An evolutionary model that satisfies detailed balance

Jüri Lember*, Chris Watkins†

March 1, 2019

Abstract

We propose a class of evolutionary models that involves an arbitrary exchangeable process as the breeding process and different selection schemes. In those models, a new genome is born according to the breeding process, and then a genome is removed according to the selection scheme that involves fitness. Thus the population size remains constant. The process evolves according to a Markov chain, and, unlike in many other existing models, the stationary distribution – so called mutation-selection equilibrium – can be easily found and studied. The behaviour of the stationary distribution when the population size increases is our main object of interest. Several phase-transition theorems are proved.

Keywords: Markov chain Monte Carlo, Dirichlet distribution, de Finetti theorem, weak convergence of probability measures.

AMS classification: 60J10, 60B10

1 Introduction

We introduce a probability model of evolution that has three purposes: as an abstract model of biological evolution; as a class of efficiently implementable genetic algorithms that are easy to analyse theoretically; and as a link between genetic algorithms and Bayesian non-parametric MCMC methods. The stationary distribution of the model can be expressed in closed form for arbitrary fitness functions: this enables us to investigate the behaviour of the model for different population sizes, mutation rates, and fitness scalings. We find two phase transitions that occur for all fitness functions. Our approach is most applicable to evolution by sexual rather than asexual reproduction.

In any model of evolution under constant conditions, there is a temporal sequence of possibly overlapping populations. In the transition from each population to the next, one

*Institute of Mathematics and Statistics, University of Tartu, Estonia
†Department of Computer Science, Royal Holloway, University of London, UK

or more new individuals are 'born', and the same number of individuals are removed from the population, or 'die'. Each new generation is a population that depends only upon the previous parent population, so that the sequence of populations is a Markov chain. In a model with mutation, the Markov chain is irreducible and has a unique stationary distribution that is also known as the mutation-selection equilibrium. Typically we wish to know what the mutation-selection equilibrium is. Unfortunately the mutation-selection equilibrium is notoriously hard to characterize, even for apparently simple and elegant models of breeding, mutation, and selection. The reason is that in previous models of evolution with sexual reproduction, the Markov chain of populations is irreversible, so that there is no obvious method of finding the stationary distribution other than attempting to compute eigenvectors of the transition matrix directly, as done in [18], but these calculations are neither easy nor revealing for arbitrary fitness functions.

For a reversible Markov chain, the stationary distribution may be found by verifying detailed balance conditions. In our model, introduced in [19], we start by writing the mutation-selection equilibrium in a convenient closed form, and then exhibit MCMC kernels that implement reversible Markov chains with this stationary distribution, and for which the proposal and acceptance algorithms are plausible abstract models of breeding, mutation, and selection.

Each generation starts with a population of n individuals. One new individual is 'bred' – that is, sampled conditionally on the existing population – to produce an expanded population of $n+1$ individuals; from this expanded population, one individual is selected to be discarded, leaving a new population of size n to start the next generation. This might appear similar to the Moran Process [12] but in fact it is very different, as explained in section 3.6

In previous evolutionary models such as [4], and in genetic algorithms such as [11], the reproductive fitness of a genome is modelled as the genome's rate of breeding, and not as its probability of death. In the breeding phase of each generation, fit genomes are chosen to breed more frequently than unfit genomes: in these models, discarding individuals – or 'death' – is modelled as random deletion from the population.

Here we model breeding as conditional sampling in which all existing members of the population are treated equally, regardless of their fitness. We discard individuals according to their fitness: less-fit individuals are more likely to be discarded, so that they remain in the population for a shorter time, and they are therefore sampled less as a result of their shorter lifetime, and so contribute less to the new individuals that are 'bred'. Thus differences in reproductive fitness are modelled as differences in longevity. This is a significant design choice in our models because breeding and selection are modelled separately, and this turns out to greatly simplify the analysis.

As explained in detail in sections 2 and 3, careful choices in modelling breeding as conditional sampling, and in modelling fitness by stochastic rejection of the less fit, enable us to construct a reversible Markov chain of populations that is a form of Metropolis-Hastings process.

Throughout the paper, we suppose there is a finite set $\mathcal{X}$ of possible genomes, and $\xi_1, \xi_2, \ldots$ denotes an exchangeable $\mathcal{X}$-valued stochastic process. For any population of n

genomes $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, we define

$$P(\xi_1 = x_1, \ldots, \xi_n = x_n) =: P_\xi(x_1, \ldots, x_n) =: P_\xi(\mathbf{x}) \tag{1.1}$$

By definition of exchangeability, we have that for any permutation σ,

$$P_\xi(x_1, \ldots, x_n) = P_\xi(x_{\sigma(1)}, \ldots, x_{\sigma(n)}).$$

Given a population of genomes $\mathbf{x} = (x_1, \ldots, x_n)$, we breed an $n+1$'th genome by sampling x_{n+1} conditionally:

$$x_{n+1} \sim P_\xi(\cdot \mid x_1, \ldots, x_n).$$

The 'fitness' of a genome x is denoted by $w(x) > 0$, where w is an arbitrary strictly positive function over $\mathcal{X}$. In the evolutionary models we propose, in each generation one individual is 'bred' by conditional sampling from the existing population, and then an individual is discarded in a fitness-biased way, so that less-fit individuals are more likely to be discarded. The stationary distribution of populations factorises into the form:

$$\underbrace{P_n(x_1, \ldots, x_n)}_{\text{stationary distribution}} = \frac{1}{Z_n} \underbrace{P_\xi(x_1, \ldots, x_n)}_{\text{breeding term}} \underbrace{w(x_1) \cdots w(x_n)}_{\text{fitness term}}. \tag{1.2}$$

The process is similar to, but not the same as non-parametric Bayesian MCMC, and we make this connection explicit in section 3.5.

A genetic algorithm has three basic parameters: population size, mutation rate, and fitness scaling. The most important results of this paper characterise the interacting effect of these parameters as population size $n \to \infty$.

We particularly consider exchangeable processes ξ that are (collections of) Polya urn processes, also known as Dirichlet-categorical processes because these admit a notion of 'mutation'. Conditional sampling from these processes is directly interpretable as a simplified model of sexual breeding with mutation, as explained in section 2. We denote the parameters defining the Dirichlet prior(s) as α. Since fitness $w > 0$, we often write w in terms of a function $\phi : \mathcal{X} \to \mathbb{R}$ such that $w(x) = \exp(-\phi(x))$. We study the interaction of n, α, and ϕ on the stationary distribution by letting population size $n \to \infty$ while the fitness of each genome x scales with n as $\exp(n^{-\lambda}\phi(x))$, and α scales with n as $n^{1-\lambda}\alpha$; in sections 5 and 6 we use the de Finetti representation of ξ to derive limiting forms of the marginal distributions over $\mathcal{X}$ of

$$P_n(x_1, \ldots, x_n \mid n^{1-\lambda}\alpha; n^{-\lambda}\phi) \quad \text{as } n \to \infty.$$

We establish that with this rescaling, for any fitness function ϕ over $\mathcal{X}$, and for any Dirichlet-categorical process ξ, there are three different non-trivial population distributions in the limit as $n \to \infty$:

Constant mutation-rate limit : $\lambda = 0$, so that α scales as $n\alpha$ and ϕ is unscaled;

Low mutation, weak selection limit : $\lambda \in (0, 1]$, so that α scales as $n^{1-\lambda}\alpha$ and ϕ as $n^{-\lambda}\phi$;

Critical low mutation limit : $\lambda = 1$, so that α is unscaled, and ϕ scales as $\frac{\phi}{n}$.

The phase transitions at $\lambda = 0$ and $\lambda = 1$ are sharp. We characterise the limiting distributions in the case that there is a unique maximal element of the posterior distribution. As far as we are aware, these are the first results that give exact explicit expressions for stationary distributions of genetic algorithms with arbitrary fitness functions.

The paper is organised as follows. In section 2 we give examples of exchangeable conditional sampling procedures that can be regarded as abstract models of breeding, and section 3 gives examples of selection procedures that, together with any of the breeding procedures, will produce a reversible Markov chain of populations with the stationary distribution of equation 1.2. Section 3.4 establishes that the stationary distributions are invariant to multiplicative fitness noise; section 3.5 shows that special cases of this process are forms of Bayesian inference by MCMC. This family of evolutionary algorithms might appear similar to the Moran Process – but section 3.6 explains that the Moran process is quite different. Next, section 4 establishes basic conditions on the convergence of the stationary distribution to a limit distribution as population size tends to infinity. Sections 5 and 6 characterise the limiting forms of the stationary distribution for large populations. These are the main results of this paper. We present computational experiments demonstrating our results in section 7. Finally we discuss implications of our results for genetic algorithms and evolutionary modelling in 8.

2 Breeding and mutation

We model breeding as Gibbs sampling [7] from an exchangeable distribution; exchangeable Gibbs sampling is a standard technique in statistical nonparametric MCMC methods, for example [14, 10], but in that context it is not of course regarded as a model of breeding. The property of exchangeability of P_ξ will be used in two ways: first, in section 3 we will use it to establish detailed balance for several selection procedures, which establishes that the stationary distribution of populations is indeed as given in equation 1.2. Second, in section 4 we use the de Finetti integral representation of ξ to establish limit properties of the stationary distribution as $n \to \infty$.

We now give examples of conditional sampling that can be regarded as plausible models of breeding.

Dirichlet-Categorical Process. In the simplest case, each 'genome' consists of only one 'gene' which can be one of K possible alleles, that we denote by $\{1, \ldots, K\}$. The exchangeable process ξ is the well known Polya urn model for a Dirichlet process with discrete base distribution. We recall the definition of this process. Let $\alpha = (\alpha_1, \ldots, \alpha_K)$, $\alpha_i > 0$ be the prior parameters of the base distribution; we write $|\alpha| := \alpha_1 + \cdots + \alpha_K$. Let $\xi_1, \xi_2, \ldots$ be a random process over $\{1, \ldots, K\}$. Let $n \geq 0$. Given $\mathbf{x} = x_1, \ldots, x_n$, we

denote the number of k-s in the sequence by $n_k(\mathbf{x})$:

$$n_k = \sum_{i=1}^{n} I_{\{k\}}(x_i). \tag{2.1}$$

Thus

$$P_\xi(\xi_{n+1} = k \mid x_1, \ldots, x_n, \alpha) := \frac{n_k + \alpha_k}{n + |\alpha|}.$$

It follows that

$$P_\xi(x_1, \ldots, x_n \mid \alpha) = \frac{\alpha_1(\alpha_1+1)\cdots(\alpha_1+n_1-1)\cdots\alpha_K(\alpha_K+1)\cdots(\alpha_K+n_K-1)}{|\alpha|(|\alpha|+1)\cdots(|\alpha|+n-1)}$$

and $\xi_1, \xi_2, \ldots$ is infinitely exchangeable by inspection. By de Finetti's theorem

$$P_\xi(x_1, \ldots, x_n \mid \alpha) = \int_{\mathcal{P}} q(x_1)\cdots q(x_n)\pi(dq),$$

where $\mathcal{P} = \{(q_1, \ldots, q_K) : q_i \geq 0, \sum q_i = 1\}$ is the set of all probability vectors (simplex) and the prior measure π is Dirichlet distribution, i.e. $\pi = \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$. This process is in a sense the central object of our study.

Concentration parameter and mutation rate. The concentration parameter $|\alpha| = \alpha_1 + \cdots + \alpha_K$ may be viewed as determining a mutation rate that depends on n. With n balls in the urn, when a new ball is sampled, there is a probability $\frac{|\alpha|}{n+|\alpha|}$ that the ball will be sampled from the collection of 'prior' balls, rather than the actual balls in the urn. This probability is independent of the colors of n actual balls present: since a new colour may be introduced in this way, we regard this as analogous to a mutation. The mutation rate u is

$$u = \frac{|\alpha|}{|\alpha| + n} \quad \text{or equivalently} \quad |\alpha| = \frac{nu}{1-u}$$

In our evolutionary processes, one new genome is sampled, and one genome is then discarded at each generation, so that n remains constant. To define processes with the same mutation rate and different values of n, α must be adjusted to depend on n.

Note that in this model, mutations only occur at birth, and – importantly – the distribution of mutations does not depend on the frequencies of different colours that are currently in the population; mutations are distributed according to a fixed prior distribution determined by the frequencies of 'magic' balls in the urn. This is strictly less general than an alternative model in which breeding occurs as follows: a ball x is sampled from the urn, and then a mutated ball $x^\dagger \sim P_M(\cdot \mid x)$ is conditionally sampled according to some mutation distribution P_M, and then x and the new mutated ball $x^\dagger$ are both returned to the urn. But this alternative model would not necessarily be reversible, and we do not consider it further.

In our model, we count as a mutation any ball which results from a draw of a 'magic' ball: this definition is consistent with an extension of our model to Dirichlet processes with continuous base distributions, which we intend to consider in future work.

2.1 Complex genomes: direct product of Dirichlet processes

More complex evolutionary models such as genetic algorithms require more complex genomes. Suppose that each genome is a vector of L genes, $x_i = (y_i^1, \ldots, y_i^L)$. Let ξ be a direct product of L independent exchangeable processes $\xi = (\xi^1, \ldots, \xi^L)$. Then

$$P_\xi(x_1, \ldots, x_n) := \prod_{j=1}^{L} P_{\xi^j}(y_1^j, \ldots, y_n^j) \tag{2.2}$$

which clearly is exchangeable as well. Exchangeable sampling from $P_\xi(\cdot \mid x_1, \ldots, x_n)$, where each x_i is a vector of discrete values, can be viewed as a model of sexual reproduction with the assumption of linkage equilibrium. A new vector $(y_{n+1}^1, \ldots, y_{n+1}^L)$ is sampled by, for $1 \le j \le L$, sampling the j-th component $y_{n+1}^j \sim P_{\xi^j}(\cdot \mid y_1^j, \ldots, y_n^j)$ independently from the rest of the components. In words, each new element of the vector x_{n+1} is either a copy of the corresponding element of a randomly chosen member of the existing population, or else a mutation. Instead of a new 'child' genome being constructed by random recombination of two parent genomes, it is instead a random recombination of all n existing genomes in the population, with mutations.

