Mathematical Analysis of Evolutionary Algorithms for Optimization

Heinz Mühlenbein
GMD – Schloss Birlinghoven
53754 Sankt Augustin, Germany
muehlen@gmd.de

Thilo Mahnig
GMD – Schloss Birlinghoven
53754 Sankt Augustin, Germany
mahnig@gmd.de

Abstract

Simulating evolution as seen in nature has been identified as one of the key computing paradigms for the new decade. Today evolutionary algorithms have been successfully used in a number of applications. These include discrete and continuous optimization problems, synthesis of neural networks, synthesis of computer programs from examples (also called genetic programming) and even evolvable hardware. But in all application areas problems have been encountered where evolutionary algorithms performed badly. In this survey we concentrate on the analysis of evolutionary algorithms for optimization. We present a mathematical theory based on probability distributions. It gives the reasons why evolutionary algorithms can solve many difficult multi-modal functions and why they fail on seemingly simple ones. The theory also leads to new sophisticated algorithms for which convergence is shown.

1 Introduction

We first introduce the most popular algorithm, the *simple genetic algorithm*. This algorithm has many degrees of freedom, especially in the recombination scheme used. We show that all genetic algorithms behave very similar, if recombination is done without selection a sufficient number of times before the next selection step. We correct the classical schema analysis of genetic algorithm. We show why the usual schema theorem folklore is mathematically wrong. We approximate genetic algorithms by a conceptual algorithm. This algorithm we call the *Univariate Marginal Distribution Algorithm* $UMDA$, which is analyzed in Section 3. We compute the difference equation for the univariate marginal distributions under the assumption of proportionate selection. This equation has been proposed in populations genetics by Sewall Wright as early as 1937 [Wri70]. This is an independent confirmation of our claim that $UMDA$ approximates any genetic algorithm. Using *Wright's equation* we show that $UMDA$ solves a *continuous optimization problem*. The function to be optimized is given by the average fitness of the population.

Proportionate selection is far too weak for optimization. This has been recognized very early in breeding of livestock. *Artificial selection* as done by breeders is a much better model for optimization than *natural selection* modelled by proportionate selection. Unfortunately an exact mathematical analysis of efficient artificial selection schemes seems impossible. Therefore breeders have developed an approximate theory, using the concepts of regression of offspring to parent, heritability and response to selection. This theory is discussed in Section 4. At the end of the section numerical results are shown which show the strength and the weakness of $UMDA$ as a numerical optimization method.

$UMDA$ optimizes very efficient some difficult optimization problems, but it fails on some simple problems. For these problems higher order marginal distributions are necessary which capture the nonlinear dependency between variables. In Section 5.2 $UMDA$ is extended to the *Factorized Distribution Algorithm FDA.* We prove convergence of the algorithm to the global optima if *Boltzmann selection* is used. The theory of factorization connects FDA with the theory of *graphical models* and *Bayesian networks.* We derive a new adaptive Boltzmann selection schedule SDS using ideas from the science of breeding.

In Section 6.1 we use results from the theory of Bayesian networks for the *Learning Factorized Distribution Algorithm* $LFDA$, which learns a factorization from the data. We make a preliminary comparison between the efficiency of FDA and $LFDA$.

In Section 7 we describe the *system dynamics approach to optimization*. The difference equations obtained for $UMDA$ are iterated until convergence. Thus the continuous optimization problem is mathematically solved without using a population of points at all. We present numerical results for three different system dynamics equations. They consists of Wright's equation, the *diversified replicator equation* and a modified version of Wright's equation which converges faster.

In the final section we classify the different evolutionary computation methods presented. The classification criterion is whether a microscopic or a macroscopic model is used for selection and/or recombination.

2 Analysis of the Simple Genetic Algorithm

In this section we investigate the standard genetic algorithm, also called the Simple Genetic Algorithm (SGA). The algorithm is described by Holland [Hol92] and Goldberg [Gol89]. It consists of

- fitness proportionate selection
- recombination/crossover
- mutation

Here we will analyze selection and recombination only. Mutation is considered to be a background operator. It can be analyzed by known techniques from stochastics [MSV94, Müh97].

There have been many claims concerning the optimization power of SGA. Most of them are based on a rather qualitative application of the *schema theorem*. We will show the shortcomings of this approach. Our analysis is based on techniques used in population genetics. The analysis reveals that an exact mathematical analysis of SGA is possible for small problems only. For a binary problem of size n the exact analysis needs the computation of 2^n equations. But we propose an approximation often used in population genetics. The approximation assumes that the gene frequencies are in *linkage equilibrium*. The main result is that *any genetic algorithm can be approximated by an algorithm using n parameters only, the univariate marginal gene frequencies.*

2.1 Definitions

Let $\mathbf{x} = (x_1, \ldots, x_n)$ denote a binary vector. For notational simplicity we restrict the discussion to binary variables $x_i \in \{0, 1\}$. We use the following conventions. Capital letters X_i denote variables, small letters x_i assignments.

Definition 2.1. *Let a function $f : \boldsymbol{X} \to R^{\geq 0}$ be given. We consider the optimization problem*

$$\mathbf{x}_{opt} = \text{argmax}\, f(\mathbf{x}) \tag{2.1}$$

We will use $f(\mathbf{x})$ as the fitness function for the SGA. We will investigate two widely used recombination/crossover schemes.

Definition 2.2. *Let two strings* $\mathbf{x}$ *and* $\mathbf{y}$ *be given. In* one-point crossover *the string* $\mathbf{z}$ *is created by randomly choosing a crossover point* $0 < l < n$ *and setting* $z_i = x_i$ *for* $i \leq l$ *and* $z_i = y_i$ *for* $i > l$. *In* uniform crossover z_i *is randomly chosen with equal probability from* $\{x_i, y_i\}$.

Definition 2.3. *Let* $p(\mathbf{x}, t)$ *denote the probability of* $\mathbf{x}$ *in the population at generation* t. *Then* $p_i(x_i, t) = \sum_{\mathbf{x}, X_i = x_i} p(\mathbf{x}, t)$ *defines a univariate marginal distribution.*

We often write $p_i(x_i)$ if just one generation is discussed. In this notation the average fitness of the population and the variance is given by

$$\begin{aligned} \bar{f}(t) &= \sum_x p(\mathbf{x}, t) f(\mathbf{x}) \\ V(t) &= \sum_x p(\mathbf{x}, t) \left(f(\mathbf{x}) - \bar{f}(t)\right)^2 \end{aligned}$$

The *response to selection* $R(t)$ is defined by

$$R(t) = \bar{f}(t+1) - \bar{f}(t) \tag{2.2}$$

2.2 Proportionate Selection

Proportionate selection changes the probabilities according to

$$p(\mathbf{x}, t+1) = p(\mathbf{x}, t) \frac{f(x)}{\bar{f}(t)} \tag{2.3}$$

Lemma 2.1. *For proportionate selection the response is given by*

$$R(t) = \frac{V(t)}{\bar{f}(t)} \tag{2.4}$$

Proof: We have

$$R(t) = \sum_x p(x, t) \frac{f^2(x)}{\bar{f}(t)} - \bar{f}(t) = \frac{V(t)}{\bar{f}(t)} \tag{2.5}$$

□

With proportionate selection the average fitness never decreases. This is true for every rational selection scheme.

2.3 Recombination

For the analysis of recombination we introduce a special distribution.

Definition 2.4. *Robbins' proportions are given by the distribution* π

$$\pi(x, t) := \prod_{i=1}^{n} p_i(x_i, t) \tag{2.6}$$

A population in Robbins' proportions is also called to be in **linkage equilibrium**.

Geiringer [Gei44] has shown that all reasonable recombination schemes lead to the same limit distribution.

Theorem 2.1 (Geiringer). *Recombination does not change the univariate marginal frequencies, i.e.* $p_i(x_i, t+1) = p_i(x_i, t)$. *The limit distribution of any complete recombination scheme is Robbins' proportions* $\pi(\mathbf{x})$.

Complete recombination means that for each subset S of $\{1, \ldots, n\}$, the probability of an exchange of genes by recombination is greater than zero. Convergence to the limit distribution is very fast. We have to mention an important fact. In a finite population linkage equilibrium cannot be exactly achieved. We take the uniform distribution as example. Here linkage equilibrium is given by $p(\mathbf{x}) = 2^{-n}$. This value can only be obtained if the size of the population N is substantial larger than 2^n! For a population of $N = 1000$ the minimum deviation DSQ_{min} from Robbins' proportions is already achieved after four generations, then DSQ slowly increases due to stochastic fluctuations by *genetic drift*. Ultimately the population will consist of one genotype only. Genetic drift has been analyzed by Asoh and & Mühlenbein [AM94b]. It will not be considered here.

2.4 Selection and Recombination

We have shown that the average $\bar{f}(t)$ never decreases after selection and that any complete recombination scheme moves the genetic population to Robbins' proportions. Now the question arises: What happens if recombination is applied *after* selection. The answer is very difficult. The problem still puzzles populations genetics [Nag92].

Formally the difference equations can be easily written. Let a recombination distribution R be given. $R_{x,yz}$ denotes the probability that y and z produce x after recombination. Then

$$p(\mathbf{x}, t+1) = \sum_{y,z} R_{x,yz} p^s(\mathbf{y}) p^s(\mathbf{z}) \tag{2.7}$$

$p^s(x)$ denotes the probability of string x after selection. For n loci the recombination distribution R consists of $2^n * 2^n$ parameters. Recently Christiansen and Feldman [CF98] have written a survey about the mathematics of selection and recombination from the viewpoint of population genetics. A new technique to obtain the equations has been developed by Vose [Vos99]. In both frameworks one needs a computer program to compute the equations for a given fitness function.

A mathematical analysis of the mathematical properties of n loci systems is difficult. For a problem of size n we have 2^n equations. Furthermore the equations depend on the recombination operator used! If the gene frequencies remain in linkage equilibrium, then only n equations are needed for the marginal frequencies. Thus the crucial question is: Does the optimization process gets worse because of this simplification? The answer is no. We provide evidence for this statement by citing a theorem from [Müh97]. It shows that the univariate marginal frequencies are the same for all recombination schemes if applied to the same distribution $p(\mathbf{x}, t)$.

Theorem 2.2. *For* any *complete recombination/crossover scheme used after proportionate selection the univariate marginal frequencies are determined by*

$$p(x_i, t+1) = \sum_{\mathbf{x}|X_i=x_i} \frac{p(\mathbf{x}, t) f(\mathbf{x})}{\bar{f}(t)}. \tag{2.8}$$

Proof: After selection the univariate marginal frequencies are given by

$$p^s(x_i, t) = \sum_{\mathbf{x}|X_i=x_i} p^s(\mathbf{x}, t) = \sum_{\mathbf{x}|X_i=x_i} \frac{p(\mathbf{x}, t) f(\mathbf{x})}{\bar{f}(t)}.$$

Now the selected individuals are randomly paired. Therefore

$$p_i(x_i, t+1) = p_i^s(x_i, t).$$

2.5 Schema Analysis Demystified

Theorem 2.2 can be formulated in the terms of Holland's schema theory [Hol92]. Let $H(x_i) = (*, \ldots, *, x_i, *, \ldots, *)$ be a first-order schema at locus i. This schema includes all strings where the gene at locus i is fixed to x_i. The univariate marginal frequency $p(x_i, t)$ is obviously identical to the frequency of schema $H(x_i)$. The fitness of the schema at generation t is given by

$$f(H(x_i), t) = \frac{1}{p_i(x_i, t)} \sum_{\mathbf{x}|X_i=x_i} p(\mathbf{x}, t) f(\mathbf{x}) \tag{2.9}$$

From Theorem 2.2 we obtain:

Corollary 2.1 (First-order schema theorem). *For a genetic algorithm with proportionate selection using any complete recombination the frequency of first-order schemata changes according to*

$$p_i(x_i, t+1) = p_i(x_i, t) \frac{f(H(x_i), t)}{\bar{f}(t).} \tag{2.10}$$

We now extend the analysis to general schemata.

Definition 2.5. *Let* $\mathbf{x}_s = (x_{s_1}, \ldots, x_{s_i}) \subset \{x_1, \ldots, x_n\}$. *Thus* $\mathbf{x}_s$ *denotes a subvector of* $\mathbf{x}$ *defined by the indices* $s_1, \ldots, s_i$. *Then the probability of schema* $H(s)$ *is defined by*

$$p(H(s), t) = \sum_{X|X_s=x_s} p(\mathbf{x}, t) \tag{2.11}$$

The summation is done by fixing the values of $\mathbf{x}_s$. Thus the probability of a schema is just the corresponding marginal distribution $p(\mathbf{x}_s)$. If $\mathbf{x}_s$ consists of a single element only, we have a univariate marginal distribution.

