On the Behavior of (mu/mui, lambda)-ES Optimizing Functions Disturbed by Generalized Noise

Hans-Georg Beyer, Markus Olhofer, Bernhard Sendhoff

2002

Preprint:

Foundations of Genetic Algortihms, 2002, in print.

On the Behavior of $(\mu/\mu_I, \lambda)$-ES Optimizing Functions Disturbed by Generalized Noise

Hans-Georg Beyer[*]
Dept. of Computer Science XI
Universität Dortmund
Dortmund, Germany

Markus Olhofer[†] **Bernhard Sendhoff**[‡]
Future Technology Research Division
Honda R&D Europe GmbH
Offenbach, Germany

1 Introduction

The performance analysis of the $(\mu/\mu_I, \lambda)$-ES in real-valued search spaces $\mathbb{R}^N$ has been mainly performed on two simple fitness models: the deterministic sphere model [6] and the ridge functions [14, 5]. Recently, progress has been made in incorporating the effect of noisy fitness evaluations in the analysis of the sphere model [1]. However, these "standard" noise models are additive and the noise term is normally distributed. In particular for real-world optimisation problems this is not sufficient. Indeed the influence of the noise is usually "hidden" in the fitness function itself. A concrete application example is the optimisation of aerodynamic structures [13, 12], where we can distinguish two types of variation. Firstly, variations of the objective parameters, e.g. control points describing a three dimensional spline structure. Secondly, variations of external constraints, e.g. inflow conditions of the air stream. Therefore, analysis based on a more general noise model is needed in order to draw conclusions which are applicable to this type of real-world problems. Recent approaches to "robust optimisation" [8, 16] can be seen as special cases of a general noise model. Special in the sense that a particular fitness measure, usually the average fitness [8], is assumed. As we will point out in Section 3 more generally, two types of measures can be used: threshold measures and statistical momentum measures.

In this paper, we focus on two test functions. In the first one noise is added directly to the objective variables for the simple sphere function. The second test function has been

[*]email: beyer@ls11.cs.uni-dortmund.de
[†]email: markus_olhofer@de.hrdeu.com
[‡]email: bernhard_sendhoff@de.hrdeu.com

constructed in such way that a trade-off between maximising the average and minimising the variance exists, a case which is of particular interest for the application described above.

Deriving progress rate formula for such fitness models from scratch, however, is usually a rather expensive task, especially if one is interested in N-dependent properties of the problems considered. Therefore, it is the intention of this paper to present a method for extending results obtained from the performance analysis of the (μ/μ_I)-ES on a specific noise model analyzed in [1, 2].

The material is organized as follows. First, we will briefly recapitulate the standard noise model usually used in ES theory [1, 2] and discuss its possible shortcomings giving rise to other noise models and two classes of robustness measures. In Section 5 and 6, we apply our approach to the two test functions and compare the theoretical analysis with simulations.

2 ES Analysis for the Standard Noise Model

2.1 The Standard Fitness Noise Model

The standard noise model for the performance analysis of $(\mu/\mu_I, \lambda)$-ES on the sphere model is an additive one assuming normally distributed noise. That is, the fitness model considered in the $\mathbb{R}^N$ search space is given by

$$F_{\text{nsp}}(\mathbf{x}) := Q_{\text{sp}}(\mathbf{x}) + \delta, \tag{1}$$

where

$$Q_{\text{sp}}(\mathbf{x}) := \beta\|\mathbf{x}\|^\alpha, \qquad \alpha > 0, \quad \beta > 0 \tag{2}$$

(minimization assumed) and

$$\delta \sim \mathcal{N}(0, \sigma_\delta^2) \tag{3}$$

has zero mean and variance σ_δ^2. The standard deviation σ_δ, also referred to as noise strength, is assumed to be constant within a *single* generation. That is, σ_δ^2 is allowed to change over the generations g, however, given a parental state $\mathbf{x}_{\text{p}}^{(g)}$, all offspring $\tilde{\mathbf{x}}_l^{(g)}$ generated have fitnesses $F_l^{(g)} = F_{\text{nsp}}(\tilde{\mathbf{x}}_l^{(g)})$ with the same noise strength $\sigma_\delta^{(g)}$.

2.2 A Short Survey on the Progress Rate Theory

Based on this fitness model, a progress rate theory has been developed that correctly predicts the behavior of the $(\mu/\mu_I, \lambda)$-ES on the qualitative level (i.e., asymptotically $N \to \infty$ [1]) as well as on the quantitative level (taking the search space dimensionality N into account [2]). Writing R for the parental distance to the optimum,

$$R^{(g)} := \|\mathbf{x}_{\text{p}}^{(g)}\|, \tag{4}$$

the progress rate is defined by the expected value of the parental distance change from generation time g to $g+1$

$$\varphi := \mathrm{E}[R^{(g)} - R^{(g+1)}]. \tag{5}$$

It is intuitively clear that φ at time g depends on the current parental state $\mathbf{x}_{\text{p}}^{(g)}$, the exogenous strategy parameters μ (number of parents) and λ (number of offspring), the strength σ of the isotropic mutations (standard deviation of a single object parameter component),

and the noise strength σ_δ. Since the mutations as well as the fitness model are isotropic, the parental state can be lumped together by considering the parental distance $R = R^{(g)}$ (dropping the generation counter for brevity). Introducing the normalized quantities (again dropping the generation counter)

$$\varphi^* = \varphi\frac{N}{R}, \qquad \sigma^* = \sigma\frac{N}{R}, \qquad \sigma_\delta^* = \sigma_\delta\frac{N}{\alpha|\beta|R^\alpha}, \tag{6}$$

one can derive the asymptotically correct progress rate expression [1]

$$\varphi^* \simeq \sigma^{*2}\left[\frac{c_{\mu/\mu,\lambda}}{\sqrt{\sigma^{*2}+\sigma_\delta^{*2}}} - \frac{1}{2\mu}\right], \tag{7}$$

where $c_{\mu/\mu,\lambda}$ is the so-called progress coefficient (for its definition, see e.g. [6, p. 247]).

Due to the definition (5), convergence of the ES is given for positive φ. The actual evolution of the real ES algorithm is, however, determined by the σ-dynamics. Since there is up to now no theory predicting the σ-evolution quantitatively, one has to resort to stability criteria in order to characterize the general ES behavior. Such a N-dependent criterion, also called *evolution criterion*, has been derived [2], it reads

$$\frac{\sigma_\delta^{*2} + \sigma^{*2}(1+\sigma^{*2}/2N)}{(1+\sigma^{*2}/2\mu N)^2} < (2\mu c_{\mu/\mu,\lambda})^2. \tag{8}$$

For small σ^* or $N \to \infty$ this criterion reduces to the asymptotically correct

$$\sigma_\delta^{*2} + \sigma^{*2} < (2\mu c_{\mu/\mu,\lambda})^2 \tag{9}$$

that can also be derived from (7) considering the inequality $\varphi > 0$. As one can easily show, criterion (8) necessarily implies

$$\sigma_\delta^* < 2\mu c_{\mu/\mu,\lambda} \tag{10}$$

in order to ensure local convergence in mean. If the "$<$" is replaced by "$=$", we have the steady-state case: The expected value R remains constant, i.e., it cannot further reduce and one observes a *residual localization error* R_∞ (expected value) of the optimum state. The violation of condition (10) determines implicitly a lower bound on R_∞. Its predictive quality will be high if σ^* is sufficiently small. This is usually the case in standard ES implementations using σ-self adaptation (σSA) or the cumulative step-size adaptation (CSA) technique [10].

The necessary evolution condition (10) can also be used to discuss ES efficiency aspects concerning the residual localization error R_∞ given a constant noise strength σ_δ. Here it is assumed that it is the aim to obtain an R_∞ as small as possible. We will summarize the most important results (see [4, 1, 2] for details):

(A) Given a fixed offspring number λ, the μ should be chosen such that $\mu = \lambda/2$ in order to get minimal R_∞.

(B) Given a fixed noise strength σ_δ, it is more efficient (with respect to the number of function evaluations needed in order to ensure the validity of the evolution criterion (10)) to increase the offspring number by a factor k instead of resampling the F-fitness k times.