This method of constructing new genomes by 'n-way recombination' is a widely used approach in genetic algorithms, as used by [1, 2, 3] and others, and sexual reproduction with full linkage equilibrium is a standard simplified model of sexual reproduction in population genetics theory [4, 6].

The extension to Cartesian products of Dirichlet processes might appear rather simple because each component of a new genome is sampled independently of the others; however, this extension can lead to models of great complexity because the fitness function w, or equivalently ϕ, can be an arbitrary function on $\mathcal{X}^L$, so that the stationary distribution need not be a product distribution. In genetic language, the fitness function can have arbitrary epistasis.

Note that there are other discrete exchangeable distributions based on Dirichlet distributions that could also be used as P_ξ. A notable example is the discrete fragmentation-coagulation sequence process introduced in [5]; this was intended as statistical model for imputing phasing in genetic analysis, but it could also be used as a breeding distribution for our purposes.

3 Selection

Several MCMC sampling methods give the factorized stationary distribution given by equation (1.2) and at the same time are models of sexual reproduction that are as plausible as those used in evolutionary computation or simplified models in population genetics.

We suppose that each element $x \in \mathcal{X}$ has a strictly positive weight $w(x)$. In context, we will denote the weights $w(x_1), \ldots, w(x_n)$ as $w_1, \ldots, w_n$.

3.1 Single tournament selection

We suppose that when a new genome x_{n+1} is 'born' and added to the population, it competes to survive by having a tournament with another randomly selected member of the population, x_i say. The probability that x_{n+1} wins the tournament and ejects x_i from the population is $\frac{w(x_{n+1})}{w(x_{n+1})+w(x_i)}$. This probability is always well defined since w is strictly positive. An equally valid tournament winning probability is that x_{n+1} wins with probability $\min\{1, \frac{w(x_{n+1})}{w(x_i)}\}$. These two tournament winning probabilities are simply different formulations of the Metropolis-Hastings acceptance rule. The proof below establishes detailed balance for the first winning rule. The algorithm for performing one generation of breeding, mutation, and selection is:

1. Sample $x_{n+1} \sim P_\xi(\cdot \mid x_1, \ldots, x_n)$
2. Sample i randomly from $\{1, \ldots, n\}$
3. With probability $\frac{w_{n+1}}{w_i+w_{n+1}}$ replace x_i with x_{n+1} and discard x_i, otherwise discard x_{n+1}.

Let $\mathbf{x}$, $\mathbf{x}'$ be populations defined as

$$\mathbf{x} = x_1, \ldots, x_n \quad \mathbf{x}' = x_1, \ldots, x_{i-1}, x_{n+1}, x_{i+1}, \ldots, x_n$$

Recall the measure $P_n(\mathbf{x})$ defined in (1.2). We now show that this measure satisfies detailed balance. By exchangeability of ξ, we have:

$$P_\xi(\mathbf{x})P_\xi(x_{n+1} \mid \mathbf{x}) = P_\xi(x_1, \ldots, x_{n+1}) = P_\xi(\mathbf{x}')P_\xi(x_i \mid \mathbf{x}') \tag{3.1}$$

Note that x_{n+1} has tournament against x_i with probability $\frac{1}{n}$ and wins with probability $\frac{w_{n+1}}{w_i+w_{n+1}}$, so:

$$\begin{aligned}
P_n(\mathbf{x})P(\mathbf{x} \to \mathbf{x}') &= P_n(\mathbf{x}) \cdot P_\xi(x_{n+1} \mid \mathbf{x}) \cdot \frac{1}{n}\frac{w_{n+1}}{w_i + w_{n+1}} \\
&= \frac{1}{Z_n}P_\xi(\mathbf{x})w_1 \cdots w_n \cdot P_\xi(x_{n+1} \mid \mathbf{x}) \cdot \frac{1}{n}\frac{w_{n+1}}{w_i + w_{n+1}} \\
&= \frac{1}{Z_n}P_\xi(x_1, \ldots, x_{n+1}) \cdot \frac{1}{n}\frac{w_1 \cdots w_{n+1}}{w_i + w_{n+1}} \\
&= \frac{1}{Z_n}P_\xi(\mathbf{x}')P_\xi(x_i \mid \mathbf{x}') \cdot \frac{1}{n}\frac{w_1 \cdots w_{n+1}}{w_i + w_{n+1}} \\
&= \frac{1}{Z_n}P_\xi(\mathbf{x}')w_1 \cdots w_{i-1}w_{n+1}w_{i+1} \cdots w_n \cdot P_\xi(x_i \mid \mathbf{x}') \cdot \frac{1}{n}\frac{w_i}{w_i + w_{n+1}} \\
&= P_n(\mathbf{x}')P(\mathbf{x}' \to \mathbf{x}).
\end{aligned}$$

3.2 Inverse fitness selection: limit of many tournaments

Suppose that many tournaments are fought, and each time the loser of the previous tournament fights another randomly chosen genome from the population. After many tournaments, and at a stopping time, the current loser is ejected. The current loser evolves according to a irreducible aperiodic Markov chain, and the limiting distribution of ejection is the stationary distribution of that chain:

$$P(\text{ eject } i\ ) = \frac{\frac{1}{w_i}}{\frac{1}{w_1} + \cdots + \frac{1}{w_{n+1}}}. \tag{3.2}$$

The algorithm for performing one generation of breeding and selection is then:

1. Sample $x_{n+1} \sim P_\xi(\cdot \mid x_1, \ldots, x_n)$
2. Sample i from a discrete p.d. over $\{1, \ldots, n+1\}$ with probabilities proportional to $\{\frac{1}{w_1}, \ldots, \frac{1}{w_{n+1}}\}$
3. Discard x_i

This process too satisfies detailed balance. With $\mathbf{x}$ and $\mathbf{x}'$ defined as above, note that:

$$\begin{aligned}
P_n(\mathbf{x})P(\mathbf{x} \to \mathbf{x}') &= P_n(\mathbf{x}) \cdot P_\xi(x_{n+1} \mid \mathbf{x}) \cdot \frac{\frac{1}{w_i}}{\frac{1}{w_1} + \cdots + \frac{1}{w_{n+1}}} \\
&= \frac{1}{Z_n} P_\xi(\mathbf{x}) w_1 \cdots w_n \cdot P_\xi(x_{n+1} \mid \mathbf{x}) \cdot \frac{\frac{1}{w_i}}{\frac{1}{w_1} + \cdots + \frac{1}{w_{n+1}}} \\
&= \frac{1}{Z_n} P_\xi(x_1, \ldots, x_{n+1}) \frac{w_1 \cdots w_{n+1}}{w_{n+1}} \frac{\frac{1}{w_i}}{\frac{1}{w_1} + \cdots + \frac{1}{w_{n+1}}} \\
&= \frac{1}{Z_n} P_\xi(x_1, \ldots, x_{n+1}) \frac{w_1 \cdots w_{n+1}}{w_i} \frac{\frac{1}{w_{n+1}}}{\frac{1}{w_1} + \cdots + \frac{1}{w_{n+1}}} \\
&= P_n(\mathbf{x}') \cdot P_\xi(x_i \mid \mathbf{x}') \cdot \frac{\frac{1}{w_{n+1}}}{\frac{1}{w_1} + \cdots + \frac{1}{w_{n+1}}} \\
&= P_n(\mathbf{x}')P(\mathbf{x}' \to \mathbf{x}).
\end{aligned}$$

Note in passing that

$$\mathbf{E}\{\text{weight of rejected genome} \mid w_1, \ldots, w_{n+1}\} = \frac{1}{\frac{1}{w_1} + \cdots + \frac{1}{w_{n+1}}}. \tag{3.3}$$

That is, the expected weight of the rejected genome is the harmonic mean of the weights of the genomes in the current population, including the newly added genome x_{n+1}.

The case $n = 1$ reduces to Metropolis-Hastings: with a population of size 1, we have proposal distribution $P_\xi(x' \mid x)P_\xi(x) = P_\xi(x, x') = P_\xi(x \mid x')P_\xi(x')$ and acceptance probability of x' given x of $\frac{w(x')}{w(x)+w(x')}$, which gives a stationary distribution

$$P_1(x) = \frac{P_\xi(x)w(x)}{\sum_x P_\xi(x)w(x)}.$$

3.3 Exchangeable breeding of many offspring

Another MCMC process which is also interpretable as an evolutionary algorithm breeds some arbitrary number m of offspring to give a population of size $n + m$, and then from these selects n genomes to form the next generation. The algorithm for a single generation is as follows:

1. Pick a random number m of offspring to breed, and a random number t of tournaments to conduct. Both m and t should be independent of the current population $x_1, \ldots, x_n$
2. Breed $x_{n+1}, \ldots, x_{n+m}$ by sequential exchangeable sampling; that is, let $x_{n+i} \sim P_\xi(\cdot \mid x_1, \ldots, x_{n+i-1})$ for $1 \leq i \leq m$
3. Assign 'survival tickets' to each of $x_1, \ldots, x_n$; the newly bred genomes $x_{n+1}, \ldots, x_{n+m}$ have as yet no survival tickets.
4. Repeat t times:
   (a) Uniformly sample from the population a genome x_i which currently has a survival ticket, and a genome x_j which currently does not have a survival ticket.
   (b) Hold a tournament between x_i and x_j; x_i wins with probability $\frac{w(x_i)}{w(x_i)+w(x_j)}$.
   (c) The winner of the tournament gets the survival ticket; after the tournament, the winner has the ticket and the loser does not.
5. After t tournaments have taken place, the n genomes currently holding survival tickets are selected to be the new population $x'_1, \ldots, x'_n$; the genomes that are not holding tickets are discarded.

The arguments of section 3.1 can readily be extended to show that this algorithm has the same stationary distribution (1.2).

3.4 Invariance to fitness noise

In real life, survival depends not just on fitness but also on luck. Suppose that when each new 'genome' is 'born', it combines its own intrinsic fitness with an independent random amount of luck. It keeps this amount of luck, unchanged, throughout its life. Does individual luck at birth alter the stationary distribution?

To formalise the question, let $\psi_1, \psi_2, \ldots, \psi_n$ be discrete random variables that represent 'luck'. We consider the ψ_i as discrete random variables because the formal derivations become far simpler. The random variable ψ_i can be interpreted as the individual luck that multiplies the fitness of i-th individual. The log fitness function of i-th individual is $\phi(x) + \psi_i$.

$$P_n(\mathbf{x}|\psi_1, \ldots, \psi_n) = \frac{1}{Z'_n} P_\xi(\mathbf{x}) \exp\Big[-\sum_{i=1}^{n} \big(\phi(x_i) + \psi_i\big)\Big]$$

The joint probability

$$\begin{aligned} P_n(\mathbf{x}; \psi_1, \ldots, \psi_n) &= \frac{1}{Z'_n} P_\xi(\mathbf{x}) \exp\Big[-\sum_{i=1}^{n} \big(\phi(x_i) + \psi_i\big)\Big] P(\psi_1, \ldots, \psi_n) \\ &= \frac{1}{Z'_n} P_\xi(\mathbf{x}) \exp\Big[-\sum_{i=1}^{n} \phi(x_i)\Big] \exp\Big[-\sum_{i} \psi_i\Big] P(\psi_1, \ldots, \psi_n). \end{aligned}$$

We see that the sum factorizes, and so $Z'_n = Z_n \cdot Z_n^\psi$, where

$$Z_n = \sum_{\mathbf{x}} P_\xi(\mathbf{x}) \exp\Big[-\sum_{i=1}^{n} \phi(x_i)\Big], \quad Z_n^\psi = \sum_{\psi_1, \ldots, \psi_n} \exp\Big[-\sum_{i} \psi_i\Big] \prod_{i} P(\psi_1, \ldots, \psi_n)$$

Thus the joint measure factorizes

$$P_n(\mathbf{x}; \psi_1, \ldots, \psi_n) = P_n(\mathbf{x}) P_n^\psi(\psi_1, \ldots, \psi_n),$$

where

$$P_n^\psi(\psi_1, \ldots, \psi_n) = \frac{1}{Z_n^\psi} P(\psi_1, \ldots, \psi_n) \exp[-\sum_{i} \psi_i]$$

is a probability measure. Thus, after summing $\psi_1, \ldots, \psi_n$ out, we end with $P_n(\mathbf{x})$:

$$\sum_{\psi_1, \ldots, \psi_n} P_n(\mathbf{x}; \psi_1, \ldots, \psi_n) = P_n(\mathbf{x}).$$

It follows that the stationary distribution is unchanged by multiplicative noise that is independent of the genomes, for all population sizes. Also, considering $\psi_1, \ldots, \psi_n$ as the prior, we see that posterior measure is $P_n(\psi_1, \ldots, \psi_n)$ that is independent of $\mathbf{x}$.

3.5 Fitness as likelihood: a connection with non-parametric Bayesian MCMC

In Dirichlet Process mixture models, as described for example in [14, 16], items of data $d_1, \ldots, d_n$ are given, together with a likelihood function $l(x, d) = P(d \mid x)$, where $x \in \mathcal{X}$ is a discrete latent variable. The prior distribution over latent variables is given by the exchangeable process ξ, each item of data is associated with its corresponding latent variable, and the aim of MCMC fitting is to sample latent variables $x_1, \ldots, x_n$ from the distribution

$$P_n(x_1, \ldots, x_n) = \frac{1}{Z_n} P_\xi(x_1, \ldots, x_n) P(d_1 \mid x_1) \cdots P(d_n \mid x_n) \tag{3.4}$$

where Z_n is an appropriate normalizing constant. If we write

$$w_j(x_i) := P(d_j \mid x_i)$$

then this distribution (3.4) is produced by the MCMC algorithm of section 3.1 with the slight modification that the tournament between the new 'genome' x_{n+1} and the randomly selected x_j is a victory for x_{n+1} with probability $\frac{w_j(x_{n+1})}{w_j(x_j)+w_j(x_{n+1})}$.

One might construct an evolutionary "Just-So" story as follows. On a rock in the ocean, there are n niches, in each of which one member of the species ξ can live; the fitness of x living in niche j is $w_j(x)$. Evolution occurs when a new individual, bred by 'n-way recombination' of all n parents, then challenges the occupant of a randomly chosen niche by fighting a tournament, after which the victor survives and takes over the niche. This gives a (fanciful) evolutionary interpretation to Bayesian latent variable models with exchangeable priors.