SGA uses fitness proportionate selection, i.e. the probability of $\mathbf{x}$ being selected is given by

$$p^s(\mathbf{x}, t) = p(\mathbf{x}, t) \frac{f(\mathbf{x})}{\bar{f}(t)} \tag{2.12}$$

$\bar{f}(t) = \sum_x p(\mathbf{x}, t) f(\mathbf{x})$ is the average fitness of the population. Let us now assume that we have an algorithm which generates new points according to the distribution of selected points, i.e.

$$p(\mathbf{x}, t+1) = p(\mathbf{x}, t) \frac{f(\mathbf{x})}{\bar{f}(t)} \tag{2.13}$$

We will later address the problem which probability distribution SGA really generates.

Definition 2.6. *The fitness of schema* $H(s_i)$ *is defined by*

$$f(H(s_i), t) = \sum_{X|X_s=x_s} \frac{p(\mathbf{x}, t)}{p(H(s), t)} f(\mathbf{x}) \tag{2.14}$$

Theorem 2.3 (Schema Theorem). *The probability of schema* $H(s)$ *is given by*

$$p(H(s), t+1) = p(H(s, t)) \frac{f(H(s), t)}{\bar{f}(t)} \tag{2.15}$$

Holland ([Hol92] Theorem 6.2.3) computed for SGA an inequality

$$p(H(\boldsymbol{s}), t+1) \geq (1-\delta) p(H(\boldsymbol{s}), t) \frac{f(H(\boldsymbol{s}), t)}{\bar{f}(t)} \tag{2.16}$$

δ is a small factor. The inequality complicates the application. But a careful investigation of Holland's analysis [Hol92] reveals the fact, that the application is much simpler if equation (2.14) is used instead of the inequality. Thus equation (2.14 is obviously the *ideal starting point* for Holland's schema analysis.

2.6 Schema Analysis Folklore

The mathematical difficulty of using the inequality (2.16) to estimate the distribution of schemata lies in the fact that the fitness of a schema depends on $p(\mathbf{x}, t)$, i.e the distribution of the genotypes of the population. This is a defining fact of Darwinian *natural selection*. The fitness is always relative to the current population. To cite a proverb: *the one-eyed is the king of the blinds.*

Thus an application of the inequality (2.16) is not possible without computing $p(\mathbf{x}, t)$. Goldberg [Gol89] circumvented this problem by assuming

$$p(H(\boldsymbol{s}), t) \geq (1+c)\bar{f}(t) \tag{2.17}$$

With this assumption we estimate $p(H(\boldsymbol{s}), t) \geq (1+c)^t p(H(\boldsymbol{s}), 0)$. But the assumption can never be fulfilled for all t. When approaching an optimum, the fitness of all schemata in the population will be only $1 \pm \epsilon$ away from the average fitness. Here proportionate selection gets difficulties.

The typical folklore which arose from the schema analysis is nicely summarized by Ballard. He is not biased towards or against genetic algorithms. He just cites the commonly used arguments ([Bal97], p.270).

- *Short schemata have a high probability of surviving the genetic operations.*
- *Focusing on short schemata that compete shows that, over the short run, the fittest are increasing at an exponential rate.*
- *Ergo, if all of the assumptions hold (we cannot tell whether they do, but we suspect they do), GAs are optimal.*

We will not investigate the optimality argument, because we will show that the basic conclusion of exponential increasing schemata does not hold.

2.7 Correct Schema Analysis

It turns out that equation 2.13 for proportionate selection admits an analytical solution.

Theorem 2.4 (Convergence). *The distribution $p(\mathbf{x}, t)$ for proportionate selection is given by*

$$p(\mathbf{x}, t) = \frac{p(\mathbf{x}, 0) f(\mathbf{x})^t}{\sum_y p(\mathbf{y}, 0) f(\mathbf{y})^t} \tag{2.18}$$

Let $\mathcal{M}$ be the set of global optima, then

$$\lim_{t \to \infty} p(\mathbf{x}, t) = \begin{cases} 1/|\mathcal{M}| & \mathbf{x} \in \mathcal{M} \\ 0 & \textit{else} \end{cases} \tag{2.19}$$

Proof: The proof is by induction. The assumption is fulfilled for $t = 1$. Then

$$\begin{aligned} p(\mathbf{x}, t+1) &= \frac{p(\mathbf{x}, 0) f(\mathbf{x})^{t+1}}{\sum_y p(\mathbf{y}, 0) f(\mathbf{y})^{t+1}} \\ &= \frac{p(\mathbf{x}, 0) f(\mathbf{x})^t}{\bar{f}(t)} \frac{f(\mathbf{x})}{\sum_y \frac{p(\mathbf{y}, 0) f(\mathbf{y})^t \cdot f(\mathbf{y})}{\bar{f}(t)}} \\ &= \frac{p(\mathbf{x}, 0) f(\mathbf{x})^{t+1}}{\sum_y p(\mathbf{y}, 0) f(\mathbf{y})^{t+1}} \end{aligned}$$

Let $\mathbf{x}_{max} \in \mathcal{M}$ and $f(\mathbf{x}) < f(\mathbf{x}_{max})$. Then

$$\frac{p(\mathbf{x}, t)}{p(\mathbf{x}_{max}, t)} = \frac{p(\mathbf{x}, 0) f(\mathbf{x})^t}{p(\mathbf{x}_{max}, 0) f(\mathbf{x}_{max})^t} \to 0$$

QED.

This shows that our algorithm is ideal in the sense that it even converges to the set of global optima.

2.8 Application of the Schema Equation

By using equation (2.18) we can make a correct schema analysis. We compute the probabilities of all schemata. We just discuss the interesting case of a *deceptive function*. We take the 3-bit deceptive function defined by

$$\begin{aligned} \mathit{decep}(\mathbf{x}) = {} & 0.9 - 0.1(x_1 + x_2 + x_3) \\ & - 0.7(x_1 x_2 + x_2 x_3 + x_1 x_3) + 2.5 x_1 x_2 x_3 \end{aligned}$$

The function is called deceptive because the global optimum $(1, 1, 1)$ is isolated, whereas the local optimum $(0, 0, 0)$ is surrounded by strings of high fitness. We now look at the behavior of some schemata.

Definition 2.7. *A schema is called optimal if its defining string s is contained in an optimal string.*

In our example $H(X_1 = 1)$ and $H(X_1 = X_2 = 1)$ are optimal schemata. They are displayed in Figure 1. We see that the probability of the optimal schema $p(H(X_1 = 1)$ decreases for about 8 generations, then it increases fairly slowly. This behavior is contrary to the simple interpretation of the evolution of schemata. Schema $H(X_1 = X_2 = 1)$ decreases even dramatically at the first generation. Then its probability is almost identical to the probability of the optimum $(1, 1, 1)$.

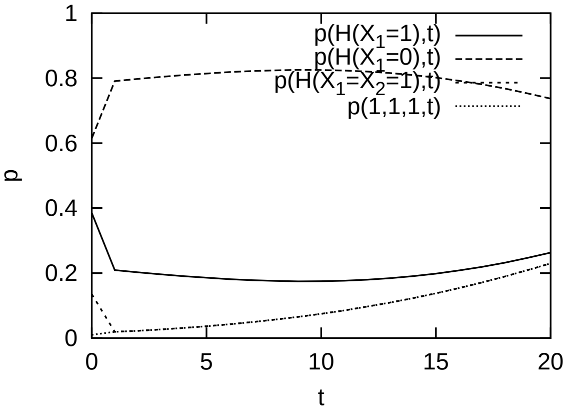

Figure 1: Evolution of some schemata

Remark: *Even the probability of optimal schemata can decrease for a long time. It depends on the probability distribution, if the optimal schemata will increase. Any argument about an exponential increase of optimal schemata is mathematically wrong.*

3 The Univariate Marginal Distribution Algorithm $UMDA$

The univariate marginal distribution algorithm $UMDA$ generates new points according to $p(\mathbf{x},t) = \prod_{i=1}^{n} p_i^s(x_i,t)$. Thus $UMDA$ keeps the gene frequencies in linkage equilibrium. This makes a mathematical analysis possible. We derive a difference equation for proportionate selection. This equation has already been proposed by Sewall Wright in 1937 [Wri70]. Wright's equation shows that $UMDA$ is trying to solve a continuous optimization problem. The continuous function to be optimized is the average fitness of the population $W(\mathbf{p})$. The variables are the univariate marginal distributions. In a fundamental theorem we show the relation between the attractors of the continuous problem and the local optima of the fitness function $f(\mathbf{x})$.

3.1 Definition of $UMDA$

Instead of performing recombination a number of times in order to converge to linkage equilibrium, one can achieve this in one step by *gene pool recombination* [MV96]. In gene pool recombination a new string is computed by randomly taking for each loci a gene from the distribution of the selected parents. This means that gene x_i occurs with probability $p_i^s(x_i)$ in the next population. $p_i^s(x_i)$ is the distribution of x_i in the selected parents. New strings $\mathbf{x}$ are generated according to the distribution

$$p(\mathbf{x},t+1) = \prod_{i=1}^{n} p_i^s(x_i,t) \tag{3.1}$$

One can simplify the algorithm still more by directly computing the univariate marginal frequencies from the data. Then Equation 3.1 can be used to generate new strings. This method is used by $UMDA$.

UMDA

- **STEP 0:** Set $t \Leftarrow 1$. Generate $N \gg 0$ points randomly.
- **STEP 1:** Select $M \leq N$ points according to a selection method. Compute the marginal frequencies $p_i^s(x_i,t)$ of the selected set.
- **STEP 2:** Generate N new points according to the distribution $p(\mathbf{x},t+1) = \prod_{i=1}^{n} p_i^s(x_i,t)$. Set $t \Leftarrow t+1$.
- **STEP 3:** If termination criteria are not met, go to STEP 1.

For proportionate selection we need the average fitness of the population $\bar{f}(t)$. We consider $\bar{f}(t)$ as a function which depends on $p(x_i)$. To emphasize this dependency we write

$$W(p_1(X_1=0), p(X_1=1), \ldots, p_n(X_n=1)) := \bar{f}(t) \tag{3.2}$$

W formally depends on $2n$ parameters. $p_i(X_i = 1)$ and $p_i(X_i = 0)$ are considered as two independent parameters despite the constraint $p_i(X_i = 0) = 1 - p_i(X_i = 1)$. We abbreviate $p_i := p_i(X_i = 1)$. If we insert $1 - p_i$ for $p_i(X_i = 0)$ into W, we obtain $\tilde{W}$. $\tilde{W}$ depends on n parameters. Now we can formulate the main theorem.

Theorem 3.1. *For infinite populations and proportionate selection the difference equations for the gene frequencies used by UMDA are given by*

$$p_i(x_i,t+1) = p_i(x_i,t)\frac{\bar{f}_i(x_i,t)}{W(t)} = p_i(x_i,t)\frac{\frac{\partial W}{\partial p_i(x_i)}}{W(t)} \tag{3.3}$$

where $\bar{f}_i(x_i,t) = \sum_{\mathbf{x},X_i=x_i} f(\mathbf{x}) \prod_{j\neq i}^{n} p(x_j,t)$. *The equation can also be written as*

$$p_i(t+1) = p_i(t) + p_i(t)(1-p_i(t))\frac{\frac{\partial \tilde{W}}{\partial p_i}}{\tilde{W}(t)} \tag{3.4}$$

The reponse is approximately given by

$$R(t) = \frac{V_A(t)}{\tilde{W}} + \frac{1}{2}\sum_{i\neq j}\frac{\alpha_i * \alpha_j}{\tilde{W}^2}\frac{\partial^2 W}{\partial p_i \partial p_j} + ... \tag{3.5}$$

$$VA(t) = \sum_i p_i(1,t)(f_i(1,t)-\tilde{W})^2 + p_i(0,t)(f_i(0,t)-\tilde{W})^2 \tag{3.6}$$

$$\alpha_i = p_i(t)(1-p_i(t))\frac{\partial \tilde{W}}{\partial p_i}$$

$VA(t)$ *is called the* **additive genetic variance**. *Furthermore the average fitness never decreases.*

Proof: Equation 3.3 has been proven in [Müh97]. We have to prove Equation 3.4. Note that

$$p_i(t+1) - p_i(t) = p_i(t)\frac{\bar{f}_i(x_i = 1, t) - \tilde{W}(t)}{\tilde{W}(t)}$$

Obviously we have

$$\frac{\partial \tilde{W}}{\partial p_i} = \bar{f}(x_i = 1, t) - \bar{f}(x_i = 0, t)$$

From $p_i(t)\bar{f}_i(x_i = 1, t) + (1 - p_i(t))\bar{f}_i(x_i = 0, t) = \tilde{W}(t)$, we obtain

$$\begin{aligned} &\bar{f}(x_i = 1, t) - \tilde{W}(t) - (1 - p_i(t))\bar{f}(x_i = 1, t) \\ &\quad + (1 - p_i(t))\bar{f}(x_i = 0, t) = 0 \end{aligned}$$

This gives

$$\bar{f}_i(x_i = 1, t) - \tilde{W}(t) = (1 - p_i(t))\frac{\partial \tilde{W}}{\partial p_i}$$

Inserting this equation into the difference equation gives Equation 3.4.

