Bias in Standard Self-Adaptive Evolution Strategies

Amir Omeradzic and Hans-Georg Beyer
Research Center Business Informatics
Vorarlberg University of Applied Sciences
Dornbirn, Austria
{amir.omeradzic, hans-georg.beyer}@fhv.at

***Abstract*—The mutation strength (σ) adaptation of a multi-recombinative self-adapting Evolution Strategy is investigated on the Rastrigin test function by theoretical and experimental means. Sampling σ from a log-normal distribution reveals the occurrence of an undesired steady-state under high multimodality, which halts the σ-adaptation and prevents convergence. It is shown that an inherent bias is the reason for this steady-state when sampling log-normal mutations. Therefore, sampling from a normal distribution is introduced as an alternative. Normal sampling does not exhibit the steady-state behavior and it is more stable optimizing functions under high multimodality.**

***Index Terms*—Evolution Strategy, Rastrigin Function, Self-Adaptation, Multimodality**

I. Introduction

Evolution Strategies (ES) have shown to be successful in optimizing highly multimodal test functions with adequate global structure, if suitable strategy parameters are chosen [1]. For the optimization under high multimodality, e.g., on the Rastrigin function, well-adjusted adaptation of the mutation strength σ is crucial to find the global optimum with high probability without getting trapped in a local minimum [2], [3]. State-of-the-art mutation adaptation schemes are self-adaptation (SA) [4]–[6], and cumulative step-size adaptation (CSA) [7]–[9]. CSA controls σ by evaluating the path-length of selected mutation vectors over many generations and comparing it to expected results under random selection. Self-adaptation uses individual sampling of mutation strengths and implicit selection of suitable σ-values based on information of the current generation.

Under self-adaptation each offspring is sampling (denoted by $\widetilde{\ }$) its mutation strength as $\widetilde{\sigma} = \sigma\xi$, where σ is the parental mutation strength and ξ a random variable. ξ can be chosen to attain values from a continuous distribution (two cases considered in this paper) or to establish a discrete operator (see [4], [6]). Sampling from a log-normal distribution was introduced in [5] and can be regarded as the default choice for self-adaptation. The alternative of sampling from a normal distribution was discussed in [6], but not further analyzed.

Two basic properties can be identified by evaluating $\mathrm{median}[\xi]$ and $\mathrm{E}[\xi]$. The property $\mathrm{median}[\xi] = 1$ ensures sampling larger and smaller values with equal probabilities $\Pr\{\xi \leq 1\} = \Pr\{\xi \geq 1\}$. This ensures a symmetric sampling process, which is desired. Additionally, in order to have *unbiased* sampling, the second property $\mathrm{E}[\xi] = 1$ is required. As explained in Sec. II, the log-normal operator introduces a bias due to $\mathrm{E}[\xi] > 1$. This bias is usually regarded as beneficial since it increases the search steps in flat or noisy fitness landscapes [10], thus allowing to leave non-promising regions quickly in contrast to the CSA-ES that may exhibit problems [11]. However, this bias has also a downside, which is being reported here for the first time. It prevents the ES from being convergent in the limit of high multimodality on the Rastrigin function, resulting in a steady-state behavior. This observation motivates the investigation of the unbiased normal operator. The performance of both operators is studied in more detail throughout the paper. To this end, the Rastrigin function and noisy sphere will be considered. The reason for choosing these two functions for the subsequent analysis is that theoretical results can be used to explain the behavior observed in experiments.

For an N-dimensional search vector $\mathbf{y} = [y_1, ..., y_N]$ the Rastrigin test function (ras) is defined in terms of oscillation strength A and frequency α as

$$f_{\text{ras}}(\mathbf{y}) \coloneqq \sum_{i=1}^{N} \left[y_i^2 + A(1 - \cos(\alpha y_i))\right]. \tag{1}$$

Self-adaptation using log-normal mutations on f_{ras} was investigated in [2], [3], [12]. However, these studies have not detected the steady-state since comparably small A- and α-values were chosen. The noisy sphere (ns) subjected to constant normally distributed noise with variance σ_ϵ^2 is defined as

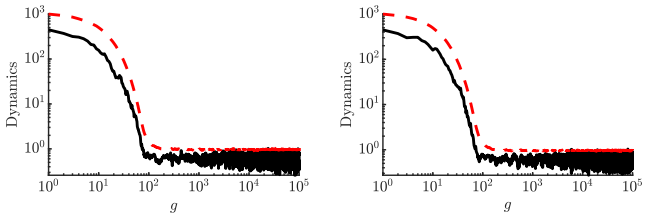
$$f_{\text{ns}}(\mathbf{y}) \coloneqq \sum_{i=1}^{N} y_i^2 + \mathcal{N}(0, \sigma_\epsilon^2). \tag{2}$$

As will be shown later, for sufficiently large α, the ES will share similar properties regarding the self-adaptation on f_{ras} and f_{ns}.

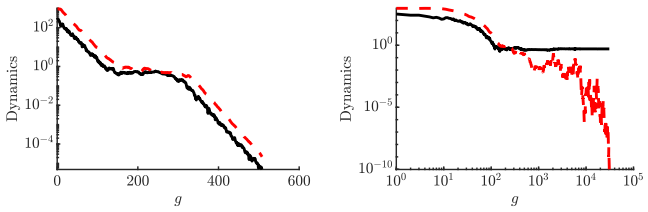
As the performance measure of global progress we use the residual distance $R \coloneqq \sqrt{\sum_i y_i^2}$. Furthermore, the sphere normalization $\sigma^* = \sigma N/R$ (the asterisk * denoting normalized quantities) of the mutation strength σ will be needed later. The rescaling of σ by its residual distance R ensures scale-invariance on the sphere and will be useful later.

Figure 1 shows an introductory example illustrating the problem of undesired steady-state behavior. Figure 1a displays the σ-self-adaptive (σSA-)ES using (default) log-normal mutations, while Fig. 1b shows the CSA-ES (see e.g. [9,

This work was supported by the Austrian Science Fund (FWF) under grant P33702-N. The authors gratefully thank Lisa Schönenberger for providing valuable feedback.



(a) σSA-ES with log-normal mutations using $\tau = 1/\sqrt{2N}$.



(b) CSA-ES with cumulation constant $c = 1/\sqrt{N}$ and damping parameter $D = \sqrt{N}$.

Fig. 1: Dynamics over generation g of $(1000/1000_I, 2000)$-ES using two σ-adaptation methods (a) and (b). The left column shows Rastrigin (1) with $N = 20$, $A = 20$, $\alpha = 2\pi$. The right column shows the noisy sphere (2) with $N = 20$ and $\sigma_\epsilon^2 = 4000$. The black line shows the R-dynamics, the red dashed line the σ-dynamics. The runs are initialized at $\mathbf{y}^{(0)} = 100 \cdot \mathbf{1}$ with $\sigma^{(0)} = 1000$.

p. 12]) as a first comparison. After the initialization phase has passed, the σSA-ES on the Rastrigin function (left) shows a constant σ-level, which leads to a constant residual distance in expectation (fluctuations can be observed). This occurs over a very large number of generations (see generation counter) and the ES does not converge locally. The CSA-ES on the other hand (Fig. 1b left) solves the Rastrigin problem easily, which indicates an issue of the standard σSA-ES.

