Toward a Theory of Evolution Strategies: Self-Adaptation

Hans-Georg Beyer,
University of Dortmund,
Department of Computer Science,
Systems Analysis Research Group,
D-44221 Dortmund,
Germany

beyer@LS11.informatik.uni-dortmund.de

Abstract

This paper analyzes the Self-Adaptation (SA) algorithm widely used to adapt strategy parameters of the Evolution Strategy (ES) in order to obtain maximal ES-performance. The investigations are concentrated on the adaptation of one general mutation strength σ (called σSA) in $(1, \lambda)$ ESs. The hypersphere serves as the fitness model.

Starting from an introduction into the basic concept of self-adaptation, a framework for the analysis of σSA is developed on two levels: a microscopic level concerning the description of the stochastic changes from one generation to the next, and a macroscopic level describing the evolutionary dynamics of the σSA over the time (generations). The σ-SA requires the fixing of a new strategy parameter, the so-called learning parameter. The influence of this parameter on the ES performance is investigated and rules for its tuning are presented and discussed. The results of the theoretical analysis are compared with ES experiments and it will be shown that applying Schwefel's τ-scaling rule guarantees the linear convergence order of the ES.

Keywords

ES-adaptation rules, evolutionary dynamics, linear convergence order, mutative step size control, self-adaptation

1 Introduction

The power of an Evolution Strategy (ES) is mainly based upon its ability to perform a 'second order' evolutionary process. This process adapts internal strategy parameters, especially the mutation strength, in such a way that the whole algorithm exhibits near optimal performance. That is, an ES *drives itself* into its optimal working regime.

There are different possibilities of constructing such strategies. The simplest one is the *1/5 - success probability rule* (Rechenberg, 1973) theoretically analyzed in Beyer (1993). Many scientists do not regard this rule of thumb as an adequate representation of what the spirit of evolutionary strategies really is. It simply adapts the mutation strength by a deterministic and rigid control mechanism. Therefore state-of-the-art ESs do not use it anymore. However, it is still of certain theoretical interest.

To date, there are two techniques in common use that utilize the ES paradigm in order to tune strategy parameters for optimal ES performance. The σ-Self-Adaptation (σSA) developed by Schwefel (1974, 1995) and the meta-ES (mES), a hierarchically organized, population based ES involving isolation periods (Herdy, 1992)[1].

Conceptually, the mES is - simply speaking - the application of a meta level ES onto a function optimizing ES (lower level), interpreting the latter as a function of the strategy parameter(s) to be optimized. Though this seems to be a very straightforward concept, it contains open strategy parameters at the meta level such as the isolation period and the number of subpopulations. mES are up to now a playground for empirical studies and will continue to be as long as a theoretical treatment is unavailable due to lack of a general progress rate theory (for the differences between the *normal progress* and the *hard progress* definition, see Beyer (1994)).

This article is devoted to the other paradigm - the σSA-ES. Unlike the mES, σ-self-adapting ESs perform the adaptation of the strategy parameter(s) at the same hierarchy level. That is, the evolution of the strategy parameters is closely coupled with that of the object parameters. Learning acts on the level of the individuals and is a *local* process (if the lifetime of the individuals is restricted to one generation, as usual for (μ, λ) strategies). This simplifies the theoretical analysis considerably. Thus, it will be possible to investigate theoretically the influence of the *learning parameter* τ - the only new (and exogenous) parameter to be introduced in the next section.

This paper is organized as follows. Section 2 provides the basic σSA-ES algorithm. In section 3 the framework for the analysis of the σSA will be developed that splits into two parts: the evolutionary dynamics, describing the evolution over the time (generation number), and the analysis of the driving forces for this process - the microscopic aspects. This analysis is restricted to the $(1, \lambda)$ ES. However, it can be easily extended to multirecombinant intermediate $(\mu/\mu, \lambda)$ strategies (Beyer, 1995b).

Section 4 implements the ideas developed in the previous part concerning the microscopic aspects. The results will be used to build up the basis for the understanding of the evolutionary dynamics to be developed in section 5. There, two models (with different approximation quality) of the σSA process will be investigated: the 'first order dynamics' which neglects the fluctuations and the 'noisy iterated map' that incorporates fluctuations by assuming Gaussian noise.

The steady-state of the mutation strength $\sigma^\star$ is of special interest. Therefore, the influence of the learning parameter τ on the steady-state is investigated as well as the transient time behavior. The question of optimal steady-state ES-performance leads us to Schwefel's $\tau \propto 1/\sqrt{N}$-rule.

The results to be provided so far are valid for $(1, \lambda)$ strategies. However, some of them can be easily transferred to recombinant ESs. In the concluding section 6 an idea will be presented on how the σSA process in recombinant ESs could be explained and how the performance could be further improved.

For this paper, it is expected that the reader is familiar with the basics of the ES theory. The notations used are in accord with former articles of the author (Beyer, 1994, 1995a, 1995b).

2 The σ-Self-Adaptation

2.1 General Aspects

The σSA has been developed by Schwefel (1974). In its most general form it can be found in Schwefel & Rudolph (1995). The key idea of σSA is the *individual* coupling of the object

[1] These two paradigms do not exclude each other. A mES may perform σSA and a σSA can include a lifetime mechanism allowing a variable lifespan of certain individuals (Schwefel & Rudolph, 1995).

parameters and of the (evolvable, endogenous) strategy parameters.[2] That is, each individual has its own set of strategy parameters. If an individual is selected with respect to its fitness, the corresponding strategy parameter set survives, too. Thus, the "strategy parameter genes", which are responsible for the mutation of the object parameters, drive the individuals (hopefully) into a regime of maximal quality gain (expected fitness change per generation, cp. Beyer (1994)).

The $(1, \lambda)$ algorithm we are going to analyze contains one evolvable strategy parameter[3] - the common mutation strength σ. The N-dimensional object parameter vector $\mathbf{y}$ to be optimized with respect to a given fitness $F(\mathbf{y})$ with optimum $\hat{F} = F(\hat{\mathbf{y}})$ is changed from generation (g) to $(g+1)$ by isotropic Gaussian mutations $\mathbf{Z}$

$$\tilde{\mathbf{y}}_l := \mathbf{y}^{(g)} + \mathbf{Z}, \qquad l = 1, \dots \lambda.$$

I. e., each of the N components z_i of $\mathbf{Z}$ is iid (independently, identically) normally distributed[4]

$$\mathbf{Z} := \sigma\ (\mathcal{N}(0,1),\ \dots\ \mathcal{N}(0,1))^T.$$

Since the mutation strength has to change over the generations, σ itself has to be subject to mutations

$$\tilde{\sigma}_l := \boldsymbol{\Xi}\left[\sigma^{(g)}\right]. \tag{1}$$

The $\boldsymbol{\Xi}[.]$-mutation operator performs *multiplicative* mutations in contrast to the additive $\mathbf{Z}$ of the object variables. This is because of the expected scaling behavior of the σ-change per generation which should be (on average) proportional to the actual σ-value. In order to see this, the reader is pointed to results from Beyer (1993, 1995a, 1995b): Optimal performance is achieved for a certain $\hat{\sigma}^\star$-value that maximizes the *normalized* progress rate $\varphi^\star = \varphi^\star(\sigma^\star)$. Let R be the expectation of the (local) radius of curvature, then this normalization reads

$$\sigma^\star := \sigma\frac{N}{R} \quad \text{and} \quad \varphi^\star := \varphi\frac{N}{R}. \tag{2}$$

If we assume an optimal working σSA, then for each generation

$$\hat{\sigma}^\star = \sigma^{(g)}\frac{N}{R^{(g)}} = \sigma^{(g+1)}\frac{N}{R^{(g+1)}} \tag{3}$$

should hold. Taking the definition of φ into account

$$\varphi := R^{(g)} - R^{(g+1)}, \tag{4}$$

one obtains from equations (3) and (2)

$$\sigma^{(g+1)} = \sigma^{(g)}\frac{R^{(g+1)}}{R^{(g)}} = \sigma^{(g)}\left(1 - \frac{\varphi^\star(\hat{\sigma}^\star)}{N}\right). \tag{5}$$

Thus, it becomes clear that σ should be changed by multiplicative mutations. This can be done by multiplication of the parental σ with a random number ξ. Equation (1) can be written as

$$\tilde{\sigma}_l := \xi\sigma^{(g)}, \quad l = 1, \dots \lambda. \tag{6}$$

[2]There are also exogenous strategy parameters that may be fixed during the evolution process. For example, usually the number of offspring λ is kept constant throughout the generations.

[3]For possible generalizations including more than one σ, see section 6.

[4]In this article $\mathcal{N}(0,1)$ denotes a Gaussian (standard) normal variate, i. e., $\overline{\mathcal{N}(0,1)} = 0$ and $\overline{(\mathcal{N}(0,1))^2} = 1$.

Procedure σSA-$(1,\lambda)$-ES;	line #
Begin	1
$g := 0$;	2
`initialize` $(\mathbf{y}^{(g)}, \sigma^{(g)})$	3
Repeat	4
For $l := 1$ **To** λ **Do Begin**	5
$\tilde{\sigma}_l := \xi \sigma^{(g)}$;	6
$\tilde{\mathbf{y}}_l := \mathbf{y}^{(g)} + \tilde{\sigma}_l \, (\mathcal{N}(0,1), \; \ldots \; \mathcal{N}(0,1))^T$;	7
$\tilde{F}_l := F(\tilde{\mathbf{y}}_l)$	8
End;	9
$l_p :=$ `selection`$_{1;\lambda}(\tilde{F}_1, \tilde{F}_2, \ldots \tilde{F}_\lambda)$;	10
$\sigma^{(g+1)} := \tilde{\sigma}_{l_p}$;	11
$\mathbf{y}^{(g+1)} := \tilde{\mathbf{y}}_{l_p}$;	12
$g := g + 1$;	13
Until stop-criterion	14
End	15

Table 1: Pseudo code of the σSA-$(1,\lambda)$-ES.

Since the expected σ-change is small (as indicated by (5), because $\frac{\varphi^\star(\hat{\sigma}^\star)}{N} \ll 1$), the expectation of ξ should *not* deviate too much from 1

$$E\{\xi\} \approx 1. \tag{7}$$

In section 5.1.1 a necessary condition will be derived that must be imposed on the $p_\sigma(\xi)$ distribution in order to be a suitable candidate for the σSA algorithm.

2.2 The σSA Algorithm

With the general statements made in the last section, it is not difficult to write down the pseudo code of the σSA algorithm displayed in Table 1.

As can be seen in Table 1, line 10, selection is performed according to "$1;\lambda$", i. e., the index value of the fittest offspring is determined. The corresponding $\tilde{\mathbf{y}}_l$-vector and the $\tilde{\sigma}_l$-value of this offspring serve as the parent for the new generation. Note, the "$1;\lambda$" and "$m;\lambda$" symbols are used to refer to the best and mth best individual, respectively, according to their fitness (we will make use of this notation, introduced in Beyer (1995b), in the theory part of this article).

2.3 How to Mutate the Mutation Strength

Though the condition (7) on ξ is not very restrictive, there are only a few $p_\sigma(\xi)$ distributions in practical use. The most prominent and widely used one is the *lognormal* distribution proposed by Schwefel in his pioneer work (1974).

$$p_\sigma(\xi) = \frac{1}{\sqrt{2\pi}\tau} \frac{1}{\xi} \mathrm{e}^{-\frac{1}{2}\left(\frac{\ln \xi}{\tau}\right)^2}. \tag{8}$$

ξ can be easily generated by sampling from the $\mathcal{N}(0,1)$ distribution and subsequent exponential transformation

$$\xi := \mathrm{e}^{\tau \mathcal{N}(0,1)}. \tag{9}$$

Given a certain $\sigma_0 = \sigma^{(g)}$, the output σ of the mutation rule (6) obeys the lognormal distribution

$$p_\sigma(\sigma) = \frac{1}{\sqrt{2\pi}\tau} \frac{1}{\sigma} \mathrm{e}^{-\frac{1}{2}\left(\frac{\ln(\sigma/\sigma_0)}{\tau}\right)^2} =: p_{\mathtt{ln}}(\sigma). \tag{10}$$

In (8 – 10), a new exogenous strategy parameter is introduced - the *learning parameter* τ. It determines how fast and accurate, respectively, the σSA is performed. How to choose τ will be a central theme and one of the goals of this paper.

Another simple mutation rule is given by the symmetrical two-point distribution[5]

$$p_\sigma(\sigma) = \frac{1}{2}\left[\delta(\sigma - \alpha\sigma_0) + \delta(\sigma - \sigma_0/\alpha)\right] =: p_{\mathtt{II}}(\sigma), \tag{11}$$

$$\alpha = 1 + \beta, \qquad 0 < \beta \lessapprox 0.3 \tag{12}$$

with the learning parameter α or β, respectively. Thus, an implementation of $\boldsymbol{\Xi}$, equation (1), could read (see, e. g. Rechenberg (1994), p. 47)

$$\sigma := \begin{cases} \sigma_0\,(1+\beta), & \text{if} \quad \mathtt{u}(0,1] \le 1/2 \\ \sigma_0/(1+\beta), & \text{if} \quad \mathtt{u}(0,1] > 1/2 \end{cases}, \tag{13}$$

with $\mathtt{u}(0,1]$ as a sampling from the random uniform $(0,1]$ distribution.

One can even generalize equation (11) to

$$p_\sigma(\sigma) = (1-\gamma)\,\delta(\sigma - (1+\beta_+)\sigma_0) + \gamma\,\delta(\sigma - (1-\beta_-)\sigma_0) =: p_\beta(\sigma) \tag{14}$$

by introduction of three exogenous learning parameters

$$0 < \beta_+, \qquad 0 < \beta_-, \qquad \text{and} \quad 0 < \gamma < 1.$$

This can open the possibility for a more refined tuning of the learning behavior of the σSA. However, its dynamical consequences will not be investigated here.

As a last mutation variant to be introduced, the one used in "meta-EP" (Fogel, 1992) should be mentioned. It can be easily obtained by Taylor expansion of (9) breaking off after the linear term

$$\xi := 1 + \tau\,\mathcal{N}(0,1). \tag{15}$$

Therefore, the newly generated σ-offspring is normally distributed

$$p_\sigma(\sigma) = \frac{1}{\sqrt{2\pi}\tau} \frac{1}{\sigma_0} \mathrm{e}^{-\frac{1}{2}\left(\frac{\sigma/\sigma_0 - 1}{\tau}\right)^2} =: p_{\mathcal{N}}(\sigma). \tag{16}$$

It will be an interesting result of the analysis to be presented that the effects of the mutation rules (9), (13), and (15) are comparable, if τ and β, respectively, become sufficiently small.

[5] Dirac's δ-function used, not to be mixed with the "SAR" (Self-Adaptation-Response) function $\delta^{(k)}(\varsigma^\star)$ to be introduced in section 3.1.2.

3 Basics of the Self-Adaptation Theory

In this section, the analysis of the σSA will be split into two parts. The first part (3.1) deals with the dynamics of the evolutionary process. That builds the framework for the second part (3.2) which is devoted to the driving forces of the ES-dynamics. We will call this the "microscopic" aspects of the σSA; whereas the "macroscopic" aspects refere to the ES-dynamics.

The analysis will be done for the spherical model; however, it holds for all fitness landscapes that can be approximated by a *local radius of curvature* as introduced in Beyer (1994) and generalized in Beyer (1995b).

3.1 The Evolution Dynamics

The state and the dynamics of the $(1, \lambda)$ ES can be definitely described by observation of the parent's local radius of curvature r and its inherited mutation strength ς. Note, in order to emphasize the stochastic character of the process, we use $(\varsigma, r)^T$ instead of $(\sigma, R)^T$. The latter will be reserved in order to describe the *mean value dynamics* of the ES, whereas the random variates ς and r describe the evolution in the "space of probability".

One may interpret the ES dynamics as a stochastic mapping or general Markov process

$$\left\{ \begin{matrix} \varsigma^{(g)} \\ r^{(g)} \end{matrix} \right\} \quad \mapsto \quad \left\{ \begin{matrix} \varsigma^{(g+1)} \\ r^{(g+1)} \end{matrix} \right\}.$$

The probability density of $(\varsigma, r)^T$ is governed by the so-called Chapman-Kolmogorov equation (see Fisz (1971)). However, the formulation of the theory in terms of Markov processes and Chapman-Kolmogorov equations at this stage is of really minor interest, because they cannot be solved analytically. Even the closed formulation of the integral equation describing the $\varsigma^\star$-evolution is excluded, due to the non-existence of an analytical expression for the $\varsigma^\star$ transition probability (cf. section 3.2.2). Therefore, we will take a step-by-step approximation procedure extracting the important features of the σSA process. The discussion concerning the Chapman-Kolmogorov equation is postponed to section 5.2 (cf. equations (93), (94)). Up to that point the approximations are to derive which will be the basis for the treatment of the stochastic dynamics.

3.1.1 The r-Evolution

Let the system at generation (g) be in the state $(\varsigma^{(g)}, r^{(g)})^T$. The transition to the new r state at generation $(g+1)$ can be described by

$$r^{(g+1)} = r^{(g)} - \varphi(\varsigma^{(g)}, r^{(g)}) + \epsilon_R(\varsigma^{(g)}, r^{(g)}). \tag{17}$$

Here, φ is the well-known progress rate describing the expected r-change under given parental conditions $(\varsigma^{(g)}, r^{(g)})^T$

$$\varphi(\varsigma^{(g)}, r^{(g)}) := E\left\{ r^{(g)} - r^{(g+1)} \mid \varsigma^{(g)}, r^{(g)} \right\} \tag{18}$$

and ϵ_R is the fluctuation term, necessarily obeying $E\left\{\epsilon_R(\varsigma^{(g)}, r^{(g)})\right\} = 0$. The ϵ_R-term carries the stochastics of the r-evolution. As a first approximation, it is assumed that ϵ_R is a Gaussian random variate

$$\epsilon_R = D_\varphi(\varsigma^{(g)}, r^{(g)})\, \mathcal{N}(0, 1) \tag{19}$$

with the conditional variance $D_\varphi^2 = E\left\{\epsilon_R^2 \,|\, \varsigma^{(g)}, r^{(g)}\right\}$

$$\begin{aligned} D_\varphi^2 &= E\left\{\left[-(r^{(g)} - r^{(g+1)}) + \varphi\right]^2\right\} \\ D_\varphi^2 &= E\left\{(r^{(g)} - r^{(g+1)})^2 \mid \varsigma^{(g)}, r^{(g)}\right\} \;-\; \varphi^2(\varsigma^{(g)}, r^{(g)}). \end{aligned} \tag{20}$$

The assumption (19) is not a critical one, because we are interested in the mean value dynamics $R = \overline{r}$, and ϵ_R influences R via the ς-dynamics only.

