A Survey of Evolution Strategies

Thomas Bäck* **Frank Hoffmeister†** **Hans–Paul Schwefel‡**

University of Dortmund
Department of Computer Science XI
P.O. Box 50 05 00 · D–4600 Dortmund 50 · Germany

Abstract

Similar to *Genetic Algorithms*, *Evolution Strategies* (ESs) are algorithms which imitate the principles of natural evolution as a method to solve parameter optimization problems. The development of Evolution Strategies from the first mutation–selection scheme to the refined (μ,λ)–ES including the general concept of self–adaptation of the strategy parameters for the mutation variances as well as their covariances are described.

1 Introduction

The idea to use principles of organic evolution processes as rules for optimum seeking procedures emerged independently on both sides of the Atlantic ocean more than two decades ago. Both approaches rely upon imitating the collective learning paradigm of natural populations, based upon Darwin's observations and the modern synthetic theory of evolution.

In the USA Holland introduced *Genetic Algorithms* in the 60ies, embedded into the general framework of adaptation [Hol75]. He also mentioned the applicability to parameter optimization which was first realized in the work of De Jong [Jon75].

This article focuses on the German development called *Evolution Strategies* (ESs), introduced by Rechenberg at Berlin in the 60ies as well [Rec73], and further developed by Schwefel [Sch75b]. ESs were applied first to experimental optimization problems with more or less continuously changeable parameters only. The first numerical applications were performed by Hartmann [Har74] and Höfler [Höf76], and a first attempt towards extending this strategy in order to solve discrete or even binary parameter optimization problems was made by Schwefel [Sch75a].

*baeck@lumpi.informatik.uni-dortmund.de

†iwan@lumpi.informatik.uni-dortmund.de

‡uin005@ddohrz11.bitnet

The aim of this paper is to give an overview of the development of ESs, beginning with the first simple mutation–selection mechanism with two individuals per generation only and stopping at the (μ,λ)–ES as used nowadays on single processor computers. The important idea of the on–line adaptation of the strategy parameters during the search process by incorporating them into the genetic representation of the individuals is also explained.

First a short introduction to the basic terminology concerning the parameter optimization problem is given. The overall goal of a parameter optimization problem $f : M \subseteq \mathbb{R}^n \to \mathbb{R}$, $M \neq \emptyset$, where f is called the *objective function*, is to find a vector $x^* \in M$ such that:

$$\forall x \in M \ : \ f(x) \geq f(x^*) = f^* \tag{1}$$

where f^* is called a *global minimum*; x^* is the *minimum location (point or set).*

$$M = \{x \in \mathbb{R}^n \mid g_j(x) \geq 0 \ \forall j \in \{1, \ldots, q\}\} \tag{2}$$

is the set of feasible points for a problem with inequality constraints $g_j : \mathbb{R}^n \to \mathbb{R}$. For an unconstrained problem $M = \mathbb{R}^n$.

Since $\max\{f(x)\} = -\min\{-f(x)\}$, the restriction to minimization is without loss of generality. In the following a minimization problem is assumed. In general the optimization task is complicated by the existence of non–linear objective functions with multiple local optima. A *local minimum* $\hat{f} = f(\hat{x})$ is defined by the condition (3).

$$\exists \epsilon > 0 \ \forall x \in M \ : \ \|x - \hat{x}\| < \epsilon \Rightarrow \hat{f} \leq f(x) \tag{3}$$

Even if there is only one local optimum, it may be difficult to find a path towards it in case of discontinuities of the objective function or its derivatives. Simplification, e.g. linearization, may help to make things easier, but it can lead to results which are far away from the true optimum.

2 The Two Membered ES

According to Rechenberg [Rec73], the first efforts towards an evolution strategy took place in 1964 at the Technical University of Berlin (TUB). Then, the idea to imitate principles of organic evolution was applied in the field of experimental parameter optimization. The applications dealt with hydrodynamical problems like shape optimization of a bended pipe and of a flashing nozzle, or with control problems like the optimization of a PID regulator within a highly nonlinear system. Besides of simulating different versions of the strategy on the first available digital computer at the TUB, a Zuse Z23 [Sch65], computers soon were also used to solve numerical optimization problems by means of the first versions of simple ESs [Har74, Höf76].

