Condensation in stochastic systems with selection and mutation

Peter Mörters



based on joint work with

Steffen Dereich (Münster)

Setup of the talk

- (1) A branching model with selection and mutation
- (2) A condensation result and some open problems
- (3) A related mean field model
- (4) Shape of the condensation wave
- (5) Universality of wave shapes?

Our model has two parameters

- a mutation probability $\beta \in [0, 1]$,
- a mutant fitness distribution q, which is a probability measure on [0, 1].

Our model has two parameters

- a mutation probability $\beta \in [0, 1]$,
- a mutant fitness distribution q, which is a probability measure on [0, 1].

The model is a branching process in continuous time.

Our model has two parameters

- a mutation probability $\beta \in [0, 1]$,
- a mutant fitness distribution q, which is a probability measure on [0, 1].

The model is a branching process in continuous time.

• The initial particle has a random fitness chosen according to q.

Our model has two parameters

- a mutation probability $\beta \in [0, 1]$,
- a mutant fitness distribution q, which is a probability measure on [0, 1].

The model is a branching process in continuous time.

- The initial particle has a random fitness chosen according to q.
- Particles with fitness f live forever and produce single offspring with rate f.

Our model has two parameters

- a mutation probability $\beta \in [0, 1]$,
- a mutant fitness distribution q, which is a probability measure on [0, 1].

The model is a branching process in continuous time.

- The initial particle has a random fitness chosen according to q.
- Particles with fitness f live forever and produce single offspring with rate f.
- Every particle born either
 - ▶ inherits the fitness of the parent with probability 1β , or
 - mutates with probability β in which case its fitness is drawn from q.

Our model has two parameters

- a mutation probability $\beta \in [0, 1]$,
- a mutant fitness distribution q, which is a probability measure on [0, 1].

The model is a branching process in continuous time.

- The initial particle has a random fitness chosen according to q.
- Particles with fitness f live forever and produce single offspring with rate f.
- Every particle born either
 - \blacktriangleright inherits the fitness of the parent with probability $1-\beta,$ or
 - mutates with probability β in which case its fitness is drawn from q.

This is a stochastic house-of-cards model for a population with a balance of genetic selection and mutation.

This is a multitype Galton-Watson process with uncountable type space [0, 1].

This is a multitype Galton-Watson process with uncountable type space [0, 1]. Why is it hard to analyse?

This is a multitype Galton-Watson process with uncountable type space [0, 1]. Why is it hard to analyse?

Key to the martingale analysis is the eigenfunction corresponding to the principal eigenvalue of the operator $A: C[0,1] \rightarrow C[0,1]$ given by

$$Af(x) = x\big((1-\beta)f(x) + \beta \int f(y)q(dy)\big).$$

We have

This is a multitype Galton-Watson process with uncountable type space [0, 1]. Why is it hard to analyse?

Key to the martingale analysis is the eigenfunction corresponding to the principal eigenvalue of the operator $A: C[0,1] \rightarrow C[0,1]$ given by

$$Af(x) = x((1-\beta)f(x) + \beta \int f(y)q(dy)).$$

We have

 $Af = \lambda f$

This is a multitype Galton-Watson process with uncountable type space [0, 1]. Why is it hard to analyse?

Key to the martingale analysis is the eigenfunction corresponding to the principal eigenvalue of the operator $A: C[0,1] \rightarrow C[0,1]$ given by

$$Af(x) = x((1-\beta)f(x) + \beta \int f(y)q(dy)).$$

We have

$$Af = \lambda f \Leftrightarrow f(x) = rac{eta x}{\lambda - (1 - eta)x} \int f \, dq$$

This is a multitype Galton-Watson process with uncountable type space [0, 1]. Why is it hard to analyse?

Key to the martingale analysis is the eigenfunction corresponding to the principal eigenvalue of the operator $A: C[0,1] \rightarrow C[0,1]$ given by

$$Af(x) = x((1-\beta)f(x) + \beta \int f(y)q(dy)).$$

We have
$$Af = \lambda f \Leftrightarrow f(x) = \frac{\beta x}{\lambda - (1 - \beta)x} \int f \, dq$$

 $\Rightarrow \exists \lambda^* \ge 1 - \beta \text{ with } 1 = \beta \int \frac{x}{\lambda^* - (1 - \beta)x} \, q(dx)$

This is a multitype Galton-Watson process with uncountable type space [0, 1]. Why is it hard to analyse?

