Genetic Programming and Evolvable Machines manuscript No.
(will be inserted by the editor)

Self-Adaptation of Evolution Strategies under Noisy Fitness Evaluations

Hans-Georg Beyer[1], Silja Meyer-Nieberg[2]

[1] Department of Computer Science, Vorarlberg University of Applied Sciences, Hochschulstr. 1, A-6850 Dornbirn, Austria, e-mail: hans-georg.beyer@fhv.at
[2] Department of Computer Science, Universität der Bundeswehr München, D-85557 Neubiberg, Germany, e-mail: silja.meyer-nieberg@unibw.de

Received: date / Revised version: date

Abstract This paper investigates the self-adaptation behavior of $(1, \lambda)$-evolution strategies (ES) on the noisy sphere model. To this end, the stochastic system dynamics is approximated on the level of the mean value dynamics. Being based on this "microscopic" analysis, the steady state behavior of the ES for the scaled noise scenario and the constant noise strength scenario will be theoretically analyzed and compared with real ES runs. An explanation will be given for the random walk like behavior of the mutation strength in the vicinity of the steady state. It will be shown that this is a peculiarity of the $(1, \lambda)$-ES and that intermediate recombination strategies do not suffer from such behavior.

1 Introduction

The performance of evolution strategies in continuous search spaces depends sensitively on the correct choice of the mutations' sizes. In the simplest case, the mutation operator produces iid normally distributed mutation components each of which with standard deviation σ resulting in an isotropic mutation distribution. The mutation parameter σ controls the size of the mutations and has to be adapted online to the shape of the local fitness landscape: A small value of σ will generally lead to a high probability of generating better offspring. The expected improvement w.r.t. the step size towards the optimum, however, will be minor. In contrast to this, a relatively large mutation strength may be connected with a low probability of reaching the success region.

Various methods to control the mutation parameter have been developed. Among these there are for example Rechenberg's 1/5th rule [12], the self-adaptation of Rechenberg and Schwefel [12,14], or the cumulative path-length control by Hansen and Ostermeier [11].

This paper focuses on σ-self-adaptation (σSA) as introduced by Rechenberg and Schwefel. The adaptation of the mutation strength is directly controlled by the evolution strategy itself. The σ-self-adaptation of an $(1, \lambda)$-ES without noise was already analyzed in [3]. There, it has been shown using a dynamical systems approach that the $(1, \lambda)$-ES is able to achieve linear convergence order on the sphere model test function $F(\mathbf{y}) = f(\|\mathbf{y} - \hat{\mathbf{y}}\|)$ (f – strictly monotonic function, $\hat{\mathbf{y}}$ – location of the optimizer). Furthermore, it has been shown that choosing the correct scaling law for the learning parameter τ (for its definition, see below), the ES is able to travel with the maximal speed toward the optimizer.

For a time period of approximate five years [3] remained the only theoretical analysis on σSA. In 2002 and the following years, however, other researchers have also developed alternative approaches investigating and proving, respectively, the convergence behavior of the σSA-ES: Semenov and Terkel [15,16] presented an approach based on stochastic Lyapunov function inspired by Lyapunov's "second method" in stability theory of dynamical systems. Hart et al. [8] gave a rigorous proof of convergence for a special one-dimensional version of the $(1, \lambda)$-ES. Bienvenüe and François [7] provided a Markov chain analysis. By a similar approach, Auger [2] showed the linear convergence order for the standard $(1, \lambda)$-σSA-ES using Monte-Carlo simulations.

All these analyses done so far consider deterministic fitness functions. We will consider an $(1, \lambda)$-Evolution Strategy trying to optimize a fitness function F disturbed by noise. Noisy optimization is a field of growing interest [9]. Its importance in simulation optimization (e.g. Monte-Carlo simulations) and in robust optimization (using injected noise to test robustness) is self-evident. $(1, \lambda)$-σSA-ES exhibit some interesting dynamical behaviors which are not well understood up to now. It is the aim of this paper to shed some light on these behaviors.

The analysis will be restricted to fitness functions of the form $F : \mathbb{R}^N \to \mathbb{R}$ with $F(\mathbf{y}^{(g)}) = c\|\mathbf{y}^{(g)} - \hat{\mathbf{y}}\|^\alpha$, and constant $c > 0$, $\alpha \in \mathbb{R}$, $\alpha > 0$, and $\hat{\mathbf{y}}$ the optimal point of F. Considering basically the $(1, \lambda)$-ES in this paper, we have to deal with only one parent. Therefore there is no recombination and the new generation is created in the following way. Based on the parental object vector $\mathbf{y}^{(g)}$, the offspring population is created by adding normally distributed mutation vectors $\mathbf{z}_l$. The components have zero mean and standard deviation σ (the so-called mutation strength). The seemingly best of the λ offspring is then chosen as the parent in the next generation. Since we consider noisy fitness evaluations the selection is not based on the actual fitness value but on the *observed* fitness

$$\tilde{Q}(\mathbf{x}) = F(\mathbf{x}) + \epsilon. \tag{1}$$

That is, we will investigate an additive noise model. The noise term ϵ is assumed to be a normally distributed random variable with zero mean and standard deviation σ_ϵ, also referred to as noise strength. The model allows to consider two cases of special interest:

a) the constant noise problem: $\sigma_\epsilon = \text{const.}$ and
b) the scaled noise problem: $\sigma_\epsilon = g(\|\mathbf{y} - \hat{\mathbf{y}}\|)$.

As known from empirical investigations and theoretical analyses, case a) causes a behavior with a dynamics approaching a final residual distance of the parents to the global optimizer $\hat{\mathbf{y}}$. That is, the ES is not able to locate the optimizer arbitrarily exact. Case b) can exhibit convergence and divergence, respectively, depending on the noise scaling law considered. It is the aim of this work to quantitatively investigate these behaviors. To this end, we will use the techniques developed in [4] on the level of the mean value dynamics: On the "microscopic" level this concerns the determination of the so-called self-adaptation response (SAR) function and on the "macroscopic" level the solution of the evolution equations. The former will be done in Section 2 while Section 3 is devoted to the solution of the evolution equations in the deterministic approximation. Finally, the concluding Section 4 gives a summary and an outlook.

2 How to Model the σSA-ES

2.1 Self-Adaptation Operators

As already mentioned, in self-adaptive ES the adaptation of the mutation strength σ is evolutionary controlled by the ES itself, i.e., the mutation strength is subject to mutation and selection: The mutation strength which is connected with the seemingly best offspring survives and serves as the mutation strength "parent" in the next generation. The change of the mutation strength, i.e., in this context its mutation, is the first step in an offspring's creation. The operator performing the mutation of the strategy parameter should fulfill several conditions. Among these are the concept of reachability, unbiasedness, and scalability. For a detailed discussion see [6, 5]. The concept of scalability ensures that the ES is able to adapt to the characteristics of the local fitness landscape. With respect to the mutation operator this leads to the requirement that the mutation is to be realized by a multiplication with a random variable ζ. In general, ζ has to fulfill $\mathrm{E}[\zeta] \approx 1$.

There are several operators [4] that fulfill the conditions above. Among these there are for example the "two-point rule" by Rechenberg [13] where the offspring's mutation strength is generated according to

$$\sigma_l := \begin{cases} \sigma\alpha, & \text{if } u(0,1] \leq \frac{1}{2} \\ \sigma/\alpha, & \text{if } u(0,1] > \frac{1}{2} \end{cases}, \tag{2}$$

with $\alpha > 0$ and u standing for the uniform distribution. Probably the most frequently used rule is the "log-normal rule"

$$\sigma_l := \sigma \mathrm{e}^{\tau \mathcal{N}(0,1)}. \tag{3}$$

In this article, we will concentrate on the latter. The density function [4, p.262] introduced by this operator reads

$$p_\sigma(\varsigma|\sigma) = \frac{1}{\sqrt{2\pi}\tau\varsigma} \mathrm{e}^{-\frac{1}{2}\frac{(\ln(\varsigma/\sigma))^2}{\tau^2}}. \tag{4}$$

The parameter τ, $\tau \geq 0$, is referred to as the learning rate.

2.2 Modeling the Evolution Dynamics

Generally, one is interested in the behavior of the ES over the course of the generations g which is referred to as "evolution dynamics" in the sequel. For the sphere model considered, the main dynamical quantities of interest are the mutation strength and the distance to the optimizer $r^{(g)} = \|\mathbf{r}^{(g)}\| := \|\mathbf{y}^{(g)} - \hat{\mathbf{y}}\|$. They fully characterize the system's behavior.

Their evolution can be described by so-called evolution equations. These difference equations model the change of the variables from one generation to the next and are made up of two parts: One represents the expected change of the variable, i.e., the deterministic part of the equation, the other covers the stochastic fluctuations

$$r^{(g+1)} = r^{(g)} - \varphi(r^{(g)}, \varsigma^{(g)}, \sigma_\epsilon, N) + \epsilon_R(r^{(g)}, \varsigma^{(g)}, \sigma_\epsilon, N), \tag{5}$$

$$\varsigma^{(g+1)} = \varsigma^{(g)}(1 + \psi(r^{(g)}, \varsigma^{(g)}, \sigma_\epsilon, N)) + \epsilon_\sigma(r^{(g)}, \varsigma^{(g)}, \sigma_\epsilon, N). \tag{6}$$

The variables ϵ_R and ϵ_σ represent the perturbation parts of the evolution equations. The local performance measures φ and ψ which appear in (5) and (6) denote the expected change of the parameters considered. In the case of the distance $r^{(g)}$, the so-called progress rate φ is defined by

$$\varphi(r^{(g)}, \varsigma^{(g)}) := \mathrm{E}[r^{(g)} - r^{(g+1)}|r^{(g)}, \varsigma^{(g)}], \tag{7}$$

that is, it gives the expected one-generation change of the distance to the optimizer. In contrast to the additive change of the distance, the mutation strength is changed by a multiplication with a random variable. Therefore, we are interested in the relative change of the mutation strength from one generation to the next. The resulting expected relative change is denoted by the measure ψ also referred to as "self-adaptation response" (SAR). It is defined by

$$\psi(r^{(g)}, \varsigma^{(g)}) := \mathrm{E}\left[\frac{\varsigma^{(g+1)} - \varsigma^{(g)}}{\varsigma^{(g)}}|r^{(g)}, \varsigma^{(g)}\right]. \tag{8}$$

Both quantities, the progress rate and the SAR, depend on the noise strength $\sigma_\epsilon^{(g)}$ or its normalized version $\sigma_\epsilon^{*(g)}$ (see below) at generation g.

We will now introduce normalizations to simplify the analysis. Let N denote the search space's dimensionality. Setting

$$\varphi^{*(g)} = \varphi N/r^{(g)}, \quad \varsigma^{*(g)} = \varsigma N/r^{(g)}, \quad \sigma_\epsilon^{*(g)} = \sigma_\epsilon N/(r^{(g)} F^{(g)\prime}), \tag{9}$$

with $F' = \frac{\mathrm{d}}{\mathrm{d}R}F(R)$ and

$$\epsilon^{*\,(g)}_R = \epsilon_R N/(r^{(g)}), \quad \epsilon^{*\,(g)}_\sigma = \epsilon_\sigma/(\varsigma^{*(g)}) \tag{10}$$

we arrive at

$$r^{(g+1)} = r^{(g)}\left(1 - \frac{1}{N}\varphi^*(\varsigma^{*(g)}, \sigma^{*(g)}_\epsilon, N)\right) + \frac{r^{(g)}}{N}\epsilon^{*\,(g)}_R \tag{11}$$

and

$$\varsigma^{*(g+1)} = \varsigma^{*(g)}\frac{1 + \psi(\varsigma^{*(g)}, \sigma^{*(g)}_\epsilon, N) + \epsilon^{*\,(g)}_\sigma}{1 - \frac{1}{N}\varphi^*(\varsigma^{*(g)}, \sigma^{*(g)}_\epsilon, N) + \frac{\epsilon^{*\,(g)}_R}{N}}. \tag{12}$$

If the noise strength σ_ϵ is assumed to be constant, the *normalized* noise strength changes according to

$$\begin{aligned}
\sigma^{*(g+1)}_\epsilon &= \sigma_\epsilon \frac{N}{r^{(g+1)}{F^{(g+1)}}'} = \sigma_\epsilon \frac{N}{c\alpha(r^{(g+1)})^\alpha} \\
&= \sigma_\epsilon \frac{N}{c\alpha\left(r^{(g)}\left(1 - \frac{1}{N}\varphi^*(\varsigma^{*(g)}, \sigma^{*(g)}_\epsilon, N)\right) + \frac{r^{(g)}}{N}\epsilon^{*\,(g)}_R\right)^\alpha} \\
&= \frac{\sigma^{*(g)}_\epsilon}{\left(1 - \frac{1}{N}\varphi^*(\varsigma^{*(g)}, \sigma^{*(g)}_\epsilon, N) + \frac{\epsilon^{*\,(g)}_R}{N}\right)^\alpha}.
\end{aligned} \tag{13}$$

Note, the $r^{(g)}$-evolution is governed by the evolution of the noise strength and of the mutation strength, whereas their dynamics – as we will see later on – are independent of $r^{(g)}$. Assuming in turn σ^*_ϵ as a constant leads to the equation

$$\sigma_\epsilon{}^{(g+1)} = \sigma_\epsilon{}^{(g)}\left(1 - \frac{1}{N}\varphi^*(\varsigma^{*(g)}, \sigma^*_\epsilon) + \frac{\epsilon^{*\,(g)}_R}{N}\right)^\alpha. \tag{14}$$

In this paper, we will neglect the stochastic perturbation parts of (11) and (12) leading to deterministic evolution equations. This can be regarded as a first approximation, the resulting dynamics is also referred to as mean value dynamics.

2.3 The Progress Rate

In order to investigate the mean value dynamics, the progress rate (7) must be determined. Using the normalization (9), the expression

$$\varphi^* = N\,\mathrm{E}[1 - r^{(g+1)}/r^{(g)}|\varsigma^{*(g)}, r^{(g)}]. \tag{15}$$

has to be calculated. In this paper we will make the assumption that the learning parameter τ in (3), (4) is sufficiently small so that we can neglect its influence on φ^*. This is in accordance with findings on the non-noisy case where τ should scale with the search space dimension N, i.e., $\tau \propto 1/\sqrt{N}$ [4]. That is, as $N \to \infty$, one can assume $\tau = 0$ (but only for φ, not for ψ) and use the asymptotically exact progress rate formulae from [4, p.88]:

$$\varphi^*(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)}) \simeq c_{1,\lambda} \frac{(\varsigma^{*(g)})^2}{\sqrt{(\varsigma^{*(g)})^2 + (\sigma_\epsilon^{*(g)})^2}} - \frac{(\varsigma^{*(g)})^2}{2}. \tag{16}$$

The coefficient $c_{1,\lambda} := d_{1,\lambda}^{(1)}$ is a special case of the so-called progress coefficients $d_{1,\lambda}^{(k)}$ (see [4, p. 119] or Eq. (26), respectively) and equals the expectation of the λth order statistics of the standard normal distribution.