One main goal of the "microscopic" part of the theory will be the determination of the expectations from (18) and (20). With the definition of the *higher order progress rates* $\varphi^{(k)}$ ($\varphi^{(k)}$ not to be mixed with the generation counter $(.)^{(g)}$!)

$$\boxed{\varphi^{(k)}(\varsigma^{(g)}, r^{(g)}) := E\left\{(r^{(g)} - r^{(g+1)})^k \mid \varsigma^{(g)}, r^{(g)}\right\} = \int_{r=0}^{\infty} (r^{(g)} - r)^k \, p_{1;\lambda}(r \mid \varsigma^{(g)}, r^{(g)})\, dr,} \tag{21}$$

(17) and (20) can be written as

$$r^{(g+1)} = r^{(g)} - \varphi^{(1)}(\varsigma^{(g)}, r^{(g)}) + \epsilon_R(\varsigma^{(g)}, r^{(g)}) \tag{22}$$

and

$$D_\varphi^2 = \varphi^{(2)}(\varsigma^{(g)}, r^{(g)}) - \left(\varphi^{(1)}(\varsigma^{(g)}, r^{(g)})\right)^2. \tag{23}$$

The transition density $p_{1;\lambda}(r \mid \varsigma^{(g)}, r^{(g)})$ will be derived in section 3.2.2. Later on, it will be shown that $p_{1;\lambda}$ can be formulated with normalized quantities similar to (2)

$$\varphi^{\star(g)} := \varphi^{(g)} \frac{N}{r^{(g)}}, \quad \varsigma^{\star(g)} := \varsigma^{(g)} \frac{N}{r^{(g)}}, \quad \epsilon_R^\star := \epsilon_R \frac{N}{r^{(g)}}, \qquad \text{and} \quad D_\varphi^\star := D_\varphi \frac{N}{r^{(g)}}. \tag{24}$$

Applying the normalization to the evolution equation (17) yields

$$r^{(g+1)} = r^{(g)} \left(1 - \frac{1}{N}\, \varphi^\star(\varsigma^{(g)})\right) \;+\; \frac{r^{(g)}}{N}\, D_\varphi^\star(\varsigma^{(g)})\, \mathcal{N}(0,1). \tag{25}$$

This equation will be used to derive the R-value dynamics in section 5.

3.1.2 The ς-Evolution

The ς-evolution of the best offspring (which is also the parent for the next generation) can be expressed by the decomposition

$$\varsigma^{(g+1)} := \varsigma^{(g)} + \varsigma^{(g)}\, \delta^{(1)}(\varsigma^{(g)}, r^{(g)}) + \epsilon_\sigma(\varsigma^{(g)}, r^{(g)}). \tag{26}$$

This decomposition reflects the σ-perturbation mechanism in the σSA-algorithm (Table 1, line 6, see also equations (5 – 7)). It takes into account that the parental σ-value is only slightly changed by the multiplicative mutations. These mutations are around the value one. Intuitively one would expect (provided that the σSA does work) that the mutation which produces the best offspring should be greater than one, if the parental mutation strength σ is too low. And in the opposite case one would expect a mutation smaller than one. That is, given a certain parental σ, there is a "statistical tendency" of the σ change, that can be decomposed into its expected value $\delta^{(1)}$ and a fluctuation term ϵ_σ.

As in the case (17) it is postulated that the ϵ_σ-term carries the stochastics of the process and $E\left\{\epsilon_\sigma \mid \varsigma^{(g)}, r^{(g)}\right\} = 0$. Therefore, the $\delta^{(1)}$-function (not to be mixed with Dirac's δ) determines the expected relative σ-change

$$\delta^{(1)}(\varsigma^{(g)}, r^{(g)}) := E\left\{ \frac{\varsigma^{(g+1)} - \varsigma^{(g)}}{\varsigma^{(g)}} \,\middle|\, \varsigma^{(g)}, r^{(g)} \right\}. \tag{27}$$

The author will call $\delta^{(1)}$ the first order *Self-Adaptation-Response function (SAR)*. The generalization to kth order SARs is straightforward

$$\boxed{\delta^{(k)}(\varsigma^{(g)}, r^{(g)}) := \int_{\varsigma=0}^{\infty} \left(\frac{\varsigma - \varsigma^{(g)}}{\varsigma^{(g)}} \right)^k p_{1;\lambda}(\varsigma \mid \varsigma^{(g)}, r^{(g)})\, d\varsigma.} \tag{28}$$

Again, this integral contains a transition density $p_{1;\lambda}$ to be determined in section 3.2.2. It is interesting to notice that the SAR functions are invariant against the normalization (24). As will be shown in 3.2.3, $\delta^{(k)}$ depends on $\varsigma^{\star(g)}$ only

$$\delta^{(k)}(\varsigma^{(g)}, r^{(g)}) = \delta^{(k)}(\varsigma^{\star(g)}, N).$$

The fluctuation term ϵ_σ is approximated by Gaussian noise $\epsilon_\sigma = D_\sigma(\varsigma^{(g)}, r^{(g)})\, \mathcal{N}(0,1)$. Its variance $D_\sigma^2 = E\left\{\epsilon_\sigma^2 \mid \varsigma^{(g)}, r^{(g)}\right\}$ is obtained from (26)

$$\begin{aligned} D_\sigma^2 &= E\left\{ \left[(\varsigma^{(g+1)} - \varsigma^{(g)}) - \varsigma^{(g)}\delta^{(1)} \right]^2 \right\} = (\varsigma^{(g)})^2\, E\left\{ \left(\frac{\varsigma^{(g+1)} - \varsigma^{(g)}}{\varsigma^{(g)}} \right)^2 \right\} - (\varsigma^{(g)})^2 (\delta^{(1)})^2 \\ D_\sigma &= \varsigma^{(g)} \sqrt{\delta^{(2)}(\varsigma^{\star(g)}) - (\delta^{(1)}(\varsigma^{\star(g)}))^2} \quad =: \quad \varsigma^{(g)} D_\delta. \end{aligned} \tag{29}$$

In order to obtain the evolution equation for the normalized ς one has to express $\varsigma^{(g+1)}$ by the state $(\varsigma^{(g)}, r^{(g)})^T$. Taking (24) and (25) into account one gets for the left hand side of (26)

$$\varsigma^{(g+1)} = \varsigma^{\star(g+1)} \frac{r^{(g+1)}}{N} = \varsigma^{\star(g+1)}\, \frac{r^{(g)}}{N} \left[\left(1 - \frac{\varphi^\star}{N} \right) + \frac{D_\varphi^\star}{N} \mathcal{N}(0,1) \right]. \tag{30}$$

For the right hand side of (26) one finds

$$\varsigma^{(g+1)} = \varsigma^{\star(g)}\, \frac{r^{(g)}}{N} \left[(1 + \delta^{(1)}) + D_\delta\, \mathcal{N}(0,1) \right]. \tag{31}$$

The comparison of (30) with (31) yields[6]

$$\varsigma^{\star(g+1)} = \varsigma^{\star(g)}\, \frac{1 + \delta^{(1)}(\varsigma^{\star(g)}, N) + D_\delta(\varsigma^{\star(g)}, N)\, \mathcal{N}_\sigma(0,1)}{1 - \frac{1}{N}\varphi^\star(\varsigma^{\star(g)}, N) + \frac{1}{N} D_\varphi^\star(\varsigma^{\star(g)}, N)\, \mathcal{N}_\varphi(0,1)}. \tag{32}$$

This is a remarkable result showing that the $\varsigma^\star$-evolution is independent from the r-evolution. Equation (32) is decoupled from equation (25), whereas the r-evolution (25) is *governed* by the σ evolution (32).

[6] In order to indicate the independent random processes the $\mathcal{N}$ variates have been labeled with σ and φ.

Equation (32) can be simplified, if $\varphi^\star/N$ and $D^\star_\varphi/N$ are small compared to one. This usually holds for a strategy working with sufficiently small $\varsigma^\star$ values such that $\varphi^\star \gtrapprox 0$ holds. The opposite case can be only observed for the first few generations: An initially chosen too large $\varsigma^\star$ produces a large negative $\varphi^\star$. Thus, $\varsigma^\star$ is immediately reduced due to the large denominator in (32) (see also point 5.1.3). If $\varphi^\star$ is sufficiently small, then the Taylor expansion of the denominator yields for (32)

$$\varsigma^{\star(g+1)} = \varsigma^{\star(g)}\,\left(1+\delta^{(1)}+D_\delta\,\mathcal{N}_\sigma\right)\left(1+\frac{\varphi^\star}{N}-\frac{D^\star_\varphi}{N}\mathcal{N}_\varphi\right). \tag{33}$$

The further treatment of (33) will be a topic of section 5.

3.2 The Microscopic Aspects

The Goal of the "microscopics" is to derive expressions for the transition probability density of $r^{(g+1)}$ and $\varsigma^{(g+1)}$ and their first two moments, the progress rates $\varphi^{(1)}$, $\varphi^{(2)}$ and the SAR-functions $\delta^{(1)}$, $\delta^{(2)}$, respectively. The following subsections (3.2.1 and 3.2.2) are devoted to the transition probabilities and the formal definition of certain moments (in 3.2.3), whereas the calculation of the moments, especially the derivation of approximations, is postponed to section 4.

The aim of deriving the transition probabilities which describe the transition from the state $(\varsigma^{(g)}, r^{(g)})^T$ to the next generation is managed stepwise. At first, the single offspring distribution $p_{1;1}(r)$ will be derived (point 3.2.1). Then, this single offspring density is used to derive the densities for ς and r of the best offspring, $p_{1;\lambda}(r)$ and $p_{1;\lambda}(\varsigma)$, which are the densities of the next generation parent (therefore they describe the parental transition from g to $g+1$).

3.2.1 The Single Offspring Distribution $p_{1;1}(r)$

The first step concerns the density $p_{1;1}(r \mid \varsigma^{(g)}, r^{(g)})$ of a *single* offspring generated from the parental state $(\varsigma^{(g)}, r^{(g)})^T$. Due to the σSA algorithm, Table 1, line 6 & 7, the ς component is mutated first with density $p_\sigma(\varsigma \mid \varsigma^{(g)})$[7]. The ς produced serves as the mutation strength in line 7 (Table 1). There, an offspring with a new r is generated having the probability density $p_r(r \mid \varsigma, r^{(g)})$. This density has been derived in Beyer (1995a), it reads

$$p_r(r \mid \varsigma, r^{(g)}) = \frac{1}{\sqrt{2\pi}\tilde{\sigma}(\varsigma)} \exp -\frac{1}{2}\left(\frac{r-\sqrt{(r^{(g)})^2+\varsigma^2 N}}{\tilde{\sigma}(\varsigma)}\right)^2 \quad \text{with} \quad \tilde{\sigma}(\varsigma) := \varsigma\sqrt{\frac{(r^{(g)})^2+\frac{\varsigma^2 N}{2}}{(r^{(g)})^2+\varsigma^2 N}}. \tag{34}$$

Equation (34) can be interpreted as a pdf (probability density function) under condition ς generated with density $p_\sigma(\varsigma \mid \varsigma^{(g)})$. The integration over the ς-states yields

$$p_{1;1}(r \mid \varsigma^{(g)}, r^{(g)}) = \int_{\varsigma=0}^{\infty} p_r(r \mid \varsigma, r^{(g)})\, p_\sigma(\varsigma \mid \varsigma^{(g)})\, d\varsigma. \tag{35}$$

The cdf (cumulative distribution function) of (35), denoted by $P_{1;1}(r)$, is easily obtained

$$P_{1;1}(r) := P\left(\|\tilde{\mathbf{y}}_l\| < r \mid \varsigma^{(g)}, r^{(g)}\right) = \int_{\tilde{r}=0}^{\tilde{r}=r} p_{1;1}(\tilde{r} \mid \varsigma^{(g)}, r^{(g)})\, d\tilde{r}. \tag{36}$$

[7] Examples for p_σ are given by the equations (10), (11), (14), and (16) with $\sigma = \varsigma$ and $\sigma_0 = \varsigma^{(g)}$.

With (34) and (35) one gets from (36)

$$P_{1;1}(r) = P\left(\|\tilde{\mathbf{y}}_l\| < r \mid \varsigma^{(g)}, r^{(g)}\right) = \int_{\varsigma=0}^{\infty} \left[\frac{1}{2} + \Phi_0\left(\frac{r - \sqrt{(r^{(g)})^2 + \varsigma^2 N}}{\tilde{\sigma}(\varsigma)}\right)\right] p_\sigma(\varsigma \mid \varsigma^{(g)})\, d\varsigma. \quad (37)$$

Here $\Phi_0(.)$ denotes the Gauss-integral (cp. Beyer (1995a), equation (12)). As a standard approximation $\Phi_0\left(\sqrt{(r^{(g)})^2 + \varsigma^2 N}/\tilde{\sigma}(\varsigma)\right) \to \frac{1}{2}$ has been used. Equation (37) can be expressed by the normalized ς variate $\varsigma := \varsigma^\star r^{(g)}/N$ (cp. (24))

$$P_{1;1}(r) = P\left(\|\tilde{\mathbf{y}}_l\| < r \mid \varsigma^{\star(g)}, r^{(g)}\right) = \int_{\varsigma^\star=0}^{\infty} \left[\frac{1}{2} + \Phi_0\left(N\,\frac{r/r^{(g)} - \sqrt{1 + (\varsigma^\star)^2/N}}{\tilde{\sigma}^\star(\varsigma^\star)}\right)\right] p_\sigma^\star(\varsigma^\star \mid \varsigma^{\star(g)})\, d\varsigma^\star \quad (38)$$

where $\tilde{\sigma}^\star$ is obtained from (34)

$$\tilde{\sigma}^\star(\varsigma^\star) = \varsigma^\star \sqrt{\frac{1 + (\varsigma^\star)^2/2N}{1 + (\varsigma^\star)^2/N}} \quad (39)$$

and the 'p_σ-collection' given in section 2.3 is transformed to

$$p_\sigma^\star(\varsigma^\star \mid \varsigma^{\star(g)}) = \frac{1}{\sqrt{2\pi}\,\tau}\,\frac{1}{\varsigma^\star}\,\exp -\frac{1}{2}\left(\frac{\ln(\varsigma^\star/\varsigma^{\star(g)})}{\tau}\right)^2 =: p_{\mathtt{ln}}^\star(\varsigma^\star), \quad (40)$$

$$p_\sigma^\star(\varsigma^\star \mid \varsigma^{\star(g)}) = \frac{1}{2}\left[\delta\left(\varsigma^\star - \alpha\varsigma^{\star(g)}\right) + \delta\left(\varsigma^\star - \varsigma^{\star(g)}/\alpha\right)\right] =: p_{\mathtt{II}}^\star(\varsigma^\star), \quad (41)$$

$$p_\sigma^\star(\varsigma^\star \mid \varsigma^{\star(g)}) = (1-\gamma)\,\delta\left(\varsigma^\star - (1+\beta_+)\varsigma^{\star(g)}\right) + \gamma\,\delta\left(\varsigma^\star - (1-\beta_-)\varsigma^{\star(g)}\right) =: p_\beta^\star(\varsigma^\star), \quad (42)$$

$$p_\sigma^\star(\varsigma^\star \mid \varsigma^{\star(g)}) = \frac{1}{\sqrt{2\pi}\,\tau}\,\frac{1}{\varsigma^{\star(g)}}\,\exp -\frac{1}{2}\left(\frac{\varsigma^\star/\varsigma^{\star(g)} - 1}{\tau}\right)^2 =: p_{\mathcal{N}}^\star(\varsigma^\star). \quad (43)$$

3.2.2 The Transition Densities $p_{1;\lambda}(r)$ and $p_{1;\lambda}(\varsigma)$

Given the single offspring density $p_{1;1}(r \mid \varsigma^{(g)}, r^{(g)})$, it is a simple task to determine the r-transition density, i. e. the density of the best offspring out of λ (see Beyer (1993), p. 171)

$$\boxed{p_{1;\lambda}(r \mid \varsigma^{(g)}, r^{(g)}) = \lambda\, p_{1;1}(r \mid \varsigma^{(g)}, r^{(g)}) \left[1 - P_{1;1}(r \mid \varsigma^{(g)}, r^{(g)})\right]^{\lambda-1}.} \quad (44)$$

In general there is not a closed analytical expression for (44). This is because of the complicated integrand in (37), (38). Only the case of the two-point σ-mutation densities is tractable. The formula can be found in Beyer (1995c).

The ς-transition density $p_{1;\lambda}(\varsigma)$ is the ς-density of the best offspring, i. e. $p_{1;\lambda}(\varsigma) = p(\varsigma_{1;\lambda})$. However, $\varsigma_{1;\lambda}$ is not directly selected from the λ elements of $\{\varsigma_l\}$ (cf. lines 10/11 in Table 1). Let ς be a (one) random sample from the $p_\sigma(\varsigma \mid \varsigma^{(g)})$ density generated in line 6 of Table 1. In order to survive from the λ trials, the r-value generated by the mutation in line 6 (Table 1) has to be the best. This is fulfilled, if the r-value is smaller than the best (smallest) r-value of the remaining $(\lambda-1)$ trials which is denoted by $r_{1;\lambda-1}$. The probability for this may be denoted by $P\left[r<r_{1;\lambda-1} \mid \varsigma\right]$. Thus, the density for the ς-value of the best offspring reads

$$p_{1;\lambda}(\varsigma)=p_\sigma(\varsigma \mid \varsigma^{(g)})\, P\left[r<r_{1;\lambda-1} \mid \varsigma\right]. \tag{45}$$

The probability $P\left[r<r_{1;\lambda-1} \mid \varsigma\right]$ can be derived by the following arguments. Let r be the random variate of the (best) trial generated with the mutation strength ς. Its density $p_r(r \mid \varsigma, r^{(g)})$ is given by equation (34). To be the "best" means, that the remaining $(\lambda-1)$ trials are worse than r, i. e., $r<\|\tilde{\mathbf{y}}\|$ must hold (cf. line 7, Table 1). For a single trial, this occurs with probability $P(r<\|\tilde{\mathbf{y}}\|)=1-P(\|\tilde{\mathbf{y}}\| \leq r)$. There are $(\lambda-1)$ independent trials which are to fulfill the condition $r<\|\tilde{\mathbf{y}}\|$, therefore, the probability is given by $[1-P(\|\tilde{\mathbf{y}}\| \leq r)]^{\lambda-1}$. Thus, the density for a given, fixed r to be the best reads

$$p_r(r \mid \varsigma, r^{(g)})\ \ [1-P(\|\tilde{\mathbf{y}}\| \leq r)]^{\lambda-1}.$$

This density holds for each single constellation. Since there are λ equivalent but excluding possibilities of "being the best individual", the factor λ completes the density formula. Integration over the r-domain taking (36) into account yields for the probability of the best individual (given a fixed ς)

$$P\left[r<r_{1;\lambda-1} \mid \varsigma\right]=\lambda \int_{r=0}^{\infty} p_r(r \mid \varsigma, r^{(g)})\ \left[1-P_{1;1}(r \mid \varsigma^{(g)}, r^{(g)})\right]^{\lambda-1} dr.$$

Finally one obtains from (45) the desired transition density

$$\boxed{p_{1;\lambda}(\varsigma \mid \varsigma^{(g)}, r^{(g)})=\lambda p_\sigma(\varsigma \mid \varsigma^{(g)}) \int_{r=0}^{\infty} p_r(r \mid \varsigma, r^{(g)})\ \left[1-P_{1;1}(r \mid \varsigma^{(g)}, r^{(g)})\right]^{\lambda-1} dr.} \tag{46}$$

As can be seen, due to the complicated integrand in (46), analytical expressions for the $p_{1;\lambda}(\varsigma)$-transition density cannot be obtained. Not even for the $p_\sigma(\varsigma)=p_{\mathrm{II}}$ case, equations (11) and (41), if $\lambda>2$. Therefore one has to rely on approximations or on numerical calculations to be presented in section 4.

