Probability of fixation of an advantageous allele in the "Bolthausen– Sznitman model".

Etienne Pardoux

Aix–Marseille Université

with B. Bah and S. Grusea

Contents

A formula in terms of the continuous time selection parameter α

The Wright–Fisher model with selection

- Consider the classical Wright–Fisher Markov chain which describes the proportion of an advantageous allele in a population of fixed size made of two types, say A and a, and A has a per generation selective advantage s over a.
- If N denotes the size of the population, $\alpha = s\,N$, and X_t^N denotes the proportion of the advantageous allele A in the population at generation $[Nt]$, then as $N \to \infty$, $X^N \Rightarrow X$, where X solves the SDE

$$X_t = x + \alpha \int_0^t X_s(1 - X_s)ds + \int_0^t \sqrt{X_s(1 - X_s)}dB_s.$$

The Wright–Fisher model with selection

- Consider the classical Wright–Fisher Markov chain which describes the proportion of an advantageous allele in a population of fixed size made of two types, say A and a, and A has a per generation selective advantage s over a.
- If N denotes the size of the population, $\alpha = s\,N$, and X_t^N denotes the proportion of the advantageous allele A in the population at generation $[Nt]$, then as $N \to \infty$, $X^N \Rightarrow X$, where X solves the SDE

$$X_t = x + \alpha \int_0^t X_s(1 - X_s)ds + \int_0^t \sqrt{X_s(1 - X_s)}dB_s.$$

Malécot's formula

- In this model one of the two alleles fixates in finite time, i.e.

$$T_{\text{fix}} = \inf\{t > 0,\ X_t \in \{0, 1\}\} < \infty \text{ a.s.}$$

and the probability that type A fixates is given by Malécot's formula (usually attributed to Kimura)

$$\mathbb{P}(X_{T_{\text{fix}}} = 1) = \frac{1 - e^{-2\alpha x}}{1 - e^{-2\alpha}}.$$

- Note that the genealogy of the population in this model is Kingman's coalescent, and the fact that one of the two alleles fixates in finite time is due to the fact that Kingman's coalescent "*comes down from infinity*", i.e. a countable number of individuals has finitely many ancestors at any time $t > 0$ back in time. Equivalently, forward in time soon or later the progeny of a single individual will invade the whole population.

Malécot's formula

- In this model one of the two alleles fixates in finite time, i.e.

$$T_{\text{fix}} = \inf\{t > 0,\ X_t \in \{0, 1\}\} < \infty \text{ a.s.}$$

and the probability that type A fixates is given by Malécot's formula (usually attributed to Kimura)

$$\mathbb{P}(X_{T_{\text{fix}}} = 1) = \frac{1 - e^{-2\alpha x}}{1 - e^{-2\alpha}}.$$

- Note that the genealogy of the population in this model is Kingman's coalescent, and the fact that one of the two alleles fixates in finite time is due to the fact that Kingman's coalescent "*comes down from infinity*", i.e. a countable number of individuals has finitely many ancestors at any time $t > 0$ back in time. Equivalently, forward in time soon or later the progeny of a single individual will invade the whole population.

Another model

- It is possible to choose a variant of the above model, such that the limiting proportion as $N \to \infty$ of advantageous alleles A obeys the less classical SDE

$$X_t = x + \alpha \int_0^t X_s(1 - X_s)ds + \int_{[0,t]\times(0,1)^2} p\Psi(u, X_{s^-})\bar{M}(ds, du, dp),$$

- where

$$\Psi(u, r) = \mathbf{1}_{u \le r} - r,$$
$$\bar{M}(ds, du, dp) = M(ds, du, dp) - p^{-2}\, ds\, du\, dp,$$

- and M is a Poisson Random Measure on $\mathbb{R}_+ \times (0,1)^2$ with mean $\nu(ds, du, dp) = p^{-2}\, ds\, du\, dp$, i.e. a random collection of points such that the number of points in any subset $A \subset \mathbb{R}_+ \times (0,1)^2$ is Poisson distributed, with parameter $\int_A p^{-2}\, ds\, du\, dp$, and the number of points in disjoint subsets are mutually independent.

Another model

- It is possible to choose a variant of the above model, such that the limiting proportion as $N \to \infty$ of advantageous alleles A obeys the less classical SDE

$$X_t = x + \alpha \int_0^t X_s(1 - X_s)ds + \int_{[0,t]\times(0,1)^2} p\Psi(u, X_{s^-})\bar{M}(ds, du, dp),$$

- where

$$\Psi(u, r) = \mathbf{1}_{u \leq r} - r,$$
$$\bar{M}(ds, du, dp) = M(ds, du, dp) - p^{-2}\, ds\, du\, dp,$$

- and M is a Poisson Random Measure on $\mathbb{R}_+ \times (0,1)^2$ with mean $\nu(ds, du, dp) = p^{-2}\, ds\, du\, dp$, i.e. a random collection of points such that the number of points in any subset $A \subset \mathbb{R}_+ \times (0,1)^2$ is Poisson distributed, with parameter $\int_A p^{-2}\, ds\, du\, dp$, and the number of points in disjoint subsets are mutually independent.