2.4 The Self-Adaptation Response

The self-adaptation response (SAR), Eq. (8), is the expectation of the relative change of the mutation strength from one generation to the next. After normalizing the expressions using $r^{(g)}/N$, the SAR is given by

$$\psi(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)}, r^{(g)}) = \int_0^\infty \left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}} \right) p_{1,\lambda}(\varsigma^*) \, \mathrm{d}\varsigma^*. \tag{17}$$

The density $p_{1,\lambda}(\varsigma^*)$ is the density of the mutation strength associated with the seemingly best of the λ offspring. Using the concept of induced order statistics [1,4], $p_{1,\lambda}(\varsigma^*)$ consists of the density that is given by the distribution of the mutation strength $p_\sigma^*(\varsigma^*|\varsigma^{*(g)})$ and of the probability that the mutation strength ς^* leads to the seemingly best of λ candidates, i.e.,

$$\begin{aligned} p_{1,\lambda}(\varsigma^*) = p_\sigma^*(\varsigma^*|\varsigma^{*(g)}) \lambda \int_{-\infty}^\infty & p(\tilde{Q}|\varsigma^*, r^{(g)}) \\ & \times \left(1 - P(\tilde{Q}|\varsigma^{*(g)}, r^{(g)})\right)^{\lambda-1} \mathrm{d}\tilde{Q}. \end{aligned} \tag{18}$$

The SAR (17) can thus be expressed by the double integral

$$\begin{aligned} \psi(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)}, r^{(g)}) = \lambda \int_0^\infty \left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}} \right) p_\sigma^*(\varsigma^*|\varsigma^{*(g)}) \int_{-\infty}^\infty & p(\tilde{Q}|\varsigma^*, r^{(g)}) \\ \times \left(1 - P(\tilde{Q}|\varsigma^{*(g)}, r^{(g)})\right)^{\lambda-1} \mathrm{d}\tilde{Q} \, \mathrm{d}\varsigma^*. & \end{aligned} \tag{19}$$

The probability density function (pdf) $p(\tilde{Q}|\varsigma^*, r^{(g)})$ remains to be determined. To this end, let us consider the observed fitness $\tilde{Q} = F + \epsilon$ (1). The noise term ϵ in (1) is assumed to be normally distributed. Therefore, the pdf of the observed fitness $\tilde{Q}$ of an offspring given the offspring's real (i.e., hidden) fitness value F is $p(\tilde{Q}|F) = 1/(\sigma_\epsilon\sqrt{2\pi}) \exp(-[\tilde{Q} - F]^2/[2\sigma_\epsilon^2])$.

The fitness value F in turn is a function of the offspring's distance vector $\mathbf{R}$ to the optimizer. The vector is given by the distance vector of the parent $\mathbf{r}$ and a mutation vector, i.e., $\mathbf{R} = \mathbf{r} + \mathbf{z}$. The corresponding fitness $F = f(\|\mathbf{R}\|) = f(R)$ can be expressed as $F = f(r\sqrt{(1-x/r)^2 + \|\mathbf{h}\|^2/(r)^2})$ (see Appendix A). The x- and $\|\mathbf{h}\|$-term stem from a decomposition of $\mathbf{r}$ into a part pointing into the direction of the optimizer and a perpendicular part. As a result, one finally obtains (for details, see Appendix A)

$$p(\tilde{Q}|\varsigma^*, \sigma_\epsilon^{*(g)}, r^{(g)}) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left[\frac{\tilde{Q} - F^{(g)} - \frac{\sigma_\epsilon}{2\sigma_\epsilon^{*(g)}}\varsigma^{*2}}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 \varsigma^{*2}}}\right]^2\right) \times \frac{1}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 \varsigma^{*2}}} \quad (20)$$

with $F^{(g)} = F(\mathbf{y}^{(g)})$.

In the following, we will only give a short sketch of the derivation of the self-adaptation response. A detailed version can be found in Appendix B. Note, that we will consider the limit case of $N \to \infty$. As we will see later, the approximation quality of the thus obtained SAR is quite good for finite dimensional search spaces.

The expressions in (19) generally do not allow for an analytical solution, therefore some simplifications have to be introduced. The most important is the consideration of $\tau \ll 1$. It has been shown in [4] that the cumulative distribution function (cdf) $P(\tilde{Q}|\varsigma^{*(g)}, r^{(g)})$ in (19), which is given by

$$P(\tilde{Q}|\varsigma^{*(g)}, \sigma_\epsilon^{*(g)}, r^{(g)}) = \int_0^\infty \Phi\left(\frac{\tilde{Q} - F^{(g)} - \frac{\sigma_\epsilon}{2\sigma_\epsilon^{*(g)}}\varsigma^{*2}}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 \varsigma^{*2}}}\right) \times p_\sigma^*(\varsigma^*|\varsigma^{*(g)})\,\mathrm{d}\varsigma^*, \quad (21)$$

can be approximated with

$$P(\tilde{Q}|\varsigma^{*(g)}, \sigma_\epsilon^{*(g)}, r^{(g)}) = \Phi\left(\frac{\tilde{Q} - F^{(g)} - \frac{\sigma_\epsilon}{2\sigma_\epsilon^{*(g)}}(\varsigma^{*(g)})^2}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 (\varsigma^{*(g)})^2}}\right) \quad (22)$$

if log-normal or two-point distributions are considered. Inserting (22) and (20) into (19) leads to

$$\psi(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)}, r^{(g)}) = \frac{\lambda}{\sqrt{2\pi}}\int_0^\infty \int_{-\infty}^\infty \left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}}\right) p_\sigma^*(\varsigma^*|\varsigma^{*(g)})$$

$$\times \left(1 - \Phi\left[\frac{\tilde{Q} - F^{(g)} - \frac{\sigma_\epsilon}{2\sigma_\epsilon^{*(g)}}(\varsigma^{*(g)})^2}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 (\varsigma^{*(g)})^2}}\right]\right)^{\lambda-1}$$

$$\times \exp\left(-\frac{1}{2}\left[\frac{\tilde{Q} - F^{(g)} - \frac{\sigma_\epsilon}{2\sigma_\epsilon^{*(g)}}\varsigma^{*2}}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 \varsigma^{*2}}}\right]^2\right)$$

$$\times \frac{1}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 \varsigma^{*2}}}\, \mathrm{d}\tilde{Q}\, \mathrm{d}\varsigma^*. \qquad (23)$$

In order to simplify the expressions in (23) further, we introduce a new variable $-t$ for the term inside the Φ-function. Rewriting it for $\tilde{Q}$ leads to

$$-t = \frac{\tilde{Q} - F^{(g)} - \frac{\sigma_\epsilon}{2\sigma_\epsilon^{*(g)}}(\varsigma^{*(g)})^2}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 (\varsigma^{*(g)})^2}}$$

$$\Rightarrow \tilde{Q} = -t\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 (\varsigma^{*(g)})^2} + F^{(g)} + \frac{\sigma_\epsilon}{2\sigma_\epsilon^{*(g)}}(\varsigma^{*(g)})^2$$

and to

$$\psi(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)}) = \frac{\lambda}{\sqrt{2\pi}} \int_0^\infty \int_{-\infty}^\infty \left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}}\right) p_\sigma^*(\varsigma^*|\varsigma^{*(g)})$$

$$\times \sqrt{\frac{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 (\varsigma^{*(g)})^2}{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 \varsigma^{*2}}}$$

$$\times \exp\left(-\frac{1}{2}\left[\sqrt{\frac{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 (\varsigma^{*(g)})^2}{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 \varsigma^{*2}}}\,t \right.\right.$$

$$\left.\left. - \frac{\frac{\sigma_\epsilon}{2\sigma_\epsilon^{*(g)}}\left((\varsigma^{*(g)})^2 - \varsigma^{*2}\right)}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 \varsigma^{*2}}}\right]^2\right)$$

$$\times \Phi(t)^{\lambda-1}\, \mathrm{d}t\, \mathrm{d}\varsigma^*$$

$$= \frac{\lambda}{\sqrt{2\pi}} \int_0^\infty \int_{-\infty}^\infty \left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}}\right) p_\sigma^*(\varsigma^*|\varsigma^{*(g)})$$

$$\times g(\varsigma^*) h(\varsigma^*, t) \Phi(t)^{\lambda-1}\, \mathrm{d}t\, \mathrm{d}\varsigma^*, \qquad (24)$$

with h standing for the exponential and g for the root function.

The consideration of $\tau \ll 1$ allows further simplifications. Basically, we will expand g and h into Taylor series around the parental mutation strength $\varsigma^{*(g)}$. Since the learning parameter controls the distance between these two mutation strengths, it can be assumed that with a very high probability $|\varsigma^* - \varsigma^{*(g)}| \ll 1$ is fulfilled, thus, allowing us to break off the Taylor series after the first terms without introducing severe approximation errors. As a result, the functions g and h can be represented by polynomials in $(\varsigma^* - \varsigma^{*(g)})^k$.

We will first consider the inner integral in (24), i.e., $\int_{-\infty}^{\infty} h(\varsigma^*, t)\, \Phi(t)^{\lambda-1}\, \mathrm{d}t$. First, h is substituted by its Taylor series T_h. Due to the form of h, it is possible to regroup the terms of T_h into terms of $t^k \exp(-t^2/2)$, leading to the general form $T_h(t) = \exp(-t^2/2) \sum_{k=0}^{4} h_k t^k$. The coefficients h_k depend on $(\varsigma^* - \varsigma^{*(g)})^m$ and are derived in Appendix B, Eq. (78). Thus, the integral over t in (24) can be approximated with

$$I_t = \lambda \sum_{k=0}^{4} h_k \int_{-\infty}^{\infty} t^k \frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \Phi(t)^{\lambda-1}\, \mathrm{d}t. \tag{25}$$

Comparing I_t with the so-called progress coefficients [4, p.119]

$$d_{1,\lambda}^{(k)} = \frac{\lambda}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^k \mathrm{e}^{-\frac{t^2}{2}} \Phi(t)^{\lambda-1}\, \mathrm{d}t. \tag{26}$$

it can be re-written as $I_t = h_0 + \sum_{k=1}^{4} h_k d_{1,\lambda}^{(k)}$. Since the integration over ς^* remains to be done, re-ordering I_t in terms of $(\varsigma^* - \varsigma^{*(g)})^m$ is the next step. Thus, I_t (25) can be given by the general expression $T(\varsigma^*) = 1 + A_t(\varsigma^* - \varsigma^{*(g)}) + B_t(\varsigma^* - \varsigma^{*(g)})^2/2 + \mathcal{O}((\varsigma^* - \varsigma^{*(g)})^3)$. Again, the coefficients may be found in Appendix B, Eq. (81). Similar to h, the root

$$g(\varsigma^*) = \sqrt{\frac{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 (\varsigma^{*(g)})^2}{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^{*(g)}}\right)^2 \varsigma^{*2}}}, \tag{27}$$

is approximated by a Taylor series with the general form $T_g(\varsigma^*) = 1 - A_g(\varsigma^* - \varsigma^{*(g)}) + B_g(\varsigma^* - \varsigma^{*(g)})^2/2 + \mathcal{O}((\varsigma^* - \varsigma^{*(g)})^3)$. For the coefficients, see Appendix B, Eq. (74). Thus, the SAR (24) becomes

$$\begin{aligned}\psi(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)}) &= \frac{\lambda}{\sqrt{2\pi}} \int_0^{\infty} \int_{-\infty}^{\infty} \frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}} p_\sigma^*(\varsigma^* | \varsigma^{*(g)}) \\ &\quad \times g(\varsigma^*) h(\varsigma^*, t) \Phi(t)^{\lambda-1}\, \mathrm{d}t\, \mathrm{d}\varsigma^* \\ &= \int_0^{\infty} \frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}} T_g(\varsigma^*) T(\varsigma^*) p_\sigma^*(\varsigma^* | \varsigma^{*(g)})\, \mathrm{d}\varsigma^*. \end{aligned} \tag{28}$$

Computing the product of T_g and T and neglecting the higher order terms leads finally to

$$\begin{aligned}\psi(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)}) = \int_0^\infty \left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}}\right)\Bigg(1 + (A_t - A_g)\varsigma^{*(g)}\left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}}\right) \\ + \left(\frac{B_g + B_t}{2} - A_g A_t\right)(\varsigma^{*(g)})^2\left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}}\right)^2 \\ + \mathcal{O}\Big((\varsigma^* - \varsigma^{*(g)})^3\Big)\Bigg) p_\sigma^*(\varsigma^* | \varsigma^{*(g)})\, \mathrm{d}\varsigma^*. \end{aligned} \tag{29}$$

Since the expectation of a log-normally distributed variable is given by $\overline{\varsigma^{*k}} = (\varsigma^{*(g)})^k \exp(\tau^2 k^2/2)$, the $\overline{((\varsigma^* - \varsigma^{*(g)})/\varsigma^{*(g)})^k}$-terms can be calculated easily. In addition, due to considering $\tau \to 0$, it is possible to break off the exponential series $\exp(\tau^2 k^2/2) = \sum_{n=0}^{\infty} 2^{-n}(k\tau)^{2n}/n!$ after τ^2. Thus, we finally arrive at the asymptotical SAR

$$\begin{aligned}\psi(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)}) \simeq \tau^2 \Bigg(\frac{1}{2} + \big(d_{1,\lambda}^{(2)} - 1\big)\frac{(\varsigma^{*(g)})^2}{(\sigma_\epsilon^{*(g)})^2 + (\varsigma^{*(g)})^2} \\ - \frac{c_{1,\lambda}(\varsigma^{*(g)})^2}{\sqrt{(\sigma_\epsilon^{*(g)})^2 + (\varsigma^{*(g)})^2}}\Bigg). \end{aligned} \tag{30}$$

Equation (30) was obtained assuming $\tau \ll 1$ and $N \to \infty$ which enabled several simplifications of the original SAR. Equation (30) is compared to the results of experiments in Fig. 1. As can be seen, for the $(1, \lambda)$-ES and noise strengths considered, the prediction quality seems to be good if $\varsigma^{*(g)}$ is sufficiently small.

3 Deterministic Approximation

The evolution of the σSA-ES is fully described by the system of stochastic evolution equations (11), (12), and (13). Due to the stochasticity, the general solution would be given by a time-dependent pdf $p(r, \varsigma^*, \sigma_\epsilon^*)^{(g)}$ to be obtained by solving the corresponding Chapman-Kolmogorov-Equations. In this section, we will abstain from trying to solve these equations by means of analytical approximations in general. Instead, we will only consider the stationary state (also referred to as steady state) which is observed for a sufficiently large generation time g, i.e., in the limit $g \to \infty$. Furthermore, we will not search for the steady state pdf, but rather for its first moment assuming that the fluctuating parts in the evolution equations (11), (12), and (13) can be neglected. This is a rather crude approximation, therefore it will be compared with simulations.

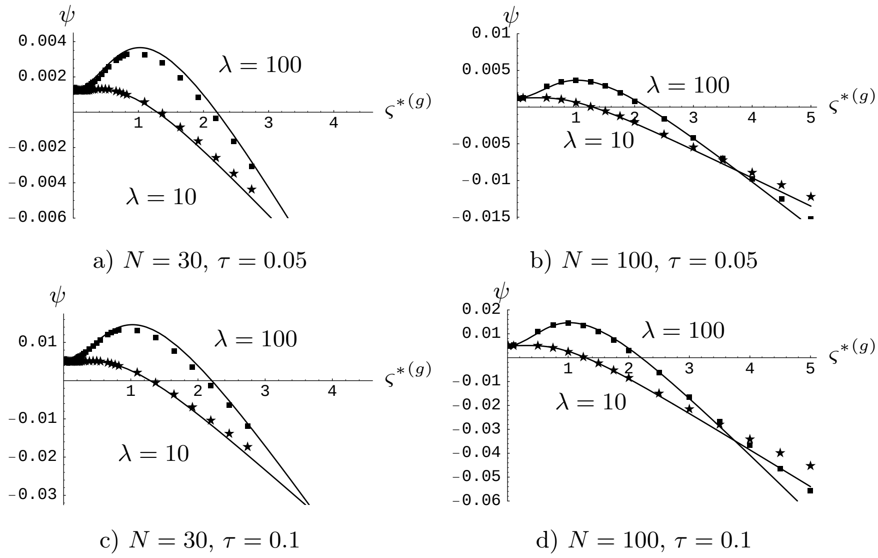


Fig. 1 The first-order self-adaptation response function ψ (30) for some choices of τ. The noise strength is $\sigma^*_\epsilon = 1$. The points denote the results of one-generation experiments on the quadratic sphere and each was obtained by averaging over 250,000 trials.