The observed steady-state very much resembles the steady-state on the noisy sphere (2) (see Fig. 1a right). Choosing a large noise variance σ_ϵ^2 (details discussed below (32)), the dynamics of Rastrigin and the noisy sphere are practically the same showing a steady residual distance $R \approx 0.5$. The CSA-ES (Fig. 1b right) also fluctuates at $R \approx 0.5$, but one observes a continuous σ-decrease over many generations. The similar behavior of the ES on Rastrigin and the noisy sphere will be discussed later again.

The remainder of the paper is structured as follows. In Sec. II the log-normal and normal operators are introduced and their differences are discussed. The undesired steady-state is investigated for both operators by theoretical and experimental means in Secs. III and IV, respectively. Additional studies regarding the stability of the adaptation are done in Sec. IV. Finally, conclusions are drawn in Sec. V.

II. Normal Operator for Self-Adaptation

The multi-recombinative ES, see Algorithm 1, consists of μ parents and λ offspring (truncation ratio $\vartheta := \mu/\lambda$). The σSA-ES samples mutation strengths independently for each

Algorithm 1 $(\mu/\mu_I, \lambda)$-σSA-ES with (log-)normal operator

1: $g \leftarrow 0$
2: initialize $\left(\mathbf{y}^{(0)}, \sigma^{(0)}, \tau\right)$
3: **repeat**
4: **for** $l = 1, ..., \lambda$ **do**
5: $\tilde{\sigma}_l \leftarrow \begin{cases} \sigma\text{SA}_L: & \sigma^{(g)} \mathrm{e}^{\tau \mathcal{N}_l(0,1)} \\ \sigma\text{SA}_N: & \sigma^{(g)}(1 + \tau \mathcal{N}_l(0,1)) \end{cases}$
6: $\tilde{\mathbf{y}}_l \leftarrow \mathbf{y}^{(g)} + \tilde{\sigma}_l \mathcal{N}_l(\mathbf{0}, \mathbf{1})$
7: $\tilde{f}_l \leftarrow f(\tilde{\mathbf{y}}_l)$
8: **end for**
9: $(\tilde{f}_{1;\lambda}, ..., \tilde{f}_{m;\lambda}, ..., \tilde{f}_{\mu;\lambda}) \leftarrow \mathrm{sort}(\tilde{f}_1, ..., \tilde{f}_\lambda)$
10: $\mathbf{y}^{(g+1)} \leftarrow \frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{\mathbf{y}}_{m;\lambda}$
11: $\sigma^{(g+1)} \leftarrow \frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{\sigma}_{m;\lambda}$
12: $g \leftarrow g + 1$
13: **until** termination criterion

offspring, see Alg. 1 (Line 5). The sampling is parameterized using learning parameter τ (default value on the sphere was derived in [13] as $\tau = 1/\sqrt{2N}$). The sampled σ-values are then used to generate isotropic mutations. Selection of the $m = 1, ..., \mu$ best individuals (denoted by subscript "$m;\lambda$") is based on their respective fitness values, which were attained using the sampled mutation strength. The selection of individuals therefore implicitly selects suitable mutation strengths.

A standard self-adaptive ES samples log-normally distributed (subscript L) σ-values [5], [6] according to

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

Note that $\widetilde{\sigma}_L > 0$ is satisfied by definition. The variate $\xi = \mathrm{e}^{\tau \mathcal{N}(0,1)}$ is log-normally distributed with $\mathrm{median}[\xi] = 1$ and $\mathrm{E}[\xi] = \mathrm{e}^{\frac{\tau^2}{2}}$. The expected value of (3) yields simply

$$\mathrm{E}[\widetilde{\sigma}_L] = \sigma \mathrm{e}^{\tau^2/2}. \tag{4}$$

As will be shown later, the characteristic bias of the log-normal distribution, i.e., $\mathrm{E}[\widetilde{\sigma}_L] > \sigma$, will enable a steady-state behavior.

Instead of sampling log-normally distributed values, one can introduce a normal (subscript N) sampling scheme as

$$\widetilde{\sigma}_N = \sigma(1 + \tau \mathcal{N}(0,1)). \tag{5}$$

It was originally introduced in the context of evolutionary programming in [14] and referred to as the Meta-EP-operator. In [6] it is discussed in the context of self-adaptation for ES as an alternative to log-normal mutations. The operator $\xi = 1 + \tau \mathcal{N}(0,1)$ yields $\mathrm{median}[\xi] = 1$ and $\mathrm{E}[\xi] = 1$, such that one has

$$\mathrm{E}[\widetilde{\sigma}_N] = \sigma, \tag{6}$$

which yields no bias (in contrast to (4)). Equation (5) can be obtained from (3) using Taylor-expansion and neglecting $O(\tau^2)$, which suggests similar adaptation characteristics for small τ. However, sampling (3) will show a different steady-state behavior compared to (5). Furthermore, it should be noted that (5) may yield negative samples for the mutation strength.

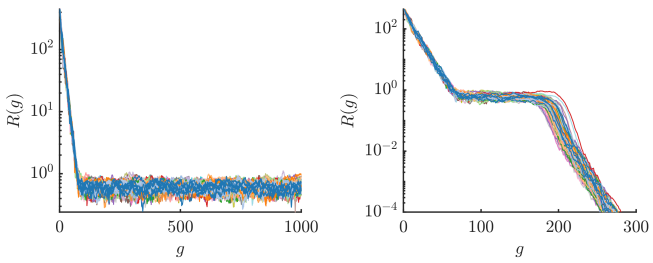


Fig. 2: The $R(g)$-dynamics is shown for 100 individual trials (left: σSA_L, right: σSA_N) on Rastrigin $A = 20$, $\alpha = 2\pi$, $N = 20$, $(1000/1000_I, 2000)$-ES with $\tau = 1/\sqrt{2N}$. They are initialized at $\mathbf{y}^{(0)} = 100 \cdot \mathbf{1}$ with $\sigma^{(0)} = 1000$. Each run is colored differently for better visibility.

This does not pose an issue from an algorithmic perspective, as $\pm\sigma\mathcal{N}(0,1)$ yields the same distribution.

The self-adaptive ES using log-normal sampling will be denoted as σSA_L, while the ES using normal sampling will be abbreviated as σSA_N.

In Fig. 2 a first evaluation of both σSA_L and σSA_N on the same configuration as in Fig. 1 is done. To this end, repeated evaluations of Alg. 1 are performed and the residual distance dynamics $R(g)$ is displayed for a convergence analysis. All runs of σSA_L (left) are trapped in the steady-state, while all runs of σSA_N are converging globally. One would expect that the steady-state of σSA_L could be avoided, if a larger μ or smaller learning parameter τ are chosen facilitating a more robust search. However, this is not the case and will be analyzed later. Figure 2 is a first example showing that the normal operator does remove the undesired steady-state. However, an open question is what the strategy and fitness parameter dependencies are, such that the steady-state is observed (or not). To this end, a few theoretical results will be stated in Sec. III. Then, in Sec. III-C, a more detailed comparison between σSA_L and σSA_N will be given.