The algorithm used in these applications was a simple mutation–selection scheme called *two membered* ES. It is based upon a "population" consisting of one parent individual (a real–valued vector), and one descendant, created by means of adding normally distributed random numbers. The better of both individuals then serves as the ancestor of the following iteration/generation. Such a (1+1)–ES can be described as the following 8–tuple:

$$(1{+}1)\text{–ES} = (P^0, m, s, c_d, c_i, f, g, t) \tag{4}$$

where

P^0	$=$	$(x^0, \sigma^0) \in I$	population $I = \mathrm{IR}^n \times \mathrm{IR}^n$
m	:	$I \rightarrow I$	mutation operator
s	:	$I \times I \rightarrow I$	selection operator
c_d, c_i	$\in$	IR	step-size control
f	:	$\mathrm{IR}^n \rightarrow \mathrm{IR}$	objective function
g_j	:	$\mathrm{IR}^n \rightarrow \mathrm{IR}$	constraint functions $j \in \{1, \ldots, q\}$
t	:	$I \times I \rightarrow \{0, 1\}$	termination criterion

P^0 denotes the initial "population" consisting of a single parent which produces by means of mutation a single offspring resulting in

$$\begin{aligned} P'^t &= (a_1'^t, a_2'^t) \in I \times I \\ a_1'^t &= P^t = (x^t, \sigma^t) \\ a_2'^t &= m(P^t) = (x'^t, \sigma^t) \end{aligned} \tag{5}$$

The mutation operator is applied to all components of the object parameter x^t. According to the biological observation that offspring are similar to their parents and that smaller changes occur more often than larger ones, mutation is realized by normally distributed random numbers:

$$x'^t = x^t + \mathbf{N_0}(\sigma^t) \tag{6}$$

where $\mathbf{N_0}$ denotes a vector of independent Gaussian random numbers with zero mean and standard deviations σ_i^t $(i = 1, \ldots, n)$. The selection operator then determines the fitter individual to become the parent of the next "generation":

$$P^{t+1} = s(P'^t) = \begin{cases} a_2'^t \text{ if } \begin{cases} f(x'^t) \leq f(x^t) \wedge \\ g_j(x'^t) \geq 0 \\ \forall j \in \{1, \ldots, q\} \end{cases} \\ a_1'^t = P^t \text{ else} \end{cases} \tag{7}$$

in case of minimization. The iteration process $P^t \rightarrow P^{t+1}$ stops when the termination criterion $t(a_1'^t, a_2'^t) = 1$ holds. Function t depends on the implementation and may utilize elapsed CPU time, elapsed number of generations, absolute or relative progress per generation, etc.

In the description presented so far the standard deviations $\sigma^t \in \mathrm{IR}^n$ remain constant over time. For theoretical considerations all components of σ^t are identical, i.e. $\forall\, i, j \in \{1, \ldots, n\} : \sigma_i^t = \sigma_j^t =: \sigma$. For a (1+1)–ES and a *regular* optimization problem a convergence property can be shown (see [Bor78]). The regularity of the optimization problem is specified by the criteria given in definition 1.

Definition 1

The optimization problem (1) is called *regular*, iff the following conditions are satisfied:

1. f is continuous.
2. M is a closed set.
3. $\forall \tilde{x} \in M \;:\; \tilde{M} := \{x \in M \mid f(x) \leq f(\tilde{x})\}$ is a closed set.
4. $\forall \epsilon > 0 \;:\; L^0_{f^*+\epsilon} := \mathrm{int}\,(L_{f^*+\epsilon}) \neq \emptyset$, where int denotes the set of all internal points and $L_{f^*+\epsilon} := \{x \in M \mid f(x) \leq f^* + \epsilon\}$ is a *level set* of f.

Then the global convergence of the (1+1)–ES can be shown:

Theorem 1

For $\sigma > 0$ and a regular optimization problem (1) with $f^* > -\infty$

$$p\left\{\lim_{t\rightarrow\infty} f(x^t) = f^*\right\} = 1$$

holds, i.e. the global optimum is found with probability 1 for sufficiently long search times.

The longish proof is omitted here; it can be found in [Bor78]. The basic idea is to use the monotone sequence $f(x^0) \geq f(x^1) \geq \ldots \geq f(x^i) \geq \ldots$ of objective function values generated by the process. Then, for the limit value $f^* \leq \tilde{f} = \lim_{t\rightarrow\infty} f(x^t)$ under the assumption $\tilde{f} > f^*$ a contradiction emerges, such that $\tilde{f} = f^*$ must be valid.

As is well known from similar theorems for *Simulated Annealing* [AK89] and *Genetic Algorithms* [EAH91], such results are not of much practical relevance due to the unlimited time condition. In fact, we are interested in the expectation of the *convergence rate* φ, which is given by the quotient of the distance covered towards the optimum and the number of trials needed for this distance.

