Haldane's formula in Cannings models with moderate selection

Cornelia Pokalyuk
joint work in progress with Florin Boenkost,
Adrian González Casanova, Anton Wakolbinger

GOETHE
UNIVERSITÄT
FRANKFURT AM MAIN

Haldane's formula

Haldane's formula is a rough-and-ready rule
for the fixation probability π of a slightly advantageous allele
initially present in a single individual in a large population:

$$\pi \approx \frac{s}{v/2}$$

with s the selective advantage,
and v the individual offspring variance per generation.

Spreading like a branching process

In a large population an advantageous allele spreads initially like a supercritical branching process. The survival probability of a (slightly) supercritical branching process obeys Haldane's formula:

Theorem (Athreya 1992)

Let $(Z^{(N)})_{N\geq 1}$ be a sequence of supercritical Galton-Watson processes with offspring mean $1 + s_N \to 1$, offspring variance $v_N \to v$ and uniformly bounded third moment, starting with a single individual in generation 0.
Then the survival probability π_N of $Z^{(N)}$ fulfills

$$\pi_N \sim \frac{s_N}{v/2}.$$

Diffusion setting

$s_N = \frac{\alpha}{N}$, x initial frequency of advantageous allele

Kimura's formula

$$\pi = \frac{1 - e^{-2\alpha x}}{1 - e^{-2\alpha}}.$$

Set $x = \frac{1}{N}$

$$\frac{1 - e^{-2\alpha x}}{1 - e^{-2\alpha}} = \frac{1 - e^{-2s_N}}{1 - e^{-2\alpha}} \sim \frac{2s_N}{1 - e^{-2\alpha}}$$

Today

- **Moderate selection $s_N \sim N^{-b}$ for $0 < b < 1$**
  Covers a broad range of selection strengths
- **Discrete time models**
  Model populations with seasonal behaviour, experimental evolution, ...
  Cannings models with selection

Neutral Cannings model

- constant population size N
- discrete generations $g = 1, 2, \ldots$
- exchangeable offspring numbers
  - each child chooses a parent in the previous generation
  - $\xi_i^{(g)}$ - the number of children of the i-th parent in generation g, $i = 1, \ldots, N$
  - $\sum_{i=1}^{N} \xi_i^{(g)} = N$
  - $\xi_1^{(g)}, \ldots, \xi_N^{(g)}$ exchangeable.
- time homogeneous: $(\xi^{(g)})_{g=1,2,\ldots}$ are iid

Most prominent Cannings model: Wright-Fisher model

Parents are chosen independently and uniformly among all individuals in the previous generation

$$(\xi_1^{(g)}, \ldots, \xi_N^{(g)}) \stackrel{d}{=} \text{Multinom}(N; 1/N, \ldots, 1/N).$$

A subclass of neutral Cannings models

For each g let $\mathcal{W}^{(g)} = (W_1^{(g)}, \ldots, W_N^{(g)})$ be a random exchangeable *mass partition of* 1, that is

$$W_i^{(g)} \geq 0 \qquad \text{and} \qquad \sum_{i=1}^{N} W_i^{(g)} = 1.$$

Assume $(\mathcal{W}^{(g)})_{g=0,1,\ldots}$ are iid.

$$(\xi_1^{(g)}, \ldots, \xi_N^{(g)}) \stackrel{d}{=} \text{Multinom}(N; W_1^{(g)}, \ldots, W_N^{(g)}).$$

These models mimic e.g. the following reproduction cycle

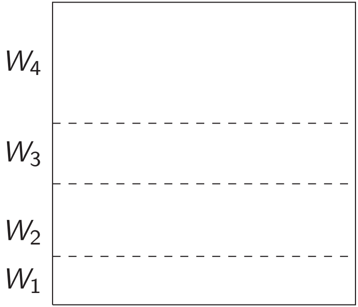
- Assume parents $i = 1, ..., N$ have large iid random numbers $Y_1, ..., Y_N$ of offspring in generation g.
- Only N of these offspring survive till generation $g + 1$.
- The individuals (the children) which survive are drawn randomly from all offspring.
- If the number of offspring is large, drawing with and without replacement is approximately the same.
- Hence, the probability that the i-th child is an offspring parent j is approximately

$$W_j = \frac{Y_j}{Y_1 + ... + Y_N}$$

for $i = 1, ..., N$.