Key to the martingale analysis is the eigenfunction corresponding to the principal eigenvalue of the operator $A: C[0,1] \rightarrow C[0,1]$ given by

$$Af(x) = x((1-\beta)f(x) + \beta \int f(y)q(dy)).$$

We have

$$Af = \lambda f \Leftrightarrow f(x) = \frac{\beta x}{\lambda - (1 - \beta)x} \int f \, dq$$

$$\Rightarrow \exists \lambda^* \ge 1 - \beta \text{ with } 1 = \beta \int \frac{x}{\lambda^* - (1 - \beta)x} \, q(dx)$$

$$\Leftrightarrow \beta \int \frac{1}{1 - x} \, q(dx) \ge 1 \qquad (1)$$

This is a multitype Galton-Watson process with uncountable type space [0, 1]. Why is it hard to analyse?

Key to the martingale analysis is the eigenfunction corresponding to the principal eigenvalue of the operator $A: C[0,1] \rightarrow C[0,1]$ given by

$$Af(x) = x((1-\beta)f(x) + \beta \int f(y)q(dy)).$$

We have

$$Af = \lambda f \Leftrightarrow f(x) = \frac{\beta x}{\lambda - (1 - \beta)x} \int f \, dq$$

$$\Rightarrow \exists \lambda^* \ge 1 - \beta \text{ with } 1 = \beta \int \frac{x}{\lambda^* - (1 - \beta)x} \, q(dx)$$

$$\Leftrightarrow \beta \int \frac{1}{1 - x} \, q(dx) \ge 1 \qquad (1)$$

Only under assumption (1) can we perform a martingale analysis.

Let

 $X_t = #\{$ particles alive at time $t\}$

and Ξ_t be the empirical fitness distribution at time t given by

 $\Xi_t(A) = \frac{\#\{\text{particles with fitness in } A \text{ at time } t\}}{\#\text{particles alive at time } t}.$

Let

 $X_t = #\{$ particles alive at time $t\}$

and Ξ_t be the empirical fitness distribution at time t given by

$$\Xi_t(A) = \frac{\#\{\text{particles with fitness in } A \text{ at time } t\}}{\#\text{particles alive at time } t}.$$

Problems:

Let

 $X_t = #\{$ particles alive at time $t\}$

and Ξ_t be the empirical fitness distribution at time t given by

$$\Xi_t(A) = \frac{\#\{\text{particles with fitness in } A \text{ at time } t\}}{\#\text{particles alive at time } t}.$$

Problems:

(1) How fast does X_t grow?

Let

 $X_t = #\{$ particles alive at time $t\}$

and Ξ_t be the empirical fitness distribution at time t given by

$$\Xi_t(A) = \frac{\#\{\text{particles with fitness in } A \text{ at time } t\}}{\#\text{particles alive at time } t}.$$

Problems:

(1) How fast does X_t grow?

- If (1) holds, then X_t grows exponentially with rate λ^* .
- If (1) fails, the exponential rate of growth is 1 − β, but the growth is not strictly exponential and finding the actual speed of growth is a rather difficult open problem.

Let

 $X_t = #\{$ particles alive at time $t\}$

and Ξ_t be the empirical fitness distribution at time t given by

$$\Xi_t(A) = \frac{\#\{\text{particles with fitness in } A \text{ at time } t\}}{\#\text{particles alive at time } t}.$$

Problems:

(1) How fast does X_t grow?

- If (1) holds, then X_t grows exponentially with rate λ^* .
- If (1) fails, the exponential rate of growth is 1 − β, but the growth is not strictly exponential and finding the actual speed of growth is a rather difficult open problem.

(2) Does the empirical fitness distribution Ξ_t converge and what is the limit?

Let

 $X_t = #\{$ particles alive at time $t\}$

and Ξ_t be the empirical fitness distribution at time t given by

$$\Xi_t(A) = \frac{\#\{\text{particles with fitness in } A \text{ at time } t\}}{\#\text{particles alive at time } t}.$$

Problems:

(1) How fast does X_t grow?