Another model

- It is possible to choose a variant of the above model, such that the limiting proportion as $N \to \infty$ of advantageous alleles A obeys the less classical SDE

$$X_t = x + \alpha \int_0^t X_s(1 - X_s)ds + \int_{[0,t]\times(0,1)^2} p\Psi(u, X_{s^-})\bar{M}(ds, du, dp),$$

- where

$$\Psi(u, r) = \mathbf{1}_{u \leq r} - r,$$
$$\bar{M}(ds, du, dp) = M(ds, du, dp) - p^{-2}\, ds\, du\, dp,$$

- and M is a Poisson Random Measure on $\mathbb{R}_+ \times (0, 1)^2$ with mean $\nu(ds, du, dp) = p^{-2}\, ds\, du\, dp$, i.e. a random collection of points such that the number of points in any subset $A \subset \mathbb{R}_+ \times (0, 1)^2$ is Poisson distributed, with parameter $\int_A p^{-2}\, ds\, du\, dp$, and the number of points in disjoint subsets are mutually independent.

Bolthausen–Sznitman

- The genealogy of a population associated to the above SDE is not described by Kingman's coalescent, but by the Bolthausen–Sznitman coalescent.
- The Bolthausen–Sznitman coalescent has the property that while there are k lineages in the genealogy, any subset of ℓ of them coalesces at rate

$$\lambda_{k,\ell} = \int_0^1 p^{\ell-2}(1-p)^{k-\ell} dp.$$

- The Bolthausen–Sznitman coalescent belongs to the class of the Λ–coalescents, which are defined in the exact same way, except that we repalce dp by $\Lambda(dp)$, where Λ is an arbitrary measure on $[0,1)$, Kingman's coalescent being the special case where $\Lambda = c\delta_0$, in which case only binary merges happen.

Bolthausen–Sznitman

- The genealogy of a population associated to the above SDE is not described by Kingman's coalescent, but by the Bolthausen–Sznitman coalescent.
- The Bolthausen–Sznitman coalescent has the property that while there are k lineages in the genealogy, any subset of ℓ of them coalesces at rate

$$\lambda_{k,\ell} = \int_0^1 p^{\ell-2}(1-p)^{k-\ell} dp.$$

- The Bolthausen–Sznitman coalescent belongs to the class of the Λ–coalescents, which are defined in the exact same way, except that we repalce dp by $\Lambda(dp)$, where Λ is an arbitrary measure on $[0,1)$, Kingman's coalescent being the special case where $\Lambda = c\delta_0$, in which case only binary merges happen.

Bolthausen–Sznitman

- The genealogy of a population associated to the above SDE is not described by Kingman's coalescent, but by the Bolthausen–Sznitman coalescent.
- The Bolthausen–Sznitman coalescent has the property that while there are k lineages in the genealogy, any subset of ℓ of them coalesces at rate

$$\lambda_{k,\ell} = \int_0^1 p^{\ell-2}(1-p)^{k-\ell} dp.$$

- The Bolthausen–Sznitman coalescent belongs to the class of the Λ–coalescents, which are defined in the exact same way, except that we repalce dp by $\Lambda(dp)$, where Λ is an arbitrary measure on $[0,1)$, Kingman's coalescent being the special case where $\Lambda = c\delta_0$, in which case only binary merges happen.

Our first formula

- The Bolthausen–Sznitman coalescent does not come down from infinity. As a consequence, in our new model we do not have fixation in finite time. In other words, $T_{\text{fix}} = +\infty$ a.s.
- However, it is easy to see that $X_t \to X_\infty$ a.s., where $X_\infty \in \{0, 1\}$. But $0 < X_t < 1$ a.s. for all $t > 0$ (of course provided that $0 < x < 1$).
- Our first formula reads

$$\mathbb{P}_{BS}(X_\infty = 1) = \frac{xe^\alpha}{xe^\alpha + (1-x)}.$$

Our first formula

- The Bolthausen–Sznitman coalescent does not come down from infinity. As a consequence, in our new model we do not have fixation in finite time. In other words, $T_{\text{fix}} = +\infty$ a.s.
- However, it is easy to see that $X_t \to X_\infty$ a.s., where $X_\infty \in \{0, 1\}$. But $0 < X_t < 1$ a.s. for all $t > 0$ (of course provided that $0 < x < 1$).
- Our first formula reads

