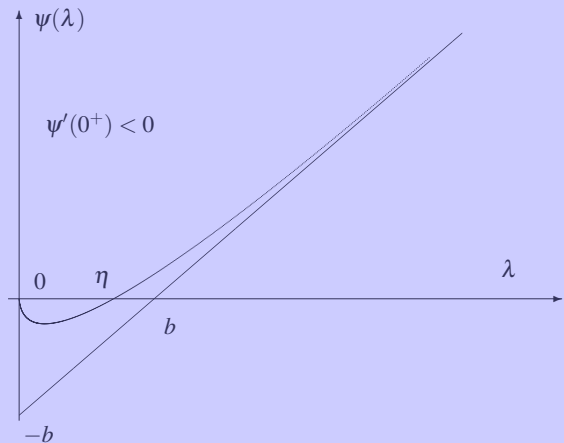
Then Y is a Lévy process with no negative jumps and Laplace exponent ψ

$$\psi(\lambda) := \lambda - \int_0^\infty (1 - \exp(-\lambda r)) \Lambda(dr) \quad \lambda \geq 0,$$

that is,

$$\mathbb{E}_0(\exp(\lambda Y_t)) = \exp(-t\psi(\lambda)).$$

(Supercritical) Laplace exponent



The Lévy process Y (2)

- The geometric **growth rate** m of the BGW process \mathcal{Z} is given by

$$\psi'(0+) = 1 - m$$

- The exponential **growth rate** η , or **Malthusian parameter**, of the CMJ process Ξ , is given by

$$\psi(\eta) = 0,$$

so in particular $\eta < b$, and $\eta = b$ in the Yule case.

- The **scale function** W is the positive function with Laplace transform $1/\psi$. For any $0 \leq x \leq a$,

$$P_x(T_0 < T_{(a,+\infty)}) = W(a-x)/W(a).$$

The Lévy process Y (2)

- The geometric **growth rate** m of the BGW process \mathcal{Z} is given by

$$\psi'(0+) = 1 - m$$

- The exponential **growth rate** η , or **Malthusian parameter**, of the CMJ process Ξ , is given by

$$\psi(\eta) = 0,$$

so in particular $\eta < b$, and $\eta = b$ in the Yule case.

- The **scale function** W is the positive function with Laplace transform $1/\psi$. For any $0 \leq x \leq a$,

$$P_x(T_0 < T_{(a,+\infty)}) = W(a-x)/W(a).$$

The Lévy process Y (2)

- The geometric **growth rate** m of the BGW process \mathcal{Z} is given by

$$\psi'(0+) = 1 - m$$

- The exponential **growth rate** η , or **Malthusian parameter**, of the CMJ process Ξ , is given by

$$\psi(\eta) = 0,$$

so in particular $\eta < b$, and $\eta = b$ in the Yule case.

- The **scale function** W is the positive function with Laplace transform $1/\psi$. For any $0 \leq x \leq a$,

$$P_x(T_0 < T_{(a,+\infty)}) = W(a-x)/W(a).$$

Examples

- Yule** process with (birth) rate c

$$W(x) = e^{cx}$$

- Noncritical birth–death** processes with birth rate b , death rate d , growth rate $r := b - d$

$$W(x) = 1 + \frac{b}{r}(e^{rx} - 1)$$

- Critical birth–death** processes with birth/death rate c

$$W(x) = 1 + cx$$

- Lifetimes in the domain of a **stable law** of index $\alpha \in (1, 2]$

$$W(x) = 1 + cx^{\alpha-1}$$

Examples

- Yule** process with (birth) rate c

$$W(x) = e^{cx}$$

- Noncritical birth–death** processes with birth rate b , death rate d , growth rate $r := b - d$

$$W(x) = 1 + \frac{b}{r}(e^{rx} - 1)$$

- Critical birth–death** processes with birth/death rate c

$$W(x) = 1 + cx$$

- Lifetimes in the domain of a **stable law** of index $\alpha \in (1, 2]$

$$W(x) = 1 + cx^{\alpha-1}$$

Examples

- Yule** process with (birth) rate c

$$W(x) = e^{cx}$$

- Noncritical birth–death** processes with birth rate b , death rate d , growth rate $r := b - d$

$$W(x) = 1 + \frac{b}{r}(e^{rx} - 1)$$

- Critical birth–death** processes with birth/death rate c

$$W(x) = 1 + cx$$

- Lifetimes in the domain of a **stable law** of index $\alpha \in (1, 2]$

$$W(x) = 1 + cx^{\alpha-1}$$

Examples

- **Yule** process with (birth) rate c

$$W(x) = e^{cx}$$

- **Noncritical birth–death** processes with birth rate b , death rate d , growth rate $r := b - d$

$$W(x) = 1 + \frac{b}{r}(e^{rx} - 1)$$

- **Critical birth–death** processes with birth/death rate c

$$W(x) = 1 + cx$$

- Lifetimes in the domain of a **stable law** of index $\alpha \in (1, 2]$

$$W(x) = 1 + cx^{\alpha-1}$$

One–dimensional marginals

Corollary

The probability of extinction is $\mathbb{P}_\chi(\text{Ext}) = e^{-\eta\chi}$.

The one-dimensional marginal of the CMJ process Ξ_τ is given by

$$\mathbb{P}_\chi(\Xi_\tau = 0) = W(\tau - \chi)/W(\tau) = P_\chi(T_0 < T_{(\tau, +\infty)}),$$

and conditional on being nonzero, Ξ_τ has a geometric distribution with success probability

$$1 - 1/W(\tau) = P_\tau(T_0 > T_{(\tau, +\infty)}).$$

In particular, $\mathbb{E}_\chi(\Xi_\tau | \Xi_\tau \neq 0) = W(\tau)$.

One–dimensional marginals

Corollary

The probability of extinction is $\mathbb{P}_\chi(\text{Ext}) = e^{-\eta\chi}$.