Equation 3.5 is just the beginning of a multi-dimensional Taylor expansion. The first term follows from

$$\begin{aligned} \sum_i (p_i(t+1) - p_i(t))\frac{\partial \tilde{W}}{\partial p_i} &= \sum_i p_i(t)(1 - p_i(t))\left(\frac{\partial \tilde{W}}{\partial p_i}\right)^2 \\ &= \sum_i p_i(t)(f_i(1,t) - \tilde{W})(f_i(1,t) - \tilde{W} + \tilde{W} - f_i(0,t)) \\ &= \sum_i p_i(t)(f_i(1,t) - \tilde{W})^2 + (1 - p_i(t))(f_i(0,t) - \tilde{W}) \\ &= VA(t) \end{aligned}$$

The above equations completely describe the dynamics of UMDA with proportionate selection. Mathematically UMDA performs gradient ascent in the landscape defined by W or $\tilde{W}$.

Equation 3.4 is especially suited for the theoretical analysis. It is called *Wright's equation* because it has been proposed by Wright in 1937. Wright's [Wri70] remarks are still valid today:

> The appearance of this formula is deceptively simple. Its use in conjunction with other components is not such a gross oversimplification in principle as has sometimes been alleged ... Obviously calculations can be made only from rather simple models, involving only a few loci or simple patterns of interaction among many similarly behaving loci... Apart from application to simple systems, the greatest significance of the general formula is that its form brings out properties of systems that would not be apparent otherwise.

The restricted application lies in the following fact. In general the difference equations need the evaluation of 2^n terms. The computational complexity can be drastically reduced if the fitness function has a special form.

Example 3.1. $f(x) = \sum_i a_i x_i, \quad x_i \in \{0, 1\}$
After some tedious manipulations one obtains:

$$\begin{aligned} W(\mathbf{p}) &= \sum_i a_i p_i(1) \\ \frac{\partial W}{\partial p_i(1)} &= a_i + \sum_{j \neq i} a_j p_j(1) \end{aligned}$$

This gives the difference equation

$$\Delta p_i(1) = p_i(1,t)(1 - p_i(1,t))\frac{a_i}{\sum_i a_i p_i(1,t)} \tag{3.7}$$

Noting that $\frac{\partial \tilde{W}}{\partial p_i(1)} = a_i$ we have proving nothing else than Wright's equation. This equation has been approximately solved in [MM99a].

This example shows that the expressions for W and its derivatives can be surprisingly simple. $W(\mathbf{p})$ can be obtained from $f(\mathbf{x})$ by exchanging x_i with $p_i(1)$. But the formal derivation of $W(\mathbf{p})$ cannot be obtained from the simplified $W(\mathbf{p})$ expression.

We will investigate the computation of W and its gradient in the following section.

3.2 Computing the Average Fitness

Wright is also the originator of the landscape metaphor now popular in evolutionary computation and population genetics. Unfortunately Wright used two quite different definitions for the landscape, apparently without realizing the fundamental distinction between them. The first landscape describes the relation between the genotypes and their fitness, while the second describes the relation between the allele frequencies in a population and its mean fitness.

The first definition is just the fitness function $f(\mathbf{x})$ used in evolutionary computation, the second one is the average fitness $\tilde{W}(\mathbf{p})$. The second definition is much more useful, because it lends to a quantitative description of the evolutionary process, i.e. Wright's equation.

For notational simplicity we only derive the relation between $f(\mathbf{x})$ and $\tilde{W}$ for binary alleles. Let $\alpha = (\alpha_1, \ldots, \alpha_n)$ with $\alpha_i \in \{0, 1\}$ be a multi-index. We define with $0^0 := 1$:

$$\mathbf{x}^\alpha := \prod_i x_i^{\alpha_i}$$

Lemma 3.1. $\tilde{W}(\mathbf{p}) := \bar{f}(t)$ *is an extension of* $f(x)$ *to* S. *There exist two representations for* $\tilde{W}(p)$. *These are given by*

$$\begin{aligned} \tilde{W}(\mathbf{p}) =& f(0, \ldots, 0)(1 - p_1) \cdots (1 - p_n) + \cdots \\ &+ f(1, \ldots, 1)p_1 \cdots p_n \\ \tilde{W}(\mathbf{p}) =& \sum_\alpha a_\alpha p^\alpha \end{aligned}$$

The above lemma can rigorously be proven by Moebius inversion. If the the order of the function is bounded by a constant independent of n, $\tilde{W}(\mathbf{p})$ can be computed in polynomial time. The equation can also be used to compute the derivative of $\tilde{W}$, which is needed for Wright's equation. It is given by

$$\frac{\partial \tilde{W}(p)}{\partial p_i(1)} = \sum_{\alpha|\alpha_i=1} a_\alpha p^{\alpha'} \tag{3.8}$$

with $\alpha'_i = 0, \alpha'_j = \alpha_j$.
We will now characterize the attractors of UMDA. Let $S_i = \{q_i | \sum_{k\in\{0,1\}} q_i(x_k) \leq 1; \quad 0 \leq q_i(x_k) \leq 1\}$ and $S = \prod_i S_i$ the Cartesian product. Then $S = [0,1]^n$ is the unit cube.

Theorem 3.2. *The stable attractors of Wright's equation are at the corners of S, i.e $p_i \in \{0,1\} \quad i = 1,\ldots,n$. In the interior there are only saddle points or local minima where grad $W(p)) = 0$. The attractors are local maxima of $f(x)$ according to one bit changes. Wright's equation solves the continuous optimization problem* $\operatorname{argmax}\{\tilde{W}(\mathbf{p})\}$ *in S by gradient ascent.*

Proof: W is linear in p_i, therefore it cannot have any local maxima in the interior. Points with *grad* $W(p) = 0$ are unstable fixpoints of UMDA.

We next show that boundary points which are not local maxima of $f(x)$ cannot be attractors. We prove the conjecture indirectly. Without loss of generality, let the boundary point be $\hat{p} = (1,\ldots,1)$. We now consider an arbitrary neighbor, i.e $p^* = (0,1,\ldots,1)$. The two points are connected at the boundary by

$$p(z) = (1-z, 1, \ldots, 1) \qquad z \in [0,1]$$

We know that $\tilde{W}$ is *linear* in the parameters p_i. Because $\tilde{W}(p^*) = f(0,1,\ldots,1)$ and $\tilde{W}(\hat{p}) = f(1,\ldots,1)$ we have

$$\tilde{W}(p(z)) = f(1,\ldots,1) + z \cdot [f(0,1,\ldots,1) - f(1,\ldots,1)]. \tag{3.9}$$

If $f(0,1,\ldots,1) > f(1,\ldots,1)$ then $\hat{p}$ cannot be an attractor of UMDA. The mean fitness increases with z. □

The extension of the above lemma to multiple alleles and multivariate distributions is straightforward, but the notation becomes difficult.

4 The Science of Breeding

Fitness proportionate selection is the undisputed selection method in population genetics. It is considered to be a model for *natural selection*. But for proportionate selection the following problem arises. When the population approaches an optimum, selection gets weaker and weaker because the fitness values become similar. Therefore breeders of livestock use other selection methods. These are called *artificial selection*. For large populations they mainly apply *truncation selection*. It works as follows. A truncation threshold τ is fixed. Then the τN best individuals are selected as parents for the next generation. These parents are then randomly mated.

The science of breeding is the domain of *quantitative genetics*. The theory is based on macroscopic variables. Because an exact mathematical analysis is impossible, many statistical techniques are used. In fact, the concepts of regression, correlation, heritability and decomposition of variance have been developed and applied in quantitative genetics for the first time.

4.1 Single Trait Theory

For a single trait the theory can be easily summarized. Starting with the fitness distribution, the *selection differential* $S(t)$ is introduced. It is the difference between the average of the selected parents and the average of the population.

$$S(t) = W(\mathbf{p}^s(t+1)) - W(\mathbf{p}(t)) \tag{4.1}$$

Similarly the response $R(t)$ is defined

$$R(t) = W(\mathbf{p}(t+1)) - W(\mathbf{p}(t)) \tag{4.2}$$

Next a linear regression is done

$$R(t) = b(t)S(t) \tag{4.3}$$

$b(t)$ is called *realized heritability*. The most difficult part of applying the theory is to predict $b(t)$. The first estimate uses the *regression of offspring to parent*. Let f_i, f_j be the phenotypic values of parents i and j, then

$$\bar{f}_{i,j} = \frac{f_i + f_j}{2}$$

is called the mid-parent value. Let the stochastic variable $\bar{F}$ denote the mid-parent value.

Theorem 4.1. *Let $P(t) = (f_1,\ldots,f_N)$ be the population at generation t, where f_i denotes the phenotypic value of individual i. Assume that an offspring generation $O(t)$ is created by random mating, without selection. If the regression equation*

$$o_{ij}(t) = a(t) + b_{\bar{P}O}(t) \cdot \frac{f_i + f_j}{2} + \epsilon_{ij} \tag{4.4}$$

with

$$E(\epsilon_{ij}) = 0$$

is valid, where o_{ij} is the fitness value of the offspring of i and j, then

$$b_{\bar{P}O}(t) \approx b(t) \tag{4.5}$$

Proof: From the regression equation we obtain for the expected averages

$$E(O(t)) = a(t) + b_{\bar{P}O}(t)M(t)$$

Because the offspring generation is created by random mating without selection, the expected average fitness remains constant

$$E(O(t)) = M(t)$$

Let us now select a subset as parents. The parents will be randomly mated, producing the offspring generation. If the subset is large enough, we may still use the regression equation and obtain for the averages

$$M(t+1) = a(t) + b_{\bar{P}O}(t) \cdot M_s(t)$$

Here $M(t+1)$ is the average fitness of the offspring generation produced by the selected parents. Subtracting the above equations we obtain

$$M(t+1) - M(t) = b_{\bar{P}O}(t) \cdot (M_s(t) - M(t))$$

This proves $b_{\bar{P}O}(t) = b(t)$.

The importance of regression for estimating the heritability was discovered by Galton and Pearson at the end of the 19th century. They computed the regression coefficient rather intuitively by scatter diagrams of mid-parent and offspring. The problem of computing a good regression coefficient is mathematically solved by the theorem of Gauss-Markov. We just cite the theorem. The proof can be found in any textbook on statistics [Rao73].

Theorem 4.2. *A good estimate for the regression coefficient of mid-parent and offspring is given by*

$$b_{\bar{P}O}(t) = \frac{cov(O(t), \bar{P}(t))}{var(\bar{P}(t))} \tag{4.6}$$

The covariance of O and $\bar{P}$ is defined by

$$cov(O(t), \bar{P}(t)) = \frac{1}{N}\sum_{i,j}(o_{i,j} - av(O(t))) \cdot (\bar{f}_{i,j} - av(\bar{P}(t)))$$

av denotes the average and var the variance. Closely related to the regression coefficient is the correlation coefficient $cor(\bar{F}, O)$. It is given by

$$cor(\bar{P}(t), O(t)) = b_{\bar{P}O}(t) \cdot (\frac{var(\bar{P}(t))}{var(O(t))})^{1/2}$$

The concept of covariance is restricted to parents producing offspring. It cannot be used for UMDA. Here the *analysis of variance* helps. We will decompose the fitness value $f(\mathbf{x})$ recursively into an additive part and interaction parts. We recall the definition of conditional probability.

Definition 4.1. *Let $p(\mathbf{x})$ denote the probability of $\mathbf{x}$. Then the conditional probability $p(\mathbf{x}|y)$ of $\mathbf{x}$ given y is defined by*

$$p(\mathbf{x}|y) = \frac{p(\mathbf{x}, y)}{p(y)} \tag{4.7}$$

The proof of the next theorem can be found in [AM94a].