3.6 Differences from previous models

The well known Moran Process was introduced by [12], but since then the term has broadened to include many other related genetic models. Our model has significant differences from the Moran process and its subsequent variants, which we now describe.

First, our model implicitly includes mutation; the Moran model did not and the mutation had to be incorporated independently. To show the difference more clearly, consider the Polya urn scheme with an $\alpha_i > 0$ "prior" of initial balls of color i. Unlike in the Moran model, in our model the urn is not the same as the population. A ball is randomly chosen and returned to the urn with another ball of the same color. This procedure is repeated n times, and after that there are $n + |\alpha|$ balls in the urn, but n elements in the population. Now, according to our breeding rule, a random ball is chosen again and returned with another ball of the same color. Observe that the prior balls are involved in the breeding process as much as the rest of the balls. Whenever a prior ball is chosen during the breeding process, we call it a mutation. For example, if all n population balls are currently white, and a black prior ball is chosen, then a black ball enters the population. Or it might be that a color (type) that has never observed before (inside the n balls taken out)

suddenly appears. Or a mutation may produce a colour that is already common in the population. The main difference from the Moran model is that the prior balls do not take part in the selection step. Thus, before tournament(s) start(s), $|\alpha|$ balls (α_i balls from i-th color) are removed from the urn, and so they will never be discarded. There are now $n+1$ balls in the urn and in the population, one of them will be ejected according to our selection rule. Now the breeding starts, but before that the previously removed $|\alpha|$ prior balls are put back to the urn. In this way, the Markov chain does not have absorbing states, and no fixation occurs.

Second, in the Moran model and in variants we know of, differences in fitness are differences in the probability of being selected to breed, whereas in our model fitness is related to the probability of being selected for discard. This decision avoids complicating the model of breeding by including arbitrary fitness, which simplifies the analysis.

Third, our model applies to any exchangeable distributions over finite sets of possible genomes. The most important instantiation we give here is the product of Dirichlet processes with finite support, but other complex finitely supported exchangeable distributions are possible. This means that our model can be applied to non-trivial genetic algorithms. The Moran model, in contrast, concerns the fixation probabilities of individual alleles.

Finally, the Moran model required one individual to be born and one to die in each generation, whereas our model also to applies to exchangeable breeding of arbitrary numbers of offspring, as described in section 3.3 above.

Our model is a population generalisation of Metropolis Hastings, and is more closely related to nonparametric Bayes inference with distributions based on Dirichlet processes, than it is to the Moran model.

Forms of reversible genetic algorithm, or 'evolutionary MCMC' that satisfy detailed balance were previously suggested by [15], [17], and others, but in these models the genetic operators are not biologically relevant and although they are described as 'evolutionary', they have no relevance to modelling biological evolution; rather, they are proposal heuristics for Metropolis Hastings. These methods are normally applied to continuous domains, but if they were applied to a discrete vector space, the stationary distribution in our notation would be:

$$P_n(x_1, \ldots, x_n) = \frac{1}{Z_n} w(x_1) \cdots w(x_n)$$

which differs from our stationary distribution in equation 1.2 in that the breeding term is absent or equivalently that P_ξ is the uniform distribution over $\mathcal{X}$. In our examples, and in genetic models, P_ξ is very far from the uniform distribution.

3.7 Designing a reversible evolutionary model

Our larger aim is to develop a model of evolution that is sufficiently realistic to capture some of the computational power of natural evolution, but which is also simple and tractable for analysis. To ensure that our model satisfies detailed balance and has a stationary distribution that factorises into a breeding and a selection term, we have made

the following simplifying assumptions in addition to the usual simplifications of population genetics or genetic algorithms:

Overlapping populations : We believe that overlapping populations are necessary for reversibility. If full replacement of the population is enforced at each generation, there can be do guarantee that the population at time t could be easily bred from the population at time $t+1$. In MCMC, state changes typically occur through proposing changes that may or may not be accepted; in an overlapping generations model, if a proposed change is not accepted, we continue with the same population as before.

n-way recombination : in the product of Dirichlet processes breeding system, each new genome is bred from n parents rather than from two parents selected from the population. We conjecture that this is necessary for exact reversibility because, with long genomes and many mutations, if a child is bred from two parents, then the child will be more similar to each of its parents than to other individuals in the population, so that 'triples' of two parents and one child will be identifiable even in the stationary distribution. This breaks reversibility since the direction of time can be determined observing evolution in the stationary distribution.

Mutation as sampling : We consider mutation as sampling from a base distribution of possible alleles. This model of mutation is not as general as those found in biology, where mutation probabilities are not symmetric or reversible.

Fitness as lifetime : All members of the population 'breed' at the same rate, and differences in fitness affect only the expected lifetime of an individual. This clearly differs from many types of natural selection, but it is also well known that many organisms continue to produce offspring throughout their lives, so that their total reproductive success depends on their lifetime, as well as on other factors.

It is beyond the scope of this article to argue further whether our model successfully abstracts some essential computational aspects of evolution with sexual reproduction: we present it merely as an abstraction of sexual evolution which is significantly more tractable to analyse than other apparently simple models.

4 The measure P_n

The measure P_n is our main object of interest. In this section we show that the marginal distributions P_n over the set of genotypes converge as the population size $n \to \infty$; in the next section we characterize these limits.

Since ξ is exchangeable, by de Finetti's theorem there exists a prior measure π on the set of all probability measures on $\mathcal{X}$ (simplex) $\mathcal{P}$ such that for every $\mathbf{x} \in \mathcal{X}^n$

$$P_\xi(\mathbf{x}) = \int_{\mathcal{P}} \prod_{i=1}^{n} q(x_i)\pi(dq) \tag{4.1}$$

so that we can write P_n as follows

$$P_n(\mathbf{x}) = \frac{1}{Z_n}\prod_{i=1}^n w(x_i)\int_{\mathcal{P}}\prod_{i=1}^n q(x_i)\pi(dq) = \frac{1}{Z_n}\int_{\mathcal{P}}\prod_{i=1}^n \big(q(x_i)w(x_i)\big)\pi(dq), \tag{4.2}$$

where Z_n is the normalizing constant. In order to analyze the measure, it is convenient to rewrite it as follows. First, let us introduce some notation

$$\langle q, w\rangle := \sum_{k=1}^K q(k)w(k), \quad r_q(k) := \frac{w(k)q(k)}{\langle q, w\rangle}, \quad k = 1, \dots, K.$$

Note that since $w(k) > 0$ for all k, r_q is correctly defined for every $q \in \mathcal{P}$. Thus $\langle q, w\rangle$ is the expected weight (under q-measure) and r_q is a probability measure on $\mathcal{P}$. Now

$$P_n(\mathbf{x}) = \frac{1}{Z_n}\int_{\mathcal{P}}\prod_{i=1}^n \big(q(x_i)w(x_i)\big)\pi(dq) = \frac{1}{Z_n}\int_{\mathcal{P}}\langle q, w\rangle^n\prod_{i=1}^n r_q(x_i)\pi(dq). \tag{4.3}$$

Since r_q is a probability measure, it holds that $\sum_{\mathbf{x}}\prod_{i=1}^n r_q(x_i) = 1$ and so the normalization constant for P_n is

$$Z_n = \sum_{\mathbf{x}\in\mathcal{X}^n}\int_{\mathcal{P}}\langle q, w\rangle^n\prod_{i=1}^n r_q(x_i)\pi(dq) = \int_{\mathcal{P}}\langle q, w\rangle^n\pi(dq).$$

Finally note that (4.3) can be rewritten more neatly by defining the measure

$$d\bar{\pi}_n := \frac{\langle q, w\rangle^n}{Z_n}d\pi. \tag{4.4}$$

With $\bar{\pi}_n$, we have

$$P_n(\mathbf{x}) = \int_{\mathcal{P}}\prod_{i=1}^n r_q(x_i)\bar{\pi}_n(dq) \quad \forall \quad \mathbf{x}\in\mathcal{X}^n. \tag{4.5}$$

From(4.5), it is easy to find all marginal distributions, namely for any $m = 1, \dots, n$

$$P_n(x_1, \dots, x_m) = \int_{\mathcal{P}}\prod_{i=1}^m r_q(x_i)\bar{\pi}_n(dq) = \frac{1}{Z_n}\int_{\mathcal{P}}\langle q, w\rangle^{n-m}\prod_{i=1}^m q(x_i)w(x_i)\pi(dq). \tag{4.6}$$

In particular, when $(X_1, \dots, X_n) \sim P_n$, then

$$P(X_i = k) = \int_{\mathcal{P}} r_q(k)\bar{\pi}_n(dq) = \frac{1}{Z_n}\int\langle q, w\rangle^{n-1}w(k)q(k)\pi(dq)$$
$$P(X_i = k, X_j = l) = \int_{\mathcal{P}} r_q(k)r_q(l)\bar{\pi}_n(dq) = \frac{1}{Z_n}\int\langle q, w\rangle^{n-2}w(k)q(k)w(l)q(l)\pi(dq)$$

and so on. It is important to observe that $P_n(x_1, \dots, x_m)$ depends on n.

The limit process. We have defined for every n the measure (4.5) that describes the genotype distribution of a n-element population. Now the natural question is: do these measures converge (in some sense) if the population size n grows? First we have to define the sense of convergence. Since every measure P_n is defined on different domain ($\mathcal{X}^n$), we cannot speak about standard (weak) convergence of measures. Instead, we ask about the existence of a limiting stochastic process. To explain the sense of convergence, consider that we have defined a triangular array of random variables:

$$\begin{aligned}
&X_{1,1} \sim P_1\\
&(X_{2,1}, X_{2,2}) \sim P_2\\
&(X_{3,1}, X_{3,2}, X_{3,3}) \sim P_3\\
&\dots\\
&(X_{n,1}, X_{n,2}, \dots, X_{n,n}) \sim P_n\\
&\dots
\end{aligned}$$

We also know that the joint distribution of the first m variables in every row depends on n. Therefore we ask: is there a stochastic process $X_1, X_2, \dots$ so that for every m the following convergence holds

$$(X_{1,n}, \dots X_{m,n}) \Rightarrow (X_1, \dots X_m)? \tag{4.7}$$

According to Kolmogorov's existence theorem, the existence of a stochastic process is equivalent to the existence of (finite dimensional) measures P^*_m on set $\mathcal{X}^m$, $m = 1, 2, \dots$ that satisfy the following consistency conditions: for every m and for every $(x_1, \dots, x_m) \in \mathcal{X}^m$, it holds that

$$\sum_{x_{m+1}} P^*_{m+1}(x_1, \dots, x_m, x_{m+1}) = P^*_m(x_1, \dots, x_m).$$

If we also want (4.7) to be true, then for every m and for every $(x_1, \dots, x_m) \in \mathcal{X}^m$ the following convergences must hold:

$$P_n(x_1, \dots, x_m) \to P^*(x_1, \dots, x_m), \quad \forall m, \quad \forall (x_1, \dots, x_m) \in \mathcal{X}^m. \tag{4.8}$$

We now present a general lemma that guarantees the convergence (4.8). To achieve the full generality, we let w also depend on n. Thus, we have weights w_n, and we define the measures $r_{q,n}$ as follows

$$r_{q,n}(k) := \frac{w_n(k)q(k)}{\langle q, w_n \rangle} \quad \forall k \in \mathcal{X}, .$$

We start with the following observation, proven in appendix.

Claim 4.1 *If $w_n(i) \to w(i)$ $\forall i \in \mathcal{X}$, and $r_{q,n}$ and r_q are defined with respect of w_n and w, respectively, then the following uniform convergence holds.*

$$\sup_{q \in \mathcal{P}} |r_{q,n}(k) - r_q(k)| \to 0. \tag{4.9}$$

In the following lemma, $\bar{\pi}_n$ is an arbitrary probability measures on $\mathcal{P}$, not necessarily as in (4.4). The measure $\bar{\pi}_n$ define P_n as in (4.5).

Lemma 4.1 *Let $w_n(k) \to w(k)$ for every $k \in \mathcal{X}$. If there exists an probability measure $\bar{\pi}$ such that $\bar{\pi}_n \Rightarrow \bar{\pi}$, then for every m there exists a probability measure P_m^* on $\mathcal{X}^m$ so that (4.8) holds. Moreover, for every $(x_1, \ldots, x_m) \in \mathcal{X}^m$*

$$P_m^*(x_1, \ldots, x_m) = \int_{\mathcal{P}} \prod_{i=1}^{m} r_q(x_i) \bar{\pi}(dq), \quad \text{where} \quad r_q(k) = \frac{w(k)q(k)}{\langle w, q \rangle}, \quad k = 1, \ldots, K$$

and the measures P_m^*, $m = 1, 2, \ldots$ satisfy consistency conditions.