A graphical description of the Cannings model

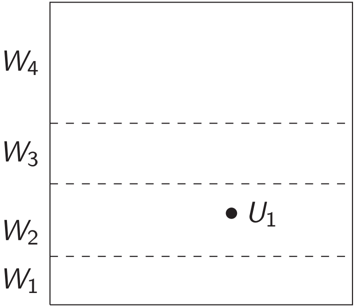
$\mathbb{P}($ Child i chooses j as parent $|\mathcal{W}) = W_j$, for $i = 1, ..., N$ independently.



with $U_i \sim$Unif$[0,1]^2$, $i = 1, ..., N$, and independent.

A graphical description of the Cannings model

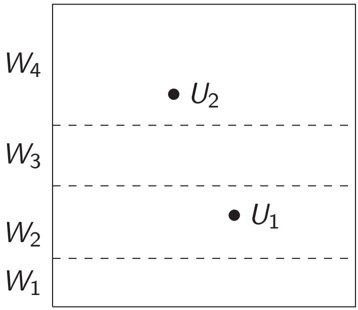
$\mathbb{P}($ Child i chooses j as parent $|\mathcal{W}) = W_j$, for $i = 1, ..., N$ independently.



with $U_i \sim$Unif$[0,1]^2$, $i = 1, ..., N$, and independent.

A graphical description of the Cannings model

$\mathbb{P}($ Child i chooses j as parent $|\mathcal{W}) = W_j$, for $i = 1, ..., N$ independently.



with $U_i \sim$Unif$[0,1]^2$, $i = 1, ..., N$, and independent.

A Wright-Fisher model with selection

- Constant population size N with 2 types: wild type and beneficial type
- s_N selection strength, $0 < s_N < 1$.
- give beneficial individuals weight 1
- give wild type individuals weight $1 - s_N$

Cannings models with selection

- Constant population size N with 2 types wildtype and beneficial type.
- s_N selection strength, $0 < s_N < 1$.
- Let $\{W^{(g)}\}_{g \geq 0} = (W_1^{(g)}, ..., W_N^{(g)})_{g \geq 0}$ be an iid exchangeable mass partition of 1.
- Distinguish a neutral part of mass

$$(1 - s_N)W_i$$

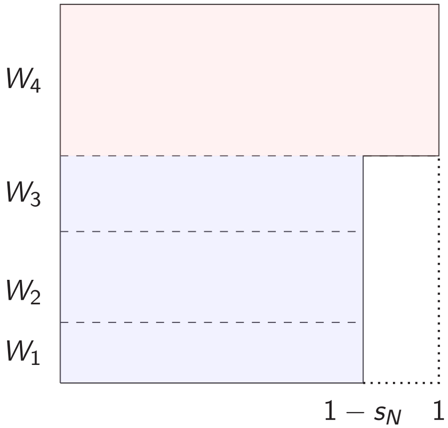
and a selective part of mass

$$s_N W_i$$

for each weight W_i.

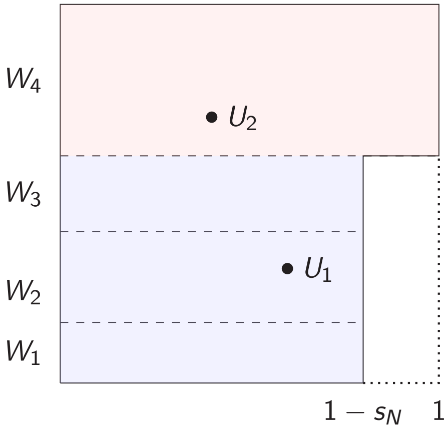
Cannings models with selection: forward construction

Assume there are $k \leq N$ wild type individuals.
$U_1, U_2, ..., U_N$ are uniformly distributed on the coloured part.



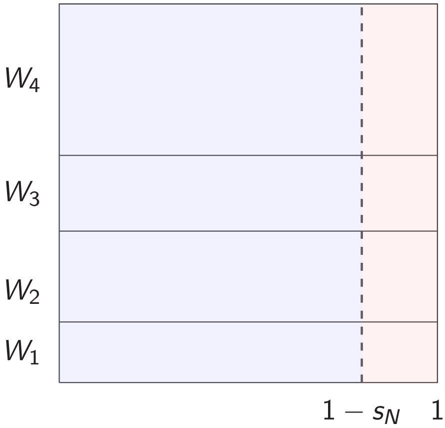
Cannings models with selection: forward construction

Assume there are $k \leq N$ wild type individuals.
$U_1, U_2, ..., U_N$ are uniformly distributed on the coloured part.



