12 Evolutionary Systems I

There are many misconceptions about what is understood by the word "evolution". Even the definitions given in the Oxford Concise Science Dictionary, Chambers, or Webster's are wrong, since they associate evolution with progress. However, according to the evolutionary biologists, evolution is *any* change in the frequency of characteristics in the gene pool from one generation to the next (see, for instance, Futujama 86 or Curtis and Barnes 89).

As with many scientific fields, also the history of evolutionary theories dates back to the Greek philosophers. The first that are known to have thought about evolution and how to pass on characteristics from one generation to the next were Anaxagoras (500-428 b.c.) and Aristoteles (384-322 b.c.). This was followed by more or less obscure theories about evolution for about 1,500 years until Charles Darwin proposed his theory in 1859 that evolution is caused by natural selection: Within a species, random variants occur. These variants introduce new characteristics that will influence the frequency of characteristics in a species depending on how they will affect the production of offspring and their survival. His view is widely accepted today as the basic explanation for evolution. Modern biochemistry and genetics were able to refine this explanation and to identify the basic mechanisms for passing on characteristics from one generation to the next.

In essence, all genetic information (that is, all information that can be passed on from one generation to the next) is contained in the DNA (or desoxyribonuclein acid). This DNA is stored in every cell of an individual. In the case of bacteria, it consists of a single, ring-shaped chromosome, and in the case of higher organisms, it may consist of up to 23 pairs of chromosomes. A chromosome is a string of molecules that encode the genetic information of an organism.

There are basically two different ways of passing genetic information from one generation to the next: the asexual (as it used by bacteria) and the sexual (as it is used by mammals) creation of offspring. Figure 1 shows the basic phases in this process.

In bacteria, basically only mutation (i.e., the random change of genetic information) can cause the development of new variants. For higher organisms, an additional, powerful mechanism called *recombination* is used. The recombination consists of two parts: the *intrachromosomal recombination*, and the *interchromosomal recombination*. During the intrachromosomal recombination, each pair of chromosomes may exchange some parts of their strings within themselves, and during the interchromosomal recombination, one chromosome is independently selected from each pair to form in the case of male organisms two spermatides, and in the case of female organisms one ovum. In both cases, the resulting cells contain only half of the usual amount of chromosomes (one instead of two chromosomes of each type). The combination of a spermatide and an ovum will then bring back the chromosomes to a complete set of pairs.

Since the 1950s, people tried to use evolutionary mechanisms to develop algorithms for hard optimization problems. Methods that only use mutations are, for instance, the metropolis algorithm (Metropolis et al 1953), simulated annealing (Kirkpatrick 1982), and evolutionary programming (Fogel, Owens, and Walsh 1966). Methods that also use recombination are, for instance, evolution strategies (Bienert, Rechenberg, and Schwefel 1970) and genetic algorithms (Holland 1975). Evolutionary algorithms aim at solving an optimization problem by using models of natural evolution. For this they use a population of individuals, where each individual represents a point in the space of feasible solutions. The initial population is usually selected by a problem-specific method, and it evolves afterwards towards successively better regions