Theorem 4.3. *Let the population be in linkage equilibrium i.e.*

$$p(\mathbf{x}) = \prod_{i=1}^{n} p_i(x_i) \tag{4.8}$$

Then the variance of the population is given by

$$V = V_1 + V_2 + \cdots + V_{n-1} + V_n \tag{4.9}$$

The covariance of mid-parent and offspring can be computed from

$$cov(\bar{P}, O) = \frac{1}{2}V_1 + \frac{1}{4}V_2 + \cdots + \frac{1}{2^n}V_n = \sum_{k=1}^{n}\frac{1}{2^k}V_k \tag{4.10}$$

We now compare the estimates for heritability. For proportionate selection we have from Theorem 3.1

$$R_{UMDA}(t) = \frac{V_A(t)}{V(t)}S(t) + error_1(t).$$

For Two-Parent-Recombination (TPR) Mühlenbein (1997) has shown for $n = 2$ loci

$$R_{TPR}(t) = 2\frac{cov(\bar{P}(t), O(t))}{V(t)}S(t) + \frac{1}{2}error_2(t)$$

If the population is in linkage equilibrium we have $error_1 = error_2$ Using the covariance decomposition we can write

$$R_{TPR}(t) = \frac{VA(t)}{V(t)}S(t) + \frac{1}{2}\frac{V_2(t)}{V(t)}S(t) + \frac{1}{2}error(t)$$

Thus the first term of the expansion is identical to the $UMDA$ term. This shows again the similarity between two parent recombination and the $UMDA$ method. Breeders usually use the expression $b(t) = VA(t)/V(t)$ as an estimate. It is called *heritability in the narrow sense* [Fal81]. But note that the variance decomposition seems to be only true for Robbins' proportions.

The selection differential is not suited for mathematical analysis. For truncation selection it can be approximated by

$$S(t) \approx I_\tau V^{\frac{1}{2}}(t) \tag{4.11}$$

where I_τ is called the *selection intensity*. Combining the two equations we obtain the *famous equation for the response to selection.*

$$R(t) = b(t) I_\tau V^{\frac{1}{2}}(t) \tag{4.12}$$

These equations are in depth discussed in [Müh97]. The theory of breeding uses macroscopic variables, the average and the variance of the population. But we have derived only one equation, the response to selection equation. We need a second equation connecting the average fitness and the variance in order to be able to compute the time evolution of the average fitness and the variance. There have been many attempts in population genetics to find a second equation. But all equations assume that the variance of the population continuously decreases. This is not the case for arbitrary fitness functions. Recently Prügel-Bennet and Shapiro [PBS97] have independently proposed to use moments for describing genetic algorithms. They apply methods of statistical physics to derive equations for higher moments for special fitness functions.

Results for tournament selection and analytical solutions for linear functions can be found in [MM00].

We next present numerical results for some popular fitness functions.

4.2 Numerical Results for UMDA

This section solves the problem put forward by Mitchell et al. [MHF94]: to understand the class of problems for which genetic algorithms are most suited, and in particular, for which they will outperform other search algorithm. The famous Royal-Road function is analyzed in [MM00].

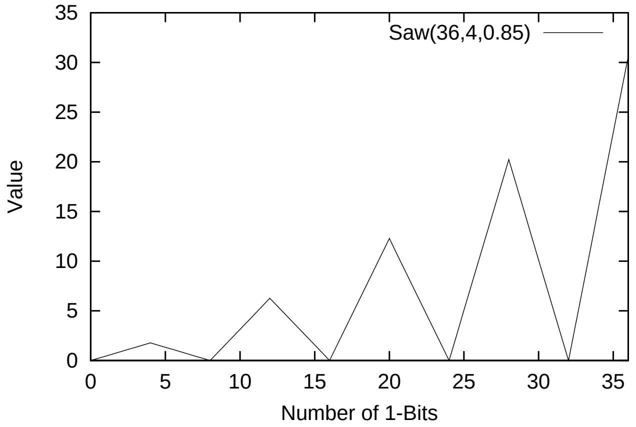

Figure 2: Definition of Saw(36,4,0.85)

Equation 3.4 shows that UMDA performs a gradient ascent in the landscape given by W. This helps our search for functions best suited for UMDA. We take the Saw landscape as a spectacular example. The definition of the function can be extrapolated from Figure 2. In Saw(n, m, k), n denotes the number of bits and $2m$ the distance from one peak to the next. The highest peak is multiplied by k (with $k \leq 1$), the second highest by k^2, then k^3 and so on. The landscape is very rugged. In order to get from one local optimum to another one, one has to cross a deep valley.

But again the transformed landscape $W(\mathbf{p})$ is fairly smooth. An example is shown in Figure 3. Whereas $f(\mathbf{x})$ has 5 isolated peaks, $W(\mathbf{p})$ has three plateaus, a local peak and the global peak. We will use $UMDA$ with truncation selection. We have not been able to derive precise analytical expressions. In Figure 3 the results are displayed.

In the simulation two truncation thresholds, $\tau = 0.05$ and $\tau = 0.01$, have been used. For $\tau = 0.05$ the probability p stops at the local maximum for $\tilde{W}(\mathbf{p})$. It is approximately $p = 0.78$. For $\tau = 0.01$ $UMDA$ is able to converge to the optimum $p = 1$. It does so by even going downhill!

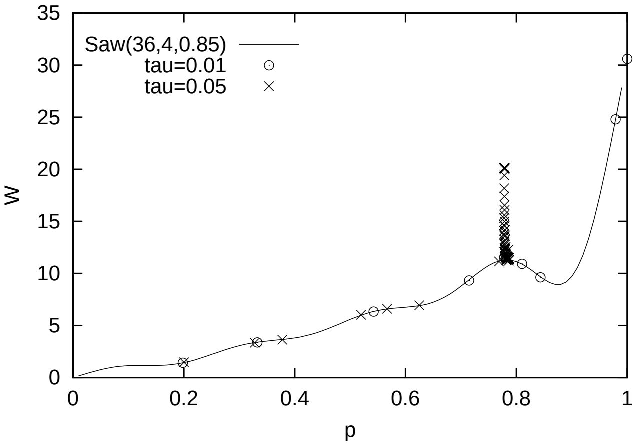

Figure 3: Results with normal and strong selection.

This example confirms in a nutshell our theory. *$UMDA$ transforms the original fitness landscape defined by $f(\mathbf{x})$ into a fitness landscape defined by $\tilde{W}(\mathbf{p})$. This transformation smoothes the rugged fitness landscape $f(\mathbf{x})$. $UMDA$ might find the global optimum, if there is a tendency towards the global optimum.*

This example shows that UMDA can solve difficult multimodal optimization problems. It is obvious that any search method using a single search point like the $(1+1)$-algorithm needs an almost exponential number of function evaluations. We next show how the science of breeding can be used for controlling $UMDA$.

4.3 Numerical Investigations of the Science of Breeding

The application of the science of breeding needs the computation of the average fitness $\bar{f}(t)$, the variance $V(t)$ and the additive genetic variance $VA(t)$. The first two terms are standard statistical terms. The computation of VA needs $\bar{f}_i(x_i)$ and $p_i(x_i)$. The computation of the first term only poses some difficulties. It can be approximated by

$$\bar{f}_i(X_i = 1, t) = \sum_{x, X_i = 1} \frac{p(\mathbf{x})}{p_i(X_i = 1)} f(\mathbf{x}) \approx \frac{1}{N} \sum_{k=1}^{N} \frac{f(\zeta_i^k)}{p_i(x_i)} \tag{4.13}$$

ζ_i^k are those $\mathbf{x}$ values in the population which contain $x_i = 1$.

Linear functions are the ideal case for the theory. The heritability $b(t)$ is 1 and the additive genetic variance is identical to the variance. We skip this trivial case and start with a multiplicative fitness function $f(\mathbf{x}) = \prod_i (1-s)^{1-x_i}$. For a multiplicative function we also have $R(t) = S(t)$.

Figure 4 confirms the theoretical results from Section 2 (VA and Var are multiplied by 10 in this figure). Additive genetic variance is almost identical to the variance and the heritability is 1. The function is highly nonlinear of order n,

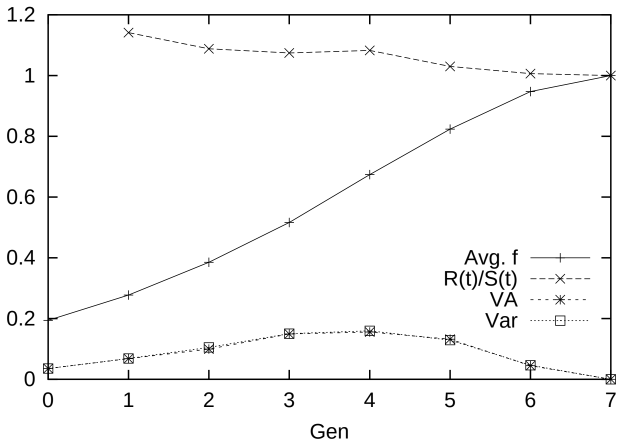

Figure 4: Heritability and Variance for a multiplicative function (variance and VA multiplied by 10); $s = 0.1$, $n = 32$

but nevertheless it is easy to optimize. The function has also been investigated by Rattray and Shapiro [RS99]. But their calculations are very difficult.

We have shown that $UMDA$ can optimize difficult multimodal functions, thus explaining the success of genetic algorithms in optimization. We have also shown that $UMDA$ can easily be deceived by simple functions called deceptive functions. These functions need more complex search distributions. This problem is investigated next.

5 Graphical Models and Optimization

The simple product distribution of $UMDA$ cannot capture dependencies between variables. If these dependencies are necessary to find the global optimum, $UMDA$ and simple genetic algorithms fail. We take an extreme case as example, the *needle in a haystack problem.* The fitness function is everywhere one, except for a single $\mathbf{x}$ where it is 10. All x_i values have to be set in the right order to obtain the optimum. Of course, there exist no clever search method for this problem. But there is a continuum of increasing complexity from the simple $OneMax$ function to the needle in a haystack. For complex problems we need a complex search distribution. A good candidate for a search distribution for optimization is the Boltzmann distribution.

Definition 5.1. *For $\beta \geq 0$ define the* weighted Boltzmann distribution *of a function $f(\mathbf{x})$ as*

$$p_{\beta,f}(\mathbf{x}) := \frac{p_0(\mathbf{x})e^{\beta f(\mathbf{x})}}{\sum_y p_0(\mathbf{y})e^{\beta f(\mathbf{y})}} := \frac{p_0(\mathbf{x})e^{\beta f(\mathbf{x})}}{Z_f(\beta, p_0)} \qquad (5.1)$$

where $Z_f(\beta, p_0)$ is the partition function. To simplify the notation β and/or f can be omitted. $p_0(\mathbf{x})$ is the distribution for $\beta = 0$.

The Boltzmann distribution concentrates the search around good fitness values. Thus it is theoretically a very good candidate for a search distribution used for optimization. The problem lies in the efficient computation of the Boltzmann distribution. The theory presented in this section unifies simulated annealing and population based algorithms with the general theory of estimating distributions.

x

5.1 Boltzmann selection

Closely related to the Boltzmann distribution is Boltzmann selection. An early study about this selection method can be found in [dlMT93].

Definition 5.2. *Given a distribution p and a selection parameter γ, **Boltzmann selection** calculates the distribution of the selected points according to*

$$p^s(\mathbf{x}) = \frac{p(\mathbf{x})e^{\gamma f(\mathbf{x})}}{\sum_y p(\mathbf{y})e^{\gamma f(\mathbf{y})}} \qquad (5.2)$$

This allows us to define the $BEDA$ (Boltzmann Estimated Distribution Algorithm).

BEDA – Boltzmann Estimated Distribution Algorithm

- **STEP 0:** $t \Leftarrow 0$. Generate N points according to the $p(\mathbf{x}, 0) = p_0(\mathbf{x})$.
- **STEP 1:** With a given $\Delta\beta(t) > 0$, let
$$p^s(\mathbf{x}, t) = \frac{p(\mathbf{x}, t)e^{\Delta\beta(t)f(\mathbf{x})}}{\sum_y p(\mathbf{y}, t)e^{\Delta\beta(t)f(\mathbf{y})}}.$$
- **STEP 2:** Generate N new points according to the distribution $p(x, t+1) = p^s(x, t)$.
- **STEP 3:** $t \Leftarrow t + 1$.
- **STEP 4:** If stopping criterion not met go to STEP 1

$BEDA$ is a conceptional algorithm, because the calculation of the distribution requires to compute the sum of exponentially many terms. The following convergence theorem is easily proven.