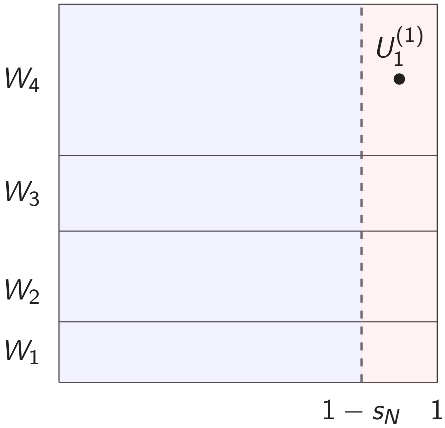
Cannings models with selection: backward construction

- Determine the (exchangeable) weights $W^{(g)} = (W_1^{(g)}, ..., W_N^{(g)})$.
- Don't know types of parents.
- For each individual $1 \leq i \leq N$ throw iid uniforms $(U_i^{(1)}, U_i^{(2)}, ...)$ into $[0,1]^2$ until one falls into the neutral part.
- This leads to a $\text{Geom}(1 - s_N)$ distributed number of potential parents for each individual.



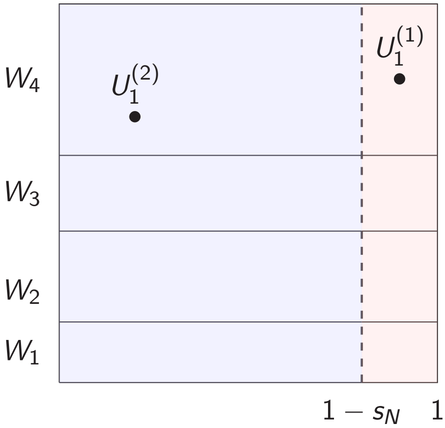
Cannings models with selection: backward construction

- Determine the (exchangeable) weights $W^{(g)} = (W_1^{(g)}, \dots, W_N^{(g)})$.
- Don't know types of parents.
- For each individual $1 \le i \le N$ throw iid uniforms $(U_i^{(1)}, U_i^{(2)}, \dots)$ into $[0,1]^2$ until one falls into the neutral part.
- This leads to a $\mathrm{Geom}(1 - s_N)$ distributed number of potential parents for each individual.



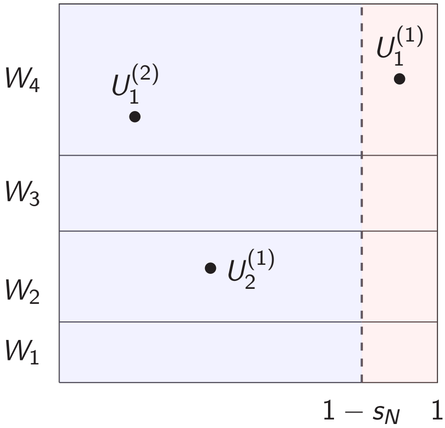
Cannings models with selection: backward construction

- Determine the (exchangeable) weights $W^{(g)} = (W_1^{(g)}, ..., W_N^{(g)})$.
- Don't know types of parents.
- For each individual $1 \leq i \leq N$ throw iid uniforms $(U_i^{(1)}, U_i^{(2)}, ...)$ into $[0, 1]^2$ until one falls into the neutral part.
- This leads to a Geom$(1 - s_N)$ distributed number of potential parents for each individual.



Cannings models with selection: backward construction

- Determine the (exchangeable) weights $W^{(g)} = (W_1^{(g)}, ..., W_N^{(g)})$.
- Don't know types of parents.
- For each individual $1 \leq i \leq N$ throw iid uniforms $(U_i^{(1)}, U_i^{(2)}, ...)$ into $[0,1]^2$ until one falls into the neutral part.
- This leads to a $\mathrm{Geom}(1 - s_N)$ distributed number of potential parents for each individual.



Cannings Ancestral Selection Process

- Proceed further by determining potential ancestors of the potential parents the till generation 0.
- The process $(A_g)_{g\geq 0}$ counts the number of potential parents g generations in the past.
- Colour the individuals in generation 0.
- Propagate types further.

A child has the beneficial type if and only, if
at least one of its potential parents has the beneficial type.