3.1 The Stationary State

As already mentioned, we will neglect the stochastic perturbation parts of the evolution equations (11), (12), and (13). Applying thus a deterministic approach, the equations simplify to

$$r^{(g+1)} = r^{(g)} \left(1 - \frac{1}{N}\varphi^*(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)})\right) \tag{31}$$

$$\varsigma^{*(g+1)} = \varsigma^{*(g)} \frac{1 + \psi(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)})}{\left(1 - \frac{1}{N}\varphi^*(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)})\right)} \tag{32}$$

$$\sigma_\epsilon^{*(g+1)} = \frac{\sigma_\epsilon^{*(g)}}{\left(1 - \frac{1}{N}\varphi^*(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)})\right)^\alpha}. \tag{33}$$

The progress rate and the SAR in (31) – (33) are given by (16) and (30). As one can see, the r-evolution, Eq. (31), is governed by the evolution of the mutation (32) and the noise strength(33). However, (32) and (33) do *not* depend on (31). That is why we only have to consider the system (32), (33) whereas the r-dynamics is fully controlled by the solution of (32) and (33).

We will consider two typical cases of noisy environments. First, we assume that the noise strength σ_ϵ scales with the distance to the optimizer resulting in a $\sigma_\epsilon^* =$ const. condition. Second, we consider the constant noise strength scenario.

3.1.1 Evolution under Constant Normalized Noise σ_ϵ^* From a correctly working σ-self-adaptation, one expects the normalized mutation strength to reach a stationary value, i.e., $\varsigma^{*(g)} \to \varsigma^*_{stat}$, because ensuring ς^*_{stat} in the vicinity of the maximizer of Eq. (16) ensures maximal progress toward the optimum. Considering the evolution equation (32), this leads to the steady state condition

$$-\psi(\varsigma^*_{stat}, \sigma_\epsilon^{*(g)}) = \varphi^*(\varsigma^*_{stat}, \sigma_\epsilon^{*(g)})/N. \tag{34}$$

It is well known that for sufficiently small normalized noise level σ_ϵ^* the normalized progress rate φ^*, Eq. (16), has two zeros. The φ^*-function is positive in the interval $[0, \sqrt{4c_{1,\lambda}^2 - (\sigma_\epsilon^{*(g)})^2}]$. Thus, we have to require

$$\varsigma^*_{stat} < \sqrt{4c_{1,\lambda}^2 - (\sigma_\epsilon^{*(g)})^2} =: \varsigma^*_{\varphi_0} \tag{35}$$

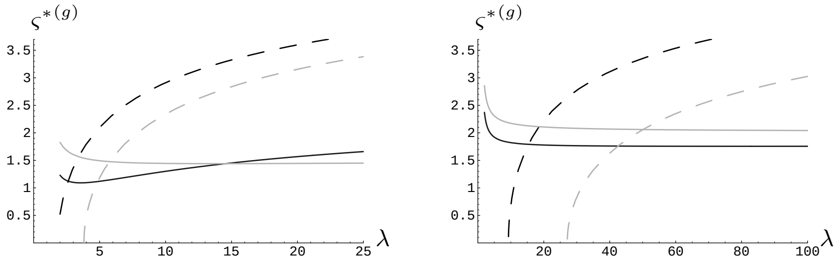
for the stationary mutation strength. Note, a positive progress rate can only be obtained if the normalized noise strength is smaller than $\sigma^*_{\epsilon\max} = 2c_{1,\lambda}$.

Because of (34) the zero of the SAR (30) $\varsigma^*_{\psi_0}$ has to be smaller than ς^*_{stat} in order to allow progress in the steady state. A necessary evolution condition is therefore $0 \le \varsigma^*_{\psi_0} < \varsigma^*_{stat} < \varsigma^*_{\varphi_0}$. Let us first consider the zeros of the SAR (30)

$$\begin{aligned}
&\psi(\varsigma^*, \sigma_\epsilon^*) = 0 \\
&\quad\Longleftrightarrow 0 = \frac{1}{2} + \frac{\varsigma^{*2}}{\varsigma^{*2} + \sigma_\epsilon^{*2}}(d_{1,\lambda}^{(2)} - 1) - \frac{\varsigma^{*2}}{\sqrt{\varsigma^{*2} + \sigma_\epsilon^{*2}}} c_{1,\lambda} \\
&\quad\Longleftrightarrow 0 = \varsigma^{*6} - \frac{(2d_{1,\lambda}^{(2)} - 1)^2 - 4c_{1,\lambda}^2 \sigma_\epsilon^{*2}}{4c_{1,\lambda}^2} \varsigma^{*4} \\
&\qquad\qquad - \sigma_\epsilon^{*2} \frac{2d_{1,\lambda}^{(2)} - 1}{2c_{1,\lambda}^2} \varsigma^{*2} - \frac{\sigma_\epsilon^{*4}}{4c_{1,\lambda}^2}.
\end{aligned} \tag{36}$$

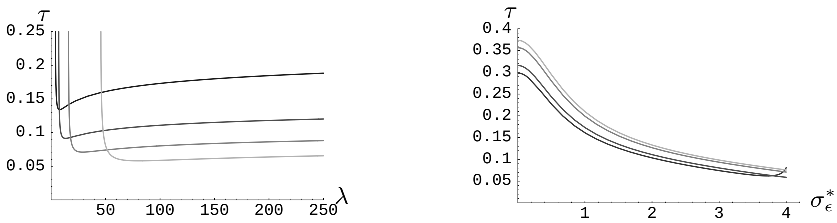
The equation obtained is a third-order polynomial in ς^{*2} so that the solutions (see Appendix D) are not very informative. The results of the numerical comparisons between the zero of the SAR and the second zero of the progress rate are shown in Fig. 2. If the offspring population size is chosen sufficiently large with respect to the normalized noise strength, the positive real solution of $\psi = 0$ is far smaller than $\varsigma^*_{\varphi_0}$ and the first part of the necessary evolution condition is fulfilled.

Trying to determine the general stationary mutation strength using (34) leads to a fourth-order polynomial in ς^{*2}. This equation will not be considered further since we are basically interested in the mutation strength that



a) $\sigma_\epsilon^* = 1$, $\sigma_\epsilon^* = 2$ (gray lines) b) $\sigma_\epsilon^* = 3$, $\sigma_\epsilon^* = 4$ (gray lines)

Fig. 2 The zero of the SAR and the second zero of the progress rate as functions of λ for some noise strengths. The progress rates are denoted by the dashed lines. The zeros (SAR and progress rate) for the higher noise strengths in a) and in b) are plotted in gray.



a) τ_{opt} as a function of λ b) τ_{opt} as a function of σ_ϵ^*

Fig. 3 Optimal choice of τ ($N = 100$) as given by (38). The values for $\sigma_\epsilon^* = 1$, $\sigma_\epsilon^* = 2$, $\sigma_\epsilon^* = 3$, and $\sigma_\epsilon^* = 4$ are shown from top to bottom in figure a). The figure on the right shows the results for $\lambda = 50$ (bottom), $\lambda = 100$, $\lambda = 500$, and $\lambda = 1000$ (top).

maximizes the progress rate φ^* (16). Setting the derivative of φ^* to zero leads to a polynomial of third-order in ς^{*2}

$$\left(\varsigma^{*2} + \sigma_\epsilon^{*2}\right)^3 = c_{1,\lambda}^2\left(\varsigma^{*4} + 4\varsigma^{*2}\sigma_\epsilon^{*2} + 4\sigma_\epsilon^{*4}\right). \tag{37}$$

Its solution ς_{opt}^* is provided in Appendix D. The optimal choice of the τ-value is then obtained by inserting ς_{opt}^* into the steady-state condition and resolving it for τ

$$\tau_{opt} = \frac{1}{\sqrt{N}}\sqrt{\frac{\frac{{\varsigma_{opt}^*}^2}{2} - \frac{{\varsigma_{opt}^*}^2}{\sqrt{{\varsigma_{opt}^*}^2+\sigma_\epsilon^{*2}}}c_{1,\lambda}}{\frac{1}{2} + \frac{{\varsigma_{opt}^*}^2}{\varsigma^{*2}+\sigma_\epsilon^{*2}}(d_{1,\lambda}^{(2)} - 1) - \frac{{\varsigma_{opt}^*}^2}{\sqrt{{\varsigma_{opt}^*}^2+\sigma_\epsilon^{*2}}}c_{1,\lambda}}}. \tag{38}$$

The optimal learning rate scales with $1/\sqrt{N}$ which is due to the form of the steady state condition (34) and due to the form of the SAR (30). Figure 3 shows the optimal τ-value for some combinations of noise strengths and offspring population sizes λ. As one can see, as long as λ is sufficiently large in relation to σ_ϵ^*, an optimal choice of τ exists. Therefore, for a constant σ_ϵ^*, the ES is able to adapt the mutation strength so that it progresses with maximal speed – provided that the offspring population size is large enough. Generally, the parameter τ_{opt} grows with increasing λ and decreases with increasing noise.

3.1.2 Evolution under Permanent Noise σ_ϵ Let us now consider the case of a constant noise strength σ_ϵ. The normalized noise strength defined in (10) $\sigma_\epsilon^{*(g)} = \sigma_\epsilon N/(c\alpha(r^{(g)})^\alpha)$ gradually increases during the course of the evolution until no progress is possible anymore and the evolution of the $r^{(g)}$ comes to a hold (on average).

Three phases can be distinguished. As long as the system is far away from the optimum, the influence of the normalized noise strength can be neglected. As a consequence, the steady state formula

$$\varsigma_{stat}^* = c_{1,\lambda}(1 - N\tau^2) + \sqrt{c_{1,\lambda}^2(1 - N\tau^2)^2 + N\tau^2(2d_{1,\lambda}^{(2)} - 1)}, \qquad (39)$$

obtained in [4], holds. And considering the maximizer $\varsigma^* = c_{1,\lambda}$ of the noise-free progress rate, one gets $\tau = c_{1,\lambda}/\sqrt{N[2c_{1,\lambda}^2 + 1 - 2d_{1,\lambda}^{(2)}]}$ for the optimal τ-value.

As the ES progresses and the normalized noise strength increases, $\varsigma^* = c_{1,\lambda}$ does not fulfill the steady state condition anymore. The former steady state is lost. The increasing noise strength $\sigma_\epsilon^{*(g)}$ influences the equations more and more and leads to a continuously changing mutation strength. As a result, the r-dynamics converges to a stationary state which is characterized by $\varphi^*(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)}) = 0$.

Recall, there are two qualitatively different zeros of φ^* (16), $\varsigma_1^* = 0$ (associated ideally with $\sigma_\epsilon^* = 2c_{1,\lambda}$) and $\varsigma_2^* = \sqrt{4c_{1,\lambda}^2 - \sigma_\epsilon^{*2}}$. Demanding stationarity of the ς^*-evolution, i.e., $\psi = 0$, the latter condition can be used to determine a stationary mutation strength ς_∞^* and thus the corresponding noise strength $\sigma_{\epsilon\,\infty}^*$. Setting $\psi(\varsigma_\infty^*) = 0$, we arrive at

$$\begin{aligned} 0 &= \frac{1}{2} + \frac{(\varsigma_\infty^*)^2}{(\varsigma_\infty^*)^2 + (\sigma_{\epsilon\,\infty}^*)^2}(d_{1,\lambda}^{(2)} - 1) - \frac{c_{1,\lambda}(\varsigma_\infty^*)^2}{\sqrt{(\varsigma_\infty^*)^2 + (\sigma_{\epsilon\,\infty}^*)^2}} \\ \Rightarrow 0 &= \frac{1}{2} + (\varsigma_\infty^*)^2 \frac{d_{1,\lambda}^{(2)} - 1 - 2c_{1,\lambda}^2}{4c_{1,\lambda}^2} \\ \Rightarrow \varsigma_\infty^{*A} &= 2c_{1,\lambda}\frac{1}{\sqrt{2\big(2c_{1,\lambda}^2 + 1 - d_{1,\lambda}^{(2)}\big)}}. \end{aligned} \qquad (40)$$

This leads to the stationary noise strength

$$\sigma_{\epsilon\,\infty}^{*A} = 2c_{1,\lambda}\sqrt{1 - \frac{1}{2\big(2c_{1,\lambda}^2 + 1 - d_{1,\lambda}^{(2)}\big)}} \qquad (41)$$

and using ${\sigma_\epsilon^*}^{(g)} = \sigma_\epsilon N/(c\alpha(r^{(g)})^\alpha)$ to a residual location error

$$R_\infty^A = \sqrt[\alpha]{\frac{\sigma_\epsilon N}{c\alpha 2c_{1,\lambda}}\sqrt{\frac{2\big(2c_{1,\lambda}^2+1-d_{1,\lambda}^{(2)}\big)}{2\big(2c_{1,\lambda}^2+1-d_{1,\lambda}^{(2)}\big)-1}}} \tag{42}$$

defined for $2c_{1,\lambda}^2 + 1 - d_{1,\lambda}^{(2)} > 1/2$.

As explained above, the r-evolution is determined by those of the mutation strength and the noise strength. Therefore, the r-evolution will not be considered further. There are two different pairs of equilibrium points of the evolution equations (32) and (33). The first with $\mathbf{e}_1 = (0, w)^{\mathrm{T}}$ with $w \in \mathbb{R}$ and ideally $w = 2c_{1,\lambda}$ and the second at $\mathbf{e}_2 = (s_2, w_2)^{\mathrm{T}}$ with s_2 given by (40) and w_2 by (41). The question arises which of these pairs is locally stable, i.e., stable w.r.t. small disturbances.

Let us consider a difference equation $\mathbf{x}^{(g+1)} = f(\mathbf{x}^{(g)})$. Let $\mathbf{y}$ be an equilibrium or fixed point with $\mathbf{y} = f(\mathbf{y})$. The effect of small disturbances $\mathbf{w}^{(g)} = \mathbf{y} + \Delta\mathbf{w}^{(g)}$ can be examined using the Taylor series of f at $\mathbf{y}$ and neglecting the higher order terms [10]. We obtain the new difference equation $\Delta\mathbf{w}^{(g+1)} = Df|_{\mathbf{x}=\mathbf{y}}\Delta\mathbf{w}^{(g)}$, where $Df|_{\mathbf{x}=\mathbf{y}}$ is the Jacobian of f at $\mathbf{y}$. In the case of a function $f : \mathbb{R}^N \to \mathbb{R}^N$, $f = (f_1, \ldots, f_N)^{\mathrm{T}}$, it is given as

$$Df = \begin{pmatrix} \frac{\partial}{\partial x_1} f_1 & \cdots & \frac{\partial}{\partial x_N} f_1 \\ \vdots & & \vdots \\ \frac{\partial}{\partial x_1} f_N & \cdots & \frac{\partial}{\partial x_N} f_N \end{pmatrix}. \tag{43}$$

The fixed point $\mathbf{y}$ is then called asymptotically (locally) stable if $\lim_{g\to\infty} \Delta\mathbf{w}^{(g)} = 0$. The question, whether $\mathbf{y}$ is a stable fixed point can be solved by determining the eigenvalues of $Df|_{\mathbf{x}=\mathbf{y}}$. If an eigenvalue λ_i exists with $|\lambda_i| > 1$, then $\mathbf{y}$ is unstable [10]. Thus, we have to determine the solutions of $\det(Df|_{\mathbf{x}=\mathbf{y}} - \lambda^{\mathrm{T}}\mathbf{E}) = 0$, where $\mathbf{E}$ is the unity matrix.