3 Shortcomings of the Standard Fitness Noise Model in the Context of Robustness

As we already pointed out in the introduction, the additive noise model of the last section is not sufficient for a large class of realistic optimization problems. Leaving aside those cases where the F-fluctuations cannot be well described by Gaussian noise, there are cases where the underlying driving random forces are indeed (approximately) normally distributed:

a) Relative F-measurement errors, the noise strength σ_δ scales with the (unperturbed) F-values, i.e.

$$\sigma_\delta = \text{const.}|Q(\mathbf{x})|. \tag{11}$$

As a result, each offspring individual l has its own individual noise strength $\sigma_{\delta l}$.

b) Systematic object parameter noise, the externally given object parameters are randomly perturbed in $Q(\mathbf{x})$, i.e., one has

$$F(\mathbf{x}) = Q(\mathbf{x}+\mathbf{z}) \quad \text{with} \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \epsilon_z^2 \mathbf{1}) \tag{12}$$

and F becomes a random variate for constant $\mathbf{x}$. This kind of noise could also be called *actuator noise*.

c) General noise case (including all other cases), here the fitness function F depends on the object parameters $\mathbf{x}$ and a secondary set of randomly fluctuating parameters $\mathbf{z}$ yielding the noisy fitness model ($\mathbf{C}$ – covariance matrix)

$$F(\mathbf{x}) = Q(\mathbf{x}, \mathbf{z}) \quad \text{with} \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}). \tag{13}$$

In the following we will develop a simple approach for handling these kinds of noise models in an approximative manner by transforming them to the standard noise model.

In order to analyze more general noise models in the context of robustness, it is necessary to define the fitness criterion which is to be used for selection. Coarsely speaking, robustness refers to solutions which are of high quality not just at one particular point in the design space but over a larger domain often called the working range of the solution. Thus, instead of using one realisation of the random variables in the general noise model, we define probabilistic evaluation criteria. Two classes can be distinguished:

Threshold Measures: The idea is to define "minimality" (maximality can be defined analogously) in the following sense. We demand that the probability of F-realizations smaller than a certain threshold q should be maximal

$$\Pr[F \le q \,|\, \mathbf{x}] \to \max. \tag{14}$$

Statistical Momentum Measures: The idea is to define "minimality" with respect to the kth moments of F, $\overline{F^k} = \mathrm{E}[F^k \,|\, \mathbf{x}]$, and demanding

$$\mathrm{E}[F^k | \, \mathbf{x}] \to \min. \tag{15}$$

For maximization, the first moment should be maximized. When we take the first and second moments into account one might also consider a definition that demands maximization of the average and minimization of the variance.

4 Basic Idea of the Novel Approach

In this section, we consider the general case c) where the fitness F depends on a set of external object parameters $\mathbf{x}$ and a set of internal random parameters $\mathbf{z}$ independent of $\mathbf{x}$.

Consider now the $(\mu/\mu, \lambda)$-ES. Selection of individuals is solely based on the *observed*, i.e. noisy, fitness values. The F-function appears as a black box; the ES algorithm does not "know" how the F values are internally generated. This is an important property because it opens up the way for a new approach to the general noise case: *The original $F(\mathbf{x}) = Q(\mathbf{x}, \mathbf{z})$ is additively decomposed into a deterministic part and a stochastic part.* To this end, we consider $F(\mathbf{x})$ as a conditional random function having the conditional probability density function $p(F(\mathbf{z})|\mathbf{x})$. Taking the conditional expectation

$$\mathrm{E}[F(\mathbf{x}, \mathbf{z})|\mathbf{x}] =: \overline{F}(\mathbf{x}) \tag{16}$$

we have got a function $\overline{F}$ that depends only on the object parameters. Therefore, it is a deterministic function and the original noisy function can be expressed as

$$F(\mathbf{x}, \mathbf{z}) = \overline{F}(\mathbf{x}) + \underbrace{(F(\mathbf{x}, \mathbf{z}) - \overline{F}(\mathbf{x}))}_{\Delta_z(\mathbf{x})}. \tag{17}$$

By this decomposition, Δ_z carries all the stochastics and

$$\overline{\Delta_z} = \mathrm{E}[\Delta_z|\mathbf{x}] = 0. \tag{18}$$

Since the ES is a black box optimization algorithm, it does no matter how the F-values are generated. Thus, one can use the noise model (17) instead of the original one for theoretical investigations.

While the decomposition (17) holds for all kinds of noise, the practical application of this technique is necessarily restricted to cases where the fitness model $\overline{F}(\mathbf{x})$ with the noise density $p(\Delta_z(\mathbf{x}))$ has already been analyzed. With respect to our sphere model analysis, recapitulated in Section 2.1, the method can be applied if one is able to express or approximate:

1. $\overline{F}(\mathbf{x})$ by a (deterministic) sphere model

$$\overline{F}(\mathbf{x}) \simeq Q_{\mathrm{sp}}(\mathbf{x}) \tag{19}$$

   and

2. $p(\Delta_z(\mathbf{x}))$ by a normal distribution

$$\Delta_z(\mathbf{x}) \simeq \mathcal{N}(0, \sigma_\delta^2(\|\mathbf{x}\|)), \tag{20}$$

where the variance σ_δ^2 is obtained by

$$\sigma_\delta^2 = \mathrm{Var}[F \mid \|\mathbf{x}\|] = \mathrm{E}[\Delta_z^2 \mid \|\mathbf{x}\|]. \tag{21}$$

If such a model transformation or approximation is possible, all results obtained for the noisy sphere model can be transferred to the new situations. This also includes the ES efficiency considerations (A) and (B) from the end of Section 2.2.

We will demonstrate the technique on two example functions below.

5 Example Function f_1

Function f_1 is of type b) from the list in Section 3. It is defined as

$$F(\mathbf{x}) = f_1(\mathbf{x}) := (\mathbf{x} + \mathbf{z})^2, \qquad \mathbf{x}, \mathbf{z} \in \mathbb{R}^N, \tag{22}$$

where $\mathbf{z}$ is a normally distributed random vector each component of which having variance ε^2

$$\mathbf{z} \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \varepsilon^2 \mathbf{1}). \tag{23}$$

5.1 Threshold Measure for Example Function f_1

We start with the discussion of the threshold criterion, Eq. (14). Since $\mathbf{x} = (x_1, \ldots, x_N)$, Eq. (22) can be written as

$$F = \sum_{i=1}^{N} (x_i + \varepsilon \mathcal{N}_i(0,1))^2 = \varepsilon^2 \sum_{i=1}^{N} \left(\frac{x_i}{\varepsilon} + \mathcal{N}_i(0,1) \right)^2. \tag{24}$$

According to [11, p. 130ff] the sum in the right-hand side of (24) obeys a noncentral χ^2 distribution with $\nu = N$ degrees of freedom and noncentrality parameter $\zeta = \sum_{i=1}^{N} (x_i/\varepsilon)^2$

$$\zeta = \mathbf{x}^2/\varepsilon^2. \tag{25}$$

The cumulative distribution function (cdf) of this sum, denoted by $\chi'^2_N(\zeta)$, can be expressed by a series of cdfs of central χ^2_{N+2j} distributions. Writing formally $\Pr[\chi^2_{N+2j} \le t]$ for the cdf of the χ^2_{N+2j} variate, the cdf of $\chi'^2_N(\zeta)$ reads [11, p. 132]

$$\Pr[\chi'^2_N(\zeta) \le t] = \mathrm{e}^{-\frac{1}{2}\zeta} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{1}{2}\zeta \right)^j \Pr[\chi^2_{N+2j} \le t]. \tag{26}$$

With $t := q/\varepsilon^2$, we have

$$\Pr[F \le q \,|\, \mathbf{x}] = \Pr\left[\frac{F}{\varepsilon^2} \le \frac{q}{\varepsilon^2} \right] = \Pr\left[\frac{F}{\varepsilon^2} \le t \right] = \Pr\left[\chi'^2_N(\zeta) \le t \right]$$

and the criterion (14) can be expressed by

$$\Pr[F \le q \,|\, \mathbf{x}] = \mathrm{e}^{-\frac{1}{2}\frac{\mathbf{x}^2}{\varepsilon^2}} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{1}{2}\frac{\mathbf{x}^2}{\varepsilon^2} \right)^j \Pr\left[\chi^2_{N+2j} \le \frac{q}{\varepsilon^2} \right] \to \max. \tag{27}$$

Given an arbitrary $q > 0$, this probability is maximized for $\mathbf{x}^2 \to 0$, i.e.

$$\Pr[F \le q \,|\, \mathbf{x}] \to \max \quad \Leftrightarrow \quad \mathbf{x} = \mathbf{0}. \tag{28}$$

This can be easily proven by showing that

$$\forall \mathbf{x}, \|\mathbf{x}\| > 0 : \quad \Pr[F < q|\mathbf{0}] > \Pr[F < q|\mathbf{x}].$$

Due to space limitation the proof must be omitted here.

Let us summarize the findings on example function f_1: Given the criterion (14), the performance of an optimization algorithm on the model function (22) can be evaluated by considering the dynamics of $R := \sqrt{\mathbf{x}^2}$ which is also the residual distance to the optimum for vanishing noise ($\varepsilon^2 = 0$).