Rechenberg calculated the convergence rates for the model functions

$$\begin{aligned} f_1(x) &= \mathrm{F}(x_1) = c_0 + c_1 x_1 \\ & \quad \forall i \in \{2,\ldots,n\} : -b/2 \le x_i \le b/2 \\ f_2(x) &= \sum_{i=1}^{n} x_i^2 \end{aligned} \tag{8}$$

where $x = (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathbb{R}^n$. f_1 is called the *corridor model* and represents a simple linear function with inequality constraints. Improvement of this objective function is only accomplished by moving along the first axis of the search space inside a corridor of width b. f_2 is called the *sphere model*. It comprises the simplest kind of non-linear, unimodal function. For these model functions the expectations of the rates of convergence are [Rec73]

$$\begin{aligned} \varphi_1 &= \frac{\sigma}{\sqrt{2\pi}} \left(1 - \sqrt{\frac{2}{\pi}} \frac{\sigma}{b}\right)^{n-1} \\ & \quad \text{for } n \gg 1 \\ \varphi_2 &= \frac{\sigma}{\sqrt{2\pi}} \left(\exp\left(-\left(\frac{n\sigma}{\sqrt{8}r}\right)^2\right)\right) \\ &- \frac{\sigma}{\sqrt{2\pi}} \left(\sqrt{\pi}\frac{n\sigma}{\sqrt{8}r}\left(1 - \mathrm{erf}\left(\frac{n\sigma}{\sqrt{8}r}\right)\right)\right) \\ & \quad \text{for } n \gg 1 \end{aligned} \tag{9}$$

where $\mathrm{erf}(x)$ refers to the well–known error function. The rate of convergence for the sphere model depends on the current location within the search space where r denotes the current euclidean distance from the optimum. From (9) it is possible to determine the optimum standard deviations σ_i^{opt} $(i = 1, 2)$ according to $\left.\frac{\mathrm{d}\varphi_i}{\mathrm{d}\sigma_i}\right|_{\sigma_i^{\mathrm{opt}};\varphi_i^{\max}} = 0$:

$$\begin{aligned} \sigma_1^{\mathrm{opt}} &= \sqrt{\frac{\pi}{2}} \cdot \frac{b}{n} \quad ; \quad \varphi_1^{\max} = \frac{1}{2e} \cdot \frac{b}{n} \\ \sigma_2^{\mathrm{opt}} &\approx 1.224 \cdot \frac{r}{n} \quad ; \quad \varphi_2^{\max} \approx 0.2025 \cdot \frac{r}{n} \end{aligned} \tag{10}$$

It should be noted, that in both cases the step size σ_i^{opt} is inversely proportional to the number of object variables n. Hence, the maximum rate of convergence is also inversely proportional to n.

The optimum standard deviations can be combined with the probabilities for a successful mutation:

$$\begin{aligned} p_1^t &= \frac{1}{2}\left(1 - \sqrt{\frac{2}{\pi}}\frac{\sigma}{b}\right)^{n-1} \quad \text{for } n \gg 1 \\ p_2^t &= \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{n\sigma}{\sqrt{8}r}\right)\right) \quad \text{for } n \gg 1 \end{aligned} \tag{11}$$

For optimum step sizes these probabilities turn to the values $p_1^{\mathrm{opt}} = 1/(2e) \approx 0.184$ and $p_2^{\mathrm{opt}} \approx 0.270$. From these findings Rechenberg postulated his *1/5 success rule*:

> *The ratio of successful mutations to all mutations should be 1/5. If it is greater than 1/5, increase the variance; if it is less, decrease the mutation variance.*

Though, in general, problems of interest may have characteristics different from those of the model functions used above, the following heuristic often helps to dynamically adjust the σ^t – not individually, but all at the same time, only. Hence, the mutation operator m is extended by the following equation:

$$\sigma^{t+n} = \begin{cases} c_d \cdot \sigma^t & , \text{ if } p_s^t < 1/5 \\ c_i \cdot \sigma^t & , \text{ if } p_s^t > 1/5 \\ \sigma^t & , \text{ if } p_s^t = 1/5 \end{cases} \tag{12}$$

where p_s^t is the frequency of successful mutations, measured e.g. over intervals of $10n$ trials. Schwefel [Sch81] gives reasons to use the factors $c_d = 0.82$ and $c_i = 1/0.82$ for the adjustment, which should take place every n mutations. It should be noted that with (6) and (12) the operator m consists of a random and a deterministic component, now.

As explained in [Sch81] the 1/5 success rule is a measure to increase the efficiency at the cost of effectiveness or robustness. It may lead the (1+1)–ES to premature termination even in the case of unimodal functions if there are discontinuities or active restrictions.