The one-dimensional marginal of the CMJ process Ξ_τ is given by

$$\mathbb{P}_\chi(\Xi_\tau = 0) = W(\tau - \chi)/W(\tau) = P_\chi(T_0 < T_{(\tau, +\infty)}),$$

and conditional on being nonzero, Ξ_τ has a geometric distribution with success probability

$$1 - 1/W(\tau) = P_\tau(T_0 > T_{(\tau, +\infty)}).$$

In particular, $\mathbb{E}_\chi(\Xi_\tau | \Xi_\tau \neq 0) = W(\tau)$.

One–dimensional marginals

Corollary

The probability of extinction is $\mathbb{P}_\chi(\text{Ext}) = e^{-\eta\chi}$.

The one-dimensional marginal of the CMJ process Ξ_τ is given by

$$\mathbb{P}_\chi(\Xi_\tau = 0) = W(\tau - \chi)/W(\tau) = P_\chi(T_0 < T_{(\tau, +\infty)}),$$

and conditional on being nonzero, Ξ_τ has a geometric distribution with success probability

$$1 - 1/W(\tau) = P_\tau(T_0 > T_{(\tau, +\infty)}).$$

In particular, $\mathbb{E}_\chi(\Xi_\tau | \Xi_\tau \neq 0) = W(\tau)$.

One–dimensional marginals

Corollary

The probability of extinction is $\mathbb{P}_\chi(\text{Ext}) = e^{-\eta\chi}$.

The one-dimensional marginal of the CMJ process Ξ_τ is given by

$$\mathbb{P}_\chi(\Xi_\tau = 0) = W(\tau - \chi)/W(\tau) = P_\chi(T_0 < T_{(\tau, +\infty)}),$$

and conditional on being nonzero, Ξ_τ has a geometric distribution with success probability

$$1 - 1/W(\tau) = P_\tau(T_0 > T_{(\tau, +\infty)}).$$

In particular, $\mathbb{E}_\chi(\Xi_\tau | \Xi_\tau \neq 0) = W(\tau)$.

One–dimensional marginals

Corollary

The probability of extinction is $\mathbb{P}_\chi(\text{Ext}) = e^{-\eta\chi}$.

The one-dimensional marginal of the CMJ process Ξ_τ is given by

$$\mathbb{P}_\chi(\Xi_\tau = 0) = W(\tau - \chi)/W(\tau) = P_\chi(T_0 < T_{(\tau, +\infty)}),$$

and conditional on being nonzero, Ξ_τ has a geometric distribution with success probability

$$1 - 1/W(\tau) = P_\tau(T_0 > T_{(\tau, +\infty)}).$$

In particular, $\mathbb{E}_\chi(\Xi_\tau | \Xi_\tau \neq 0) = W(\tau)$.

Conditioning on extinction

Corollary

In the supercritical case, set $\mathbb{P}^\natural := \mathbb{P}(\cdot \mid \text{Ext})$. Then the JCCP under \mathbb{P}^\natural is a Lévy process with Laplace exponent ψ^\natural

$$\psi^\natural(\lambda) = \psi(\lambda + \eta) = \lambda - \int_0^\infty (1 - e^{-\lambda r}) e^{-\eta r} \Lambda(dr) \quad \lambda \geq 0.$$

As a consequence, the supercritical splitting tree **conditioned on its extinction** has the same law as the **subcritical** splitting tree with lifespan measure $e^{-\eta r} \Lambda(dr)$. In particular its birth rate equals $b - \eta$.

Asymptotic behaviour

Corollary

(i) (Yaglom's distribution) In the subcritical case,

$$\lim_{\tau \rightarrow \infty} \mathbb{P}(\Xi_\tau = n \mid \Xi_\tau \neq 0) = m^{n-1}(1-m) \quad n \geq 1.$$

(ii) In the critical case, provided that $\int^\infty r^2 \Lambda(dr) < \infty$,

$$\lim_{\tau \rightarrow \infty} \mathbb{P}(\Xi_\tau / \tau > x \mid \Xi_\tau \neq 0) = \exp(-\psi''(0+) x/2) \quad x \geq 0.$$

(iii) In the supercritical case, conditional on $\{\text{Ext}^c\}$ and provided that $\int^\infty r \log(r) \Lambda(dr) < \infty$,

$$\lim_{\tau \rightarrow \infty} e^{-\eta \tau} \Xi_\tau = \xi \quad \text{a.s.},$$

where ξ is an exponential variable with parameter $1/\psi'(\eta)$.

Asymptotic behaviour

Corollary

(i) (Yaglom's distribution) In the **subcritical** case,

$$\lim_{\tau \rightarrow \infty} \mathbb{P}(\Xi_\tau = n \mid \Xi_\tau \neq 0) = m^{n-1}(1-m) \quad n \geq 1.$$

(ii) In the **critical** case, provided that $\int^\infty r^2 \Lambda(dr) < \infty$,

$$\lim_{\tau \rightarrow \infty} \mathbb{P}(\Xi_\tau / \tau > x \mid \Xi_\tau \neq 0) = \exp(-\psi''(0+) x/2) \quad x \geq 0.$$

(iii) In the supercritical case, conditional on $\{\text{Ext}^c\}$ and provided that $\int^\infty r \log(r) \Lambda(dr) < \infty$,

$$\lim_{\tau \rightarrow \infty} e^{-\eta \tau} \Xi_\tau = \xi \quad \text{a.s.},$$

where ξ is an exponential variable with parameter $1/\psi'(\eta)$.