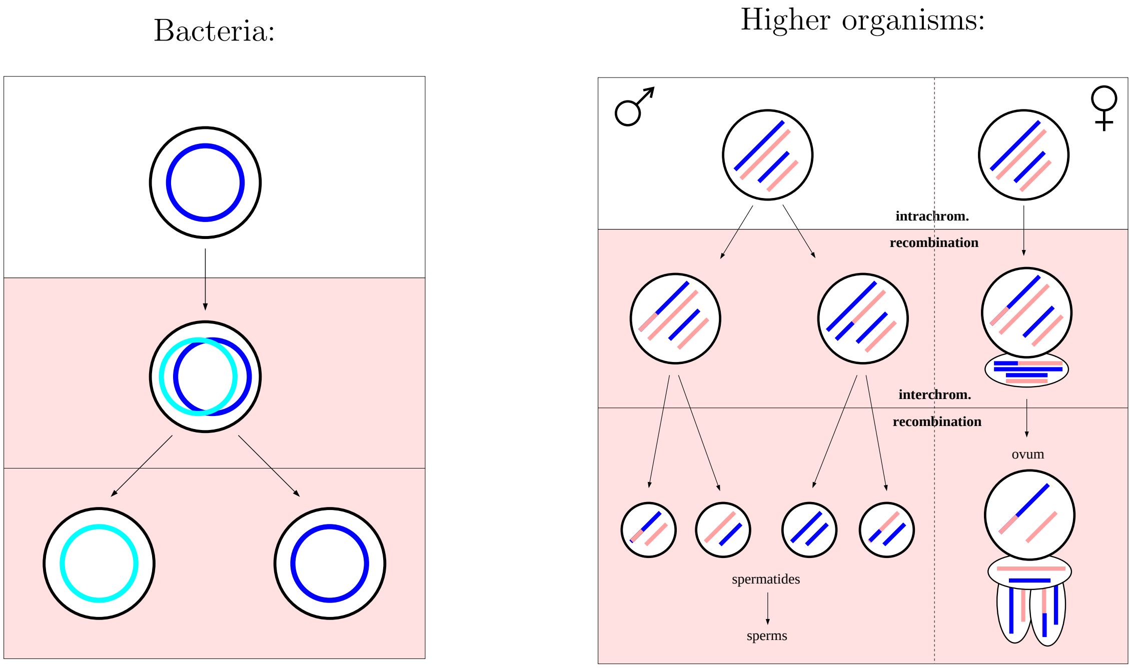


Figure 1: Heredity of Genetic Information.

of the search space by means of mechanisms such as recombination, mutation, and selection. The environment (representing the optimization problem) gives a quality information (or fitness value) for new search points, and the selection process favors those individuals of higher quality. The recombination mechanism allows the mixing of parental information while passing it to their descendants, and mutation introduces new features to the population.

Today, algorithms based on evolutionary methods form the most important class of heuristics for finding good solutions to hard optimization problems. An important question has been, how powerful these algorithmic approaches are. In order to answer this for particular cases, we will introduce the concept of dynamical systems.

12.1 Dynamical systems

A dynamical system consists of:

- a finite set S of *states*, $s = |S|$,
- an initial *potential vector* $\boldsymbol{q}^{(0)} = (q_j)_{1 \leq j \leq s} \in \mathbb{R}^s$ and
- a *transition function* $f : \mathbb{R}^s \to \mathbb{R}^s$

The system works in *rounds*. Initially, the potential vector is given by $\boldsymbol{q}^{(0)}$. Given a potential vector of $\boldsymbol{q}^{(t)}$ after t rounds, we obtain a potential vector of $\boldsymbol{q}^{(t+1)}$ after $t+1$ rounds by

$$\boldsymbol{q}^{(t+1)} = f(\boldsymbol{q}^{(t)})$$

An important aspect in the area of dynamical systems is the *trajectory* of a point $q \in \mathrm{IR}^s$, i.e. the sequence of points $\boldsymbol{q}$, $f(\boldsymbol{q})$, $f(f(\boldsymbol{q}))$,... There are many interesting questions concerning these trajectories:

What is the geometric or topological structure of the points or sets against which the system converges if such structures exist? These structures are usually called *attractors*. How sensitive is the system or the choice of the attractor against deviations? How long does it take until the trajectory has reached an attractor (sufficiently closely)?

We will deal with these questions in the following for the special cases of linear and quadratic dynamical systems.

Definition 12.1 *A dynamical system is called* linear *if f is of the form*

$$f(\boldsymbol{q}) = (g_1(\boldsymbol{q}), \ldots, g_s(\boldsymbol{q}))$$

with the property that every g_j is linear, that is, of the form

$$g_j(\boldsymbol{q}) = a_{0,j} + \sum_{1 \leq i \leq s} q_i \cdot a_{i,j}$$

with $a_{i,j} \in \mathrm{IR}$ for all i, j. Thus, instead of a mapping $\boldsymbol{q}^{(t+1)} = f(\boldsymbol{q}^{(t)})$ we can use a vector $\boldsymbol{a} = (a_{0,j})_{1 \leq j \leq s}$ and a matrix $\boldsymbol{A} = (a_{i,j})_{1 \leq i,j \leq s}$ with

$$\boldsymbol{q}^{(t+1)} = \boldsymbol{a} + \boldsymbol{q}^{(t)} \cdot \boldsymbol{A}$$

for all t.