5.2 Calculating the First Moments of f_1

Since we need $\overline{F}$ and $\mathrm{Var}[F|\mathbf{x}]$ for further calculations, we will derive them here.

We will calculate the first moment (i.e. the expected value) of F. Using (22) one obtains

$$\begin{aligned}\overline{F} &= \mathrm{E}[(\mathbf{x}+\mathbf{z})^2|\mathbf{x}] = \mathrm{E}\left[\sum_{i=1}^{N}(x_i^2+2x_iz_i+z_i^2)\,\middle|\,\mathbf{x}\right] \\ &= \sum_{i=1}^{N}(x_i^2+2x_i\mathrm{E}[z_i]+\mathrm{E}[z_i^2]) = \sum_{i=1}^{N}x_i^2+2\sum_{i=1}^{N}x_i\overline{z_i}+\sum_{i=1}^{N}\overline{z_i^2}. \end{aligned} \tag{29}$$

Here, we have written $\overline{z_i^k}=\mathrm{E}[z_i^k]$ for brevity. Recalling the moment expressions of Gaussian noise (23) with zero mean and variance ε^2

$$\overline{z_i}=0, \qquad \overline{z_i^2}=\varepsilon^2, \qquad \overline{z_i^3}=0, \qquad \overline{z_i^4}=3\varepsilon^4 \tag{30}$$

one obtains from (29)

$$\overline{F} = \sum_{i=1}^{N}x_i^2+\sum_{i=1}^{N}\varepsilon^2 = R^2+N\varepsilon^2 \tag{31}$$

(recall that $R^2=\sum_{i=1}^{N}x_i^2$). Due to the statistical independence of the z_i-components, the conditional variance of F is obtained as

$$\mathrm{Var}[F|\mathbf{x}]=\mathrm{Var}\left[\sum_{i=1}^{N}(x_i+z_i)^2\right]=\sum_{i=1}^{N}\mathrm{Var}[(x_i+z_i)^2|x_i]. \tag{32}$$

Using $\mathrm{Var}[t]=\overline{t^2}-\overline{t}^2$ and (30) one gets

$$\begin{aligned}\mathrm{Var}[(x_i+z_i)^2|x_i] &= \mathrm{Var}[2x_iz_i+z_i^2\,|\,x_i] \\ &= \mathrm{E}[(2x_iz_i+z_i^2)^2\,|\,x_i]-\mathrm{E}[2x_iz_i+z_i^2\,|\,x_i]^2 \\ &= 4x_i^2\varepsilon^2+2\varepsilon^4 \end{aligned} \tag{33}$$

and for (32)

$$\mathrm{Var}[F|R]=4R^2\varepsilon^2+2N\varepsilon^4. \tag{34}$$

Considering (31), one sees that the optimizer state of the expected value of F corresponds to the minimum of the case without noise. That is, the minimum is obtained for $R=0$ or equivalently $\mathbf{x}=\mathbf{0}$. As one can see, both the threshold measure from Section 5.1 (cf. Eq. (28)) and the statistical momentum measure considered here have led to the same conclusion.

Interestingly, the state $\mathbf{x}=\mathbf{0}$ is also that state for which the fitness variance of F gets minimal. Thus, decreasing R reduces $\overline{F}$ and $\mathrm{Var}[F]$. Therefore, the state $\mathbf{x}=\mathbf{0}$ may be regarded as a *robust* optimizer state.

5.3 Estimating the Residual Localization Error Bound R_∞

When considering the variance expression (34) one notices that for $R \to 0$ there remains a variance $2N\varepsilon^4 > 0$. According to Section 2.2 this implies a residual localization error for the optimizer. We will derive a lower bound R_∞ for this residual error. To this end, we make contact to the sphere model approximation idea developed in Section 4. Using the results (31) and (34), the sphere model (1) – (3) becomes

$$Q_{\mathrm{sp}}(\mathbf{x}) = \|\mathbf{x}\|^2 + N\varepsilon^2, \tag{35}$$

where the additive term $N\varepsilon^2$ is of no relevance[1] and the noise strength σ_δ becomes

$$\sigma_\delta = 2\varepsilon\sqrt{\|\mathbf{x}\|^2 + N\varepsilon^2/2}. \tag{36}$$

Now, the necessary evolution criterion (10) can be applied. Using the normalization (6), one obtains (writing $R = \|\mathbf{x}\|$)

$$\frac{N\varepsilon\sqrt{R^2 + N\varepsilon^2/2}}{2R^2} < \mu c_{\mu/\mu,\lambda}. \tag{37}$$

This inequality can be resolved for R yielding a condition on the (parental) $R = \|\mathbf{x}\|$-values for which the ES exhibits local convergence. A simple calculation yields the biquadratic inequality

$$0 \; < \; R^4 - \frac{N^2\varepsilon^2}{4\mu^2 c^2_{\mu/\mu,\lambda}}\, R^2 - \frac{N^3\varepsilon^4}{8\mu^2 c^2_{\mu/\mu,\lambda}}. \tag{38}$$

Resolving for R^2 yields

$$R^2 > \frac{N^2\varepsilon^2}{8\mu^2 c^2_{\mu/\mu,\lambda}} + \sqrt{\frac{N^4\varepsilon^4}{64\mu^4 c^4_{\mu/\mu,\lambda}} + \frac{N^3\varepsilon^4}{8\mu^2 c^2_{\mu/\mu,\lambda}}} \tag{39}$$

and finally the lower bound R_∞ (taking the square root in (39))

$$R > R_\infty = \frac{\varepsilon N}{\sqrt{8}\mu c_{\mu/\mu,\lambda}} \sqrt{1 + \sqrt{1 + \frac{8\mu^2 c^2_{\mu/\mu,\lambda}}{N}}}. \tag{40}$$

As we will see in the simulations section, the predictive power of R_∞ as an estimate for the observed average steady-state R is astonishingly good. Therefore, Eq. (40) can be used to discuss the influence of the strategy parameters on the ES performance with respect to the average steady-state R.

5.4 Simulations for Function f_1

Extensive simulations have been performed in order to assess the behavior of the ES on the objective function (22), (23) and to test the validity of (40). The $(\mu/\mu_I, \lambda)$-ES as well as the mutation strength control techniques, the σ-selfadaptation (σSA) and the cumulative step-size adaptation (CSA), are described and explained in Appendix A.

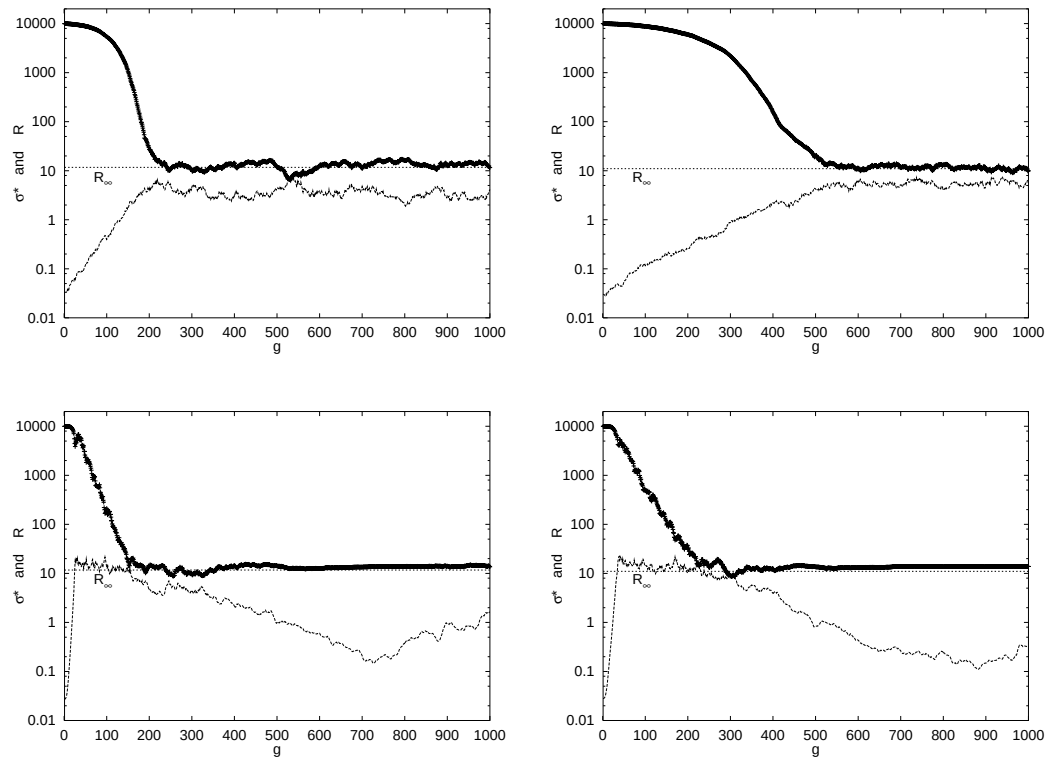

Figure 1: Dynamics of $(\mu/\mu_I, \lambda)$-ES on the quadratic $N = 30$ dimensional sphere model with systematic noise (cf. (22), (23)) $\varepsilon = 6$. The R-dynamics of the parental centroid are displayed by "+" symbols (upper curves in the graphs, appearing as bold lines) and the σ^*-dynamics appear as zigzagging thin lines (lower curves). The figures on top are from runs using standard σSA technique, the bottom figures are the results of the CSA technique. The figures on the left-hand side display the $(15/15_I, 50)$-ES, those on the right-hand side the $(30/30_I, 50)$-ES.