Proof. For every $x_1, \ldots, x_m$ from (4.9), it follows that

$$\sup_{q \in \mathcal{P}} |\prod_{i=1}^{m} r_{q,n}(x_i) - \prod_{i=1}^{m} r_q(x_i)| \to 0. \tag{4.10}$$

Since the functions

$$q \mapsto \prod_{i=1}^{m} r_{q,n}(x_i), \quad q \mapsto \prod_{i=1}^{m} r_q(x_i)$$

are bounded (by 1) continuous functions, from uniform convergence, it holds

$$P_n(x_1, \ldots, x_m) = \int_{\mathcal{P}} \prod_{i=1}^{m} r_{q,n}(x_i) \bar{\pi}_n(dq) \to \int_{\mathcal{P}} \prod_{i=1}^{m} r_q(x_i) \bar{\pi}(dq) = P_m^*(x_1, \ldots, x_m).$$

Clearly P_m^* are probability measures. The consistency condition trivially holds, because

$$\sum_{m+1} P_{m+1}^*(x_1, \ldots, x_{m+1}) = \int_{\mathcal{P}} \sum_{x_{m+1}} \prod_{i=1}^{m+1} r_q(x_i) \bar{\pi}(dq) = \int_{\mathcal{P}} \prod_{i=1}^{m} r_q(x_i) \bar{\pi}(dq) = P_m^*(x_1, \ldots, x_m).$$

■

4.1 Frequencies: the measure Q_n

We now consider how to express the limit measure P^* in terms of a limiting measure on the simplex $\mathcal{P}$. Recall $n_k(\mathbf{x})$ defined in (2.1) and let

$$\mathbf{n}(\mathbf{x}) := (n_1, \ldots, n_K), \quad \text{where } n_i = n_i(\mathbf{x}), \text{ for } i = 1, \ldots, K$$

Since ξ is exchangeable, the probability $P_\xi(\mathbf{x})$ depends on the counts $\mathbf{n}(\mathbf{x})$ only, so

$$P_\xi(\mathbf{x}) =: g(n_1, \cdots, n_K).$$

We may now write:

$$g(n_1, \ldots, n_K) = P_\xi(\mathbf{x}) = \int_{\mathcal{P}} \prod_{i=1}^{n} q(x_i) \pi(dq) = \int_{\mathcal{P}} \prod_{k=1}^{K} q(k)^{n_k} \pi(dq). \tag{4.11}$$

In what follows, let us denote

$$\mathbb{N}_n := \{(n_1, \dots, n_K) : \sum_i n_i = n\}.$$

Observe that

$$\sum_{(n_1,\dots,n_k)\in\mathbb{N}_n} \frac{n!}{n_1! \cdots n_K!} g(n_1, \cdots, n_K) = 1.$$

Therefore, the measure P_n can be defined on the set $\mathbb{N}_n$ as follows:

$$P_n(\mathbf{n}) := \frac{1}{Z_n} \frac{n!}{n_1! \cdots n_K!} g(n_1, \dots, n_K) \prod_{k=1}^{K} \big(w_n(k)\big)^{n_k} = \frac{n!}{n_1! \cdots n_K!} \int_{\mathcal{P}} \prod_{k=1}^{K} \big(r_{q,n}(k)\big)^{n_k} \bar{\pi}_n(dq). \tag{4.12}$$

Considering the frequencies instead of counts, we can define the corresponding measure on the simplex $\mathcal{P}$. Let us denote that measure as Q_n, so that with $\mathbf{n}/n := (n_1/n, \dots n_k/n)$

$$Q_n\left(\frac{\mathbf{n}}{n}\right) := P_n(\mathbf{n}), \quad \forall \mathbf{n} \in \mathbb{N}_n. \tag{4.13}$$

Thus Q_n is a discrete measure

$$Q_n = \sum_{\mathbf{n}\in\mathbb{N}_n} P_n(\mathbf{n}) \delta_{\frac{\mathbf{n}}{n}}.$$

The advantage of Q_n over the measure P_n on $\mathbb{N}_n$ is that for any n, Q_n is defined on the same domain $\mathcal{P}$, and so one can speak about the weak convergence of Q_n. Essentially, obviously, the measure P_n on $\mathcal{X}^n$, the measure P_n on $\mathbb{N}_n$ and Q_n on $\mathcal{P}$ are all the same, just the domains are different.

Since the measures Q_n are defined on the same space (simplex), it is now natural to ask, whether there exists a probability measure Q^* so that $Q_n \Rightarrow Q^*$? It turns out the if the assumption of Lemma 4.1 holds, i.e. $\pi_n \Rightarrow \bar{\pi}$ and $w_n \to w$ (pointwise), then the limit measure is actually $\bar{\pi} r^{-1}$, where

$$r : \mathcal{P} \mapsto \mathcal{P}, \quad r(q) = r_q$$

and r is defined with respect to limit weight function w. Thus for a measurable $E \subset \mathcal{P}$,

$$\bar{\pi} r^{-1}(E) = \bar{\pi}\big(r^{-1}(E)\big).$$

For example, if $\bar{\pi} = \delta_{q^*}$ (the measure is concentrated on one point), then

$$\bar{\pi} r^{-1} = \delta_{r(q^*)},$$

because

$$\delta_{q^*} r^{-1}(E) = 1 \quad \Leftrightarrow \quad q^* \in r^{-1}(E) \quad \Leftrightarrow \quad r(q^*) \in E.$$

The following lemma is the counterpart of Lemma 4.1. Again, $\bar{\pi}_n$ is an arbitrary sequence of probability measures on $\mathcal{P}$, P_n are defined via $\bar{\pi}_n$ by (4.12) and Q_n via P_n as in (4.13).

Lemma 4.2 *Let $w_n(k) \to w(k)$ for every $k \in \mathcal{X}$. If there exists a probability measure $\bar{\pi}$ such that $\bar{\pi}_n \Rightarrow \bar{\pi}$, then $Q_n \Rightarrow \bar{\pi} r^{-1}$.*

Proof. Let $f : \mathcal{P} \to \mathbb{R}$ be a (K-variable) bounded continuous function. By definition of the weak convergence, it suffices to show that

$$\int f(q) Q_n(dq) \to \int f(q) \bar{\pi} r^{-1}(dq) = \int f(r_q) \bar{\pi}(dq), \tag{4.14}$$

where the last equality holds by the change of variable formula. Note that

$$\begin{aligned}
\int f(q) Q_n(dq) &= \sum_{\mathbf{n} \in \mathbb{N}_n} f\Big(\frac{\mathbf{n}}{n}\Big) P_n(\mathbf{n}) = \sum_{\mathbf{n} \in \mathbb{N}_n} f\Big(\frac{\mathbf{n}}{n}\Big) \frac{n!}{n_1! \cdots n_K!} \int \prod_{k=1}^{K} \big(r_{q,n}(k)\big)^{n_k} \bar{\pi}_n(dq) \\
&= \int \Big(\sum_{\mathbf{n} \in \mathbb{N}_n} f\Big(\frac{\mathbf{n}}{n}\Big) \frac{n!}{n_1! \cdots n_K!} \prod_{k=1}^{K} \big(r_{q,n}(k)\big)^{n_k} \Big) \bar{\pi}_n(dq) \\
&= \int f_n(r_{q,n}(1), \ldots, r_{q,n}(K)) \bar{\pi}_n(dq) = \int f_n(r_{q,n}) \bar{\pi}_n(dq),
\end{aligned}$$

where

$$f_n(r_{q,n}) =: f_n(r_{q,n}(1), \ldots, r_{q,n}(K)) := \sum_{(n_1, \ldots, n_K) \in \mathbb{N}_n :} f\Big(\frac{n_1}{n}, \ldots, \frac{n_1}{n}\Big) \frac{n!}{n_1! \cdots n_K!} \prod_{k=1}^{K} \big(r_{q,n}(k)\big)^{n_k}$$

is the Bernstein polynomial evaluated at $r_{q,n} = (r_{q,n}(1), \ldots, r_{q,n}(K))$. It is easy to see and well known that for any vector $r \in \mathcal{P}$, $f_n(r) \to f(r)$, moreover, the convergence is uniform over $\mathcal{P}$:

$$\sup_{r \in \mathcal{P}} \Big| f_n(r) - f(r) \Big| \to 0. \tag{4.15}$$

Since for every q, $r_{q,n}$ is a probability vector, then

$$\frac{n!}{n_1! \cdots n_K!} \prod_{k=1}^{K} \big(r_{q,n}(k)\big)^{n_k} \leq 1,$$

and since f is bounded, we see that for every n, $q \mapsto f_n(r_{q,n}) =: b_n(q)$ is a bounded continuous function. Also the function $q \mapsto f(r_q) =: b(q)$ is a bounded continuous function. Then

$$\sup_q |b_n(q) - b(q)| \leq \sup_q |f_n(r_{q,n}) - f(r_{q,n})| + \sup_q |f(r_{q,n}) - f(r_q)|.$$

By (4.9), $\sup_q |r_{q,n}(i) - r_q(i)| \to 0$ for every i. Then also $\sup_q \|r_{q,n} - r_q\| \to 0$. A continuous function on compact space is uniformly continuous, so

$$\sup_q |f(r_{q,n}) - f(r_q)| \to 0.$$

By (4.15),

$$\sup_q |f_n(r_{q,n}) - f(r_{q,n})| \leq \sup_r |f_n(r) - f(r)| \to 0.$$

Therefore, we have shown that $\sup_q |b_n(q) - b(q)|$ implying that

$$\int f(q)Q_n(dq) = \int b_n(q)\bar{\pi}_n(dq) \to \int b(q)\bar{\pi}(dq) = \int f(r_q)\bar{\pi}(dq).$$

■

5 P^* and Q^* in the large population limit

In what follows, let us rewrite fitnesses in terms of $\phi(k) := -\ln(w(k))$, so that for any genome $k \in \mathcal{X}$, $w(k) = \exp(-\phi(k))$. Moreover, in order to increase the influence of prior, we let weights w_n depend on n in the following way:

$$w_n(k) = \exp[-\frac{\phi(k)}{n^\lambda}], \quad k = 1, \ldots, K \tag{5.1}$$

where $\lambda \geq 0$ and $0 \leq \phi(1) < \phi(2) < \cdots < \phi(K)$. The case $\lambda = 0$ corresponds to fitness that is constant in that it does not vary with n. Clearly, for every k, $w_n(k) \to w(k)$ and λ controls the speed of that convergence. When $\lambda > 0$, then $w(k) = 1$ implying that in this case the mapping r is identity, i.e. for every q, $r_q = q$.

Let us return to our original $\bar{\pi}_n$, defined as in (4.4) with w_n:

$$\bar{\pi}_n(E) = \int_E \frac{\langle q, w_n\rangle^n}{Z_n}\pi(dq), \quad Z_n = \int \langle q, w\rangle^n \pi(dq). \tag{5.2}$$

In this section we consider the case where the prior measure π is independent of n, and the support of π is the whole simplex $\mathcal{P}$. Since by assumption $w(1) > w(2)$, clearly the function $q \mapsto \langle q, w\rangle$ has unique maximizer $q^* := (1, 0, \ldots, 0)$. The following theorem states that phase transition occurs: when $0 \leq \lambda < 1$, then $\bar{\pi}_n \Rightarrow \delta_{q^*}$ and $P_n(x_1, \ldots, x_m) \to P^*(x_1, \ldots, x_m)$, where

$$P^*(x_1, \ldots, x_m) = \prod_{i=1}^m r_{q^*}(x_i) = \prod_{i=1}^m q^*(x_i) = \begin{cases} 1, & \text{if for every } i,\ x_i = 1\ ; \\ 0, & \text{else.} \end{cases},$$

because in both cases (i.e. $\lambda = 0$ and $\lambda \in (0, 1)$), it holds that $r_{q^*} = q^*$. Thus the limit process $X_1, X_2, \ldots$ has only one realization: $1, 1, \ldots$. In this case also $Q_n \Rightarrow \delta_{q^*}$. Thus, when $\lambda \in [0, 1)$ then only the fittest genotype survives, no one else has any change, no matter what the prior says. In other words, the influence of the prior vanishes.
When $\lambda = 1$, then $\bar{\pi}_n$ as well Q_n converges to a nondegenerate distribution, specified below, and also the limit measure P^* is non-degenerate.
And finally, when $\lambda > 1$, then $\bar{\pi}_n \Rightarrow \pi$, $Q_n \Rightarrow \pi$ and the measure P^* is the law of birth process ξ. In this case the influence of fitness vanish and only the prior matters – the limit process equals to the breeding one.

Theorem 5.1 *Let the fitness function be defined as in (5.1) and assume that the support of the prior π is $\mathcal{P}$. Then the following convergences hold:*

1) *If $\lambda \in [0,1)$, then $\bar{\pi}_n \Rightarrow \delta_{q^*}$, $Q_n \Rightarrow \delta_{q^*}$ and (4.8) holds with*

$$P^*(x_1,\dots,x_m) = \prod_{i=1}^{m} q^*(x_i), \text{ where } q^* = (1,0,\dots,0).$$

2) *If $\lambda = 1$, then $\bar{\pi}_n \Rightarrow \bar{\pi}$, $Q_n \Rightarrow \bar{\pi}$ and (4.8) holds with*

$$P^*(x_1,\dots,x_m) = \int \prod_{i=1}^{m} q(x_i)\bar{\pi}(dq),$$

where for every $E \subset \mathcal{P}$,

$$\bar{\pi}(E) = \frac{1}{Z}\int_E \exp[-\langle \phi, q\rangle]\pi(dq), \quad Z = \int \exp[-\langle \phi, q\rangle]\pi(dq).$$

3) *If $\lambda > 1$, then $\bar{\pi}_n \Rightarrow \pi$, $Q_n \Rightarrow \pi$ and (4.8) holds with*

$$P^*(x_1,\dots,x_m) = P_\xi(x_1,\dots,x_m).$$

5.1 Proof of Theorem 5.1

Before proving the theorem, let us state a very useful preliminary result. Recall that simplex $\mathcal{P}$ is a compact set. Let $f_n, f : \mathcal{P} \to \mathbb{R}^+$ be continuous, hence bounded measurable functions so that $f_n \to f$ uniformly and let $m_n \to \infty$ be an increasing sequence. We are given a measure π on $\mathcal{P}$, and we are interested in the asymptotic behavior of the measure ν_n, where

$$\nu_n(E) := \int_E h_n(q)\pi(dq), \quad h_n(q) := \frac{f_n^{m_n}(q)}{\int f_n^{m_n}(q)\pi(dq)} = \Big(\frac{f_n(q)}{\|f_n\|_{m_n}}\Big)^{m_n}.$$

Here we assume that $\int f_n^{m_n} d\pi < \infty$ for every n. If π is a finite measure, then the conditions automatically holds due to the boundedness of f_n. In what follows, let

$$\mathcal{S}^* := \{q \in \mathcal{S} : f(q) = \|f\|_\infty\}, \quad \mathcal{S}^*_\delta := \{q \in \mathcal{S} : f(q) > \|f\|_\infty - \delta\}. \tag{5.3}$$

Here $\|f\|_\infty$ is the essential supremum of f with respect to the π-measure. If f is continuous and the support of π is $\mathcal{P}$, then $\|f\|_\infty = \sup_q f(q)$. The proof of the following Proposition 5.1 is given in appendix.