III. Theoretical Model of the Steady State

The steady-state is characterized by vanishing progress of the ES (on average, see R-dynamics in Fig. 1a) and by a halting σ-adaptation (see constant σ-level in Fig. 1a). The theoretical analysis of the steady-state requires evaluating the R-dependent progress rate and self-adaptation response of the σ-adaptation. The second order progress rate φ_R^{II} is defined as

$$\varphi_R^{\text{II}} \coloneqq \mathrm{E}\left[(R^{(g)})^2 - (R^{(g+1)})^2 \,\middle|\, R^{(g)}, \sigma^{(g)} \right]. \tag{7}$$

The self-adaptation response (SAR) function is defined as the expected value over the relative σ-change

$$\psi \coloneqq \mathrm{E}\left[(\sigma^{(g+1)} - \sigma^{(g)})/\sigma^{(g)} \,\middle|\, R^{(g)}, \sigma^{(g)} \right]. \tag{8}$$

For the steady-state the following conditions hold (no R-progress and σ-change in expectation)

$$\text{Steady State:} \quad \varphi_R^{\text{II}} = 0 \quad \text{and} \quad \psi = 0. \tag{9}$$

Quantities (7) and (8) were already derived for the Rastrigin function in [12] and [3], respectively. φ_R^{II} was derived for slow self-adaptation in the limit $\tau \to 0$ (necessary condition to obtain a closed-form solution) and contains no τ-dependence. ψ was derived as a function of τ using the log-normal operator. In Sec. III-B the derivation of ψ is modified to account for the normal operator. First, all relevant quantities for the theoretical analysis will be introduced. Finally, the steady-state (9) will be evaluated.

A. Progress Rate and Quality Gain

The progress rate measures the (expected) search space progress on Rastrigin and was derived in [12]. After normalization ($\varphi_R^{\text{II},*} = \varphi_R^{\text{II}} N/2R^2$), the Rastrigin progress rate yields in terms of functions g and h (see [12, (B.5)])

$$\varphi_R^{\text{II},*} = \frac{c_\vartheta \sigma^* (1+g)}{\sqrt{1 + \frac{\sigma^{*2}}{2N} + h}} - \frac{\sigma^{*2}}{2\mu}. \tag{10}$$

Progress coefficient c_ϑ is given in (21). The parameters g and h in (10), respectively, are defined as

$$g \coloneqq \frac{\alpha^2 A}{2} \mathrm{e}^{-\frac{(\alpha R)^2}{2}\left(\frac{\sigma^{*2}}{N^2} + \frac{1}{N}\right)} \tag{11}$$

$$\begin{aligned} h \coloneqq{}& \frac{N^2}{4R^4\sigma^{*2}} \\ &\times \left\{ \frac{NA^2}{2} \left[1 - \mathrm{e}^{-\left(\frac{\alpha R \sigma^*}{N}\right)^2}\right] \left[1 - \mathrm{e}^{-(\alpha R)^2\left[\left(\frac{\sigma^*}{N}\right)^2 + \frac{2}{N}\right]}\right] \right. \\ &\left. + 2NA\alpha^2 R^4 \left(\frac{\sigma^*}{N}\right)^2 \left[\left(\frac{\sigma^*}{N}\right)^2 + \frac{2}{N}\right] \mathrm{e}^{-\frac{(\alpha R)^2}{2}\left[\left(\frac{\sigma^*}{N}\right)^2 + \frac{1}{N}\right]} \right\}. \end{aligned} \tag{12}$$

For $g = h = 0$, (10) yields the well-known sphere progress rate with gain term $c_\vartheta \sigma^* / \sqrt{1 + \sigma^{*2}/2N}$ and loss term $-\sigma^{*2}/2\mu$ due to unsuccessful mutations, see [12, (39)]. On Rastrigin $g, h > 0$ yield correction terms accounting for local attraction.

The derivation of progress rate (10) requires the evaluation of the quality gain Q due to random mutation $\mathbf{x} \sim \sigma\mathcal{N}(\mathbf{0}, \mathbf{1})$

$$Q \coloneqq f(\mathbf{y} + \mathbf{x}) - f(\mathbf{y}). \tag{13}$$

The normal approximation is used as a probabilistic model of the quality gain distribution function P_Q [15]

$$P_Q(q) = \Phi\left(\frac{q - \mathrm{E}[Q]}{\sqrt{\mathrm{Var}[Q]}} \right), \tag{14}$$

with Φ denoting the normal distribution function. The expected value $E_Q \coloneqq \mathrm{E}[Q]$ of the quality gain is given in [16, (1.74)]

$$E_Q = N\sigma^2 + NA\mathrm{e}^{-\frac{1}{2}\frac{(\alpha R)^2}{N}} \left(1 - \mathrm{e}^{-\frac{(\alpha\sigma)^2}{2}}\right). \tag{15}$$

The variance $D_Q^2 \coloneqq \mathrm{Var}[Q]$ was derived in [12, (36)]

$$\begin{aligned} D_Q^2 = {}& 4R^2\sigma^2 + 2N\sigma^4 + \\ &+ \frac{NA^2}{2}\left(1 - \mathrm{e}^{-(\alpha\sigma)^2}\right)\left(1 - \mathrm{e}^{-\alpha^2\left(\sigma^2 + 2\frac{R^2}{N}\right)}\right) \\ &+ 2NA\alpha^2\sigma^2 \mathrm{e}^{-\frac{\alpha^2}{2}\left(\sigma^2 + \frac{R^2}{N}\right)}\left(\sigma^2 + 2\frac{R^2}{N}\right). \end{aligned} \tag{16}$$

Note that both E_Q and D_Q^2 will reappear during the derivation of ψ. Now variance (16) is rearranged using scale-invariant mutations $\sigma = \sigma^* R/N$ in terms of h from (12)

$$D_Q^2 = 4R^4 \left(\sigma^*/N\right)^2 \left[1 + \sigma^{*2}/2N + h\right], \tag{17}$$

where h contains Rastrigin-specific terms. This will be useful later.

B. Self-Adaptation Response

The self-adaptation response (SAR) function is defined as the expected value over the relative σ-change, see (8). In [3] it was derived for $\tau \to 0$ and $\mu \to \infty$ (constant μ/λ) using mutation strength density p_σ as

$$\psi \simeq \frac{1}{\vartheta} \int_{s_l}^{s_u} \left(\frac{s-\sigma}{\sigma}\right) p_\sigma(s) \times \Phi\left(\frac{E_Q(\sigma) + D_Q(\sigma)\Phi_\vartheta^{-1} - E_Q(s)}{D_Q(s)}\right) \mathrm{d}s\,. \tag{18}$$

Note that p_σ is the density of either log-normally or normally distributed mutations with $s \in (0, \infty)$ or $s \in (-\infty, \infty)$, respectively. The abbreviation $\Phi_\vartheta^{-1} \coloneqq \Phi^{-1}(\vartheta)$ is used with Φ^{-1} denoting the quantile function of the normal distribution. To provide closed-form solutions of (18) $\Phi(s, \sigma)$ is Taylor-expanded in terms of factor $\Delta_s^k \coloneqq (s-\sigma)^k/\sigma^k$ and truncated for small Δ_s by assuming $\tau \to 0$. Abbreviating $\Phi_\sigma^{(k)} \coloneqq \partial^k \Phi / \partial s^k \,|_{s=\sigma}$ one expands (18)

$$\begin{aligned}\psi &= \frac{1}{\vartheta} \int_{s_l}^{s_u} \Delta_s p_\sigma(s) \sum_{k=0}^{\infty} \frac{\sigma^k}{k!} \left.\frac{\partial^k \Phi}{\partial s^k}\right|_{s=\sigma} \Delta_s^k \,\mathrm{d}s \\ &= \frac{1}{\vartheta} \int_{s_l}^{s_u} p_\sigma(s) \left[\Phi_\sigma^{(0)} \Delta_s + \sigma \Phi_\sigma^{(1)} \Delta_s^2 + O\left(\Delta_s^3\right)\right] \mathrm{d}s \\ &\simeq \frac{1}{\vartheta} \left[\Phi_\sigma^{(0)}\, \mathrm{E}\left[\Delta_s\right] + \sigma \Phi_\sigma^{(1)}\, \mathrm{E}\left[\Delta_s^2\right] + O\left(\Delta_s^3\right)\right].\end{aligned} \tag{19}$$