Theorem 5.1 (Convergence). *Let $\Delta\beta(t)$ be an annealing schedule, i.e. for every t increase the inverse temperature β by $\Delta\beta(t)$. Then for $BEDA$ the distribution at time t is given by*

$$p(\mathbf{x}, t) = \frac{p_0(\mathbf{x})e^{\beta(t)f(\mathbf{x})}}{Z_f(\beta(t), p_0)} \qquad (5.3)$$

with the inverse temperature

$$\beta(t) = \sum_{\tau=1}^{t} \Delta\beta(\tau). \qquad (5.4)$$

Let $\mathcal{M}$ be the set of global optima. If $\beta(t) \to \infty$, then

$$\lim_{t\to\infty} p(x, t) = \begin{cases} 1/|\mathcal{M}| & x \in \mathcal{M} \\ 0 & else \end{cases} \qquad (5.5)$$

Proof: Let $x^m \in \mathcal{M}$ be a point with optimal fitness and $x \notin \mathcal{M}$ a point with $f(\mathbf{x}) < f(x^m)$. Then

$$p(x,t) = \frac{p_0(\mathbf{x})e^{\beta(t)f(\mathbf{x})}}{\sum_y p_0(y)e^{\beta(t)f(y)}} \leq \frac{e^{\beta(t)f(\mathbf{x})}}{|\mathcal{M}| \cdot C \cdot e^{\beta(t)f(x^m)}}$$
$$\leq \frac{1}{|\mathcal{M}| \cdot C \cdot e^{\beta(t)[f(x^m)-f(\mathbf{x})]}}$$

As $\beta(t) \to \infty$, $p(x,t)$ converges (exponentially fast) to 0. Because $p(x,t) = p(y,t)$ for all $x^m, y^m \in \mathcal{M}$, the limit distribution is the uniform distribution on the set of optima. □

We next transform $BEDA$ into a practical algorithm. This means the reduction of the parameters of the distribution and the computation of an adaptive schedule.

5.2 Factorization of the distribution

In this section we describe a method for computing a factorization of the probability, given an additive decomposition of the function:

Definition 5.3. *Let $s_1, \ldots, s_m$ be index sets, $s_i \subseteq \{1, \ldots, n\}$. Let f_{s_i} be functions depending only on the variables x_j with $j \in s_i$. These variables we denote as x_{s_i} Then*

$$f(\mathbf{x}) = \sum_{i=1}^{m} f_{s_i}(\mathbf{x}) = f_i(x_{s_i}) \tag{5.6}$$

is an additive decomposition *of the fitness function f.*

Definition 5.4. *Given $s_1, \ldots, s_m$, we define for $i = 1, \ldots, m$ the sets d_i, b_i and c_i:*

$$d_i := \bigcup_{j=1}^{i} s_j, \quad b_i := s_i \setminus d_{i-1}, \quad c_i := s_i \cap d_{i-1} \tag{5.7}$$

We set $d_0 = \emptyset$.

In the theory of decomposable graphs, d_i are called *histories, b_i residuals* and c_i *separators* [Lau96]. We recall the following definition.

Definition 5.5. *The conditional probability $p(\mathbf{x}|\mathbf{y})$ is defined as*

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{y})} \tag{5.8}$$

In [MMO99], we have shown the following theorem.

Theorem 5.2 (Factorization Theorem). *Let $p(\mathbf{x})$ be a Boltzmann distribution with*

$$p(\mathbf{x}) = \frac{e^{\beta f(\mathbf{x})}}{Z_f(\beta)} \tag{5.9}$$

and $f(\mathbf{x}) = \sum_{i=1}^{m} f_{s_i}(\mathbf{x})$ be an additive decomposition. If

$$b_i \neq \emptyset \quad \forall i = 1, \ldots, l; \quad d_l = \tilde{X}, \tag{5.10}$$
$$\forall i \geq 2 \; \exists j < i \;\; \text{such that } c_i \subseteq s_j \tag{5.11}$$

then

$$p(\mathbf{x}) = \prod\nolimits_{i=1}^{m} p(x_{b_i}|x_{c_i}) \tag{5.12}$$

The constraint defined by Equation (5.11) is called the *running intersection property* [Lau96].

FDA – Factorized Distribution Algorithm

- **STEP 0:** Calculate b_i and c_i from the decomposition of the function.
- **STEP 1:** Generate an initial population with N individuals.
- **STEP 2:** Select N individuals using Boltzmann selection.
- **STEP 3:** Estimate the conditional probabilities $p(x_{b_i}|x_{c_i}, t)$ from the selected points.
- **STEP 4:** Generate new points according to $p(\mathbf{x}, t+1) = \prod_{i=1}^{m} p(x_{b_i}|x_{c_i}, t)$.
- **STEP 5:** If not stopping criterion reached: $t \Leftarrow t+1$ Go To STEP2

With the help of the factorization theorem, we can turn the conceptional algorithm $BEDA$ into FDA, the Factorized Distribution Algorithm. As the factorized distribution is identical to the Boltzmann distribution if the conditions of the factorization theorem are fulfilled, the convergence proof of $BEDA$ also applies to FDA.

Not every additive decomposition leads to a factorization using the factorization theorem. In these cases, more sophisticated methods have to be used. But FDA can also be used with an approximate factorization.

We discuss a simple example.

Example 5.1. *Functions with a chain-like interaction can also be factorized:*

$$Chain(\mathbf{x}) = \sum_{i=2}^{n} f_i(x_{i-1}, x_i) \tag{5.13}$$

Here the factorization is

$$p(\mathbf{x}) = p(x_1) \prod_{i=2}^{n} p(x_i|x_{i-1}) \tag{5.14}$$

FDA can be used with any selection scheme, but then the convergence proof is no longer valid. We think that Boltzmann selection is an essential part in using the FDA. In order to obtain a practical algorithm, we still have to solve two problems: To find a good annealing schedule for Boltzmann selection and to determine a reasonable sample size (population size). These two problems will be investigated next.

5.3 A new annealing schedule for the Boltzmann distribution

Boltzmann selection needs an annealing schedule. But if we anneal too fast, the approximation of the Boltzmann due to the sampling error can be very bad. For an extreme case, if the annealing parameter is very large, the second generation should consist only of the global maxima.

5.3.1 Taylor expansion of the average fitness

In order to determine an adaptive annealing schedule, we will make a Taylor expansion of the average fitness of the Boltzmann distribution.

Definition 5.6. *The **average fitness** of a fitness function and a distribution is*

$$W_f(p) = \sum_x f(\mathbf{x})p(\mathbf{x}) \tag{5.15}$$

For the Boltzmann distribution, we use the abbreviation $W_f(\beta) := W_f(p_{\beta,f})$.

Theorem 5.3. *The average fitness of the Boltzmann distribution $W_f(\beta)$ has the following expansion in β:*

$$W_f(\tilde{\beta}) = W_f(\beta) + \sum_{i\geq 1} \frac{(\tilde{\beta}-\beta)^i}{i!} M^c_{i+1}(\beta) \tag{5.16}$$

where M^c_i are the centred moments

$$M^c_i(\beta) := \sum_x \left[f(\mathbf{x}) - W_f(\beta)\right]^i p(\mathbf{x}) \tag{5.17}$$

They can be calculated using the derivatives of the partition function:

$$M^c_{i+1}(\beta) = \left(\frac{Z'_f(\beta)}{Z_f(\beta)}\right)^{(i)} \qquad \text{for } i \geq 1, \quad M^c_1 = 0 \tag{5.18}$$

Proof: The k-th derivative of the partition function obeys for $k \geq 0$:

$$Z^{(k)}_f(\beta) = \sum_x f(\mathbf{x})^k e^{\beta f(\mathbf{x})} \tag{5.19}$$

Thus the moments for $k \geq 1$ can be calculated as

$$M_k(\beta) := \sum_x f(\mathbf{x})^k p(\mathbf{x}) = \frac{Z^{(k)}_f(\beta)}{Z_f(\beta)} \tag{5.20}$$

and thus

$$W_f(\beta) = M_1(\beta) = Z'_f(\beta)/Z_f(\beta). \tag{5.21}$$

Direct evaluation of the derivatives of W leads to complicate expressions. The proof is rather technical by induction. We omit it here. □

We can now derive an adaptive annealing schedule. The variance (and the higher moments) can be estimated from the generated points. As long as the approximation is valid, one can choose a desired increase in the average fitness and set $\beta(t+1)$ accordingly. So we can set

$$\Delta\beta(t) := \beta(t+1) - \beta(t) = \frac{W^{\text{new}}_f(t) - W_f(\beta(t))}{\sigma^2_f(\beta(t))} \tag{5.22}$$

As truncation selection has proven to be a robust and efficient selection scheme, we can try to approximate the behaviour of this method.

Definition 5.7. *The standard deviation schedule (SDS) is defined by*

$$\Delta\beta(t) = \frac{c}{\sigma_f(\beta(t))} \tag{5.23}$$

Note that this annealing schedule cannot be used for simulated annealing, as the estimation of the variance of the distribution requires samples that are independently drawn. But the sequence of samples generated by simulated annealing are not independent.
We next turn to the fixation problem in finite populations.

5.4 Finite Populations

In finite populations convergence of $UMDA$ or FDA can only be probabilistic. Since $UMDA$ a simple FDA algorithm, it is sufficient to discuss FDA. This section is extracted from [MM99b].
Definition 5.8. *Let ϵ be given. Let $P_{conv}(N)$ denote the probability that FDA with a population size of N converges to the optima. Then the critical population size is defined as*

$$N^*(\epsilon) = \min_N P_{conv}(N) \geq 1-\epsilon \tag{5.24}$$

If FDA with a finite population does not convergence to an optimum, then at least one gene is fixed to a wrong value. The probability of fixation is reduced if the population size is increased. We obviously have for FDA

$$P_{conv}(N_1) \leq P_{conv}(N_2) \quad N_1 \leq N_2$$

The critical question is: How many sample points are necessary to reasonably approximate the distribution used by FDA. A general estimate from Vapnik [Vap98] can be a guideline. One should use a sample size which is about 20 times larger than the number of free parameters.

We discuss the problem with a special function called *Int*. *Int*$(\mathbf{x})$ gives the integer value of the binary representation.

$$Int(n) = \sum_{i=1}^{n} 2^{i-1} x_i \tag{5.25}$$

The fitness distribution of this function is not normal distributed. The function has 2^n different fitness values. We show the cumulative fixation probability in Table 1 for

	$\tau = 0.25$	$\tau = 0.5$	$Boltz.$	SDS
t	$N = 80$	$N = 60$	$N = 700$	$N = 100$
1	0.0	0.0	0.0885	0.0
2	0.0025	0.0095	0.1110	0.0
3	0.0165	0.0205	0.1275	0.0
4	0.0355	0.0325	0.1375	0.002
5	0.0575	0.0490	0.1455	0.002
6	0.0695	0.0630	0.1510	0.008
7	0.0740	0.0715	0.1555	0.018
8	0.0740	0.0780	0.1565	0.030
9	0.0740	0.0806	0.1575	0.036
14				0.084

Table 1: Cumulative fixation probability for *Int*(16). Truncation selection vs. Boltzmann selection with $\Delta\beta = 0.01$ and Boltzmann SDS; N denotes size of population.

Int(16). The fixation probability is larger for stronger selection. For a given truncation selection the maximum fixation probability is at generation 1 for very small N. For larger values of N the fixation probability increases until a maximum is reached and then decreases again. This behaviour has been observed for many fitness distributions.

For truncation selection with $\tau = 0.25$ we have for $N = 80$ a fixation probability of about 0.075. A larger τ reduces the fixation probability. But this advantage is set off by the larger number of generations needed to converge. The problem of an optimal population size for truncation selection is investigated in [MM99b]. Boltzmann selection with $\Delta\beta = 0.01$ is still very strong for the fitness distribution given by *Int*(16). For $N = 700$ the largest fixation probability is still at the first generation. Therefore the critical population size for Boltzmann selection for $\Delta\beta = 0.01$ is very high ($N^* > 700$). In comparison, the adaptive Boltzmann schedule SDS has a total fixation probability of 0.084 for a population size of $N = 100$. This is almost as small as truncation selection.

This example shows that Boltzmann selection in finite populations critically depends on a good annealing schedule. Normally we run FDA with truncation selection. This selection method is a good compromise. But Boltzmann selection with SDS schedule is of comparable performance.