Considering the evolution equations (32) and (33), we have to compute the Jacobian matrix at $(\varsigma_\infty^*, {\sigma_\epsilon^*}_\infty)^{\mathrm{T}}$ of

$$f\begin{pmatrix}\varsigma^* \\ \sigma_\epsilon^*\end{pmatrix} = \begin{pmatrix} \varsigma^* \frac{1+\psi(\varsigma^*,\sigma_\epsilon^*)}{1-\varphi^*(\varsigma^*,\sigma_\epsilon^*)/N} \\ \sigma_\epsilon^* \frac{1}{(1-\varphi^*(\varsigma^*,\sigma_\epsilon^*)/N)^\alpha} \end{pmatrix}. \tag{44}$$

In general, it is given by

$$Df\begin{pmatrix}\varsigma^* \\ \sigma_\epsilon^*\end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial\varsigma^*} f_1 & \frac{\partial}{\partial\sigma_\epsilon^*} f_1 \\ \frac{\partial}{\partial\varsigma^*} f_2 & \frac{\partial}{\partial\sigma_\epsilon^*} f_2 \end{pmatrix} \tag{45}$$

with

$$\frac{\partial}{\partial\varsigma^*} f_1 = \frac{1+\psi(\varsigma^*,\sigma_\epsilon^*)}{1-\varphi^*(\varsigma^*,\sigma_\epsilon^*)/N} + \varsigma^* \left(\frac{\frac{\partial}{\partial\varsigma^*}\psi(\varsigma^*,\sigma_\epsilon^*)}{1-\varphi^*(\varsigma^*,\sigma_\epsilon^*)/N} \right.$$

$$
\begin{aligned}
&\qquad + \frac{\partial}{\partial \varsigma^*}\varphi^*(\varsigma^*,\sigma_\epsilon^*)\frac{1+\psi(\varsigma^*,\sigma_\epsilon^*)}{N(1-\varphi^*(\varsigma^*,\sigma_\epsilon^*)/N)^2}\Bigg)\\
&\frac{\partial}{\partial \varsigma^*} f_2 = \sigma_\epsilon^* \frac{\alpha \frac{\partial}{\partial \varsigma^*}\varphi^*(\varsigma^*,\sigma_\epsilon^*)}{N(1-\varphi^*(\varsigma^*,\sigma_\epsilon^*)/N)^{\alpha+1}}\\
&\frac{\partial}{\partial \sigma_\epsilon^*} f_1 = \varsigma^* \Bigg(\frac{\frac{\partial}{\partial \sigma_\epsilon^*}\psi(\varsigma^*,\sigma_\epsilon^*)}{1-\varphi^*(\varsigma^*,\sigma_\epsilon^*)/N}\\
&\qquad + \frac{\partial}{\partial \sigma_\epsilon^*}\varphi^*(\varsigma^*,\sigma_\epsilon^*)\frac{1+\psi(\varsigma^*,\sigma_\epsilon^*)}{N(1-\varphi^*(\varsigma^*,\sigma_\epsilon^*)/N)^2})\Bigg)\\
&\frac{\partial}{\partial \sigma_\epsilon^*} f_2 = \frac{1}{(1-\varphi^*(\varsigma^*,\sigma_\epsilon^*)/N)^\alpha} + \sigma_\epsilon^* \frac{\alpha \frac{\partial}{\partial \sigma_\epsilon^*}\varphi^*(\varsigma^*,\sigma_\epsilon^*)}{N(1-\varphi^*(\varsigma^*,\sigma_\epsilon^*)/N)^{\alpha+1}}.
\end{aligned}
\tag{46}
$$

The derivations of the progress rate and the SAR are given by

$$
\begin{aligned}
\frac{\partial}{\partial \varsigma^*}\varphi^*(\varsigma^*,\sigma_\epsilon^*) &= \frac{c_{1,\lambda}\varsigma^*}{\sqrt{\varsigma^{*2}+\sigma_\epsilon^{*2}}}\Big(2-\frac{\varsigma^{*2}}{\varsigma^{*2}+\sigma_\epsilon^{*2}}\Big) - \varsigma^*\\
\frac{\partial}{\partial \varsigma^*}\psi^*(\varsigma^*,\sigma_\epsilon^*) &= \frac{\tau^2\varsigma^*}{\sqrt{\varsigma^{*2}+\sigma_\epsilon^{*2}}}\Bigg(\frac{2(d_{1,\lambda}^{(2)}-1)}{\sqrt{\varsigma^{*2}+\sigma_\epsilon^{*2}}}(1-\frac{\varsigma^{*2}}{\varsigma^{*2}+\sigma_\epsilon^{*2}})\\
&\qquad + \frac{c_{1,\lambda}\varsigma^{*2}}{\varsigma^{*2}+\sigma_\epsilon^{*2}} - 2c_{1,\lambda}\Bigg)\\
\frac{\partial}{\partial \sigma_\epsilon^*}\varphi^*(\varsigma^*,\sigma_\epsilon^*) &= -\frac{c_{1,\lambda}\sigma_\epsilon^*\varsigma^{*2}}{\sqrt{\varsigma^{*2}+\sigma_\epsilon^{*2}}^3}\\
\frac{\partial}{\partial \sigma_\epsilon^*}\psi^*(\varsigma^*,\sigma_\epsilon^*) &= \frac{\tau^2\sigma_\epsilon^*\varsigma^{*2}}{\sqrt{\varsigma^{*2}+\sigma_\epsilon^{*2}}^3}\Big(c_{1,\lambda} - 2\frac{d_{1,\lambda}^{(2)}-1}{\sqrt{\varsigma^{*2}+\sigma_\epsilon^{*2}}}\Big).
\end{aligned}
\tag{47}
$$

Let us now consider the first equilibrium point $\mathbf{e}_1 = (0, w)^{\mathrm{T}}$ with $w \in \mathbb{R}$. The Jacobian at $\mathbf{e}_1$ is easily calculated as

$$
Df = \begin{pmatrix} 1+\frac{\tau^2}{2} & 0 \\ 0 & 1 \end{pmatrix}
\tag{48}
$$

leading to the equation $(1+\tau^2/2-\lambda_1)(1-\lambda_2) = 0$ for the eigenvalues. Since $\lambda_1 > 1$, a disturbance that affects the mutation strength is intensified and $\varsigma^{*(g)}$ grows. Thus, a point with $\varsigma^*_\infty = 0$ is unstable.

The stability of the second equilibrium point can be determined by inserting (40) and (41) into the Jacobian. The expression obtained is rather clumsy, therefore, we provide a numerically obtained plot of the eigenvalues for $\alpha = 2$ and a range of λ-values in Fig. 4. As one can see, the larger of both eigenvalues is less than the critical value of one. Generally, the larger eigenvalue approaches 1 if $\tau \to 0$ and decreases if the learning parameter increases. This is a reasonable result: If $\tau = 0$, the mutation operator (2)

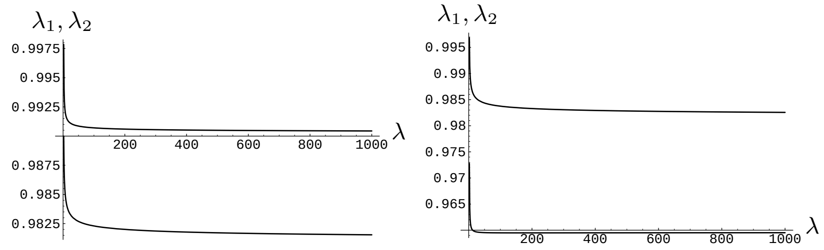


Fig. 4 Numerically obtained eigenvalues of the Jacobian for the fixed point $\mathbf{e}_2$, i.e. the mutation strength given by (40) and the noise strength by (41). The search space dimension was set to $N = 100$ and additionally $\alpha = 2$ was chosen.

does not change the mutation strength. That is, the mapping is neither contracting nor expanding. In finite dimensional search spaces and for $\tau > 0$, one can conclude that the second fixed point, where the mutation strength is given by (40) and the noise strength by (41), is locally stable – at least for the sphere. Figure 5 illustrates the behavior of the equilibrium points if small disturbances occur.

Interestingly, the distance $R_\infty^B = \sqrt[\alpha]{\sigma_\epsilon N/(2c\alpha c_{1,\lambda})}$ obtained as an ideal case for a vanishing mutation strength and for a noise strength ${\sigma^*_\epsilon}_\infty = 2c_{1,\lambda}$ does not differ much from (42) (see Fig. 6). If the size of the offspring population is sufficiently large, the difference is negligible. This means in turn, any mutation strength between zero and (40) leads to similar residual location errors.

3.2 Simulations

Due to space limitations and because of its higher practical relevance, we will focus on the constant noise strength case of Section 3.1.2. We will compare the predicted stationary mutation strength (40) and the residual location error (42) with the results of experiments. The quadratic sphere was chosen as test function in all experiments.

Figure 6 compares the predicted expected r-value at the steady state with simulations of real ES runs depending on the number of offspring individuals. As one can see, the predictive quality of (42) is rather good, however, one observes some randomly appearing small deviations of some data points from the curve. There is a deeper reason for this behavior which can be traced back to the ς^*-evolution.

Figure 7 a) presents the long-term ς^*-dynamics of a typical run of an $(1, 100)$-ES on a sphere with constant noise strength. After approaching the vicinity of the steady state (within a few hundred generations) the initial steady state is lost again. Unlike the prediction of the *deterministic* approximation, the ES is generally not able to regain the predicted

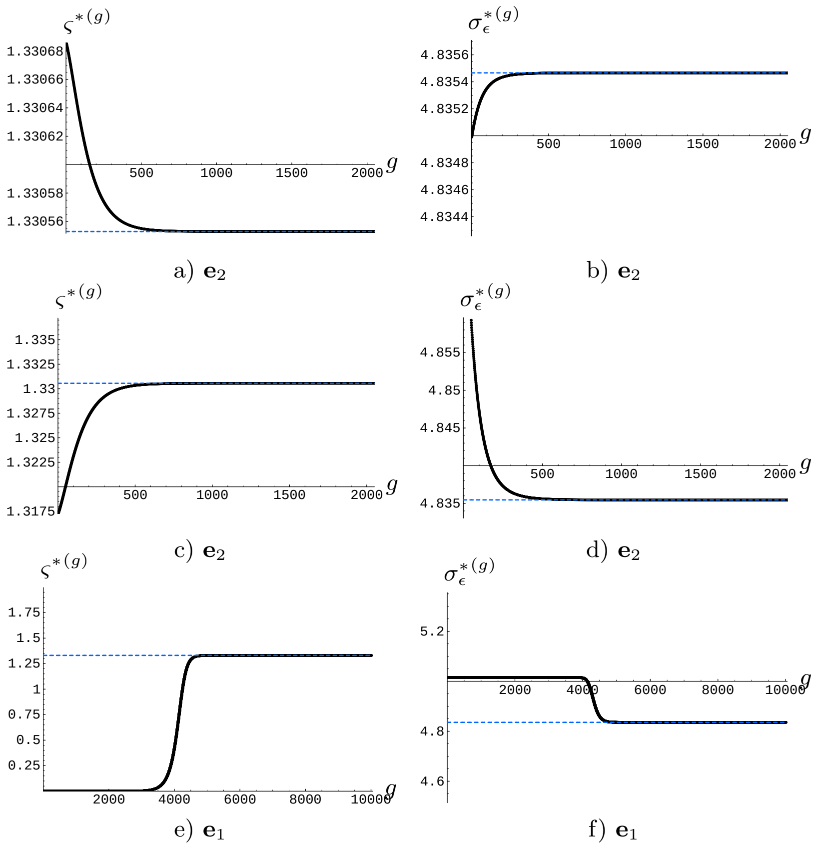
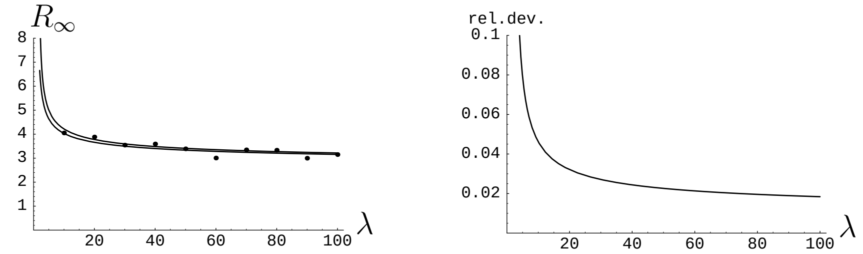


Fig. 5 Behavior of the evolution equations (30) and (31) close to the fixed points. As parameters $\lambda = 100$, $N = 100$, $\tau = 0.1$, and $\alpha = 2$ were chosen. The dashed lines represent the steady state mutation strength (40) and the noise strength (41), respectively.

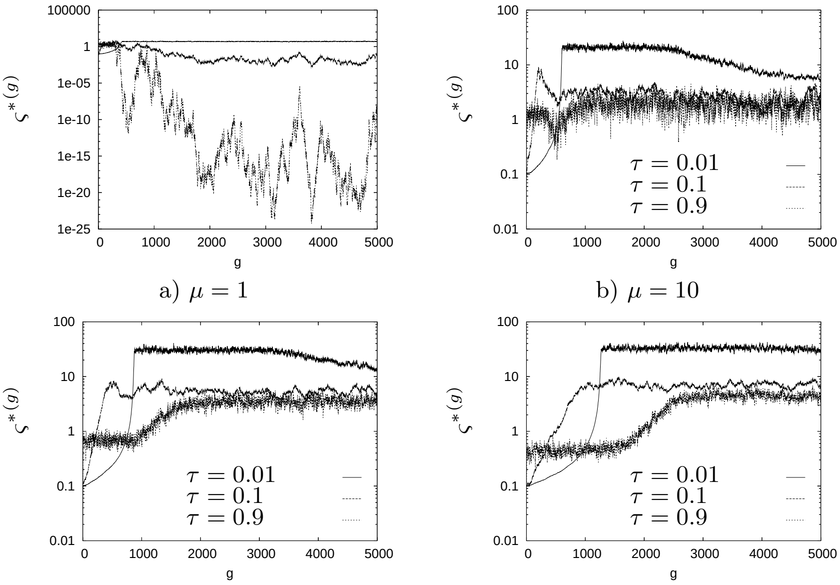
steady state ς^* (40). Sometimes short nearly stationary phases exist, but they appear only sporadically. The only observable tendency seems to be a general preference of small mutation strengths. That is, the predicted stationary mutation strength (40) cannot be observed after reaching the vicinity of R_∞. However, the resulting effect on the finally observed steady state r is rather small: Since any mutation strength between zero and (40) leads to nearly the same residual location error, both estimates (42) and $R_\infty^B = \sqrt[\alpha]{\sigma_\epsilon N/(2c\alpha c_{1,\lambda})}$ serve relatively well as predictors of the final R_∞ which can be seen in Fig. 6.

Interestingly, it can be seen in Fig. 7 that the non-existence of a final stationary state of the mutation strength seems to occur only in the case of $(1,\lambda)$-ES. If intermediate recombinative $(\mu/\mu_I,\lambda)$-ES are used, the behavior changes qualitatively: The mutation strength fluctuates very stably around



a) R_∞^A (42) and R_∞^B b) Relative deviation

Fig. 6 Final residual location errors as obtained by (42) and $R_\infty^B = \sqrt[\alpha]{\sigma_\epsilon N/(2c\alpha c_{1,\lambda})}$. The parameters were set to $N = 100$, $\tau = 0.1$, $\sigma_\epsilon = 1$, $c = 1$, and $\alpha = 2$. The points denote the results of $(1, \lambda)$-ES runs and each was averaged over $500,000$ generations. Figure b) shows the relative deviation of (42) from R_∞^B.



a) $\mu = 1$ b) $\mu = 10$

c) $\mu = 30$ d) $\mu = 50$

Fig. 7 The ς^*-evolution of some typical $(\mu/\mu_I, 100)$-ES runs ($N = 100$) on the sphere ($\alpha = 2$). Shown are the results for $\tau = 0.01$ (topmost curve), $\tau = 0.1$, and $\tau = 0.9$ (lowest curve). The duration of the initial steady state for $\varsigma^{*(g)}$ depends on τ and thus on the convergence velocity of the r-variable towards the final steady state.

a stationary value. We will discuss this interesting phenomenon in the next section.