3.2.3 Progress Rates $\varphi^{(k)}$ and SAR-Functions $\delta^{(k)}$

From the definition (21) one obtains with the transition density (44)

$$\varphi^{(k)}(\varsigma^{(g)}, r^{(g)})=\lambda \int_{r=0}^{\infty}(r^{(g)}-r)^k p_{1;1}(r \mid \varsigma^{(g)}, r^{(g)})\left[1-P_{1;1}(r \mid \varsigma^{(g)}, r^{(g)})\right]^{\lambda-1} dr.$$

By the introduction of the normalization

$$\varphi^{\star(k)}:=\varphi^{(k)}\left(\frac{N}{r^{(g)}}\right)^k$$

$\varphi^{\star(k)}$ becomes

$$\varphi^{\star(k)}(\varsigma^{\star(g)})=N^k \lambda \int_{t=0}^{\infty}(1-t)^k\, p_{1;1}^{\star}(t \mid \varsigma^{\star(g)})\left[1-P_{1;1}^{\star}(t \mid \varsigma^{\star(g)})\right]^{\lambda-1} dt \tag{47}$$

with

$$P^{\star}_{1;1}(t \mid \varsigma^{\star(g)}) = \int_{\varsigma^\star=0}^{\infty} \left[\frac{1}{2} + \Phi_0\left(\frac{t - \sqrt{1+(\varsigma^\star)^2/N}}{\tilde{\sigma}^\star(\varsigma^\star)/N}\right)\right] p^\star_\sigma(\varsigma^\star \mid \varsigma^{\star(g)})\, d\varsigma^\star. \tag{48}$$

The further treatment of $\varphi^{\star(k)}$ will be done in section 4.1.

Now, let us derive the $\delta^{(k)}$-integrals. From the definition (28) we obtain with (46) after the change of the integration order

$$\delta^{(k)}(\varsigma^{(g)}, r^{(g)}) = \lambda \int_{r=0}^{\infty} \left[1 - P_{1;1}(r \mid \varsigma^{(g)}, r^{(g)})\right]^{\lambda-1} \int_{\varsigma=0}^{\infty} \left(\frac{\varsigma - \varsigma^{(g)}}{\varsigma^{(g)}}\right)^k p_r(r \mid \varsigma, r^{(g)})\, p_\sigma(\varsigma \mid \varsigma^{(g)})\, d\varsigma\, dr.$$

As in the case of the $\varphi^{(k)}$ functions, normalized quantities are to introduce. This gives with (38) and the substitution $r = r^{(g)}t$

$$\delta^{(k)}(\varsigma^{\star(g)}) = \lambda \int_{t=0}^{\infty} \left[1 - P^\star_{1;1}(t \mid \varsigma^{\star(g)})\right]^{\lambda-1} \int_{\varsigma^\star=0}^{\infty} \left(\frac{\varsigma^\star - \varsigma^{\star(g)}}{\varsigma^{\star(g)}}\right)^k p^\star_r(t \mid \varsigma^\star)\, p^\star_\sigma(\varsigma^\star \mid \varsigma^{\star(g)})\, d\varsigma^\star\, dt. \tag{49}$$

Here, $P^\star_{1;1}$ is given by (48), $p^\star_\sigma$, is taken from (40 – 43), and $p^\star_r$ is transformed from (34), i. e.,

$$p^\star_r(t \mid \varsigma^\star) = \frac{N}{\sqrt{2\pi}\tilde{\sigma}^\star(\varsigma^\star)} \exp -\frac{1}{2}\left(\frac{t - \sqrt{1+(\varsigma^\star)^2/N}}{\tilde{\sigma}^\star(\varsigma^\star)/N}\right)^2. \tag{50}$$

Closed analytical solutions to the progress rate integral (47), (48) and the SAR-integral (49), (50), (48) are hard to find. The only exception known to date is for $p_\sigma(\sigma) = p_{\mathrm{II}}(\sigma)$ (the two-point distribution (11), (41)) with $\lambda = 2$ (see Beyer (1995c)). All the more interesting variants have to rely on numerical integration techniques or approximate solutions holding for limited $\varsigma^\star$ parameter ranges. This will be discussed in the next section.

4 Determination of the Progress Rates and the SAR-Functions

This section is devoted to the calculation of certain moments of the transition densities describing the generational change $(\varsigma^{\star(g)}, r^{(g)})^T \mapsto (\varsigma^{\star(g+1)}, r^{(g+1)})^T$. This will be done by numerical integration techniques and analytical approximations as well.

Current ES-analyses would stop at the end of this section, because knowing $\varphi^\star$ provides the information how to choose the mutation strength $\sigma^\star$ for optimum ES-performance. However, the problem of tuning σ such that $\sigma^\star$ reaches the optimum working regime has been neglected. But in this work, the results of this section are only intermediary ones. They are the completion of the "microscopics" part which is the basis for the evolutionary dynamics to be investigated in section 5.

4.1 The Progress Rates $\varphi^{\star(k)}$

4.1.1 A Numerical Example for $\varphi^{\star(1)}$

In general, the calculation of the progress rates $\varphi^{\star(k)}$, given by equation (47) with equation (48), $p^\star_{1;1}(t) = \frac{d}{dt}P^\star_{1;1}(t)$, and the mutation density $p^\star_\sigma(\varsigma^\star)$ (equations (40 – 43)) can only be done by numerical integration techniques. As examples, first order progress rate graphs ($k = 1$) from the $(1, 10)$ ES with parameter space dimension $N = 30$ are displayed. In Fig. 1 the left picture

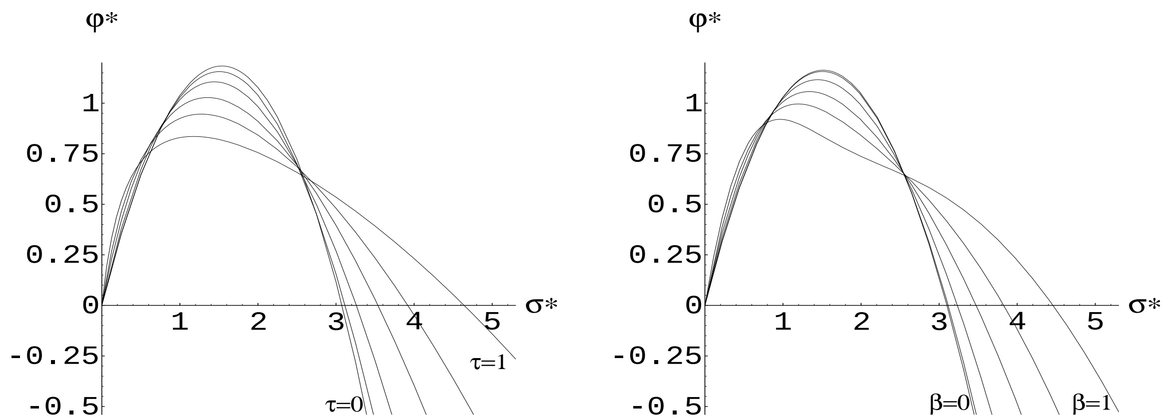

Figure 1: The (first order) progress rate $\varphi^{*(1)}(\varsigma^*)$ for the $(1,10)$-ES, $N = 30$, for different learning parameters τ and β, respectively. Left picture: The mutation strength σ is mutated by lognormal distributed random numbers. The learning parameters are $\tau = 0$, 0.1, 0.3, 0.5, 0.7, and 1.0. Right picture: The mutation strength σ is mutated by the symmetrical two-point distribution. Graphs for the learning parameter $\alpha := 1+\beta$ with $\beta = 0$, 0.1, 0.3, 0.5, 0.7, and 1.0 are displayed (the cases $\beta = 0$ and $\beta = 0.1$ are very close together).

refers to the lognormal distribution case (10), (40) $p^\star_\sigma(\varsigma^\star) = p^\star_{\mathtt{ln}}(\varsigma^\star)$, whereas the right picture is obtained for the symmetrical two-point distribution $p^\star_\sigma(\varsigma^\star) = p^\star_{\mathtt{II}}(\varsigma^\star)$ equations (11), (12), (41).

$\varphi^\star = \varphi^{\star(1)}$ is one of the central quantities of the ES-theory. Given a certain state $(\varsigma^{(g)}, r^{(g)})^T$, $\varphi^{(1)}$ predicts the expected (average) change Δr toward the optimum. Positive φ-values indicate (local) convergence. It is interesting to notice that the normalized progress rate $\varphi^\star$ depends on the $\varsigma^\star = \varsigma^{\star(g)}$-state only (there is no r-dependency). It is quite clear that the learning parameter τ (or β, respectively) influences the progress rate. From Figure 1 it can be seen that large learning parameters should be avoided though they provide a wider σ^*-range for convergence. As a rule of thumb

$$0 < \tau \overset{<}{\approx} 0.3; \qquad 0 < \beta \overset{<}{\approx} 0.3 \quad (\text{or } 1 < \alpha \overset{<}{\approx} 1.3)$$

may serve. For such parameter settings $\varphi^\star$ can be roughly approximated as if the learning parameter were zero. This special case has been already treated in Beyer (1995a)

$$\varphi^\star(\varsigma^\star)|_{\tau=0} = N\left(1 - \sqrt{1 + \frac{(\varsigma^\star)^2}{N}}\right) \; + \; c_{1,\lambda}\,\varsigma^\star \sqrt{\frac{1 + \frac{(\varsigma^\star)^2}{2N}}{1 + \frac{(\varsigma^\star)^2}{N}}} \tag{51}$$

($c_{1,\lambda}$ - progress coefficient, see Table 2 below). The construction of approximate solutions $\tau > 0$, however, remains a future task. (For the tractable case $p^\star_\sigma = p_{\mathtt{II}}$, $\lambda = 2$ see Beyer (1995c).)

One important remark should be added in order to avoid confusion about the message of Figure 1. Some readers might misinterpret the information from Figure 1 in a way that $\tau = \beta = 0$ should be the best choice, because it would appear to provide the largest possible $\varphi^\star$. However, this is a rather static picture which *does not* (and cannot) display the dynamic change of the actually fluctuating mutation strength $\varsigma^\star$. Using a $\tau > 0$ is necessary to keep pace with the

$\varsigma^\star$-evolution. To put it another way, $\tau = \beta = 0$ means the σSA is *switched off*! Under such a condition, the ES cannot converge at all, because the mutation strength σ remains constant (i. e. $\xi \equiv 1$, equation 6, section 2. 1). And, again, we would be left with the question how to control σ. In this sense, the analysis of a real ES-algorithm *must not* stop at the calculation of the 'microscopic' progress rate $\varphi^\star$. What is of interest here is the *expected* $\varphi^\star$ value. However, this requires the analysis of the whole ES dynamics (to be done in the rest of this paper). The calculation of $\varphi^\star$ is a necessary but "intermediate" step. But, there is even a grain of truth in the "$\tau = \beta = 0$-best-choice-misinterpretation": the curves in Figure 1 provide the upper bound of the ES-performance (concerning the $(1, \lambda)$-ES). From this point of view, τ should be chosen *as small as possible*. There are other aspects to be discussed in section 5 which demand the opposite choice. Thus, finding the right compromise is one goal of this paper.

4.1.2 The $\tau \to 0$ Approximation for $D^\star_\varphi$

The first order progress rate $\varphi^\star = \varphi^{\star(1)}$ predicts the expected r-change from generation (g) to $(g+1)$. Because of the random nature of the mutation process, the real r change fluctuates around its expectation (see (17), (22)). The variance of this fluctuations has been denoted by D^2_φ, equation (23), and the normalized standard deviation (see (24)) reads

$$D^\star_\varphi = \sqrt{\varphi^{\star(2)} - (\varphi^{\star(1)})^2}. \tag{52}$$

Since $\varphi^{\star(k)}$ depends on the learning parameter (τ, β, etc.) the standard deviation $D^\star_\varphi$, too, depends on it. In order to get manageable analytical expressions we will neglect this dependency. I. e., it will be assumed that $\tau, \beta \to 0$. In this case $\varphi^{\star(1)}$ is given by (51) and the remaining task concerns $\varphi^{\star(2)}$. With (47), (48) the integral to be calculated becomes (writing $s = \varsigma^{\star(g)}$)

$$\varphi^{\star(2)}(s) = \frac{N^2\lambda}{\sqrt{2\pi}} \int_{t=0}^{\infty} \frac{(1-t)^2}{\tilde{\sigma}^\star(s)/N} \exp -\frac{1}{2}\left(\frac{t-\sqrt{1+s^2/N}}{\tilde{\sigma}^\star(s)/N}\right)^2 \left[\frac{1}{2} - \Phi_0\left(\frac{t-\sqrt{1+s^2/N}}{\tilde{\sigma}^\star(s)/N}\right)\right]^{\lambda-1} dt.$$

After the substitution $x = -N\left(t - \sqrt{1+s^2/N}\right)/\tilde{\sigma}^\star(s)$ one obtains for sufficiently large N (and shifting the lower t limit to $-\infty$)

$$\varphi^{\star(2)}(s) = \frac{N^2\lambda}{\sqrt{2\pi}} \int_{t=-\infty}^{\infty} \left(1 - \sqrt{1+s^2/N} + \frac{\tilde{\sigma}^\star(s)}{N}\, t\right)^2 \mathrm{e}^{-\frac{1}{2}t^2} \left[\frac{1}{2} + \Phi_0(t)\right]^{\lambda-1} dt$$

and further

$$\varphi^{\star(2)}(s) = N^2 \left[\left(1 - \sqrt{1+s^2/N}\right)^2 + 2\left(1 - \sqrt{1+s^2/N}\right)\frac{\tilde{\sigma}^\star(s)}{N}\, c_{1,\lambda} + \left(\frac{\tilde{\sigma}^\star(s)}{N}\right)^2 d^{(2)}_{1,\lambda}\right]. \tag{53}$$

Apart from the well-known progress coefficient $c_{1,\lambda}$ (see Beyer (1995a)) a second integral parameter $d^{(2)}_{1,\lambda}$ - the so-called *second order progress coefficient* (Beyer, 1994) - is used. The $d^{(k)}_{1,\lambda}$-progress coefficients, introduced by the author, are defined as follows

$$\boxed{d^{(k)}_{1,\lambda} := \frac{\lambda}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^k \mathrm{e}^{-\frac{1}{2}t^2} \left[\frac{1}{2} + \Phi_0(t)\right]^{\lambda-1} dt.}$$

Table 2 gives a collection of $d^{(k)}_{1,\lambda}$-values. Now, $(D^\star_\varphi)^2$ can be completed. From (52), (51), and

λ	$c_{1,\lambda}$	$d^{(2)}_{1,\lambda}$	$d^{(3)}_{1,\lambda}$	$d^{(4)}_{1,\lambda}$	$d^{(5)}_{1,\lambda}$	$d^{(6)}_{1,\lambda}$	$\sqrt{d^{(2)}_{1,\lambda} - c^2_{1,\lambda}}$	$\varsigma^\star_{\delta_0}$
2	0.5642	1.0000	1.4105	3.0000	6.0650	15.000	0.8256	0.8862
3	0.8463	1.2757	2.1157	4.1945	9.0976	21.800	0.7480	0.9165
4	1.0294	1.5513	2.7004	5.3891	11.881	28.599	0.7012	1.0213
5	1.1630	1.8000	3.2249	6.5234	14.539	35.227	0.6690	1.1179
6	1.2672	2.0217	3.7053	7.5974	17.095	41.681	0.6449	1.2009
7	1.3522	2.2203	4.1497	8.6170	19.558	47.974	0.6260	1.2722
8	1.4236	2.3995	4.5636	9.5879	21.939	54.117	0.6106	1.3343
9	1.4850	2.5626	4.9512	10.515	24.243	60.119	0.5978	1.3890
10	1.5388	2.7121	5.3158	11.403	26.477	65.990	0.5868	1.4376
20	1.8675	3.7632	8.1298	18.736	45.876	118.92	0.5251	1.7474
30	2.0428	4.4187	10.097	24.326	61.677	164.28	0.4958	1.9183
40	2.1608	4.8969	11.629	28.915	75.215	204.52	0.4775	2.0349
50	2.2491	5.2740	12.892	32.841	87.166	240.97	0.4644	2.1227
60	2.3193	5.5856	13.970	36.292	97.929	274.48	0.4545	2.1928
70	2.3774	5.8512	14.914	39.382	107.76	305.61	0.4465	2.2509
80	2.4268	6.0827	15.755	42.187	116.84	334.77	0.4399	2.3005
90	2.4697	6.2880	16.514	44.762	125.28	362.26	0.4343	2.3436
100	2.5076	6.4724	17.207	47.145	133.20	388.32	0.4294	2.3817
200	2.7460	7.7015	22.077	64.733	194.31	597.55	0.4009	2.6225
300	2.8778	8.4310	25.164	76.580	237.80	754.07	0.3865	2.7559

Table 2: Progress coefficients of the $(1, \lambda)$-ES and "related parameters". The values have been obtained by numerical integration.

(53) one finally obtains the desired standard deviation (writing $\varsigma^\star = s$)

$$D^\star_\varphi(\varsigma^\star) = \tilde{\sigma}^\star(\varsigma^\star)\sqrt{d^{(2)}_{1,\lambda} - (c_{1,\lambda})^2}. \tag{54}$$

The square root in (54) is a slowly decreasing function of λ, as can be seen in Table 2. If (39) is taken into account, then it can be stated that $D^\star_\varphi(\varsigma^\star)$ is roughly proportional to the normalized (parental) mutation strength $\varsigma^\star$. Furthermore, the fluctuations decrease with increasing offspring number.