Haldane's formula

- (moderate selection) $s_N \sim N^{-b}$ for $b \in (0,1)$.
  - moderately weak selection $\frac{1}{2} < b < 1$.
  - moderately strong selection $0 < b < \frac{1}{2}$.

Conditions **C**

$$\mathbb{E}\left[W_1^2\right] = \frac{\rho^2}{N^2},\ \rho \geq 1$$

$$\Rightarrow \text{Offspring variance } v_N = \rho^2(1+o(1))$$

$$\mathbb{E}\left[W_1^3\right] = O(N^{-3})$$

- These conditions imply Möhle's criteria for convergence of the genealogies to Kingman's coalescent.

Haldane's formula for $\frac{1}{2} < b < 1$

Let X_g the frequency of wildtype individuals in generation g.
Assume $X_0 = 1 - \frac{1}{N}$.

Theorem

Let $\frac{1}{2} < b < 1$ and the Conditions **C** be fulfilled.
Set

$$T_{fix} = \inf\left\{g \geq 0 : X_g \in \{0,1\}\right\}.$$

Then

$$\pi_N := \mathbb{P}(X_{T_{fix}} = 0) = \frac{2s_N}{\rho^2} + o(s_N).$$

Idea of proof, $b > 1/2$

Sampling Duality

Theorem

Denote by K_g the number of wildtype individuals at time $g \geq 0$ and by A_g the number of potential ancestors g generations back in time. Then

$$\mathbb{E}\left[\frac{K_g(K_g-1)\cdots(K_g-n+1)}{N(N-1)\cdots(N-n+1)}\Big|K_0=k\right]$$
$$=\mathbb{E}\left[\frac{k(k-1)\cdots(k-A_g+1)}{N(N-1)\cdots(N-A_g+1)}\Big|A_0=n\right].$$

Sampling Duality

Theorem

Denote by K_g the number of wildtype individuals at time $g \geq 0$ and by A_g the number of potential ancestors g generations back in time. Then

$$\mathbb{E}\left[\frac{K_g(K_g-1)\cdots(K_g-n+1)}{N(N-1)\cdots(N-n+1)}\big|K_0=k\right]$$
$$=\mathbb{E}\left[\frac{k(k-1)\cdots(k-A_g+1)}{N(N-1)\cdots(N-A_g+1)}\big|A_0=n\right].$$

Apply the sampling duality to $k=N-1, n=N$ yields

$$\mathbb{P}(K_g=N|K_0=N-1)=1-\mathbb{E}\left[\frac{A_g}{N}\right]$$

Apply sampling duality

Taking the limit $g \to \infty$ yields

$$\pi_N = \lim_{g \to \infty} 1 - \mathbb{P}(K_g = N | K_0 = N - 1) = \mathbb{E}\left[\frac{A_{eq}}{N}\right].$$

Apply sampling duality

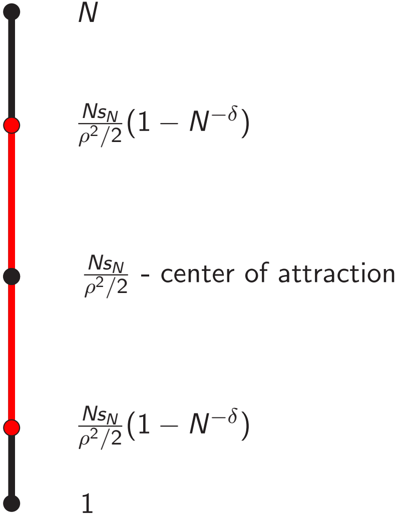
Taking the limit $g \to \infty$ yields

$$\pi_N = \lim_{g\to\infty} 1 - \mathbb{P}(K_g = N | K_0 = N - 1) = \mathbb{E}\left[\frac{A_{eq}}{N}\right].$$