3.3 On the Erratic Behavior of the $(1, \lambda)$-ES and a Possible Remedy

In order to discuss the steady state behavior of the ES, we should recall that the ES is operating in the high-noise regime. After having reached the vicinity of R_∞, the noise with strength σ_ϵ = const. is so large that it totally overshadows the actual fitness information. Thus, the selection process becomes nearly random, i.e., the ς^*-evolution is basically driven by random samples from a log-normal distribution with parameter τ. Under this condition, the probability of an in- or decrease of the mutation strength equals 1/2

$$\Pr\left(\varsigma^{*(g+1)} \leq \varsigma^{*(g)}\right) = \int_0^{\varsigma^{*(g)}} \frac{\mathrm{e}^{-\frac{(\ln(\varsigma^*/\varsigma^{*(g)}))^2}{2\tau^2}}}{\tau\varsigma^*\sqrt{2\pi}}\,\mathrm{d}\varsigma^*$$

$$= \int_{-\infty}^{0} \frac{\mathrm{e}^{-\frac{t^2}{2\tau^2}}}{\tau\sqrt{2\pi}}\,\mathrm{d}t = \Phi_{0,\tau^2}(0) = \frac{1}{2}. \qquad (49)$$

Put it another way, the ς^*-evolution of the $(1, \lambda)$-σSA-ES performs a biased random walk: It probabilistically accepts any ς^*-decrease, however, it punishes large ς^* values due to its selective disadvantage. As a result, the $(1, \lambda)$-σSA-ES has a slight tendency towards smaller mutation strengths. This is a clear disadvantage of the standard version of $(1, \lambda)$-σSA-ES. A possible remedy would be to increase the probability of ς^*-increases slightly. We will come back to this idea below.

The question arises why recombinative strategies exhibit a qualitatively different behavior. For sake of simplicity, we consider the case of an infinite number of parents. Without loss of generality, let $\varsigma^{*(g)} = 1$. Since the mutation strengths Y_i of the μ parents are independently identically distributed random variables with mean $m = \exp(\tau^2/2)$ and variance $s^2 = \exp(\tau^2)[\exp(\tau^2) - 1]$, the sum $1/\mu\sum_{i=1}^{\mu} Y_i$ converges to a normally distributed random variable $S \sim \mathcal{N}(m, s^2/\mu)$. If μ is sufficiently large, the probability that the mutation strength decreases can be estimated using the cdf of the normal distribution. The probability of $1/\mu\sum_{i=1}^{\mu} Y_i \leq 1$ becomes

$$Pr\left(\frac{1}{\mu}\sum_{i=1}^{\mu} Y_i \leq 1\right) \to \Phi\left(\sqrt{\mu}\frac{1 - \mathrm{e}^{\frac{\tau^2}{2}}}{\sqrt{\mathrm{e}^{\tau^2}(\mathrm{e}^{\tau^2} - 1)}}\right) \qquad (50)$$

which is smaller than 1/2 if $\tau > 0$. Actually, this preference for ς^*-increases can also be shown for the smallest parental population size $\mu = 2$. Therefore, an intermediate recombinative strategy possesses a natural tendency to provide more ς^*-increases than decreases.

As to the $(1, \lambda)$-ES, this suggests the introduction of a slight preference for ς^*-increases in the mutation operator by using a log-normal distribution

$$p_\sigma^*(\varsigma^*|\varsigma^{*(g)}) = \frac{1}{\varsigma^*\tau\sqrt{2\pi}}\exp\left(-\frac{\left(\ln(\varsigma^*/\varsigma^{*(g)}) - \beta\right)^2}{2\tau^2}\right) \qquad (51)$$

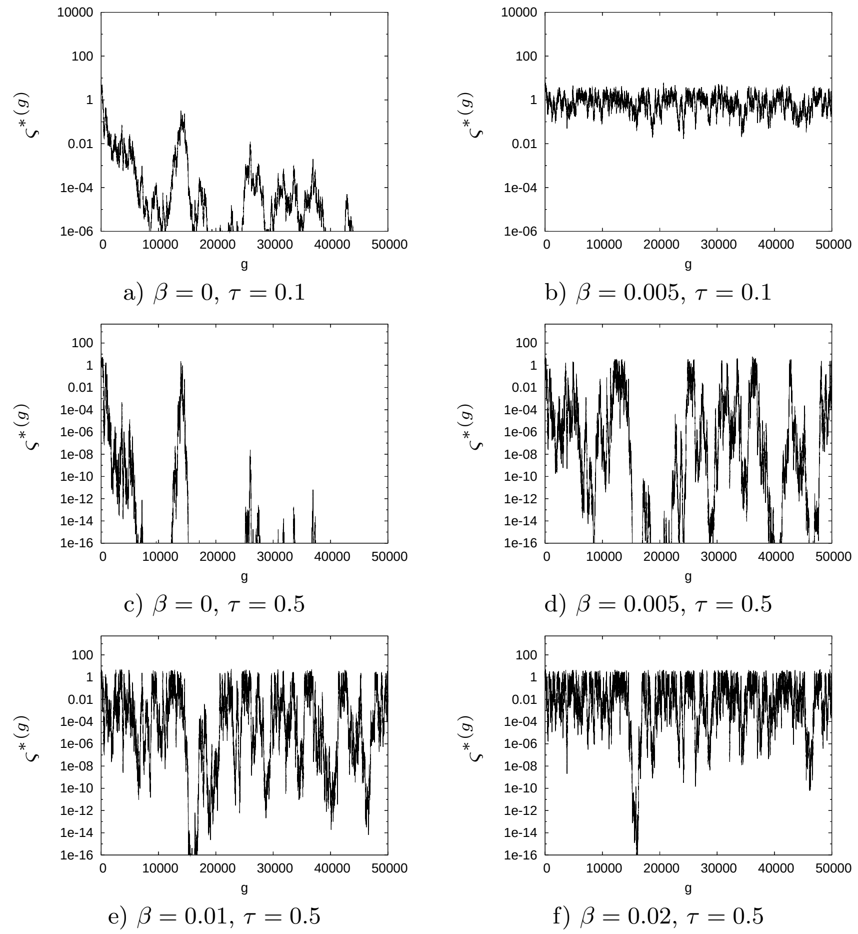


Fig. 8 Dynamics of the normalized mutation strength of $(1, \lambda)$-ES. Shown are the results of typical ES runs on the quadratic sphere. The dimension of the search space is $N = 100$ and the noise strength is set to $\sigma_\epsilon = 1$.

with a bias $\beta > 0$. The question remains how to choose β. On the one hand, it has to be sufficiently large to induce a trend towards higher mutation strengths. One the other hand considering the change $\sigma_l = \sigma^{(g)}\zeta$, the $\mathrm{E}[\zeta] \approx 1$ condition still has to be fulfilled.

Figure 8 shows the results of some ES-runs with different choices of β. The effect of the bias β also depends on the learning parameter. If τ is relatively high, the ES tends towards smaller values and shows irregular patterns. An increase of β changes the behavior. Larger learning rates seem to require larger biases in turn. One the other hand, a learning parameter that is too small may lead to divergent behavior.

In order to investigate this behavior theoretically, one can apply the techniques developed in this paper. In what follows, we will only sketch the derivations. Introducing $\beta > 0$ changes the raw moments of the log-normal distribution to $\overline{\varsigma^{*k}} = (\varsigma^{*(g)})^k \exp(k\beta)\exp(k^2\tau^2/2)$. Thus, if β is chosen sufficiently small, the approximations of the functions in (24) with their Taylor series are still valid. Therefore, the derivation of the SAR remains the same. The only change occurs in the last step of the calculation leading to (30), because the expectations of $[(\varsigma^* - \varsigma^{*(g)})/\varsigma^{*(g)}]^k$ w.r.t. the log-normal density with bias $\beta = 0$ must be replaced. Finally SAR (30) becomes

$$\begin{aligned}\psi(\varsigma^{*(g)}) &= \frac{\tau^2}{2}\mathrm{e}^{\beta} + (\varsigma^{*(g)})(A_t - A_g)\mathrm{e}^{2\beta}\tau^2 \\ &= \tau^2 \mathrm{e}^{\beta}\Big[\frac{1}{2} + \mathrm{e}^{\beta}\big(d^{(2)}_{1,\lambda} - 1\big)\frac{(\varsigma^{*(g)})^2}{(\sigma^*_\epsilon)^2 + (\varsigma^{*(g)})^2} \\ &\qquad - \frac{\mathrm{e}^{\beta} c_{1,\lambda}(\varsigma^{*(g)})^2}{\sqrt{(\sigma^*_\epsilon)^2 + (\varsigma^{*(g)})^2}}\Big]. \end{aligned} \tag{52}$$

We will now determine the stationary points, i.e., the solutions of $\varphi^* = 0$ and $\psi = 0$ using (16) and (52). The condition $\varphi^* = 0$ gives $(\varsigma^{*(g)})^2 + (\sigma^{*(g)}_\epsilon)^2 = 4c^2_{1,\lambda}$. Inserting this into (52) leads to the stationary mutation strength

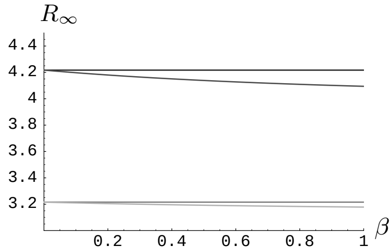
$$\begin{aligned} 0 &= \frac{1}{2} + \mathrm{e}^{\beta}\big(d^{(2)}_{1,\lambda} - 1\big)\frac{\varsigma^{*\,2}_{\infty}}{\sigma^{*2}_{\epsilon\,\infty} + \varsigma^{*\,2}_{\infty}} - \frac{\mathrm{e}^{\beta} c_{1,\lambda}\varsigma^{*\,2}_{\infty}}{\sqrt{\sigma^{*2}_{\epsilon\,\infty} + \varsigma^{*\,2}_{\infty}}} \\ \Rightarrow 0 &= \frac{1}{2} + \mathrm{e}^{\beta}\varsigma^{*\,2}_{\infty}\left(\frac{\big(d^{(2)}_{1,\lambda} - 1\big)}{4c^2_{1,\lambda}} - \frac{1}{2}\right) \\ \Rightarrow \varsigma^*_\infty &= \frac{2c_{1,\lambda}\mathrm{e}^{-\frac{\beta}{2}}}{\sqrt{2\big(2c^2_{1,\lambda} + 1 - d^{(2)}_{1,\lambda}\big)}}. \end{aligned} \tag{53}$$

Finally, the associated noise strength $\sigma^*_{\epsilon\,\infty} = 2c_{1,\lambda}\sqrt{1 - \frac{\mathrm{e}^{-\beta}}{2\left(2c^2_{1,\lambda}+1-d^{(2)}_{1,\lambda}\right)}}$ gives an estimate of the residual location error

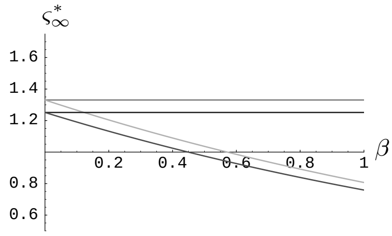
$$R^{\beta}_{\infty} = \sqrt[\alpha]{\frac{\sigma_\epsilon N}{2c\alpha c_{1,\lambda}}\sqrt{\frac{1}{1 - \frac{\mathrm{e}^{-\beta}}{2\left(2c^2_{1,\lambda}+1-d^{(2)}_{1,\lambda}\right)}}}}. \tag{54}$$

As can be shown numerically (see Fig. 9), as long as β is sufficiently small, the estimates (53) and (54) do not differ significantly from (40) and (42) obtained for $\beta = 0$.

Several caveats must be added here. It seems to be difficult to find a value of β that on the one hand raises the mutation strength sufficiently and on the other hand does not lead to a deterioration of the residual location error. In addition, the estimates only hold for sufficiently small β-values and they do not account for the interplay with the learning parameter τ. Considering the results of the experiments (see Fig. 10), one finds that in the case of larger β-values, i.e., here already for $\beta \geq 0.01$, the predicted mutation strength (53) is lower than the experimentally observed one. Also, the ES shows a significant higher sensitivity w.r.t. choice of β than predicted by (53). These deviations clearly indicate the limits of the deterministic analysis presented.



a) residual location error



b) mutation strength

Fig. 9 Comparison of the predictions of the stationary mutation strength and the residual location error. Figure a) shows the prediction obtained by R^A_∞ (42) and R^β_∞ (54). Figure b) compares the mutation strengths (53) and (40). The dimension is $N = 100$, the noise strength $\sigma_\epsilon = 1$, and $\alpha = 2$. The gray lines indicate the results for $\lambda = 100$ whereas the black stand for $\lambda = 10$.

4 Conclusions

In this paper, we have investigated the self-adaptation of $(1, \lambda)$-ES on the noisy sphere model. The evolution of the distance to the optimal point and of the mutation strength can be modeled by evolution equations – difference equations describing the one-generation change.

In general, they can be divided into stochastic and deterministic parts. The latter denote the expected change leading in the case of the distance to the progress rate. The mutation strength is generally changed multiplicatively. Thus, the so-called self-adaptation response is the expected relative change of the mutation parameter.

After obtaining equations describing the self-adaptation response and the progress rate, a deterministic approach was applied and the stochastic parts of the evolution equations were neglected.

The evolution equations can be used to characterize the steady state and can be used to analyze the ES's behavior. It has been shown that the optimal learning rate scales with $1/\sqrt{N}$ and depends on the normalized noise strength.

In the case of a constant noise strength σ_ϵ, three different phases of the evolution have been identified. As long as the system is still far away from

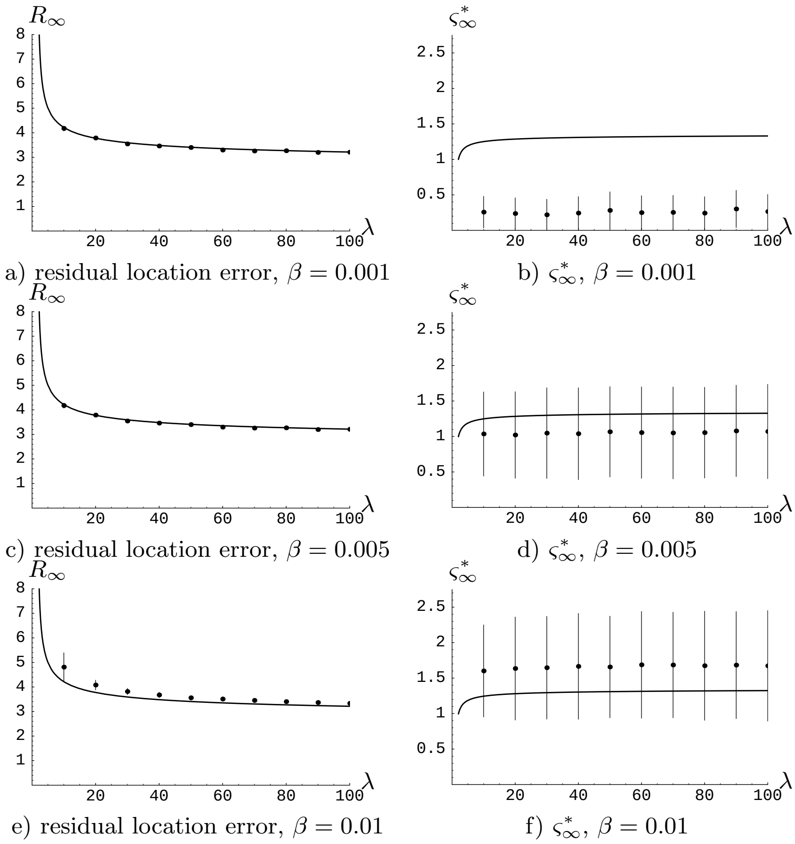


Fig. 10 Comparison of the predictions of the stationary mutation strength (53) and the residual location error (54) with the results of experiments on the sphere function for some choices of β. The search space dimension is $N = 100$, the noise was set to $\sigma_\epsilon = 1$, and $\tau = 0.1$ was chosen as the learning parameter. Each data point was averaged over $500,000$ generations. The vertical bars indicate the measured standard deviations.

the optimum, the influence of the noise can be neglected. As a result, the ES reaches a similar stationary mutation strength as in the noise free case and the same recommendations for choosing the learning parameter apply.