Definition 12.2 *A dynamical system is called* quadratic *if f can be expressed as*

$$f(\boldsymbol{q}) = (g_1(\boldsymbol{q}), \ldots, g_s(\boldsymbol{q}))$$

with the property that every g_j is at most quadratic, that is, of the form

$$g_j(\boldsymbol{q}) = a_{0,j} + \sum_{1 \leq i \leq s} q_i \cdot a_{i,j} + \sum_{1 \leq i,k \leq s} q_i \cdot q_k \cdot b_{i,k,j}$$

with $a_{i,j}, b_{i,j,k} \in \mathrm{IR}$ for all i, j, k.

We will see that these systems are of great interest for the study of evolutionary systems, i.e. dynamic systems based on evolutionary operations.

13 Linear Dynamical Systems

We will concentrate in this section on linear systems with a transition function f of the form

$$f(\boldsymbol{q}) = \boldsymbol{q} \cdot \boldsymbol{A}$$

for a stochastic $s \times s$ matrix $\boldsymbol{A}$.

A matrix $\boldsymbol{A} = (a_{i,j})_{1\leq i,j\leq s}$ is called *stochastic* if $a_{i,j} \in [0,1]$ for all i,j and

$$\sum_{j=1}^{s} a_{i,j} = 1$$

for all i.

If the initial vector $\boldsymbol{q}^{(0)}$ represents a probability distribution, that is, $q_j \in [0,1]$ for all j and $\sum_{j=1}^{s} q_j = 1$, and $\boldsymbol{A}$ is a stochastic matrix, a linear dynamical system is also called *Markov chain*. If $\boldsymbol{A}$ is stochastic, then we will denote it in the following by $\boldsymbol{P} = (p_{i,j})$.

To sum up, a Markov chain consists of the following parts:

- a finite set S of states, $s = |S|$,
- an initial probability distribution (or short distribution) $\boldsymbol{q}^{(0)} = (q_j)_{1\leq j\leq s}$ and
- a stochastic *transition matrix* $\boldsymbol{P} = (p_{i,j})_{1\leq i,j\leq s}$.

For all $t \geq 0$, the distribution $\boldsymbol{q}^{(t)}$ after t rounds is equal to $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)} \cdot \boldsymbol{P}$. The following result can be shown for $\boldsymbol{q}^{(t)}$:

Theorem 13.1 *If $\boldsymbol{q}^{(0)}$ is a probability distribution and $\boldsymbol{P}$ a stochastic matrix, then also $\boldsymbol{q}^{(t)}$ for all $t \geq 0$ is a probability distribution.*

The proof is an assignment. Before we give more properties of Markov chains, we first introduce some notation and motivation.

13.1 Some elementary definitions

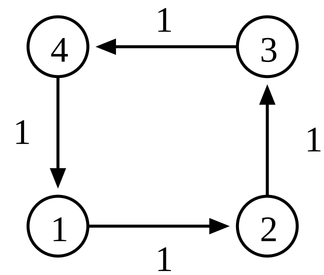
A Markov chain is called *periodic* with period $d > 1$ if the rows and columns of $\boldsymbol{P}$ can be permuted such that

$$\boldsymbol{P} = \begin{pmatrix} \mathbf{0} & \boldsymbol{A}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{A}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{A}_3 & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{A}_d & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix}$$

where the $\boldsymbol{A}_i$'s are arbitrary stochastic matrices and for any column j that belongs to matrix $\boldsymbol{A}_k$, row j belongs to matrix $\boldsymbol{A}_{k+1}$. For an example see Figure 2.

A Markov chain is called *irreducible* if there is a $t \geq 0$ so that $\boldsymbol{P}^t > \mathbf{0}$ (i.e. for $\boldsymbol{P}^t = (p_{i,j}^{(t)})$, $p_{i,j}^{(t)} > 0$ for all i,j). In this case we have the following result.