3 The First Multimembered ES

So far the population principle has not really been used. The (1+1)–ES can be designated as a kind of probabilistic gradient search technique – not, however, as a pure random or Monte Carlo method. To introduce the population concept into the algorithm, Rechenberg proposed the *multimembered* ES, where $\mu > 1$ parents can participate in the generation of one offspring individual. This has been denoted by Schwefel as $(\mu+1)$–ES and can be formalized this way:

$$(\mu+1)\text{–ES} = (P^0, \mu, r, m, s, c_d, c_i, f, g, t) \tag{13}$$

where

P^0	$=$	$(a_1^0, \ldots, a_\mu^0) \in I^\mu$	population
			$I = \mathbb{R}^n \times \mathbb{R}^n$
μ	$>$	1	number of parents
r	$:$	$I^\mu \to I$	recombination operator
m	$:$	$I \to I$	mutation operator
s	$:$	$I^{\mu+1} \to I^\mu$	selection operator

c_d, c_i	$\in$	$\mathbb{R}$	step-size control
f	:	$\mathbb{R}^n \to \mathbb{R}$	objective function
g_j	:	$\mathbb{R}^n \to \mathbb{R}$	constraint functions $j \in \{1, \dots, q\}$
t	:	$I^\mu \to \{0, 1\}$	termination criterion

With the introduction of μ parents instead of only one, the imitation of sexual reproduction is possible, which is provided by the additional recombination operator r:

$$
\begin{aligned}
r(P^t) = a' = (x', \sigma') \in I \quad x' \in \mathbb{R}^n \ , \ \sigma' \in \mathbb{R}^n \\
x'_i \;=\; \begin{cases} x_{a,i} \ , \ \mathcal{X} \leq 1/2 \\ x_{b,i} \ , \ \mathcal{X} > 1/2 \end{cases} \quad \forall i \in \{1, \dots, n\} \\
\sigma'_i \;=\; \begin{cases} \sigma_{a,i} \ , \ \mathcal{X} \leq 1/2 \\ \sigma_{b,i} \ , \ \mathcal{X} > 1/2 \end{cases} \quad \forall i \in \{1, \dots, n\}
\end{aligned}
\tag{14}
$$

where $a = (x_a, \sigma_a), b = (x_b, \sigma_b) \in I$ are two parents internally chosen by r. By convention all parents in a population have the same mating probabilities, i.e. the parents a and b are determined by uniform random numbers. $\mathcal{X}$ denotes a uniform random variable on the interval $[0, 1]$, and it is sampled anew for each component of the vectors x' and σ'. r is called a *discrete* recombination operator due to the fact that component values are just copied from on of the parents.

According to the saying "survival of the fittest" (which was not coined by Darwin, but by one of his antagonists) the selection operator s removes the least fit individual – may it be the offspring or one of the parents – from the population before the next generation starts producing a new offspring.

$$
\begin{aligned}
P'^t &= (a'^t_1, \dots, a'^t_{\mu+1}) = (a^t_1, \dots, a^t_\mu, m(r(P^t))) \\
P^{t+1} &= s(P'^t) \quad \text{such that } \forall\, a^{t+1}_i = (x, \sigma) \\
&\qquad \nexists a'^t_j = (x', \sigma') : f(x') < f(x)
\end{aligned}
\tag{15}
$$

The mutation operator m and the adjustment of σ^t is realized in the same manner as for a (1+1)–ES (4,5,12). Self–adaptation of the step sizes has not been possible within the $(\mu$+1)–ES scheme, since offspring with reduced mutation variances are always preferred.

4 $(\mu+\lambda)$–ES and (μ,λ)–ES

The motivation to extend the $(\mu$+1)–ES to a $(\mu+\lambda)$–ES and (μ,λ)–ES has been twofold [Sch77, Sch81]: first, to make use of (at that time futuristic) parallel computers, and secondly, to enable self–adaptation of strategic parameters like the (even n different) standard deviations of the mutations. Instead of changing the σ^t by an exogenous heuristic in a deterministic manner, Schwefel completely viewed σ^t as a part of the genetic information of an individual. Consequently, it is subject to recombination and mutation as well. Those individuals with better adjusted strategy parameters are expected to perform better. Thus, selection will favour them accordingly and sooner or later those individuals will dominate the population, i.e. a better parameter setting will emerge by means of self-adaptation.

As the nomenclature $(\mu+\lambda)$–ES suggests, μ parents produce λ offspring which are reduced again to the μ parents of the next generation. In all variants of $(\mu+\lambda)$–ES selection operates on the joined set of parents and offspring. Thus, parents survive until they are superseded by better offspring. It might be even possible for very well adapted individuals to survive forever. This feature gives rise to some deficiencies of a $(\mu+\lambda)$–ES:

1. On problems with an optimum moving over time a $(\mu+\lambda)$–ES gets stuck at an out-dated good location if the internal parameter setting becomes unsuitable to jump to the new field of possible improvements.
2. The same happens if the measurement of the fitness (objective) or the adjustment of the object variables are subject to noise, e.g. in experimental settings.
3. For a $(\mu+\lambda)$–ES with $\mu/\lambda \geq p^{opt}_{f(x)}$ (probability for a successful mutation) there is a deterministic selection advantage for those offspring which reduce some of their σ_i (9).