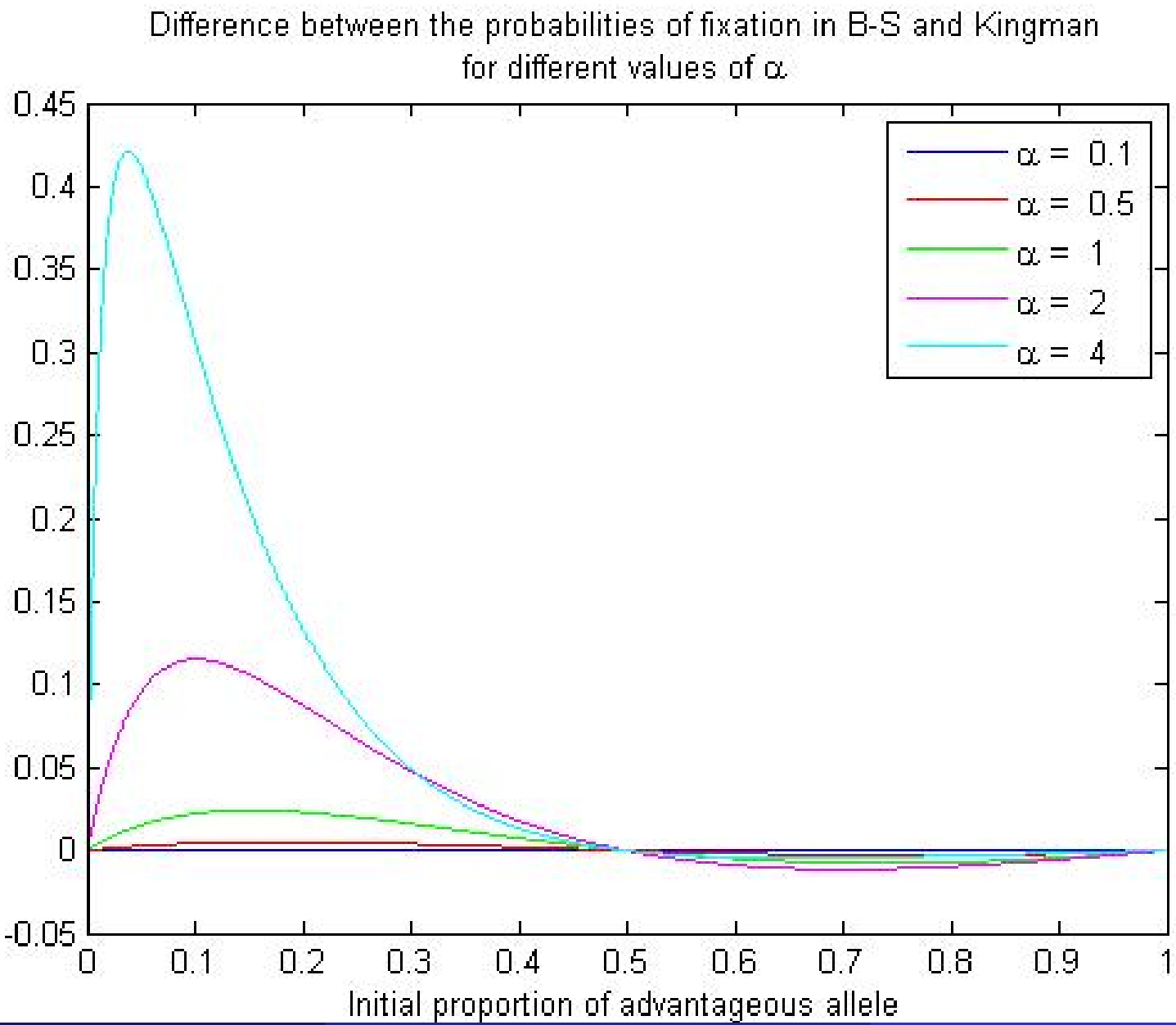
$$\mathbb{P}_{BS}(X_\infty = 1) = \frac{xe^\alpha}{xe^\alpha + (1-x)}.$$

Our first formula

- The Bolthausen–Sznitman coalescent does not come down from infinity. As a consequence, in our new model we do not have fixation in finite time. In other words, $T_{\text{fix}} = +\infty$ a.s.
- However, it is easy to see that $X_t \to X_\infty$ a.s., where $X_\infty \in \{0, 1\}$. But $0 < X_t < 1$ a.s. for all $t > 0$ (of course provided that $0 < x < 1$).
- Our first formula reads

$$\mathbb{P}_{BS}(X_\infty = 1) = \frac{xe^\alpha}{xe^\alpha + (1-x)}.$$

Comparison with Kimura's formula



Idea of the proof

- Let K_t denote the number of active lineages in the ASG at time t.
- K_t is positive recurrent, and has a unique invariant probability measure, which is the geometric distribution with parameter $e^{-\alpha}$.
- We have the duality relation (a similar relation in the case of the Kingman model can be found e.g. in Pokalyuk, Pfaffelhuber (2013)).

$$\mathbb{E}[(1-X_t)^k | X_0 = x] = \mathbb{E}[(1-x)^{K_t} | K_0 = k].$$

Note that the duality is between $Y_t = 1 - X_t$ and K_t.