Estimates for the necessary size of the population can also be found in [HCPGM99]. But they use a weaker performance definition. The goal is to have a certain percentage of the bits of the optimum in the final population. Furthermore their result is only valid for fitness function which are approximately normally distributed.

The danger of fixation can further be reduced by a technique very popular in Bayesian statistics. This is discussed in the next section.

5.5 Bayesian Networks, Population Size and Bayesian Prior

In order to derive the results of this section we will use a normalized representation of the distribution.

Theorem 5.4 (Bayesian Factorization). *Each probability can be factored into*

$$p(\mathbf{x}) = p(x_1)\prod_{i=2}^{n} p(x_i|pa_i) \tag{5.26}$$

Proof: By definition of conditional probabilities we have

$$p(\mathbf{x}) = p(x_1)\prod_{i=2}^{n} p(x_i|x_1,\cdots,x_{i-1}) \tag{5.27}$$

Let $pa_i \subset \{x_1,\cdots,x_{i-1}\}$. If x_i and $\{x_1,\cdots,x_{i-1}\}\setminus pa_i$ are conditionally independent given pa_i, we can simplify $p(x_i|x_1,\cdots,x_{i-1}) = p(x_i|pa_i)$. □

PA_i are called the parents of variable X_i. This factorization can be represented by a directed graph. In the context of graphical models the graph and the conditional probabilities are called a *Bayesian network* [Jor99, Fre98]. It is obvious that the factorization used in Theorem 5.2 can be easily transformed into a Bayesian factorization.

Usually the empirical probabilities are computed by the maximum likelihood estimator. For N samples with $m \le N$ instances of x the estimate is defined by

$$\hat{p}(x) = \frac{m}{N}$$

For $m = N$ we obtain $p(x) = 1$ and for $m = 0$ we obtain $p(x) = 0$. This leads to our gene fixation problem, because both values are attractors. The fixation problem is reduced if $\hat{p}(x)$ is restricted to an interval $0 < p_{min} \le \hat{p}(x) \le 1 - p_{min} < 1$. This is exactly what results from the *Bayesian estimation*. The estimate $\hat{p}(x)$ is the expected value of the posterior distribution after applying Bayes formula to a prior distribution and the given data. For binary variables x the estimate

$$\hat{p}(x) = \frac{m+r}{N+2r} \tag{5.28}$$

is used with $r > 0$. r is derived from a Bayesian prior. $r = 1$ is the result of the uniform Bayesian prior. The larger r, the more the estimates tend towards $1/2$. The reader interested in a derivation of this estimate in the context of Bayesian networks is referred to [Jor99].

How can we determine an appropriate value for r for our FDA application? The uniform prior gives for $m = 0$ the value $\hat{p}_{min} = 1/(N+2)$. If N is small, then p_{min} might be so large that we generate the optima with a very small probability only. This means we perform more a random search instead of converging to the optima. This consideration leads to a constraint. $1 - p_{min}$ should be so large that the optima are still generated with high probability. We now heuristically derive p_{min} under the assumption that the optimum is unique. To simplify the formulas we require that $\max \hat{p}(x_{opt}) \geq e^{-1}$.

This means that the optimum string x_{opt} should be generated more than 30% at equilibrium. This is large enough to observe equilibrium and convergence. Let us first investigate the $UMDA$ factorization $p(x) = \prod p(x_i)$. For $r = 1$ the largest probability is $p_{max} = (N+1)/(N+2)$. Obviously

$$p_{max} = 1 - \frac{1}{N+2} = 1 - p_{min}$$

The largest probability to generate the optimum is given by

$$\hat{p}(x_{opt}) = \prod_{i=1}^{n}(1 - \frac{1}{N+2}) \approx e^{-\frac{n}{N+2}}$$

If $N = O(n^{1-\alpha})$ with $\alpha > 0$, then $p(x_{opt})$ becomes arbitrarily small for large n. For $N = n$ we obtain $\hat{p}(x_{opt}) \approx e^{-1}$. This gives the following guideline, which actually is a lower bound of the population size.

Rule of Thumb: *For $UMDA$ the size of the population should be at least equal to the size of the problem, if a Bayesian prior of $r = 1$ is used.*

Bayesian priors are also defined for conditional distributions. The above heuristic derivation can also be used for general Bayesian factorizations. The Bayesian estimator is for binary variables

$$\hat{p}(x_i|pa_i) = \frac{m+r}{P+2r}$$

P is the number of occurrences of pa_i. We make the assumption that in the best case the optimum constitutes 25% of the population. This gives $P \geq N/4$. For $r = 1$ we compute as before

$$\begin{aligned}\hat{p}(x_{opt}) &= \prod_{i=1}^{n}\hat{p}(x_{opt_i}|pa_{opt_i}) = \prod_{i=1}^{n}(1 - \frac{1}{N/4+2}) \\ &\approx e^{-\frac{n}{N/4+2}}\end{aligned}$$

If we set $N = 4n$ we obtain $\hat{p}(x_{opt}) \approx e^{-1}$. Thus we obtain a lower bound for the population size:

Rule of Thumb: *For FDA using a factorization with many conditional distributions and Bayesian prior of $r = 1$, the size of the population should be about four times the size of the problem.*

These rule of thumbs have been heuristically derived. They have to be confirmed by numerical studies. Our FDA estimate is a crude lower bound. There exist more general estimates. We just cite Vapnik [Vap98]. In order to approximate a distribution with a reasonable accuracy, he proposes to use a sample size which is about 20 times larger than the number of free parameters of the distribution. For $UMDA$ this gives $20n$, i.e. 20 times our estimate.

We demonstrate the importance of using a Bayesian prior by an example. It is a deceptive function of order 4 and problem size of $n = 32$. Our convergence theorem gives convergence of FDA with Boltzmann selection and an exact factorization. The exact factorization consists of marginal distributions of size 4. We compare in Figure 5 FDA with SDS Boltzmann selection and truncation selection without Bayesian prior. We also show a run with SDS Boltzman selection and Bayesian prior.

The simulation was started at $p = 0.15$, i.e. near the local optimum $p = 0$. Nevertheless, FDA converges to the global optimum at $p = 1$. It is interesting to note that FDA at first moves into the direction of the local optimum. At the very last moment the direction of the curve is dramatically changed. SDS Boltzmann selection behaves almost identical to truncation selection with threshold $\tau = 0.35$. But both methods need a huge population size in order to converge to the optimum. In this example it is $N = 20000$. If a prior of $r = 1$ is used the population size can be reduced to $N = 200$. With this prior the curve changes direction earlier. Because of the prior the univariate marginal probabilities never reach $p = 0$ or $p = 1$. In this example p stops at about $p = 0.975$.

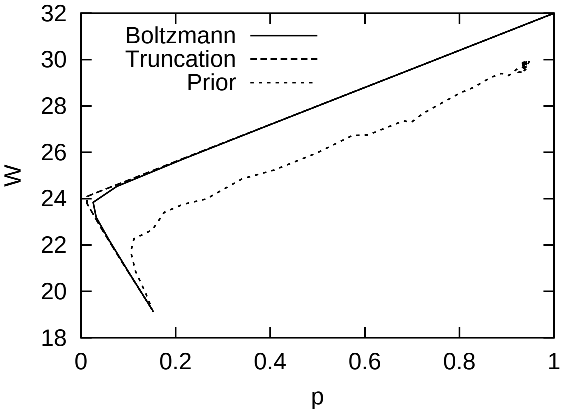

Figure 5: Average fitness $W(p)$ for FDA for Decep(32,4); population size $N = 20000$ without prior and $N = 200$ with prior $r = 1$.

Let us now summarize the results: Because FDA uses finite samples of points to estimate the conditional probabilities, convergence to the optimum will depend on the size of the samples (the population size). FDA has experimentally proven to be very successful on a number of functions where standard genetic algorithms fail to find the global optimum. In [MM99b], the scaling behaviour for various test functions has been studied. The estimation of the probabilities and the generation of new points can be done in polynomial time. If a Bayesian prior is used the influence of the population size is reduced. There is a tradeoff. If no prior is used then convergence is fast. But a large population size might be needed. If a prior is used, the population size can be much smaller. But the number of generations until convergence increases. We have not yet enough numerical results, therefore we just conjecture:

FDA with a finite population of size $N = 4n$, SDS Boltzmann selection, Bayesian prior, and a Bayesian factorization where the number of parents is restricted by k independent of n, will converge to the optimum in polynomial time with high probability.

5.6 Constraint Optimization Problems

An advantage of FDA compared to genetic algorithm is that it can handle optimization problems with constraints. Mendelian recombination or crossover in genetic algorithms often creates points which violate the constraints. If the structure of the constraints and the structure of the ADF are compatible, then FDA will generate only legal points.

Definition 5.9. *A constraint optimization problem is defined by*

$$max f(\mathbf{x}) = \sum_{i=1}^{m} f_i(\mathbf{x}_{s_i}) \tag{5.29}$$

$$s.t. C_i(\mathbf{x}_{u_i}) \tag{5.30}$$

$C_i(\mathbf{x}_{u_i})$ stands for the ith constraint function. $\mathbf{x}_{s_i}, \mathbf{x}_{u_i} \subseteq X$ are sets of variables. The constraints are locally defined. Thus they can be used to test which marginal probabilities are 0. This is technically somewhat complicated, but straightforward. For instance, if we have $C_1(x_1, x_2) = x_1 + x_2 \leq 1$, then obviously $p(X_1 = 1, X_2 = 1) = 0$. Thus the constraints are mapped to marginal distributions: if $C_i(\mathbf{x}_{u_i})$ is violated then we set $p_i(x_{u_i}) = 0$.

We can now factorize $f(\mathbf{x})$ as before. But we can also factorize the graph defined by $C_i(x_{u_i})$. Our theory can handle the two cases: the factorization of the constraints is contained in the factorization of the function, i.e. $x_{u_i} \subseteq x_{s_i}$, or the factorization of the function is contained in the factorization of the constraints, i.e. $x_{s_i} \subseteq x_{u_i}$

Let Ω_c be the set of *feasible* solutions. Then the Boltzmann distribution on Ω_c is defined as

$$p_{b,f,c}(\mathbf{x}) = \frac{p_0(x)e^{\beta f(\mathbf{x})}}{\sum_{y \in \Omega_c} p_0(\mathbf{y})e^{\beta f(\mathbf{y})}} \tag{5.31}$$

Then the following convergence theorem holds.

Theorem 5.5 (Convergence). *Let 1) the initial population be feasible. Let 2) the factorization of thetechniques. constraints and the factorization of the function be contained in the FDA factorization. Let 3) $\Delta\beta(t)$ be an annealing schedule. Then for FDA the distribution at time t is given by*

$$p(\mathbf{x}, t) = \frac{p_0(\mathbf{x})e^{\beta(t)f(\mathbf{x})}}{\sum_{y \in \Omega_c} p_0(\mathbf{y})e^{\beta(t)f(y)}} \tag{5.32}$$

with the inverse temperature

$$\beta(t) = \sum_{\tau=1}^{t} \Delta\beta(\tau). \tag{5.33}$$

Let $\mathcal{M}$ be the set of global optima. If $\beta(t) \to \infty$, then

$$\lim_{t\to\infty} p(x, t) = \begin{cases} 1/|\mathcal{M}| & x \in \mathcal{M} \\ 0 & else \end{cases} \tag{5.34}$$

The proof is almost identical to the proof of Theorem 5.1. The factorization theorem needs an analytical description of the function. But it is also possible to determine the factorization from the data sampled. This is described next.

6 Computing a Bayesian Network from Data

The FDA factorization is based on the decomposition of the fitness function. This has two drawbacks: first, the structure of the function has to be known. Second, for a given instance of the fitness function, the structure might not give the smallest factorization possible. In other words: complex structures are not necessarily connected to corresponding complex dependency structures for a given fitness function. The actual dependencies depend on the actual function values. This problem can be circumvented by computing the dependency structure from the data.

Computing the structure of a Bayesian network from data is called learning. Learning gives an answer to the question: *Given a population of selected points $M(t)$, what is a good Bayesian factorization fitting the data?* The most difficult part of the problem is to define a quality measure also called scoring measure.

A Bayesian network with more arcs fits the data better than one with less arcs. Therefore a scoring metric should give the best score to the minimal Bayesian network which fits the data. It is outside the scope of this paper to discuss this problem in more detail. The interested reader is referred to the two papers by Heckerman and Friedman et al. in [Jor99].