Theorem 13.2 *If $\boldsymbol{P}^t > \mathbf{0}$ for some $t \geq 0$, then for all pairs of states $x, y \in S$ the probability is larger than 0 to get in at most t steps from x to y.*



transition matrix:

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Figure 2: A Markov chain of period 4.

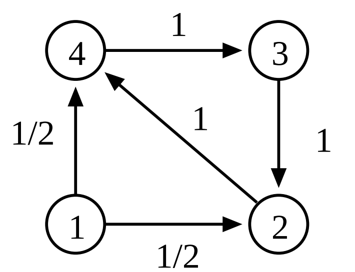


Figure 3: A reducible Markov chain.

The proof is an assignment. Let the *transition graph* of a Markov chain be defined as a directed graph $G = (S, E)$ with the property that $(x, y) \in E$ if and only if $p_{x,y} > 0$. Theorem 13.2 implies that the transition graph of an irreducible Markov chain is strongly connected. For an example of a reducible Markov chain see Figure 3.

An irreducible, aperiodic Markov chain is called *ergodic*. Ergodic Markov chains will be of particular importance, as we will see. First, some more properties of Markov chains:

Let $\boldsymbol{P}^t = (p_{i,j}^{(t)})_{1 \leq i,j \leq s}$. A state $i \in S$ is called *recurrent* or *transient*, depending on whether $\sum_{t \geq 1} p_{i,i}^{(t)}$ diverges or converges. For an example see Figure 3. It is not difficult to show that $\sum_{t \geq 1} p_{1,1}^{(t)} = 0$ and $\sum_{t \geq 1} p_{i,i}^{(t)} = \infty$ for all $i = 2, 3, 4$.

A state $i \in S$ is called *absorbing* if $p_{i,i} = 1$. If a Markov chain models a game, then absorbing states represent the success of player 1 or player 2.

Theorem 13.3 *Every Markov chain has at least one recurrent state. In an ergodic Markov chain all states are recurrent and non-absorbing.*

13.2 Applications

Suppose we are given an arbitrary combinatorial optimization problem over some finite set S of feasible solutions. In the case that we have an evolutionary algorithm that only uses mutation and local selection (such as done by the metropolis algorithm or simulated annealing method), and these two methods do not change over time, Markov chains can be used in two ways to study the performance of this kind of algorithms:

Case 1: The algorithm only acts on a population of size 1 (that is, only one solution is kept at any time). Then the parts of the Markov chain can be interpreted as follows:

- $\boldsymbol{q}^{(0)}$ represents the probability distribution for choosing the initial solution, and
- each $p_{i,j}$ in the matrix $\boldsymbol{P} = (p_{i,j})$ represents the probability of replacing solution i by solution j.

Furthermore, $\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(0)}\boldsymbol{P}^t$ represents the probability distribution for the solution after t rounds.

Case 2: The algorithm acts on a population of solutions of infinite size, and initially, every individual solution is selected independently at random according to the distribution $\boldsymbol{q}^{(0)}$. Then the parts of the Markov chain can be interpreted as follows:

- $q_i^{(0)}$ gives the initial percentage of the population that represent solution i, and
- each $p_{i,j}$ in the matrix $\boldsymbol{P} = (p_{i,j})$ denotes the percentage of the population representing solution i that will be moved to solution j.

Note that for the $p_{i,j}$ we can argue about fixed percentages rather then probability distributions around expectations, since as the size of the population goes to infinity, the deviation from the expected share of solutions moving from i to j goes to 0. Thus, $q_i^{(t)} = (\boldsymbol{q}^{(0)}\boldsymbol{P}^t)_i$ gives the percentage of the population after t rounds that represent solution i.

If one ensures that $\boldsymbol{P} = (p_{x,y})_{x,y \in S}$ has the properties that

- every state can be reached from every state in finite time and
- for every $x \in S$, the probability to remain at x in the next round is greater than 0,

then one can show that the Markov chain is ergodic, which usually significantly simplifies the analysis of the behavior of the Markov chain and therefore the algorithm represented by the Markov chain.