Asymptotic behaviour

Corollary

(i) (Yaglom's distribution) In the **subcritical** case,

$$\lim_{\tau \rightarrow \infty} \mathbb{P}(\Xi_\tau = n \mid \Xi_\tau \neq 0) = m^{n-1}(1-m) \quad n \geq 1.$$

(ii) In the **critical** case, provided that $\int^\infty r^2 \Lambda(dr) < \infty$,

$$\lim_{\tau \rightarrow \infty} \mathbb{P}(\Xi_\tau / \tau > x \mid \Xi_\tau \neq 0) = \exp(-\psi''(0+) x/2) \quad x \geq 0.$$

(iii) In the supercritical case, conditional on $\{\text{Ext}^c\}$ and provided that $\int^\infty r \log(r) \Lambda(dr) < \infty$,

$$\lim_{\tau \rightarrow \infty} e^{-\eta \tau} \Xi_\tau = \xi \quad \text{a.s.},$$

where ξ is an exponential variable with parameter $1/\psi'(\eta)$.

Asymptotic behaviour

Corollary

(i) (Yaglom's distribution) In the **subcritical** case,

$$\lim_{\tau \rightarrow \infty} \mathbb{P}(\Xi_\tau = n \mid \Xi_\tau \neq 0) = m^{n-1}(1-m) \quad n \geq 1.$$

(ii) In the **critical** case, provided that $\int^\infty r^2 \Lambda(dr) < \infty$,

$$\lim_{\tau \rightarrow \infty} \mathbb{P}(\Xi_\tau / \tau > x \mid \Xi_\tau \neq 0) = \exp(-\psi''(0+) x/2) \quad x \geq 0.$$

(iii) In the supercritical case, conditional on $\{\text{Ext}^c\}$ and provided that $\int^\infty r \log(r) \Lambda(dr) < \infty$,

$$\lim_{\tau \rightarrow \infty} e^{-\eta \tau} \Xi_\tau = \xi \quad \text{a.s.},$$

where ξ is an exponential variable with parameter $1/\psi'(\eta)$.

Supercritical case : A taste of spine decomposition

In the supercritical case, write Ξ_τ as

$$\Xi_\tau = \Xi_\tau^\infty + \Xi_\tau^f,$$

distinguishing between points whose descendance is either infinite or finite. Then conditional on Ext^c ,

$$\lim_{\tau \rightarrow \infty} e^{-\eta \tau} (\Xi_\tau^\infty, \Xi_\tau^f) \stackrel{\mathcal{L}}{=} (p\xi, (1-p)\xi),$$

where $p := 1/\psi'(\eta)$ and ξ is exponential with parameter p .

In particular, $e^{-\eta \tau} \Xi_\tau^\infty$ converges in distribution to an exponential variable with parameter 1.

Actually, $(\Xi_\tau^\infty; \tau \geq 0)$ is a Yule process with birth rate η .

Supercritical case : A taste of spine decomposition

In the supercritical case, write Ξ_τ as

$$\Xi_\tau = \Xi_\tau^\infty + \Xi_\tau^f,$$

distinguishing between points whose descendance is either infinite or finite. Then conditional on Ext^c ,

$$\lim_{\tau \rightarrow \infty} e^{-\eta\tau} (\Xi_\tau^\infty, \Xi_\tau^f) \stackrel{\mathcal{L}}{=} (p\xi, (1-p)\xi),$$

where $p := 1/\psi'(\eta)$ and ξ is exponential with parameter p .
 In particular, $e^{-\eta\tau}\Xi_\tau^\infty$ converges in distribution to an exponential variable with parameter 1.
 Actually, $(\Xi_\tau^\infty; \tau \geq 0)$ is a Yule process with birth rate η .

Supercritical case : A taste of spine decomposition

In the supercritical case, write Ξ_τ as

$$\Xi_\tau = \Xi_\tau^\infty + \Xi_\tau^f,$$

distinguishing between points whose descendance is either infinite or finite. Then conditional on Ext^c ,

$$\lim_{\tau \rightarrow \infty} e^{-\eta\tau} (\Xi_\tau^\infty, \Xi_\tau^f) \stackrel{\mathcal{L}}{=} (p\xi, (1-p)\xi),$$

where $p := 1/\psi'(\eta)$ and ξ is exponential with parameter p .
 In particular, $e^{-\eta\tau}\Xi_\tau^\infty$ converges in distribution to an exponential variable with parameter 1.
 Actually, $(\Xi_\tau^\infty; \tau \geq 0)$ is a Yule process with birth rate η .

Height process

Theorem

The generation in the discrete tree of individual visited at time t , is given by

$$H_t := \text{Card}\{s \leq t : \inf_{s \leq u \leq t} X_u = X_{s-}\}.$$

$\Rightarrow H$ is the height process of Le Gall & Le Jan (1998), Duquesne & Le Gall (2002).

Recall Z_n is the sum of lifetimes of individuals of generation n . Then Z_n is also the occupation measure of the height process :

$$Z_n = \int_0^\infty \mathbf{1}_{H_t=n} dt,$$

where one recovers the genealogy of Le Gall–Le Jan–Duquesne.

Height process

Theorem

The generation in the discrete tree of individual visited at time t , is given by

$$H_t := \text{Card}\{s \leq t : \inf_{s \leq u \leq t} X_u = X_{s-}\}.$$

$\Rightarrow H$ is the height process of Le Gall & Le Jan (1998), Duquesne & Le Gall (2002).

Recall Z_n is the sum of lifetimes of individuals of generation n . Then Z_n is also the occupation measure of the height process :

$$Z_n = \int_0^\infty \mathbf{1}_{H_t=n} dt,$$

where one recovers the genealogy of Le Gall–Le Jan–Duquesne.

Height process

Theorem

The generation in the discrete tree of individual visited at time t , is given by

$$H_t := \text{Card}\{s \leq t : \inf_{s \leq u \leq t} X_u = X_{s-}\}.$$

$\Rightarrow H$ is the height process of Le Gall & Le Jan (1998), Duquesne & Le Gall (2002).

Recall Z_n is the sum of lifetimes of individuals of generation n . Then Z_n is also the occupation measure of the height process :

$$Z_n = \int_0^\infty \mathbf{1}_{H_t=n} dt,$$

where one recovers the genealogy of Le Gall–Le Jan–Duquesne.