- If (1) holds, then X_t grows exponentially with rate λ^* .
- If (1) fails, the exponential rate of growth is 1 − β, but the growth is not strictly exponential and finding the actual speed of growth is a rather difficult open problem.

(2) Does the empirical fitness distribution Ξ_t converge and what is the limit? This problem is solved in our first theorem.

A condensation result

Theorem 1

If (1) holds there exists a unique $\lambda^* \in [1-eta,1]$ such that

$$\beta\int \frac{x}{\lambda^*-(1-\beta)x}\,q(dx)=1,$$

and if (1) fails let $\lambda^* := 1 - \beta$. Then

- the empirical mean fitness $\int_0^1 x \Xi_t(dx)$ converges almost surely to λ^* ,
- and there exists a probability measure p such that, almost surely, the empirical fitness distribution Ξ_t converges weakly to p.

A condensation result

Theorem 1

If (1) holds there exists a unique $\lambda^* \in [1-\beta,1]$ such that

$$\beta \int \frac{x}{\lambda^* - (1 - \beta)x} q(dx) = 1,$$

and if (1) fails let $\lambda^* := 1 - \beta$. Then

- the empirical mean fitness $\int_0^1 x \Xi_t(dx)$ converges almost surely to λ^* ,
- and there exists a probability measure p such that, almost surely, the empirical fitness distribution Ξ_t converges weakly to p.

The limit measure p of the empirical fitness distribution is given

(a) if (1) holds by
$$p(dx) = \frac{\beta \lambda^*}{\lambda^* - (1 - \beta)x} q(dx)$$
.

A condensation result

Theorem 1

If (1) holds there exists a unique $\lambda^* \in [1-eta,1]$ such that

$$\beta\int \frac{x}{\lambda^*-(1-\beta)x}\,q(dx)=1,$$

and if (1) fails let $\lambda^* := 1 - \beta$. Then

- the empirical mean fitness $\int_0^1 x \Xi_t(dx)$ converges almost surely to λ^* ,
- and there exists a probability measure p such that, almost surely, the empirical fitness distribution Ξ_t converges weakly to p.

The limit measure p of the empirical fitness distribution is given

(b) if (1) fails by $p(dx) = \frac{\beta}{1-x} q(dx) + \gamma(\beta)\delta_1(dx)$, where

$$\gamma(\beta):=1-\beta\int\frac{q(dx)}{1-x}>0.$$

The proof can be adapted from methods for preferential attachment networks.

The proof can be adapted from methods for preferential attachment networks. There are three methods available to do this.

• Borgs, Chayes, Daskalakis and Roch (2007) use coupling to an urn scheme, and results of Janson (2004) to study those.

The proof can be adapted from methods for preferential attachment networks. There are three methods available to do this.

- Borgs, Chayes, Daskalakis and Roch (2007) use coupling to an urn scheme, and results of Janson (2004) to study those.
- Bhamidi (2007) approximates by a fitness distribution with a cutoff near the maximal fitness.

The proof can be adapted from methods for preferential attachment networks. There are three methods available to do this.

- Borgs, Chayes, Daskalakis and Roch (2007) use coupling to an urn scheme, and results of Janson (2004) to study those.
- Bhamidi (2007) approximates by a fitness distribution with a cutoff near the maximal fitness.
- Dereich and Ortgiese (2013) use the idea of stochastic approximation as in the classical work of Robbins and Monro (1951). We adapt their approach.

The proof can be adapted from methods for preferential attachment networks.

• Dereich and Ortgiese (2013) use the idea of stochastic approximation as in the classical work of Robbins and Monro (1951). We adapt their approach.

Let

$$X_n = \frac{1}{n} \# \{ \text{individuals with fitness} \approx x \}$$

when the *n*th particle is born. Then

$$X_{n+1} - X_n = \frac{1}{n+1}F(X_n) + R_{n+1} - R_n,$$

where

$$F(X_n) = \beta q(\approx x) + (1 - \beta) \frac{x}{\bar{X}_n} X_n - X_n$$

and \bar{X}_n is the mean fitness in the system, and $R_{n+1} - R_n = X_{n+1} - \mathbb{E}[X_{n+1}|\mathcal{F}_n]$.

The proof can be adapted from methods for preferential attachment networks.