The derivatives $\Phi_\sigma^{(k)}$ yield (prime $'$ denotes $\frac{\partial(.)}{\partial s}|_{s=\sigma}$)

$$\Phi_\sigma^{(0)} = \vartheta, \;\; \Phi_\sigma^{(1)} = -\frac{\mathrm{e}^{-\frac{1}{2}[\Phi_\vartheta^{-1}]^2}}{\sqrt{2\pi}} \frac{E_Q'(\sigma) + \Phi_\vartheta^{-1} D_Q'(\sigma)}{D_Q(\sigma)}. \tag{20}$$

Given (20), one can identify asymptotic progress coefficients c_ϑ and $e_\vartheta^{1,1}$ [15] as

$$c_\vartheta = \frac{\mathrm{e}^{-\frac{1}{2}[\Phi_\vartheta^{-1}]^2}}{\sqrt{2\pi}\vartheta}, \quad e_\vartheta^{1,1} = -\frac{\Phi_\vartheta^{-1} \mathrm{e}^{-\frac{1}{2}[\Phi_\vartheta^{-1}]^2}}{\sqrt{2\pi}\vartheta}. \tag{21}$$

Finally, the expected values in (19) using log-normal mutations (3) yield $\mathrm{E}\left[\Delta_s\right] = \tau^2/2 + O\left(\tau^4\right)$ and $\mathrm{E}\left[\Delta_s^2\right] = \tau^2 + O\left(\tau^4\right)$, see also [3]. For normal mutations (5) one has $\mathrm{E}\left[\Delta_s\right] = 0$ and $\mathrm{E}\left[\Delta_s^2\right] = \tau^2$. The SAR functions therefore yield

$$\sigma\mathrm{SA}_L: \quad \psi_L \simeq \tau^2 \left(\frac{1}{2} - c_\vartheta \sigma \frac{E_Q'}{D_Q} + e_\vartheta^{1,1} \sigma \frac{D_Q'}{D_Q}\right) \tag{22}$$

$$\sigma\mathrm{SA}_N: \quad \psi_N \simeq \tau^2 \sigma \left(-c_\vartheta \frac{E_Q'}{D_Q} + e_\vartheta^{1,1} \frac{D_Q'}{D_Q}\right). \tag{23}$$

The term $1/2$ introduces a constant bias using $\sigma\mathrm{SA}_L$, which does not exists for $\sigma\mathrm{SA}_N$. This term will provide steady-state solutions for the $\sigma\mathrm{SA}_L$.

Note that the subsequent analysis will be done for $\vartheta = \mu/\lambda = 1/2$, which yields $e_\vartheta^{1,1} = 0$ ($\Phi_\vartheta^{-1} = 0$). This simplification is necessary to provide a closed-form solution for the steady-state. For better readability the expressions for E_Q' and D_Q are not inserted into (22) and (23). The derivative $\frac{\partial E_Q}{\partial \sigma}$ of (15) yields (setting $\sigma = \sigma^* R/N$)

$$\begin{aligned}E_Q' &= 2N\sigma + \alpha^2 AN\sigma \mathrm{e}^{-\frac{\alpha^2}{2}\left(\sigma^2 + \frac{R^2}{N}\right)} \\ &= 2\sigma^* R\left(1 + \frac{\alpha^2 A}{2} \mathrm{e}^{-\frac{\alpha^2 R^2}{2}\left(\frac{\sigma^{*2}}{N^2} + \frac{1}{N}\right)}\right) \\ &= 2\sigma^* R(1+g),\end{aligned} \tag{24}$$

applying definition (11) to obtain the last line. D_Q is already given in (17).

C. Steady State Analysis

Based on the results of the progress rate (10) and self-adaptation responses (22) and (23), the steady-state condition (9) on Rastrigin is investigated. Demanding vanishing (normalized) progress $\varphi_R^{\mathrm{II},*} = 0$ in (10), one has

$$\frac{c_\vartheta \sigma^*(1+g)}{\sqrt{1 + \sigma^{*2}/2N + h}} = \frac{\sigma^{*2}}{2\mu}. \tag{25}$$

Demanding $\psi = 0$ in (22) and (23) (using (17), (24), $e_\vartheta^{1,1} = 0$, $\sigma = \sigma^* R/N$) yields for the $\sigma\mathrm{SA}_L$ and $\sigma\mathrm{SA}_N$

$$\sigma\mathrm{SA}_L: \quad \frac{1}{2} = \frac{c_\vartheta \sigma^*(1+g)}{\sqrt{1 + \sigma^{*2}/2N + h}}, \tag{26}$$

$$\sigma\mathrm{SA}_N: \quad 0 = \frac{c_\vartheta \sigma^*(1+g)}{\sqrt{1 + \sigma^{*2}/2N + h}}. \tag{27}$$

Note that on Rastrigin one has $g > 0$ and $h > 0$. Equation (26) will in general have a solution satisfying $\psi = 0$ for $\sigma^* > 0$ (see visualization of $\psi = 0$ in Fig. 3a). On the other hand, (27) only vanishes if $\sigma^* = 0$. For any mutation strength $\sigma^* > 0$ no solution exists and $\psi < 0$, i.e., there is no steady-state of the σ-adaptation.

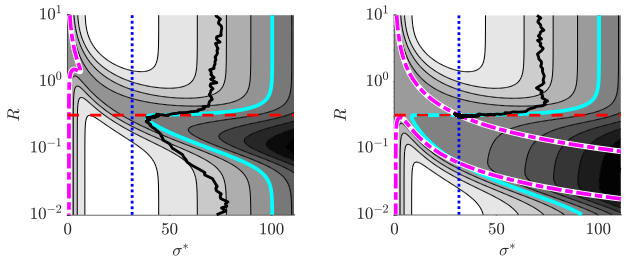
For vanishing progress (25) can be inserted into (26) giving the steady-state conditions ($\sigma^* = \sigma_{ss}^*$)

$$\sigma\mathrm{SA}_L: \quad \frac{1}{2} = \frac{\sigma_{ss}^{*2}}{2\mu} \iff \sigma_{ss,L}^* = \sqrt{\mu} \tag{28}$$

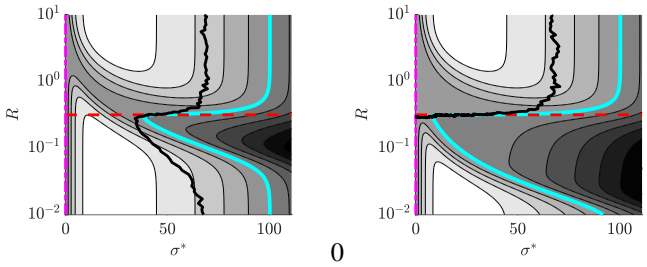
$$\sigma\mathrm{SA}_N: \quad 0 = \frac{\sigma_{ss}^{*2}}{2\mu} \iff \sigma_{ss,N}^* = 0. \tag{29}$$

Result (28) is not new and was already derived in [13] for the noisy sphere (2). However, its occurrence on multimodal functions is new and surprising. This relation is investigated further in Sec. III-D. Note that the steady-state occurs independent of the choice of τ.