Heights

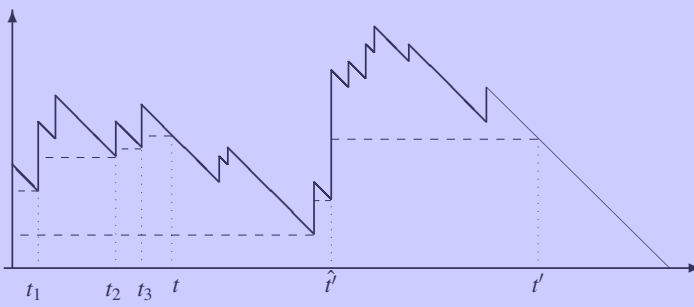


FIG.: The JCCP of some finite splitting tree with jumps in solid line. The first visits to the successive ancestors are shown.

Coalescence

Fix $\tau > 0$ and let $(x_i; 1 \leq i \leq \Xi_\tau)$ denote the points at level τ ranked in the order of the contour.

Under \mathbb{P} , the durations $H_i = \tau - a_i$ elapsed since the coalescence of points x_i and x_{i+1} has

$$\mathbb{P}(H_i \leq \sigma) = \frac{1 - 1/W(\sigma)}{1 - 1/W(\tau)} \quad \sigma \leq \tau.$$

Coalescence

Fix $\tau > 0$ and let $(x_i; 1 \leq i \leq \Xi_\tau)$ denote the points at level τ ranked in the order of the contour.

Theorem

Under \mathbb{P} , the coalescence levels between consecutive points $a_i, 1 \leq i \leq \Xi_\tau$, form a sequence of *i.i.d. r.v. killed at the first negative one*, and distributed as $\tau - \inf Y_t$, where Y is the Lévy process with Laplace exponent ψ started at τ and killed upon exiting $(0, \tau]$. As a consequence, the duration $H_i = \tau - a_i$ elapsed since the coalescence of points x_i and x_{i+1} has

$$\mathbb{P}(H_i \leq \sigma) = \frac{1 - 1/W(\sigma)}{1 - 1/W(\tau)} \quad \sigma \leq \tau.$$

Coalescence

Fix $\tau > 0$ and let $(x_i; 1 \leq i \leq \Xi_\tau)$ denote the points at level τ ranked in the order of the contour.

Theorem

Under \mathbb{P} , the coalescence levels between consecutive points $a_i, 1 \leq i \leq \Xi_\tau$, form a sequence of *i.i.d. r.v. killed at the first negative one*, and distributed as $\tau - \inf Y_t$, where Y is the Lévy process with Laplace exponent ψ started at τ and killed upon exiting $(0, \tau]$. As a consequence, the duration $H_i = \tau - a_i$ elapsed since the coalescence of points x_i and x_{i+1} has

$$\mathbb{P}(H_i \leq \sigma) = \frac{1 - 1/W(\sigma)}{1 - 1/W(\tau)} \quad \sigma \leq \tau.$$

Coalescence

Fix $\tau > 0$ and let $(x_i; 1 \leq i \leq \Xi_\tau)$ denote the points at level τ ranked in the order of the contour.

Theorem

Under \mathbb{P} , the coalescence levels between consecutive points $a_i, 1 \leq i \leq \Xi_\tau$, form a sequence of *i.i.d. r.v. killed at the first negative one*, and distributed as $\tau - \inf Y_t$, where Y is the Lévy process with Laplace exponent ψ started at τ and killed upon exiting $(0, \tau]$. As a consequence, the duration $H_i = \tau - a_i$ elapsed since the coalescence of points x_i and x_{i+1} has

$$\mathbb{P}(H_i \leq \sigma) = \frac{1 - 1/W(\sigma)}{1 - 1/W(\tau)} \quad \sigma \leq \tau.$$

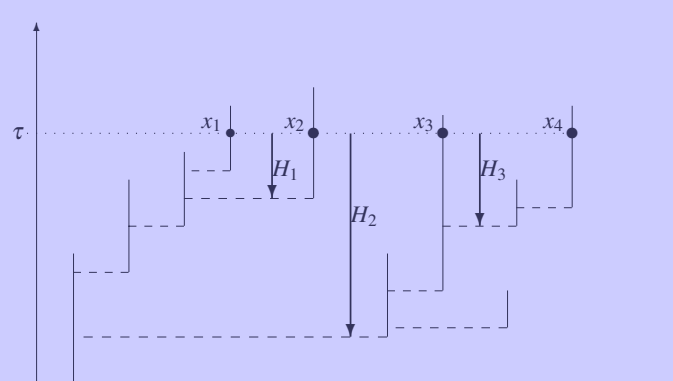
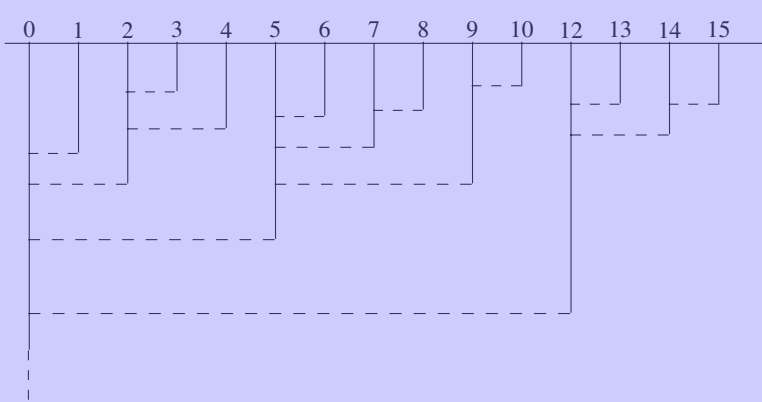


FIG.: Illustration of a splitting tree showing the durations H_1, H_2, H_3 elapsed since coalescence for each of the three consecutive pairs $(x_1, x_2), (x_2, x_3)$ and (x_3, x_4) of the $\Xi_\tau = 4$ individuals alive at level τ .