First, the dynamical behavior is considered. Figure 1 shows the residual distance dynamics $R^{(g)}$ and the evolution of the normalized mutation strength $\sigma^{*(g)}$. The initial conditions have been chosen in an ill-adapted manner in order to test the adaptive behavior of the σSA and CSA. Both techniques are able to adapt the mutation strength σ. However, the CSA does this with a much higher rate as one can infer from the steep ascent of the σ^* values in the two bottom figures. Furthermore, the σSA is not able to reach the high σ^* values obtained by the CSA during the adaptation phase of the run. This observation is in accordance with empirical findings in [9] showing that the σSA is not able to adapt the large mutations strength necessary in $(\mu/\mu_I, \lambda)$-ES in order to reach optimal performance.

After the initial phase one observes qualitatively the same behavior for both the σSA and the CSA: A steady-state regime is reached exhibiting a residual localization error for the

[1]This is so, because the ES uses (μ, λ)-selection. This selection type is invariant with respect to constant fitness transformations.

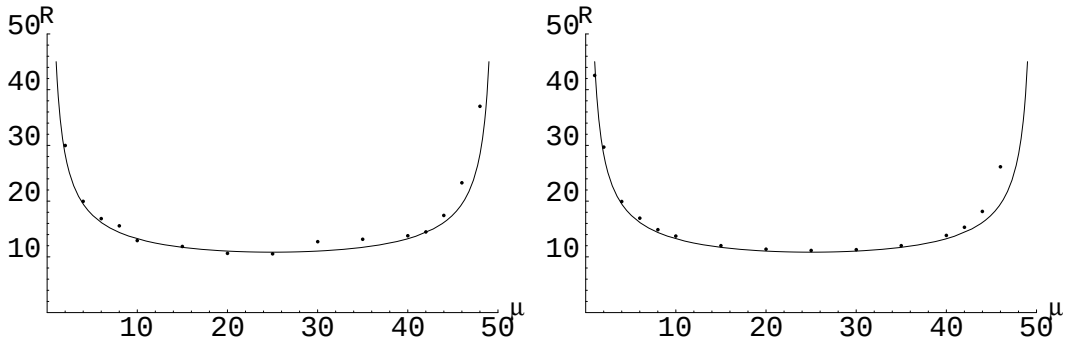

Figure 2: On the predictive quality of Eq. (40) serving as an estimate for the mean value of the residual localization error, left-hand side $(\mu/\mu_I, 50)$-CSA-ES, right-hand side $(\mu/\mu_I, 50)$-σSA-ES. The parameters of the fitness function (22), (23) are $\varepsilon = 6$, $N = 30$. The dots represent the simulation results (for $\mu = 1$, 2, 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 42, 44, 46, 48, 49) using the average over generations $g = 2001$ to 20000. The initial values are $R^{(0)} = 10000$ and $\sigma^{(0)} = 10$.

optimizer. The actual $R^{(g)}$ values fluctuate around an average value greater than zero.

The predictive power of the R_∞-formula (40) is displayed in Fig 2. We can see how well R_∞ predicts the average value of the steady-state R. It appears that in general the average steady-state R is better predicted for the ES using σSA. This can be explained by looking at the σ^*-evolution in Fig. 1. After reaching the steady-state regime, σ^* exhibits in the CSA version a random walk like behavior having very small σ^* values over a long time period. That is, having reached a certain R value, it will only change sporadicly. This is different in σSA where σ^* stays at a comparatively high level. (However, this σ^* is small enough to ensure the approximately equal condition in the evolution criterion (10).)

As we already mentioned at the end of Section 2.2, the optimal strategy parameter setting for the number of parents is $\mu = \lambda/2$. One can show that his holds for the R_∞ formula (40), too, and this is confirmed by the simulation results in Fig. 2. However, one also can see that the performance is only slightly degraded when chosing a higher selection pressure ensuring a faster convergence speed (cf. the left pictures in Fig. 1). Therefore, the previous $N \to \infty$ result of choosing $\mu \approx 0.27\lambda$ can be recommended.

5.5 Estimating the Steady-State On-Line Fitness Behavior of the Parental Centroid

In real ES runs it is customary to monitor the fitness values along the generations g. We will refer to this analysis as *on-line* behavior as opposite to the off-line behavior of the fitness values given a fixed object parameter vector. One possibility of evaluating the on-line behavior is to collect the data from the fitness values of the (parental) center of mass individuals $\mathbf{x}_{\mathrm{p}}^{(g)} = \langle \mathbf{x} \rangle^{(g)}$, i.e., those search space states which are obtained from the intermediate (μ/μ) recombination of the μ best offspring individuals (cf. Appendix A, Eq. (63))

$$F_{\mathrm{p}}^{(g)} := F(\mathbf{x}_{\mathrm{p}}). \tag{41}$$

After reaching the steady-state, $F_{\mathrm{p}}^{(g)}$ fluctuates around its steady-state mean value $\overline{F}_{\mathrm{p}}$. We will present here an idea how to determine $\overline{F}_{\mathrm{p}}$ as well as the standard deviation σ_F

$$\sigma_F := \sqrt{\mathrm{Var}[F_{\mathrm{p}}]} \tag{42}$$

of the $F_{\mathrm{p}}^{(g)}$ fluctuations.

The most striking effect of noise in the steady-state regime is the appearance of a residual localization error. That is, on average the $\mathbf{x}_{\mathrm{p}}$ states are located in a distance $R \gtrapprox R_\infty$ to the optimum. It is important to understand that this distance is *not* realized by a (nearly) fixed centroid state, but it is the result of the fluctuations of the coordinates of the centroid state. That is, the residual localization error is a statistical phenomenon where each of the coordinates contributes to. The centroid states are centered around the optimum state. Due to the spherical symmetry of the fitness function, the centroid states are uniformly distributed on hypersphere shells the expected radius R of those is approximately R_∞. This is very similar to the observation of an expected nonzero length of a vector $\mathbf{x}$ consisting of N x-components each of which fluctuates independently with a Gaussian distribution $\mathcal{N}(0, \sigma_x^2)$ with zero mean. According to (30) the expected length square of such a random vector is $\overline{\mathbf{x}^2} = N\sigma_x^2$ and $\overline{\|\mathbf{x}\|} \simeq \sigma_x\sqrt{N}$. Making the *Ansatz* $R^2 = N\sigma_x^2$ provides therefore an estimate for the variance of a single centroid component. While this approach is not rigorous in a strict mathematical sense, it is correct from the viewpoint of approximating the unknown distribution of the centroid components by a normal distribution.[2] Accepting the validity of this approximation and taking into account that $R^2 \geq R_\infty^2$, one immediately obtains

$$\sigma_x^2 \geq R_\infty^2/N \tag{43}$$

As we have seen, assuming $R \approx R_\infty$ gives a good estimate for the actually observed mean value of the steady-state R, we will use the equal sign in (43) as an estimate for the steady-state σ_x^2. Using Eq. (40), one obtains

$$\sigma_x^2 = \frac{\varepsilon^2 N}{8\mu^2 c_{\mu/\mu,\lambda}^2}\left(1 + \sqrt{1 + \frac{8\mu^2 c_{\mu/\mu,\lambda}^2}{N}}\right). \tag{44}$$

In order to calculate the expected value of $F_{\mathrm{p}}^{(g)}$, we use the definition of f_1, Eq. (22). Each component of $\mathbf{z}$ in (22) fluctuates with variance ε^2 and the x-components fluctuate with variance (44). Due to the statistical independence assumption of both processes, the variances are additive. Thus, one gets

$$F_{\mathrm{p}} = \sum_{i=1}^{N}(x_i + z_i)^2 = \sum_{i=1}^{N}\mathcal{N}(0, \underbrace{\sigma_x^2 + \varepsilon^2}_{:=\tilde{\sigma}^2})^2, \tag{45}$$

where

$$\tilde{\sigma}^2 = \varepsilon^2\left[1 + \frac{N}{8\mu^2 c_{\mu/\mu,\lambda}^2}\left(1 + \sqrt{1 + \frac{8\mu^2 c_{\mu/\mu,\lambda}^2}{N}}\right)\right]. \tag{46}$$

[2] One might also use maximum entropy technique to draw the same conclusion.