The obtained results will now be illustrated for different configurations of the Rastrigin function. As stagnation of the R- and σ^*-dynamics occurs, results of Alg. 1 will be displayed as function of $R(\sigma^*)$ (with generations g implicitly given). To this end, the median over 100 trials is taken as an aggregated measure instead of showing individual runs. Then, the results are overlaid with $\varphi_R^{\mathrm{II},*}(\sigma^*, R)$ and $\psi(\sigma^*, R)$, which predict



(a) $\sigma\mathrm{SA}_L$ showing steady-state.



(b) $\sigma\mathrm{SA}_N$ showing no steady-state.

Fig. 3: Dynamics of $(1000/1000_I, 2000)$-ES on Rastrigin $N = 20$, $A = 5$, with varying α (left: $\alpha = 2\pi$) and (right: $\alpha = 10\pi$). The bold black curve shows the median dynamics of Alg. 1 over 100 trials. $\varphi_R^{\mathrm{II},*} = 0$ from (25) is shown as bold cyan, and $\psi = 0$ from (26) and (27) as bold dash-dotted magenta. The dotted blue line marks $\sigma^* = \sqrt{\mu}$ from (28) and the dashed red line $R = R_\infty$ from (34).

progress rate and SAR as expected values at each given (σ^*, R). At the stagnation point $\varphi_R^{\mathrm{II},*} = 0$ and $\psi = 0$ are expected to occur.

Figure 3 shows an exemplary σ^*-R-landscape. The gray background indicates the progress rate $\varphi_R^{\mathrm{II},*}$ (large values are white, small (negative) values are black). The bold black curve shows the median σ^*-R-dynamics over all trials. The ES moves from top to bottom (minimizing R with positive progress rate) achieving a constant scale-invariant σ^*-level, where the Rastrigin function is virtually the sphere. A characteristic σ^*-decrease occurs when local attraction starts to dominate. For R small enough the ES reaches the global attractor increasing σ^* again optimizing a quadratic function.

In Fig. 3a (left) at small α, there is no steady-state of the $\sigma\mathrm{SA}_L$. The problem is "simple enough" for the ES to achieve positive progress without reaching the $\psi = 0$ boundary. At larger α (right) the effect of local attractors increases significantly. The boundary $\varphi_R^{\mathrm{II},*} = 0$ now reaches smaller σ^*-values, which can be attributed to local attraction enforcing smaller mutations strengths. Additionally, the boundary $\psi = 0$ degenerates occurring over a large range of σ^*-values. At the steady-state one has $\varphi_R^{\mathrm{II},*} = 0$ and $\psi = 0$ with $\sigma^* \approx \sqrt{\mu}$ from (28). The median of Alg. 1 agrees well with the prediction. Even though local attraction is present (at high frequency), the $\sigma\mathrm{SA}_L$ is not converging locally.

Figure 3b shows the same experiment using $\sigma\mathrm{SA}_N$. As expected, there is no steady-state independent of the choice of α. Instead, at $R \approx R_\infty$ (residual distance under constant noise, see (34)), the ES reduces its σ^*-level continuously until it gets trapped in a local attractor (Fig. 3b, right). This is in contrast to Fig. 3a (right), where the $\sigma\mathrm{SA}_L$ gets trapped in the steady-state at $R \approx R_\infty$ and $\sigma^* \approx \sqrt{\mu}$.

For Rastrigin, depending on α and A, the progress rate may (or may not) become zero satisfying (25). An exact solution of (25) is not possible due to the non-linearity of the underlying terms (a first attempt in approximating the smallest σ^* having zero progress is given in [3]). The same holds for ψ in (26) or (27). One can therefore not predict at which parameter configuration of N, α, and A the steady-state of the $\sigma\mathrm{SA}_L$ will occur. However, in the limit of $\alpha \to \infty$ the occurrence can be guaranteed. To understand this, an important connection of the Rastrigin function to the noisy sphere is made in the next section.

D. Steady State of the Noisy Sphere (or Rastrigin $\alpha \to \infty$)

The (normalized) progress rate of the noisy sphere (2) is given by [17]

$$\varphi^* = \frac{c_{\mu/\mu,\lambda}\sigma^*(1 + \sigma^{*2}/(2\mu N))}{\sqrt{1 + \frac{\sigma^{*2}}{\mu N}}\sqrt{1 + \frac{\sigma^{*2}}{2N} + \vartheta_\epsilon^2}} - N\left(\sqrt{1 + \frac{\sigma^{*2}}{\mu N}} - 1\right), \tag{30}$$

with progress coefficient $c_{\mu/\mu,\lambda}$, see [6, p. 216], noise-to-signal ratio $\vartheta_\epsilon^2 = (\sigma_\epsilon N/2R^2)^2/\sigma^{*2}$. For large $\mu N \gg \sigma^{*2}$ Taylor expansion $\sqrt{1 + \sigma^{*2}/\mu N} = 1 + \sigma^{*2}/2\mu N + O((\sigma^{*2}/\mu N)^2)$ yields the simplified formula for constant noise strength σ_ϵ

$$\varphi^* = \frac{c_{\mu/\mu,\lambda}\sigma^*}{\sqrt{1 + \frac{\sigma^{*2}}{2N} + \frac{N^2\sigma_\epsilon^2}{4R^4\sigma^{*2}}}} - \frac{\sigma^{*2}}{2\mu}. \tag{31}$$

Note that for $\sigma_\epsilon = 0$ one obtains the regular progress rate of the sphere. Now some relations between (31) and the Rastrigin progress rate (10) are established based on parameters g (11) and h (12). For moderately large μ, i.e., $\mu \gtrapprox 50$, one has $c_{\mu/\mu,\lambda} \simeq c_\vartheta$ and the progress coefficient can be replaced by its asymptotic form being a function of $\vartheta = \mu/\lambda$. Furthermore, applying the limit $\alpha \to \infty$ to g and h yields

$$\begin{aligned} \lim_{\alpha\to\infty} g &= 0 \\ \lim_{\alpha\to\infty} h &= \frac{N^2}{4R^4\sigma^{*2}}\frac{NA^2}{2} =: \frac{N^2\sigma_{\text{ras}}^2}{4R^4\sigma^{*2}}, \end{aligned} \tag{32}$$

with Rastrigin noise variance

$$\sigma_{\text{ras}}^2 := NA^2/2. \tag{33}$$

Inserting the limits (32) into (10), the Rastrigin progress rate approaches the noisy sphere progress rate (31) with corresponding noise strength $\sigma_\epsilon = \sigma_{\text{ras}}$ (example shown in Fig. 1). The limit makes intuitively sense, as for infinitely large frequency the multimodality degenerates to the case of having constant noise variance. This relation is useful as results from noisy sphere theory are applicable.

The zero-progress condition (25) yields for $\sigma^* \to 0$ [12, (B.11), (B.12)] the residual distance under constant noise

$$\varphi_R^{\mathrm{II},*} = 0: \quad R_\infty = \sqrt{\sigma_{\mathrm{ras}} N / 4 c_\vartheta \mu}. \tag{34}$$

Note that (34) is valid for both $\sigma\mathrm{SA}_L$ and $\sigma\mathrm{SA}_N$ (see Fig. 3), as it is derived from progress rate results independent of the self-adaptation response. For the $\sigma\mathrm{SA}_L$ $\psi = 0$ occurs for $\sigma^* > 0$ via (28). Using $\sigma_{ss}^* = \sqrt{\mu}$ with $\sigma_{ss} = \sigma_{ss}^* R_\infty / N$ one obtains

$$\psi_L = 0: \quad \sigma_{ss} = \sqrt{\sigma_{\mathrm{ras}} / 4 c_\vartheta N}. \tag{35}$$

The derived stationary σ-level could also be observed in Fig. 1a being $\sigma_{ss} \approx 1$. Result (35) will be useful in Sec. III-E to detect the steady-state.