13.3 Attractors

A state distribution $\boldsymbol{\pi}$ of a Markov chain is called *stationary* if $\boldsymbol{\pi} = \boldsymbol{\pi} \cdot \boldsymbol{P}$. The following theorem contains a central property of Markov chains.

Theorem 13.4 *Every ergodic Markov chain has a unique stationary distribution $\boldsymbol{\pi}$, that is, for all initial distributions $\boldsymbol{q}^{(0)}$ the trajectory converges towards $\boldsymbol{\pi}$. Furthermore, $\boldsymbol{\Pi} = (\boldsymbol{\pi}, \ldots, \boldsymbol{\pi})^T$ is the matrix with*

$$\lim_{t \to \infty} \boldsymbol{P}^t = \boldsymbol{\Pi} .$$

Hence, if the Markov chain is ergodic, it will be much easier to determine for the underlying evolutionary algorithm whether or not it will get sufficiently close to an optimal solution.

13.4 Sensitivity to deviations

Suppose we have two distributions $\boldsymbol{q}^{(0)}$ and $\boldsymbol{r}^{(0)}$ with the property that $|\boldsymbol{q}^{(0)} - \boldsymbol{r}^{(0)}| = \sum_{j=1}^{s} |q_j^{(0)} - r_j^{(0)}| = \epsilon$. Then we want to determine $|\boldsymbol{q}^{(t)} - \boldsymbol{r}^{(t)}|$ for all $t \geq 1$. For this we define $\boldsymbol{\delta}^{(t)}$ so that $\boldsymbol{q}^{(t)} = \boldsymbol{r}^{(t)} + \boldsymbol{\delta}^{(t)}$ for all $t \geq 0$. Initially, $|\boldsymbol{\delta}^{(0)}| = \epsilon$. Furthermore,

$$\begin{aligned} \boldsymbol{q}^{(t+1)} &= \boldsymbol{q}^{(t)} \cdot \boldsymbol{P} = (\boldsymbol{r}^{(t)} + \boldsymbol{\delta}^{(t)})\boldsymbol{P} \\ &= \boldsymbol{r}^{(t+1)} + \boldsymbol{\delta}^{(t)} \cdot \boldsymbol{P} = \boldsymbol{r}^{(t+1)} + \boldsymbol{\delta}^{(t+1)} \ . \end{aligned}$$

Since $\boldsymbol{P}$ is a stochastic matrix, we have

$$\begin{aligned} |\boldsymbol{\delta}^{(t+1)}| &= \sum_{j=1}^{s} |\sum_{i=1}^{s} \delta_i^{(t)} \cdot p_{i,j}| \leq \sum_{j=1}^{s}\sum_{i=1}^{s} |\delta_i^{(t)}| \cdot p_{i,j}^{(t)} \\ &= \sum_{i=1}^{s}\sum_{j=1}^{s} |\delta_i^{(t)}| \cdot p_{i,j}^{(t)} = \sum_{i=1}^{s} |\delta_i^{(t)}| = |\boldsymbol{\delta}^{(t)}| \ . \end{aligned}$$

Thus, $|\boldsymbol{q}^{(t)} - \boldsymbol{r}^{(t)}| \leq \epsilon$ for all $t \geq 1$. Moreover, it follows from our computation that deviations never increase, but will rather decrease.

13.5 Speed of Convergence

We describe now one of many possibilities to bound the speed of convergence towards the stationary distribution. Let the *relative pointwise distance* to the stationary distribution be defined as

$$\Delta_t = \max_{i,j \in S} \frac{|p_{i,j}^{(t)} - \pi_j|}{\pi_j} \ .$$

An ergodic Markov chain with stationary distribution $\boldsymbol{\pi}$ is called *reversible* if

$$p_{i,j} \cdot \pi_i = p_{j,i} \cdot \pi_j$$

for all i, j. Let $\lambda_1, \ldots, \lambda_n$ be all eigenvalues of $\boldsymbol{P}$, that is, all solutions to $\det(\lambda \boldsymbol{I}_n - \boldsymbol{P}) = 0$ in the order of descending values. Then it holds for all reversible Markov chains that

$$\Delta_t \leq \frac{\lambda_2^t}{\min_{j \in S} \pi_j} \ .$$

This gives an upper bound on how close $\boldsymbol{P}^t$ is to $\boldsymbol{\Pi}$. Since furthermore $\boldsymbol{q} \cdot \boldsymbol{\Pi} = \boldsymbol{\pi}$ for all probability distributions $\boldsymbol{q}$, we can determine how close the system is to the stationary distribution after t rounds.