In order to avoid these effects Schwefel investigated the properties of a (μ,λ)–ES where only the offspring undergo selection, i.e. the life time of *every* individual is limited to one generation. The limited life span allows to forget inappropriate internal parameter settings. This may result in short phases of recession, but it avoids long stagnation phases due to misadapted strategy parameters [Sch87]. The $(\mu+\lambda)$–ES and (μ,λ)–ES fit into the same formal framework with the only difference being the limited life time of individuals in (μ,λ)–ES. Thus, only a formal description of a (μ,λ)–ES is presented here:

$$
(\mu,\lambda)\text{–ES} = (P^0, \mu, \lambda; r, m, s; \Delta\sigma; f, g, t) \tag{16}
$$

where

P^0	=	$(a^0_1, \dots, a^0_\mu) \in I^\mu$	population $I = \mathbb{R}^n \times \mathbb{R}^n$
μ	$\in$	$\mathbb{N}$	number of parents
λ	$\in$	$\mathbb{N}$	number of offspring $\lambda > \mu$
r	:	$I^\mu \to I$	recombination operator
m	:	$I \to I$	mutation operator
s	:	$I^\lambda \to I^\mu$	selection operator
$\Delta\sigma$	$\in$	$\mathbb{R}$	step-size meta-control
f	:	$\mathbb{R}^n \to \mathbb{R}$	objective function
g_j	:	$\mathbb{R}^n \to \mathbb{R}$	constraint functions $j \in \{1, \dots, q\}$
t	:	$I^\mu \to \{0, 1\}$	termination criterion

In case of a $(\mu+\lambda)$–ES the selection operator s is modified to $s : I^{\mu+\lambda} \rightarrow I^{\mu}$. The major difference to the ES variants given before results from the handling of the internal strategy parameter σ^t, which now is incorporated into the genetic information of an individual $a_i^t = (x^t, \sigma^t) \in I$ and which is not controlled by some meta–level algorithm like the 1/5 success rule anymore. As a result, the mutation operator is different from before:

$$\begin{array}{rcl} a_i'^t & = & r(P^t) \\ m(a_i'^t) & = & a_i''^t = (x''^t, \sigma''^t) \\ \sigma''^t & = & \sigma'^t \exp \mathbf{N_0}(\Delta\sigma) \\ x''^t & = & x'^t + \mathbf{N_0}(\sigma''^t) \end{array} \tag{17}$$

Mutation not only works on x^t but also on σ^t. The adjustment of σ^t by the 1/5 success rule has been replaced by a random modification, where unsuitable σ''^t are removed by means of selection. The operators r and s work as in (13).

Schwefel theoretically investigated the case of a $(1,\lambda)$–ES in a similar way as Rechenberg did before with respect to the (1+1)–ES. In particular, severe analytical problems arise as soon as one has to look for the stationary distribution of a population. More details are omitted here, but may be found in [Sch77, Sch81]. Like Rechenberg, Schwefel considered the corridor model and the sphere model (8). From an approximation of the maximum rates of convergence at optimum σ^t he deduced the corresponding best values for λ, the number of offspring per generation, for a $(1,\lambda)$–ES on a SISD computer with sequential evaluations of the offspring:

$$\begin{array}{rcl} \lambda_1 & = & 6 \\ \lambda_2 & = & 5 \end{array} \tag{18}$$

On average at least one out of λ_1 (λ_2) offspring represents an improvement of the objective function, i.e. the probability for a successful offspring is approximately $1/\lambda_1$ ($1/\lambda_2$). These values are pretty close to Rechenberg's 1/5 success rule. On a MIMD computer, a $(1,\lambda)$–ES outperforms the simple (1+1)–ES by far and it has the additional advantage to be able to escape from local optima as soon as μ is increased beyond only 1 while $\mu/\lambda = \text{const}$. Like for the (1+1)–ES the maximum rate of convergence is inversely proportional to n, the number of object variables [Sch77, Sch81].

ESs which are operated with an optimum ratio of μ/λ for a maximum rate of convergence, are biased towards local search. As a result, such ESs tend to reduce their genetic diversity, i.e. the number of different alleles (specific parameter settings) in a population, as soon as they are attracted by some local optimum. In order to avoid the effect of missing alleles Born [Bor78] proposed the concept of a *genetic load* for some kind of $(\mu+1)$–ES. With genetic load a population is made up of an initially fixed, constant sub-population and a dynamic sub-population which evolves over time as described before: $P^t = P_l \cup P_a^t$ where P_l denotes the genetic load and P_a^t refers to the evolving (active) sub-population. The genetic load may be used to introduce knowledge about suitable strategy parameters and object variables which are close to the optimum. Its major task is to maintain a minimum genetic diversity and a set of alleles that need not be learned by the algorithm itself. Recombination and mutation are extended and include the genetic load component as well as the dynamic part of the population. It is important to note that in this approach the genetic load is not subject to selection, i.e. no individual of the genetic load is replaced by offspring. For this case of a $(\mu+1)$–ES with genetic load Born could also prove the global convergence analogous to theorem 1 [Bor78].