As already mentioned, one cannot actually predict for Rastrigin at which α and A the steady-state occurs using the $\sigma\mathrm{SA}_L$. However, it was shown theoretically that it exists at sufficiently high multimodality (see also experiments in Fig. 4). In this limit the σ-adaptation on Rastrigin behaves as for the sphere under constant noise. The $\sigma\mathrm{SA}_L$ is forced into a steady-state, while the $\sigma\mathrm{SA}_N$ decreases σ continuously

$$R = R_\infty: \quad \psi_N < 0 \text{ for } \sigma^* > 0. \tag{36}$$

Relation (36) is shown now by inserting E_Q' and D_Q into (23). The quality gain of the noisy sphere (2) is given by $Q = \sum_i (2 y_i x_i + x_i^2) + \sigma_\epsilon \mathcal{N}(0,1)$ [18, (29)]. Expected value $\mathrm{E}[Q]$ and its derivative yield ($\sigma = \sigma^* R / N$ is applied)

$$\mathrm{E}[Q] = E_Q = N\sigma^2, \quad E_Q' = 2N\sigma = 2\sigma^* R. \tag{37}$$

For the variance one has $\mathrm{Var}[Q] = \sum_i \mathrm{Var}\left[2 y_i x_i + x_i^2\right] + \mathrm{Var}\left[\mathcal{N}(0, \sigma_\epsilon^2)\right] = \sum_i (4 y_i^2 \sigma^2 + 2\sigma^4) + \sigma_\epsilon^2$, see [18, (31)], such that with $R^2 = \sum_i y_i^2$ one gets

$$\begin{aligned} \mathrm{Var}[Q] = D_Q^2 &= 4R^2\sigma^2 + 2N\sigma^4 + \sigma_\epsilon^2 \\ &= 4R^4 \left(\frac{\sigma^*}{N}\right)^2 \left[1 + \frac{\sigma^{*2}}{2N} + \frac{N^2 \sigma_\epsilon^2}{4R^4 \sigma^{*2}}\right]. \end{aligned} \tag{38}$$

Inserting (37) and (38) into (23) yields

$$\psi_N = -\tau^2 c_\vartheta \sigma^* \Big/ \sqrt{1 + \frac{\sigma^{*2}}{2N} + \frac{N^2 \sigma_\epsilon^2}{4R^4 \sigma^{*2}}}. \tag{39}$$

On the noisy sphere the ES approaches the minimal residual distance R_∞, see (34). As $\psi_N < 0$ in (39) and R is constant (in expectation), one observes a decrease of σ and σ^* at $R = R_\infty$. This behavior guarantees local convergence on the Rastrigin function even in the limit of $\alpha \to \infty$ without the occurrence of a steady-state. Note that the CSA has shown a similar σ-decrease in Fig. 1b (right).

E. Experimental Investigation

In this section a series of experiments is conducted. The theoretical predictions indicate that the steady-state will occur for the $\sigma\mathrm{SA}_L$ (for sufficiently large α), but not for the $\sigma\mathrm{SA}_N$. To investigate the parameter dependency A and α will be varied for two configurations of N and population size parameters. The learning parameter is kept at default $\tau = 1/\sqrt{2N}$. All runs will be initialized outside the local attraction region by setting $\mathbf{y}^{(0)} = \lceil \alpha A / 2 \rceil \cdot \mathbf{1}$ and using $\sigma^{(0)} = \sigma_{\varphi_0}^* N / \left\| \mathbf{y}^{(0)} \right\|$ with $\sigma_{\varphi_0}^* = [(N^2 + 8 N c_\vartheta^2 \mu^2)^{1/2} - N]^{1/2}$ (this initialization ensures a large initial step-size according to the global structure of Rastrigin, see [12]).

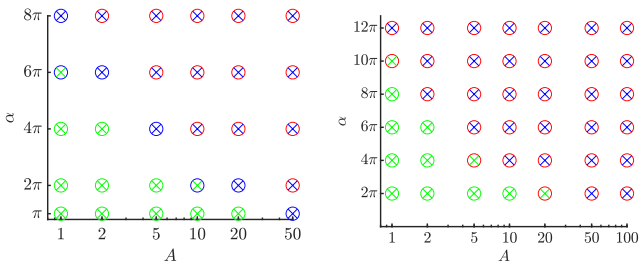


Fig. 4: Steady-state occurrence for $\sigma\mathrm{SA}_L$ ($\circ$) and $\sigma\mathrm{SA}_N$ ($\times$) with left: $N = 100$, $\mu = 1000$, $\lambda = 2000$, $F = 2 \cdot 10^8$; and right: $N = 20$, $\mu = 500$, $\lambda = 2000$, $F = 1.2 \cdot 10^8$. Green: all trials of a configuration successful; blue: mixed results with global and local convergence, but no steady-state; red: steady-state occurs for all trials.

For each configuration 10 trials are evaluated and there are three possible outcomes for each run (given a budget of function evaluations F). For $R < R_{\mathrm{stop}} = 10^{-3}$ global convergence is detected. For $\sigma < \sigma_{\mathrm{stop}} = 10^{-4}$ local convergence is detected. If the budget F is exhausted and the mean value of σ (over the last 1000 generations) is within 10% of (35), the steady-state is detected for the $\sigma\mathrm{SA}_L$. As will be seen, no steady-state is detected for $\sigma\mathrm{SA}_N$. F is determined by the average number of function evaluations the $\sigma\mathrm{SA}_N$ requires to achieve local convergence in the worst case (largest A and α), and then increased by a factor of 10.

Figure 4 displays the results. One can clearly see that no steady-state occurs for the $\sigma\mathrm{SA}_N$ (green and blue $\times$), while it does occur for the $\sigma\mathrm{SA}_L$ (red $\circ$), especially at large A or α. Both strategies are less successful for increased A or α, which was expected (population sizing results were derived in [2], [12]). Note that the theoretical results of Sec. III-C only apply to $\vartheta = 1/2$. However, the experiments in Fig. 4 (right) confirm that similar results also hold for smaller $\vartheta = 1/4$, which is closer to sphere-optimal truncation ratios.

IV. Stability Investigation

In this section a few distinct properties of the $\sigma\mathrm{SA}_L$ and $\sigma\mathrm{SA}_N$ are discussed by investigating the σ-dynamics under specific conditions. It will be shown that the σSA-ES using log-normal sampling may become unstable on a multimodal function showing similar behavior as on a random function. To this end, the following functions are defined

$$\text{Random function } f_{\mathrm{ran}}(\mathbf{y}) \coloneqq \mathcal{N}(0,1) \tag{40}$$

$$\text{Cosine function } f_{\cos}(\mathbf{y}) \coloneqq \sum_{i=1}^{N} A(1 - \cos(\alpha y_i)). \tag{41}$$

Note that functions (40) and (41) correspond to (1) and (2), respectively, with removed y_i^2-terms. Function $f_{\cos}$ can

be regarded as a very simple model of a highly-multimodal function without global structure. All local attractors are equal and at larger scales there is no preferred direction, i.e., it appears as a flat fitness landscape. Functions f_{ran} and $f_{\cos}$, albeit being completely different, will share some properties regarding the σ-dynamics over many generations.