Approaching the optimum, however, changes the situation. Due to the increasing normalized noise, the steady state of the mutation strength is lost. The progress gets smaller and smaller until the ES cannot get any closer to the optimum on average. The progress rate becomes zero. This can be used to determine the residual location error. There are two estimates that can be obtained. The first is associated with a vanishing mutation strength, the other demands stationarity of the mutation strength evolution by requiring additionally the SAR to be zero. Interestingly, both are very similar especially if large offspring population sizes are considered.

A remarkable observation is that the $(1, \lambda)$-ES is not able to stabilize the mutation strength although the deterministic approach predicts a locally stable non-zero mutation strength. Instead its behavior resembles a random walk where the mutation strength fluctuates between the non-zero mutation strength (40) and zero. A general preference of small values can be observed. Since any mutation strength between these two extremes leads nearly to the same residual location error, the estimates that were obtained lead to good predictions.

The reason for the behavior of $(1, \lambda)$-ES cannot be explained by considering the deterministic approximation. Comparing the behavior of $(1, \lambda)$-ES with that of intermediate $(\mu/\mu_I, \lambda)$-ES, one finds that the latter show a second stationary phase of the mutation strength once the system has reached the vicinity of the residual localization error. The difference in the behavior is clearly due to the missing recombination of the mutation strength. If the normalized mutation strength is considerably smaller than the normalized noise strength, the ES is nearly unable to choose the offspring on basis of the actual fitness values. Instead – concerning the mutation strength – the selection is similar to a random sampling of log-normally distributed variables.

Using intermediate recombination introduces a probabilistic preference towards an increase of the mutation strength whereas an $(1, \lambda)$-ES de- and increases the mutation strength with the same probability. Thus, $(\mu/\mu_I, \lambda)$-ES will tend to increase a small mutation strength until it is sufficiently large so that the information obtained by the fitness function is taken into account. As far as the constant noise scenario is considered this “bias” can be regarded as a desirable property of intermediate recombination.

The $(1, \lambda)$-ES on the sphere model has a slight bias towards a decrease of the mutation strength. This explains the wandering behavior of the mutation strength. It can be remedied to a certain extend by introducing a slight bias in the σ mutation operator.

While this paper has provided first insights into the mechanism of self-adaptation of the ES on the noisy sphere, the investigations are far from being complete. First, our considerations did not explicitly take into account the stochasticity of the evolutionary process. Especially in the high noise regime, the deterministic approximation leads to predictions which are not fully consonant with the observed dynamics. Therefore, incorporating fluctuations and solving the corresponding Chapman-Kolmogorov-Equations remains as a task for the future. Additionally, and even with higher priority, the case of intermediate recombination remains to be considered. As we have learned in this work, besides the genetic repair effect responsible for larger progress rates, intermediate recombination introduces an additional bias which seems to be beneficial in highly noisy environments. Therefore, the theoretical investigation of these strategies should be considered with high priority.

Acknowledgments

This work was supported by the Research Center for Process- and Product-Engineering at the Vorarlberg University of Applied Sciences and by the Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Center SFB 531 at the University of Dortmund.

A Derivation of the Density $p(\tilde{Q}|\varsigma^*, \sigma_\epsilon^{*(g)}, r^{(g)})$

The noise is modeled by a normally distributed random variable ϵ with zero mean and standard deviation σ_ϵ. The observed fitness $\tilde{Q} = F + \epsilon$ has therefore the pdf

$$p_\epsilon(\tilde{Q}|F) = \frac{1}{\sigma_\epsilon\sqrt{2\pi}} \mathrm{e}^{-\frac{(\tilde{Q}-F)^2}{2\sigma_\epsilon^2}} \tag{55}$$

given the actual fitness F. Recall that the actual fitness value is a function of the distance to the optimum r. Let $\mathbf{r} = \mathbf{y}^{(g)} - \hat{\mathbf{y}}$ be the distance vector of the parent. The distance vector $\mathbf{R}$ of an offspring is created by adding a mutation vector $\mathbf{z}$ such that $\mathbf{R} = \mathbf{r} + \mathbf{z}$. To continue, we will decompose the mutation vector such that $\mathbf{z} = -x\mathbf{e}_r + \mathbf{h}$, with $-x\mathbf{e}_r$ pointing in the direction of the optimum and $\mathbf{h}^{\mathrm{T}}\mathbf{e}_r = 0$. Thus, $F = f(R)$ can be expressed by

$$F = f\left(r\sqrt{(1 - x/r)^2 + \|\mathbf{h}\|^2/r^2}\right). \tag{56}$$

This decomposition can be used to derive the pdf of $\tilde{Q}$.

Re-ordering the terms in the square root to $1 + 2[-x/r(1 - x/r) + \|\mathbf{h}\|^2/(2r)^2)] := 1 + \delta$ enables us to substitute $\sqrt{1+\delta}$ with its Taylor series at $\delta = 0$, leading to $\sqrt{1+\delta} = 1 + \delta/2 + \mathcal{O}(\delta^2)$. Provided that $x \ll r$ and $\|\mathbf{h}\|^2 \ll r$, all higher order terms can be neglected and F (56) be approximated with

$$F = f\left(r + r\Big(-\frac{x}{r} + \frac{\|\mathbf{h}\|^2}{2r^2}\Big)\right). \tag{57}$$

The function f can in turn be substituted by its Taylor series at r, i. e.,

$$F = f(r) - f'(r)x + \frac{f'(r)}{2r}\|\mathbf{h}\|^2 + \dots. \tag{58}$$

Inserting the approximation for F (58) into the pdf of $\tilde{Q}$ (55) gives

$$p_\epsilon(\tilde{Q}|x, \|\mathbf{h}\|^2) = \frac{1}{\sigma_\epsilon\sqrt{2\pi}}\exp\left[\frac{(\tilde{Q} - f(r) - f'(r)x + \frac{f'(r)}{2r}\|\mathbf{h}\|^2)^2}{2\sigma_\epsilon^2}\right] \tag{59}$$

which still depends on x and $u = \|\mathbf{h}\|^2$. Their densities can be obtained easily. Since x can be assumed w.l.o.g. to be the first component of the mutation vector, it is simply a normally distributed random variable with zero mean and standard deviation ς. The variable $u = \|\mathbf{h}\|^2 = \sum_{i=2}^{N} h_i^2$ follows a χ^2-distribution which can be modeled using a normal distribution with mean $N\varsigma^2$ and standard deviation $\sqrt{2N}\varsigma^2$ provided that N is sufficiently large.

Let us now consider $\tilde{Q} \le \tilde{q}$. Since $\tilde{Q} = f(R) + \epsilon$, this equals $\epsilon - f'(r)x + f'(r)/(2r)u \le \tilde{q} - f(r)$ if (58) is taken into account. Since all random variables on the left hand side are normally distributed, their sum is also normally distributed with mean $f'(r)N/(2r)\varsigma^2$ and variance $\sigma_\epsilon^2 + \big(f'(r)\big)^2\varsigma^2[1 + N/(2r^2)\varsigma^2]$. Thus, we arrive at

$$p_\epsilon(\tilde{Q}|\sigma, r) = \frac{1}{\sqrt{\sigma_\epsilon^2 + \big(f'(r)\big)^2\varsigma^2\Big(1 + \frac{N}{2r^2}\varsigma^2\Big)}\sqrt{2\pi}} \times \exp\left[-\frac{(\tilde{Q} - f(r) - \frac{f'(r)N}{2r}\varsigma^2)^2}{2\Big(\sigma_\epsilon^2 + \big(f'(r)\big)^2\varsigma^2\Big(1 + \frac{N}{2r^2}\varsigma^2\Big)\Big)}\right]. \tag{60}$$

If the usual normalizations are introduced, that is, $\varsigma^* = \varsigma N/r$, and $\sigma_\epsilon^* = \sigma_\epsilon N/(rf'(r))$, the density becomes

$$p_\epsilon(\tilde{Q}|\varsigma^*) = \frac{1}{\sqrt{\sigma_\epsilon^2 + \frac{\big(f'(r)\big)^2 r^2}{N^2}\varsigma^{*2}\big(1 + \frac{\varsigma^{*2}}{2N}\big)}\sqrt{2\pi}} \times \exp\left[-\frac{(\tilde{Q} - f(r) - \frac{f'(r)r}{2N}\varsigma^{*2})^2}{2\Big(\sigma_\epsilon^2 + \frac{(f'(r))^2 r^2}{N^2}\varsigma^{*2}\big(1 + \frac{\varsigma^{*2}}{2N}\big)\Big)}\right] \tag{61}$$

and finally

$$p(\tilde{Q}|\varsigma^*, \sigma_\epsilon^*, r) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\sigma_\epsilon^2 + \Big(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\Big)^2\varsigma^{*2}\big(1 + \frac{(\varsigma^*)^2}{2N}\big)}} \times \exp\left[-\frac{1}{2}\left(\frac{\tilde{Q} - F^{(g)} - \frac{\sigma_\epsilon}{2\sigma_\epsilon^*}\varsigma^{*2}}{\sqrt{\sigma_\epsilon^2 + \Big(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\Big)^2\varsigma^{*2}\big(1 + \frac{(\varsigma^*)^2}{2N}\big)}}\right)^2\right]. \tag{62}$$

Since we will consider $N \to \infty$ in this paper, the density can be simplified to

$$p(\tilde{Q}|\varsigma^*, \sigma_\epsilon^*, r) = \frac{1}{\sqrt{2\pi\Big(\sigma_\epsilon^2 + \Big(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\Big)^2\varsigma^{*2}\Big)}} \times \exp\left[-\frac{1}{2}\left(\frac{\tilde{Q} - F^{(g)} - \frac{\sigma_\epsilon}{2\sigma_\epsilon^*}\varsigma^{*2}}{\sqrt{\sigma_\epsilon^2 + \Big(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\Big)^2\varsigma^{*2}}}\right)^2\right]. \tag{63}$$

B The SAR

To simplify the notations, let $\sigma^* := \varsigma^{*(g)}$ and $\sigma^*_\epsilon := \sigma^{*(g)}_\epsilon$. We will now consider the first order self-adaptation response function

$$\psi(\sigma^*, \sigma^*_\epsilon, r^{(g)}) = \lambda \int_0^\infty \int_{-\infty}^\infty \left(\frac{\varsigma^* - \sigma^*}{\sigma^*}\right) p^*_\sigma(\varsigma^*|\sigma^*) \\ \times \exp\left[-\frac{1}{2}\left(\frac{\tilde{Q} - F^{(g)} - \frac{\sigma_\epsilon}{2\sigma^*_\epsilon}\varsigma^{*2}}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma^*_\epsilon}\right)^2 \varsigma^{*2}}}\right)^2\right] \\ \times \frac{1}{\sqrt{2\pi}\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma^*_\epsilon}\right)^2 \varsigma^{*2}}} \, \mathrm{d}\varsigma^* \\ \times \left(1 - P(\tilde{Q}|\sigma^*, \sigma^*_\epsilon, r^{(g)})\right)^{\lambda-1} \mathrm{d}\tilde{Q} \tag{64}$$

where the cumulative distribution function $P(\tilde{Q}|\sigma^*, r^{(g)})$ (see Appendix A) is given by

$$P(\tilde{Q}|\sigma^*, \sigma^{*(g)}_\epsilon, r^{(g)}) = \int_0^\infty \Phi\left(\frac{\tilde{Q} - F^{(g)} - \frac{\sigma_\epsilon}{2\sigma^*_\epsilon}\varsigma^{*2}}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma^*_\epsilon}\right)^2 \varsigma^{*2}}}\right) p^*_\sigma(\varsigma^*|\sigma^*) \, \mathrm{d}\varsigma^*. \tag{65}$$

It has been shown in [4] that in the limit case $\tau \to 0$, the distribution function (65) can be approximated by

$$P(\tilde{Q}|\sigma^*, \sigma^*_\epsilon, r^{(g)}) = \Phi\left(\frac{\tilde{Q} - F^{(g)} - \frac{\sigma_\epsilon}{2\sigma^*_\epsilon}\sigma^{*2}}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma^*_\epsilon}\right)^2 \sigma^{*2}}}\right) \tag{66}$$

if log-normal or two-point distributions are considered. Provided that $\tau \ll 1$, (64) simplifies to

$$\psi(\sigma^*, \sigma^*_\epsilon, r^{(g)}) = \lambda \int_0^\infty \int_{-\infty}^\infty \left(\frac{\varsigma^* - \sigma^*}{\sigma^*}\right) p^*_\sigma(\varsigma^*|\sigma^*) \\ \times \left[1 - \Phi\left(\frac{\tilde{Q} - F^{(g)} - \frac{\sigma_\epsilon}{2\sigma^*_\epsilon}\sigma^{*2}}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma^*_\epsilon}\right)^2 \sigma^{*2}}}\right)\right]^{\lambda-1} \\ \times \exp\left[-\frac{1}{2}\left(\frac{\tilde{Q} - F^{(g)} - \frac{\sigma_\epsilon}{2\sigma^*_\epsilon}\varsigma^{*2}}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma^*_\epsilon}\right)^2 \varsigma^{*2}}}\right)^2\right]$$

$$\times \frac{1}{\sqrt{2\pi}\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\right)^2 \varsigma^{*2}}} \,\mathrm{d}\tilde{Q}\,\mathrm{d}\varsigma^* \tag{67}$$

At first, we simplify the expression inside the Φ-function. Substituting

$$-t = \frac{\tilde{Q} - F^{(g)} - \frac{\sigma_\epsilon}{2\sigma_\epsilon^*}\sigma^{*2}}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\right)^2 \sigma^{*2}}}, \tag{68}$$

it follows

$$\tilde{Q} = -t\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\right)^2 \sigma^{*2}} + F^{(g)} + \frac{\sigma_\epsilon}{2\sigma_\epsilon^*}\sigma^{*2} \tag{69}$$

leading to

$$\begin{aligned}
\psi(\sigma^*, \sigma_\epsilon^*) \;=\; & \frac{\lambda}{\sqrt{2\pi}} \int_0^\infty \int_{-\infty}^\infty \left(\frac{\varsigma^* - \sigma^*}{\sigma^*}\right) p_\sigma^*(\varsigma^*|\sigma^*) \sqrt{\frac{\sigma_\epsilon^2 + (\frac{\sigma_\epsilon}{\sigma_\epsilon^*})^2\sigma^{*2}}{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\right)^2 \varsigma^{*2}}} \\
& \times \exp\left[-\frac{1}{2}\left(\sqrt{\frac{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\right)^2\sigma^{*2}}{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\right)^2 \varsigma^{*2}}}\, t - \frac{\frac{\sigma_\epsilon}{2\sigma_\epsilon^*}\left(\sigma^{*2} - \varsigma^{*2}\right)}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\right)^2 \varsigma^{*2}}}\right)^2\right] \\
& \times \Phi(t)^{\lambda-1}\,\mathrm{d}t\,\mathrm{d}\varsigma^* \\
:= \; & \frac{\lambda}{\sqrt{2\pi}} \int_0^\infty \int_{-\infty}^\infty \left(\frac{\varsigma^* - \sigma^*}{\sigma^*}\right)^k p_\sigma^*(\varsigma^*|\sigma^*) \\
& \times g(\varsigma^*) h(\varsigma^*, t) \Phi(t)^{\lambda-1}\,\mathrm{d}t\,\mathrm{d}\varsigma^*.
\end{aligned} \tag{70}$$

Note, (70) does not depend on $r^{(g)}$. To derive an expression for the SAR, we need to integrate over ς^* and t. In general, the terms of (70) still do not allow for a analytical solution, so further simplifications have to be introduced.