Now, one can calculate the expected value of the centroid fitness. Taking the expected value of (45) yields immediately $\overline{F}_{\mathrm{p}} = N\tilde{\sigma}^2$ and with (46)

$$\overline{F}_{\mathrm{p}} = N\varepsilon^2 \left[1 + \frac{N}{8\mu^2 c^2_{\mu/\mu,\lambda}} \left(1 + \sqrt{1 + \frac{8\mu^2 c^2_{\mu/\mu,\lambda}}{N}} \right) \right]. \tag{47}$$

This equation predicts the actually measured mean value in ES runs well. Since there is no qualitative difference to the R curves in Fig. 2, graphics are not displayed here.

For the standard deviation of F_{p} one has to consider the variance of (45). Its calculation is similar to that of (34). One obtains *mutatis mutandis* ($R = 0$, $\varepsilon = \tilde{\sigma}$) $\mathrm{Var}[F_{\mathrm{p}}] = 2N\tilde{\sigma}^4$ and finally with (46) and (42)

$$\sigma_F = \sqrt{2N}\varepsilon^2 \left[1 + \frac{N}{8\mu^2 c^2_{\mu/\mu,\lambda}} \left(1 + \sqrt{1 + \frac{8\mu^2 c^2_{\mu/\mu,\lambda}}{N}} \right) \right]. \tag{48}$$

Figure 3 shows the predictive quality of this formula. It is in good agreement with the ES

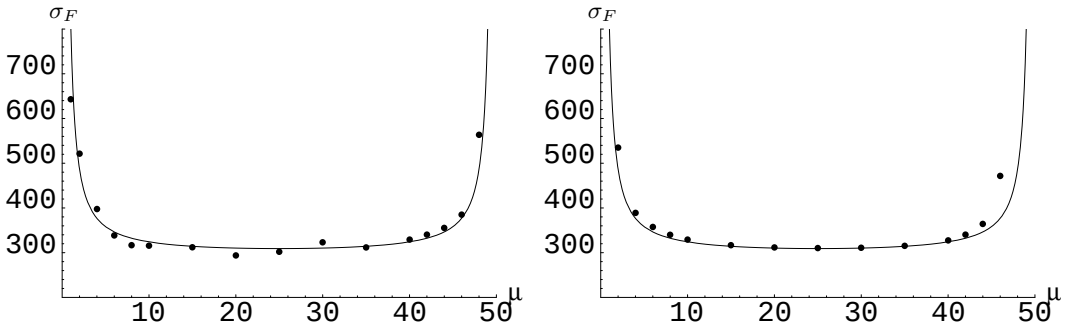

Figure 3: On the predictive quality of Eq. (48) serving as an estimate for the on-line steady-state standard deviation of the parental centroid fitness, left-hand side $(\mu/\mu_I, 50)$-CSA-ES, right-hand side $(\mu/\mu_I, 50)$-σSA-ES. The parameters of the fitness function (22), (23) are $\varepsilon = 6$, $N = 30$. The dots represent results obtained from the same ES runs as in Fig. 2 using generations $g = 2001$ to 20000.

runs. The minimum of the fluctuations is observed for $\mu = \lambda/2$. Since this population sizing also yields the smallest R_∞ and $\overline{F}_{\mathrm{p}}$, it may be regarded as the most robust strategy setting.

6 Example Function f_2

Function f_2 is of type c) from the list in Section 3. It is defined as

$$F(\mathbf{x}) = f_2(\mathbf{x}) := a - (z+1)\|\mathbf{x}\|^\alpha + bz, \qquad \mathbf{x} \in \mathbb{R}^N, \quad z \in \mathbb{R}, \tag{49}$$

where z is a normally distributed random variate with variance ε^2

$$z \sim \mathcal{N}(0, \varepsilon^2). \tag{50}$$

Clearly, the task at hand is maximization of f_2. Having a closer look at (49), one sees that F can be rearranged such that

$$F(\mathbf{x}) = a - \|\mathbf{x}\|^\alpha + (b - \|\mathbf{x}\|^\alpha)z. \tag{51}$$

Therefore, one immediately obtains without any further approximations

$$Q_{\mathrm{sp}}(\mathbf{x}) = a - \|\mathbf{x}\|^\alpha \tag{52}$$

and

$$\mathrm{Var}[F|\mathbf{x}] = \varepsilon^2(b - \|\mathbf{x}\|^\alpha)^2. \tag{53}$$

6.1 Threshold Measure for Example Function f_2

We can investigate the probabilistic maximality condition similar to (14). It reads

$$\Pr[F > f|\mathbf{x}] \to \max. \tag{54}$$

Since

$$\begin{aligned} \Pr[F > f|\mathbf{x}] &= 1 - \Pr[F \leq f|\mathbf{x}] \to \max \\ \iff & \Pr[F \leq f|\mathbf{x}] \to \min. \end{aligned} \tag{55}$$

$\Pr[F \leq f|\mathbf{x}]$ defines the cdf of F which is easily obtained since the random process is Gaussian. Using the cdf $\Phi(y)$ of the standard Gaussian variate which is defined as

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{t=-\infty}^{t=y} \mathrm{e}^{-\frac{1}{2}t^2}\,\mathrm{d}t = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{y}{\sqrt{2}}\right)\right), \tag{56}$$

one obtains writing $R = \|\mathbf{x}\|$

$$\Pr[F \leq f|\mathbf{x}] = \Phi\left(\frac{f - a + R^\alpha}{\varepsilon|b - R^\alpha|}\right) \to \min. \tag{57}$$

For small ε this function shows the expected behavior, i.e. in order to get a small probability in Eq. (57), given an arbitrary f-threshold, R must be chosen as small as possible. Thus, the optimization quality can again be measured by considering the residual R-distance: R should be as small as possible. However, when we consider larger values for ε things get more complicated, see Fig. 4. The plot on the right-hand side is especially interesting because it reveals a plateau being indifferent as to the R-values for larger f-values. Even worse, in that region there is not a monotonous dependency of Pr on R. That is, Pr can be reduced by either decreasing R or increasing R. This can be a source of instability. Furthermore, for smaller f-values one observes a minimum of Pr for $R \neq 0$. Both observations will have consequences for the amelioration behavior of the ES. However, the actual ES behavior cannot be further inferred from these plots.

6.2 Momentum Measure for Example Function f_2

The ambiguities observed with the probabilistic maximality can be observed *mutatis mutandis* using the momentum based approach similar to (15), i.e. for $\mathrm{E}[F^k|\mathbf{x}] \to \max$. Considering only the first moment ($k = 1$), one finds $R = 0$ using Eq. (52). The expected value

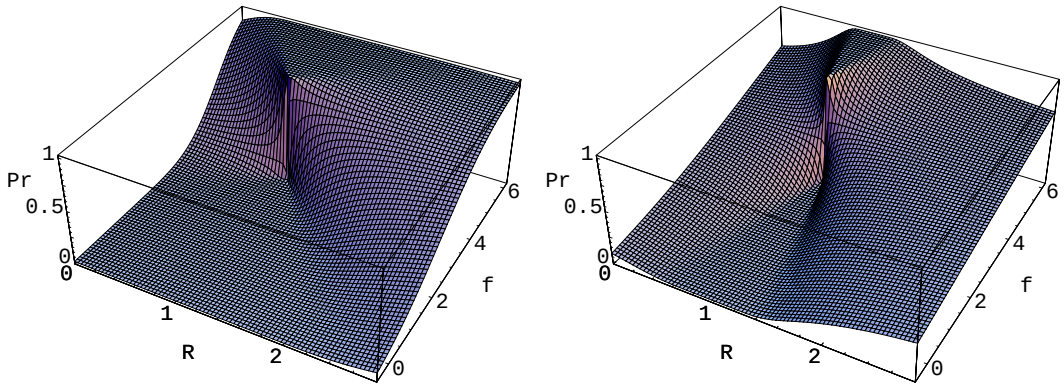

Figure 4: The probability Pr that F, Eq. (49), is less than or equal to a bound f given a residual R according to Eq. (57) with $a = 5$, $b = 1$, $\alpha = 1$. The plot on the left-hand side is for $\varepsilon = 1$ and that on right-hand side holds for $\varepsilon = 4$.