- The result follows by letting $t \to \infty$ in the above identity, hence

$$\mathbb{P}(X_\infty = 0 | X_0 = x) = \mathbb{E}((1-x)^{K_\infty}).$$

Idea of the proof

- Let K_t denote the number of active lineages in the ASG at time t.
- K_t is positive recurrent, and has a unique invariant probability measure, which is the geometric distribution with parameter $e^{-\alpha}$.
- We have the duality relation (a similar relation in the case of the Kingman model can be found e.g. in Pokalyuk, Pfaffelhuber (2013)).

$$\mathbb{E}[(1-X_t)^k | X_0 = x] = \mathbb{E}[(1-x)^{K_t} | K_0 = k].$$

Note that the duality is between $Y_t = 1 - X_t$ and K_t.

- The result follows by letting $t \to \infty$ in the above identity, hence

$$\mathbb{P}(X_\infty = 0 | X_0 = x) = \mathbb{E}((1-x)^{K_\infty}).$$

Idea of the proof

- Let K_t denote the number of active lineages in the ASG at time t.
- K_t is positive recurrent, and has a unique invariant probability measure, which is the geometric distribution with parameter $e^{-\alpha}$.
- We have the duality relation (a similar relation in the case of the Kingman model can be found e.g. in Pokalyuk, Pfaffelhuber (2013)).

$$\mathbb{E}[(1-X_t)^k | X_0 = x] = \mathbb{E}[(1-x)^{K_t} | K_0 = k].$$

Note that the duality is between $Y_t = 1 - X_t$ and K_t.

- The result follows by letting $t \to \infty$ in the above identity, hence

$$\mathbb{P}(X_\infty = 0 | X_0 = x) = \mathbb{E}((1-x)^{K_\infty}).$$

Idea of the proof

- Let K_t denote the number of active lineages in the ASG at time t.
- K_t is positive recurrent, and has a unique invariant probability measure, which is the geometric distribution with parameter $e^{-\alpha}$.
- We have the duality relation (a similar relation in the case of the Kingman model can be found e.g. in Pokalyuk, Pfaffelhuber (2013)).

$$\mathbb{E}[(1-X_t)^k | X_0 = x] = \mathbb{E}[(1-x)^{K_t} | K_0 = k].$$

Note that the duality is between $Y_t = 1 - X_t$ and K_t.

- The result follows by letting $t \to \infty$ in the above identity, hence

$$\mathbb{P}(X_\infty = 0 | X_0 = x) = \mathbb{E}((1-x)^{K_\infty}).$$

A formula in terms of the per generation selection parameter s

The Wright–Fisher model

- In the Wright–Fisher model, the per generation selection parameter s is interpreted as follows. $1+s$ is the ratio of the probability of choosing an advantageous father to the probability of choosing a non–advantageous father in the previous generation.
- As indicated above, in order to to take a continuous time limit of the Wright–Fisher model we scale time by considering a length of continuous time $1/N$ between two consecutive generations, and let $s = s_N = \frac{\alpha}{N}$.
- It is then natural to use Malécot's formula $\frac{1-e^{-2x\alpha}}{1-e^{-2\alpha}}$ with $\alpha = s\,N$, and $x = 1/N$, in order to compute the probability of fixation of a new advantageous allele.
- We get

$$\frac{1-e^{-2s}}{1-e^{-2sN}} \simeq 1-e^{-2s},$$

a formula which is well–known in population genetics.

The Wright–Fisher model

- In the Wright–Fisher model, the per generation selection parameter s is interpreted as follows. $1+s$ is the ratio of the probability of choosing an advantageous father to the probability of choosing a non–advantageous father in the previous generation.
- As indicated above, in order to to take a continuous time limit of the Wright–Fisher model we scale time by considering a length of continuous time $1/N$ between two consecutive generations, and let $s = s_N = \frac{\alpha}{N}$.
- It is then natural to use Malécot's formula $\frac{1-e^{-2x\alpha}}{1-e^{-2\alpha}}$ with $\alpha = s\,N$, and $x = 1/N$, in order to compute the probability of fixation of a new advantageous allele.
- We get

$$\frac{1-e^{-2s}}{1-e^{-2sN}} \simeq 1-e^{-2s},$$

a formula which is well–known in population genetics.