A coalescent point process for 16 individuals



Definition

Genealogy of individuals $0, 1, 2, \dots$ given by $C_{i,i+k} > 0$ (for $i, k \geq 0$),

$C_{i,i+k}$ = **coalescence time** between individuals i and $i+k$.

The collection of **branch lengths** : $H_0 = \infty$ and

$$H_{i+1} := C_{i,i+1} \quad i \geq 0,$$

forms a **coalescent point process** if

1 the coalescence times can be written as

$$C_{i,i+k} = \max\{H_{i+1}, \dots, H_{i+k}\}$$

2 the branch lengths H_1, H_2, \dots are **i.i.d.**

Notation.

$$W(x) := \frac{1}{\mathbb{P}(H > x)} \quad x > 0.$$

Definition

Genealogy of individuals $0, 1, 2, \dots$ given by $C_{i,i+k} > 0$ (for $i, k \geq 0$),

$C_{i,i+k}$ = **coalescence time** between individuals i and $i+k$.

The collection of **branch lengths** : $H_0 = \infty$ and

$$H_{i+1} := C_{i,i+1} \quad i \geq 0,$$

forms a **coalescent point process** if

1 the coalescence times can be written as

$$C_{i,i+k} = \max\{H_{i+1}, \dots, H_{i+k}\}$$

2 the branch lengths H_1, H_2, \dots are **i.i.d.**

Notation.

$$W(x) := \frac{1}{\mathbb{P}(H > x)} \quad x > 0.$$

Definition

Genealogy of individuals $0, 1, 2, \dots$ given by $C_{i,i+k} > 0$ (for $i, k \geq 0$),

$C_{i,i+k}$ = **coalescence time** between individuals i and $i+k$.

The collection of **branch lengths** : $H_0 = \infty$ and

$$H_{i+1} := C_{i,i+1} \quad i \geq 0,$$

forms a **coalescent point process** if

1 the coalescence times can be written as

$$C_{i,i+k} = \max\{H_{i+1}, \dots, H_{i+k}\}$$

2 the branch lengths H_1, H_2, \dots are **i.i.d.**

Notation.

$$W(x) := \frac{1}{\mathbb{P}(H > x)} \quad x > 0.$$

Definition

Genealogy of individuals $0, 1, 2, \dots$ given by $C_{i,i+k} > 0$ (for $i, k \geq 0$),

$C_{i,i+k}$ = **coalescence time** between individuals i and $i+k$.

The collection of **branch lengths** : $H_0 = \infty$ and

$$H_{i+1} := C_{i,i+1} \quad i \geq 0,$$

forms a **coalescent point process** if

1 the coalescence times can be written as

$$C_{i,i+k} = \max\{H_{i+1}, \dots, H_{i+k}\}$$

2 the branch lengths H_1, H_2, \dots are **i.i.d.**

Notation.

$$W(x) := \frac{1}{\mathbb{P}(H > x)} \quad x > 0.$$

Definition

Genealogy of individuals $0, 1, 2, \dots$ given by $C_{i,i+k} > 0$ (for $i, k \geq 0$),

$C_{i,i+k}$ = **coalescence time** between individuals i and $i+k$.

The collection of **branch lengths** : $H_0 = \infty$ and

$$H_{i+1} := C_{i,i+1} \quad i \geq 0,$$

forms a **coalescent point process** if

1 the coalescence times can be written as

$$C_{i,i+k} = \max\{H_{i+1}, \dots, H_{i+k}\}$$

2 the branch lengths H_1, H_2, \dots are **i.i.d.**

Notation.

$$W(x) := \frac{1}{\mathbb{P}(H > x)} \quad x > 0.$$

Definition

Genealogy of individuals $0, 1, 2, \dots$ given by $C_{i,i+k} > 0$ (for $i, k \geq 0$),

$C_{i,i+k}$ = **coalescence time** between individuals i and $i+k$.

The collection of **branch lengths** : $H_0 = \infty$ and

$$H_{i+1} := C_{i,i+1} \quad i \geq 0,$$

forms a **coalescent point process** if

1 the coalescence times can be written as

$$C_{i,i+k} = \max\{H_{i+1}, \dots, H_{i+k}\}$$

2 the branch lengths H_1, H_2, \dots are **i.i.d.**

Notation.

$$W(x) := \frac{1}{\mathbb{P}(H > x)} \quad x > 0.$$

A famous sister model : the Kingman coalescent

Recall that the genealogy of a sample of n individuals is given by the **Kingman coalescent** if PAIRS of lineages coalesce at constant rate.

The Kingman coalescent and coalescent point processes are **different** :

=> the **smallest** branch length is...

- ...in the Kingman coalescent : the **minimum** of $n(n-1)/2$ independent exponential r.v.
- ...in the coalescent point process : the **minimum** of $n-1$ independent r.v. (distributed as H).

A famous sister model : the Kingman coalescent

Recall that the genealogy of a sample of n individuals is given by the **Kingman coalescent** if PAIRS of lineages coalesce at constant rate.

The Kingman coalescent and coalescent point processes are **different** :

=> the **smallest** branch length is...

- ...in the Kingman coalescent : the **minimum** of $n(n-1)/2$ independent exponential r.v.
- ...in the coalescent point process : the **minimum** of $n-1$ independent r.v. (distributed as H).

A famous sister model : the Kingman coalescent

Recall that the genealogy of a sample of n individuals is given by the **Kingman coalescent** if PAIRS of lineages coalesce at constant rate.

The Kingman coalescent and coalescent point processes are **different** :

=> the **smallest** branch length is...

- ...in the Kingman coalescent : the **minimum** of $n(n-1)/2$ independent exponential r.v.
- ...in the coalescent point process : the **minimum** of $n-1$ independent r.v. (distributed as H).

A famous sister model : the Kingman coalescent

Recall that the genealogy of a sample of n individuals is given by the **Kingman coalescent** if PAIRS of lineages coalesce at constant rate.