• Dereich and Ortgiese (2013) use the idea of stochastic approximation as in the classical work of Robbins and Monro (1951). We adapt their approach.

Let

$$X_n = \frac{1}{n} \# \{ \text{individuals with fitness} \approx x \}$$

when the *n*th particle is born. Then

$$X_{n+1} - X_n = \frac{1}{n+1}F(X_n) + R_{n+1} - R_n,$$

where

$$F(X_n) = \beta q(\approx x) + (1 - \beta) \frac{x}{\bar{X}_n} X_n - X_n$$

and \bar{X}_n is the mean fitness in the system, and $R_{n+1} - R_n = X_{n+1} - \mathbb{E}[X_{n+1}|\mathcal{F}_n]$. Convergence

$$ar{X}_n o \lambda^*$$
 and $X_n o rac{eta q(pprox x)}{1-(1-eta)rac{x}{\lambda^*}}$

can be established simultaneously by a bootstrapping argument based on careful estimates of the stochastic error R_n .

The condensation wave

If (1) fails and selection beats mutation the branching population experiences a condensation effect and the fitness of a positive proportion of individuals is driven to maximal value.



The condensation wave

If (1) fails and selection beats mutation the branching population experiences a condensation effect and the fitness of a positive proportion of individuals is driven to maximal value.



Problem:

• What is the shape of the wave moving towards the maximal fitness?

The condensation wave

If (1) fails and selection beats mutation the branching population experiences a condensation effect and the fitness of a positive proportion of individuals is driven to maximal value.



Problem:

• What is the shape of the wave moving towards the maximal fitness?

We cannot currently answer this question for our model and instead treat the problem for a much simpler mean-field model in Theorem 2.

Kingman (1974) introduced a model for the balance of selection and mutation, which is a mean-field version of our process. It consists of a sequence of probability measures (p_n) on the unit interval [0,1] describing the distribution of fitness values in the *n*th generation of a population.

• We put $p_0 = q$.

Kingman (1974) introduced a model for the balance of selection and mutation, which is a mean-field version of our process. It consists of a sequence of probability measures (p_n) on the unit interval [0,1] describing the distribution of fitness values in the *n*th generation of a population.

- We put $p_0 = q$.
- If p_n is the fitness distribution in the *n*th generation we denote by

$$w_n=\int x\,p_n(dx)$$

the mean fitness and define

$$p_{n+1}(dx) = (1-\beta) \frac{x p_n(dx)}{w_n} + \beta q(dx).$$

Kingman (1974) introduced a model for the balance of selection and mutation, which is a mean-field version of our process. It consists of a sequence of probability measures (p_n) on the unit interval [0,1] describing the distribution of fitness values in the *n*th generation of a population.

• We put $p_0 = q$.

• If p_n is the fitness distribution in the *n*th generation we denote by

$$w_n = \int x \, p_n(dx)$$

the mean fitness and define

$$p_{n+1}(dx) = (1-\beta) \frac{x p_n(dx)}{w_n} + \beta q(dx).$$

Loosely speaking, a proportion $1 - \beta$ of the genes in the new generation are resampled from the existing population using their fitness as a selective criterion, and the rest have undergone mutation and are therefore sampled from the fitness distribution q.

Kingman (1974) introduced a model for the balance of selection and mutation, which is a mean-field version of our process. It consists of a sequence of probability measures (p_n) on the unit interval [0,1] describing the distribution of fitness values in the *n*th generation of a population.

- We put $p_0 = q$.
- If p_n is the fitness distribution in the *n*th generation we denote by

$$w_n=\int x\,p_n(dx)$$

the mean fitness and define

$$p_{n+1}(dx) = (1-\beta) \frac{x p_n(dx)}{w_n} + \beta q(dx).$$

Kingman showed that in this model $p_n \rightarrow p$ for the same limit distribution p as before, and condensation occurs if and only if (1) fails.

Theorem 2 Dereich and M (2013)

Suppose that the fitness distribution q(dx) = q(1 - x) dx fails (1), so that condensation occurs. Then there are three possibilities for the shape of the condensation wave.

Theorem 2 Dereich and M (2013)

Suppose that the fitness distribution q(dx) = q(1 - x) dx fails (1), so that condensation occurs. Then there are three possibilities for the shape of the condensation wave.