The Wright–Fisher model

- In the Wright–Fisher model, the per generation selection parameter s is interpreted as follows. $1+s$ is the ratio of the probability of choosing an advantageous father to the probability of choosing a non–advantageous father in the previous generation.
- As indicated above, in order to to take a continuous time limit of the Wright–Fisher model we scale time by considering a length of continuous time $1/N$ between two consecutive generations, and let $s = s_N = \frac{\alpha}{N}$.
- It is then natural to use Malécot's formula $\frac{1-e^{-2x\alpha}}{1-e^{-2\alpha}}$ with $\alpha = s\,N$, and $x = 1/N$, in order to compute the probability of fixation of a new advantageous allele.
- We get

$$\frac{1-e^{-2s}}{1-e^{-2sN}} \simeq 1-e^{-2s},$$

a formula which is well–known in population genetics.

The Wright–Fisher model

- In the Wright–Fisher model, the per generation selection parameter s is interpreted as follows. $1+s$ is the ratio of the probability of choosing an advantageous father to the probability of choosing a non–advantageous father in the previous generation.
- As indicated above, in order to to take a continuous time limit of the Wright–Fisher model we scale time by considering a length of continuous time $1/N$ between two consecutive generations, and let $s = s_N = \frac{\alpha}{N}$.
- It is then natural to use Malécot's formula $\frac{1-e^{-2x\alpha}}{1-e^{-2\alpha}}$ with $\alpha = s\,N$, and $x = 1/N$, in order to compute the probability of fixation of a new advantageous allele.
- We get

$$\frac{1-e^{-2s}}{1-e^{-2sN}} \simeq 1-e^{-2s},$$

  a formula which is well–known in population genetics.

Approximating the Bolthausen–Sznitman model by a generation model

- We follow a construction due to Schweinsberg (2003). We consider a population of fixed size N, whose genealogy from generation to generation is described as follows.
- Let $X_1, \ldots, X_N$ be i.i.d. $\mathbb{N}$–valued r.v.'s be such that $\mathbb{P}(X_1 \geq k) = 1/k$, $k \geq 1$. For $1 \leq i \leq N$, X_i is the number of children of i. Since $\mathbb{E}X_1 = +\infty$, with probability almost 1, $X_1 + \cdots + X_N \geq N$.
- We obtain the next generation by sampling N of the $X_1 + \cdots + X_N$ individuals uniformly without replacement. The offsprings of i in the next generation are those individuals among the X_i's which are sampled (the others die).

Approximating the Bolthausen–Sznitman model by a generation model

- We follow a construction due to Schweinsberg (2003). We consider a population of fixed size N, whose genealogy from generation to generation is described as follows.
- Let $X_1, \ldots, X_N$ be i.i.d. $\mathbb{N}$–valued r.v.'s be such that $\mathbb{P}(X_1 \geq k) = 1/k$, $k \geq 1$. For $1 \leq i \leq N$, X_i is the number of children of i. Since $\mathbb{E}X_1 = +\infty$, with probability almost 1, $X_1 + \cdots + X_N \geq N$.
- We obtain the next generation by sampling N of the $X_1 + \cdots + X_N$ individuals uniformly without replacement. The offsprings of i in the next generation are those individuals among the X_i's which are sampled (the others die).

Approximating the Bolthausen–Sznitman model by a generation model

- We follow a construction due to Schweinsberg (2003). We consider a population of fixed size N, whose genealogy from generation to generation is described as follows.
- Let $X_1, \ldots, X_N$ be i.i.d. $\mathbb{N}$–valued r.v.'s be such that $\mathbb{P}(X_1 \geq k) = 1/k$, $k \geq 1$. For $1 \leq i \leq N$, X_i is the number of children of i. Since $\mathbb{E}X_1 = +\infty$, with probability almost 1, $X_1 + \cdots + X_N \geq N$.
- We obtain the next generation by sampling N of the $X_1 + \cdots + X_N$ individuals uniformly without replacement. The offsprings of i in the next generation are those individuals among the X_i's which are sampled (the others die).

Convergence to the Bolthausen–Sznitman coalescent

- We now sample $n < N$ individuals in the present generation 0, and define a discrete–time Markov chain $\Psi_{n,N}(m)_{m\geq 0}$ with values in the set $\mathcal{P}_n$ of partitions of the set $\{1,\ldots,n\}$. i and j are in the same block of $\Psi_{n,N}(m)$ iff the individuals i and j have the same ancestor in generation $-m$.
- We have the following result from Schweinsberg (2003).

Proposition

The process $\{\Psi_{n,N}([t\log N]),\ t\geq 0\}$ converges in law as $N\to\infty$ towards $\{\Psi_{n,\infty}(t),\ t\geq 0\}$, which is the restriction to $\{1,\ldots,n\}$ of the Bolthausen–Sznitman coalescent.