Proposition 5.1 *Let $f_n \to f$ uniformly and let π be a finite measure on $\mathcal{P}$. Then for every $\delta > 0$, $\nu_n\big(\mathcal{S}^*_\delta\big) \to 1$. If $\mathcal{S}^* = \{q^*\}$, then $\nu_n \Rightarrow \delta_{q^*}$.*

Besides Proposition 5.1, the proof of Theorem 5.1 is based on the following well-known observation: when $m \to \infty$, then

$$\sup_{q \in \mathcal{P}} \left| \langle \exp[-\frac{\phi}{m}], q \rangle^m - \exp[-\langle \phi, q \rangle] \right| \to 0. \tag{5.4}$$

Proof. (Theorem 5.1)

1) For $\lambda = 0$, take $f_n(q) = f(q) = \langle w, q \rangle$. From Proposition 5.1, it follows that $\bar{\pi}_n \Rightarrow \delta_{q^*}$. Since for any weight w, $r_{q^*}(k) = q^*(k)$, from Lemma 4.1, it follows that

$$P^*(x_1, \ldots, x_m) = \prod_{i=1}^m r_{q^*}(x_i) = \prod_{i=1}^m q^*(x_i).$$

Since $\bar{\pi}_n r^{-1}(q^*) = q^*$, from Lemma 4.2, it follows that $Q_n \Rightarrow \delta_{q^*}$. So, for $\lambda = 0$, the statement is proven and we now consider the case $\lambda \in (0, 1)$. Let

$$f_n(q) := \langle \exp[-\frac{\phi}{n^\lambda}], q \rangle^{n^\lambda}, \quad f(q) := \exp[-\langle \phi, q \rangle].$$

By (5.4), $\|f_n - f\|_\infty \to 0$. Since $\lambda \in (0, 1)$, take $m_n = n^{1-\lambda}$. Then

$$h_n(q) := \frac{\langle \exp[-\frac{\phi}{n^\lambda}], q \rangle^n}{Z_n}$$

is the density of $\bar{\pi}_n$ with respect to π. Since f is continuous, the set $\mathcal{S}^*$ in (5.3) is

$$\mathcal{S}^* = \arg\max_{q \in \mathcal{S}} f(q) = \arg\min_{q \in \mathcal{S}} \langle \phi, q \rangle = \{(1, 0, \ldots, 0)\} = \{q^*\}.$$

By Proposition 5.1, $\bar{\pi}_n \Rightarrow \delta_{q^*}$. As in the case of $\lambda = 0$, it follows that $Q_n \Rightarrow \delta_{q^*}$ and $P^*(x_1, \ldots, x_m) = 1$ if and only if $x_1 = \cdots = x_m = 1$.

2) Since for any q and any n, it holds

$$\langle \exp[-\frac{\phi}{n}], q \rangle^n \leq e^{-\phi(1)} \leq 1,$$

we obtain from (5.4) and bounded convergence that for any measurable E

$$\int_E \langle \exp[-\frac{\phi}{n}], q \rangle^n \pi(dq) \to \int_E \exp[-\langle \phi, q \rangle] \pi(dq). \tag{5.5}$$

Recall

$$\bar{\pi}(E) = \frac{1}{Z} \int_E \exp[-\langle \phi, q \rangle] \pi(dq), \quad \text{where} \quad Z = \int \exp[-\langle \phi, q \rangle] \pi(dq).$$

Therefore, from (5.5), it follows that when $\lambda = 1$, we have

$$\bar{\pi}_n(E) \to \bar{\pi}(E),$$

meaning that $\bar{\pi}_n \Rightarrow \bar{\pi}$ (even in a stronger sense). Since $r_q = q$, from Lemma 4.1, it follows that the limits of $P_n(x_1, \ldots, x_m)$ are

$$P^*(x_1, \ldots, x_m) = \int \prod_{i=1}^{m} q(x_i)\bar{\pi}(dq) = \frac{1}{Z} \int \prod_{i=1}^{m} q(x_i) \exp[-\langle \phi, q \rangle]\pi(dq).$$

Since r is identity function, by Lemma 4.2 the limit measure of frequencies is $\bar{\pi}$, i.e. $Q_n \Rightarrow \bar{\pi}$.

3) Since for any q,

$$\langle \exp[-\frac{\phi}{n^\lambda}], q \rangle^n \to 1$$

by dominated convergence, again, for any measurable E

$$\int_E \langle \exp[-\frac{\phi}{n}], q \rangle^n \pi(dq) \to \pi(E).$$

Therefore $\bar{\pi}_n \Rightarrow \pi$. By Lemma 4.1, the limits of $P_n(x_1, \ldots, x_m)$ are

$$P^*(x_1, \ldots, x_m) = \int \prod_{i=1}^{m} q(x_i)\pi(dq),$$

so that the limit process is ξ. The convergence $Q_n \Rightarrow \pi$ follows from Lemma 4.2.

■

We have seen that the critical case $\lambda = 1$ is the only case where the prior and fitnesses both determine the limit measure. In this case, the limit process $X_1, X_2, \ldots$, governed by P^* has marginals

$$P(X_i = k) = P(X_1 = k) = \int q(k)\bar{\pi}(dq) = \frac{1}{Z} \int e^{-\langle \phi, q \rangle} q(k)\pi(dq),$$
$$P(X_i = k, X_j = l) = P(X_1 = k, X_2 = l) = \frac{1}{Z} \int e^{-\langle \phi, q \rangle} q(k)q(l)\pi(dq).$$

It is also interesting to point out that in the critical case $\lambda = 1$, the measure $\bar{\pi}$ satisfies

$$\bar{\pi} = \arg\min_{\pi' \in E} D(\pi \| \pi'),$$

where E is a set of probability measures on $\mathcal{P}$, namely $E := \{\pi' : \int \langle \phi, q \rangle \pi'(dq) \geq c\}$ and $c > 0$ is a constant.

6 Dirichlet prior

Also in the current section we consider the weights $w_n(k)$ as in (5.1), where $\lambda \in [0, 1]$. We already know that in the case of constant priors, the case $\lambda < 1$ means that the

fitnesses will prevail over the prior, and the limit measure is degenerate one. Therefore, it is meaningful to consider the non-constant priors so that the influence of prior increases with suitable rate. Therefore, in the present section, we consider Dirichlet' priors

$$\pi_n = \mathrm{Dir}(n^{1-\lambda}\alpha_1, \dots n^{1-\lambda}\alpha_K), \tag{6.1}$$

where $\alpha := (\alpha_1, \dots, \alpha_K)$, $\alpha_k > 0$ and $|\alpha| := \sum_k \alpha_k$. The constant $\sum_i n^{1-\lambda}\alpha_k = |\alpha| n^{1-\lambda}$ is the so called *concentration* or *precision* parameter, the bigger that parameter, the more prior is concentrated over it expectation $(\alpha_1/|\alpha|, \dots, \alpha_K/|\alpha|)$. Increasing the concentration parameter increases the influence of prior, and now it is clear the the smaller is λ, the bigger must be the prior influence. This justifies the choice of $n^{1-\lambda}$. The case $\lambda = 1$ corresponds to already studied case of constant priors, therefore we now consider the case $\lambda \in [0, 1)$ The following theorem shows that the phase transition occurs again.

Theorem 6.1 *Let the fitness function be defined be defined as in (5.1) and the prior π_n as in (6.1). Let $\bar{\pi}_n$ be defined as in (5.2) with π_n instead of π. Then the following convergences hold:*

1) *If $\lambda = 0$, then $\bar{\pi}_n \Rightarrow \delta_{q^*}$, where q^* is the unique maximizer of the following function*

$$\ln\langle e^{-\phi}, q\rangle + \sum_{k=1}^{K} \alpha_k \ln q(k). \tag{6.2}$$

Then $Q_n \Rightarrow \delta_{r^*}$, where $r^* = r_{q^*}$, so that $r^*(k) \propto q^*(k)w(k)$, and (4.8) holds with

$$P^*(x_1, \dots, x_m) = \prod_{i=1}^{m} r^*(x_i).$$

2) *If $\lambda \in (0, 1)$, then $\bar{\pi}_n \Rightarrow \delta_{q^*}$, where q^* is the unique maximizer of the following function:*

$$-\langle \phi, q\rangle + \sum_{k=1}^{K} \alpha_k \ln q(k) \tag{6.3}$$

Then $Q_n \Rightarrow \delta_{q^*}$ and (4.8) holds with

$$P^*(x_1, \dots, x_m) = \prod_{i=1}^{m} q^*(x_i).$$

Let us start with proving the uniqueness of the solutions of (6.2) and (6.3). The proof of the following lemma is in appendix.

Lemma 6.1

1) *The function (6.2) has an unique maximizer q^*, where*

$$q^*(k) = \frac{\alpha_k}{(1+|\alpha|) - \frac{w(k)}{\theta}}, \quad k = 1, \dots, K \tag{6.4}$$

where $\theta > 0$ is a parameter satisfying $\theta = \langle w, q^* \rangle$.

2) *The function (6.3) has an unique maximizer q^*, where*

$$q^*(k) = \frac{\alpha_k}{\phi(k) + |\alpha| - \theta}, \quad k = 1, \dots, K, \tag{6.5}$$

where $\theta > 0$ is the parameter satisfying $\theta = \langle \phi, q^* \rangle$.

Proof of theorem 6.1.

1) In the case $\lambda = 0$, the measure $\bar{\pi}_n$ has the following density with respect to the Lebesgue measure:

$$\bar{\pi}_n(q) = \frac{1}{Z_n} \langle e^{-\phi}, q \rangle^n \cdot \frac{1}{B(n\alpha)} \prod_{k=1}^{K} (q(k))^{n(\alpha_k - \frac{1}{n})} = \frac{f_n(q)^n}{Z'_n},$$

where

$$f_n(q) := \langle e^{-\phi}, q \rangle \prod_{k=1}^{K} (q(k))^{(\alpha_k - \frac{1}{n})}, \quad Z'_n := \int f_n^n(q) dq.$$

Clearly for every q, $f_n(q) \to f(q)$, where

$$f(q) = \langle e^{-\phi}, q \rangle \prod_{k=1}^{K} (q(k))^{\alpha_k}.$$

It is not hard to see that the convergence is uniform, i.e. $\sup_q |f_n(q) - f(q)| \to 0$. By **1)** of Lemma 6.1, the function f has unique maximizer q^* (6.4), i.e. $\mathcal{S}^* = \{q^*\}$. Now apply Proposition 5.1 with π being the Lebesgue measure on $\mathcal{P}$ (hence π is finite) and $m_n = n$ so that

$$\nu_n(E) = \int_E h_n dq = \frac{1}{Z'_n} \int_E f_n(q)^n dq = \bar{\pi}_n(E).$$

Since all assumptions are fulfilled, we have $\bar{\pi}_n \Rightarrow \delta_{q^*}$. Since $w_n = w$, by Lemma 4.1 the limit process has finite-dimensional distributions

$$P^*(x_1, \dots, x_m) = \prod_{i=1}^{m} r^*(x_i), \quad \text{where} \quad r^* = r_{q^*}$$

so that the limit process P^* corresponds to a i.i.d. sequence $X_1, X_2, \dots$ with $X_1 \sim r^*$. According to Lemma 4.2, the frequencies Q_n converge weakly to the measure r^* and this is also quite obvious by SLLN.

2) The proof is similar: $\bar{\pi}_n$ has density (with respect to Lebesgue measure)

$$\frac{f_n(q)^{m_n}}{Z'_n}, \quad \text{where} \quad f_n(q) = \langle e^{-\frac{\phi}{n^\lambda}}, q\rangle^{n^\lambda} \prod_{k=1}^{K} (q(k))^{(\alpha_k - \frac{1}{n^{1-\lambda}})}, \quad m_n = n^{(1-\lambda)}.$$

Since

$$\langle e^{-\frac{\phi}{n^\lambda}}, q\rangle^{n^\lambda} \to e^{\langle \phi, q\rangle},$$

uniformly over q, we have that sequence f_n converges uniformly to

$$f(q) = e^{-\langle \phi, q\rangle} \prod_{k=1}^{K} (q(k))^{\alpha_k}.$$

By **2)** of Lemma 6.1, the function f has unique maximizer q^* (6.5). As in the case **1)**, it is easy to see that the assumptions of Proposition 5.1 are fulfilled with π being Lebesgue measure on $\mathcal{P}$, $m_n = n^{1-\lambda}$ and so $\bar{\pi}_n \Rightarrow \delta_{q^*}$. In the present case, for every $k = 1, \ldots, K$, $w_n(k) \to 1$ and so by Lemma 4.1, the limit process P^* is i.i.d. process with distribution q^* (because r is identity function). According to Lemma 4.2, the frequencies Q_n converge weakly to the measure δ_{q^*}.

6.1 Relation between $\lambda = 0$ and $\lambda \in (0, 1)$

From (5.4), it follows:

$$\sup_{q \in \mathcal{P}} \left| m \ln\langle \exp[-\frac{\phi}{m}], q\rangle + \langle \phi, q\rangle \right| \to 0 \tag{6.6}$$

so that with

$$f_m(q) := \ln\langle \exp[-\frac{\phi}{m}], q\rangle + \sum_k \frac{\alpha_k}{m} \ln q(k), \quad f(q) = -\langle \phi, q\rangle + \sum_k \alpha_k \ln q(k),$$

we have

$$\sup_{q \in \mathcal{P}} |m f_m(q) - f(q)| \to 0.$$

Since $f(q)$ is as in (6.3), it has the unique maximizer q^* given in (6.4). On the other hand, the maximizer of $m f_m(q)$ is the same as the maximizer of $f_m(q)$, which corresponds to (6.2) where ϕ is replaced by ϕ/m and α is replaced by α/m. Let this unique maximizer be q^*_m Since the functions $m f_m(\cdot)$ and $f(\cdot)$ are continuous, uniformly convergent and having unique maximum, it follows that $q^*_m \to q^*$ (in usual sense, because $\mathcal{P}$ is compact). Thus, we have proven the following proposition.

