Progress Rate Analysis of Evolution Strategies on the Rastrigin Function: First Results

Amir Omeradzic[1,2] [0000−0003−1979−8916] (✉) and
Hans-Georg Beyer[1,3] [0000−0002−7455−8686]

[1] Vorarlberg University of Applied Sciences, Research Center Business Informatics, Hochschulstraße 1, 6850 Dornbirn, Austria
[2] amir.omeradzic@fhv.at
[3] hans-georg.beyer@fhv.at, http://homepages.fhv.at/hgb

Abstract. A first order progress rate is derived for the intermediate multi-recombinative Evolution Strategy $(\mu/\mu_I, \lambda)$-ES on the highly multimodal Rastrigin test function. The progress is derived within a linearized model applying the method of so-called noisy order statistics. To this end, the mutation-induced variance of the Rastrigin function is determined. The obtained progress approximation is compared to simulations and yields strengths and limitations depending on mutation strength and distance to the optimizer. Furthermore, the progress is iterated using the dynamical systems approach and compared to averaged optimization runs. The property of global convergence within given approximation is discussed. As an outlook, the need of an improved first order progress rate as well as the extension to higher order progress including positional fluctuations is explained.

Keywords: Evolution Strategies · Rastrigin function · Progress rate analysis · Global optimization

1 Introduction

Evolution Strategies (ES) [13,14] are well-recognized Evolutionary Algorithms suited for real-valued non-linear optimization. State-of-the-art ES such as the CMA-ES [9] or its simplification [6] are also well-suited for locating global optimizers in highly multimodal fitness landscapes. While the CMA-ES was originally mainly intended for non-differentiable optimization problems, but yet regarded as a locally acting strategy, it was already in [8] observed that using a large population size can make the ES a strategy that is able to locate the global optimizer among a huge number of local optima. This is a surprising observation when considering the ES as a strategy that acts mainly local in the search space following some kind of gradient or natural gradient [4,7,12]. As one can easily check using standard (highly) multimodal test functions such as Rastrigin, Ackley, and Griewank to name a few, this ES property is not intimately related to the covariance matrix adaptation (CMA) ES which generates non-isotropic correlated mutations, but can also be found in $(\mu/\mu_I, \lambda)$-ES with *isotropic* mutations. Therefore, if one wants to understand the underlying working principles

how the ES locates the global optimizer, the analysis of the $(\mu/\mu_I, \lambda)$-ES should be the starting point.

The question regarding why and when optimization algorithms – originally designed for local search – are able to locate global optima has gained attention in the last few years. A recurring idea comes from relaxation procedures that transform the original multimodal optimization problem into a convex optimization problem called Gaussian continuation [10]. Gaussian continuation is nothing else but a convolution of the original optimization problem with a Gaussian kernel. As has been shown in [11], using the right Gaussian, Rastrigin-like functions can be transformed into a convex optimization problem, thus making it accessible to gradient following strategies. However, this raises the question how to perform the convolution efficiently. One road followed in [15] uses high-order Gauss-Hermite integration in conjunction with a gradient descent strategy yielding surprisingly good results. The other road coming to mind is approximating the convolution by Gaussian sampling. This resembles the procedure ES do: starting from a parental state, offspring are generated by Gaussian mutations. The problem is, however, that in order to get a reliable gradient, a huge number of samples, i.e. offspring in ES must be generated in order to get reliable convolution results. The number of offspring needed to get reliable estimates seems much larger than the offspring population size needed in ES experiments conducted in [8] showing approximately a linear relation between problem dimension N and population size for the Rastrigin function. Therefore, understanding the ES performance from viewpoint of Gaussian relaxation does not seem to help much.

The approach followed in this paper will incorporate two main concepts, namely a progress rate analysis as well as its application within the so-called evolution equations modeling the transition dynamics of the ES [3]. The progress rate measure yields the expected positional change in search space between two generations depending on location, strategy and test function parameters. Aiming to investigate and understand the dynamics of globally converging ES runs, the progress rate is an essential quantity to model the expected evolution dynamics over many generations.

This paper provides first results of a scientific program that aims at an analysis of the performance of the $(\mu/\mu_I, \lambda)$-ES on Rastrigin's test function based on a first order progress rate. After a short introduction of the $(\mu/\mu_I, \lambda)$-ES, the N-dimensional first order progress will be defined and an approximation will be derived resulting in a closed form expression. The predictive power and its limitations will be checked by one-generation experiments. The progress rate will then be used to simulate the ES dynamics on Rastrigin using difference equations. This simulation will be compared with real runs of the $(\mu/\mu_I, \lambda)$-ES. In a concluding section a summary of the results and outlook of the future research will be given.