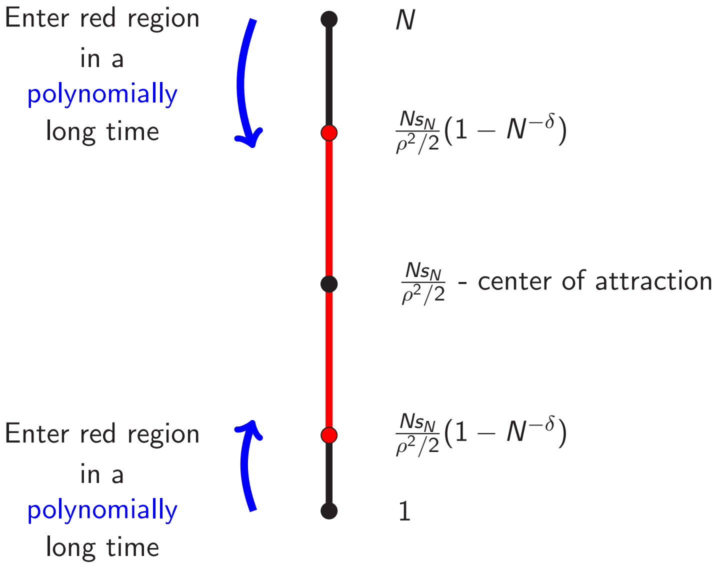
It sufficies to show

$$\mathbb{E}\left[A_{eq}\right] \sim \frac{Ns_N}{\rho^2/2}$$

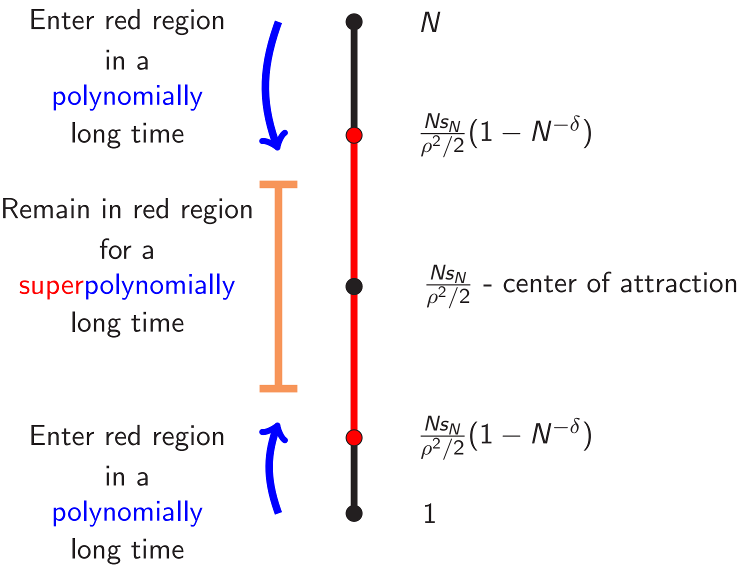
Concentration of $\mathcal{A} = (A_g)_{g \geq 0}$



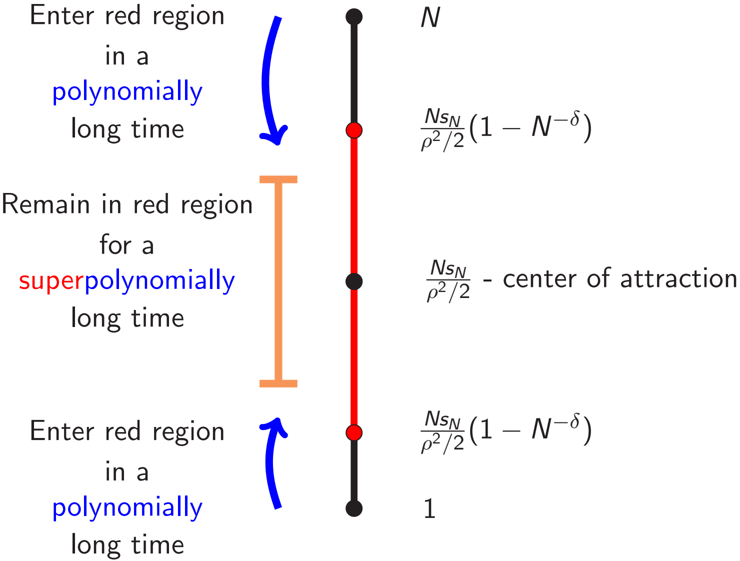
Concentration of $\mathcal{A} = (A_g)_{g \geq 0}$



Concentration of $\mathcal{A} = (A_g)_{g \geq 0}$



Concentration of $\mathcal{A} = (A_g)_{g \geq 0}$



$$\Rightarrow \mathbb{E}[A_{eq}] \sim \frac{Ns_N}{\rho^2/2}$$

Moderately strong selection, $0 < b < \frac{1}{2}$

Dirichlet-type weights

Assume the random weights $\{W_i, 1 \leq i \leq N\}$ are of the form

$$W_i = \frac{Y_i}{\sum_{j=1}^{N} Y_j},$$

with $Y_1, Y_2, \ldots$ iid on $\mathbb{R}^+$

and the following Conditions $\mathbf{C}'$ are fulfilled

- $\mathbb{E}[W_1^2] = \frac{\rho^2}{N^2} + O(N^{-3})$
- For some $h > 0$

$$\mathbb{E}[exp(hY_1)] < \infty.$$

Haldane's formula for $0 < b < \frac{1}{2}$

Let X_g be the frequency of wildtype individuals at time g with $X_0 = 1 - \frac{1}{N}$.