of F can be maximized by decreasing R to zero. However, from the viewpoint of robustness this is bought at the expense of a higher F-variance. From Eq. (53) we immediately see that the variance takes its zero minimum at $R^\alpha = \|\mathbf{x}\|^\alpha = b$. That is, for R-values less or greater than $b^{1/\alpha}$ the Var$[F]$ increases monotonously. This again is a hint for possible instabilities in the ES dynamics. Figure 5 shows such instabilities in an actual ES run using parameters chosen at the edge of stability (for the underlying theory, see Section 6.3, below). The

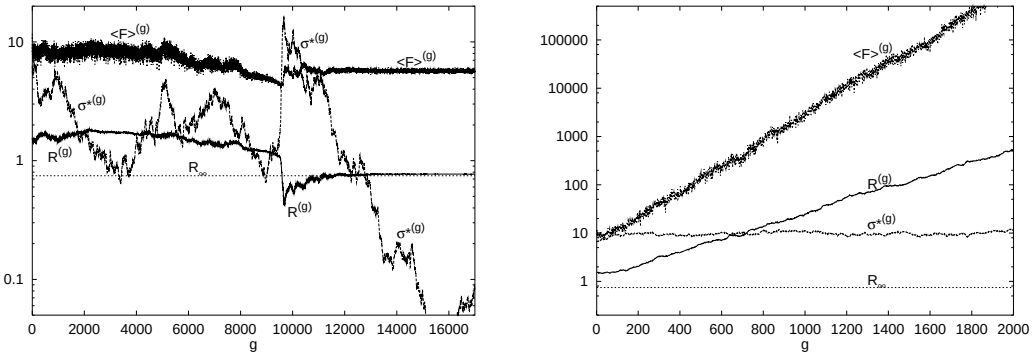

Figure 5: Dynamics of the $(100/100_I, 200)$-CSA-ES (left-hand side) and the $(100/100_I, 200)$-σSA-ES (right-hand side) operating at the edge of instability. The parameters of the fitness function f_2, defined by Eq. (49), are $N = 100$, $\varepsilon = 4$, $a = 5$, $b = 1$, $\alpha = 2$.

graphs on the right-hand side show the typical divergence behavior of a σSA-ES: The σSA realizes a steady-state σ^* significantly greater than zero. The reason for this behavior can be traced back to the bias of the mutation operator: In the case of selective neutrality (no relevant fitness information, e.g. strong fitness noise or flat fitness landscape) the operators increase on average the mutation strength σ by a constant factor (for a discussion of the

underlying SA-principle, see [7]).[3] The plots on the left-hand side of Fig. 5 show convergence behavior to a residual localization error. The CSA-ES starts with a $R^{(0)} > b = 1$. It reduces R gradually, thus it decreases the variance (53) up to that point where $R = 1$. This can be well seen in the fluctuation behavior of the $\langle F \rangle^{(g)}$ values ($\langle F \rangle^{(g)}$ is the average of the measured μ parental fitness values). At $R \approx 1$ we have approximately the noise-free case, resulting in a large progress rate φ. Therefore, one observes the abrupt decrease of R around the generation number $g = 9500$. However, this small R results in a large F-variance and R increases again up to its steady-state value near to R_∞.

6.3 ES Stability and Residual Localization Error

The qualitative behavior of the ES-dynamics on f_2 as displayed in Fig. 5 can be well explained by the stability analysis using the necessary evolution criterion (10). Using (53) for σ_δ and the normalization (6), one obtains (writing $R = \|\mathbf{x}\|$)

$$h(R^\alpha) := \frac{|b - R^\alpha|}{R^\alpha} < \frac{2\alpha\mu c_{\mu/\mu,\lambda}}{\varepsilon N}. \tag{58}$$

The meaning of this inequality is visualized in Fig. 6. As one can easily show, the auxiliary

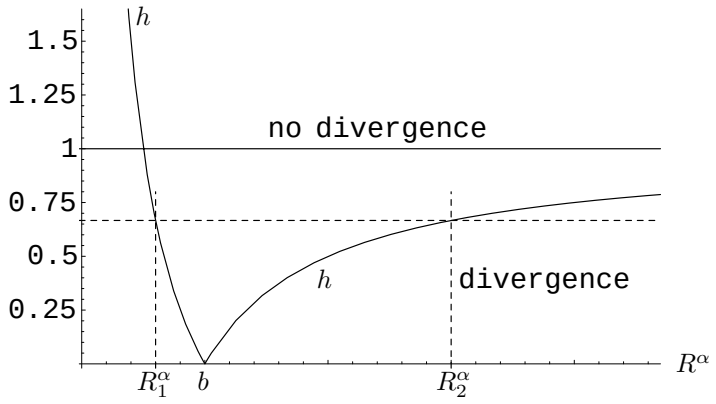

Figure 6: Visualization of the necessary evolution criterion (58) valid for function f_2 (49). For fixed parameters α, μ, λ, N, and ε the intersection of the function $h = h(R^\alpha, b)$ and the horizontal dashed line determines the domain of local convergence. The equilibrium points are labeled by R_1^α and R_2^α.

function $h(R^\alpha)$ has the asymptotic properties (note that $\alpha > 0$)

$$h(R^\alpha) \xrightarrow{R\to 0} \infty \qquad \text{and} \qquad h(R^\alpha) \xrightarrow{R\to\infty} 1. \tag{59}$$

That is, depending on the actual value of $2\alpha\mu c_{\mu/\mu,\lambda}/(\varepsilon N)$ in (58) one observes one or two equilibrium points in Fig. 6. We will discuss the different cases in detail:

[3] The behavior of the CSA-ES is different in the situation of selection neutrality: σ exhibits a random walk like behavior.

1) First case: $\frac{2\alpha\mu c_{\mu/\mu,\lambda}}{\varepsilon N} \geq 1$.

Since $h(R^\alpha) \xrightarrow{R\to\infty} 1$, there is only the equilibrium point R_1^α. That is, independent of the initial value $R^{(0)}$, the R_1^α serves as lower bound on the mean value of R at the steady-state. Since

$$R_1^\alpha \leq b, \tag{60}$$

inequality (58) can be resolved for R_1^α using the equal sign instead of the "<." One immediately obtains

$$R_1^\alpha = \frac{b}{1 + \frac{2\alpha\mu c_{\mu/\mu,\lambda}}{\varepsilon N}}.$$

Neglecting fluctuations, one therefore finds as the lower bound R_∞ of the residual localization error of the optimizer

$$R > R_\infty = \sqrt[\alpha]{\frac{b}{1 + \frac{2\alpha\mu c_{\mu/\mu,\lambda}}{\varepsilon N}}} \tag{61}$$

independent of the initial $R^{(0)} < \infty$ chosen.

2) Second case: $\frac{2\alpha\mu c_{\mu/\mu,\lambda}}{\varepsilon N} < 1$.

There are two equilibrium points R_1^α and R_2^α. The qualitative dynamical behavior depends on the initial value $R^{(0)}$:

a) $R^{(0)} > R_2$.
From Fig. 6 we infer that $h((R^{(0)})^\alpha) > \frac{2\alpha\mu c_{\mu/\mu,\lambda}}{\varepsilon N}$. The necessary evolution criterion is violated ($\varphi < 0$) with the result that the initial $R^{(0)}$ is increased in expectation. Provided that $\sigma > 0$, the ES diverges. The point of instability R_2^α can be calculated. Using in (58) the equal sign instead of the "<" and taking $R_2 > b$ into account, one gets

$$R_2^\alpha - b = R_2^\alpha \frac{2\alpha\mu c_{\mu/\mu,\lambda}}{\varepsilon N} \quad \Rightarrow \quad R_2^\alpha = \frac{b}{1 - \frac{2\alpha\mu c_{\mu/\mu,\lambda}}{\varepsilon N}}.$$

Therefore, ES divergence is observed for (again neglecting fluctuations)

$$R^{(0)} > \sqrt[\alpha]{\frac{b}{1 - \frac{2\alpha\mu c_{\mu/\mu,\lambda}}{\varepsilon N}}}. \tag{62}$$

b) $R_1 < R^{(0)} < R_2$.
Since (58) is fulfilled, the ES reduces the parental R and converges to the steady-state the mean value of which is bounded from below by R_∞ (61).

c) $R^{(0)} < R_1$.
Condition (58) is violated. The ES successively increases R up to that point where $R^{(g)} \geq R_1$. Again, the expected value is bounded from below by R_∞ (61).