- Let us comment on the time scale. In the above discrete–time model, the mean number of generations we should follow backwards for two randomly chosen individuals to find a common ancestor is $\log N$. Consequently the above scaling is such that in the limit two individuals find a common ancestor in mean time 1.

Convergence to the Bolthausen–Sznitman coalescent

- We now sample $n < N$ individuals in the present generation 0, and define a discrete–time Markov chain $\Psi_{n,N}(m)_{m\geq 0}$ with values in the set $\mathcal{P}_n$ of partitions of the set $\{1,\ldots,n\}$. i and j are in the same block of $\Psi_{n,N}(m)$ iff the individuals i and j have the same ancestor in generation $-m$.
- We have the following result from Schweinsberg (2003).

Proposition

The process $\{\Psi_{n,N}([t\log N]),\ t\geq 0\}$ *converges in law as* $N\to\infty$ *towards* $\{\Psi_{n,\infty}(t),\ t\geq 0\}$, *which is the restriction to* $\{1,\ldots,n\}$ *of the Bolthausen–Sznitman coalescent.*

- Let us comment on the time scale. In the above discrete–time model, the mean number of generations we should follow backwards for two randomly chosen individuals to find a common ancestor is $\log N$. Consequently the above scaling is such that in the limit two individuals find a common ancestor in mean time 1.

Convergence to the Bolthausen–Sznitman coalescent

- We now sample $n < N$ individuals in the present generation 0, and define a discrete–time Markov chain $\Psi_{n,N}(m)_{m\geq 0}$ with values in the set $\mathcal{P}_n$ of partitions of the set $\{1,\ldots,n\}$. i and j are in the same block of $\Psi_{n,N}(m)$ iff the individuals i and j have the same ancestor in generation $-m$.
- We have the following result from Schweinsberg (2003).

Proposition

The process $\{\Psi_{n,N}([t\log N]),\ t\geq 0\}$ *converges in law as* $N\to\infty$ *towards* $\{\Psi_{n,\infty}(t),\ t\geq 0\}$, *which is the restriction to* $\{1,\ldots,n\}$ *of the Bolthausen–Sznitman coalescent.*

- Let us comment on the time scale. In the above discrete–time model, the mean number of generations we should follow backwards for two randomly chosen individuals to find a common ancestor is $\log N$. Consequently the above scaling is such that in the limit two individuals find a common ancestor in mean time 1.

Discrete time approximation of the ASG

- We now describe a process which converges, as $N \to \infty$, towards the n–Bolthausen–Sznitman Ancestral Selection Graph. Start with n lines at time $t = 0$.
- We alternate the two following procedures at each discrete time step.
  1. With probability $s_N = \alpha / \log(N)$, each line splits into 2, independently of all other lines.
  2. The existing lines (whose number is smaller than N, since $n << N$) coalesce according to the Schweinsberg procedure, those lines being e.g. the first of the N lines in Scheinsberg construction.
- $s = s_N$ can be interpreted as a selection coefficient per generation. Indeed, the effect of the first step is that above each individual, there is one branch with probability $1 - s$, and there are two branches with probability s. The rule in the ASG is that the individual is A, i.e. carries the selective allele if at least one branch above him is A. From exchangeability, the fact that N is large and de Finetti's theorem, the probability of being A becomes after step 1

$$X_t(1-s) + sX_t(1+Y_t) = X_t + sX_tY_t = X_t + sX_t(1-X_t).$$

Discrete time approximation of the ASG

- We now describe a process which converges, as $N \to \infty$, towards the n–Bolthausen–Sznitman Ancestral Selection Graph. Start with n lines at time $t = 0$.
- We alternate the two following procedures at each discrete time step.
  1. With probability $s_N = \alpha / \log(N)$, each line splits into 2, independently of all other lines.
  2. The existing lines (whose number is smaller than N, since $n << N$) coalesce according to the Schweinsberg procedure, those lines being e.g. the first of the N lines in Scheinsberg construction.
- $s = s_N$ can be interpreted as a selection coefficient per generation. Indeed, the effect of the first step is that above each individual, there is one branch with probability $1 - s$, and there are two branches with probability s. The rule in the ASG is that the individual is A, i.e. carries the selective allele if at least one branch above him is A. From exchangeability, the fact that N is large and de Finetti's theorem, the probability of being A becomes after step 1