4.1 Recombination Types

Intermediate recombination is motivated by the following Gedankenexperiment. When a population moves up-hill along a ridge or down-hill along a narrow valley, the individuals will have positions either on one or the other side of the ridge / ravine. In this case intermediate recombination of parents on different sides of the collective pathway (which is close to the gradient) may yield extraordinary successes.

$$\begin{array}{rcll} r(P^t) & = & a' = (x', \sigma') \in I & \\ x_i' & = & \frac{1}{2}(x_{a,i} + x_{b,i}) & i = 1, \ldots, n \\ \sigma_i' & = & \frac{1}{2}(\sigma_{a,i} + \sigma_{b,i}) & \end{array} \tag{19}$$

Again, $a = (x_a, \sigma_a)$ and $b = (x_b, \sigma_b)$ are two parents chosen by r. Unfortunately, this type of recombination tends to reduce the genetic diversity of the population, but on the other hand it is a measure to avoid over–adaptation, especially with respect to the strategy parameters. In ESs the model functions suggest that the achievable rate of progress is inversely proportional to the number of object variables, hence individuals moving in a subspace of the object variables can exhibit a temporarily larger convergence rate than those moving in the full space. These individuals will reach a relative optimum only, and the whole evolution process may stagnate there for a while.

By intermediate recombination, a random but unsuitable extinction of a mutation step size σ_i is always reverted (increased again) as long as there is no mate with a similar step size adaptation. Actually, Schwefel's implementation of $(\mu+\lambda)$–ES and (μ,λ)–ES contains five types of recombination [Sch81]:

$$r(P^t) = a' = (x', \sigma') \in I$$

$$x_i' = \begin{cases} x_{a,i} & \text{(A) no recombination} \\ x_{a,i} \text{ or } x_{b,i} & \text{(B) discrete} \\ \frac{1}{2}(x_{a,i} + x_{b,i}) & \text{(C) intermediate} \\ x_{a,i} \text{ or } x_{b_i,i} & \text{(D) global, discrete} \\ \frac{1}{2}(x_{a,i} + x_{b_i,i}) & \text{(E) global, intermediate} \end{cases} \tag{20}$$

where $a, b, b_i \in P^t$ are parents chosen by r. Note that with global recombination the mating partners for the recombination of a *single* component x'_i are chosen anew from the population resulting in a higher mixing of the genetic information than in the standard case (B). An experimental comparison of most traditional, commonly used, direct optimization strategies to ESs on a set of 50 test functions, including multimodal as well as unimodal ones, restricted as well as unrestricted ones, has shown rather good results for ESs. Their convergence rate is comparable to other algorithms, but their reliability as well as their chance to find a low–dimensional global optimum was remarkably better than for the other strategies compared [Sch81]. Best results were obtained with different recombination types for the object variables (discrete) and the strategy parameters (intermediate).

4.2 Correlated Mutations

In ESs mutation realizes a kind of hill–climbing search procedure (12), when it is considered in combination with selection. With dedicated σ_i for each object variable x_i preferred directions of search can be established only along the axes of the coordinate system. In general, the best search direction (the gradient) is *not* aligned on those axes. Thus, an optimum rate of progress is achieved only by chance when suitable mutations coincide, i.e. when they are correlated. Otherwise, the trajectory of the population through the search space is zigzagging along the gradient. In order to avoid this reduction of the rate of progress, Schwefel [Sch81] extended the mutation operator to handle *correlated mutations* which require an additional strategy vector θ.

$$\begin{aligned} m(a_i'^t) &= a_i''^t = (x'', \sigma'', \theta'') \in I\ ,\ I = \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^w \\ \sigma'' &= \sigma'^t \exp \mathbf{N_0}(\Delta\sigma) \\ \theta'' &= \theta'^t + \mathbf{N_0}(\Delta\theta) \\ x'' &= x'^t + \mathbf{N_0}(\mathbf{A}) \end{aligned} \tag{21}$$

where $\mathbf{N_0}$ denotes a vector of independent Gaussian random numbers with expectation zero and standard deviations $\Delta\sigma_i$ and $\Delta\theta_i$, respectively. $\mathbf{N_0}(\mathbf{A})$ refers to a normally distributed random vector z with expectation zero and probability density

$$p(z) = \sqrt{\frac{\det \mathbf{A}}{(2\pi)^{\mathrm{n}}}}\ \exp\left(-\frac{1}{2} z^{\mathrm{T}} \mathbf{A}\, z\right) \tag{22}$$

The diagonal elements of the covariance matrix $\mathbf{A}^{-1}$ are the independent variances $\sigma_i''^2$ (squares of the mutation step sizes) of the object variables x_i^t, while the off-diagonal elements represent the covariances $c_{i,j}$ of the mutations. Schwefel restricts the space of equal probability density to the surface of n-dimensional rotating hyperellipsoids, which are realized by a set of inclination angles $\theta'' \in \mathbb{R}^w$ of the main axes of the hyperellipsoid, $w = n(n-1)/2$. This helps to keep the covariance matrix positive definite. The standard deviations σ_i'' serve as a kind of mean step size along those axes.