(a) If q is slowly varying at zero, then for x > 0,

$$\lim_{n\uparrow\infty}p_n(1-\frac{x}{n},1)=\gamma(\beta)\int_0^x e^{-y}\,dy,$$

i.e. the condensation wave has the shape of an exponential distribution.



Theorem 2 Dereich and M (2013)

Suppose that the fitness distribution q(dx) = q(1 - x) dx fails (1), so that condensation occurs. Then there are three possibilities for the shape of the condensation wave.

(b) If q is regularly varying at zero with index $\alpha > 0$, then for x > 0,

$$\lim_{n\uparrow\infty}p_n(1-\frac{x}{n},1)=\frac{\gamma(\beta)}{\Gamma(\alpha+1)}\int_0^x y^{\alpha}e^{-y}\,dy,$$

i.e. the condensation wave has the shape of a gamma distribution with shape parameter $1 + \alpha$.



Theorem 2 Dereich and M (2013)

Suppose that the fitness distribution q(dx) = q(1 - x) dx fails (1), so that condensation occurs. Then there are three possibilities for the shape of the condensation wave.

(c) If log q satisfies a mild technical condition and

$$\frac{-1}{(\log q)''(x)x^2}\downarrow 0 \text{ as } x\downarrow 0,$$

then, for sufficiently large *n*, define $y_n \downarrow 0$ and $\sigma_n \downarrow 0$ by $(\log q)'(y_n) = n$ and $\sigma_n^2 = \frac{-1}{(\log q)''(y_n)}$. Then, for a < b,

$$\lim_{n\uparrow\infty}p_n(1-y_n+a\sigma_n,1-y_n+b\sigma_n)=\frac{\gamma(\beta)}{\sqrt{2\pi}}\int_a^b e^{-\frac{y^2}{2}}\,dy,$$

i.e. the condensation wave has the shape of a normal distribution.

Remarks:

• While the shape of the bulk is a modification of q, the shape of the wave is universal, i.e. not depending on the finer details of q.

Remarks:

- While the shape of the bulk is a modification of *q*, the shape of the wave is universal, i.e. not depending on the finer details of *q*.
- The range of fitness distributions where the wave has the shape of a gamma distribution is the 'largest', comprising all q with q(x) ~ c x^α, for α > 0.

Remarks:

- While the shape of the bulk is a modification of *q*, the shape of the wave is universal, i.e. not depending on the finer details of *q*.
- The range of fitness distributions where the wave has the shape of a gamma distribution is the 'largest', comprising all q with q(x) ~ c x^α, for α > 0.
- The normal distribution however seems to be the standard shape for unbounded fitness distributions, as conjectured by Park and Krug (2007).

Remarks:

- While the shape of the bulk is a modification of *q*, the shape of the wave is universal, i.e. not depending on the finer details of *q*.
- The range of fitness distributions where the wave has the shape of a gamma distribution is the 'largest', comprising all q with q(x) ~ c x^α, for α > 0.
- The normal distribution however seems to be the standard shape for unbounded fitness distributions, as conjectured by Park and Krug (2007).

Problem:

• To what extent does the picture extend to other stochastic systems with condensation?

This is the topic of a recently started research project.

Define $W_0 := \frac{1}{\beta}$ and, for $n \ge 1$, $W_n := w_1 \cdots w_n$. Given the family $(W_n)_{n \ge 0}$ the solution can be obtained as

$$p_n(dx) = \sum_{r=0}^n \frac{W_{n-r}}{W_n} (1-\beta)^r \beta x^r q(dx).$$

Define $W_0 := \frac{1}{\beta}$ and, for $n \ge 1$, $W_n := w_1 \cdots w_n$. Given the family $(W_n)_{n \ge 0}$ the solution can be obtained as

$$p_n(dx) = \sum_{r=0}^n \frac{W_{n-r}}{W_n} (1-\beta)^r \beta x^r q(dx).$$

Hence $u_n := W_n (1 - \beta)^{1-n}$ satisfies the renewal equation

$$u_n = \frac{\beta}{1-\beta} \sum_{r=1}^n u_{n-r} \mu_r, \quad \text{for } n \ge 1,$$