4.2 The SAR-Functions $\delta^{(k)}$

4.2.1 Numerical Examples and Experiments for $\delta^{(1)}$

The central quantity of the σSA theory is the SAR-function $\delta^{(1)}(\varsigma^\star)$, because it describes the expected relative σ change per generation, that is, the *Self-Adaptation Response* (SAR). The calculation of this function is difficult (the only analytical solution known is for $p^\star_\sigma = p_{\mathtt{II}}$, $\lambda = 2$, see Beyer (1995c)). As in the case of the progress rate one has to rely on numerical integration techniques applied to (49) and (48). Example graphs of $\delta^{(1)}$ are depicted in Figure 2. The ES under investigation is a $(1, 10)$ ES with $N = 30$. Two different σ-mutation policies are displayed,

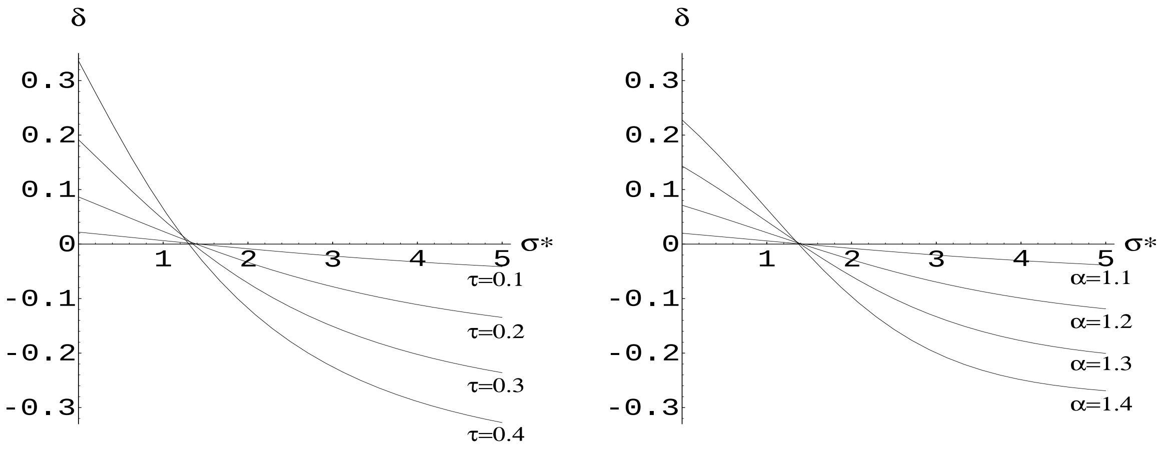

Figure 2: The (first order) self-adaptation response $\delta(\varsigma^*) = \delta^{(1)}(\varsigma^*)$ of a $(1, 10)$-ES with $N = 30$. Left picture: Lognormal σ-mutations with density $p^\star_{\mathtt{ln}}(\varsigma^\star)$ and learning parameter $\tau = 0.1$, 0.2, 0.3, and 0.4. Right picture: Symmetrical two-point σ-mutations with density $p^\star_{\mathtt{II}}(\varsigma^\star)$ and learning parameter $\alpha = 1 + \beta = 1.1$, 1.2, 1.3, and 1.4.

the lognormal and the two-point one. It is worth noticing that both policies produce similar SAR results. Furthermore, as will be shown in 4.2.3, for small τ and $\varsigma^\star$ both policies become equivalent.

From Figure 2 valuable information can be gathered. Due to the definition (26), $\delta^{(1)}$ describes the self-adaptative response of the σSA algorithm. This $\delta^{(1)}$ can be positive or negative depending on the actual (parental) $\varsigma^\star$-value. $\delta^{(1)}$ has one zero $\varsigma^\star_{\delta_0}$ (at least it should have one, and only one). If the parental $\varsigma^\star$ is greater than $\varsigma^\star_{\delta_0}$, then δ becomes negative, and thus, the expected offspring's ς-value will be decreased (on average, cp. equation 26). Whereas, in the opposite case, $\varsigma^\star < \varsigma^\star_{\delta_0}$, ς will be increased. Thus, the ς change can be "hopefully" focused on the $\varsigma^\star_{\delta_0}$-value. Now, if this zero $\varsigma^\star_{\delta_0}$ maximizes $\varphi^\star$, or at least produces positive φ-values, then the σSA will work properly. If we compare the pictures of Figure 1 and Figure 2 we can conclude that this seems to hold for the example chosen. However, the argumentation presented is a first approach only. Equation (26) describes the ς evolution which depends on the r evolution. In order to become independent from r we have to switch to the normalized quantities. This topic is deferred to section 5.

The influence of the learning parameters on δ can be observed in Figure 2. It becomes clear that τ and α, respectively, determine the "learning strength". For $\tau = 0$ (or $\alpha = 1$) there is no learning behavior at all, $\delta(\tau = 0) \equiv 0$ (or $\delta(\alpha = 1) \equiv 0$), because ς will not be mutated (cp. equation (6), (9) and (13), respectively). Generally, larger learning parameters provide a faster self-adaptation (see section 5). However, a glance at Figure 1 shows that the maximal achievable progress rate becomes smaller. There is always a trade-off between fast self-adaptation and maximal progress behavior.

Since the SAR-concept is new, it is indicated to compare its predictions with experiments. This has been done extensively. Figure 3 and Figure 4 show simulation results concerning the $(1, 10)$ ES with lognormal σ-mutations. For each data point $G = 40{,}000$ "one generation

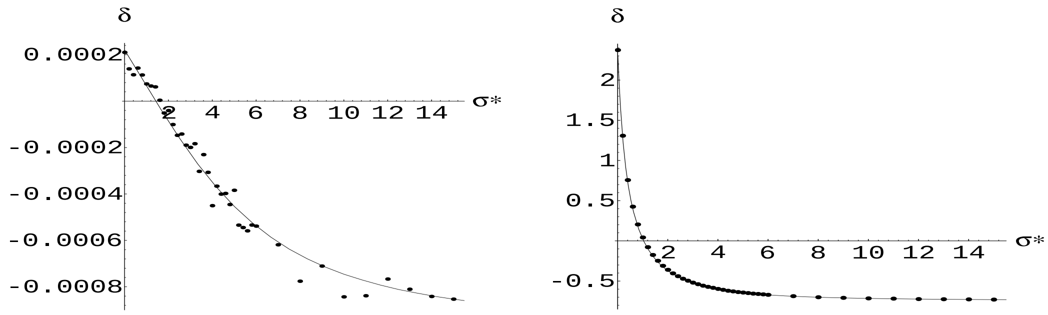

Figure 3: Comparison with experiments. Each dot represents a simulation result. Left picture: $\tau = 0.01$, $N = 30$ simulation. Right picture: $\tau = 1$, $N = 30$ simulation.

$(1, \lambda)$ experiments" have been performed. "One generation $(1, \lambda)$ experiment" refers to the algorithm, Table 1, without line 11–13. The mean value $\langle \tilde{\sigma}_{l_p} \rangle$ obtained by averaging $\tilde{\sigma}_{l_p}$ over G generations serves as an estimate for the computation of δ according to (27). In the right picture, $\tau = 1$, a very good agreement between experiment and theory is observed. In the case $\tau = 0.01$, left picture, fluctuations around the theoretical results are observed. These fluctuations can be theoretically explained. As will be shown in 4.2.3, the fluctuation around the mean is proportional to τ. Whereas δ, itself, scales with τ^2. Thus, for small τ values δ is dominated by the fluctuations.

The right picture of Fig. 3 reveals another characteristic of the $\delta^{(1)}$-function: It exhibits a (negative) saturation. For large $\varsigma^\star$-values δ becomes asymptotically constant $\lim_{\varsigma^\star \to \infty} \delta(\varsigma^\star) = \delta_\infty$. Generally, $\delta(\varsigma^\star) \geq -1$ is the lower bound on δ, independent of τ, λ, N, and ς. This is because of the definition (27). This bound can be sharpened depending on λ and τ (see Beyer (1995c)).

The saturation value δ_∞ depends on λ, N, and τ. The dependence on τ and N can be seen in Figure 4. There is a strong dependence for δ_∞ on N in case of smaller τ-values.

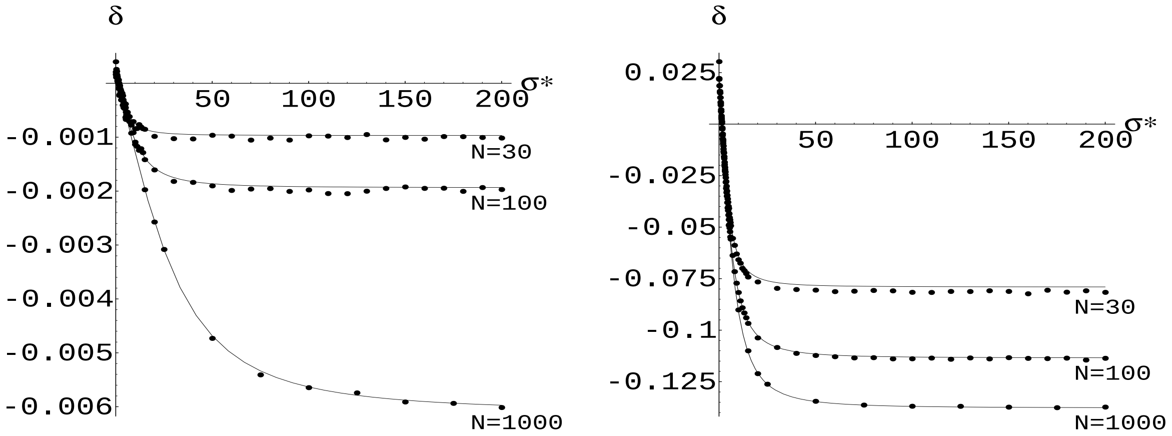

Figure 4: Further comparisons with simulations (displayed by dots) and investigations of the asymptotic behavior $\varsigma^\star \gg 0$ for different parameter space dimensions $N = 30,\ 100,\ 1000$. Left picture: lognormal σ-mutations with $\tau = 0.01$. Right picture: lognormal σ-mutations with $\tau = 0.1$.

4.2.2 The $\delta^{(k)}$-Approximations for Small $\varsigma^{\star(g)}$ and τ - General Aspects

For the investigation of the evolutionary dynamics, especially its steady-state behavior, the $\varsigma^\star \lessapprox 2c_{1,\lambda}$ range is of special interest, because it covers the positive $\varphi^\star$ values. In this range, there is only a slight dependency of $\delta^{(k)}$ and $\varphi^{\star(k)}$ on N. Furthermore, one can assume that $\delta^{(1)}$ can be approximated as a linear function of $\varsigma^\star$ controlled by τ (or β) (cp. Figures 2–4). The derivation of these approximations is cumbersome and rather long. Only a rough sketch can be presented here. For the details the reader is refered to Beyer (1995c). The approximation comprises several steps:

1. Under the condition $\varsigma^\star \ll N$, equations (48) and (50) are simplified by the assumption $\sqrt{1+(\varsigma^\star)^2/N} \approx 1 + (\varsigma^\star)^2/2N$ and $\tilde{\sigma}^\star(\varsigma^\star) \approx \varsigma^\star$. One obtains

$$P^\star_{1;1}(t \mid \varsigma^{\star(g)}) = \int_{\varsigma^\star=0}^{\infty} \left[\frac{1}{2} + \Phi_0\left(\frac{t-(1+(\varsigma^\star)^2/2N)}{\varsigma^\star/N}\right)\right] p^\star_\sigma(\varsigma^\star \mid \varsigma^{\star(g)})\, d\varsigma^\star. \tag{55}$$

$$p^\star_r(t \mid \varsigma^\star) = \frac{N}{\sqrt{2\pi}\varsigma^\star} \exp -\frac{1}{2}\left(\frac{t-(1+(\varsigma^\star)^2/2N)}{\varsigma^\star/N}\right)^2. \tag{56}$$

2. The next simplification concerns the dependency of $P^\star_{1;1}$, equation (55), on the learning parameter τ. One obtains

$$P^\star_{1;1}(t \mid \varsigma^{\star(g)}) = \frac{1}{2} + \Phi_0\left(\frac{t-(1+(\varsigma^{\star(g)})^2/2N)}{\varsigma^{\star(g)}/N}\right) \; + \; O(\tau^2). \tag{57}$$

It is interesting to notice that an analogous formula holds for the symmetrical two-point distribution $p^\star_{\mathrm{II}}(\varsigma^\star)$, as can be proved by Taylor expansion.

3. After the insertion of equation (57) and (56) into (49) and the substitution $x = -(t - (1 + s^2/2N))/(s/N)$ (writing $s = \varsigma^{\star(g)}$) one gets

$$\delta^{(k)} = \frac{\lambda}{\sqrt{2\pi}} \int_{x=-\infty}^{\infty} \left[\frac{1}{2} + \Phi_0(x)\right]^{\lambda-1} \int_{\varsigma^\star=0}^{\infty} \frac{s}{\varsigma^\star} \left(\frac{\varsigma^\star - s}{s}\right)^k h(\varsigma^\star, x, s)\, p_\sigma^\star(\varsigma^\star \mid s)\, d\varsigma^\star\, dx \qquad (58)$$

with

$$h(\varsigma^\star, x, s) := \exp -\frac{1}{2}\left(\frac{s}{\varsigma^\star}\, x \;+\; \frac{(\varsigma^\star)^2 - s^2}{2\varsigma^\star}\right)^2.$$

4. In order to approximate the inner integral of (58), h is expanded at $\varsigma^\star = s$

$$h(\varsigma^\star, x, s) = h(\varsigma^\star, x, s)|_{\varsigma^\star=s} \;+\; \left.\frac{\partial h}{\partial \varsigma^\star}\right|_{\varsigma^\star=s} (\varsigma^\star - s) \;+\; \frac{1}{2} \left.\frac{\partial^2 h}{\partial \varsigma^{\star 2}}\right|_{\varsigma^\star=s} (\varsigma^\star - s)^2 \;+\; \ldots. \qquad (59)$$

Again, the justification for this step comes from the assumed smallness of τ (or β, in case of $p_\sigma^\star = p_{\mathrm{II}}$) such that the probability mass of $p_\sigma^\star$ is concentrated around s. The inner integration over x yields

$$\begin{aligned}\delta^{(k)}(s) \;=\; & \overline{\frac{s}{\varsigma^\star}\left(\frac{\varsigma^\star - s}{s}\right)^k} + \left(d_{1,\lambda}^{(2)} - s c_{1,\lambda}\right) \overline{\frac{s}{\varsigma^\star}\left(\frac{\varsigma^\star - s}{s}\right)^{k+1}} \\ & + \frac{1}{2}\left[d_{1,\lambda}^{(4)} - 2s d_{1,\lambda}^{(3)} + (s^2 - 3) d_{1,\lambda}^{(2)} + 3 s c_{1,\lambda} - s^2\right] \overline{\frac{s}{\varsigma^\star}\left(\frac{\varsigma^\star - s}{s}\right)^{k+2}} + \ldots\end{aligned} \qquad (60)$$

with

$$\overline{\frac{s}{\varsigma^\star}\left(\frac{\varsigma^\star - s}{s}\right)^m} := \int_{\varsigma^\star=0}^{\infty} p_\sigma^\star(\varsigma^\star \mid s)\, \frac{s}{\varsigma^\star}\left(\frac{\varsigma^\star - s}{s}\right)^m d\varsigma^\star. \qquad (61)$$

The determination of (61) depends on the σ-mutation density $p_\sigma^\star$ used. In section 4.2.3 the $p_{\mathtt{ln}}^\star$ and $p_{\mathrm{II}}^\star$ variants for $k = 1$ and $k = 2$ will be investigated.

4.2.3 Approximations for $\delta^{(1)}$ and D_δ

The determination of $\delta^{(1)}$, $\delta^{(2)}$, and D_δ from (60), (61) is a straightforward task. Results on the σ mutation densities $p_{\mathtt{ln}}^\star$ and $p_{\mathrm{II}}^\star$ will be presented. The treatment of other densities, e. g. $p_\beta^\star$, equation (42), and $p_{\mathcal{N}}^\star$, equation (43), is omitted here.

For the lognormal case $p_\sigma^\star = p_{\mathtt{ln}}^\star(\varsigma^\star)$ one finally obtains from (60) (writing back $\varsigma^{\star(g)} = s$)

$$\boxed{p_\sigma = p_{\mathtt{ln}}(\sigma) \;: \qquad \delta_{1,\lambda}^{(1)}(\varsigma^{\star(g)}) = \tau^2 \left[\left(d_{1,\lambda}^{(2)} - \frac{1}{2}\right) - c_{1,\lambda}\varsigma^{\star(g)}\right] + O(\tau^4)} \qquad (62)$$

and

$$\delta_{1,\lambda}^{(2)}(\varsigma^{\star(g)}) = \tau^2 + O(\tau^4). \qquad (63)$$

With the D_δ definition (29) one further gets

$$D_\delta = \sqrt{\delta^{(2)} - (\delta^{(1)})^2} = \sqrt{\tau^2 + O(\tau^4)} = \tau + O(\tau^2). \qquad (64)$$

It is interesting to see that, within the approximations used, the self-adaptation response δ fluctuates with the order of the learning parameter τ.

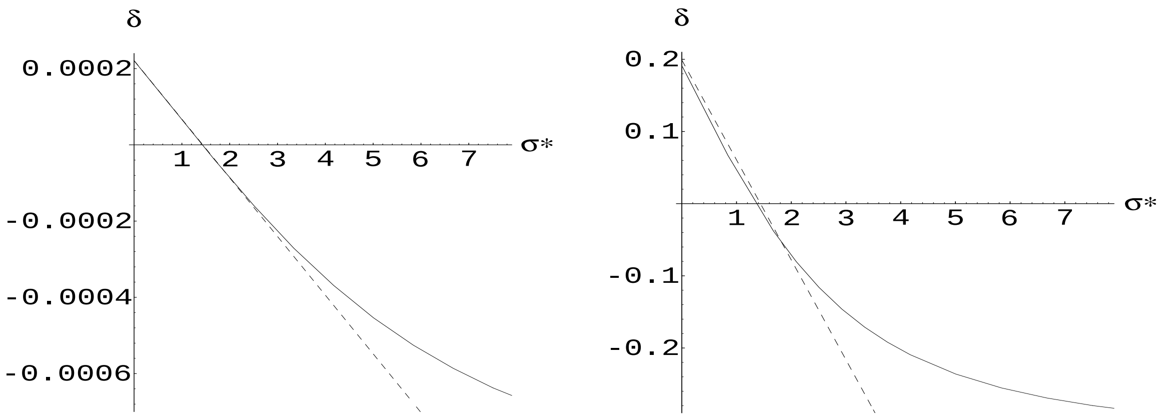

Figure 5: Comparison of the numerically obtained δ-SAR-function (solid curves) with the τ, $\varsigma^\star \to 0$ approximation (dashed lines) for the $(1, 10)$ ES, $N = 10$. Left picture: $\tau = 0.01$, right picture: $\tau = 0.3$.

As already pointed out, the results obtained hold for small τ ($\tau \lessapprox 0.3$) and small $\varsigma^\star$ values ($\varsigma^{\star(g)} \lessapprox 2c_{1,\lambda}$). The latter is mainly caused by the series break off in (59) and the $O(\tau^4)$ approximations for (61). Improvements are possible, however, the expressions become very bulky. In Figure 5 a comparison between the results of the numerical integration and the approximation (62) is displayed. Since (62) is a linear function of $\varsigma^{\star(g)}$ it is quite clear that the approximation has restricted validity. However the zero $\varsigma_0^\star$ of δ is satisfactorily predicted and up to $\varsigma^{\star(g)} \approx 2\varsigma_0^\star$ equation (62) should be applicable (see section 5).