The Kingman coalescent and coalescent point processes are **different** :

=> the **smallest** branch length is...

- ...in the Kingman coalescent : the **minimum** of $n(n-1)/2$ independent exponential r.v.
- ...in the coalescent point process : the **minimum** of $n-1$ independent r.v. (distributed as H).

Assumptions

Now mutations occur at random times.

- 1 mutations occur at **constant rate** θ on germ lines
- 2 mutations are **neutral** : they have no effect on the genealogy (birth rates, lifetimes...)

Assumptions

Now mutations occur at random times.

- 1 mutations occur at **constant rate** θ on germ lines
- 2 mutations are **neutral** : they have no effect on the genealogy (birth rates, lifetimes...)

DNA sequences

- each mutation is a point substitution occurring at a **single site** on the DNA sequence
- *assumption* : each site can mutate **at most once** (infinitely-many sites model)
- a site is **polymorphic**, or **segregating**, if at least one individual carries the wild-type and at least another one carries the mutant
- all distinct instances of DNA sequences are called **alleles** or **haplotypes**
- *consequence* (in this model) : each mutation yields a new allele (infinitely-many alleles model)

DNA sequences

- each mutation is a point substitution occurring at a **single site** on the DNA sequence
- *assumption* : each site can mutate **at most once** (infinitely-many sites model)
- a site is **polymorphic**, or **segregating**, if at least one individual carries the wild-type and at least another one carries the mutant
- all distinct instances of DNA sequences are called **alleles** or **haplotypes**
- *consequence* (in this model) : each mutation yields a new allele (infinitely-many alleles model)

DNA sequences

- each mutation is a point substitution occurring at a **single site** on the DNA sequence
- *assumption* : each site can mutate **at most once** (infinitely-many sites model)
- a site is **polymorphic**, or **segregating**, if at least one individual carries the wild-type and at least another one carries the mutant
- all distinct instances of DNA sequences are called **alleles** or **haplotypes**
- *consequence* (in this model) : each mutation yields a new allele (infinitely-many alleles model)

DNA sequences

- each mutation is a point substitution occurring at a **single site** on the DNA sequence
- *assumption* : each site can mutate **at most once** (infinitely-many sites model)
- a site is **polymorphic**, or **segregating**, if at least one individual carries the wild-type and at least another one carries the mutant
- all distinct instances of DNA sequences are called **alleles** or **haplotypes**
- *consequence* (in this model) : each mutation yields a new allele (infinitely-many alleles model)

DNA sequences

- each mutation is a point substitution occurring at a **single site** on the DNA sequence
- *assumption* : each site can mutate **at most once** (infinitely-many sites model)
- a site is **polymorphic**, or **segregating**, if at least one individual carries the wild-type and at least another one carries the mutant
- all distinct instances of DNA sequences are called **alleles** or **haplotypes**
- *consequence* (in this model) : each mutation yields a new allele (infinitely-many alleles model)

Polymorphic sites

For a sample of n individuals,

- S_n := number of **polymorphic sites**
- $S_n(k)$:= number of mutations carried by k individuals :

$$S_n = \sum_{k=1}^{n-1} S_n(k).$$

- the sequence $(S_n(1), \dots, S_n(n-1))$ is the **site frequency spectrum** of the sample.

Polymorphic sites

For a sample of n individuals,

- S_n := number of **polymorphic sites**
- $S_n(k)$:= number of mutations carried by k individuals :

$$S_n = \sum_{k=1}^{n-1} S_n(k).$$

- the sequence $(S_n(1), \dots, S_n(n-1))$ is the **site frequency spectrum** of the sample.

Polymorphic sites

For a sample of n individuals,

- S_n := number of **polymorphic sites**
- $S_n(k)$:= number of mutations carried by k individuals :

$$S_n = \sum_{k=1}^{n-1} S_n(k).$$

- the sequence $(S_n(1), \dots, S_n(n-1))$ is the **site frequency spectrum** of the sample.

Haplotypes

For a sample of n individuals,

- A_n := number of **distinct haplotypes**
- $A_n(k)$:= number of distinct haplotypes carried by k individuals :

$$A_n = \sum_{k=1}^n A_n(k) \text{ and } \sum_{k=1}^n kA_n(k) = n.$$

- the sequence $(A_n(1), \dots, A_n(n))$ is the **allele frequency spectrum** of the sample.

Remark. We always have $S_n \geq A_n - 1$: a new mutation is needed for each new haplotype.

Haplotypes

For a sample of n individuals,

- A_n := number of **distinct haplotypes**
- $A_n(k)$:= number of distinct haplotypes carried by k individuals :

$$A_n = \sum_{k=1}^n A_n(k) \text{ and } \sum_{k=1}^n kA_n(k) = n.$$

- the sequence $(A_n(1), \dots, A_n(n))$ is the **allele frequency spectrum** of the sample.

Remark. We always have $S_n \geq A_n - 1$: a new mutation is needed for each new haplotype.

Haplotypes

For a sample of n individuals,

- A_n := number of **distinct haplotypes**
- $A_n(k)$:= number of distinct haplotypes carried by k individuals :

$$A_n = \sum_{k=1}^n A_n(k) \text{ and } \sum_{k=1}^n kA_n(k) = n.$$

- the sequence $(A_n(1), \dots, A_n(n))$ is the **allele frequency spectrum** of the sample.

Remark. We always have $S_n \geq A_n - 1$: a new mutation is needed for each new haplotype.

Haplotypes

For a sample of n individuals,

- A_n := number of **distinct haplotypes**
- $A_n(k)$:= number of distinct haplotypes carried by k individuals :

$$A_n = \sum_{k=1}^n A_n(k) \text{ and } \sum_{k=1}^n kA_n(k) = n.$$

- the sequence $(A_n(1), \dots, A_n(n))$ is the **allele frequency spectrum** of the sample.