14 Quadratic Dynamical Systems

We restrict ourselves in this section to quadratic dynamical systems (or short QDS) with a transition function f of the form

$$f(\boldsymbol{q}) = \boldsymbol{q} \times \boldsymbol{q} \ ,$$

where the *mating operator* "$\times$" over a matrix $\boldsymbol{B} = (b_{(i,j),(k,\ell)})_{(i,j),(k,\ell)\in S^2}$ is defined as

$$(\boldsymbol{q} \times \boldsymbol{q})_\ell = \sum_{i,j,k} q_i \cdot q_j \cdot b_{(i,j),(k,\ell)}$$

for all $\ell \in S$.

The matrix $\boldsymbol{B} = (b_{(i,j),(k,\ell)})$ is *stochastic* if $b_{(i,j),(k,\ell)} \in [0,1]$ for all i,j,k,ℓ and

$$\sum_{(k,\ell)\in S^2} b_{(i,j),(k,\ell)} = 1$$

for all $(i,j) \in S^2$.

If the initial vector $\boldsymbol{q}$ represents a probability distribution and $\boldsymbol{B}$ is a stochastic matrix, then a QDS is called a *stochastic QDS*. If the matrix $\boldsymbol{B}$ is stochastic, we will denote it in the following by $\boldsymbol{P} = (p_{(i,j),(k,\ell)})$.

To sum up, a stochastic QDS consists of the following parts:

- a finite set S of states, $s = |S|$,
- an inital probability distribution (or short distribution) $\boldsymbol{q}^{(0)} = (q_j)_{1\leq j\leq s}$ and
- a stochastic *transition matrix* $\boldsymbol{P} = (p_{(i,j),(k,\ell)})_{1\leq i,j,k,\ell\leq s}$.

The distribution $\boldsymbol{q}^{(t)}$ after t rounds results from the formula $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)} \times \boldsymbol{q}^{(t)}$ for all $t \geq 0$. For $\boldsymbol{q}^{(t)}$ it holds analogous to Section 13:

Theorem 14.1 *If $\boldsymbol{q}^{(0)}$ is a probability distribution and $\boldsymbol{P}$ is a stochastic matrix, then also $\boldsymbol{q}^{(t)}$ for all $t \geq 0$ is a probability distribution.*

The proof is an assignment. Before we deal closer with properties of stochastic QDS's, we introduce some elementary definitions.

14.1 Some elementary definitions

A stochastic QDS resp. a stochastic matrix is called *symmetric* if for all i,j,k,ℓ it holds that

$$p_{(i,j),(k,\ell)} = p_{(j,i),(k,\ell)} = p_{(i,j),(\ell,k)} \ .$$

Furthermore, a stochastic QDS resp. a stochastic matrix is called *locally reversible* if for all i,j,k,ℓ it holds that

$$p_{(i,j),(k,\ell)} = p_{(k,\ell),(i,j)} \ .$$

A stochastic QDS resp. stochastic matrix is called *aperiodic* if $p_{(i,j),(i,j)} > 0$ for all i,j.

A trajectory $\boldsymbol{q}^{(0)}, \boldsymbol{q}^{(1)}, \ldots$ is called *irreducible* if there is a $\delta > 0$ (that may depend on $\boldsymbol{q}^{(0)}$) so that for a sufficiently large t, $q_i^{(t)} \geq \delta$ for all i.