Now, let us skip to the two-point distribution $p_\sigma^\star = p_{\mathtt{II}}^\star$. The calculations which can be found in Beyer (1995c) yield

$$\boxed{p_\sigma = p_{\mathtt{II}}(\sigma) \ : \qquad \delta_{1,\lambda}^{(1)}(\varsigma^{\star(g)}) = \beta^2(1-\beta)\left[\left(d_{1,\lambda}^{(2)} - \frac{1}{2}\right) - c_{1,\lambda}\varsigma^{\star(g)}\right] + O(\beta^4)}$$

and

$$\delta_{1,\lambda}^{(2)}(\varsigma^{\star(g)}) = \beta^2(1-\beta) + O(\beta^4).$$

Comparing these two formulae for the two-point distribution $p_{\mathtt{II}}$ with those of the lognormal distribution $p_{\mathtt{ln}}$ (62, 63) reveals a remarkable correspondence: If

$$\boxed{p_{\mathtt{ln}} \leftrightarrow p_{\mathtt{II}} \ \text{correspondence:} \qquad \tau^2 = \beta^2(1-\beta),} \tag{65}$$

then both σ mutation policies become equivalent with respect to the first and second order SAR-functions. Thus, results obtained for the lognormal distribution can be easily transferred to strategies using the two-point mutation rule (13). Due to this correspondence the further treatment of the $p_{\mathtt{II}}^\star$ case is not necessary for the rest of this paper. The results for the $p_{\mathtt{ln}}^\star$ case hold *mutatis mutandis* for the $p_{\mathtt{II}}^\star$ case.

5 Evolutionary Dynamics

The aim of this section is to provide an explanation how the σSA algorithm works in the time domain. One might call this the *macroscopic* aspect of the ES. The time evolution of the ES is modeled by the two evolution equations (25) and (32). It is really important to realize that the $\varsigma^\star$ evolution, given by (32), does *not* depend on r (the local radius of curvature), whereas the r evolution (25) is *governed* by $\varsigma^\star$ (the normalized mutation strength).

From a mathematical point of view, the system (25), (32) is a *noisy iterated map.*[8] The analytical treatment of (32) is especially difficult. Two approximations will be discussed. The simplest one yields the "first order dynamics" by neglecting the fluctuation terms in (25), (32). The incorporation of the fluctuations will be done in 5.2.

5.1 First Order Dynamics

The first order dynamics are obtained from (25), (32) by postulation of

$$D_\delta = 0, \qquad D^\star_\varphi = 0. \tag{66}$$

Thus, the system reduces to

$$r^{(g+1)} = r^{(g)} \left(1 - \frac{1}{N}\,\varphi^\star(\varsigma^{\star(g)}, N)\right), \tag{67}$$

$$\varsigma^{\star(g+1)} = \varsigma^{\star(g)}\, \frac{1 + \delta^{(1)}(\varsigma^{\star(g)}, N)}{1 - \frac{1}{N}\varphi^\star(\varsigma^{\star(g)}, N)}. \tag{68}$$

5.1.1 The Steady-State, Optimal ES Performance, and Schwefel's Rule

As already pointed out, the σSA evolution is governed by the $\varsigma^\star$ evolution (68). For large generation numbers g, the $\varsigma^\star$ evolution reaches asymptotically the steady-state $\varsigma^\star_{ss} := \lim_{g\to\infty} \varsigma^{\star(g)}$. That is, $\varsigma^{\star(g)} = \varsigma^{\star(g+1)} = \varsigma^\star_{ss}$, and from (68) one gets the condition

$$\underline{\texttt{steady-state:}} \qquad \frac{1}{N}\,\varphi^\star(\varsigma^\star_{ss}, N) = -\delta^{(1)}(\varsigma^\star_{ss}, N). \tag{69}$$

The steady-state $\varsigma^\star_{ss}$ determines the r evolution after a certain transient time (see 5.1.2). In order to have evolutionary progress, i. e. $\varphi^\star > 0$ (cp. e. g. (67)) one has to impose the *necessary evolution condition*

$$\underline{\texttt{necessary evolution condition:}} \qquad 0 < \varsigma^\star_{ss} < \varsigma^\star_{\varphi_0}$$

on $\varsigma^\star_{ss}$. Here, $\varsigma^\star_{\varphi_0}$ is the (second) zero of $\varphi^\star(\varsigma^\star)$

$$\varsigma^\star_{\varphi_0}: \quad \varphi^\star(\varsigma^\star) = 0 \quad \Leftrightarrow \quad \varsigma^\star = \varsigma^\star_{\varphi_0} \quad \wedge \quad \varsigma^\star > 0. \tag{70}$$

For large N ($N \to \infty$) the steady-state $\varsigma^\star_{ss}$ is determined by the zero $\varsigma^\star_{\delta_0}$ of $\delta^{(1)}$: Because $-\delta^{(1)}(\varsigma^\star)$ is a monotonically increasing function of $\varsigma^\star$ with zero $\varsigma^\star_{\delta_0}$ (cp. 4.2.1) and the left hand side of

[8] Up until now the notion "map" - originated in dynamical systems theory - is not common terminology in computer science and evolutionary computation because most people do not consider their algorithms as dynamical systems. However, optimization is a process which works in space and time. Therefore, methods from the dynamical systems theory or from non-equilibrium thermodynamics should gain more attention in the field of evolutionary computation.

(69) becomes zero for $N \to \infty$. Thus, it is $0 = -\delta^{(1)}(\varsigma^\star_{ss}) \Leftrightarrow \varsigma^\star_{ss} = \varsigma^\star_{\delta_0}$. Due to the monotonicity of $\delta^{(1)}$ $\varsigma^\star_{ss}$ cannot be smaller than $\varsigma^\star_{\delta_0}$. Therefore, one obtains for the steady-state $\varsigma^\star_{ss}$

$$\varsigma^\star_{\delta_0} < \varsigma^\star_{ss} < \varsigma^\star_{\varphi_0}. \tag{71}$$

The inequality (71) provides a necessary condition to be imposed on the σ mutation density $p^\star_\sigma(\varsigma^\star \,|\, \varsigma^{\star(g)})$ in order to be a suitable candidate for the σSA ES.

One can easily check that (71) does hold for $p^\star_\sigma = p^\star_{\mathtt{ln}}(\varsigma^\star \,|\, \varsigma^{\star(g)})$, at least for small τ. In this case $\varphi^\star(\varsigma^\star)$ can be approximated by (51) and even be simplified for $N \to \infty$

$$\varphi^\star(\varsigma^\star) = c_{1,\lambda}\varsigma^\star - \frac{1}{2}(\varsigma^\star)^2. \tag{72}$$

Thus, we have $\varsigma^\star_{\varphi_0} = 2c_{1,\lambda}$ and from (62) one obtains

$$\varsigma^\star_{\delta_0} = \left(d^{(2)}_{1,\lambda} - \frac{1}{2}\right) \Big/ \, c_{1,\lambda}. \tag{73}$$

Looking up the values for $\varsigma^\star_{\delta_0}$ and $c_{1,\lambda}$ in Table 2 confirms the validity of $\varsigma^\star_{\delta_0} < \varsigma^\star_{\varphi_0}$.

For the approximation (72) and (62) the equilibrium condition (69) can be solved for $\varsigma^\star_{ss}$. The solution reads

$$\varsigma^\star_{ss} = c_{1,\lambda}(1 - N\tau^2) \; + \; \sqrt{c^2_{1,\lambda}(1 - N\tau^2)^2 + N\tau^2\left(2d^{(2)}_{1,\lambda} - 1\right)}. \tag{74}$$

As can be seen, the learning parameter τ influences the steady-state $\varsigma^\star_{ss}$. Actually, the whole admissible $\varsigma^\star_{ss}$ range (71) can be controlled by τ. From (74) one obtains the extremes

$$\varsigma^\star_{ss} = \varsigma^\star_{\delta_0}, \qquad \text{if} \quad N\tau^2 \gg 1 \tag{75}$$

$$\varsigma^\star_{ss} = 2c_{1,\lambda} = \varsigma^\star_{\varphi_0}, \qquad \text{if} \quad N\tau^2 \to 0. \tag{76}$$

This raises the question how to choose the learning parameter τ. Sometimes a scaling rule $\tau \propto 1/N$ has been proposed. However, this produces $N\tau^2 \propto N(1/N)^2 = 1/N \to 0$ for large N and the steady-state would be in the vicinity of $\varsigma^\star_{\varphi_0}$ (cp. equation (76)) resulting in a nearly zero progress rate, $\varphi^\star \to 0$. Optimal ES performance, however, is achieved for that $\hat{\varsigma}^\star$ which maximizes the progress rate $\varphi^\star$

$$\hat{\varsigma}^\star : \quad \mathtt{Max}_{[\varsigma^\star]}\left\{\varphi^\star(\varsigma^\star)\right\} \quad \Leftrightarrow \quad \varsigma^\star = \hat{\varsigma}^\star,$$

i. e., $\hat{\varsigma}^\star = c_{1,\lambda}$, if approximation (72) is considered. If $\varsigma^\star_{ss} \overset{!}{=} \hat{\varsigma}^\star = c_{1,\lambda}$ is introduced into the equilibrium condition (69), using (72) and (62) one gets

$$\frac{1}{N}\frac{c^2_{1,\lambda}}{2} = -\tau^2\left[\left(d^{(2)}_{1,\lambda} - \frac{1}{2}\right) - c^2_{1,\lambda}\right].$$

Therefore, one obtains for τ

$$\tau = \frac{1}{\sqrt{N}} \frac{c_{1,\lambda}}{\sqrt{2c^2_{1,\lambda} + 1 - 2d^{(2)}_{1,\lambda}}} \qquad \Rightarrow \qquad \tau \propto \frac{1}{\sqrt{N}}. \tag{77}$$

Here, we have found Schwefel's $\tau \propto 1/\sqrt{N}$ rule (Schwefel, 1974): *The learning parameter* τ *should be chosen proportional to* $1/\sqrt{N}$. The factor of proportionality can be fixed around $c_{1,\lambda}$, as far as λ is large enough $\lambda \gtrapprox 10$. This is because of the asymptotic behavior $\lim_{\lambda\to\infty} d^{(2)}_{1,\lambda}/c^2_{1,\lambda} = 1$. It should be emphasized that the τ scaling formula

$$\boxed{\underline{\tau\text{-scaling rule:}} \qquad \lambda \gtrapprox 10: \quad \tau \approxeq \frac{c_{1,\lambda}}{\sqrt{N}}} \tag{78}$$

is a *rule*: At first, it is derived from the first order dynamics (68). Second, for small λ the proportionality factor should be larger than $c_{1,\lambda}$. I. e., the left formula of (77) may be used. However, for $\lambda < 4$ the formula becomes imaginary. This indicates that for $\lambda < 4$ the σSA algorithm cannot reach the theoretical $\varphi^\star$ maximum[9] (though, it still self-adapts the mutation strength ς). Therefore, it is always recommended that λ should be chosen not too small.

Choosing τ with respect to the optimal steady-state is one aspect of σSA. Another aspect concerns the *adaptation time*, i. e., the transient behavior of the algorithm starting from a $\varsigma^\star \neq \hat{\varsigma}^\star$.

5.1.2 The Differential Equation of the σ Evolution and the Transient Behavior for Small $\varsigma^{\star(0)} < \varsigma^\star_{ss}$

Starting from a certain state $(\varsigma^{(0)}, r^{(0)})^T$, the evolution is determined by the system (67), (68). In order to obtain insight into the evolution dynamics of self-adaptation two cases will be considered. First, in this section, the case of an initial ς value chosen too small, i. e. $\varsigma^{\star(0)} < \varsigma^\star_{ss}$, is investigated. In the next section the extreme case $\varsigma^{\star(0)} \gg \varsigma^\star_{ss}$ will be discussed.

In order to derive the differential equation of the $\varsigma^\star$-evolution, we start from equation (33) which holds for

$$|\varphi^\star(\varsigma^\star)|/N \ll 1. \tag{79}$$

Taking (66) into account and applying (72) and (62) yields

$$\varsigma^{\star(g+1)} = \varsigma^{\star(g)} + \tau^2 \varsigma^{\star(g)} \left[\left(\left(d^{(2)}_{1,\lambda} - \frac{1}{2} \right) - c_{1,\lambda}\varsigma^{\star(g)} \right) \left(1 + \frac{\varphi^\star(\varsigma^{\star(g)})}{N} \right) + \frac{1}{N\tau^2} \left(c_{1,\lambda}\varsigma^{\star(g)} - \frac{1}{2}(\varsigma^{\star(g)})^2 \right) \right]. \tag{80}$$

Again, $\varphi^\star/N$ is neglected against 1, whereas $1/N\tau^2$ may be finite, due to (78). Furthermore, it is assumed that τ^2 is small such that $\varsigma^{\star(g+1)} - \varsigma^{\star(g)}$ can be treated as a differential quotient. Thus, one obtains the *differential equation of the* σ *evolution* (writing $\varsigma^{\star(g)} = \sigma^\star(g)$)

$$\frac{d}{dg}\sigma^\star(g) = \tau^2 \sigma^\star(g) \left[\left(d^{(2)}_{1,\lambda} - \frac{1}{2} \right) - c_{1,\lambda} \left(1 - \frac{1}{N\tau^2} \right) \sigma^\star(g) - \frac{1}{2}\frac{1}{N\tau^2} (\sigma^\star(g))^2 \right]. \tag{81}$$

The differential equation can be easily solved for g by the separation method (using integral table (Bronstein & Semendjajew, 1981), p. 89). With

$$a := -\frac{1}{2}\frac{1}{N\tau^2}, \qquad b := -c_{1,\lambda}\left(1 - \frac{1}{N\tau^2}\right), \qquad c := d^{(2)}_{1,\lambda} - \frac{1}{2} \tag{82}$$

[9] Note, this statement is valid within the 'first order approximation'.

we obtain under the condition $\sigma^\star(g) < \varsigma^\star_{ss}$

$$g = \frac{1}{\tau^2} \frac{1}{2d^{(2)}_{1,\lambda} - 1} \left\{ \ln\left(\frac{\sigma^{\star 2}(g)}{\sigma^{\star 2}(0)} \frac{c + b\sigma^\star(0) + a\sigma^{\star 2}(0)}{c + b\sigma^\star(g) + a\sigma^{\star 2}(g)} \right) + \frac{b}{\sqrt{b^2 - 4ac}} \ln\left(\frac{2a\sigma^\star(g) + b + \sqrt{b^2-4ac}}{2a\sigma^\star(g) + b - \sqrt{b^2-4ac}} \frac{2a\sigma^\star(0) + b - \sqrt{b^2-4ac}}{2a\sigma^\star(0) + b + \sqrt{b^2-4ac}} \right) \right\}. \tag{83}$$

This formula can be used to estimate the number of generations necessary to reach the vicinity of the steady-state $\varsigma^\star_{ss}$.

It is interesting to investigate the N-scaling behavior of the transition time (adaptation time, number of generations g). If the scaling rule $\tau \propto 1/\sqrt{N}$ is used, then the terms in the brace of (83) become constant (given fixed $\sigma^\star(0)$ and $\sigma^\star(g)$). The number of generations g needed to reach a certain $\sigma^\star(g)$, $0 < \sigma^\star(0) < \sigma^\star(g) < \varsigma^\star_{ss}$, is inversely proportional to τ^2. Due to $\tau \propto 1/\sqrt{N}$ it becomes proportional to N

$$\texttt{adaptation time:} \qquad g \propto N\,, \qquad \text{if} \qquad \tau \propto \frac{1}{\sqrt{N}}. \tag{84}$$

This is the other side of the coin. Optimal σSA has a transient time behavior proportional to N. It is not a serious problem for current applications with, say $N < 200$. However, if larger applications are to be developed, then this problem should receive more attention. As an example the $(1,10)$ ES with $N = 10{,}000$ has been tested. The optimal τ with respect to the σSA is due to (78) $\tau \approx 0.015$. The strategy starts with $r^{(0)} = 10^4$ and a really mischosen $\sigma = 10^{-3}$ (i. e. $\sigma^\star(0) = 10^{-3}$). One ES run was recorded over $100{,}000$ generations. The first $30{,}000$ generations are displayed in the left picture of Figure 6. The smooth solid curve was computed by numerical inversion of the $g = g(\sigma^\star)$ formula (83). As can be seen, it takes the

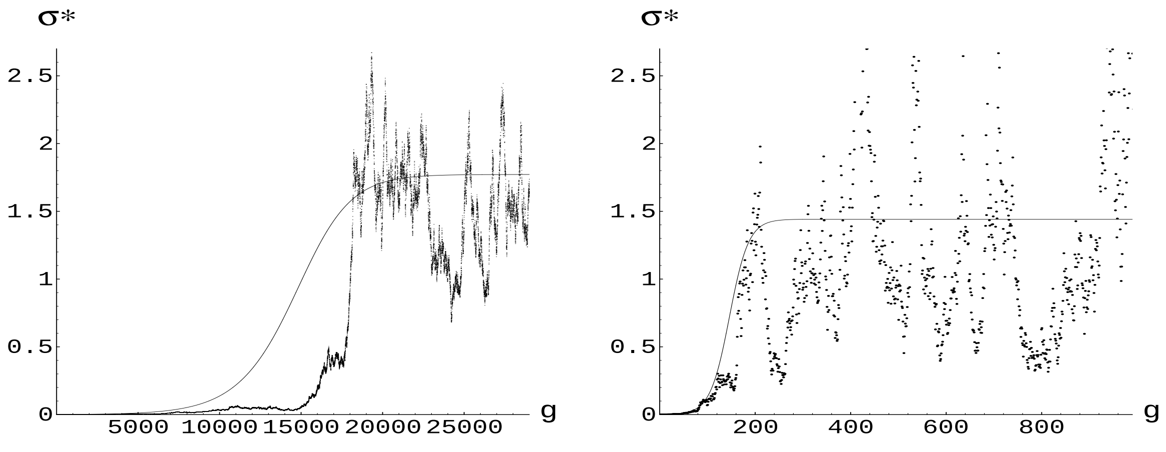

Figure 6: The self-adaptation time (number of generations) depends on the learning parameter τ. The $\sigma^*(g)$-function of the 'first order theory' is displayed in comparison with an actual ES-run (fuzzy curve, consisting of single dots). Left picture: The (almost) optimal $\tau = 0.015$ ($N = 10{,}000$, $(1,10)$ ES). Right picture: The same ES, but $\tau = 0.15$ chosen.

σSA roughly 20,000 generations to approach the steady-state. In the right picture the learning parameter $\tau = 0.15$ was chosen. As predicted by the theory (cp. (83)) an increase of τ by a factor of 10 reduces the adaptation time by a factor of 1/100. We will come back to the example after the treatment of the r evolution.