Remark. We always have $S_n \geq A_n - 1$: a new mutation is needed for each new haplotype.

Haplotypes

For a sample of n individuals,

- $A_n :=$ number of **distinct haplotypes**
- $A_n(k) :=$ number of distinct haplotypes carried by k individuals :

$$A_n = \sum_{k=1}^n A_n(k) \text{ and } \sum_{k=1}^n k A_n(k) = n.$$

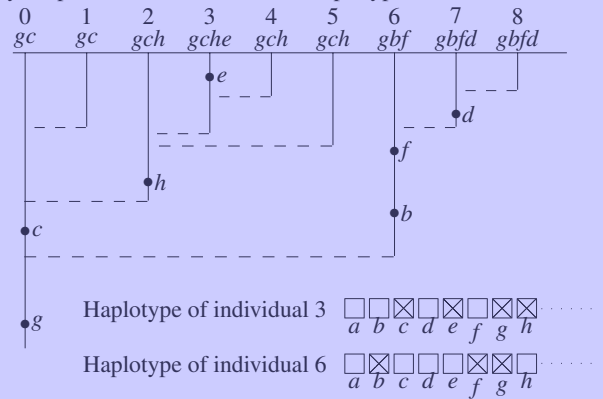
- the sequence $(A_n(1), \dots, A_n(n))$ is the **allele frequency spectrum** of the sample.

Remark. We always have $S_n \geq A_n - 1$: a new mutation is needed for each new haplotype.

A coalescent point process with mutations

Sites a and g are not polymorphic.

Polymorphic sites : $S_n = ?$ Distinct haplotypes : $A_n = ?$



The Chinese restaurant process

Customers enter successively a Chinese restaurant with many large tables. When k customers are seated,

- 1 the $(k + 1)$ -th customer chooses to sit at an **empty** table with probability $\theta / (k + \theta)$...
- 2 or to the table of any **given customer** with probability $1 / (k + \theta)$.

Theorem

The number of tables occupied by the first n customers is distributed like A_n in the Kingman coalescent.

The Chinese restaurant process

Customers enter successively a Chinese restaurant with many large tables. When k customers are seated,

- 1 the $(k + 1)$ -th customer chooses to sit at an **empty** table with probability $\theta / (k + \theta)$...
- 2 or to the table of any **given customer** with probability $1 / (k + \theta)$.

Theorem

The number of tables occupied by the first n customers is distributed like A_n in the Kingman coalescent.

The Chinese restaurant process

Customers enter successively a Chinese restaurant with many large tables. When k customers are seated,

- 1 the $(k + 1)$ -th customer chooses to sit at an **empty** table with probability $\theta / (k + \theta)$...
- 2 or to the table of any **given customer** with probability $1 / (k + \theta)$.

Theorem

The number of tables occupied by the first n customers is distributed like A_n in the Kingman coalescent.

Classical results for the Kingman coalescent

Theorem

For the **Kingman coalescent** with rescaled mutation rate θ , as $n \rightarrow \infty$,

$$S_n \sim \theta \ln(n) \quad \text{and} \quad A_n \sim \theta \ln(n),$$

with convergence rate $\sqrt{\ln(n)}$. In addition,

$$\lim_{n \rightarrow \infty} S_n(k) \stackrel{\mathbb{P}}{=} \frac{\theta}{k} \quad \text{and} \quad \lim_{n \rightarrow \infty} A_n(k) \stackrel{\mathcal{L}}{=} Y_k,$$

where Y_k denotes a Poisson r.v. with parameter θ/k .

Segregating sites : Case when $\mathbb{E}(H)$ is finite

Law of large numbers and central limit theorem

For the **coalescent point process**...

Theorem

If $\mathbb{E}(H) < \infty$, then

$$\lim_{n \rightarrow \infty} n^{-1} S_n = \theta \mathbb{E}(H) \quad \text{a.s. and in } L^1.$$

If in addition $\mathbb{E}(H^2) < \infty$, then

$$\sqrt{n} (n^{-1} S_n - \theta \mathbb{E}(H))$$

converges in distribution to a centered normal variable with variance $\theta \mathbb{E}(H) + \theta^2 \text{Var}(H)$ (mutational variance + evolutionary variance).

Segregating sites : Case when $\mathbb{E}(H)$ is finite

Expected site frequency spectrum

Theorem

For all $1 \leq k \leq n-1$,

$$\mathbb{E}(S_n(k)) = \theta \int_0^\infty dx \left(1 - \frac{1}{W(x)}\right)^{k-1} \left(\frac{n-k-1}{W(x)^2} + \frac{2}{W(x)}\right),$$

which is **finite** if and only if $\mathbb{E}(H) < \infty$. Then in particular,

$$\lim_{n \rightarrow \infty} n^{-1} \mathbb{E}(S_n(k)) = \theta \int_0^\infty \frac{dx}{W(x)^2} \left(1 - \frac{1}{W(x)}\right)^{k-1}.$$

Segregating sites : Case when $\mathbb{E}(H_1 \wedge H_2)$ is finite

Site frequency spectrum of large samples

Assume that $\mathbb{E}(\min(H_1, H_2)) < \infty$, that is, $1/W^2$ is integrable.

Theorem

For all $k \geq 1$, the following convergence holds a.s. (and in L^1 as well if $\mathbb{E}(H) < \infty$)

$$\begin{aligned} \lim_{n \rightarrow \infty} n^{-1} S_n(k) &= \theta \mathbb{E}((\min\{H_1, H_{k+1}\} - \max\{H_2, \dots, H_k\})^+) \\ &= \theta \int_0^\infty \frac{dx}{W(x)^2} \left(1 - \frac{1}{W(x)}\right)^{k-1}. \end{aligned}$$

Yule case.

$$\lim_{n \rightarrow \infty} n^{-1} S_n(k) = \frac{\theta/c}{k(k+1)}$$

Segregating sites : Case when $\mathbb{E}(H)$ is infinite

Untreated case : Critical birth–death process

Particular case of $\mathbb{E}(H) = \infty$: the **critical birth–death** process.