Proposition 6.1 *Let*

$$q^*_m = \arg\max_q \Big(\ln\langle \exp[-\frac{\phi}{m}], q\rangle + \sum_k \frac{\alpha_k}{m} \ln q(k) \Big)$$

and let q^* be the maximizer of (6.3). Let r^*_m be the corresponding r measure, i.e. $r^*_m(k) \propto q^m(k) \exp[-\frac{\phi(k)}{m}]$. Then $q^*_m \to q^*$ and $r^*_m \to q^*$.

6.2 Product of Dirichlet priors

Recall the setup in Subsection 2.1. The set of genomes is now $\mathcal{X}^L = \overbrace{\mathcal{X} \times \cdots \times \mathcal{X}}^{L}$ and the breeding process $\xi = (\xi^1, \ldots, \xi^L)$, where ξ^l are independent exchangeable processes. We now assume that the prior of ξ^l is $\pi^l = \mathrm{Dir}(\alpha^l)$, where $\alpha^l = (a_1^l, \ldots, \alpha_K^l)$, $l = 1, \ldots, L$. In this model, L different Polya urns are run independently. Let $\mathcal{P}^L$ be the set of L-fold product measures:

$$\mathcal{P}^L := \{q^1 \times \cdots \times q^L : q^j \in \mathcal{P}\},$$

where $\mathcal{P}$, as previously, stands for the $(K-1)$-dimensional simplex. Observe that $\mathcal{P}^L$ is a compact subset of the set of all possible probability measures on $\mathcal{X}^L$. Since the components of ξ are independent, the prior π of ξ is the product of Dirichlet measures $\pi = \pi^1 \times \cdots \times \pi^L$. This means that the support of π is $\mathcal{P}^L$ and for every element $q = q^1 \times \cdots \times q^L \in \mathcal{P}^L$, the density is (with slight abuse of notation, π stands for the measure as well as for its density)

$$\pi(q) = \prod_{l=1}^{L} \pi^l(q^l) = \frac{1}{B} \prod_{l=1}^{L} \prod_{k=1}^{K} \left(q(k)^l\right)^{\alpha_k^l - 1}, \quad B := \prod_{l=1}^{L} B(\alpha^l).$$

The function ϕ is now defined on the set $\mathcal{X}^L$, and so for any $q \in \mathcal{P}^L$,

$$\langle \phi, q \rangle = \sum_{(k_1, \ldots, k_L) \in \mathcal{X}^L} \phi(k_1, \ldots, k_L) q^1(k_1) \cdots q^L(k_L)$$

and $\langle e^{-\phi}, q \rangle$ is defined similarly. When $\lambda = 0$, the measure $\bar{\pi}_n$ has density $f_n(q)^n / Z_n'$, where

$$f_n(q) = \langle e^{-\phi}, q \rangle \prod_{l=1}^{L} \prod_{k=1}^{K} (q^l(k))^{(\alpha_k^l - \frac{1}{n})}, \quad Z_n' := \int f_n^n(q) dq.$$

Clearly $f_n(q)$ converges uniformly to

$$f(q) := \langle e^{-\phi}, q \rangle \prod_{l=1}^{L} \prod_{k=1}^{K} (q^l(k))^{(\alpha_k^l)}. \tag{6.7}$$

Similarly, when $\lambda \in (0, 1)$ the measure $\bar{\pi}_n$ has density $f_n(q)^{m_n} / Z_n'$, where $m_n = n^{1-\lambda}$,

$$f_n(q) = \langle e^{\frac{-\phi}{n^\lambda}}, q \rangle^{n^\lambda} \prod_{l=1}^{L} \prod_{k=1}^{K} (q^l(k))^{(\alpha_k^l - \frac{1}{n^{1-\lambda}})}, \quad Z_n' := \int f_n^{m_n}(q) dq.$$

Again, $f_n(q)$ converges uniformly to

$$f(q) := e^{\langle -\phi, q \rangle} \prod_{l=1}^{L} \prod_{k=1}^{K} (q^l(k))^{(\alpha_k^l)}. \tag{6.8}$$

When (6.7) (resp. (6.8)) have unique solution q^*, then the statements of Theorem 6.1 hold (the proof is the same):

1) Suppose $\lambda = 0$ and (6.7) has unique maximizer $q^* = q^1 \times \cdots \times q^L$. Then $\bar{\pi}_n \Rightarrow \delta_{q^*}$, and $Q_n \Rightarrow \delta_{r^*}$, where

$$r^*(k_1, \ldots, k_L) \propto w(k_1, \ldots, k_L) q^1(k_1) \cdots q^L(k_L),$$

and $w(k_1, \ldots, k_L) = \exp[-\phi(k_1, \ldots, k_L)]$. Observe that the measure r^* is not necessarily a product measure. Then also (4.8) holds with

$$P^*(x_1, \ldots, x_m) = \prod_{i=1}^{m} r^*(x_i), \quad x_i \in \mathcal{X}^L.$$

2) Suppose $\lambda \in (0, 1)$ and (6.8) has unique maximizer $q^* = q^1 \times \cdots \times q^L$. Then $\bar{\pi}_n \Rightarrow \delta_{q^*}$, $Q_n \Rightarrow \delta_{q^*}$ and (4.8) holds with

$$P^*(x_1, \ldots, x_m) = \prod_{i=1}^{m} q^*(x_i), \quad x_i \in \mathcal{X}^L.$$

In the case $L > 1$, the maximizer of (6.7) and (6.8) is not always unique. Whether it is unique or not depends on ϕ and vectors α^l. Indeed, maximizing (6.7) is equivalent to minimizing

$$-\ln\left(\langle e^{-\phi}, q\rangle\right) - \sum_{l=1}^{L}\sum_{k=1}^{K} \alpha_k^l \ln(q^l(k)), \tag{6.9}$$

and $-\sum_{l=1}^{L}\sum_{k=1}^{K} \alpha_k^l \ln(q^l(k))$ is always a convex function. So, when the parameters α_k^l are big enough, then the whole function (6.9) becomes convex. The same argument holds for (6.8). We shall present some sufficient conditions for convexity of (6.7) and (6.8) for the case $K = L = 2$ below.

Recall: when a positive continuous function $f(q)$ has one maximizer q^*, then for any sequence $m_n \to \infty$, the measures ν_n with densities proportional to $f^{m_n}(q)$ converge weakly to δ_{q^*}. When the function has, say, two maximizers, q_1^* and q_2^*, then by Proposition 5.1, for all disjoint open balls B_1 and B_2 so that $q_i^* \in B_i$, it still holds that $\nu_n(B_1) + \nu_n(B_2) \to 1$. Thus, when the measures ν_n are weakly convergent, then the limit measure is concentrated on $\{q_1^*, q_2^*\}$ so that the limit measure must be $p\delta_{q_1^*} + (1-p)\delta_{q_2^*}$, for some $p \in [0, 1]$. In this case $\nu_n(B_1) \to p$. However, the function f might be so that the limits $\nu_n(B_i)$ do not exist. And even if they do exist (i.e. the measures ν_n are weakly convergent), the limits p and $1 - p$ might be arbitrary real numbers, and hard to determine. Therefore the following theorem adapted from [9] (Theorem 5.7 and a remark after it) might be very useful.

Theorem 6.2 *Suppose $K \subset \mathbb{R}^k$ is a compact non-empty subset and let $g : K \to [0, \infty)$ be a twice continuously differentiable function with finitely many minimum points $\{a_1, \ldots, a_r\}$ all located in the interior of K. Let, for every $i = 1, \ldots r$ the Hessian of g at a_i be positive definite. Given any increasing sequence m_n, define the sequence of measures*

$$\nu_n(E) := \frac{1}{Z_n} \int_E \exp[-m_n \cdot g(x)]dx, \quad Z_n := \int \exp[-m_n \cdot g(x)]dx.$$

Then $\nu_n \Rightarrow \nu$, where

$$\nu = \sum_{i=1}^{r} p_i \delta_{a_i}, \qquad \textit{with } p_i \propto \frac{1}{\sqrt{\det H(a_i)}}$$

and $\det H(a_i)$ is a determinant of Hessian evaluated at a_i.

To apply the theorem in our case, let us first note that any $K-1$ dimensional simplex can be considered as a $K-1$-dimensional non-empty compact set

$$\mathcal{P}_K = \{(q(1), \dots, q(K-1)) : q(k) \geq 0, \sum_{k}^{K-1} q(k) \leq 1\}.$$

Therefore, our search space $\mathcal{P}^L$ can be considered as a subset in $\mathbb{R}^{L(K-1)}$. This subset has non-empty interior. Clearly any solution of (6.7) and (6.8) has all components strictly positive so that all maximizers of maximizer of (6.7) are interior points and the same holds for (6.8). We take $g(q) = -\ln f(q)$, where f is as in (6.7) or (6.8). Thus, for any m, $\exp[-mg(y)] = f^m(q)$ so that the measure defined in the statement of theorem is

$$\nu_n(E) \propto \int_E f^{m_n}(q) dq.$$

However, even when the measures ν_n converge weakly to a limit, it does not automatically follow that the measures $\bar{\pi}_n$ converge to the same limit even if f_n converges to f uniformly. This convergence might depend on the speed of the uniform convergence, and we leave it for the further studies and proceed with an example instead.

6.2.1 The case $K = L = 2$

Let us analyze more closely the case $K = 2$ and $L = 2$. Denote $q^1(1) =: z_1$ and $q^2(1) =: z_2$. Also denote $\alpha_k^1 = \alpha_k$ and $\alpha_k^2 = \beta_k$. The function (6.9) is

$$g(z_1, z_2) = -\ln\big(\langle w, z\rangle\big) - \alpha_1 \ln z_1 - \alpha_2 \ln(1 - z_1) - \beta_1 \ln z_2 - \beta_2 \ln(1 - z_2), \qquad (6.10)$$

where

$$\langle w, z\rangle = w(1,1)z_1z_2 + w(1,2)z_1(1-z_2) + w(2,1)(1-z_1)z_2 + w(2,2)(1-z_1)(1-z_2).$$

Thus with $w^* := w(1,1) - w(1,2) - w(2,1) + w(2,2)$ and

$$\begin{aligned} \theta_1^1 &:= w(1,1)z_2 + w(1,2)(1-z_2), \quad \theta_2^1 = w(2,1)z_2 + w(2,2)(1-z_2) \\ \theta_1^2 &:= w(1,1)z_1 + w(2,1)(1-z_1), \quad \theta_2^2 := w(1,2)z_1 + w(2,2)(1-z_1). \end{aligned}$$

we obtain the Hessian

$$\begin{pmatrix} \frac{\partial^2 g}{\partial z_1^2} & \frac{\partial^2 g}{\partial z_1 \partial z_2} \\ \frac{\partial^2 g}{\partial z_2 \partial z_1} & \frac{\partial^2 g}{\partial z_2^2} \end{pmatrix} = \begin{pmatrix} \frac{\alpha_1}{z_1^2} + \frac{\alpha_2}{(1-z_1)^2} + \frac{(\theta_1^1 - \theta_2^1)^2}{\langle w,z\rangle^2} & \frac{(\theta_1^1 - \theta_2^1)(\theta_1^2 - \theta_2^2) - w^*\langle w,z\rangle}{\langle w,z\rangle^2} \\ \frac{(\theta_1^1 - \theta_2^1)(\theta_1^2 - \theta_2^2) - w^*\langle w,z\rangle}{\langle w,z\rangle^2} & \frac{\beta_1}{z_2^2} + \frac{\beta_2}{(1-z_2)^2} + \frac{(\theta_1^2 - \theta_2^2)^2}{\langle w,z\rangle^2} \end{pmatrix}.$$

The elements in the main diagonal are strictly positive and therefore the matrix is positive definite if the determinant is positive, i.e.

$$\Big(\frac{\alpha_1}{z_1^2}+\frac{\alpha_2}{(1-z_1)^2}+\frac{(\theta_1^1-\theta_2^1)^2}{\langle w,z\rangle^2}\Big)\Big(\frac{\beta_1}{z_2^2}+\frac{\beta_2}{(1-z_2)^2}+\frac{(\theta_1^2-\theta_2^2)^2}{\langle w,z\rangle^2}\Big)>\Big(\frac{(\theta_1^1-\theta_2^1)(\theta_1^2-\theta_2^2)}{\langle w,z\rangle^2}-\frac{w^*}{\langle w,z\rangle}\Big)^2.$$

Observe: $\langle w,z\rangle \geq \min_{i,j} w(i,j) > 0$ and so

$$\Big|\frac{w^*}{\langle w,z\rangle}\Big| \leq \frac{|w^*|}{\min_{i,j} w(i,j)}.$$

On the other hand, for any $(z_1,z_2)\in[0,1]\times[0,1]$,

$$\frac{\alpha_1}{z_1^2}+\frac{\alpha_2}{(1-z_1)^2}\geq(\alpha_1^{\frac{1}{3}}+\alpha_2^{\frac{1}{3}})^3,\quad \frac{\beta_1}{z_2^2}+\frac{\beta_2}{(1-z_2)^2}\geq(\beta_1^{\frac{1}{3}}+\beta_2^{\frac{1}{3}})^3.$$

Therefore, it is not hard to see that when

$$(\alpha_1^{\frac{1}{3}}+\alpha_2^{\frac{1}{3}})^3(\beta_1^{\frac{1}{3}}+\beta_2^{\frac{1}{3}})^3>\Big(\frac{w^*}{\min_{i,j} w(i,j)}\Big)^2, \tag{6.11}$$

then the Hessian is always positive definite, i.e. the function g in (6.10) is strictly convex, and the minimum unique. In particular, the condition holds if $w^*=0$. In particular, if $\alpha_1=\beta_1$, $\alpha_2=\beta_2$ and $w(1,2)=w(2,1)$, then under (6.11) the unique solution z_1,z_2 satisfies $z_1=z_2$ (by symmetry). Indeed, in symmetric case it holds: if (z_1,z_2) is a solution, then so must be (z_2,z_1), and if the solution is unique, then it must be that $z_1=z_2$. If (6.11) fails, then the function might be non-convex, and for small α_k and β_k values it is (given $w^*\neq 0$), but evaluated at the minimums, the Hessian might still be positive definite and so Theorem 6.2 might apply.