$$X_t(1-s) + sX_t(1+Y_t) = X_t + sX_tY_t = X_t + sX_t(1-X_t).$$

Discrete time approximation of the ASG

- We now describe a process which converges, as $N \to \infty$, towards the n–Bolthausen–Sznitman Ancestral Selection Graph. Start with n lines at time $t = 0$.
- We alternate the two following procedures at each discrete time step.
  1. With probability $s_N = \alpha / \log(N)$, each line splits into 2, independently of all other lines.
  2. The existing lines (whose number is smaller than N, since $n << N$) coalesce according to the Schweinsberg procedure, those lines being e.g. the first of the N lines in Scheinsberg construction.
- $s = s_N$ can be interpreted as a selection coefficient per generation. Indeed, the effect of the first step is that above each individual, there is one branch with probability $1 - s$, and there are two branches with probability s. The rule in the ASG is that the individual is A, i.e. carries the selective allele if at least one branch above him is A. From exchangeability, the fact that N is large and de Finetti's theorem, the probability of being A becomes after step 1

$$X_t(1-s) + sX_t(1+Y_t) = X_t + sX_tY_t = X_t + sX_t(1-X_t).$$

Convergence to the ASG

- The above construction produces an $\mathbb{N}$–valued Markov chain $\{\tilde{\Psi}_{n,N}(\ell),\ \ell \geq 0\}$. Define $K_t^N = |\tilde{\Psi}_{n,N}([t \log N])|$, the number of lineages in $\tilde{\Psi}_{n,N}([t \log N])$.
- Recall the process K_t which counts the number of active branches in the ASG and has been defined above.
- We have

Proposition

$\{K_t^N,\ t \geq 0\}$ converges in law, as $N \to \infty$, towards $\{K_t,\ t \geq 0\}$.

The proof is easy, given Schweinsberg's Proposition.

Convergence to the ASG

- The above construction produces an $\mathbb{N}$–valued Markov chain $\{\tilde{\Psi}_{n,N}(\ell),\ \ell \geq 0\}$. Define $K_t^N = |\tilde{\Psi}_{n,N}([t \log N])|$, the number of lineages in $\tilde{\Psi}_{n,N}([t \log N])$.
- Recall the process K_t which counts the number of active branches in the ASG and has been defined above.
- We have

Proposition

$\{K_t^N,\ t \geq 0\}$ converges in law, as $N \to \infty$, towards $\{K_t,\ t \geq 0\}$.

The proof is easy, given Schweinsberg's Proposition.

Convergence to the ASG

- The above construction produces an $\mathbb{N}$–valued Markov chain $\{\tilde{\Psi}_{n,N}(\ell),\ \ell \geq 0\}$. Define $K_t^N = |\tilde{\Psi}_{n,N}([t \log N])|$, the number of lineages in $\tilde{\Psi}_{n,N}([t \log N])$.
- Recall the process K_t which counts the number of active branches in the ASG and has been defined above.
- We have

Proposition

$\{K_t^N,\ t \geq 0\}$ converges in law, as $N \to \infty$, towards $\{K_t,\ t \geq 0\}$.

The proof is easy, given Schweinsberg's Proposition.

Probability of fixation of a new mutation

- We consider that at time $t = 0$ a unique individual in the population carries the new mutation. Consequently, with $s = \alpha / \log N$, we have

Theorem

Consider a population of size N which evolves according to the Bolthausen–Sznitman model, where a mutation occurs in one individual, which confers him a selective advantage s, then the probability of fixation of that new mutation is approximately given by

$$\mathbb{P}_{BS}(X_\infty = 1 | X_0 = 1/N) = \frac{1}{1 + N^{1-s} - N^{-s}} \simeq \frac{1}{1 + N^{1-s}}.$$

- Proof : This formula from the first with $x = 1/N$ and $\alpha = s \log N$:

$$\mathbb{P}_{BS}(X_\infty = 1 | X_0 = 1/N) = \frac{N^{-1} e^{s \log(N)}}{N^{-1} e^{s \log(N)} + 1 - N^{-1}} = \frac{N^{s-1}}{N^{s-1} + 1 - N^{-1}}.$$

Probability of fixation of a new mutation

- We consider that at time $t = 0$ a unique individual in the population carries the new mutation. Consequently, with $s = \alpha / \log N$, we have

Theorem

Consider a population of size N which evolves according to the Bolthausen–Sznitman model, where a mutation occurs in one individual, which confers him a selective advantage s, then the probability of fixation of that new mutation is approximately given by

$$\mathbb{P}_{BS}(X_\infty = 1 | X_0 = 1/N) = \frac{1}{1 + N^{1-s} - N^{-s}} \simeq \frac{1}{1 + N^{1-s}}.$$

- Proof : This formula from the first with $x = 1/N$ and $\alpha = s \log N$:

$$\begin{aligned}\mathbb{P}_{BS}(X_\infty = 1 | X_0 = 1/N) &= \frac{N^{-1} e^{s \log(N)}}{N^{-1} e^{s \log(N)} + 1 - N^{-1}} \\ &= \frac{N^{s-1}}{N^{s-1} + 1 - N^{-1}}.\end{aligned}$$