An updated $\sigma^{(g+1)}$ is obtained after selection of the m-th best sampled value $\widetilde{\sigma}_{m;\lambda}$ and intermediate recombination (see Line 11 of Alg. 1)

$$\sigma^{(g+1)} = \frac{1}{\mu}\sum_{m=1}^{\mu}\widetilde{\sigma}_{m;\lambda}. \tag{42}$$

On a flat fitness landscape there is no fitness preference, c.f. analysis in [19]. This also holds on the random function in expectation. In both cases one has

$$\mathrm{E}\left[\sigma^{(g+1)}\right] = \frac{1}{\mu}\sum_{m=1}^{\mu}\mathrm{E}\left[\widetilde{\sigma}_{m;\lambda}\right] = \mathrm{E}\left[\widetilde{\sigma}\right]. \tag{43}$$

Applying expected value (4) ($\sigma\mathrm{SA}_L$) over multiple generations g one may predict the mean σ-growth as (the notation indicating recombination was dropped, as it has no influence)

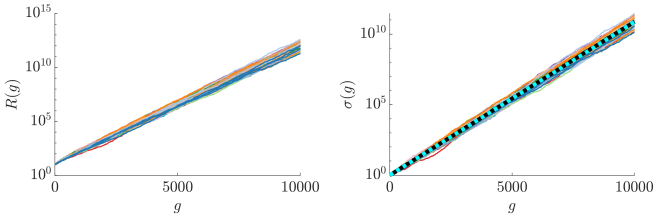
$$\begin{aligned}\mathrm{E}\left[\sigma_L^{(g)}\right] &= \mathrm{E}\left[\sigma^{(0)}\mathrm{e}^{\tau\mathcal{N}(0,1)}\cdots\mathrm{e}^{\tau\mathcal{N}(0,1)}\right]\\ &= \sigma^{(0)}\,\mathrm{E}\left[\mathrm{e}^{\tau\mathcal{N}(0,1)}\right]\cdots\mathrm{E}\left[\mathrm{e}^{\tau\mathcal{N}(0,1)}\right]\\ &= \sigma^{(0)}\mathrm{e}^{\frac{\tau^2}{2}}\cdots\mathrm{e}^{\frac{\tau^2}{2}} = \sigma^{(0)}\mathrm{e}^{\frac{\tau^2}{2}g}.\end{aligned} \tag{44}$$

The $\sigma\mathrm{SA}_L$ therefore exhibits a positive bias towards larger mutations on the random function. For $\sigma\mathrm{SA}_N$, according to expected value (6), σ does not change (on average), such that

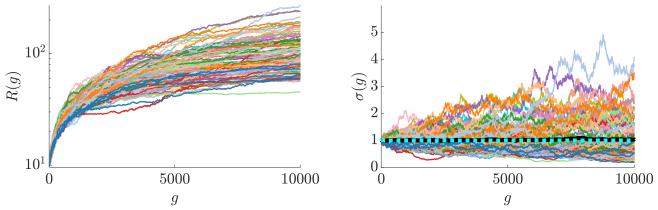
$$\begin{aligned}&\mathrm{E}\left[\sigma_N^{(g)}\right] = \mathrm{E}\left[\sigma^{(0)}(1+\tau\mathcal{N}(0,1))\cdots(1+\tau\mathcal{N}(0,1))\right]\\ &= \sigma^{(0)}\,\mathrm{E}\left[1+\tau\mathcal{N}(0,1)\right]\cdots\mathrm{E}\left[1+\tau\mathcal{N}(0,1)\right] = \sigma^{(0)}.\end{aligned} \tag{45}$$

Figure 5 shows the dynamics of Alg. 1 on the random function (40). The $\sigma\mathrm{SA}_L$ (top row) shows the expected behavior of exponential σ-growth (44), which in turn leads to exponential growth of $R(g)$. Note that R is used as a pure distance measure relative to the initial value, as there is no global optimizer. On the other hand, $\sigma\mathrm{SA}_N$ (bottom row) exhibits a stable σ-dynamics, as expected from (45), with the mean-level remaining close to $\sigma^{(0)} = 1$. The mean $R(g)$-dynamics exhibits a random walk under constant σ (in expectation) obeying the $\sqrt{g}$ law of Gaussian diffusion processes. The stability of $\sigma^{(g)}$ is useful for the subsequent experiment.

Figure 6 shows the dynamics of Alg. 1 on the cosine function (41). The initial $\sigma^{(0)} = 0.5$ was chosen slightly above an escape threshold $\sigma_{\text{esc}}^{(0)} \approx 0.48$. By choosing $\sigma^{(0)} < \sigma_{\text{esc}}^{(0)}$ all runs would converge locally. Therefore, a larger value $\sigma^{(0)} > \sigma_{\text{esc}}^{(0)}$ is taken to prevent immediate local convergence and investigate the stability of the adaptation. $\sigma\mathrm{SA}_L$ again shows a prominent σ-growth, as the oscillation (without global structure) behaves in a way similar to the random function. On a large scale the cosine function may be regarded as a flat landscape. The observed growth agrees well with (44). Due

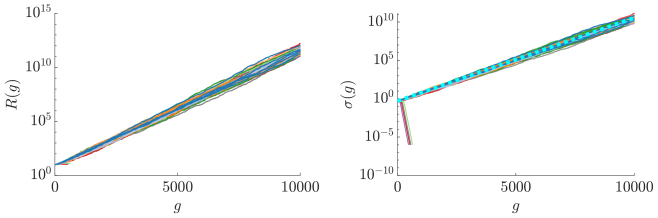


(a) $\sigma\mathrm{SA}_L$: σ-dynamics (right) predicted by (44) (dotted cyan).

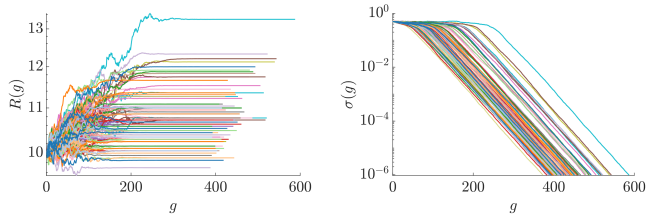


(b) $\sigma\mathrm{SA}_N$: σ-dynamics (right) predicted by (45) (dotted cyan).

Fig. 5: Dynamics of 100 runs using $(100/100_I, 200)$-ES with $\mathbf{y}^{(0)} = \mathbf{1}$ and $\sigma^{(0)} = 1$ on random function (40) ($N = 100$) showing $R(g)$ (left column) and $\sigma(g)$ (right column) with $\tau = 1/\sqrt{2N}$. The bold black curves show the mean σ-value over all trials.



(a) $\sigma\mathrm{SA}_L$: only 4 runs (out of 100) are converging locally and 96 runs exhibit divergence. σ-dynamics predicted by (44) (dotted cyan).



(b) $\sigma\mathrm{SA}_N$: all runs converging locally.