2 Rastrigin Function and Local Quality Change

The real-valued minimization problem defined for an N-dimensional search vector $\mathbf{y} = (y_1, ..., y_N)$ is performed on the Rastrigin test function f given by

$$f(\mathbf{y}) = \sum_{i=1}^{N} f_i(y_i) = \sum_{i=1}^{N} y_i^2 + A - A\cos(\alpha y_i), \tag{1}$$

with A denoting the oscillation amplitude and $\alpha = 2\pi$ the corresponding frequency. The quadratic term with superimposed oscillations yields a finite number of local minima M for each dimension i, such that the overall number of minima scales exponentially as M^N posing a highly multimodal minimization problem. The global optimizer is at $\hat{\mathbf{y}} = \mathbf{0}$.

For the progress rate analysis in Sec. 4 the local quality function $Q_{\mathbf{y}}(\mathbf{x})$ at $\mathbf{y}$ due to mutation vector $\mathbf{x} = (x_1, ..., x_N)$ is needed. In order to reuse results from noisy progress rate theory it will be formulated for the *maximization* case of $F(\mathbf{y}) = -f(\mathbf{y})$ with $F_i(y_i) = -f_i(y_i)$, such that local quality change yields

$$Q_{\mathbf{y}}(\mathbf{x}) = F(\mathbf{y} + \mathbf{x}) - F(\mathbf{y}) = f(\mathbf{y}) - f(\mathbf{y} + \mathbf{x}). \tag{2}$$

$Q_{\mathbf{y}}(\mathbf{x})$ can be evaluated for each component i independently giving

$$Q_{\mathbf{y}}(\mathbf{x}) = \sum_{i=1}^{N} Q_i(x_i) = \sum_{i=1}^{N} f_i(y_i) - f_i(y_i + x_i) \tag{3}$$

$$= \sum_{i=1}^{N} -\left(x_i^2 + 2y_i x_i + A\cos(\alpha y_i)(1 - \cos(\alpha x_i)) + A\sin(\alpha y_i)\sin(\alpha x_i)\right). \tag{4}$$

A closed form solution of the progress rate appears to be obtainable only for a linearized expression of $Q_i(x_i)$. A first approach taken in this paper is based on a Taylor expansion for the mutation x_i and discarding higher order terms

$$Q_i(x_i) = F_i(y_i + x_i) - F_i(y_i) = \frac{\partial F_i}{\partial y_i} x_i + O(x_i^2) \tag{5}$$

$$\approx (-2y_i - \alpha A\sin(\alpha y_i))\, x_i =: -f_i' x_i, \tag{6}$$

using the following derivative terms

$$k_i = 2y_i \quad \text{and} \quad d_i = \alpha A\sin(\alpha y_i), \quad \text{such that} \quad \frac{\partial f_i}{\partial y_i} = f_i' = k_i + d_i. \tag{7}$$

A second approach is to consider only the linear term of Eq. (4) and neglect all non-linear terms denoted by $\delta(x_i)$ according to

$$Q_i(x_i) = -2y_i x_i - x_i^2 - A\cos(\alpha y_i)(1 - \cos(\alpha x_i)) - A\sin(\alpha y_i)\sin(\alpha x_i) \tag{8}$$

$$= -2y_i x_i + \delta(x_i) \approx -2y_i x_i = -k_i x_i. \tag{9}$$

The linearization using f_i' is a local approximation of the function incorporating oscillation parameters A and α. Using only k_i (setting $d_i = 0$) discards oscillations by approximating the quadratic term via $k_i = \partial(y_i^2)/\partial y_i = 2y_i$ with negative sign due to maximization. Both approximations will be evaluated later.

3 The $(\mu/\mu_I, \lambda)$-ES with Normalized Mutations

The Evolution Strategy under investigation consists of a population of μ parents and λ offspring ($\mu < \lambda$) per generation g. Algorithm 1 is presented below and offspring variables are denoted with overset "$\sim$".

Population variation is achieved by applying an isotropic normally distributed mutation $\mathbf{x} \sim \sigma\mathcal{N}(0, \mathbf{1})$ with strength σ to the parent recombinant in Lines 6 and 7. The recombinant is obtained using intermediate recombination of all μ parents equally weighted in Line 11. Selection of the $m = 1, ..., \mu$ best search vectors $\mathbf{y}_{m;\lambda}$ (out of λ) according to their fitness is performed in Line 10.