Basically, we will expand the more complicated functions of ς^* into Taylor series about σ^*. Since we consider the limit case $\tau \to 0$, it can be assumed that $|\varsigma^* - \sigma^*| \ll 1$ holds with an overwhelming probability. The Taylor series can soon be broken off without introducing larger errors. Let us first consider the function

$$g(\varsigma^*) = \sqrt{\frac{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\right)^2 \sigma^{*2}}{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\right)^2 \varsigma^{*2}}}. \tag{71}$$

The first derivations are given by

$$\frac{\partial}{\partial \varsigma^*} g(\varsigma^*) = -\sqrt{\frac{(\sigma_\epsilon^*)^2 + \sigma^{*2}}{\left[(\sigma_\epsilon^*)^2 + \varsigma^{*2}\right]^3}}\,\varsigma^* \tag{72}$$

and

$$\frac{\partial^2}{\partial\varsigma^{*2}} g(\varsigma^*) = \sqrt{\frac{(\sigma_\epsilon^*)^2 + \sigma^{*2}}{\left[(\sigma_\epsilon^*)^2 + \varsigma^{*2}\right]^3}} \left[\frac{3\varsigma^{*2}}{(\sigma_\epsilon^*)^2 + \varsigma^{*2}} - 1\right]. \tag{73}$$

The Taylor series around σ^* reads

$$\begin{aligned} T_g(\varsigma^*) \;=\;& 1 - \frac{\sigma^*}{(\sigma_\epsilon^*)^2 + \sigma^{*2}}\big(\varsigma^* - \sigma^*\big) + \frac{\big(\varsigma^* - \sigma^*\big)^2}{2\left((\sigma_\epsilon^*)^2 + \sigma^{*2}\right)} \\ &\times \left[\frac{3\sigma^{*2}}{(\sigma_\epsilon^*)^2 + \sigma^{*2}} - 1\right] + \mathcal{O}\Big((\varsigma^* - \sigma^*)^3\Big) \\ :=\;& 1 - A_g\big(\varsigma^* - \sigma^*\big) + \frac{B_g}{2}\big(\varsigma^* - \sigma^*\big)^2 + \mathcal{O}\Big((\varsigma^* - \sigma^*)^3\Big). \end{aligned} \tag{74}$$

Let h denote the exponential function

$$h(\varsigma^*, t) = \exp\left[-\frac{1}{2}\left(\sqrt{\frac{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\right)^2\sigma^{*2}}{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\right)^2\varsigma^{*2}}}\, t - \frac{\frac{\sigma_\epsilon}{2\sigma_\epsilon^*}\left(\sigma^{*2} - \varsigma^{*2}\right)}{\sqrt{\sigma_\epsilon^2 + \left(\frac{\sigma_\epsilon}{\sigma_\epsilon^*}\right)^2\varsigma^{*2}}}\right)^2\right].$$

The function $h(\varsigma^*, t)$ will also be expanded into a Taylor series around σ^*. The calculations are rather lengthy. The use of a computer algebra system is recommended. One obtains

$$h(\sigma^*, t) = \mathrm{e}^{-\frac{t^2}{2}} \tag{75}$$

$$\frac{\partial}{\partial\varsigma^*} h(\sigma^*, t) = \frac{\mathrm{e}^{-\frac{t^2}{2}}\sigma^*}{\sqrt{(\sigma_\epsilon^*)^2 + \sigma^{*2}}}\left[\frac{t^2}{\sqrt{(\sigma_\epsilon^*)^2 + \sigma^{*2}}} - t\right] \tag{76}$$

$$\begin{aligned} \frac{\partial^2}{\partial\varsigma^{*2}} h(\sigma^*, t) = \mathrm{e}^{-\frac{t^2}{2}}\Bigg[& \frac{-\sigma^{*2}}{(\sigma_\epsilon^*)^2 + \sigma^{*2}} \\ & - \frac{t}{\sqrt{(\sigma_\epsilon^*)^2 + \sigma^{*2}}}\left[1 - \frac{4\sigma^{*2}}{(\sigma_\epsilon^*)^2 + \sigma^{*2}}\right] \\ & + \frac{t^2}{(\sigma_\epsilon^*)^2 + \sigma^{*2}}\left[1 + \sigma^{*2} - \frac{4\sigma^{*2}}{(\sigma_\epsilon^*)^2 + \sigma^{*2}}\right] \\ & - \frac{2\sigma^{*2}}{\sqrt{(\sigma_\epsilon^*)^2 + \sigma^{*2}}^3} t^3 + \frac{\sigma^{*2}}{\left[(\sigma_\epsilon^*)^2 + \sigma^{*2}\right]^2} t^4\Bigg]. \end{aligned} \tag{77}$$

If we regroup the Taylor series $T_h(\varsigma^*, t)$ into terms of t^k, we get

$$T_h(\varsigma^*, t) \;=\; \mathrm{e}^{-\frac{t^2}{2}}\Bigg\{1 - \frac{\big(\varsigma^* - \sigma^*\big)^2}{2}\frac{\sigma^{*2}}{(\sigma_\epsilon^*)^2 + \sigma^{*2}}$$

$$
\begin{aligned}
&+\Bigg[\big(\varsigma^*-\sigma^*\big)\Bigg(-\frac{\sigma^*}{\sqrt{(\sigma_\epsilon^*)^2+\sigma^{*2}}}\Bigg)\\
&-\frac{\big(\varsigma^*-\sigma^*\big)^2}{2}\Bigg(\frac{1}{\sqrt{(\sigma_\epsilon^*)^2+\sigma^{*2}}}\Bigg(1-\frac{4\sigma^{*2}}{(\sigma_\epsilon^*)^2+\sigma^{*2}}\Bigg)\Bigg)\Bigg]t\\
&+\Bigg[\big(\varsigma^*-\sigma^*\big)\Bigg(\frac{\sigma^*}{(\sigma_\epsilon^*)^2+\sigma^{*2}}\Bigg)\\
&+\frac{\big(\varsigma^*-\sigma^*\big)^2}{2}\Bigg(\frac{1+\sigma^{*2}}{(\sigma_\epsilon^*)^2+\sigma^{*2}}-\frac{4\sigma^{*2}}{\big[(\sigma_\epsilon^*)^2+\sigma^{*2}\big]^2}\Bigg)\Bigg]t^2\\
&-\frac{\big(\varsigma^*-\sigma^*\big)^2}{2}\Bigg[\frac{2\sigma^{*2}}{\sqrt{(\sigma_\epsilon^*)^2+\sigma^{*2}}^3}\Bigg]t^3\\
&+\frac{\big(\varsigma^*-\sigma^*\big)^2}{2}\Bigg[\frac{\sigma^{*2}}{\big[(\sigma_\epsilon^*)^2+\sigma^{*2}\big]^2}\Bigg]t^4+\dots\Bigg\}\\
:=\ &\mathrm{e}^{-\frac{t^2}{2}}\Big[h_0+h_1t+h_2t^2+h_3t^3+h_4t^4+\dots\Big]. \qquad (78)
\end{aligned}
$$

The neglected terms stem from higher-order terms of the Taylor series and are of order $\mathcal{O}([\varsigma^*-\sigma^*]^3)$. Substituting functions g and h in (70) with their Taylor series (74) and (78) leads to the SAR

$$
\begin{aligned}
\psi(\sigma^*,\sigma_\epsilon^*) &= \frac{\lambda}{\sqrt{2\pi}}\int_0^\infty\int_{-\infty}^\infty\Bigg(\frac{\varsigma^*-\sigma^*}{\sigma^*}\Bigg)p_\sigma^*(\varsigma^*|\sigma^*)\\
&\quad\times g(\varsigma^*)h(\varsigma^*,t)\Phi(t)^{\lambda-1}\,\mathrm{d}t\,\mathrm{d}\varsigma^*\\
&=\int_0^\infty\Bigg(\frac{\varsigma^*-\sigma^*}{\sigma^*}\Bigg)^k\Big[1-A_g\big(\varsigma^*-\sigma^*\big)\\
&\quad+\frac{B_g}{2}\big(\varsigma^*-\sigma^*\big)^2\Big]p_\sigma^*(\varsigma^*|\sigma^*)\\
&\quad\times\frac{\lambda}{\sqrt{2\pi}}\int_{-\infty}^\infty\mathrm{e}^{-\frac{t^2}{2}}\Phi(t)^{\lambda-1}\Big[h_0+h_1t\\
&\quad+h_2t^2+h_3t^3+h_4t^4+\dots\Big]\,\mathrm{d}t\,\mathrm{d}\varsigma^*. \qquad (79)
\end{aligned}
$$

Performing the integration over t and taking the definition of the progress coefficients $d_{1,\lambda}^{(k)}$, Eq. (26), into account leads to

$$
\begin{aligned}
I_t &= \frac{\lambda}{\sqrt{2\pi}}\int_{-\infty}^\infty\mathrm{e}^{-\frac{t^2}{2}}\Phi(t)^{\lambda-1}\\
&\quad\times\Big[h_0+h_1t+h_2t^2+h_3t^3+h_4t^4+\dots\Big]\,\mathrm{d}t\\
&= h_0+h_1c_{1,\lambda}+h_2d_{1,\lambda}^{(2)}+h_3d_{1,\lambda}^{(3)}+h_4d_{1,\lambda}^{(4)}+\dots. \qquad (80)
\end{aligned}
$$

The coefficient $c_{1,\lambda} := d^{(1)}_{1,\lambda}$ denotes a special case of the progress coefficients. The integration over ς^* remains to be done. To this end, I_t will be re-ordered into terms of $(\varsigma^* - \sigma^*)^k$

$$
\begin{aligned}
T(\varsigma^*) \;=\; & 1 + \big(\varsigma^* - \sigma^*\big)\left[\frac{d^{(2)}_{1,\lambda}\sigma^*}{(\sigma^*_\epsilon)^2 + \sigma^{*2}} - \frac{c_{1,\lambda}\sigma^*}{\sqrt{(\sigma^*_\epsilon)^2 + \sigma^{*2}}}\right] \\
& + \frac{\big(\varsigma^* - \sigma^*\big)^2}{2}\Bigg[- \frac{\sigma^{*2}}{(\sigma^*_\epsilon)^2 + \sigma^{*2}} \\
& - \frac{c_{1,\lambda}}{\sqrt{(\sigma^*_\epsilon)^2 + \sigma^{*2}}}\left(1 - \frac{4\sigma^{*2}}{(\sigma^*_\epsilon)^2 + \sigma^{*2}}\right) \\
& + d^{(2)}_{1,\lambda}\left(\frac{1 + \sigma^{*2}}{(\sigma^*_\epsilon)^2 + \sigma^{*2}} - \frac{4\sigma^{*2}}{\big[(\sigma^*_\epsilon)^2 + \sigma^{*2}\big]^2}\right) \\
& - \frac{2d^{(3)}_{1,\lambda}\sigma^{*2}}{\sqrt{(\sigma^*_\epsilon)^2 + \sigma^{*2}}^3} + \frac{d^{(4)}_{1,\lambda}\sigma^{*2}}{\big[(\sigma^*_\epsilon)^2 + \sigma^{*2}\big]^2}\Bigg] + \mathcal{O}\big((\varsigma^* - \sigma^*)^3\big) \\
:= & \; 1 + A_t(\varsigma^* - \sigma^*) + B_t\frac{(\varsigma^* - \sigma^*)^2}{2} + \mathcal{O}\Big((\varsigma^* - \sigma^*)^3\Big). \qquad (81)
\end{aligned}
$$

Using the Taylor series of g (74) and the result (81) for the integral (79), the SAR (70) changes to

$$
\begin{aligned}
\psi(\sigma^*, \sigma^*_\epsilon) &= \frac{\lambda}{\sqrt{2\pi}}\int_0^\infty \int_{-\infty}^\infty \left(\frac{\varsigma^* - \sigma^*}{\sigma^*}\right) p^*_\sigma(\varsigma^*|\sigma^*) \\
&= \times g(\varsigma^*) h(\varsigma^*, t) \Phi(t)^{\lambda-1}\,\mathrm{d}t\,\mathrm{d}\varsigma^* \\
&= \int_0^\infty \left(\frac{\varsigma^* - \sigma^*}{\sigma^*}\right)\Bigg(\left[1 - A_g\big(\varsigma^* - \sigma^*\big) + \frac{B_g}{2}\big(\varsigma^* - \sigma^*\big)^2\right] \\
&\quad \times\left[1 + A_t\big(\varsigma^* - \sigma^*\big) + \frac{B_t}{2}\big(\varsigma^* - \sigma^*\big)^2\right] \\
&\quad + \mathcal{O}\Big((\varsigma^* - \sigma^*)^3\Big)\Bigg) p^*_\sigma(\varsigma^*|\sigma^*)\,\mathrm{d}\varsigma^* \\
&= \int_0^\infty \left(\frac{\varsigma^* - \sigma^*}{\sigma^*}\right)\Bigg(1 + (A_t - A_g)\big(\varsigma^* - \sigma^*\big) \\
&\quad + \Big(\frac{B_t + B_g}{2} - A_t A_g\Big)\big(\varsigma^* - \sigma^*\big)^2 \\
&\quad + \mathcal{O}\Big((\varsigma^* - \sigma^*)^3\Big)\Bigg) p^*_\sigma(\varsigma^*|\sigma^*)\,\mathrm{d}\varsigma^* \\
&= \int_0^\infty \left(\left(\frac{\varsigma^* - \sigma^*}{\sigma^*}\right) + \sigma^*(A_t - A_g)\left(\frac{\varsigma^* - \sigma^*}{\sigma^*}\right)^2\right.
\end{aligned}
$$

$$+\sigma^{*2}\Big(\frac{B_t+B_g}{2}-A_tA_g\Big)\left(\frac{\varsigma^*-\sigma^*}{\sigma^*}\right)^3$$
$$+\mathcal{O}\Big((\varsigma^*-\sigma^*)^4\Big)\Bigg)p_\sigma^*(\varsigma^*|\sigma^*)\,\mathrm{d}\varsigma^*$$
$$=\overline{\left(\frac{\varsigma^*-\sigma^*}{\sigma^*}\right)}+\sigma^*(A_t-A_g)\overline{\left(\frac{\varsigma^*-\sigma^*}{\sigma^*}\right)^2}$$
$$+\sigma^{*2}\Big(\frac{B_t+B_g}{2}-A_tA_g\Big)\overline{\left(\frac{\varsigma^*-\sigma^*}{\sigma^*}\right)^3}$$
$$+\mathcal{O}\Big(\overline{(\varsigma^*-\sigma^*)^4}\Big). \qquad (82)$$

In the following, we will concentrate on the log-normal operator. Its expectation $\overline{\varsigma^{*k}}$ is given by $\overline{\varsigma^{*k}}=\sigma^{*k}\exp(\tau^2k^2/2)$. The expected values of $\left(\frac{\varsigma^*-\sigma^*}{\sigma^*}\right)^k$ can be easily computed (see Appendix C). Using the Taylor expansion of the resulting exponential function up to the fourth power of τ, (82) becomes

$$\psi(\sigma^*,\sigma^*_\epsilon)=\Big(\frac{\tau^2}{2}+\frac{\tau^4}{8}\Big)+\sigma^*(A_t-A_g)\Big(\tau^2+\frac{7\tau^4}{4}\Big)$$
$$+\sigma^{*2}\Big(\frac{B_t+B_g}{2}-A_tA_g\Big)\Big(\frac{9\tau^4}{2}\Big)+\mathcal{O}\Big(\tau^4\Big). \qquad (83)$$

Considering $\tau\to 0$, all terms with τ^4 are neglected. The coefficients A_g and A_t are given in (74) and (81). The SAR is finally obtained as

$$\psi(\sigma^*,\sigma^*_\epsilon)\simeq\frac{\tau^2}{2}+\sigma^*(A_t-A_g)\tau^2$$
$$=\tau^2\Big(\frac{1}{2}+\big(d^{(2)}_{1,\lambda}-1\big)\frac{\sigma^{*2}}{(\sigma^*_\epsilon)^2+\sigma^{*2}}-\frac{c_{1,\lambda}\sigma^{*2}}{\sqrt{(\sigma^*_\epsilon)^2+\sigma^{*2}}}\Big). \qquad (84)$$

To derive the first order self-adaptation response function, several simplifications were introduced mainly considering $N\to\infty$ and $\tau\to 0$. The eligibility of this approach depends on the consideration of $\tau\to 0$. This allows the truncation of several Taylor series without introducing too large errors.