From the discussion presented in the previous section one can infer that stable working of the ES on f_2 always implies steady-state residual localization errors $R_{\mathrm{ss}} < b^{1/\alpha}$. Noting that states $R^\alpha = b$ result in vanishing fitness noise (53), it becomes clear that the ES on f_2 does not prefer fitness variance minimal states. Instead, the steady-state is in between the fitness maximal $R = 0$ and the fitness variance minimal $R = b^{1/\alpha}$. The actual steady-state can be controlled within these bounds by choosing μ and λ appropriately. Maximizing fitness and minimizing variance in f_2 are two conflicting goals. Whether one might call the ES's behavior *robust optimization* is therefore a question of definition.

6.4 Simulations for Function f_2

6.4.1 Dynamical Behavior

Figures 7 – 9 display actual ES runs on f_2. These simulations support the theoretical findings presented in the pervious section. Figure 7 shows typical convergence behavior. The CSA-

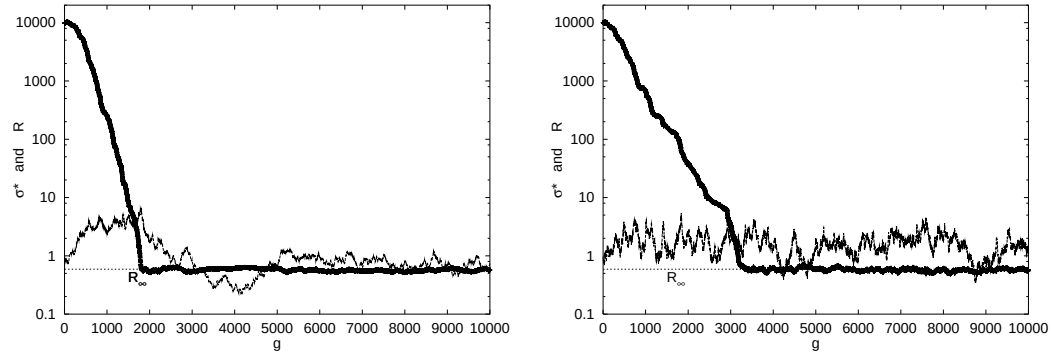

Figure 7: Dynamics of the $(4/4_I, 15)$-CSA-ES (left-hand side) and the $(4/4_I, 15)$-σSA-ES (right-hand side). The R-dynamics and the σ^*-dynamics are shown as in Fig. 1. The parameters of the fitness function are $N = 100$, $\varepsilon = 0.1$, $a = 5$, $b = 1$, $\alpha = 2$ and the initial condition $R^{(0)} = 10000$, $\sigma^{(0)} = 100$. The R_∞ predicts well the steady-state average R.

ES provides the faster convergence compared to the σSA-ES. Both strategies approach the steady-state regime. The mean value of the steady-state R is well predicted by R_∞ (for a more detailed investigation, see below).

Figure 8 shows the behavior of the ES at the edge of instability. The variance ε has been chosen in such a way that the theory still predicts convergence. In this special situation, using additionally an initial σ much too small, the σSA-ES performs better than the CSA-ES: The σSA-ES approaches the steady-state $R \to R_\infty$. This is so, because the σSA-ES increases the mutation strength in expectation (thus, generating offspring with R and F values of selective relevance), whereas the CSA-ES exhibits a random walk like behavior for σ (nearly selective neutrality of the F-values). Things are changing when considering variance values where the theory predicts divergence. As one can see in Fig. 9, the σSA-ES diverges exponentially fast, while the CSA-ES exhibits only a moderate increase of the initial R value. Whether this behavior is a desired one, depends on the application the user has in mind. Both behaviors may be detrimental: Facing an unknown objective function, the CSA-ES behavior might signal convergence to an optimum state even though it is merly

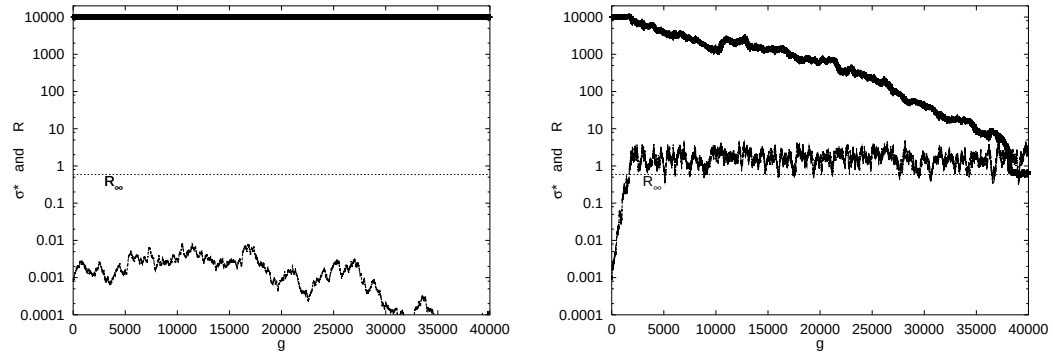

Figure 8: Dynamics of the $(4/4_I, 15)$-CSA-ES (left-hand side) and the $(4/4_I, 15)$-σSA-ES (right-hand side) at the edge of instability. The R-dynamics and σ^*-dynamics are shown as in Fig. 1. The parameters of the fitness function are $N = 100$, $a = 5$, $b = 1$, $\alpha = 2$ and the initial condition $R^{(0)} = 10000$, $\sigma^{(0)} = 0.1$, however, $\varepsilon = 0.15$ has been chosen (cf. Fig. 7).

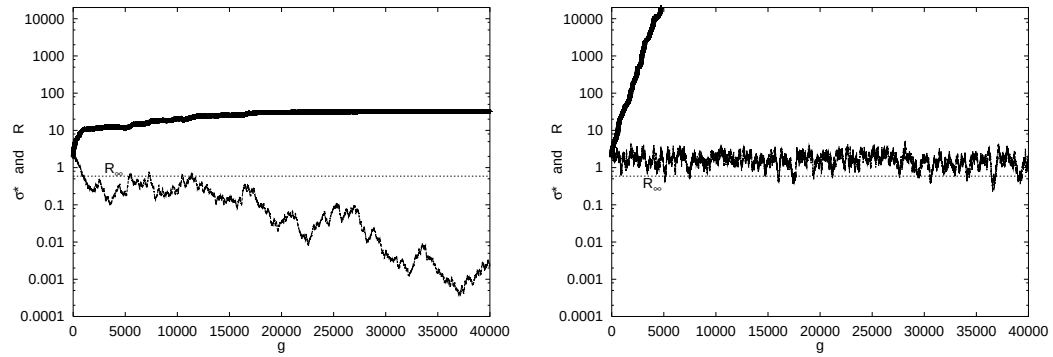

Figure 9: Dynamics of the $(4/4_I, 15)$-CSA-ES (left-hand side) and the $(4/4_I, 15)$-σSA-ES (right-hand side) showing divergence behavior for $\varepsilon = 0.28$. The R-dynamics and σ^*-dynamics are shown as in Fig. 1. The parameters of the fitness function are $N = 100$, $a = 5$, $b = 1$, $\alpha = 2$ and the initial condition $R^{(0)} = 2$, $\sigma^{(0)} = 0.1$.

a state with a selectively neutral neighborhood. Conversely, the σSA-ES leaves this state exponentially fast even though there could be an optimum state hidden in its neighborhood. Therefore, in uncertain cases it is recommended to use both adaptation techniques.

6.4.2 Steady-State Behavior

Provided that the ES does not diverge for f_2, it exhibits a steady-state behavior with a residual distance R of the parents to the optimum state. A lower bound R_∞ for the mean value of the fluctuating R values is given by Eq. (61). Its predictive quality has been tested by extensive simulations. The basic results and observations are summarized in Fig. 10 for parameter space dimensionality $N = 100$. One observes a relatively good predictive quality which becomes slightly worse for lower dimensionalities, plots are not shown. The test cases for $\alpha = 3$ exhibit on average the largest deviations between theory and simulation. Recalling