Note that the ES in Algorithm 1 operates under constant normalized mutation σ^* in Lines 3 and 12 using the spherical normalization

$$\sigma^* = \frac{\sigma^{(g)} N}{\left\|\mathbf{y}^{(g)}\right\|} = \frac{\sigma^{(g)} N}{R^{(g)}}. \tag{10}$$

This property ensures global convergence of the algorithm as the mutation strength $\sigma^{(g)}$ decreases if and only if the residual distance $\left\|\mathbf{y}^{(g)}\right\| = R^{(g)}$ decreases. While σ^* is not known during black-box optimizations, it is used here to investigate the dynamical behavior of the ES using the first order progress rate approach to be developed in this paper. Incorporating self-adaptation of σ or cumulative step-size adaptation remains for future research.

Algorithm 1 $(\mu/\mu_I, \lambda)$-ES with constant σ^*

1: $g \leftarrow 0$
2: $\mathbf{y}^{(0)} \leftarrow \mathbf{y}^{(\text{init})}$
3: $\sigma^{(0)} \leftarrow \sigma^* \left\|\mathbf{y}^{(0)}\right\| / N$
4: **repeat**
5: **for** $l = 1, ..., \lambda$ **do**
6: $\tilde{\mathbf{x}}_l \leftarrow \sigma^{(g)} \mathcal{N}_l(0, \mathbf{1})$
7: $\tilde{\mathbf{y}}_l \leftarrow \mathbf{y}^{(g)} + \tilde{\mathbf{x}}_l$
8: $\tilde{f}_l \leftarrow f(\tilde{\mathbf{y}}_l)$
9: **end for**
10: $(\tilde{\mathbf{y}}_{1;\lambda}, \ldots, \tilde{\mathbf{y}}_{\mu;\lambda}) \leftarrow \text{sort}\left(\tilde{\mathbf{y}} \text{ w.r.t. ascending } \tilde{f}\right)$
11: $\mathbf{y}^{(g+1)} \leftarrow \frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{\mathbf{y}}_{m;\lambda}$
12: $\sigma^{(g+1)} \leftarrow \sigma^* \left\|\mathbf{y}^{(g+1)}\right\| / N$
13: $g \leftarrow g + 1$
14: **until** termination criterion

4 Progress Rate

4.1 Definition

Having introduced the Evolution Strategy, we are interested in the expected one-generation progress of the optimization on the Rastrigin function (1) before investigating the dynamics over multiple generations.

A first order progress rate φ_i for the i-th component between two generations $g \to g+1$ can be defined as the expectation value over the positional difference of the parental components

$$\varphi_i = \mathrm{E}\left[y_i^{(g)} - y_i^{(g+1)} \,\middle|\, \sigma^{(g)}, \mathbf{y}^{(g)} \right] = y_i^{(g)} - \mathrm{E}\left[y_i^{(g+1)} \,\middle|\, \sigma^{(g)}, \mathbf{y}^{(g)} \right], \tag{11}$$

given mutation strength $\sigma^{(g)}$ and the position $\mathbf{y}^{(g)}$. First, an expression for $\mathbf{y}^{(g+1)}$ is needed, see Alg. 1, Line 11. It is the result of mutation, selection and recombination of the $m = 1, ..., \mu$ offspring vectors yielding the highest fitness, such that $\mathbf{y}^{(g+1)} = \frac{1}{\mu}\sum_{m=1}^{\mu} \tilde{\mathbf{y}}_{m;\lambda} = \frac{1}{\mu}\sum_{m=1}^{\mu} (\mathbf{y}^{(g)} + \mathbf{x})_{m;\lambda}$. Considering the i-th component, noting that $\mathbf{y}^{(g)}$ is the same for all offspring and setting $(\mathbf{x}_{m;\lambda})_i = x_{m;\lambda}$ one has

$$y_i^{(g+1)} = \frac{1}{\mu}\sum_{m=1}^{\mu} (y_i^{(g)} + x_{m;\lambda}) = y_i^{(g)} + \frac{1}{\mu}\sum_{m=1}^{\mu} x_{m;\lambda}. \tag{12}$$