For Bayesian networks two quality measures are most frequently used - the *Bayes Dirichlet* (BDe) score and the *Minimal Description Length* (*MDL*) score. We concentrate on the *MDL* principle. This principle is motivated by universal coding. Suppose we are given a set D of instances, which we would like to store. Naturally, we would like to conserve space and save a compressed version of D. One way of compressing the data is to find a suitable model for D that the encoder can use to produce a compact version of D. In order to recover D we must also store the model used by the encoder to compress D. Thus the total description length is defined as the sum of the length of the compressed version of D and the length of the description of the model. The *MDL* principle postulates that the optimal model is the one that minimizes the total description length.

6.1 LFDA - Learning a Bayesian Factorization

In the context of learning Bayesian networks, the model is a network B describing a probability distribution p over the instances appearing in the data. Several authors have approximately computed the *MDL* score. Let $M = |D|$ denote the size of the data set. Then *MDL* is approximately given by

$$\mathit{MDL}(B,D) = -\operatorname{ld}(P(B)) + M \cdot H(B,D) + \tfrac{1}{2} PA \cdot \operatorname{ld}(M) \tag{6.1}$$

with $\operatorname{ld}(x) := \log_2(x)$. $P(B)$ denotes the prior probability of network B, $PA = \sum_i 2^{|pa_i|}$ gives the total number of probabilities to compute. $H(B,D)$ is defined by

$$H(B,D) = -\sum_{i=1}^{n} \sum_{pa_i} \sum_{x_i} \frac{m(x_i,pa_i)}{M} \operatorname{ld} \frac{m(x_i,pa_i)}{m(pa_i)} \tag{6.2}$$

where $m(x_i,pa_i)$ denotes the number of occurrences of x_i given configuration pa_i. $m(pa_i) = \sum_{x_i} m(x_i,pa_i)$. If $pa_i = \emptyset$, then $m(x_i,\emptyset)$ is set to the number of occurrences of x_i in D.

The formula has an interpretation which can be easily understood. If no prior information is available, $P(B)$ is identical for all possible networks. For minimizing, this term can be left out. $0.5PA \cdot \operatorname{ld}(M)$ is the length required to code the parameter of the model with precision 1/M. Normally one would need $PA \cdot \operatorname{ld}(M)$ bits to encode the parameters. However, the central limit theorem says that these frequencies are roughly normally distributed with a variance of $M^{-1/2}$. Hence, the higher $0.5 \operatorname{ld}(M)$ bits are not very useful and can be left out. $-M \cdot H(B,D)$ has two interpretations. First, it is identical to the logarithm of the maximum likelihood ($\operatorname{ld}(L(B|D))$). Thus we arrive at the following principle:

Choose the model which maximizes $\operatorname{ld}(L(B|D)) - \frac{1}{2}PA \cdot \operatorname{ld}(M)$.

The second interpretation arises from the observation that H(B,D) is the conditional entropy of the network structure B, defined by PA_i, and the data D. The above principle is appealing, because it has no parameter to be tuned. But the formula has been derived under many simplifications. In practice, one needs more control about the quality vs. complexity tradeoff. Therefore we use a weight factor α. Our measure is defined by BIC.

$$BIC(B,D,\alpha) = -M \cdot H(B,D) - \alpha PA \cdot \operatorname{ld}(M) \tag{6.3}$$

This measure with $\alpha = 0.5$ has been first derived by Schwarz [Sch78] as *Bayesian Information Criterion.* Therefore we abbreviate our measure as $BIC(\alpha)$.

To compute a network B^* which maximizes BIC requires a search through the space of all Bayesian networks. Such a search is more expensive than to search for the optima of the function. Therefore the following greedy algorithm has been used. k_{max} is the maximum number of incoming edges allowed.

```
                    $\mathbf{BN}(\alpha, \mathbf{k_{max}})$

• STEP 0: Start with an arc-less network.

• STEP 1: Add the arc $(x_i, x_j)$ which gives the maximum increase of BIC($\alpha$) if $|PA_j| \le k_{max}$ and adding the arc does not introduce a cycle.

• STEP 2: Stop if no arc is found.
```

Checking whether an arc would introduce a cycle can be easily done by maintaining for each node a list of parents and ancestors, i.e. parents of parents etc. Then $(x_i \to x_j)$ introduces a cycle if x_j is ancestor of x_i.

The BOA algorithm of Pelikan [PGCP00] uses the BDe score. This measure has the following drawback. It is more sensitive to coincidental correlations implied by the data than the *MDL* measure. As a consequence, the BDe measure will prefer network structures with more arcs over simpler networks [Bou94].

Given the BIC score we have several options to extend FDA to $LFDA$ which learns a factorization. Due to limitations of space we can only show results of an algorithm which computes a Bayesian network at each generation using algorithm $BN(0.5, k_{max})$. FDA and $LFDA$ should behave fairly similar, if $LFDA$ computes factorizations which are in probability terms very similar to the FDA factorization. FDA uses the same factorization for all generations, whereas $LFDA$ computes a new factorization at each step which depends on the given data M.

We have applied $LFDA$ to many problems [MM99b]. The results are encouraging. The numerical result indicates that control of the weight factor α can substantially reduce the amount of computation. For Bayesian network we have not yet experimented with control strategies. We have intensively studied the problem in the context of neural networks [ZOM97].

7 The System Dynamics Approach to Optimization

We have shown that Wright's equations converge to some local optima of the fitness function at the boundary. We might ask ourselves: Why not using the difference equations directly, without generating a population? This approach is called the systems dynamics approach to optimization. We just discuss a few examples which are connected with our theory.

7.1 The Replicator Equation

In this section we investigate the relation between Wright's equation and a popular equation called *replicator equation.* Replicator dynamics is a standard model in evolutionary biology to describe the dynamics of growth and decay of a number of species under selection. Let $S = \{1, 2, \ldots, s\}$ be a set of species, p_i the frequency of species i in a fixed population of size N. Then the replicator equation is defined on a simplex $S^s = \{p : \sum p_i = 1, 0 \le p_i \le 1\}$

$$\frac{dp_i}{dt} = p_i(t)\left(f_i(\mathbf{p}) - \sum_{i=1}^{s} p_i(t)f_i(\mathbf{p})\right) \qquad (7.1)$$

f_i gives the fitness of species i in relation to the others. The replicator equation is discussed in detail in [HS98]. For the replicator equation a maximum principle can be shown.

Theorem 7.1. *If there exists a potential V with $\partial V/\partial p_i = f_i(\mathbf{p})$, then $dV/dt \geq 0$, i.e the potential V increases using the replicator dynamics.*

If we want to apply the replicator equation to a binary optimization problem of size n, we have to set $s = 2^n$. Thus the number of species is exponential in the size of the problem. The replicator equation can be used for small size problems only.

Voigt [Voi89] had the idea, to generalize the replicator equation by introducing continuous variables $0 \leq p_i(x_k) \leq 1$ with $\sum_k p_i(x_k) = 1$. Thus $p_i(x_k)$ can be interpreted as univariate probabilities. Voigt [Voi89] proposed the following discrete equation.

Definition 7.1. *The* Discrete Diversified Replicator Equation *DDRP is given by*

$$\begin{aligned} p_i(x_k)(t+1) &= p_i(x_k)(t) + p_i(x_k)(t) \\ & \frac{f_{ik}(\mathbf{p}) - \sum_{x_k} p_i(x_k) f_{ik}(\mathbf{p})}{\sum_{x_k} p_i(x_k) f_{ik}(\mathbf{p})} \end{aligned}$$

The name Discrete Diversified Replicator Equation was not a good choice. The DDRP is more similar to Wright's equation than to the replicator equation. This is the content of the next theorem.

Theorem 7.2. *If the average fitness $W(\mathbf{p})$ is used as potential, then Wright's equation and the Discrete Diversified Replicator Equation are identical.*

We recently discovered that Baum and Eagon [BE67] have proven a discrete maximum principle for certain instances of the DDRP.

Theorem 7.3 (Baum-Eagon). *Let $V(\mathbf{p})$ be a polynomial with nonnegative coefficients homogeneous of degree d in its variables $p_i(x_j)$ with $p_i(x_j) \geq 0$ and $\sum_{x_j} p_i(x_j) = 1$. Let $\mathbf{p}(t+1)$ be the point given by*

$$p_i(x_j, t+1) = \frac{p_i(x_j,t)\frac{\partial V}{\partial p_i(x_j)}}{\sum_{x_k} p_i(x_k)\frac{\partial V}{\partial p_i(x_k)}} \qquad (7.2)$$

The derivatives are taken at $\mathbf{p}(t)$. Then $V(\mathbf{p}(t+1)) > V(\mathbf{p}(t)$ unless $\mathbf{p}(t+1) = \mathbf{p}(t)$

Equation 7.2 is exactly the DDRP with a potential V. Thus the DDRP could be called the Baum-Eagon equation. From the above theorem the discrete maximum principle for Wright's equation follows by setting $V = W$ and $d = n$. Thus the potential is the average fitness, which is homogeneous of degree n.

7.2 Some System Dynamics Equations for Optimization

The theorem of Baum Eagon shows that both, Wright's equation and the DDRP, maximize some potential. This means that both equations can be used for maximization. But there is a problem: both equations are deterministic. For difficult optimization problems, there exists a large number of attractors, each with a corresponding attractor region. If the iteration starts at a point within the attractor region, it will converge to the corresponding attractor at the boundary. But if the iteration starts at points which lie at the boundary of two or more attractors, i.e on the separatrix, the iteration will be confined to the separatrix. The deterministic system cannot decide for one of the attractors.

$UMDA$ with a finite population does not have a sharp boundary between attractor regions. We model this behavior by introducing randomness. The new value $p_i(x_j, t+1)$ is randomly chosen from the interval

$$[(1-c)p'_i(x_j, t+1), (1+c)p'_i(x_j, t+1)]$$

$p'_i(x_j, t+1)$ is determined by the deterministic equation. c is a small number. For $c = 0$ we obtain the deterministic equation. In order to use the difference equation optimally, we do not allow the boundary values $p_i = 0$ or $p_i = 1$. We use $p_i = p_{min}$ and $p_i = 1 - p_{min}$ instead.

A second extension concerns the determination of the solution. All dynamic equations presented use variables, which can be interpreted as probabilities. Thus instead of waiting that the dynamic system converges to some boundary point, we terminate the iteration at a suitable time and generate a set of solutions. Thus, given the values for $p_i(x_j)$ we generate points x according to the $UMDA$ distribution $p(\mathbf{x}) = \prod_{i=1}^n p_i(x_i)$.

We can now formulate a family of optimization algorithms, based on difference equations (*DIFFOPT*).

DIFFOPT

- **STEP 0:** Set $t \Leftarrow 0$ and $p_i(x_j, 0) = 0.5$ Input p_{min}.
- **STEP 1:** Compute $p'_i(x_j, t+1)$ according to a dynamic difference equation. If $p'_i(x_j, t+1) < p_{min}$ then $p'_i(x_j, t+1) = p_{min}$. If $p'_i(x_j, t+1) > 1 - p_{min}$ then $p'_i(x_j, t+1) = 1 - p_{min}$
- **STEP 2:** Compute randomly $p_i(x_j, t+1)$ in the interval $(1-c)p'_i(x_j, t+1), (1+c)p'_i(x_j, t+1)$. Set $t \Leftarrow t+1$
- **STEP 3:** If termination criteria are not met, go to STEP 1.
- **STEP4:** Generate N solutions according to $p(\mathbf{x}, t) = \prod_{i=1}^n p_i(x_i, t)$ and compute $max f(\mathbf{x})$ and $argmax f(\mathbf{x})$

$DIFFOPT$ is not restricted to Wright's equation or DDRP. The numerical efficiency of $DIFFOPT$ needs additional study.

8 Three Royal Roads to Optimization

In this section we will try to classify the different approaches presented. Population search methods are based on two components at least – selection and reproduction with variation. In our research we have transformed genetic algorithms to a family of algorithms using search distributions instead of recombination/mutation of strings. The simplest algorithm of this family is the univariate marginal distribution algorithm $UMDA$.

Wright's equation describes the behavior of $UMDA$ using an infinite population and proportionate selection. The equation shows that $UMDA$ does *not* primarily optimize the *fitness function* $f(\mathbf{x})$, but the *average fitness* of the population $W(\mathbf{p})$ which depends on the continuous marginal frequencies $p_i(x_i)$. Thus the important landscape for population search is *not* the landscape defined by the fitness function $f(\mathbf{x})$, but the landscape defined by $W(\mathbf{p})$.