The r evolution within the first order theory (neglecting fluctuations) is described by equation (67). Under the condition (79), equation (67) can be approximated by the differential equation

$$\frac{d}{dg}\, r(g) = -r(g)\, \frac{1}{N}\, \varphi^\star(\sigma^\star(g)). \tag{85}$$

Since the $\sigma^\star$ evolution is fully determined by (81), the r differential equation can be formally solved for $r(g)$

$$r(g) = r(0) \exp\left\{ -\frac{1}{N} \int_{g'=0}^{g'=g} \varphi^\star(\sigma^\star(g'))\, dg' \right\}. \tag{86}$$

However, an analytical expression can be derived for the steady-state $\varphi^\star_{ss} = \varphi^\star(\varsigma^\star_{ss}) = const.$ only. In this case one obtains with (72)[10]

$$r(g) = r(g_1) \exp\left\{ -\frac{g - g_1}{N} \left(c_{1,\lambda} - \frac{1}{2}\varsigma^\star_{ss} \right) \varsigma^\star_{ss} \right\} \tag{87}$$

or on a logarithmic scale

$$\boxed{\lg(r(g)) = \lg(r(g_1)) \;-\; \lg(\mathsf{e})\, \frac{\varsigma^\star_{ss}}{N} \left(c_{1,\lambda} - \frac{1}{2}\varsigma^\star_{ss} \right)\; (g - g_1).} \tag{88}$$

As can be seen, the steady-state is characterized by a constant negative slope $-\frac{1}{N}\,\varphi^\star(\varsigma^\star_{ss})$ of the logarithmic $r(g)$ function. That is, the σSA with a learning parameter chosen according to (78) exhibits a *linear convergence order.* In Figure 7 the protocol of the ES experiment described above is displayed, again. But now, over the whole range of 100,000 generations. Apart from the $\varsigma^\star$-evolution the graph of $\lg(r) = f(g)$ is displayed. One clearly observes the linear convergence behavior of the r-evolution. In the upper picture of Figure 7 one measures a steady-state progress rate of $\varphi^\star_\infty \approx 1.03$ (from $g = 20,000$ to $g = 100,000$), whereas the lower yields $\varphi^\star_\infty \approx 0.90$ only. However, due to the larger adaptation time of the upper variant ($\tau = 0.015$) it will take more than 100,000 generations to outperform the $\tau = 0.15$ variant (starting from $\sigma(0) = 10^{-3}$).

Summing up the theoretical results obtained so far, one solution to the trade-off problem affecting the optimal σSA performance for large N ($N \gtrapprox 1,000$) might be the *introduction* of a *time dependent learning parameter* τ: In the initial phase, say the first 1,000 generations, τ should be chosen independently from N, e. g., $\tau \approx 0.3$. Within this time interval the σSA should have approached the vicinity of the steady-state. After that, τ is to decrease according to Schwefel's $1/\sqrt{N}$ rule, i. e., (78) is to apply.

5.1.3 Approaching the Steady-State from $\varsigma^\star \gg \varsigma^\star_{ss}$

Unlike the $\varsigma^\star \lessapprox \varsigma^\star_{ss}$ case, the differential equation (81) cannot be used directly, because of the violation of (79). Therefore, one has to start from (67), (68). If $\varsigma^\star$ is very large

$$\varsigma^{\star(0)} \gg N, \tag{89}$$

[10] It is assumed that after $g = g_1$ generations the σSA has reached the vicinity of the steady-state $\varsigma^\star_{ss}$.

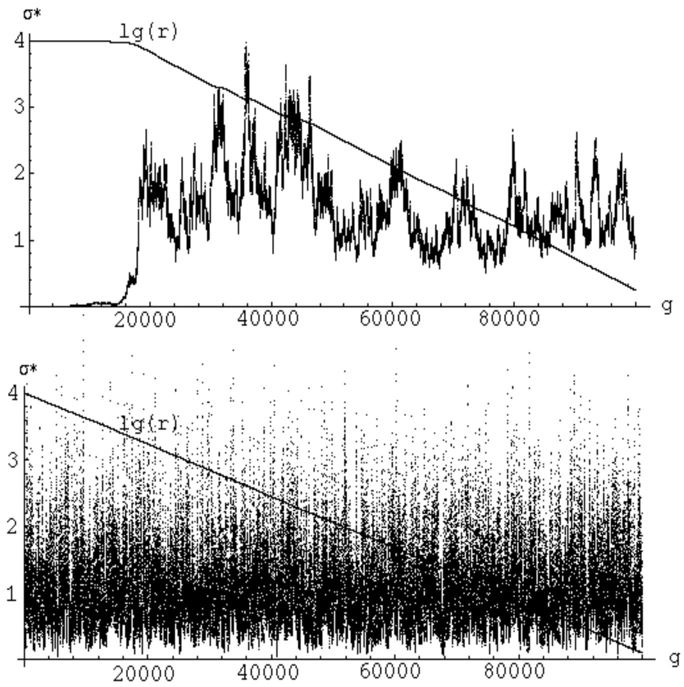

Figure 7: The evolutionary dynamics of the $(1, 10)$ ES experiments with $N = 10,000$, $\sigma(0) = 10^{-3}$, $r(0) = 10^4$, upper picture $\tau = 0.015$, lower picture $\tau = 0.15$.

then the asymptotic expressions for $\delta^{(1)}$ and $\varphi^\star$ can be used

$$\delta^{(1)}(\varsigma^\star) = \delta_\infty = const. \tag{90}$$

$$\varphi^\star(\varsigma^\star) \sim -\sqrt{N}\,\varsigma^\star. \tag{91}$$

The latter can be easily derived from (51). Introducing (90, 91) into (68) and taking (89) into account yields

$$\varsigma^{\star(1)} \sim (1 + \delta_\infty)\sqrt{N}.$$

This is remarkable in that in only one generation the $\varsigma^\star$ value is decreased to a value smaller than $\sqrt{N}$ independent from the initial $\varsigma^{\star(0)} \gg N$ chosen. However, this 'fast adaptation' is bought at the price of a drastically increased r. From (67) and (68) we have $r^{(1)} \sim r^{(0)}\sqrt{N}\,\varsigma^{\star(0)}$. That is, by the first evolution step, the individuals are driven away from the optimum by a factor of (roughly) $\sqrt{N}\,\varsigma^{\star(0)}$. One might interpret this as a launch into global search, desirable if nothing is known about the relative quality of the initially state chosen. However, if the initial starting point was chosen by prior information, then it will be better to use a small initial mutation strength $\varsigma^{\star(0)}$.

Let us come back to the evolutionary dynamics. As already explained, there is only one 'big' $\varsigma^\star$ step. After the first generation the $\varsigma^\star$ changes become smaller. From generation $^{(1)}$ to generation $^{(2)}$, e. g., the relative $\varsigma^\star$ change is roughly proportional to $1/\sqrt{2}$ and $|\varphi(\varsigma^\star)/N| < 1$

holds. Therefore, the differential equation (81) may serve as a (very) rough approximation of the self-adaptation process with the initial condition $\sigma^\star(0) = \sqrt{N}$. Its solution is similar to (83). Again, an important feature is the N-scaling behavior (84) that holds independently from the initial σ value chosen.

5.2 The Influence of Fluctuations

The σSA process is strongly influenced by fluctuations. These fluctuations are of such a kind that they do shift considerably the steady-state $\varsigma^\star_{ss}$ obtained from the 'first order theory'. See, for example, the simulation experiment discussed in 5.1.2 and displayed in Figures 6/7. Consider the case $\tau = 0.015$. For the (first order) steady-state one obtains from (74) $\varsigma^\star_{ss} \approx 1.77$. However, the actually observed one is $\sigma^\star_\infty \approx 1.50$ (obtained from the ES-run by averaging over the generations $g = 20,000$ to $g = 100,000$). For the case $\tau = 0.15$, equation (74) predicts $\varsigma^\star_{ss} \approx 1.44$, whereas the experiment yields $\sigma^\star_\infty \approx 1.08$.

Another interesting observation is the large magnitude of the $\varsigma^\star$-fluctuations. As one can intuitively infer from Figures 6/7, the $\varsigma^\star$-fluctuations around the $\sigma^\star_\infty$ value (the mean value after the transient time regime) are large compared to the fluctuations of the driving force ('microscopic fluctuations').[11] Due to the magnitude of the $\varsigma^\star$-fluctuations, they cannot be neglected.

The real important observation is that the induced stochastic $\varsigma^\star$ process produces a mean value $\sigma^\star_\infty$ which is shifted from the first order steady-state $\varsigma^\star_{ss}$. This indicates that the dynamics of the stochastic process is nonlinear. The aim of this section will be the extraction of the essential nonlinearities which influence the progress rate of the steady-state. In 5.2.1 the governing Chapman-Kolmogorov equations for the noisy $\{\varsigma^\star, r\}$ map are provided with approximated transition densities. The treatment of these equations using the moment method will be done in 5.2.2 – 5.2.5.

5.2.1 The Noisy Map

The evolution of the σSA ES is described by the noisy (iterated) map (25), (32). Note, even this system is an approximation of the real process, since it was assumed that the microscopic fluctuations are Gaussian shaped. Improvements can be easily introduced by Hermite polynomials as has been done in Beyer (1995a). However, the further treatment becomes very cumbersome, even in the case of the Gaussian microscopic fluctuations.

Let the σSA ES at time (generation) g be in the state $(\varsigma^{\star(g)}, r^{(g)})^T$ with (probability) density $p(\varsigma^{\star(g)})$ and $p(r^{(g)})$, respectively. The transition to the new state at $(g+1)$ can be easily written down, if the transition densities $p_{1;\lambda}(r \mid \varsigma^{\star(g)}, r^{(g)})$ and $p_{1;\lambda}(\varsigma^\star \mid \varsigma^{\star(g)})$ (cp. 3.2.2) are known. For the r evolution we have

$$\boxed{p(r^{(g+1)}) = \int_{\varsigma^{\star(g)}=0}^{\infty} \int_{r^{(g)}=0}^{\infty} p_{1;\lambda}(r^{(g+1)} \mid \varsigma^{\star(g)}, r^{(g)}) p(r^{(g)}) p(\varsigma^{\star(g)})\, dr^{(g)}\, d\varsigma^{\star(g)}} \tag{92}$$

and for the $\varsigma^\star$ evolution

$$\boxed{p(\varsigma^{\star(g+1)}) = \int_{\varsigma^{\star(g)}=0}^{\infty} p_{1;\lambda}(\varsigma^{\star(g+1)} \mid \varsigma^{\star(g)}) p(\varsigma^{\star(g)})\, d\varsigma^{\star(g)}.} \tag{93}$$

[11] The standard deviation of the 'microscopic fluctuations' is proportional to D_δ, which has been determined to be of order τ (see equation (64)).

These equations are often referred to as Chapman-Kolmogorov equations (see Fisz (1971)). Writing down these equations is a straightforward task. However, the determination of the transition densities from scratch is the real challenge. A large part of this article has served as the preparation for this final step.

First, the $p_{1;\lambda}(r^{(g+1)} \,|\, \varsigma^{\star(g)}, r^{(g)})$ transition density will be derived. It is obtained from the conditional cdf (cumulative distribution function) $P(r^{(g+1)} < r \mid \varsigma^{\star(g)}, r^{(g)})$ by differentiation with respect to r

$$p_{1;\lambda}(r^{(g+1)} \,|\, \varsigma^{\star(g)}, r^{(g)}) = \frac{d}{dr}\, P\left(r^{(g+1)} < r \mid \varsigma^{\star(g)}, r^{(g)}\right)\Big|_{r=r^{(g+1)}}. \tag{94}$$

From (25) we obtain

$$r^{(g+1)} = r^{(g)}\left(1 - \frac{1}{N}\,\varphi^\star(\varsigma^{\star(g)})\right) + \frac{r^{(g)}}{N}\, D^\star_\varphi(\varsigma^{\star(g)})\mathcal{N}(0,1) \;<\; r$$

which is resolved for $\mathcal{N}(0,1)$

$$\mathcal{N}(0,1) \;<\; \frac{N}{D^\star_\varphi(\varsigma^{\star(g)})r^{(g)}}\left[r - r^{(g)}\left(1 - \frac{1}{N}\,\varphi^\star(\varsigma^{\star(g)})\right)\right] \tag{95}$$

yielding the cdf

$$P(r^{(g+1)} < r \mid \varsigma^{\star(g)}, r^{(g)}) = \frac{1}{2} \;+\; \Phi_0\left(\frac{N}{D^\star_\varphi(\varsigma^{\star(g)})r^{(g)}}\left[r - r^{(g)}\left(1 - \frac{1}{N}\,\varphi^\star(\varsigma^{\star(g)})\right)\right]\right). \tag{96}$$

Thus, one gets from (94) with (96)

$$\begin{aligned} &p_{1;\lambda}(r^{(g+1)} \,|\, \varsigma^{\star(g)}, r^{(g)}) \\ &= \frac{1}{\sqrt{2\pi}}\,\frac{N}{D^\star_\varphi(\varsigma^{\star(g)})r^{(g)}}\exp -\frac{1}{2}\left\{\frac{N}{D^\star_\varphi(\varsigma^{\star(g)})r^{(g)}}\left[r^{(g+1)} - r^{(g)}\left(1 - \frac{1}{N}\,\varphi^\star(\varsigma^{\star(g)})\right)\right]\right\}^2. \end{aligned} \tag{97}$$

For the derivation of $p_{1;\lambda}(\varsigma^{(g+1)} \,|\, \varsigma^{\star(g)})$ we start from a formula similar to (94)

$$p_{1;\lambda}(\varsigma^{\star(g+1)} \,|\, \varsigma^{\star(g)}) = \frac{d}{d\varsigma^\star}\, P\left(\varsigma^{\star(g+1)} < \varsigma^\star \,|\, \varsigma^{\star(g)}\right)\Big|_{\varsigma^\star=\varsigma^{\star(g+1)}}.$$

The random variate $\varsigma^{\star(g+1)}$ is given by (32). In order to get manageable expressions, the smallness assumptions $E\{|\,\varphi^\star(\varsigma^\star)\,|/N\} \ll 1$ and $E\{D^\star_\varphi/N\} \ll 1$ are introduced as in the case of the 'first order approximation' (cp. section 5.1.2). Thus, the simplified equation (33) can be used

$$\begin{aligned} \varsigma^{\star(g+1)} \;=\; \varsigma^{\star(g)}\Bigg[&\left(1 + \delta(\varsigma^{\star(g)})\right)\left(1 + \frac{\varphi^\star(\varsigma^{\star(g)})}{N}\right) \;-\; \left(1 + \delta(\varsigma^{\star(g)})\right)\frac{D^\star_\varphi(\varsigma^{\star(g)})}{N}\,\mathcal{N}_\varphi(0,1) \\ &+ \left(1 + \frac{\varphi^\star(\varsigma^{\star(g)})}{N} - \frac{D^\star_\varphi(\varsigma^{\star(g)})}{N}\,\mathcal{N}_\varphi(0,1)\right) D_\delta \mathcal{N}_\sigma(0,1)\Bigg] \;+\; \ldots. \end{aligned}$$

Further simplifications can be applied to the $\mathcal{N}_{(.)}(0,1)$ random terms. If $|\delta| \ll 1$ is assumed (which holds for sufficiently small τ), then one obtains

$$\varsigma^{\star(g+1)} = \varsigma^{\star(g)}\left[1 + \delta(\varsigma^{\star(g)}) + \frac{\varphi^\star(\varsigma^{\star(g)})}{N} - \frac{D^\star_\varphi(\varsigma^{\star(g)})}{N}\mathcal{N}_\varphi(0,1) + D_\delta \mathcal{N}_\sigma(0,1)\right] \;+\; \ldots. \tag{98}$$

The two normally fluctuating variates $\mathcal{N}_\varphi$ and $\mathcal{N}_\sigma$ can be collected into a new Gaussian variate $\mathcal{N}$ with variance $D^2 = D_\delta^2 + (D_\varphi^\star(\varsigma^{\star(g)}))^2/N^2$. Taking (54) and (64) into account gives

$$D^2 = \tau^2 \left(1 + \frac{\tilde{\sigma}^2 \left(d_{1,\lambda}^{(2)} - c_{1,\lambda}^2\right)}{\tau^2 N^2}\right). \tag{99}$$

If τ is chosen according to Schwefel's rule or larger, $\tau \gtrapprox 1/\sqrt{N}$, then the fraction in (99) can be neglected for sufficiently small $\varsigma^{\star(g)}$ (cp. equation (39) and Table 2) and $D = \tau$ remains. Finally (98) becomes

$$\varsigma^{\star(g+1)} = \varsigma^{\star(g)} \left[1 + \delta(\varsigma^{\star(g)}) + \frac{\varphi^\star(\varsigma^{\star(g)})}{N} + \tau\mathcal{N}(0,1)\right]. \tag{100}$$

In order to derive $p_{1;\lambda}(\varsigma^{\star(g+1)} \,|\, \varsigma^{\star(g)})$ the cdf for $\varsigma^{\star(g+1)} < \varsigma^\star$ is determined from (100). Solving the inequality for $\mathcal{N}(0,1)$ yields

$$\mathcal{N}(0,1) \;<\; \frac{1}{\tau\varsigma^{\star(g)}} \left[\varsigma^\star - \varsigma^{\star(g)} \left(1 + \delta(\varsigma^{\star(g)}) + \frac{\varphi^\star(\varsigma^{\star(g)})}{N}\right)\right].$$

By comparison of the steps from (95) to (97) one finds analogously the transition density

$$p_{1;\lambda}(\varsigma^{\star(g+1)} \,|\, \varsigma^{\star(g)}) = \frac{1}{\sqrt{2\pi}} \frac{1}{\tau\varsigma^{\star(g)}} \exp -\frac{1}{2} \frac{1}{\left(\tau\varsigma^{\star(g)}\right)^2} \left[\varsigma^{\star(g+1)} - \varsigma^{\star(g)} \left(1 + \delta(\varsigma^{\star(g)}) + \frac{\varphi^\star(\varsigma^{\star(g)})}{N}\right)\right]^2. \tag{101}$$

5.2.2 The Mean Value Dynamics of the r Evolution

The system (92), (93) with the transition densities (97) and (101) does completely describe the σSA process. Given an initial $(\varsigma^{\star(0)}, r^{(0)})^T$-distribution $p(\varsigma^{\star(0)}) = \delta(\varsigma^{\star(0)} - \sigma_0^\star)$ and $p(r^{(0)}) = \delta(r^{(0)} - R_0)$ (Dirac's δ used), the evolution in the probability space is determined by (92), (93) forming path integrals. Of practical use, however, is first of all the evolution of the expected value $R(g) := E\{r^{(g)}\}$ defined by

$$R(g) := \int_{r^{(g)}=0}^{\infty} r^{(g)} p(r^{(g)})\, dr^{(g)} =: \overline{r^{(g)}}. \tag{102}$$

For $R(g+1)$ one obtains from (92) and (97)

$$\begin{aligned} R(g+1) &= \int_{\varsigma^{\star(g)}=0}^{\infty} \int_{r^{(g)}=0}^{\infty} r^{(g)} \left(1 - \frac{1}{N}\varphi^\star(\varsigma^{\star(g)})\right) p(r^{(g)})\, p(\varsigma^{\star(g)})\, dr^{(g)}\, d\varsigma^{\star(g)}, \\ R(g+1) &= R(g) \left(1 - \frac{1}{N}\overline{\varphi^\star(\varsigma^{\star(g)})}\right) \end{aligned} \tag{103}$$

with
$$\overline{\varphi^\star(\varsigma^{\star(g)})} = \int_{\varsigma^{\star(g)}=0}^{\infty} \varphi^\star(\varsigma^{\star(g)})\, p(\varsigma^{\star(g)})\, d\varsigma^{\star(g)}. \tag{104}$$

As in the case of the first order theory 5.1.2, the mean value dynamics of r can be approximated as a *linear* differential equation

$$\frac{dR(g)}{dg} = -R(g)\, \frac{\overline{\varphi^\star(\varsigma^{\star(g)})}}{N}$$

controlled by the $\varsigma^\star$ evolution. Its formal solution can be obtained in analogy to (86).