Like the strategy parameter σ^t, θ^t is also incorporated into the genetic representation of an individual and is modified in a similar way, i.e. the recombination operator is extended to work on the inclination angles θ^t, as it has been done before for the mutation step size σ^t (12). This way the ES may adapt to any preferred direction of search by means of self-learning. Due to the additional meta-parameter $\Delta\theta$ the signature of an ES with correlated mutations is defined as

$$(\mu,\lambda)\text{–ES} = (P^0, \mu, \lambda; r, m, s; \Delta\sigma, \Delta\theta; f, g, t) \tag{23}$$

The step sizes σ_i of an individual make up an ellipsoid of equal probability density to place an offspring if these step sizes are applied to x^t of the individual itself. The left part of the illustration in figure 1 shows some individuals with their corresponding ellipsoids if the step sizes are *not* correlated (simple mutations), while the illustration on the right hand shows the same individuals with correlated mutations. In long, narrow valleys the step sizes with simple mutations must be smaller than with correlated mutations, where a single step size may reach far into the valley if it is oriented appropriately. In such situations the resulting rate of convergence is much higher.

5 Conclusion

Evolution Strategies went through a long period of stepwise development since the formulation of the basic ideas in the middle of the 60ies. Some important milestones in their development were

- the analytical convergence rate calculations, which led to the development of the 1/5 success rule;
- the introduction of a population instead of one single individual, which also allowed for the sexual recombination process;
- the self–learning process of strategy parameters by means of incorporating them into the set of genetically inherited variables;
- the (μ,λ)–ES, which introduces a forgetting principle and is important in changing environments as well as a measure against over–adaptation, especially of the strategy parameters;
- the introduction of additional strategy parameters to allow for correlated mutations and thus self–learning of simple "natural laws" in the topological environment.

The third and fifth points establish a two–level learning process, since not only the object–variable population adapts according to the response surface, but

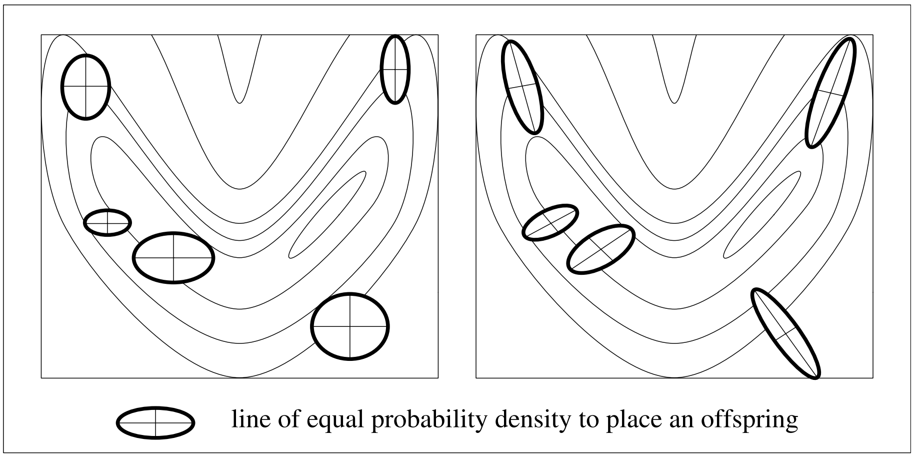

Figure 1: Searching with simple and correlated mutations

also the strategy parameters are changed with respect to the actual topolocigal requirements. Schwefel has pointed to the difficulties as well as opportunities of the two–level collective learning in [Sch87]. The strategy parameters make up an internal model of the objective function, which is learned on–line during the optimum seeking without any exogenous controlling instance or additional measure of fitness.

Besides of the different levels of genotypic/phenotypic information representation and of different selection mechanisms the two–level learning in ESs is the most striking difference between ESs and Genetic Algorithms [HB90, HB91].

Current research concerning ESs deals with applications like the travelling salesman problem [Her91, Rud91], girder–bridge optimization [Loh91], neural networks [Sal91], vector optimization [Kur91], and parameter optimization in general [BBK84, Bor78, BB79].

Furthermore the scalable parallelism of such an Evolutionary Algorithm [Hof91] is investigated with the help of implementations on MIMD–computers at the University of Dortmund, especially on a small Transputer network with 30 T800 processors [Bor89, Rud91].

References

[AK89] Emile H. L. Aarts and Jan Korst. *Simulated Annealing and Boltzmann Machines.* Wiley, Chichester, 1989.