The two components of population based search methods — selection and reproduction with variation — can work on a microscopic (individual) or a macroscopic (population) level. The level can be different for selection and reproduction. It is possible to classify the different approaches according to the level the components work. The following table shows three classes of evolutionary algorithms, each with a representative member.

Algorithm	Selection	Reproduction
Genetic Algorithm	microscopic	microscopic
UMDA	microscopic	macroscopic
System Dynamics	macroscopic	macroscopic

A genetic algorithm uses a population of individuals. Selection and recombination is done by manipulating individual strings. $UMDA$ uses marginal distributions to create individuals. These are macroscopic variables. Selection is done on a population of individuals, as genetic algorithms do. In the system dynamics approach selection is modeled by a specific dynamic difference equation for macroscopic variables. We believe that a fourth class — macroscopic selection and microscopic reproduction — makes no sense.

Each of the approaches have their specific pros and cons. Genetic algorithms are very flexible, but the standard recombination operator has limited capabilities. $UMDA$ can use any kind of selection method which is also used by genetic algorithm. $UMDA$ be extended to an algorithm which uses a more complex factorization of the distribution. This is done by the factorized distribution algorithm FDA. Selection is very difficult to model on a macroscopic level. Wright's equation are valid for proportionate selection only. Other selection schemes lead to very complicated system dynamics equations.

Thus for proportionate selection and gene pool recombination all methods will behave similarly. But each of the methods allows extensions which cannot be modeled with an approach using a different level.

Mathematically especially interesting is the extension of $UMDA$ to FDA with an adaptive Boltzmann annealing schedule. For this algorithm convergence for a large class of discrete optimization problems has been shown.

8.1 Boltzmann Selection and the Replicator Equation

Wright's equation transforms the discrete optimization problem into a continuous one. Thus mathematically we can try to optimize $W(\mathbf{p})$ instead of $f(\mathbf{x})$. For FDA with Boltzmann selection we even have a closed solution for the probability $p(\mathbf{x},t)$. It is given by

$$p_{\beta,p_0}(\mathbf{x},t) = \frac{p_0(\mathbf{x})e^{\beta f(\mathbf{x})}}{\sum_y p_0(\mathbf{y})e^{\beta f(\mathbf{y})}} \tag{8.1}$$

If we differentiate this equation we obtain after some computation

$$\frac{dp_{\beta,p_0}(\mathbf{x},t)}{dt} = \frac{d\beta}{dt} p_{\beta,p_0}(\mathbf{x},t)\left(f(\mathbf{x}) - \sum_y p_{\beta,p_0}(\mathbf{y},t)f(y)\right) \tag{8.2}$$

For $\beta' = 1$ we obtain a special case of the replicator equation 7.1. We just have to set $f(\mathbf{p}) = f_i$.

Theorem 8.1. *The dynamics of Boltzmann selection with* $\Delta\beta(t) = 1$ *is given by the replicator equation.*

From the convergence theorem 5.1 we know that the global optima are the only stable attractors of the replicator equation. Thus the replicator equation is an ideal starting point for a system dynamics approach to optimization discussed in Section 7. Unfortunately the replicator equation consists of 2^n different equations for a problem of size n.

Thus we are lead to the same problem encountered when analyzing the Boltzmann distribution. We have to factorize the probability $p(\mathbf{x})$ if we want to use the equation numerically.

Example 8.1. *Linear function* $f(\mathbf{x}) = \sum_i^n \alpha_i x_i$. *In this case the* $UMDA$ *factorization is valid* $p(\mathbf{x}) = \prod_{i=1}^n p_i(x_i)$. *By summation we obtain from equation 8.2 after some manipulation*

$$\frac{dp_i}{dt} = \frac{d\beta}{dt} p_i(1-p_i)\alpha_i \tag{8.3}$$

For $\beta' = 1$ *this is just Wright's equation without the denominator* $\tilde{W}$.

if we extend this equation we obtain another proposal for the systems dynamics approach to optimization

$$\frac{dp_i}{dt} = \frac{d\beta}{dt} p_i(1-p_i)\frac{\partial \tilde{W}}{\partial p_i} \tag{8.4}$$

This equation needs further numerical studies. The speed of convergence can be controlled by setting β'. With

$$\frac{d\beta}{dt} = \frac{c}{\sigma} \tag{8.5}$$

an interesting alternative to Wright's equation is obtained. Further numerical studies are needed.

9 Conclusion and Outlook

This chapter describes a complete mathematical analysis of evolutionary methods for optimization. The optimization problem is defined by a fitness function with a given set of variables. Part of theory consists of an adaptation of classical population genetics and the science of breeding to optimization problems. The theory is extended to general population based search methods by introducing search distributions instead of doing recombination of strings. This theory can be also used for continuous variables, a mixture of continuous and discrete variables as well as constraint optimization problems. The theory combines *learning* and *optimization* into a common framework based on *graphical models*.

We have presented three approaches to optimization. We believe that the optimization methods based on search distributions (UMDA,FDA,LFDA) have the greatest optimization power. The dynamic equations derived for $UMDA$ with proportionate selection are fairly simple. For $UMDA$ with truncation or tournament selection and FDA with conditional marginal distributions, the dynamic equations can become very complicated. FDA with Boltzmann selection SDS is an extension of simulated annealing to a population of points. It shares with simulated annealing the convergence property, but convergence is much faster.

Ultimately our theory leads to a synthesis problem: finding a good factorization for a search distribution defined by a finite sample. This is a central problem in probability theory. One approach to this problem uses Bayesian networks. For Bayesian networks numerically efficient algorithms have been developed. Our $LFDA$ algorithm computes a Bayesian network by minimizing the Bayesian Information Criterion.

The computational effort of both FDA and $LFDA$ is substantially higher than that of $UMDA$. Thus $UMDA$ should be the first algorithm to be tried in a practical problem. Next the Multi-Factorization $LFDA$ should be applied.

Our theory is defined for optimization problems which are defined by quantitative variables. The optimization problem can be defined by a cost function or a complex process to be simulated. The theory is not applicable if either the optimization problem is qualitatively defined or the *problem solving method is non-numeric*. A popular example of a non-numeric problem solving method is *genetic programming*. In genetic programming we try to find a program which optimizes the problem, not an optimal solution. Understanding these kind of problem solving methods will be a challenge for the new decade.

Theoretical biology faces the same problem. The most succcessful biological model is the foundation of classical population genetics. It is based on Mendel's laws and a simple model of Darwinan selection. The model has also been used in his paper, it is purely quantitative. But this model is a too great simplification of natural systems. The organism is not represented in the model. As such the model has lead to Ultra-Darwinism with the concept of *selfish genes.* What is urgently needed is a model of the organism. This model has to incorporate all three aspects of organisms - structure, function, and development or *being, acting, and becoming.* Karl Popper has formulated it the following way: *The whole life is problem solving.* Gene frequencies cannot explain the many problem solving capabilities found in nature.

Bibliography

[AM94a] H. Asoh and H. Mühlenbein. Estimating the heritability by decomposing the genetic variance. In Y. Davidor, H.-P. Schwefel, and R. Männer, editors, *Parallel Problem Solving from Nature*, Lecture Notes in Computer Science 866, pages 98–107. Springer-Verlag, 1994.

[AM94b] H. Asoh and H. Mühlenbein. On the mean convergence time of evolutionary algorithms without selection and mutation. In Y. Davidor, H.-P. Schwefel, and R. Männer, editors, *Parallel Problem Solving from Nature*, Lecture Notes in Computer Science 866, pages 88–97. Springer-Verlag, 1994.

[Bal97] D.H. Ballard. *An Introduction to Natural Computation*. MIT Press, Cambridge, 1997.

[BE67] L.E. Baum and J.A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bull. Am. Math. Soc.*, 73:360–363, 1967.

[Bou94] R.R. Bouckaert. Properties of bayesian network learning algorithms. In R. Lopez de Mantaras and D. Poole, editors, *Proc. Tenth Conference on Uncertainty in Artificial Intelligence*, pages 102–109, San Francisco, 1994. Morgan Kaufmann.

[CF98] F.B. Christiansen and M.W. Feldman. Algorithms, genetics and populations: The schemata theorem revisited. *Complexity*, 3:57–64, 1998.

[dlMT93] M. de la Maza and B. Tidor. An analysis of selection procedures with particular attention paid to proportional and boltzmann selection. In S. Forrest, editor, *Proc. of the Fifth Int. Conf. on Genetic Algorithms*, pages 124–131, San Mateo, CA, 1993. Morgan Kaufman.

[Fal81] D. S. Falconer. *Introduction to Quantitative Genetics*. Longman, London, 1981.

[Fre98] B.J. Frey. *Graphical Models for Machine Learning and Digital Communication*. MIT Press, Cambrigde, 1998.

[Gei44] H. Geiringer. On the probability theory of linkage in mendelian heredity. *Annals of Math. Stat.*, 15:25–57, 1944.

[Gol89] D.E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, 1989.

[HCPGM99] G. Harik, E. Cantu-Paz, D.E. Goldberg, and B.L. Miller. The gambler's ruin problem, genetic algorithms, and the sizing of populations. *Evolutionary Computation*, 7:231–255, 1999.

[Hol92] J.H. Holland. *Adaptation in Natural and Artificial Systems*. Univ. of Michigan Press, Ann Arbor, 1975/1992.

[HS98] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, 1998.

[Jor99] M.I. Jordan. *Learning in Graphical Models*. MIT Press, Cambrigde, 1999.

[Lau96] St. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.

[MHF94] M. Mitchell, J.H. Holland, and St. Forrest. When will a genetic algorithm outperform hill climbing? *Advances in Neural Information Processing Systems*, 6:51–58, 1994.

[MM99a] H. Mühlenbein and Th. Mahnig. Convergence theory and applications of the factorized distribution algorithm. *Journal of Computing and Information Technology*, 7:19–32, 1999.

[MM99b] H. Mühlenbein and Th. Mahnig. FDA – a scalable evolutionary algorithm for the optimization of additively decomposed functions. *Evolutionary Computation*, 7(4):353–376, 1999.

[MM00] H. Mühlenbein and Th. Mahnig. Evolutionary algorithms: From recombination to search distributions. In L. Kallel, B. Naudts, and A. Rogers, editors, *Theoretical Aspects of Evolutionary Computing*, Natural Computing, pages 137–176. Springer Verlag, 2000.

[MMO99] H. Mühlenbein, Th. Mahnig, and A. Rodriguez Ochoa. Schemata, distributions and graphical models in evolutionary optimization. *Journal of Heuristics*, 5:215–247, 1999.

[MSV94] H. Mühlenbein and D. Schlierkamp-Voosen. The science of breeding and its application to the breeder genetic algorithm. *Evolutionary Computation*, 1:335–360, 1994.

[Müh97] H. Mühlenbein. The equation for the response to selection and its use for prediction. *Evolutionary Computation*, 5(3):303–346, 1997.

[MV96] H. Mühlenbein and H.-M. Voigt. Gene pool recombination in genetic algorithms. In J.P. Kelly and I.H Osman, editors, *Metaheuristics: Theory and Applications*, pages 53–62, Norwell, 1996. Kluwer Academic Publisher.

[Nag92] T. Nagylaki. *Introduction to Theoretical Population Genetics*. Springer, Berlin, 1992.

[PBS97] A. Prügel-Bennet and J.L. Shapiro. An analysis of a genetic algorithm for simple random ising systems. *Physica D*, 104:75–114, 1997.

[PGCP00] M. Pelikan, D.E. Goldberg, and E. Cantu-Paz. Linkage problem, distribution estimation, and bayesian network. *Evolutionary Computation*, 8:311–341, 2000.

[Rao73] C.R. Rao. *Linear Statisticcal Inference and Its Application*. Wiley, New York, 1973.

[RS99] L. M. Rattray and J.L. Shapiro. Cumulant dynamics in a finite population. *Theoretical Population Biology*, 1999. to be published.

[Sch78] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 7:461–464, 1978.

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[Voi89] H.-M. Voigt. *Evolution and Optimization*. Akademie-Verlag, 1989.

[Vos99] M. Vose. *The Simple Genetic Algorithm: Foundations and Theory*. MIT Press, Cambridge, 1999.

[Wri70] S. Wright. Random drift and the shifting balance theory of evolution. In K. Kojima, editor, *Mathematical Topics in Population Genetics*, pages 1–31, Berlin, 1970. Springer Verlag.

[ZOM97] Byoung-Tak Zhang, Peter Ohm, and Heinz Mühlenbein. Evolutionary induction of sparse neural trees. *Evolutionary Computation*, 5:213–236, 1997.