Comment on the second formula

- We note that when $s = 0$ the above probability equals $1/N$, as it should. For all $0 \leq s < 1$ (resp. $s > 1$), that probability tends to 0 (resp. to 1) as $N \to \infty$, while it tend to $1/2$ if $s = 1$. We see that the only non trivial value for large N is obtained when s is close to 1.
- The behaviour for $s > 1$ can be compared to the following result of Foucart (2013), Griffiths (2014). If we consider a model associated to a general Λ–coalescent, fixation of the advantageous allele happens with probability one, irrespective of its initial proportion, iff $\alpha \geq \alpha^*$, where

$$\alpha^* = -\int_0^1 \frac{\log(1-p)}{p^2} \Lambda(dp).$$

Note that in the BS (Bolthausen–Sznitman) model as well as in the K (Kingman) model, $\alpha^* = +\infty$.

Comment on the second formula

- We note that when $s = 0$ the above probability equals $1/N$, as it should. For all $0 \leq s < 1$ (resp. $s > 1$), that probability tends to 0 (resp. to 1) as $N \to \infty$, while it tend to $1/2$ if $s = 1$. We see that the only non trivial value for large N is obtained when s is close to 1.
- The behaviour for $s > 1$ can be compared to the following result of Foucart (2013), Griffiths (2014). If we consider a model associated to a general Λ–coalescent, fixation of the advantageous allele happens with probability one, irrespective of its initial proportion, iff $\alpha \geq \alpha^*$, where

$$\alpha^* = -\int_0^1 \frac{\log(1-p)}{p^2} \Lambda(dp).$$

Note that in the BS (Bolthausen–Sznitman) model as well as in the K (Kingman) model, $\alpha^* = +\infty$.

Conclusion

- If we believe that we should compare the BS model and the K model for the same value of α, and if the initial proportion of the advantageous allele is smaller than $1/2$, then the probability of fixation is significantly larger in the BS model. However, if we believe that the parameter which is the same in both models is rather s, then α in the BS model would be much smaller for large N than in the K model, which would make the above comparison meaningless.
- In the BS model, the probability of fixation in terms of the per generation selective advantage s depends heavily upon N, in contradiction with the K model. A mutation with a small s fixates much easier in the K model, while a mutation with a large effect ($s > 1$) fixates with a probability close to 1 in the BS model. This may have important impact concerning the evolution of parasites and pathogens. A population which follows the BS model is likely to see few small effect mutations, but relatively many mutations with large effect.

Conclusion

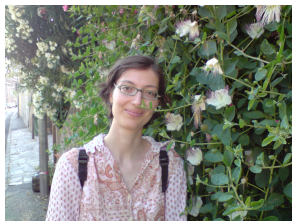
- If we believe that we should compare the BS model and the K model for the same value of α, and if the initial proportion of the advantageous allele is smaller than $1/2$, then the probability of fixation is significantly larger in the BS model. However, if we believe that the parameter which is the same in both models is rather s, then α in the BS model would be much smaller for large N than in the K model, which would make the above comparison meaningless.
- In the BS model, the probability of fixation in terms of the per generation selective advantage s depends heavily upon N, in contradiction with the K model. A mutation with a small s fixates much easier in the K model, while a mutation with a large effect ($s > 1$) fixates with a probability close to 1 in the BS model. This may have important impact concerning the evolution of parasites and pathogens. A population which follows the BS model is likely to see few small effect mutations, but relatively many mutations with large effect.

Acknowledgements

- My coauthors

  Boubacar Bah (AIMS Cameroun)

  Simona Grusea (INSA Toulouse)



- Michael Kopp (AMU), who asked the question about the per generation selective advantage, and helped us understand the content of the second formula.

Bibliography

- B. Bah and E. Pardoux, The Λ–lookdown model with selection, *Stochastic Processes and Their Applications* **125** 1089–1126, 2015.
- C. Foucart, The impact of selection in the Λ–Wright–Fisher model, *Electron. Commun. Probab.*, **18**, 1–10, 2013.
- R. Griffiths, The Lambda-Fleming-Viot process and a connection with Wright-Fisher diffusion, *Adv. in Appl. Probab.* **4**, 1009–1035, 2014.
- G. Malécot , Les processus stochastiques et la méthode des fonctions génératrices ou caractéristiques, *Publ. Inst. Stat. Univ. Paris* **1**, 3, 1–16, 1952.
- C. Pokalyuk, P. Pfaffelhuber, The ancestral selection graph under strong directional selection, *Theoretical Population Biology* **87** 25–33, 2013.
- J. Schweinsberg, Coalescent processes obtained from supercritical Galton–Watson processes, *Stochastic Process. Appl.* **106**, 107–139, 2003.

THANK YOU FOR YOUR ATTENTION !