[BB79] Joachim Born and Klaus Bellmann. *Numerische Parameteroptimierung in mathematischen Modellen mittels einer Evolutionsstrategie*, volume 18 of *Lecture Notes in Control and Information Sciences*, pages 157–167. Springer, Berlin, 1979.

[BBK84] U. Bernutat-Buchmann and J. Krieger. Evolution strategies in numerical optimization on vector computers. In Feilmeier, Joubert, and Schendel, editors, *Parallel Computing 83*, pages 99–105. Elsevier, Amsterdam, 1984.

[Bor78] Joachim Born. *Evolutionsstrategien zur numerischen Lösung von Adaptationsaufgaben.* Dissertation A, Humboldt-Universität, Berlin, 1978.

[Bor89] Andreas Bormann. *Parallelisierungsmöglichkeiten für direkte Optimierungsverfahren auf Transputersystemen.* Master thesis, University of Dortmund, Germany, April 1989.

[EAH91] A. E. Eiben, E. H. L. Aarts, and K. M. Van Hee. Global convergence of genetic algorithms: an infinite markov chain analysis. In Schwefel and Männer [SM91], pages 4–12.

[Har74] Dietrich Hartmann. *Optimierung balkenartiger Zylinderschalen aus Stahlbeton mit elastischem und plastischem Werkstoffverhalten.* PhD thesis, University of Dortmund, July 1974.

[HB90] Frank Hoffmeister and Thomas Bäck. Genetic algorithms and evolution strategies: Similarities and differences. Technical Report “Grüne Reihe” No. 365, Department of Computer Science, University of Dortmund, November 1990.

[HB91] Frank Hoffmeister and Thomas Bäck. Genetic algorithms and evolution strategies: Similarities and differences. In Schwefel and Männer [SM91], pages 455–470.

[Her91] Michael Herdy. Application of the ‘Evolutionsstrategie’ to discrete optimization problems. In Schwefel and Männer [SM91], pages 188–192.

[Höf76] A. Höfler. *Formoptimierung von Leichtbaufachwerken durch Einsatz einer Evolutionsstrategie.* PhD thesis, Technical University of Berlin, June 1976. Dept. Verkehrswesen.

[Hof91] Frank Hoffmeister. Parallel evolutionary algorithms. In Alexander N. Antamoshkin, editor, *Random Search as a Method for Adaptation and Optimization of Complex Systems*, pages 90–94, Divnogorsk, USSR, March 1991. Krasnojarsk Space Technology University.

[Hol75] John H. Holland. *Adaptation in natural and artificial systems.* The University of Michigan Press, Ann Arbor, 1975.

[Jon75] Kenneth De Jong. *An analysis of the behaviour of a class of genetic adaptive systems.* PhD thesis, University of Michigan, 1975. Diss. Abstr. Int. 36(10), 5140B, University Microfilms No. 76–9381.

[Kur91] Frank Kursawe. A variant of Evolution Strategies for vector optimization. In Schwefel and Männer [SM91], pages 193–197.

[Loh91] Reinhard Lohmann. Application of evolution strategy in parallel populations. In Schwefel and Männer [SM91], pages 198–208.

[Rec73] Ingo Rechenberg. *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution.* Frommann–Holzboog Verlag, Stuttgart, 1973.

[Rud91] Günter Rudolph. Global optimization by means of distributed evolution strategies. In Schwefel and Männer [SM91], pages 209–213.

[Sal91] R. Salomon. Improved convergence rate of back–propagation with dynamic adaptation of the learning rate. In Schwefel and Männer [SM91], pages 269–273.

[Sch65] Hans-Paul Schwefel. *Kybernetische Evolution als Strategie der experimentellen Forschung in der Strömungstechnik.* Diploma thesis, Technical University of Berlin, March 1965.

[Sch75a] Hans-Paul Schwefel. Binäre Optimierung durch somatische Mutation. Technical report, Technical University of Berlin and Medical University of Hannover, May 1975.

[Sch75b] Hans-Paul Schwefel. *Evolutionsstrategie und numerische Optimierung.* Dissertation, Technische Universität Berlin, May 1975.

[Sch77] Hans-Paul Schwefel. *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie*, volume 26 of *Interdisciplinary systems research.* Birkhäuser, Basel, 1977.

[Sch81] Hans-Paul Schwefel. *Numerical Optimization of Computer Models.* Wiley, Chichester, 1981.

[Sch87] Hans-Paul Schwefel. Collective phenomena in evolutionary systems. In *Preprints of the 31st Annual Meeting of the International Society for General System Research, Budapest*, volume 2, pages 1025–1033, June 1987.

[SM91] Hans-Paul Schwefel and Reinhard Männer, editors. *Parallel Problem Solving from Nature*, volume 496 of *Lecture Notes in Computer Science.* Springer, Berlin, 1991.