Provided that the $\varsigma^\star$ evolution approaches for sufficiently large $g \ge g_1$ a steady-state, or at least some (first) moments of $\varsigma^\star$ become stationary such that $\overline{\varphi^\star}$ becomes stationary, $\overline{\varphi^\star} \to \varphi^\star_\infty$, then the steady-state solution reads

$$R(g) = R(g_1)\ \mathrm{e}^{-\frac{g-g_1}{N}\ \varphi^\star_\infty}. \tag{105}$$

I. e., as in the case of the first order dynamics, *the R evolution exhibits a linear convergence order*. The slope on the logarithmic R scale is given by $-\varphi^\star_\infty/N$. The only difference to the results of 5.1.2 (below equation (88)) is due to the mean value of $\varphi^\star$. Eq. (105) can be used for the estimation of the generation number G needed to obtain a certain relative R-improvement R/R_0

$$G = \frac{N}{\varphi^\star_\infty}\ \ln\left(\frac{R_0}{R}\right). \tag{106}$$

E. g., reducing the distance to the optimum by a factor of *1/10* needs (roughly) $G \approx 2.3\ N/\varphi^\star_\infty$ generations. (For $\varphi^\star_\infty$ equation (128) may be used as a first approximation, see section 5.1.5.)

If the averaging (104) is performed using (72) one gets

$$\overline{\varphi^\star(\varsigma^{\star(g)})} = c_{1,\lambda}\overline{\varsigma^{\star(g)}} - \frac{1}{2}\,\overline{(\varsigma^{\star(g)})^2} \tag{107}$$

with

$$\overline{(\varsigma^{\star(g)})^k} := \int_{\varsigma^{\star(g)}=0}^{\infty} (\varsigma^{\star(g)})^k\, p(\varsigma^{\star(g)})\, d\varsigma^{\star(g)}.$$

Since the variance of $\varsigma^\star$ is

$$D^2\{\varsigma^{\star(g)}\} = \overline{(\varsigma^{\star(g)})^2} - (\sigma^\star(g))^2 \qquad \text{with} \qquad \sigma^\star(g) := \overline{\varsigma^{\star(g)}},$$

equation (107) can be expressed in terms of

$$\boxed{\overline{\varphi^\star(\varsigma^{\star(g)})} = c_{1,\lambda}\sigma^\star(g) - \frac{1}{2}\,(\sigma^\star(g))^2 - \frac{1}{2}\,D^2\{\varsigma^{\star(g)}\}.} \tag{108}$$

Here, $\sigma^\star(g)$ has been introduced as the expected value of $\varsigma^{\star(g)}$. The result obtained is very important. It provides the explanation for the observation that the maximal progress rate obtained by experiments is always smaller than the theoretical maximum $\hat{\varphi}^\star = \frac{1}{2}\,c^2_{1,\lambda}$. This can be easily seen, if $D^2\{\varsigma^\star\} > 0$ is assumed

$$\overline{\varphi^\star} = c_{1,\lambda}\sigma^\star - \frac{(\sigma^\star)^2}{2} - \frac{D^2\{\varsigma^\star\}}{2}\ <\ c_{1,\lambda}\sigma^\star - \frac{(\sigma^\star)^2}{2}\ \le\ \mathtt{Max}_{[\sigma^\star]}\left\{c_{1,\lambda}\sigma^\star - \frac{(\sigma^\star)^2}{2}\right\} = \frac{c^2_{1,\lambda}}{2}. \tag{109}$$

The $\varsigma^\star$ variance decreases the progress rate. Actually the measured progress rate $\varphi^\star$ is well predicted by (108), if $\sigma^\star_\infty$ and $D_\infty = D\{\varsigma^\star_\infty\}$ are known. This can be verified by experiments. For example, let us consider the numerical experiment introduced in section 5.1.2 (see especially the related Figures 6 and 7). For the $\tau = 0.015$ case one measures the steady-state values $\sigma^\star_\infty \approx 1.503$ and $D_\infty = D\{\varsigma^\star_\infty\} \approx 0.548$ (by starting the data collection at generation $g_1 = 20{,}000$). If one introduces these numerically obtained results into (108) one gets $\varphi^\star_\infty \approx 1.03$. This value is in good agreement with the graphically measured one. The same holds for the case $\tau = 0.15$. Here one measures $\sigma^\star_\infty \approx 1.081$ and $D_\infty = D\{\varsigma^\star_\infty\} \approx 0.594$. This yields $\varphi^\star_\infty \approx 0.90$.

Further discussions concerning (108) and how to improve the progress rate will be postponed to the concluding section 6.

5.2.3 Mean Value Dynamics of the $\varsigma^\star$ Evolution

In order to determine $\sigma^\star_\infty$ the equations describing the mean value dynamics are derived. However, unlike the $R(g)$ dynamics, which is described by only one equation (103), the moment method yields an *infinite* system of equations for the moments of $\varsigma^\star$.

For the first moment $\sigma^\star(g) = \overline{\varsigma^{\star(g)}}$ one obtains from (93) and (101)

$$\sigma^\star(g+1) = \int_{\varsigma^{\star(g)}=0}^{\infty} \varsigma^{\star(g)} \left(1 + \delta(\varsigma^{\star(g)}) + \frac{\varphi^\star(\varsigma^{\star(g)})}{N}\right) p(\varsigma^{\star(g)})\, d\varsigma^{\star(g)},$$

$$\sigma^\star(g+1) = \sigma^\star(g) \;+\; \overline{\varsigma^{\star(g)}\delta(\varsigma^{\star(g)})} \;+\; \frac{1}{N}\,\overline{\varsigma^{\star(g)}\varphi^\star(\varsigma^{\star(g)})}. \tag{110}$$

The substitution of (62) and (72) into (110) yields

$$\sigma^\star(g+1) = \sigma^\star(g)\left[1 + \tau^2\left(d^{(2)}_{1,\lambda} - \frac{1}{2}\right)\right] \;-\; \overline{(\varsigma^{\star(g)})^2}\, c_{1,\lambda}\tau^2\left(1 - \frac{1}{\tau^2 N}\right) \;-\; \overline{(\varsigma^{\star(g)})^3}\,\frac{1}{2}\frac{1}{N}. \tag{111}$$

As can be seen, $\sigma^\star(g+1)$ depends on higher moments of $\varsigma^{\star(g)}$.
For the second moment one obtains from (93)

$$\overline{(\varsigma^{\star(g+1)})^2} = \int_{\varsigma^{\star(g)}=0}^{\infty}\int_{\varsigma^{\star(g+1)}=0}^{\infty} (\varsigma^{\star(g+1)})^2 p_{1;\lambda}(\varsigma^{\star(g+1)}\,|\,\varsigma^{\star(g)}) p(\varsigma^{\star(g)})\, d\varsigma^{\star(g+1)}\, d\varsigma^{\star(g)}. \tag{112}$$

The substitution of (101) into (112) yields after a simple calculation

$$\begin{aligned}\overline{(\varsigma^{\star(g+1)})^2} \;&=\; \overline{(\varsigma^{\star(g)})^2}\,(1+\tau^2) \;+\; 2\,\overline{(\varsigma^{\star(g)})^2\left(\delta(\varsigma^{\star(g)}) + \frac{\varphi^\star(\varsigma^{\star(g)})}{N}\right)} \\ &\quad + \overline{(\varsigma^{\star(g)})^2\left(\delta(\varsigma^{\star(g)}) + \frac{\varphi^\star(\varsigma^{\star(g)})}{N}\right)^2}.\end{aligned} \tag{113}$$

Taking (62) and Schwefel's τ scaling rule $\tau \gtrapprox 1/\sqrt{N}$ into account, the third term in (113) is of order $O(\tau^4)$ and can be neglected

$$\overline{(\varsigma^{\star(g+1)})^2} = \overline{(\varsigma^{\star(g)})^2}\left(1 + 2\tau^2 d^{(2)}_{1,\lambda}\right) - \overline{(\varsigma^{\star(g)})^3}\,\tau^2 c_{1,\lambda}\,2\left(1 - \frac{1}{\tau^2 N}\right) - \overline{(\varsigma^{\star(g)})^4}\,\frac{1}{N}. \tag{114}$$

Again, just like the case of the first order moment, the second order moment depends on moments of higher order. This makes the solution of the $\sigma^\star(g)$ dynamics a hard problem, even for the steady-state case $\sigma^\star_\infty = \sigma^\star(g\to\infty)$. Rough approximations will be presented in the next section. However, the transient time behavior is similar to the first order dynamics (81). This can be easily seen if (111) and (114) are rewritten

$$\sigma^\star(g+1) - \sigma^\star(g) \;=\; \tau^2\left[\left(d^{(2)}_{1,\lambda} - \frac{1}{2}\right)\sigma^\star(g) \;-\; c_{1,\lambda}\,(1-T)\,\overline{(\varsigma^{\star(g)})^2} \;-\; \frac{T}{2}\,\overline{(\varsigma^{\star(g)})^3}\right] \tag{115}$$

$$\overline{(\varsigma^{\star(g+1)})^2} - \overline{(\varsigma^{\star(g)})^2} \;=\; 2\,\tau^2\left[d^{(2)}_{1,\lambda}\,\overline{(\varsigma^{\star(g)})^2} \;-\; c_{1,\lambda}\,(1-T)\,\overline{(\varsigma^{\star(g)})^3} \;-\; \frac{T}{2}\,\overline{(\varsigma^{\star(g)})^4}\right] \tag{116}$$

with

$$T := \frac{1}{\tau^2 N}. \tag{117}$$

Therefore, as in the case of the first order approximation in 5.1.2, the adaptation time is (roughly) proportional to $1/\tau^2$.

5.2.4 Approximative Steady-State Equations for $\sigma^\star_\infty$

In principle it is possible to calculate the $p(\varsigma^{\star(g)})$ distribution by successive integration of the path integral which is constituted by iteration of the Chapman-Kolmogorov equation (93). The steady-state $\sigma^\star_\infty$ can be obtained from the stationary distribution density $p(\varsigma^\star_\infty)$, provided that such a function $p(\varsigma^\star_\infty)$ does exist. From the mathematical point of view this is equivalent to the existence of an eigensolution of the Fredholm-type integral equation

$$c\, p_\infty(\varsigma^\star) = \int_{\varsigma^{\star'}=0}^{\infty} p_{1,\lambda}(\varsigma^\star \mid \varsigma^{\star'})\, p_\infty(\varsigma^{\star'})\, d\varsigma^{\star'} \tag{118}$$

to the eigenvalue $c = 1$, as can be easily seen, if for the stationary case $g \to \infty$, $p(\varsigma^{\star(g)}) = p(\varsigma^{\star(g+1)}) = p(\varsigma^\star_\infty)$, is inserted into (93).

One way of finding the stationary density $p(\varsigma^\star_\infty)$ is given by numerical integration of the path integral. That is, the integration of (93) is numerically performed and iterated starting from an initial $p(\varsigma^{\star(0)}) = \delta(\varsigma^{\star(0)} - \sigma^\star(0))$ (Dirac) density (with, e. g. $\sigma^\star(0) = c_{1,\lambda}$). This has been done for the example in section 5.1.2, $\tau = 0.15$. The result obtained after 218 iterations is displayed in Figure 8. The comparison between the numerical results (solid curve) and the

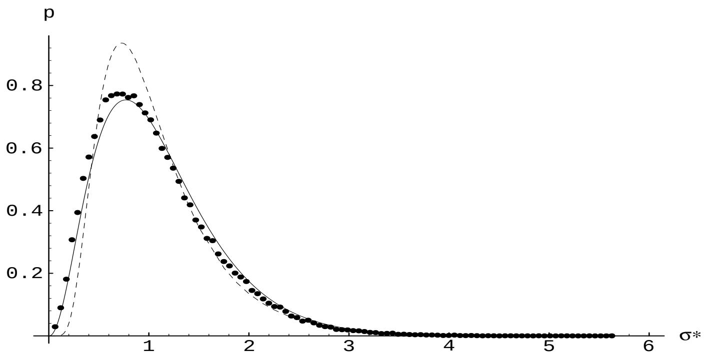

Figure 8: The stationary $p(\varsigma^\star_\infty)$ density (solid curve) for the (1,10) ES with $N = 10{,}000$, $\tau = 0.15$ obtained by numerical integrations iteratively performed. The dots are from the experiment (lower picture in Figure 7). They were obtained by collecting the actual $\varsigma^\star$ values in 100 $\varsigma^\star$-classes. The dashed curve is a lognormal fit to the actual $p(\varsigma^\star_\infty)$ density.

experiment (dots) shows a satisfactory agreement indicating that the approximation (101) of the $\varsigma^\star$ transition density is applicable.

Obtaining the stationary density $p(\varsigma^\star_\infty)$ by numerical integration and iteration is an expensive approach applicable for relatively large τ values only (note, if τ is reduced by a factor of 0.1 the number of iterations increases roughly by a factor of 100).

An alternative approximation method is given by expansion of $p(\varsigma^\star_\infty)$ into a series of orthogonal functions. However, there seems to be no tractable way to perform analytically the integration (118) for any set of basis functions. Therefore, one comes back to the momentum method. Even this approach has its own problems. One possible way to cope with the infinite number of momentum equations is the use of an *ansatz* for $p(\varsigma^\star_\infty)$. The *ansatz* contains a number of parameters to be determined by the momentum equations. If the *ansatz* contains two parameters, then the steady-state is determined by the system

$$\left(d^{(2)}_{1,\lambda} - \frac{1}{2}\right) \sigma^\star_\infty \;=\; c_{1,\lambda}\,(1-T)\;\overline{(\varsigma^\star_\infty)^2} \;+\; \frac{T}{2}\;\overline{(\varsigma^\star_\infty)^3}, \tag{119}$$

$$d^{(2)}_{1,\lambda}\;\overline{(\varsigma^\star_\infty)^2} \;=\; c_{1,\lambda}\,(1-T)\;\overline{(\varsigma^\star_\infty)^3} \;+\; \frac{T}{2}\;\overline{(\varsigma^\star_\infty)^4} \tag{120}$$

which has been obtained from (115), (116) by the steady-state condition

$$g \to \infty : \qquad \sigma^\star(g+1) = \sigma^\star(g) = \sigma^\star_\infty, \qquad \overline{(\varsigma^{\star(g+1)})^2} = \overline{(\varsigma^{\star(g)})^2} = \overline{(\varsigma^\star_\infty)^2}.$$

It is quite clear that the quality of the approximation depends strongly on the *ansatz* chosen. The lognormal distribution with the parameters $\sigma^\star_0$ and s seems to be a suitable candidate

$$p(\varsigma^\star_\infty) = \frac{1}{\sqrt{2\pi}\;s}\,\frac{1}{\varsigma^\star_\infty}\,\exp -\frac{1}{2}\left(\frac{\ln(\varsigma^\star_\infty/\sigma^\star_0)}{s}\right)^2. \tag{121}$$

Figure 8 shows its approximation quality (dashed curve). Obviously, this is a very rough approximation. Finding better ones remains as a future task. For the treatment of (119), (120) the moments of the lognormal distribution (121) up to the fourth order are needed (see Johnson & Kotz (1970)):

$$\sigma^\star_\infty = \sigma^\star_0\,\mathrm{e}^{\frac{1}{2}s^2}, \quad \overline{(\varsigma^\star_\infty)^2} = (\sigma^\star_0)^2\,\mathrm{e}^{2s^2}, \quad \overline{(\varsigma^\star_\infty)^3} = (\sigma^\star_0)^3\,\mathrm{e}^{\frac{9}{2}s^2}, \quad \overline{(\varsigma^\star_\infty)^4} = (\sigma^\star_0)^4\,\mathrm{e}^{8s^2}. \tag{122}$$

The substitution into (119), (120) yields

$$\left(d^{(2)}_{1,\lambda} - \frac{1}{2}\right) \;=\; c_{1,\lambda}\,(1-T)\,\sigma^\star_\infty\,\mathrm{e}^{s^2} \;+\; \frac{1}{2}T\,(\sigma^\star_\infty)^2\,\mathrm{e}^{3s^2}, \tag{123}$$

$$d^{(2)}_{1,\lambda} \;=\; c_{1,\lambda}\,(1-T)\,\sigma^\star_\infty\,\mathrm{e}^{2s^2} \;+\; \frac{1}{2}T\,(\sigma^\star_\infty)^2\,\mathrm{e}^{5s^2}. \tag{124}$$

Now, we are left with a nonlinear system of equations which in general cannot be solved analytically for $\sigma^\star_\infty$ and s.

5.2.5 Discussion of the Steady-State Behavior and Schwefel's Rule

Facing the problem of solving (123), (124) analytically, some special cases will be discussed. But first, results numerically obtained are presented in comparison with ES experiments. The simplest numerical approach is given by iteration of (115), (116) using the momentum formulae (122). The function $\sigma^\star_\infty = f(\tau, N)$ as well as the corresponding $\varphi^\star_\infty = \overline{\varphi^\star(\varsigma^\star_\infty)}$ (obtained from (107)) are displayed in Figure 9. The numerical results (displayed by curves) are comparable with the experiments as far as $N \gtrapprox 100$ is fulfilled. The larger deviations for $N = 30$ and small τ values are due to the stronger influence of the $\varphi^\star(\varsigma^\star)/N$ part in (110), (114). The main observation concerning the $\varphi^\star_\infty$ values is in agreement with the results of the 'first order dynamics'. That is, there is a certain τ value (roughly determined by the rule (78)) that maximizes $\varphi^\star_\infty$. Note, this maximum is *not* symmetrical with respect to τ. Choosing τ too small is much more dangerous than using $\tau > c_{1,\lambda}/\sqrt{N}$, as can be seen for $N = 1,000$ in Figure 9. There is only a moderate performance degradation if $\tau > c_{1,\lambda}/\sqrt{N}$ is chosen. Actually this accounts for the power of the σSA. If the performance of the σSA would be more sensitive to τ, then the σSA could not exhibit such a robust behavior observed in practice.