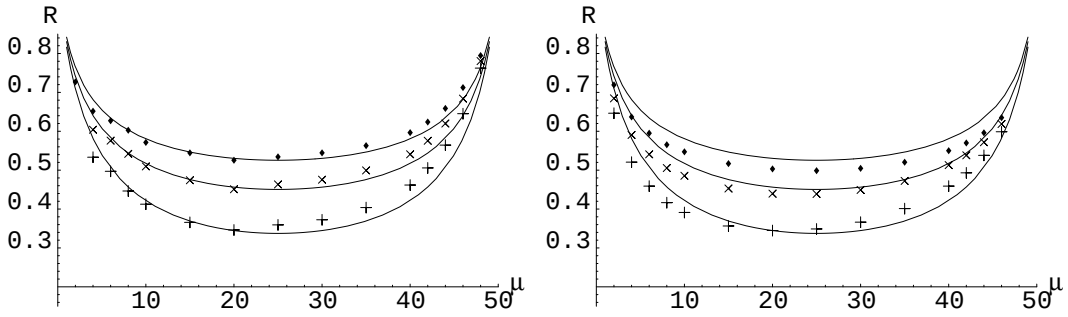

Figure 10: On the predictive quality of Formula (61) serving as an estimate for the mean value of the residual localization error, left-hand side $(\mu/\mu_I, 50)$-CSA-ES, right-hand side $(\mu/\mu_I, 50)$-σSA-ES (curves from bottom to top: $\alpha = 1$, 2, 3). The data points represent the simulation results (for $\mu = 1$, 2, 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 42, 44, 46, 48, 49) using the average over generations $g = 2001$ to 40000. The parameters of the fitness function (49) are $N = 100$, $\varepsilon = 0.2$, $a = 5$, $b = 1$. The data points for $\alpha = 1$ are displayed by "+", for $\alpha = 2$ by "×", and for $\alpha = 3$ by filled diamonds. The initial values are $R^{(0)} = 1$ and $\sigma^{(0)} = 0.1$.

that the progress rate theory was originally derived using approximations for $\alpha = 2$ and $\alpha = 1$, respectively, (see [1, 6]) and the other cases were treated by Taylor expansions [3], these deviations do not appear as a surprise.

While Eq. (61) predicts a symmetric $R_\infty(\mu)$ behavior, as can be seen in the figures, the simulations suggest a slight dissymmetry which is not yet fully understood. Also observed divergences in simulations using $\mu = 1$ (in the case of the CSA-ES) and for μ values near λ (here, for $\mu = 48$ and $\mu = 49$) should be subject of further investigations. However, a deeper theoretical understanding does necessarily imply the formulation of a dynamic theory of adaptation for both the σSA-ES and the CSA-ES.

7 Summary

In this paper, we highlighted the shortcomings of the standard noise model in particular in the framework of robustness as an evaluation criterion. Based on the analysis of typical cases from real-world optimization problems, e.g. aerodynamic design optimisation, we proposed a novel approach to get some insight into the behavior of evolution strategies for more general noise models and two classes of robustness measures. The two example functions which we analyzed are chosen in such a way that results from the standard noise model can still be applied, that relevant application cases are covered and that interesting results are obtained. On the one hand, these results are constructive, e.g. on appropriate or sub-optimal choices of parent to offspring relations. On the other hand, qualitative results have been reached for example on the explanation of possible divergent behavior of evolution strategies.

References

[1] D. V. Arnold and H.-G. Beyer. Local Performace of the $(\mu/\mu_I, \lambda)$-ES in a Noisy Environment. In W. Martin and W. Spears, editors, *Foundations of Genetic Algorithms, 6*, pages 127–141, San Francisco, CA, 2001. Morgan Kaufmann.

[2] D. V. Arnold and H.-G. Beyer. Performance Analysis of Evolution Strategies with Multi-Recombination in High-Dimensional $\mathbb{R}^N$-Search Spaces Disturbed by Noise. *Theoretical Computer Science*, 2002. in print.

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

[4] H.-G. Beyer. Evolutionary algorithms in noisy environments: theoretical issues and guidelines for practice. *Computer Methods in Applied Mechanics and Engineering*, 186(2–4):239–267, 2000.

[5] H.-G. Beyer. On the Performance of $(1, \lambda)$-Evolution Strategies for the Ridge Function Class. *IEEE Transactions on Evolutionary Computation*, 5(3):218–235, 2001.

[6] H.-G. Beyer. *The Theory of Evolution Strategies*. Natural Computing Series. Springer, Heidelberg, 2001. ISBN 3-540-67297-4.

[7] 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.

[8] J. Branke. Efficient evolutionary algorithms for searching robust solutions. In I.C. Parmee, editor, *Adaptive Computing in Design and Manufacture (ACDM)*, pages 275–285. Springer Verlag, 2000.

[9] L. Grünz and H.-G. Beyer. Some Observations on the Interaction of Recombination and Self-Adaptation in Evolution Strategies. In P.J. Angeline, editor, *Proceedings of the CEC'99 Conference*, pages 639–645, Piscataway, NJ, 1999. IEEE.

[10] N. Hansen and A. Ostermeier. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[11] N. L. Johnson and S. Kotz. *Continuous Univariate Distributions – 2*. Houghton Mifflin Company, Boston, 1970.

[12] M. Olhofer, T. Arima, T. Sonoda, M. Fischer, and B. Sendhoff. Aerodynamic shape optimisation using evolution strategies. In *Optimisation in Industry III*, 2001.

[13] M. Olhofer, T. Arima, T. Sonoda, and B. Sendhoff. Optimisation of a stator blade used in a transonic compressor cascade with evolution strategies. In I.C. Parmee, editor, *Adaptive Computing in Design and Manufacture (ACDM)*, pages 45–54. Springer Verlag, 2000.

[14] A. I. Oyman and H.-G. Beyer. Analysis of the $(\mu/\mu, \lambda)$-ES on the Parabolic Ridge. *Evolutionary Computation*, 8(3):267–289, 2000.

[15] H.-P. Schwefel. *Numerical Optimization of Computer Models*. Wiley, Chichester, 1981.

[16] S. Tsutsui and A. Gosh. Genetic algorithms with a robust solution searching scheme. *IEEE Transactions on Evolutionary Computation*, 1(3):201–208, 1997.

A Description of the ESs Used

For the simulation of the dynamic behavior of $(\mu/\mu_I, \lambda)$-ES under systematic noise the ES must control the endogenous strategy parameter σ. We used the two standard approaches to this control problem: the σ self-adaptation [15] and alternatively the cumulative step-size adaptation [10].

Using the notation

$$\langle \mathbf{a} \rangle^{(g)} := \frac{1}{\mu} \sum_{m=1}^{\mu} \mathbf{a}_{m;\lambda}^{(g)} \tag{63}$$

for intermediate recombination (centroid calculation, i.e., averaging over the $\mathbf{a}$ parameters of the μ best offspring individuals), the $(\mu/\mu_I, \lambda)$-σSA-ES can be expressed in "offspring notation"

$$\forall l = 1, \ldots, \lambda : \left\{ \begin{array}{l} \sigma_l^{(g+1)} := \langle \sigma \rangle^{(g)} \mathrm{e}^{\tau \mathcal{N}_l(0,1)} \\ \mathbf{y}_l^{(g+1)} := \langle \mathbf{y} \rangle^{(g)} + \sigma_l^{(g+1)} \boldsymbol{\mathcal{N}}_l(\mathbf{0}, \mathbf{1}). \end{array} \right. \tag{64}$$

That is, each offspring individual (indexed by l) gets its own mutation strength σ. And this mutation strength is used as mutation parameter for producing the offspring's object parameter. We used the log-normal update rule for mutating the mutation strength with $\tau = 1/\sqrt{N}$.

The cumulative step-size adaptation uses a single mutation strength parameter σ per generation to produce all the offspring. This σ is updated by a deterministic rule which is controlled by certain statistics gathered over the course of generations. The statistics used is the so-called (normalized) cumulative path-length $\mathbf{s}$. If $\|\mathbf{s}\|$ is greater than the expected length of a random path, σ is increased. In the opposite situation, σ is decreased. The update rule reads

$$\left. \begin{array}{l} \forall l = 1, \ldots, \lambda : \mathbf{y}_l^{(g+1)} := \langle \mathbf{y} \rangle^{(g)} + \sigma^{(g)} \boldsymbol{\mathcal{N}}_l(\mathbf{0}, \mathbf{1}) \\ \mathbf{s}^{(g+1)} := (1-c)\mathbf{s}^{(g)} + \sqrt{(2-c)c} \frac{\sqrt{\mu}}{\sigma^{(g)}} \left(\langle \mathbf{y} \rangle^{(g+1)} - \langle \mathbf{y} \rangle^{(g)} \right) \\ \sigma^{(g+1)} := \sigma^{(g)} \exp\left(\frac{\|\mathbf{s}^{(g+1)}\| - \overline{\chi}_N}{D \overline{\chi}_N} \right) \end{array} \right\}, \tag{65}$$

where $\mathbf{s}^{(0)} = \mathbf{0}$ is chosen initially. The recommended standard settings for the cumulation parameter c and the damping constant D are used, i.e., $c = 1/\sqrt{N}$ and $D = \sqrt{N}$. For the expected length of a random vector comprising N standard normal components, the approximation $\overline{\chi}_N = \sqrt{N}(1 - 1/4N + 1/21N^2)$ was used.