Fig. 6: Dynamics of 100 runs using $(100/100_I, 200)$-ES with $\mathbf{y}^{(0)} = \mathbf{1}$ and $\sigma^{(0)} = 0.5$ on the cosine function (41) ($N = 100$) showing $R(g)$ (left column) and $\sigma(g)$ (right column) with $\tau = 1/\sqrt{2N}$.

to the positive bias most of the runs exhibit σ-divergence and only 4 runs converge (see runs with decreasing σ-dynamics). Even though the cosine function is not a random function, the bias generates positive (increasing) feedback on the σ-growth and most runs diverge. On the other hand, using $\sigma\mathrm{SA}_N$ all runs converge. There is no bias, such that σ is kept closer to its initial value. The ES has more time to find and descend into a local attractor, which is observed for all runs.

One could argue that the bias is useful to escape a cer-

tain region of local attraction. However, it may also prevent convergence of the ES. The resulting behavior on multimodal functions may resemble the dynamics on the random function (40) or on the noisy sphere (2). The obtained theoretical and experimental results therefore suggest choosing the normal operator for self-adaptation on highly multimodal test functions. Based on the structure of the self-adaptation response (22) and (23), the removed bias does not change the derivation of the default learning parameter recommendation $\tau = 1/\sqrt{2N}$ (see also [13, (4.116)]). This property is desirable, as it makes the change from the log-normal to the normal operator straightforward.

V. Conclusions and Outlook

A self-adaptive Evolution Strategy adapts its mutation strength σ by sampling from a distribution and (implicitly) selecting successful values from the offspring population. In this paper, the σ-sampling from a log-normal and normal distribution were investigated on the Rastrigin function and the noisy sphere. In the limit of high multimodality, an undesired steady-state using log-normal sampling has been observed which prevents convergence of the algorithm. This observation motivated to introduce sampling from a normal distribution.

For the theoretical investigation, a steady-state condition was defined demanding vanishing progress and a halting of the σ-adaptation. It was shown that log-normally sampled mutations have solutions satisfying the steady-state condition. On the other hand, assuming $\mu/\lambda = 1/2$, it could be shown that normal sampling does not exhibit a steady-state behavior on Rastrigin as the self-adaptation response never vanishes. The theoretical results were validated using experiments on Rastrigin under high multimodality by testing large values for amplitude A and frequency α. Furthermore, the analysis has revealed remarkable similarities between the noisy sphere and Rastrigin in the limit of infinitely high frequency.

The observed results when sampling log-normal mutations can be explained by the bias towards larger mutations due to $\mathrm{E}\left[\widetilde{\sigma}\right] > \sigma$. Normal sampling on the other hand is unbiased with $\mathrm{E}\left[\widetilde{\sigma}\right] = \sigma$. The latter has shown to be more suitable for the optimization under high multimodality since no halting of the σ-adaptation occurs. This should hold for test functions that share multimodal properties with Rastrigin. First tests have shown similar results on Bohachevsky and Griewank functions, removing undesired steady-state behavior. However, there are cases in which the bias is beneficial, such as on the noisy sphere. With active population control, μ can be gradually increased to decrease the distance to the optimizer. The bias helps to achieve larger mutation strengths and improve the signal-to-noise ratio. Furthermore, the bias may help to escape regions of local attraction and improve global convergence in some cases (observed on Bohachevsky). A detailed experimental survey on a larger set of noisy and multimodal test functions is left for future research.

References

[1] N. Hansen and S. Kern, "Evaluating the CMA Evolution Strategy on Multimodal Test Functions," in *Parallel Problem Solving from Nature 8*, X. Yao et al., Ed. Berlin: Springer, 2004, pp. 282–291.

[2] L. Schönenberger and H.-G. Beyer, "On a Population Sizing Model for Evolution Strategies Optimizing the Highly Multimodal Rastrigin Function," in *Proceedings of the Genetic and Evolutionary Computation Conference*, ser. GECCO '23, New York, NY, USA, 2023, p. 848–855. [Online]. Available: https://doi.org/10.1145/3583131.3590451

[3] A. Omeradzic and H.-G. Beyer, "Self-Adaptation of Multi-Recombinant Evolution Strategies on the Highly Multimodal Rastrigin Function," 2023, Submitted to IEEE Transactions on Evolutionary Computation for possible publication. [Online]. Available: http://dx.doi.org/10.22541/au.170267088.89365626/v1

[4] I. Rechenberg, *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Stuttgart: Frommann-Holzboog Verlag, 1973.

[5] H.-P. Schwefel, *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie*, ser. Interdisciplinary systems research; 26. Basel: Birkhäuser, 1977.

[6] H.-G. Beyer, *The Theory of Evolution Strategies*, ser. Natural Computing Series. Heidelberg: Springer, 2001, DOI: 10.1007/978-3-662-04378-3.

[7] A. Ostermeier, "Schrittweitenadaptation in der Evolutionsstrategie mit einem entstochastisierten Ansatz," Doctoral thesis, Technical University of Berlin, Berlin, 1997.

[8] N. Hansen and A. Ostermeier, "Completely Derandomized Self-Adaptation in Evolution Strategies," *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, 2001, https://doi.org/10.1162/106365601750190398.

[9] D. Arnold, *Noisy Optimization with Evolution Strategies*. Dordrecht: Kluwer Academic Publishers, 2002.

[10] M. Hellwig and H.-G. Beyer, "On the steady state analysis of covariance matrix self-adaptation evolution strategies on the noisy ellipsoid model," *Theoretical Computer Science*, 2018, https://doi.org/10.1016/j.tcs.2018.05.016.

[11] H.-G. Beyer, M. Olhofer, and B. Sendhoff, "On the Impact of Systematic Noise on the Evolutionary Optimization Performance – A Sphere Model Analysis," *Genetic Programming and Evolvable Machines*, vol. 5, no. 4, pp. 327–360, 2004.

[12] A. Omeradzic and H.-G. Beyer, "Convergence Properties of the $(\mu/\mu_I,\ \lambda)$-ES on the Rastrigin Function," in *Proceedings of the 17th ACM/SIGEVO Conference on Foundations of Genetic Algorithms*, ser. FOGA '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 117–128. [Online]. Available: https://doi.org/10.1145/3594805.3607126

[13] S. Meyer-Nieberg, "Self-Adaptation in Evolution Strategies," Ph.D. dissertation, University of Dortmund, CS Department, Dortmund, Germany, 2007.

[14] D. B. Fogel, "Evolving Artificial Intelligence," Ph.D. dissertation, University of California, San Diego, 1992.

[15] A. Omeradzic and H.-G. Beyer, "Progress Analysis of a Multi-Recombinative Evolution Strategy on the Highly Multimodal Rastrigin Function," *Theoretical Computer Science*, vol. 978, 2023. [Online]. Available: https://doi.org/10.1016/j.tcs.2023.114179

[16] ——, "Rastrigin Function: Quality Gain and Progress Rate for $(\mu/\mu_I, \lambda)$-ES," https://opus.fhv.at/frontdoor/index/index/docId/5151, Vorarlberg University of Applied Sciences, Report, 2023.

[17] D. 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*, vol. 289, pp. 629–647, 2002.

[18] A. Melkozerov and H.-G. Beyer, "Towards an Analysis of Self-Adaptive Evolution Strategies on the Noisy Ellipsoid Model: Progress Rate and Self-Adaptation Response," in *GECCO'15: Proceedings of the Genetic and Evolutionary Computation Conference*. New York: ACM, 2015, pp. 297–304, DOI: 10.1145/2739480.2754800.

[19] H.-G. Beyer and K. Deb, "On Self-Adaptive Features in Real-Parameter Evolutionary Algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 5, no. 3, pp. 250–270, 2001.