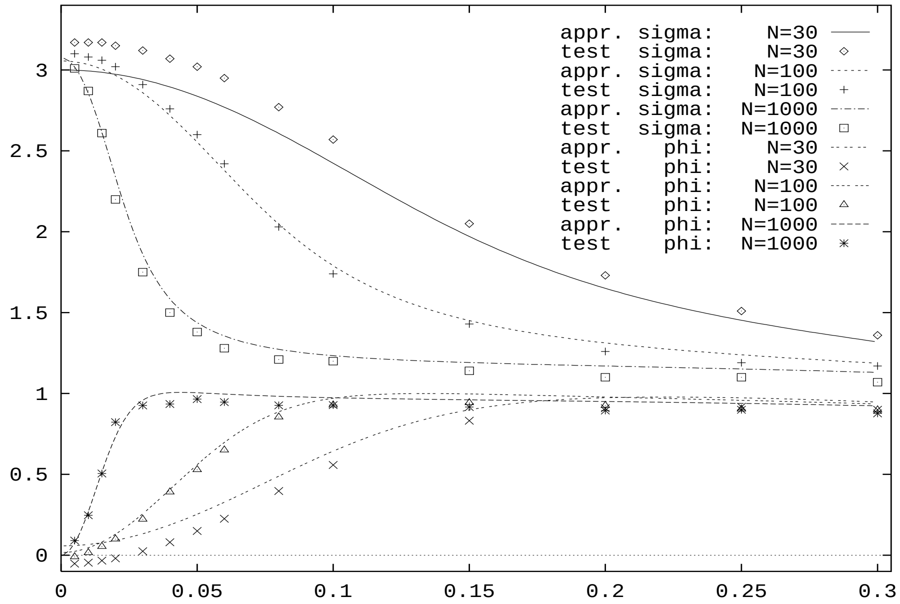

Figure 9: Comparison of the numerical results (denoted as "appr.") with ES experiments (denoted as "test") performing an $(1, 10)$ ES for parameter space dimensions $N = 30$, $N = 100$, and $N = 1000$. Each data point of the "test" was obtained by averaging over more than $90,000$ generations. The steady-state progress rate $\overline{\varphi^\star}$ and $\sigma^\star_\infty$ are displayed (vertical axis) versus the learning parameter τ (horizontal axis).

Now, let us investigate the influence of τ and N on the steady-state $\sigma^\star_\infty$. Assume, s would be known, then (123) can be resolved for $\sigma^\star_\infty$. With (117) the solution reads

$$\sigma^\star_\infty = \left[c_{1,\lambda}(1-\tau^2 N) \; + \; \sqrt{c^2_{1,\lambda}(1-\tau^2 N)^2 \; + \; \tau^2 N (2 d^{(2)}_{1,\lambda} - 1)\mathrm{e}^{s^2}} \right] \mathrm{e}^{-2s^2}. \tag{125}$$

This solution is similar to (74). Actually, the first order results can be recovered for $s = 0$. The formula (125) is well suited for explaining the '$\varsigma^\star_{ss} \to \sigma^\star_\infty$-shift' observed in experiments (cf. the first paragraph of 5.2). Without fluctuations the steady-state $\varsigma^\star_{ss}$ is given by (74) and $s = 0$ in (125), respectively. If the fluctuations are "switched on", i. e., $s^2 > 0$, then the mean value of the steady state is decreased by a factor of e^{-2s^2} (roughly). This is due to the nonlinearity of the $\varsigma^\star$ mapping which can be traced back to the $\overline{(\varsigma^\star)^2}$-term in (111).

As already mentioned, analytical solutions to (123), (124) are difficult to find. However, the special case $T \to 0$ can be easily solved for: The division of (123) by (124) yields $(d^{(2)}_{1,\lambda} - \frac{1}{2})/d^{(2)}_{1,\lambda} = \mathrm{e}^{-s^2}$ and from (123) one finds $\sigma^\star_\infty = \mathrm{e}^{-s^2}(d^{(2)}_{1,\lambda} - \frac{1}{2})/c_{1,\lambda}$. Thus, the steady-state is determined by

$$T \to 0: \qquad \sigma^\star_\infty = \frac{\left(d^{(2)}_{1,\lambda} - \frac{1}{2}\right)^2}{d^{(2)}_{1,\lambda} c_{1,\lambda}}. \tag{126}$$

Due to (117), the condition $T \to 0$ is equivalent to $\tau^2 N \to \infty$ and (126) should be applicable for $\tau^2 N \gtrapprox 10$ as far as τ is small enough ($\tau \lessapprox 0.3$) to ensure the validity of the $\delta(\varsigma^\star)$ approximation

(62). This is met for the example $N = 10,000$, $\tau = 0.15$, in section 5.1.2. The measured steady-state value was $\sigma^\star_\infty \approx 1.08$ (cf. the first paragraph of 5.2). Applying (126) yields $\sigma^\star_\infty \approx 1.17$, this is a relative error of 8%. For the steady-state progress rate $\varphi^\star_\infty$ one gets from (107), (122)

$$\varphi^\star_\infty = c_{1,\lambda}\sigma^\star_\infty - \frac{1}{2}(\sigma^\star_\infty)^2 \mathtt{e}^{s^2} \tag{127}$$

and with (126) one finally obtains the $T \to 0$ case

$$T \to 0: \qquad \varphi^\star_\infty = \frac{\left(d^{(2)}_{1,\lambda} - \frac{1}{2}\right)^2}{d^{(2)}_{1,\lambda}} \left[1 - \frac{d^{(2)}_{1,\lambda} - \frac{1}{2}}{2c^2_{1,\lambda}}\right]. \tag{128}$$

This formula can serve as a first (and very rough) approximation for the expected steady-state progress rate (provided that N is sufficiently large). For the example mentioned above one finds $\varphi^\star_\infty \approx 0.96$, whereas the experiment yields $\varphi^\star_\infty \approx 0.90$.

As has been seen by experiments, choosing $T \to 0$ is not the best policy to obtain a maximal progress rate. Schwefel's $\tau \propto 1/\sqrt{N}$ rule is to apply, instead. This can be verified, e. g., in Figure 9, using the scaling rule $\tau = c_{1,\lambda}/\sqrt{N}$. The question arises whether Schwefel's rule can be derived from the noisy map theory. Within the frame of the lognormal approximation (123), (124), (127) it is indeed possible:

Provided that λ is fixed, then the steady-state solutions of (123), (124) are functions of the parameter T *only*, i. e, $\sigma^\star_\infty = \sigma^\star_\infty(T)$ and $s = s(T)$. If introduced into (127) the steady-state progress rate becomes a function of T $\quad \varphi^\star_\infty = \varphi^\star_\infty(T)$. That is, $\varphi^\star_\infty$ is *controlled* by T. Maximum performance of the ES is achieved if $\varphi^\star_\infty$ is maximal

$$\hat{\varphi}^\star_\infty = \mathtt{Max}_{[T]}\{\varphi^\star_\infty(T)\} \qquad \Leftrightarrow \qquad T = \hat{T} < \infty.$$

In other words, by choosing $T = \hat{T}$ the ES can be tuned for maximum performance (provided that a maximum does exist). Thus, by virtue of (117) one gets

$$\varphi^\star_\infty = \hat{\varphi}^\star_\infty \quad \Leftrightarrow \quad T = \hat{T} \qquad \Rightarrow \qquad \hat{T} = \frac{1}{\tau^2 N} \qquad \Rightarrow \qquad \tau = \frac{1}{\sqrt{\hat{T}}}\frac{1}{\sqrt{N}}$$

and therefore Schwefel's rule: $\boxed{\tau \propto \frac{1}{\sqrt{N}}}$.

6 Conclusions

The analysis of the $(1,\lambda)$ σSA ES presented so far contains a lot of new information that requires summing up at this point.

First, the σSA works for different σ mutation rules. The most prominent one is Schwefel's lognormal mutation rule (6), (9) (Schwefel, 1974). However, symmetric two-point mutations (13) can be applied as well. Actually, the two rules become equivalent with respect to their first and second order SAR-functions, if the learning parameters τ and β, respectively, are sufficiently small. The correspondence $\tau = \beta\sqrt{1-\beta}$ holds (cp. (65)). Furthermore, Fogel's "meta-EP", rule (15), belongs to this class (again, if τ is small). All these rules have a 'symmetry' property in common: Their median is always 1. This is not fulfilled for the generalized two-point case (14) which can be analyzed by the methods presented. However, the investigations of the evolutionary dynamics has not yet been done. The same holds for experimental comparisons of the different mutation rules.

Second, choosing the learning parameter τ according to Schwefel's rule, e. g. $\tau = c_{1,\lambda}/\sqrt{N}$, yields nearly optimum performance for the steady-state case. That is, the steady-state progress rate will be in the vicinity of its maximum. In the case of the two-point rule (11) – (13) the same results can be achieved (theoretically, numerical experiments are yet to be done), if α is chosen according to $\alpha \approx 1 + c_{1,\lambda}/\sqrt{N}$. This formula is approximatively obtained from (65) with (78) under the assumption $\tau \approx \beta$.

Generally one observes only a weak dependence of the steady-state progress rate $\varphi^\star_\infty$ on the learning parameters as long as τ, $\beta \gtrapprox 1/\sqrt{N}$ is fulfilled. Furthermore, applying the scaling rule $\tau = c_{1,\lambda}/\sqrt{N}$ *ensures* the *linear convergence* of the ES-algorithm. This has been proved for the first order approximation (see section 5.1.2) and for the noisy map approximation (section 5.2.2 and 5.2.5) as well. Note, these results have been obtained for the spherical model. However, it is an interesting empirical observation that the linear convergence is often obtained in non-spherical and even multimodal fitness landscapes, too. The reason for this - at first glance astonishing - behavior (that accounts for the power of ES) can be well understood, if one takes into account that the ES works locally. That is, the ES performs most of the time local search with respect to the neighborhood structure of the fitness landscape induced by the mutation operator.[12] However, the local structure of the fitness landscape can be very often approximated by a spherical model (Beyer, 1994; Beyer, 1995b). Thus, the spherical model can be locally applied and the linear convergence behavior becomes understandable.

Third, the learning parameter influences the learning time (adaptation time), i. e., the transient time which is necessary to reach the steady-state regime. The generations needed are inversely proportional to τ^2. If $\tau \propto 1/\sqrt{N}$ is chosen, the transition time becomes proportional to the parameter space dimension N of the optimization problem (cp. equation (84)). This might be a problem for very large N ($N > 1{,}000$). In such cases τ should be controlled according to a schedule that uses a large τ, say $\tau = 0.3$, for an initial number of generations and then switches back to the $\tau = c_{1,\lambda}/\sqrt{N}$ rule.

Fourth, due to fluctuations, the maximal achievable progress rate $\varphi^\star_\infty$ is always smaller than the theoretical value $c^2_{1,\lambda}/2$. From (108), (109) it is known that this deterioration is controlled by the variance $D^2\{\varsigma^\star\}$ of the $\varsigma^\star$ process. Thus, each method reducing $D^2\{\varsigma^\star\}$ can improve the ES-performance. As has been seen, e. g., by experiments in section 5.1.2 (Figure 7), $D^2\{\varsigma^\star\}$ depends only weakly on the learning parameter τ. Therefore, decreasing τ in order to reduce $D^2\{\varsigma^\star\}$ is not very suitable. Within the $(1, \lambda)$ ES a *moving average* over the parental σ values using information from the last κ generations seems to be a powerful $D^2\{\varsigma^\star\}$ reduction mechanism. A more elaborate variant using some kind of 'fading memory', i. e. a weighted moving average, has been proposed by Ostermeier at al. (1994) for the adaptation of N σ_i values (each coordinate direction has its own mutation strength). This method should work for a general σ value as well. The method of moving averages is *not* likely to have an analogy in biological systems. However, from technical point of view it is an efficacious $D^2\{\varsigma^\star\}$ reduction technique. A more biology-like method may be the *intermediate multirecombination*, to be discussed in the next paragraph.

Fifth, if one leaves the $(1, \lambda)$ ES for multi-parent ESs, then another $D^2\{\varsigma^\star\}$ reduction mechanism seems to be feasible - the 'μ/μ_I-recombination' applied to the strategy parameter σ. The advantages of multirecombination in the field of object parameters has been proved in Beyer

[12] It is important to realize, that the notion "local" refers to the effect of the mutation operator. If the mutation strength is large and the search space is bounded, then one might also call its effect "global" search.

(1995b). As the basic working principle the *genetic repair (GR)* has been identified. By the GR the influence of the harmful components of the mutations are reduced. Performing the intermediate μ-parents recombination, i. e., averaging the σ values of the μ parents, should provide a similar effect in the space of the strategy parameter σ. The generational variance of σ is reduced by a factor of $\approx 1/\mu$. The quantitative influence on $D^2\{\varsigma^\star\}$, however, is still a pending problem, which remains to be solved by a theoretical analysis of the $(\mu/\mu_I, \lambda)$ σSA algorithm. Since only one σ parameter is concerned, such analysis should be possible.

Sixth, up to now the self-adaptation has been discussed for the learning of one strategy parameter only - the general mutation strength σ. The SA-algorithm works for the adaptation of individual σ_i values as well (Schwefel, 1987). The theory developed so far cannot give any quantitative recommendations how to choose the different learning parameters. However, some results obtained for the $(1, \lambda)$ ES can be transferred to more general self-adaptation algorithms. Especially the $D^2\{\varsigma^\star\}$ reduction concept holds for each kind of $(\mu/\rho, \lambda)$ ES. It is obvious that the progress rate $\varphi^\star$ is a function of the strategy parameters $\varphi^\star = \varphi^\star(\varsigma_1^\star, \ldots \varsigma_N^\star)$. Goal of the SA-algorithm is to derive the $\varsigma_i^\star$ into the vicinity of the maximum $\hat{\varphi}^\star = \varphi^\star(\hat{\varsigma}_1^\star, \ldots \hat{\varsigma}_N^\star)$. It is quite clear that each fluctuation around the optimum state $\hat{\varsigma}_i^\star$ reduces the average $\overline{\varphi^\star}$ such that $\overline{\varphi^\star} < \hat{\varphi}^\star$ holds. Therefore, any reduction of $D^2\{\varsigma_i^\star\}$ will improve the ES performance. One possibility of decreasing $D^2\{\varsigma_i^\star\}$ might be the averaging over the μ parents of the population as proposed above. It is interesting to notice, that just performing intermediate recombination on the strategy parameters instead of discrete recombination has been recommended by Schwefel (1974). His empirically obtained findings might be well explained by the GR-hypothesis given above. However, a thorough theoretical analysis is still pending.

7 Acknowledgments

The author would like to thank Hans-Paul Schwefel for his support. Long discussions with Heinz Mühlenbein (GMD) concerning the messages of this work and its semantics are worth to be mentioned. The author thanks Ken De Jong for his advice making the article more readable for scientists which are not quite familiar with the ES-theory and self-adaptation. This work was funded by the Deutsche Forschungsgemeinschaft, DFG grant Be 1578/1.

References

References

Beyer, H.-G. (1993). Toward a Theory of Evolution Strategies: Some Asymptotical Results from the $(1,^+ \lambda)$-Theory. *Evolutionary Computation 1*(2), 165–188.

Beyer, H.-G. (1994). Towards a Theory of 'Evolution Strategies': Progress Rates and Quality Gain for $(1,^+ \lambda)$-Strategies on (Nearly) Arbitrary Fitness Functions. In Y. Davidor, R. Männer, and H.-P. Schwefel (Eds.), *Parallel Problem Solving from Nature, 3*, Heidelberg, pp. 58–67. Springer.

Beyer, H.-G. (1995a). Toward a Theory of Evolution Strategies: The (μ, λ)-Theory. *Evolutionary Computation 2*(4), 381–407.

Beyer, H.-G. (1995b). Toward a Theory of Evolution Strategies: On the Benefit of Sex - the $(\mu/\mu, \lambda)$-Theory. *Evolutionary Computation 3*(1), 81–111.

Beyer, H.-G. (1995c). Towards a Theory of 'Evolution Strategies': The $(1, \lambda)$-Self-Adaptation. Technical Report SYS–1/95, Department of Computer Science, University of Dortmund.

Bronstein, I. N. & Semendjajew, K. A. (1981). *Taschenbuch der Mathematik.* Leipzig: BSB B. G. Teubner.

Fisz, M. (1971). *Wahrscheinlichkeitsrechnung und mathematische Statistik.* Berlin: VEB Deutscher Verlag der Wissenschaften.

Fogel, D. (1992). *Evolving Artificial Intelligence.* Ph. D. thesis, University of California, San Diego.

Herdy, M. (1992). Reproductive Isolation as Strategy Parameter in Hierarchically Organized Evolution Strategies. In R. Männer and B. Manderick (Eds.), *Parallel Problem Solving from Nature, 2*, pp. 207–217. Amsterdam: Elsevier.

Johnson, N. L. & Kotz, S. (1970). *continuous univariate distributions-1.* Boston: Houghton Mifflin Company.

Ostermeier, A., Gawelczyk, A., & Hansen, N. (1994). Step-Size Adaptation Based on Non-Local Use of Selection Information. In Y. Davidor, R. Männer, and H.-P. Schwefel (Eds.), *Parallel Problem Solving from Nature, 3*, Heidelberg, pp. 189–198. Springer.

Rechenberg, I. (1973). *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution.* Stuttgart: Frommann–Holzboog Verlag.

Rechenberg, I. (1994). *Evolutionsstrategie '94.* Stuttgart: Frommann–Holzboog Verlag.

Schwefel, H.-P. (1974). Adaptive Mechanismen in der biologischen Evolution und ihr Einfluß auf die Evolutionsgeschwindigkeit. Technical report, Technical University of Berlin. Abschlußbericht zum DFG-Vorhaben Re 215/2.

Schwefel, H.-P. (1987, 1.–5. Juni). Collective phenomena in evolutionary systems. In P. Checkland and I. Kiss (Eds.), *Problems of Constancy and Change — the Complementarity of Systems Approaches to Complexity, Papers presented at the 31st Annual Meeting of the Int'l Soc. for General System Research*, Volume 2, Budapest, pp. 1025–1033. Int'l Soc. for General System Research.

Schwefel, H.-P. (1995). *Evolution and Optimum Seeking.* New York, NY: Wiley.

Schwefel, H.-P. & Rudolph, G. (1995). Contemporary Evolution Strategies. In F. Morana, A. Moreno, J. J. Merelo, and P. Chacon (Eds.), *Advances in Artificial Life. Third ECAL Proceedings*, Berlin, pp. 893–907. Springer-Verlag.