Theorem:

Assume the Conditions **C**' are fulfilled.
Set

$$T_{fix} = \inf\left\{g \geq 0 : X_g \in \{0, 1\}\right\}.$$

Then

$$\pi_N := \mathbb{P}(X_{T_{fix}} = 0) = \frac{2s_N}{\rho^2} + o(s_N).$$

Strategy of proof

(1) Couple the forward process $(X_g)_{g \geq 0}$ to appropriate Galton-Watson processes form above and below to show that the number of beneficial individuals exceeds eventually the level $N^{b+\delta}$ with probability $\frac{2s_N}{\rho^2} + o(s_N)$.

Strategy of proof

(1) Couple the forward process $(X_g)_{g \geq 0}$ to appropriate Galton-Watson processes form above and below to show that the number of beneficial individuals exceeds eventually the level $N^{b+\delta}$ with probability $\frac{2s_N}{\rho^2} + o(s_N)$.

(2) Show that from $N^{b+\delta}$ the level ϵN is reached with probability $1 - o(s_N)$: Whenever $(X_g)_{g \geq 0}$ is between $N^{b+\delta}$ and $2\epsilon N$, the one-step variance of $(X_g)_{g \geq 0}$ is small enough, such that the exponential increase is strong enough to reach the level ϵN before 0 with probability $1 - o(s_N)$.

Strategy of proof

(1) Couple the forward process $(X_g)_{g\geq 0}$ to appropriate Galton-Watson processes form above and below to show that the number of beneficial individuals exceeds eventually the level $N^{b+\delta}$ with probability $\frac{2s_N}{\rho^2} + o(s_N)$.

(2) Show that from $N^{b+\delta}$ the level ϵN is reached with probability $1 - o(s_N)$: Whenever $(X_g)_{g\geq 0}$ is between $N^{b+\delta}$ and $2\epsilon N$, the one-step variance of $(X_g)_{g\geq 0}$ is small enough, such that the exponential increase is strong enough to reach the level ϵN before 0 with probability $1 - o(s_N)$.

(3) The (equilibrium) backward ancestral process meets with ϵN lines in at least one individual with probability $1 - o(s_N)$ (birthday problem).

Dominating by a Galton-Watson process from above

Assume k individuals are beneficial in generation g. Then since the weights are of Dirichlet-type, in generation $g+1$ the number of beneficial individuals is

$$\mathrm{Bin}\left(N, \frac{\sum_{i=1}^{k} Y_i}{\sum_{j=1}^{k} Y_j + \sum_{j=k+1}^{N}(1-s_N)Y_j}\right) \qquad (1)$$

Dominating by a Galton-Watson process from above

Assume k individuals are beneficial in generation g. Then since the weights are of Dirichlet-type, in generation $g+1$ the number of beneficial individuals is

$$\text{Bin}\left(N, \frac{\sum_{i=1}^{k} Y_i}{\sum_{j=1}^{k} Y_j + \sum_{j=k+1}^{N}(1-s_N)Y_j}\right) \qquad (1)$$

For a coupling with Galton-Watson-process need to approximate (1) by a sum of k iid summands.

- **Binom**$(N,p) \leq$ **Pois**(Np):
  $X_1 \sim Ber(p)$, $Y_1 \sim Pois(p)$
  $P(X_1 = 0) = 1 - p \leq e^{-p} = P(Y_1 = 0)$
  Hence, can couple $X \sim Binom(N,p)$ and $Y \sim Pois(Np)$, such that

$$P(X \leq Y) = 1.$$

Dominating by a Galton-Watson process from above

Assume k individuals are beneficial in generation g. Then since the weights are of Dirichlet-type, in generation $g+1$ the number of beneficial individuals is

$$\text{Bin}\left(N, \frac{\sum_{i=1}^{k} Y_i}{\sum_{j=1}^{k} Y_j + \sum_{j=k+1}^{N}(1-s_N)Y_j}\right) \qquad (1)$$

For a coupling with Galton-Watson-process need to approximate (1) by a sum of k iid summands.