Theorem

If $W(x) = 1 + cx$, the following convergence holds a.s.

$$\lim_{n \rightarrow \infty} n^{-1} S_n(k) = \frac{\theta/c}{k}$$

and the following one in probability

$$\lim_{n \rightarrow \infty} \frac{S_n}{n \ln(n)} = \theta/c.$$

Distinct haplotypes

The next branch with no extra mutations

K^θ := **first individual carrying no extra mutations** than individual 0
 $H^\theta := H_{K^\theta}$ is the corresponding branch length, with law

$$\mathbb{P}(H^\theta > x) = \frac{1}{W_\theta(x)} \quad x \geq 0.$$

The function W and W_θ are related through

$$W_\theta(x) = 1 + \int_0^x \frac{1}{W(u)} du$$

In particular $H^\theta \geq H$.

This result is straightforward for splitting trees.

Distinct haplotypes

The next branch with no extra mutations

K^θ := **first individual carrying no extra mutations** than individual 0
 $H^\theta := H_{K^\theta}$ is the corresponding branch length, with law

$$\mathbb{P}(H^\theta > x) = \frac{1}{W_\theta(x)} \quad x \geq 0.$$

The function W and W_θ are related through

$$W_\theta(x) = 1 + \int_0^x \frac{1}{W(u)} du$$

In particular $H^\theta \geq H$.

This result is straightforward for splitting trees.

The next branch with no extra mutations

K^θ := first individual carrying no extra mutations than individual 0
 $H^\theta := H_{K^\theta}$ is the corresponding branch length, with law

$$\mathbb{P}(H^\theta > x) =: \frac{1}{W_\theta(x)} \quad x \geq 0.$$

Proposition

The functions W and W_θ are related through

$$W_\theta(x) = 1 + \int_0^x W'(u) e^{-\theta u} du \quad x \geq 0.$$

In particular, $H^\theta \stackrel{stoch}{>} H$.

This result is straightforward for splitting trees.

Allele frequency spectrum for large samples

Theorem

For all $k \geq 1$, the following convergence holds a.s.

$$\lim_{n \rightarrow \infty} n^{-1} A_n(k) = \int_0^\infty dx \theta e^{-\theta x} \frac{1}{W_\theta(x)^2} \left(1 - \frac{1}{W_\theta(x)}\right)^{k-1}.$$

In particular,

$$\lim_{n \rightarrow \infty} n^{-1} A_n = \mathbb{E} \left(1 - e^{-\theta H^\theta}\right).$$

Critical birth–death process

Computations for the critical birth–death process :

Theorem

If $W(x) = 1 + cx$, the following convergences hold a.s.

$$\lim_{n \rightarrow \infty} n^{-1} A_n(k) = \frac{\theta/c}{k} (1 + \theta/c)^{-k}$$

$$\lim_{n \rightarrow \infty} n^{-1} A_n = \frac{\theta}{c} \ln(1 + c/\theta).$$

Fisher log-series of species abundance

In the Fisher log-series of species abundance,

- 1 the density of a given species is a Gamma r.v. with parameter a
- 2 given the value d of this density, X individuals of the species are spotted, where X is Poisson with parameter pd
- 3 as $a \downarrow 0$ conditional on $X \geq 1$, $\mathbb{P}(X = k)$ goes to $(1 + 1/\rho)^{-k}/k$.

Fisher log-series of species abundance

In the Fisher log-series of species abundance,

- 1 the density of a given species is a Gamma r.v. with parameter a
- 2 given the value d of this density, X individuals of the species are spotted, where X is Poisson with parameter pd
- 3 as $a \downarrow 0$ conditional on $X \geq 1$, $\mathbb{P}(X = k)$ goes to $(1 + 1/\rho)^{-k}/k$.

Fisher log-series of species abundance

In the Fisher log-series of species abundance,

- 1 the density of a given species is a Gamma r.v. with parameter a
- 2 given the value d of this density, X individuals of the species are spotted, where X is Poisson with parameter pd
- 3 as $a \downarrow 0$ conditional on $X \geq 1$, $\mathbb{P}(X = k)$ goes to $(1 + 1/\rho)^{-k}/k$.

Concluding comparisons between critical birth–death process and Kingman coalescent

In the critical **birth–death process**, as $n \rightarrow \infty$, (after rescaling θ)

$$S_n \sim \theta n \ln(n) \quad \text{and} \quad A_n \sim \theta \ln(1 + 1/\theta)n$$

whereas in the **Kingman coalescent**,

$$S_n \sim \theta \ln(n) \quad \text{and} \quad A_n \sim \theta \ln(n)$$

In the critical **birth–death process**, as $n \rightarrow \infty$,

$$S_n(k) \sim n\theta/k \quad \text{and} \quad A_n(k) \sim n\theta(1 + \theta)^{-k}/k$$

whereas in the **Kingman coalescent**,

$$S_n(k) \sim \theta/k \quad \text{and} \quad A_n(k) \sim \theta/k$$

Concluding comparisons between critical birth–death process and Kingman coalescent

In the critical **birth–death process**, as $n \rightarrow \infty$, (after rescaling θ)

$$S_n \sim \theta n \ln(n) \quad \text{and} \quad A_n \sim \theta \ln(1 + 1/\theta)n$$

whereas in the **Kingman coalescent**,

$$S_n \sim \theta \ln(n) \quad \text{and} \quad A_n \sim \theta \ln(n)$$

In the critical **birth–death process**, as $n \rightarrow \infty$,

$$S_n(k) \sim n\theta/k \quad \text{and} \quad A_n(k) \sim n\theta(1 + \theta)^{-k}/k$$

whereas in the **Kingman coalescent**,

$$S_n(k) \sim \theta/k \quad \text{and} \quad A_n(k) \sim \theta/k$$

...That's all, thanks for listening.