where

$$\mu_n=\int x^n q(dx).$$

Define $W_0 := \frac{1}{\beta}$ and, for $n \ge 1$, $W_n := w_1 \cdots w_n$. Given the family $(W_n)_{n \ge 0}$ the solution can be obtained as

$$p_n(dx) = \sum_{r=0}^n \frac{W_{n-r}}{W_n} (1-\beta)^r \beta x^r q(dx).$$

Hence $u_n := W_n (1 - \beta)^{1-n}$ satisfies the renewal equation

$$u_n = rac{eta}{1-eta} \sum_{r=1}^n u_{n-r} \mu_r, \qquad ext{for } n \geq 1,$$

where

$$\mu_n=\int x^n q(dx).$$

In the condensation case we obtain that $u_n \to 0$ and hence contributions to $p_n(dx)$ come from small values of r (bulk) and small values of n - r (wave). The asymptotic behaviour of $p_n(dx)$ near $x \approx 1$ can be obtained from that of μ_n .

For example in case (c) we have, with y := 1 - x,

$$\mu_n = \int (1-y)^n q(y) \, dy \approx \int \exp\left(-ny + \log q(y)\right) \, dy.$$

Hence the main contribution arises when $y \approx y_n$ solving

 $(\log q)'(y_n)=n.$

For example in case (c) we have, with y := 1 - x,

$$\mu_n = \int (1-y)^n q(y) \, dy \approx \int \exp\left(-ny + \log q(y)\right) \, dy.$$

Hence the main contribution arises when $y \approx y_n$ solving

 $(\log q)'(y_n)=n.$

By Taylor approximation

$$\int \exp\left(-ny + \log q(y)\right) dy$$

$$\approx \exp\left(-ny_n + \log q(y_n)\right) \int \exp\left(\frac{1}{2}(\log q)''(y_n)z^2\right) dz,$$

which shows that the contribution comes from an interval of width

$$\sigma_n = \sqrt{\frac{-1}{(\log q)''(y_n)}}$$

and the shape of the wave is normal.

Peter Mörters (Bath)

What else do we know about the shape of condensation waves?

• Dereich (2013) has shown that in a model of a random network with preferential attachment with a regularly varying fitness distribution the degree-weighted fitness distribution has a gamma-shaped condensation wave. The classical Babrabasi-Albert model is not covered by this work.

What else do we know about the shape of condensation waves?

- Dereich (2013) has shown that in a model of a random network with preferential attachment with a regularly varying fitness distribution the degree-weighted fitness distribution has a gamma-shaped condensation wave. The classical Babrabasi-Albert model is not covered by this work.
- Dereich and M (2012) following earlier work of Ercolani and Ueltschi (2011) have shown that in a model of random permutations with diverging cycle weights the empirical distribution of relative cycle lengths has an asymptotically gamma-shaped form.

What else do we know about the shape of condensation waves?

- Dereich (2013) has shown that in a model of a random network with preferential attachment with a regularly varying fitness distribution the degree-weighted fitness distribution has a gamma-shaped condensation wave. The classical Babrabasi-Albert model is not covered by this work.
- Dereich and M (2012) following earlier work of Ercolani and Ueltschi (2011) have shown that in a model of random permutations with diverging cycle weights the empirical distribution of relative cycle lengths has an asymptotically gamma-shaped form.
- It would be interesting to know the shape of the condensation wave in the spatial random permutations of Betz and Ueltschi (2009) and other toy models of Bose-Einstein condensation.

What else do we know about the shape of condensation waves?

- Dereich (2013) has shown that in a model of a random network with preferential attachment with a regularly varying fitness distribution the degree-weighted fitness distribution has a gamma-shaped condensation wave. The classical Babrabasi-Albert model is not covered by this work.
- Dereich and M (2012) following earlier work of Ercolani and Ueltschi (2011) have shown that in a model of random permutations with diverging cycle weights the empirical distribution of relative cycle lengths has an asymptotically gamma-shaped form.
- It would be interesting to know the shape of the condensation wave in the spatial random permutations of Betz and Ueltschi (2009) and other toy models of Bose-Einstein condensation.
- Nothing is known at this point for
 - models with self-organised condensation like the Tonks gas, zero-range model or inclusion models,
 - **spatial models**, for example when migration effects replace mutation.