- **Binom**$(N,p) \leq$ **Pois**(Np):
  $X_1 \sim Ber(p)$, $Y_1 \sim Pois(p)$
  $P(X_1 = 0) = 1 - p \leq e^{-p} = P(Y_1 = 0)$
  Hence, can couple $X \sim Binom(N,p)$ and $Y \sim Pois(Np)$, such that

$$P(X \leq Y) = 1.$$

- **Poisson-distribution is infinitely divisible**:
  $Pois(\lambda_1 + ... + \lambda_k) = \sum_{i=1}^{k} Pois(\lambda_i)$

Dominating by a Galton-Watson process from above

Assume k individuals are beneficial in generation g. Then since the weights are of Dirichlet-type, in generation $g+1$ the number of beneficial individuals is

$$\text{Bin}\left(N, \frac{\sum_{i=1}^{k} Y_i}{\sum_{j=1}^{k} Y_j + \sum_{j=k+1}^{N}(1-s_N)Y_j}\right) \qquad (1)$$

For a coupling with Galton-Watson-process need to approximate (1) by a sum of k iid summands.

- **Binom**$(N,p) \leq$ **Pois**(Np):
  $X_1 \sim Ber(p), Y_1 \sim Pois(p)$
  $P(X_1 = 0) = 1 - p \leq e^{-p} = P(Y_1 = 0)$
  Hence, can couple $X \sim Binom(N,p)$ and $Y \sim Pois(Np)$, such that
  $$P(X \leq Y) = 1.$$
- **Poisson-distribution is infinitely divisible**:
  $Pois(Np_1 + ... + Np_k) = \sum_{i=1}^{k} Pois(Np_i)$

Dominating by a Galton-Watson process from above

Choose the $p_1, ..., p_k$ for the coupling
As long as $k \leq N^{\frac{1}{2}+\delta}$ we can estimate for sufficiently many generations

$$\frac{\sum_{i=1}^{k} Y_i}{\sum_{j=1}^{k} Y_j + \sum_{j=k+1}^{N}(1-s_N)Y_j} \leq \sum_{i=1}^{k} Y_i \frac{(1+s_N)(1+N^{-\frac{1}{2}+\delta})}{N\mathbb{E}[Y_1]}$$

with probability $1-o(s_N)$.

Gives a coupling from above with a Galton-Watson-process with

$$\frac{(1+s_N)(1+N^{-\frac{1}{2}+\delta})}{N\mathbb{E}[Y_1]} Pois(Y)$$

offspring numbers where $Y \sim Y_1$
(until the level $N^{\frac{1}{2}+\delta}$ is reached).

Dominating by a Galton-Watson process from above

Choose the $p_1, ..., p_k$ for the coupling
As long as $k \leq N^{\frac{1}{2}+\delta}$ we can estimate for sufficiently many generations

$$\frac{\sum_{i=1}^{k} Y_i}{(1-s_N)(\sum_{j=1}^{k} Y_j + \sum_{j=k+1}^{N} Y_j)} \leq \sum_{i=1}^{k} Y_i \frac{(1+s_N)(1+N^{-\frac{1}{2}+\delta})}{N\mathbb{E}[Y_1]}$$

with probability $1 - o(s_N)$.

Gives a coupling from above with a Galton-Watson-process with

$$\frac{(1+s_N)(1+N^{-\frac{1}{2}+\delta})}{N\mathbb{E}[Y_1]} Pois(Y)$$

offspring numbers where $Y \sim Y_1$
(until the level $N^{\frac{1}{2}+\delta}$ is reached).

Dominating with a Galton-Watson-process from below

Let $\xi_i^{(g)}$ be the number of offsprings of the i-th beneficial individual in generation g.

$$\xi_1^{(g)} \sim \mathrm{Bin}\left(N, \frac{Y_1(1 + s_N + o(s_N))}{\sum_{i=1}^N Y_i}\right)$$

Dominating with a Galton-Watson-process from below

Let $\xi_i^{(g)}$ be the number of offsprings of the i-th beneficial individual in generation g.

$$\xi_1^{(g)} \sim \mathrm{Bin}\left(N, \frac{Y_1(1+s_N+o(s_N))}{\sum_{i=1}^{N} Y_i}\right)$$