Similarly, for (6.8)

$$g(z)=-\ln f(z)=\langle\phi,z\rangle-\alpha_1^1\ln z_1-\alpha_2^1\ln(1-z_1)-\alpha_1^2\ln z_2-\alpha_2^2\ln(1-z_2) \tag{6.12}$$

and so the Hessian is

$$\begin{pmatrix} \frac{\partial^2 g}{\partial z_1^2} & \frac{\partial^2 g}{\partial z_1\partial z_2} \\ \frac{\partial^2 g}{\partial z_2\partial z_1} & \frac{\partial^2 g}{\partial z_2^2}\end{pmatrix}=\begin{pmatrix} \frac{\alpha_1}{z_1^2}+\frac{\alpha_2}{(1-z_1)^2} & \phi^* \\ \phi^* & \frac{\beta}{z_2^2}+\frac{\beta_2}{(1-z_2)^2}\end{pmatrix},$$

where

$$\phi^*:=\phi(1,1)-\phi(1,2)-\phi(2,1)+\phi(2,2).$$

Since the elements of the main diagonal are are strictly positive, the matrix is positive definite if and only if the determinant is positive:

$$\Big(\frac{\alpha_1}{z_1^2}+\frac{\alpha_2}{(1-z_1)^2}\Big)\Big(\frac{\beta_1}{z_2^2}+\frac{\beta_2}{(1-z_2)^2}\Big)>(\phi^*)^2. \tag{6.13}$$

Thus, when the following inequality holds

$$(\alpha_1^{\frac{1}{3}} + \alpha_2^{\frac{1}{3}})^3(\beta_1^{\frac{1}{3}} + \beta_2^{\frac{1}{3}})^3 > (\phi^*)^2, \tag{6.14}$$

then the Hessian is always positive definite and the function (6.12) strictly convex implying that the minimum is unique. If the Hessian is not always unique, but (6.13) holds for minimums, then Theorem 6.2 applies. Again, when $\alpha = \beta$ and $\phi(1,2) = \phi(2,1)$, then under (6.14) the unique minimum is such that $z_1 = z_2$. For example, when $\alpha = \beta = (2,2)$ and $\phi(1,1) = 1, \phi(1,2) = \phi(2,1) = 2, \phi(2,2) = 3$, then $\phi^* = 0$ and, therefore, (6.14) holds. It means the minimum is unique, $z_1 = z_2$, and one can verify that

$$z_1 = z_2 = \frac{\sqrt{17} - 3}{2} \approx 0.561.$$

So the unique limit distribution in this case is $q \times q$, where $q = (z, 1-z)$. But when $\alpha = (2,3)$, $\beta = (3,2)$ and ϕ is as previously, then the minimum is again unique but since $\alpha \neq \beta$, we now have $z_1 \neq z_2$: $z_1 = \sqrt{6} - 2 \approx 0.45$, $\quad z_2 = \sqrt{7} - 2 \approx 0.645$. This means: the unique limit distribution is $q^1 \times q^2$, where $q^1 = (\sqrt{6} - 2, 3 - \sqrt{6})$, $q^2 = (\sqrt{7} - 2, 3 - \sqrt{7})$.

When $\alpha = \beta = (2,2)$, $\phi(1,1) = 4, \phi(1,2) = \phi(2,1) = 2, \phi(2,2) = 4$, then $\phi^* = 4$, but (6.14) still holds and therefore there is unique minimum: $z_1 = 0.5, z_2 = 0.5$. However, when the ϕ is as previously, but $\alpha = \beta = (0.25, 0.25)$, then (6.14) fails. It turns out that now the function is not convex and there are two minima:

$$(z_1 = \frac{2 - \sqrt{2}}{4}, z_2 = \frac{2 + \sqrt{2}}{4}), \quad (z_1 = \frac{\sqrt{2} + 2}{4}, z_2 = \frac{\sqrt{2} - 2}{4})$$

Observe that $z_2 = 1 - z_1$. Thus the limit measures are $q^1 \times q^2$ and $q^2 \times q^1$, where $q^1 = (z_1, z_2)$ and $q^2 = (z_2, z_1)$. These two product measures are different. Finally observe that in both cases (6.13) holds, so that by Theorem 6.2,

$$\nu_n \Rightarrow \frac{1}{2}\delta_{q^1 \times q^2} + \frac{1}{2}\delta_{q^2 \times q^1}.$$

The function

$$f_n(q) = \langle e^{-\frac{\phi}{n^\lambda}}, q \rangle^{n^\lambda} \prod_{l=1}^{2} \prod_{k=1}^{2} (q(k)^l)^{0.25 - \frac{1}{n^{1-\lambda}}}$$

is symmetric, i.e. $f_n(z_1, z_2) = f_n(z_2, z_1)$ and then $\bar{\pi}_n \Rightarrow \frac{1}{2}\delta_{q^1 \times q^2} + \frac{1}{2}\delta_{q^2 \times q^1}$. Since now $r_q = q$, by Lemma 4.1, (4.8) holds, with

$$P^*(x_1, \ldots, x_m) = \frac{1}{2}\prod_{i=1}^{m} q^1 \times q^2(x_i) + \frac{1}{2}\prod_{i=1}^{m} q^2 \times q^1(x_i), \quad x_i \in \mathcal{X}^L.$$

7 Experiments

Let $K = 2$, $\phi(1) = 0$, $\phi(2) = \ln 6$, $\alpha_1 = 0.3$, $\alpha_2 = 0.7$. Let us find the limit measures q^* as in Theorem 6.1 in the following cases: $\lambda = 0$, $\lambda \in (0,1)$ and $\lambda = 1$.

Case $\lambda = 0$: Then, as it can be easily checked by verifying (6.4) that $q^* = (3/5, 2/5)$. Since $\theta = \langle q^*, w\rangle = 2/3$, the measure r^* is as follows: $r^*_1 = w(1)q^*(1)/\theta = 9/10$ and $r^*(2) = w(2)q^*(2)/\theta = 1/10$. Therefore the limit process governed by P^* is an i.i.d. process with measure r^*, and so due to the weight function, the proportion of the first genotype has increased from 0.3 (according to the prior) to 0.9. Figure 1 illustrates the convergence.

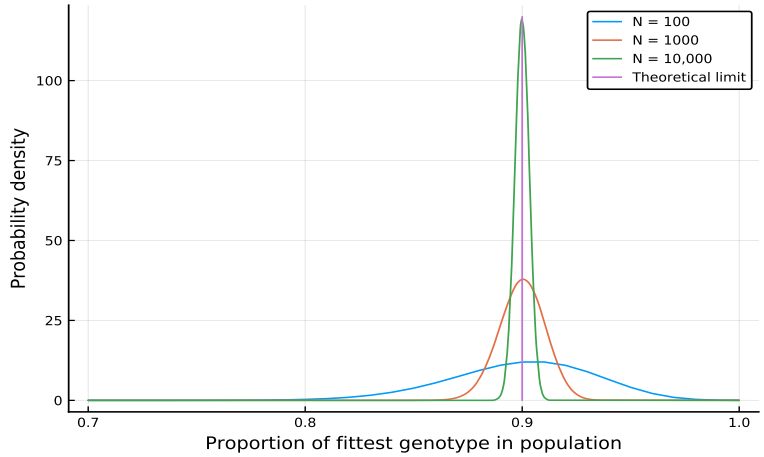

Figure 1: Density histogram of the fraction of the first type in the population, for a Dirichlet prior $\alpha = (0.3, 0.7)$, fitness $\phi = (0, \ln 6)$, and $\lambda = 0$, and three population sizes 10^2, 10^3, and 10^4. The histograms were constructed by recording the fraction of the first type in the population over 10^8 MCMC samples according to the process described in section 3.2

.

Case $\lambda \in (0, 1)$: The solution of (6.5) is

$$q^*(1) = \frac{\ln(6) - 1 + \sqrt{(\ln(6) - 1)^2 + 1.2\ln(6)}}{2\ln(6)} \approx 0.686. \tag{7.1}$$

Now $r^* = q^*$, thus we see that the limit process governed by P^* is an i.i.d. process with measure q^*, and so due to the weight function, the proportion of first genotype has increased from 0.3 (according to the prior) to 0.689. The increase is smaller than in the previous case. Figures 2, 3 and 4 illustrates the convergence for $\lambda = 0.25, 0.5, 0.75$, respectively. We see that although the limit is the same, the speed of convergence depends very much on λ. Let $q^*_m = (q_m(1), q_m(2))$ be the maximizer of

$$\ln\langle \exp[-\frac{\phi}{m}], q\rangle + \sum_k \frac{\alpha_k}{m} \ln q(k).$$

In our example $q_1(1) = 3/5$ (the case $\lambda = 0$) and

$$q_m(1) = \frac{b_m + \sqrt{b_m^2 + \frac{12}{10}(1+m)\big(1-(1/6)^{\frac{1}{m}}\big)(1/6)^{\frac{1}{m}}}}{2(1+m)\big(1-(1/6)^{\frac{1}{m}}\big)}, \text{ where } b_m = (m{+}0.3) - (1/6)^{\frac{1}{m}}(1.3{+}m).$$

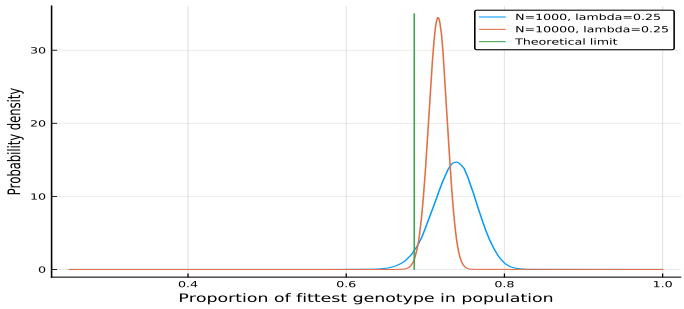

Figure 2: Case $\lambda = 0.25$: density histogram of the fraction of the first type in the population, for a Dirichlet prior $\alpha = (0.3, 0.7)$, fitness $\phi = (0, \ln 6)$, and $\lambda = 0.25$, and population sizes 10^3 and 10^4. The histograms were constructed by recording the fraction of the fitter allele in the population over 10^8 MCMC samples according to the process described in section 3.2

.

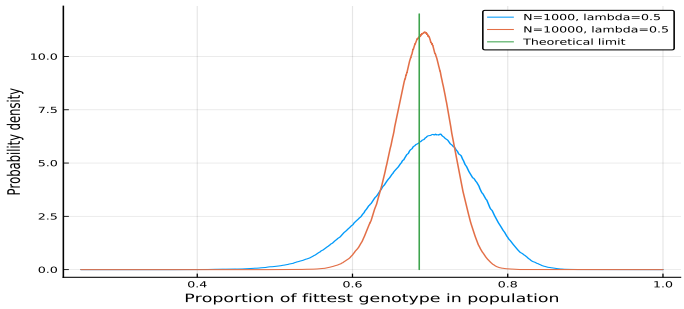

Figure 3: Case $\lambda = 0.5$: density histogram of the fraction of the first type in the population, for a Dirichlet prior $\alpha = (0.3, 0.7)$, fitness $\phi = (0, \ln 6)$, and population sizes 10^3 and 10^4. The histograms were constructed by recording the fraction of the fitter allele in the population over 10^8 MCMC samples according to the process described in section 3.2

.

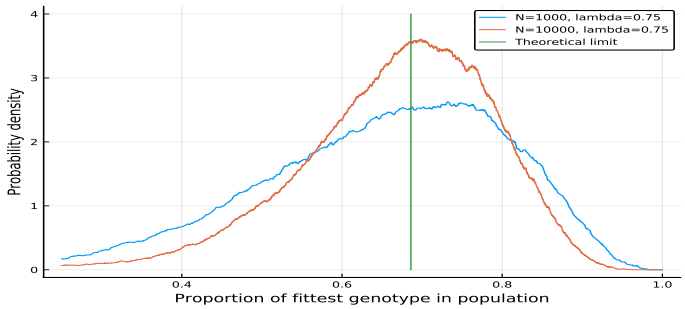

Figure 4: Case $\lambda = 0.75$: density histogram of the fraction of the first type in the population, for a Dirichlet prior $\alpha = (0.3, 0.7)$, fitness $\phi = (0, \ln 6)$, and population sizes 10^3 and 10^4. The histograms were constructed by recording the fraction of the fitter allele in the population over 10^{10} MCMC samples according to the process described in section 3.2

.

According to Proposition 6.1, in the process $m \to \infty$, $q_m(1)$ tends to $q^*(1)$ from (7.1), and one can easily verify that it is really so.

Case $\lambda = 1$: According to **2)** of Theorem 5.1, the limit measure $\bar{\pi}$ has density with respect to the Lebesgue'i measure on $[0, 1]$:

$$\bar{\pi}(q) = \frac{1}{Z} \exp[-\ln 6 \cdot q](1-q)^{0.3-1} q^{0.7-1},$$

where Z is the valuer of the moment generating function of Beta(0.7,0.3)-distributed random variable evaluated at $\ln 6$. In the density above, q stands for $q(2)$. Therefore, the limit proportion of the second genotype in the stochastic process governed by P^* is a random variable, its distribution has density $\bar{\pi}(q)$ as stated above. Figure 5 illustrates the convergence.

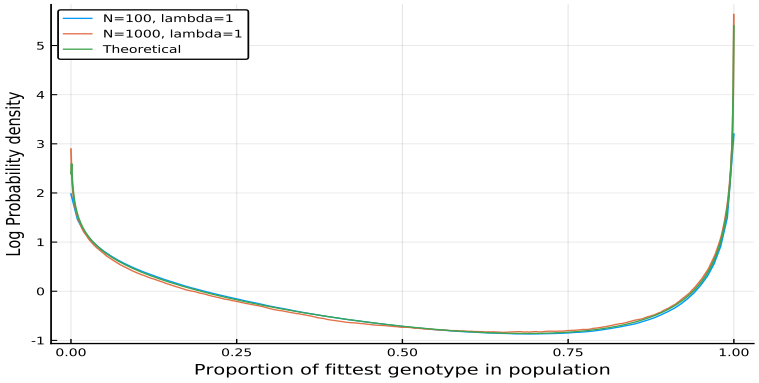

Figure 5: Case $\lambda = 1$: Density histogram (log scale) of the fraction of the first type in the population, for a Dirichlet prior $\alpha = (0.3, 0.7)$, fitness $\phi = (0, \ln 6)$, and population sizes 10^2 and 10^3. Note that because the concentration parameter is kept constant, the limit distribution is a reweighted β distribution, with infinities at 0 and 1: when $\lambda = 0$, the prior is the same for all N, and the log-fitness term is reduced by a factor of N, so that P_n never concentrates. The histograms were constructed by recording the fraction of the fitter allele in the population over 10^9 and 10^{10} MCMC samples respectively according to the process described in section 3.2

.