C Calculation of the expected values of $\left(\frac{\varsigma^*-\varsigma^{*(g)}}{\varsigma^{*(g)}}\right)^k$

The raw moments of a log-normal distribution with parameter τ are given by $\overline{(\varsigma^*)^k}=(\varsigma^{*(g)})^k\exp\big(k^2\tau^2/2\big)$. We need to derive expression for $\overline{\left(\frac{\varsigma^*-\varsigma^{*(g)}}{\varsigma^{*(g)}}\right)^k}$. Since

$$\left(\frac{\varsigma^*-\varsigma^{*(g)}}{\varsigma^{*(g)}}\right)^k=\sum_{l=0}^{k}\binom{k}{l}(\varsigma^*)^l(-1)^{k-l}(\varsigma^{*(g)})^{-l}, \qquad (85)$$

the expectation is given by

$$\overline{\left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}}\right)^k} = (-1)^k \sum_{l=0}^{k} \binom{k}{l} (-1)^l \mathrm{e}^{\frac{l^2\tau^2}{2}}. \tag{86}$$

Since $\mathrm{e}^{\frac{l^2\tau^2}{2}} = \sum_{n=0}^{\infty} \frac{l^{2n}}{n!2^n}\tau^{2n}$ we have to consider

$$\begin{aligned}\overline{\left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}}\right)^k} &= (-1)^k \sum_{n=0}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{l=0}^{k} \binom{k}{l} (-1)^l l^{2n} \\ &= (-1)^k \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{l=0}^{k} \binom{k}{l} (-1)^l l^{2n}.\end{aligned} \tag{87}$$

As can be easily shown, we have $\sum_{l=0}^{k} \binom{k}{l} (-1)^l l^{2n} = 0$ if $k \geq 2n+1$. The expected values are therefore given by

$$\overline{\left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}}\right)^k} = \begin{cases} (-1)^k \sum_{n=k/2}^{\infty} \frac{\tau^{2n}}{n!2^n} & \\ \times \sum_{l=0}^{k} \binom{k}{l} (-1)^l l^{2n} & \text{if } k = 2j \\ (-1)^k \sum_{n=(k+1)/2}^{\infty} \frac{\tau^{2n}}{n!2^n} & \\ \times \sum_{l=0}^{k} \binom{k}{l} (-1)^l l^{2n} & \text{if } k = 2j+1 \end{cases}. \tag{88}$$

As a result, if we consider the limit case of $\tau \to 0$ and do therefore break off the Taylor series after $n = n_0$, the expected values of $\left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}}\right)^k$ for $k \geq 2n_0 + 1$ do not have to be taken into account.

$$\begin{aligned}
k = 1 \Rightarrow \overline{\left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}}\right)^1} &= -\sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{l=1}^{1} \binom{1}{l} (-1)^l l^{2n} = \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} \\
&= \frac{\tau^2}{2} + \frac{\tau^4}{8} + \dots \\
k = 2 \Rightarrow \overline{\left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}}\right)^2} &= \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{l=1}^{2} \binom{2}{l} (-1)^l l^{2n} \\
&= \sum_{n=1}^{\infty} [2^n - 2^{1-n}] \frac{\tau^{2n}}{n!} = \tau^2 + \frac{7\tau^4}{4} + \dots \\
k = 3 \Rightarrow \overline{\left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}}\right)^3} &= -\sum_{n=2}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{l=1}^{3} \binom{3}{l} (-1)^l l^{2n} \\
&= \sum_{n=2}^{\infty} 3[3^{2n-1} + 1 - 2^{2n}] \frac{\tau^{2n}}{2^n n!} = \frac{9\tau^4}{2} + \dots \\
k = 4 \Rightarrow \overline{\left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}}\right)^4} &= \sum_{n=2}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{l=1}^{4} \binom{4}{l} (-1)^l l^{2n}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=2}^{\infty}[(6)2^{2n} + 2^{4n} - 4(1+3^{2n})]\frac{\tau^{2n}}{2^n n!} \\
&= 3\tau^4 + \ldots \\
k = 5 \Rightarrow \overline{\left(\frac{\varsigma^* - \varsigma^{*(g)}}{\varsigma^{*(g)}}\right)^5} &= -\sum_{n=3}^{\infty}\frac{\tau^{2n}}{n!2^n}\sum_{l=1}^{5}\binom{5}{l}(-1)^l l^{2n} \\
&= \sum_{n=3}^{\infty}[1 + 2(3^{2n}) + 5^{2n-1} - 2^{2n+1} - 2^{4n}] \\
&\quad \times \frac{5\tau^{2n}}{2^n n!} = \frac{15\tau^6}{2} + \ldots. \qquad (89)
\end{aligned}
$$

We now show $\sum_{l=0}^{k}\binom{k}{l}(-1)^l l^{2n} = 0$ if $k \geq 2n+1$. Let $m = 2n$ and start with $m = 0$. Splitting the sum into even and uneven terms and considering Pascal's triangle

$$
\begin{aligned}
\sum_{l=0}^{k}\binom{k}{l}(-1)^l &= \begin{cases} \sum_{l=0}^{(k)/2}\binom{k}{2l} - \sum_{l=0}^{k/2-1}\binom{k}{2l+1} & \text{if } k = 2j \\ \sum_{l=0}^{(k-1)/2}\binom{k}{2l} - \sum_{l=0}^{(k-1)/2-1}\binom{k}{2l+1} & \text{if } k = 2j+1 \end{cases} \\
&= 2^{k-1} - 2^{k-1} = 0. \qquad (90)
\end{aligned}
$$

Let now $m = 1$

$$
\begin{aligned}
\sum_{l=0}^{k}\binom{k}{l}(-1)^l l &= k\sum_{l=1}^{k}\binom{k-1}{l-1}(-1)^l \\
&= -k\sum_{l=0}^{k-1}\binom{k-1}{l}(-1)^l = 0. \qquad (91)
\end{aligned}
$$

Finally for $m \to m+1$, recall that l^m is of the form $l^m = \sum_{j=0}^{m} c_{m,j}\prod_{i=0}^{j-1}(l-i)$ leading to

$$
\begin{aligned}
\sum_{l=0}^{k}\binom{k}{l}(-1)^l l^{m+1} &= \sum_{l=1}^{k}\binom{k}{l}(-1)^l(l-m)l^m + m\sum_{l=1}^{k}\binom{k}{l}(-1)^l l^m \\
&= \sum_{l=1}^{k}\binom{k}{l}(-1)^l(l-m)\sum_{j=0}^{m}c_{m,j}\prod_{i=0}^{j-1}(l-i) \\
&= \sum_{l=1}^{k}\binom{k}{l}(-1)^l\sum_{j=0}^{m}c_{m,j}(l-j)\prod_{i=0}^{j-1}(l-i) \\
&\quad -\sum_{l=1}^{k}\binom{k}{l}(-1)^l\sum_{j=0}^{m}c_{m,j}(m-j)\prod_{i=0}^{j-1}(l-i) \\
&= \sum_{j=0}^{m}c_{m,j}\prod_{i=0}^{j}(k-i)\sum_{l=j+1}^{k}(-1)^l\binom{k-j-1}{l-j-1}
\end{aligned}
$$

$$
\begin{aligned}
&- \sum_{j=0}^{m} c_{m,j}(m-j) \prod_{i=0}^{j-1}(k-i) \\
&\times \sum_{l=j}^{k} (-1)^l \binom{k-j}{l-j} = 0. \qquad (92)
\end{aligned}
$$

D Solutions of $x^3 - ax^2 - bx + c = 0$

The solutions of the polynomial are given by

$$
\begin{aligned}
x_1 &= \frac{a}{3} \\
&+ \frac{\sqrt[3]{2}(a^2+3b)}{3\sqrt[3]{2a^3+9ab-27c+3\sqrt{3}\sqrt{27c^2-18abc-4a^3c-4b^3-a^2b^2}}} \\
&+ \frac{1}{3\sqrt[3]{2}} \\
&\times \sqrt[3]{2a^3+9ab-27c+3\sqrt{3}\sqrt{27c^2-18abc-4a^3c-4b^3-a^2b^2}} \\
x_2 &= \frac{a}{3} + \frac{1+i\sqrt{3}}{3\sqrt[3]{4}} \\
&\times \frac{-a^2-3b}{\sqrt[3]{2a^3+9ab-27c+3\sqrt{3}\sqrt{27c^2-18abc-4a^3c-4b^3-a^2b^2}}} \\
&- \frac{1-i\sqrt{3}}{6\sqrt[3]{2}} \\
&\times \sqrt[3]{2a^3+9ab-27c+3\sqrt{3}\sqrt{27c^2-18abc-4a^3c-4b^3-a^2b^2}} \\
x_3 &= \frac{a}{3} + \frac{1-i\sqrt{3}}{3\sqrt[3]{4}} \\
&\times \frac{-a^2-3b}{\sqrt[3]{2a^3+9ab-27c+3\sqrt{3}\sqrt{27c^2-18abc-4a^3c-4b^3-a^2b^2}}} \\
&- \frac{1+i\sqrt{3}}{6\sqrt[3]{2}} \\
&\times \sqrt[3]{2a^3+9ab-27c+3\sqrt{3}\sqrt{27c^2-18abc-4a^3c-4b^3-a^2b^2}}. \qquad (93)
\end{aligned}
$$

In the cases considered here, the positive real solution of the equation is given by x_1. Let us first give the parameters in the case of the zero of the SAR characterized by

$$
0 = \varsigma^{*6} - \frac{(2d_{1,\lambda}^{(2)}-1)^2 - 4c_{1,\lambda}^2 \sigma_\epsilon^{*2}}{4c_{1,\lambda}^2} \varsigma^{*4}
$$

$$-\sigma_\epsilon^{*2}\frac{2d_{1,\lambda}^{(2)}-1}{2c_{1,\lambda}^2}\varsigma^{*2}-\frac{\sigma_\epsilon^{*4}}{4c_{1,\lambda}^2}. \tag{94}$$

Here, we have $a=[(2d_{1,\lambda}^{(2)}-1)^2-4c_{1,\lambda}^2\sigma_\epsilon^{*2}]/[4c_{1,\lambda}^2]$, $b=\sigma_\epsilon^{*2}(2d_{1,\lambda}^{(2)}-1)/(2c_{1,\lambda}^2)$, and $c=-\sigma_\epsilon^{*4}/(2c_{1,\lambda}^2)$.

The maximum point $\varsigma_{\max}^*\neq 0$ of the progress rate has to fulfill

$$1=\frac{2c_{1,\lambda}}{\sqrt{\varsigma^{*2}+\sigma_\epsilon^{*2}}}-\frac{c_{1,\lambda}\varsigma^{*2}}{\sqrt{\varsigma^{*2}+\sigma_\epsilon^{*2}}^3}$$
$$\Rightarrow \sqrt{\varsigma^{*2}+\sigma_\epsilon^{*2}}^3=c_{1,\lambda}(\varsigma^{*2}+2\sigma_\epsilon^{*2}) \tag{95}$$

leading to $a=c_{1,\lambda}^2-3\sigma_\epsilon^{*2}$, $b=4c_{1,\lambda}^2\sigma_\epsilon^{*2}-3\sigma_\epsilon^{*4}$, and $c=\sigma_\epsilon^{*6}-4c_{1,\lambda}^2\sigma_\epsilon^{*4}$.

References

1. B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *A First Course in Order Statistics*. Wiley, New York, 1992.
2. A. Auger. Convergence Results for $(1,\lambda)$-SA-ES using the Theory of φ-irreducible Markov Chains. In T. Bäck and M. Schoenauer, editors, *Workshop on Evolutionary Algorithms, ICALP-2003*, 2003.
3. H.-G. Beyer. Toward a Theory of Evolution Strategies: Self-Adaptation. *Evolutionary Computation*, 3(3):311–347, 1996.
4. H.-G. Beyer. *The Theory of Evolution Strategies*. Natural Computing Series. Springer, Heidelberg, 2001.
5. H.-G. Beyer and K. Deb. On Self-Adaptive Features in Real-Parameter Evolutionary Algorithms. *IEEE Transactions on Evolutionary Computation*, 5(3):250–270, 2001.
6. H.-G. Beyer and H.-P. Schwefel. Evolution Strategies: A Comprehensive Introduction. *Natural Computing*, 1(1):3–52, 2002.
7. A. Bienvenüe and O. François. Global Convergence for Evolution Strategies in Spherical Problems: Some Simple Proofs and d Difficulties. *Theoretical Computer Science*, 308:269–289, 2003.
8. W.E. Hart, J.M. DeLaurentis, and L.A. Ferguson. On the Convergence of an Implicitly Self-Adaptive Evolutionary Algorithm on One-Dimensional Unimodal Problems. *IEEE Transactions on Evolutionary Computation*, 2003. To appear.
9. Y. Jin and J. Branke. Evolutionary Optimization in Uncertain Environments – A Survey. *IEEE Transactions on Evolutionary Computation*, 9(3):303–317, 2005.
10. D.G. Luenberger. *Introduction to Dynamic Systems*. Wiley, Chichester, 1979.
11. A. Ostermeier, A. Gawelczyk, and N. Hansen. A Derandomized Approach to Self-Adaptation of Evolution Strategies. *Evolutionary Computation*, 2(4):369–380, 1995.
12. I. Rechenberg. *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog Verlag, Stuttgart, 1973.

13. I. Rechenberg. *Evolutionsstrategie '94*. Frommann-Holzboog Verlag, Stuttgart, 1994.
14. H.-P. Schwefel. Adaptive Mechanismen in der biologischen Evolution und ihr Einfluß auf die Evolutionsgeschwindigkeit. Technical report, Technical University of Berlin, 1974. Abschlußbericht zum DFG-Vorhaben Re 215/2.
15. M.A. Semenov. Convergence Velocity of Evolutionary Algorithm with Self-Adaptation. In *GECCO 2002*, pages 210–213, 2002.
16. M.A. Semenov and D.A. Terkel. Analysis of Convergence of an Evolutionary Algorithm with Self-Adaptation Using a Stochastic Lyapunov Function. *Evolutionary Computation*, 11(4):363–379, 2003.