Given $\sum_{i=1}^{j-1} \xi_i^{(g)} = k$, then

$$\xi_j^{(g)} \sim \mathrm{Bin}\left(N-k, \frac{Y_j(1+s_N+o(s_N))}{\sum_{i=j}^{N} Y_i}\right)$$

Dominating with a Galton-Watson-process from below

Let $\xi_i^{(g)}$ be the number of offsprings of the i-th beneficial individual in generation g.

$$\xi_1^{(g)} \sim \mathrm{Bin}\left(N, \frac{Y_1(1+s_N+o(s_N))}{\sum_{i=1}^N Y_i}\right)$$

Given $\sum_{i=1}^{j-1} \xi_i^{(g)} = k$, then

$$\xi_j^{(g)} \sim \mathrm{Bin}\left(N-k, \frac{Y_j(1+s_N+o(s_N))}{\sum_{i=j}^N Y_i}\right)$$

Let Z_g be the number of beneficial individuals in generation g. As long as $Z_g \leq N^{\frac{1}{2}}$ this can be lower bounded by

$$\mathrm{Bin}\left(N-N^{\frac{1}{2}}, \frac{Y_j(1+s_N+o(s_N))}{\sum_{i=1}^N Y_i}\right)$$

for each beneficial individual j.

Dominating with a Galton-Watson-process from below

Let $\xi_i^{(g)}$ be the number of offsprings of the i-th beneficial individual in generation g.

$$\xi_1^{(g)} \sim \text{Bin}\left(N, \frac{Y_1(1+s_N+o(s_N))}{\sum_{i=1}^N Y_i}\right)$$

Given $\sum_{i=1}^{j-1} \xi_i^{(g)} = k$, then

$$\xi_j^{(g)} \sim \text{Bin}\left(N-k, \frac{Y_j(1+s_N+o(s_N))}{\sum_{i=j}^N Y_i}\right)$$

Let Z_g be the number of beneficial individuals in generation g. As long as $Z_g \leq N^{\frac{1}{2}}$ this can be lower bounded by

$$\text{Bin}\left(N-N^{\frac{1}{2}}, \frac{Y_j(1+s_N+o(s_N))}{\sum_{i=1}^N Y_i}\right)$$

for each beneficial individual j.

Dominating by a Galton-Watson-process from below

Let Z_g be the number of beneficial individuals in generation g and let $\xi_i^{(g)}$ be the number of offsprings of the i-th beneficial individual in generation g.

Given $\sum_{i=1}^{j-1} \xi_i^{(g)} = k$, then

$$\xi_j^{(g)} \sim \text{Bin}\left(N - k, \frac{Y_j(1 + s_N + o(s_N))}{\sum_{i=j}^{N} Y_i}\right)$$

As long as $Z_g \leq N^{\frac{1}{2}}$ this can be lower bounded with sufficiently high probability by

$$\text{Bin}\left(N - N^{\frac{1}{2}}, \frac{Y_j(1 + s_N + o(s_N))}{N\mathbb{E}[Y_1] + O(N^{\frac{1}{2}-\delta})}\right)$$

for each beneficial individual j.

Idea of proof, $0 < b < \frac{1}{2}$, lower bound

Galton-Watson-process with offspring distribution

$$\text{Bin}\left(N - N^{\frac{1}{2}}, \frac{Y_j(1 + s_N + o(s_N))}{N\mathbb{E}\left[Y_1\right] + O(N^{\frac{1}{2}+\delta})}\right)$$

reaches the level $N^{b+\delta}$ with probability $\frac{2s_N}{\rho^2}(1 + o(1))$.

Summary

For nice Cannings models with selection of strength $s_N \sim N^{-b}$ in the Kingman attraction we proved

- Haldane's formula for $0 < b < \frac{1}{2}$
  by a Galton-Watson process approximation
- Haldane's formula for $\frac{1}{2} < b < 1$
  by analysing the Cannings ancestral selection process

Summary

For nice Cannings models with selection of strength $s_N \sim N^{-b}$ in the Kingman attraction we proved

- Haldane's formula for $0 < b < \frac{1}{2}$ by a Galton-Watson process approximation
- Haldane's formula for $\frac{1}{2} < b < 1$ by analysing the Cannings ancestral selection process

F. Boenkost, A. González Casanova, C.Pokalyuk and A. Wakolbinger. Haldane's formula in Cannings models: The case of moderately weak selection.
Preprint arxiv:1907.10049