8 Conclusions

We have constructed MCMC algorithms that are similar to existing genetic algorithms. The 'breeding' consists of sampling from exchangeable distributions based on the Dirichlet distribution, and the 'selection' is essentially Metropolis-Hastings. The sequence of populations forms a reversible Markov chain that satisfies detailed balance conditions. We have exhibited two possible sampling distributions: more elaborate exchangeable sampling distributions are possible. The entire MCMC procedure is a population generalisation of

Metropolis-Hastings. As far as we are aware, this is the first plausible and general model of sexual reproduction that exactly satisfies detailed balance, and for which the stationary distribution can be written in closed form for arbitrary fitness functions.

We also explored some properties of the stationary distribution, and showed that for any fitness function there are three non-trivial limiting distributions for large population sizes, with two phase transitions. This is a first step towards a more general understanding of the interaction of the population size, fitness scaling, and mutation rate in genetic algorithms and evolutionary models.

Formulating a genetic model as a MCMC procedure opens a new research direction in using the many techniques developed in MCMC to achieve faster convergence to the stationary distribution using different MCMC kernels.

We have shown that the stationary distribution is unaffected by multiplicative noise in fitness evaluations. This has been suggested by, for example, [13], but our techniques allow a proof of this effect.

Finally there is a more general conclusion from our analysis. For many years, since [11] and [8], a widely suggested folk-motivation for genetic algorithms has been that because they are inspired by natural biological evolution, and because evolution has produced the variety of life on earth, genetic algorithms should be in some sense generally effective. Our analysis makes it clear that genetic algorithms are more closely related to conventional MCMC methods for non-parametric Bayesian inference than has previously been recognised.

Appendix

Proof of claim 4.1 **Proof.** First note that

$$\sup_{q\in\mathcal{P}} |\langle w_n, q\rangle - \langle w, q\rangle| \le \|w_n - w\|\|q\| \le \|w_n - w\| \to 0.$$

Now use the fact that if f_n, f, g_n, g are nonnegative functions such that $\sup_x |f_n(x) - f(x)| \to 0$, $\sup_x |g_n(x) - g(x)| \to 0$, $\inf_x g(x) = g_* > 0$ and f, g are bounded above, then with $g^* = \sup_x g(x)$ and $f^* = \sup_x f(x)$

$$\begin{aligned}
\sup_x \Big|\frac{f_n(x)}{f(x)} - \frac{g_n(x)}{g(x)}\Big| &= \sup_x \Big|\frac{f_n(x)g(x) - g_n(x)f(x)}{g_n(x)g(x)}\Big| \\
&\le \sup_x \Big|\frac{(f_n(x) - f(x))g(x)}{g_n(x)g(x)}\Big| + \sup_x \Big|\frac{(g_n(x) - g(x))f(x)}{g_n(x)g(x)}\Big| \\
&\le \sup_x \Big|\frac{(f_n(x) - f(x))g^*}{g_n(x)g(x)}\Big| + \sup_x \Big|\frac{(g_n(x) - g(x))f^*}{g_n(x)g(x)}\Big| \to 0,
\end{aligned}$$

because for n big enough $g_n(x)g(x) \ge \frac{g_*^2}{2}$ for every x. Take $x = q$, $f_n(q) = w_n(k)q(k)$, $f(q) = w(k)q(k)$, $g_n(q) = \langle w_n, q\rangle$ and $g(q) = \langle w, q\rangle$. Then $g_* = w(K) > 0$, $f^* = g^* = w(1)$ and so (4.9) follows. ■

8.1 Proof of Proposition 5.1

Recall that f_n and f are continuous and bounded functions on $\mathcal{P}$ so that $\|f_n\|_\infty < \infty$ and $\|f\|_\infty < \infty$. By assumption, π is a finite measure. Since f_n converges to f uniformly, it follows that $\|f_n - f\|_\infty \to 0$ and so $\|f_n\|_\infty \to \|f\|_\infty$. For every m,

$$|\|f_n\|_m - \|f\|_m| \le \|f_n - f\|_m \le \pi(\mathcal{P})^{\frac{1}{m}} \|f_n - f\|_\infty \to 0.$$

Since $\|f\|_{m_n} \to \|f\|_\infty$, we have

$$\begin{aligned}\big|\|f_n\|_{m_n} - \|f\|_\infty\big| &\le \big|\|f_n\|_{m_n} - \|f\|_{m_n}\big| + \big|\|f\|_{m_n} - |f\|_\infty\big| \\ &\le \|f_n - f\|_{m_n} + \big|\|f\|_{m_n} - |f\|_\infty\big| \le \pi(\mathcal{P})^{\frac{1}{m_n}} \|f_n - f\|_\infty + \big|\|f\|_{m_n} - |f\|_\infty\big| \to 0.\end{aligned}$$

Now fix $\delta > 0$. Since $\mathcal{S}^*_\delta := \{q : f(q) > \|f\|_\infty - \delta\}$, we have $\mathcal{S} - \mathcal{S}^*_\delta = \{q : f(q) \le \|f\|_\infty - \delta\}$. Define $\delta' := \delta / \|f\|_\infty$. Then

$$\begin{aligned}\sup_{q \in \mathcal{S} - \mathcal{S}^*_\delta} \frac{f_n(q)}{\|f_n\|_{m_n}} &= \sup_{q \in \mathcal{S} - \mathcal{S}^*_\delta} \frac{f(q) + (f_n(q) - f(q))}{\|f_n\|_{m_n}} = \sup_{q \in \mathcal{S} - \mathcal{S}^*_\delta} \frac{f(q) + (f_n(q) - f(q))}{\|f\|_\infty} \frac{\|f\|_\infty}{\|f_n\|_{m_n}} \\ &\le \sup_{q \in \mathcal{S} - \mathcal{S}^*_\delta} \frac{f(q)}{\|f\|_\infty} \frac{\|f\|_\infty}{\|f_n\|_{m_n}} + \frac{\|f_n - f\|_\infty}{\|f_n\|_{m_n}} \le 1 - \frac{\delta'}{2},\end{aligned}$$

provided n is big enough. Thus,

$$\sup_{q \in \mathcal{S} - \mathcal{S}^*_\delta} h_n(q) \le \Big(1 - \frac{\delta'}{2}\Big)^{m_n} \to 0,$$

so that $\nu_n(\mathcal{S}^*_\delta) \to 1$. We now argue that for any $\epsilon > 0$ there exists $\delta > 0$ so that

$$\mathcal{S}^*_\delta \subset B(q^*, \epsilon), \tag{8.1}$$

where $B(q^*, \epsilon)$ is a ball in Euclidean sense. If, for an $\epsilon > 0$, such a $\delta > 0$ would not exists, then there would exist a sequence $q_n \to q$ such that $f(q_n) \nearrow f(q)$, but $\|q_n - q\| \ge \epsilon$. Since $\mathcal{P}$ is compact, along a subsequence $q_{n'} \to q$ and by continuity $f(q_{n'}) \to f(q)$. On the other hand $\|q - q^*\| > \epsilon$ and that would contradict the uniqueness of q^*. Therefore (8.1) holds and so for any $\epsilon > 0$, it holds that $\nu_n\big(B(q^*, \epsilon)\big) \to 1$. From the definition of the weak convergence, it now follows that $\nu_n \Rightarrow \delta_{q^*}$.

8.2 Proof of Lemma 6.1

1) To find

$$q^* = \arg\max_{q \in \mathcal{P}} \ [\ln\langle e^{-\phi}, q\rangle + \sum_k \alpha_k \ln q(k)], \tag{8.2}$$

we define Lagrangian

$$L(q, \beta) = \ln\langle e^{-\phi}, q\rangle + \sum_k \alpha_k \ln q(k) - \beta(\sum_k q(k) - 1)$$

(here β is a scalar) and maximize $L(q,\beta)$ over $q>0$ (all entries are positive). Taking partial derivatives with respect to $q(k)$, we have

$$\frac{e^{-\phi(k)}}{\langle e^{-\phi},q\rangle}+\frac{\alpha_k}{q(k)}=\beta,\quad\Rightarrow\quad\frac{e^{-\phi(k)}q(k)}{\langle e^{-\phi},q\rangle}+\alpha_k=q(k)\beta\quad\forall k.$$

With $|\alpha|=\sum_k\alpha_k$, we have thus $\beta=1+|\alpha|$ and so the solution q^* satisfies the set of equalities

$$q^*(k)=\frac{1}{1+|\alpha|}\Big(\frac{e^{-\phi(k)}q^*_k}{\langle e^{-\phi},q^*\rangle}+\alpha_k\Big),\quad\forall k. \tag{8.3}$$

Now with $w(k)=e^{-\phi(k)}$ define parameter $\theta:=\langle w,q^*\rangle$ and rewrite (8.3) as follows

$$q^*(k)=\frac{\alpha_k}{(1+|\alpha|)-\frac{w(k)}{\theta}}\quad k=1,\dots,K. \tag{8.4}$$

We see that amongst the probability vectors satisfying $\langle w,q^*\rangle=\theta$, the solution is unique. Since $\alpha_k>0$ for every k, it is easy to see that there is only one parameter θ such that the right hand side of (8.4) would be a probability measure: if $\theta'>\theta$, then for every k, we have

$$\frac{\alpha_k}{(1+|\alpha|)-\frac{w(k)}{\theta}}>\frac{\alpha_k}{(1+|\alpha|)-\frac{w(k)}{\theta'}}.$$

Therefore a solution of (8.2) is unique vector q^* given by (8.4), where $\theta=\langle w,q^*\rangle$.

2) To find

$$q^*=\arg\max_{q\in\mathcal{P}}[-\langle\phi,q\rangle+\sum_k\alpha_k\ln q(k)], \tag{8.5}$$

we define Lagrangian

$$L(q,\beta)=-\langle\phi,q\rangle+\sum_k\alpha_k\ln q(k)-\beta(\sum_k q(k)-1).$$

Partial derivatives with respect to $q(k)$ give us the equalities

$$-\phi(k)+\frac{\alpha_k}{q(k)}=\beta\quad\forall k\quad\Rightarrow\quad-\langle\phi,q\rangle+|\alpha|=\beta.$$

Therefore, the inequalities for $q^*(k)$ are

$$q^*(k)=\frac{\phi(k)q^*(k)-\alpha_k}{\langle\phi,q^*\rangle-|\alpha|}=\frac{\phi(k)q^*(k)-\alpha_k}{\theta-|\alpha|},\quad\theta:=\langle\phi,q^*\rangle. \tag{8.6}$$

After rewriting (8.6), we obtain

$$q^*(k)=\frac{\alpha_k}{\phi(k)+|\alpha|-\theta},\quad k=1,\dots,K,$$

Thus, there cannot be two solutions having the same θ. As in the case **1)**, it is easy to see that when $\alpha_k>0$ there is only one θ so that (8.2) sums up to one. Therefore, the solution to the problem (8.5) is unique. Note that the solution is independent of λ.

Acknowledgments

The research is supported by Estonian Institutional research funding IUT34-5.

References

[1] S. Baluja and R. Caruana. Removing the genetics from the standard genetic algorithm. In *ICML*, pages 38–46. Morgan Kauffman Publishers, Inc., 1995.

[2] Shumeet Baluja. Genetic algorithms and explicit search statistics. In *Advances in Neural Information Processing Systems*, pages 319–325, 1997.

[3] E.B. Baum, D. Boneh, and C. Garrett. Where genetic algorithms excel. *Evolutionary Computation*, 9(1):93–124, 2001.

[4] James Franklin Crow and Motoo Kimura. *An introduction to population genetics theory.* New York, Evanston and London: Harper & Row, Publishers, 1970.

[5] Lloyd Elliott and Yee Whye Teh. Scalable imputation of genetic data with a discrete fragmentation-coagulation process. In *Advances in Neural Information Processing Systems 25*, pages 2861–2869, 2012.

[6] W.J. Ewens. *Mathematical population genetics: theoretical introduction*, volume 1. Springer Verlag, 2004.

[7] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

[8] D.E. Goldberg. *Genetic algorithms in search, optimization, and machine learning.* Addison-wesley, 1989.

[9] Heikki Haario and Eero Saksman. Simulated annealing process in general state space. *Advances in Applied Probability*, 23(4):866–893, 1991.

[10] Nils Lid Hjort, Chris Holmes, Peter Muller, and Stephen Walker. *Bayesian nonparametrics.* Number 28. Cambridge University Press, 2010.

[11] J.H. Holland. *Adaptation in natural and artificial systems.* University of Michigan press, 1975.

[12] P.A.P. Moran. *The statistical processes of evolutionary theory.* Clarendon Press; Oxford University Press., 1962.

[13] Gregory Morse and Kenneth O Stanley. Simple evolutionary optimization can rival stochastic gradient descent in neural networks. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, pages 477–484. ACM, 2016.

[14] R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, pages 249–265, 2000.

[15] M.J.A. Strens. Evolutionary MCMC sampling and optimization in discrete spaces. In *ICML*, volume 20, page 736, 2003.

[16] Yee Whye Teh and Michael I Jordan. Hierarchical Bayesian nonparametric models with applications. In *Bayesian nonparametrics*, pages 158–207. Cambridge University Press, 2010.

[17] Cajo JF Ter Braak. A Markov Chain Monte Carlo version of the genetic algorithm differential evolution: easy Bayesian computing for real parameter spaces. *Statistics and Computing*, 16(3):239–249, 2006.

[18] M.D. Vose. *The simple genetic algorithm: foundations and theory*. The MIT Press, 1999.

[19] Chris Watkins and Yvonne Buttkewitz. Sex as Gibbs sampling: a probability model of evolution. *arXiv preprint arXiv:1402.2704*, 2014.