14.2 Applications

Suppose again that we are given an arbitrary combinatorial optimization problem over some set S of feasible solutions. In the case that we have an evolutionary algorithm that uses (at most) pairwise recombination, mutation, local selection, and uniform mating (explained below), and these methods do not change over time, a quadratic dynamical systems can be used to study the performance of this kind of algorithm when acting on an infinitely large population of solutions. Candidates for such an algorithm are evolution strategies and genetic algorithms. The parts of the QDS can be then be interpreted as follows:

- $q_i^{(0)}$ gives the initial percentage of the population that represent solution i,
- each $p_{(i,j),(k,\ell)}$ in the matrix $\boldsymbol{P} = (p_{(i,j),(k,\ell)})$ denotes the percentage of the pairs (i,j) that are transformed into pairs (k,ℓ),
- and the operation
$$q_\ell^{(t+1)} = \sum_{i,j,k} q_i^{(t)} \cdot q_j^{(t)} \cdot p_{(i,j),(k,\ell)}$$
represents the percentage of solutions of type ℓ after $t+1$ rounds if for every (i,j), the percentage of solutions of type i mating with solutions of type j is $q_i^{(t)} \cdot q_j^{(t)}$ (i.e. each solution picks its mating partner uniformly with at random, or in other words, we have a uniform mating process).

Furthermore, $q_i^{(t)}$ gives the percentage of the population after t rounds that represent solution i.

14.3 Attractors

A state distribution $\boldsymbol{\pi}$ of a stochastic QDS is called a *stationary distribution* if $\boldsymbol{\pi} = \boldsymbol{\pi} \times \boldsymbol{\pi}$. The following theorem contains a central property of stochastic QDS's.

Theorem 14.2 *For every symmetric, locally reversible and aperiodic stochastic matrix and every initial distribution $\boldsymbol{q}^{(0)}$ whose trajectory is irreducible it holds that $\boldsymbol{q}^{(t)}$ converges towards a stationary distribution $\boldsymbol{\pi}$.*

The proof of the theorem can be done by showing that the entropy

$$H(\boldsymbol{q}) = -\sum_{i=1}^{s} q_i \log q_i$$

for all non-stationary distributions $\boldsymbol{q}$ monotonically increases when using a symmetric, locally reversible and aperiodic matrix. That is,

$$H(\boldsymbol{q} \times \boldsymbol{q}) > H(\boldsymbol{q}) \ .$$

By this it can be shown that non-reducible trajectories end in a point of (locally) maximum entropy. For more information see the paper by Rabinovich et al. The symmetry property in Theorem 14.2 is not only sufficient, but also necessary, since it is known (see Rabinovich et al) that asymmetric QDS's can not only have points but also cycles and more complex objects as attractors. It is important to note here that a symmetric, locally reversible and aperiodic QDS can have in contrast to an ergodic Markov chain many different stationary distributions. Hence, two different initial distributions may not converge to the same stationary distribution. These stationary distributions can be precisely characterized.

Theorem 14.3 *A distribution* $\boldsymbol{q}$ *is* stationary *for a symmetric, locally reversible and aperiodic QDS if and only if*

$$q_i \cdot q_j = q_k \cdot q_\ell$$

for all i, j, k, ℓ *with* $p_{(i,j),(k,\ell)} > 0$.

Furthermore, methods have been developed that can determine whether two initial distributions will converge towards the same stationary distribution (see Rabinovich et al).

14.4 Sensitivity to deviations

Suppose we have two distributions $\boldsymbol{q}^{(0)}$ and $\boldsymbol{r}^{(0)}$ with the property that $|\boldsymbol{q}^{(0)} - \boldsymbol{r}^{(0)}| = \sum_{j=1}^{s} |q_j^{(0)} - r_j^{(0)}| = \epsilon$. Then we will determine $|\boldsymbol{q}^{(t)} - \boldsymbol{r}^{(t)}|$ for all $t \geq 1$. For this we define $\boldsymbol{\delta}^{(t)}$ so that $\boldsymbol{q}^{(t)} = \boldsymbol{r}^{(t)} + \boldsymbol{\delta}^{(t)}$ for all $t \geq 0$. Initially, it holds that $|\boldsymbol{\delta}^{(0)}| = \epsilon$. Furthermore,

$$\begin{aligned}
\boldsymbol{q}^{(t+1)} &= \boldsymbol{q}^{(t)} \times \boldsymbol{q}^{(t)} = (\boldsymbol{r}^{(t)} + \boldsymbol{\delta}^{(t)}) \times (\boldsymbol{r}^{(t)} + \boldsymbol{\delta}^{(t)}) \\
&= \boldsymbol{r}^{(t)} \times \boldsymbol{r}^{(t)} + \boldsymbol{r}^{(t)} \times \boldsymbol{\delta}^{(t)} + \boldsymbol{\delta}^{(t)} \times \boldsymbol{r}^{(t)} + \boldsymbol{\delta}^{(t)} \times \boldsymbol{\delta}^{(t)} \\
&= \boldsymbol{r}^{(t)} \times \boldsymbol{r}^{(t)} + \boldsymbol{\delta}^{(t)} \times (\boldsymbol{r}^{(t)} + (\boldsymbol{r}^{(t)} + \boldsymbol{\delta}^{(t)})) \\
&= \boldsymbol{r}^{(t)} \times \boldsymbol{r}^{(t)} + \boldsymbol{\delta}^{(t)} \times (\boldsymbol{r}^{(t)} + \boldsymbol{q}^{(t)}) = \boldsymbol{r}^{(t+1)} + \boldsymbol{\delta}^{(t+1)} \,.
\end{aligned}$$

Since $\boldsymbol{P}$ is a stochastic matrix and $\boldsymbol{q}^{(t)}$ and $\boldsymbol{r}^{(t)}$ are probability distributions, it holds that

$$\begin{aligned}
|\boldsymbol{\delta}^{(t+1)}| &= \sum_{\ell=1}^{s} \Big| \sum_{1 \leq i,j,k \leq s} \delta_i^{(t)} (\boldsymbol{r}^{(t)} + \boldsymbol{q}^{(t)})_j \cdot p_{(i,j),(k,\ell)} \Big| \\
&\leq \sum_{\ell=1}^{s} \sum_{1 \leq i,j,k \leq s} |\delta_i^{(t)}| \cdot (\boldsymbol{r}^{(t)} + \boldsymbol{q}^{(t)})_j \cdot p_{(i,j),(k,\ell)} \\
&= \sum_{1 \leq i,j \leq s} |\delta_i^{(t)}| \cdot (\boldsymbol{r}^{(t)} + \boldsymbol{q}^{(t)})_j \sum_{1 \leq k,\ell \leq s} p_{(i,j),(k,\ell)} \\
&= \sum_{1 \leq i,j \leq s} |\delta_i^{(t)}| \cdot (\boldsymbol{r}^{(t)} + \boldsymbol{q}^{(t)})_j \\
&= \sum_{i=1}^{s} |\delta_i^{(t)}| \cdot 2 = 2 \cdot |\boldsymbol{\delta}^{(t)}|
\end{aligned}$$

Hence, $|\boldsymbol{q}^{(t)} - \boldsymbol{r}^{(t)}| \leq 2^t \cdot \epsilon$ for all $t \geq 1$. As we will see, this deviation bound is in principle best possible, that is, deviations can have an exponential influence on later distributions. Stochastic QDS's can therefore be in contrast to Markov chains very sensitive to deviations.

14.5 Speed of Convergence

So far, no general results have been obtained concerning the speed of convergence for stochastic QDS's. However, some special cases have been solved successfully (see, for example, Rabinovich et al).

14.6 References

- D. Aldous and J. Fill. Reversible Markov Chains and Random Walks on Graphs. Monograph in preparation. See http://stat-www.berkeley.edu/users/aldous/book.html

- T. Bäck. *Evolutionary Algorithms in Theory and Practice.* Oxford University Press, 1996.

- P.J.M. van Laarhoven and E.H.L. Aarts. *Simulated Annealing: Theory and Applications.* Mathematics and its Applications. Kluwer, 1987.

- Y. Rabinovich, A. Sinclair, and A. Wigderson. Quadratic dynamical systems. In *Proc. of the 33rd Symp. on Foundations of Computer Science*, pp. 304–313, 1992.

- A. Sinclair. Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combinatorics, Probability and Computing* 1:351–370, 1992.

- A. Sinclair and M. Jerrum. Approximate counting, unform generation and rapidly mixing Markov chains. Information & Computation 82:93-133, 1989.