Taking the expectation $\mathrm{E}\left[y_i^{(g+1)}\right]$, setting $x = \sigma z = \sigma\mathcal{N}(0,1)$ and inserting the expression back into (11) yields

$$\varphi_i = -\frac{1}{\mu}\,\mathrm{E}\left[\sum_{m=1}^{\mu} x_{m;\lambda} \middle| \sigma^{(g)}, \mathbf{y}^{(g)}\right] = -\frac{\sigma}{\mu}\,\mathrm{E}\left[\sum_{m=1}^{\mu} z_{m;\lambda} \middle| \sigma^{(g)}, \mathbf{y}^{(g)}\right]. \tag{13}$$

Therefore progress can be evaluated by averaging over the expectations of μ selected mutation contributions. In principle this task can be solved by deriving the induced order statistic density $p_{m;\lambda}$ for the m-th best individual and subsequently solving the integration over the i-th component

$$\varphi_i = -\frac{1}{\mu}\sum_{m=1}^{\mu} \int_{-\infty}^{\infty} x_i \; p_{m;\lambda}(x_i | \sigma^{(g)}, \mathbf{y}^{(g)})\, \mathrm{d}x_i \,. \tag{14}$$

However, the task of computing expectations of sums of order statistics under noise disturbance has already been discussed and solved by Arnold in [2]. Therefore the problem of Eq. (13) will be reformulated in order to apply the solutions provided by Arnold.

4.2 Expectations of Sums of Noisy Order Statistics

Let z be a random variate with density $p_z(z)$ and zero mean. The density is expanded into a Gram-Charlier series by means of its cumulants κ_i ($i \geq 1$) according to [2, p. 138, D.15]

$$p_z(z) = \frac{1}{\sqrt{2\pi\kappa_2}} \mathrm{e}^{-\frac{z^2}{2\kappa_2}} \left(1 + \frac{\gamma_1}{6} \mathrm{He}_3\left(\frac{z}{\sqrt{\kappa_2}}\right) + \frac{\gamma_2}{24} \mathrm{He}_4\left(\frac{z}{\sqrt{\kappa_2}}\right) + ...\right), \quad (15)$$

with expectation $\kappa_1 = 0$, variance κ_2, skewness $\gamma_1 = \kappa_3/\kappa_2^{3/2}$, excess $\gamma_2 = \kappa_4/\kappa_2^2$ (higher order terms not shown) and He_k denoting the k-th order probabilist's Hermite polynomials. For the problem at hand, see Eq. (13), the mutation variate $z \sim \mathcal{N}(0,1)$ with $\kappa_2 = 1$ and $\kappa_i = 0$ for $i \neq 2$ yielding a standard normal density.

Furthermore, let $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ model additive noise disturbance, such that resulting observed values are $v = z + \epsilon$. Selection of the m-th largest out of λ values yields

$$v_{m;\lambda} = (z + \mathcal{N}(0, \sigma_\epsilon^2))_{m;\lambda}, \quad (16)$$

and the distribution of selected source terms $z_{m;\lambda}$ follows a noisy order statistic with density $p_{m;\lambda}$. Given this definition and a linear relation between $z_{m;\lambda}$ and $v_{m;\lambda}$ the method of Arnold is applicable.

In our case the i-th mutation component $x_{m;\lambda}$ of Eq. (13) is related to selection via the quality change defined in Eq. (3). Maximizing the fitness $F_i(y_i + x_i)$ conforms to maximizing quality $Q_i(x_i)$ with $F_i(y_i)$ being a constant offset.

Aiming at an expression of form (16) and starting with (3), we first isolate component Q_i from the remaining $N-1$ components denoted by $\sum_{j\neq i} Q_j$. Then, approximations are applied to both terms yielding

$$Q_{\mathbf{y}}(\mathbf{x}) = Q_i(x_i) + \sum_{j\neq i} Q_j(x_j) \quad (17)$$

$$\approx -f_i' x_i + \mathcal{N}(E_i, D_i^2), \quad (18)$$

with linearization (6) applied to $Q_i(x_i)$. Additionally, $\sum_{j\neq i} Q_j \simeq \mathcal{N}(E_i, D_i^2)$, as the sum of independent random variables asymptotically approaches a normal distribution in the limit $N \to \infty$ due to the Central Limit Theorem. This is ensured by Lyapunov's condition provided that there are no dominating components within the sum due to largely different values of y_j. The corresponding Rastrigin quality variance $D_i^2 = \mathrm{Var}\big[\sum_{j\neq i} Q_j(x_j)\big]$ is calculated in the Appendix. As the expectation $E_i = \mathrm{E}\big[\sum_{j\neq i} Q_j(x_j)\big]$ is only an offset to $Q_{\mathbf{y}}(\mathbf{x})$ it has no influence on the selection and its calculation can be dropped.

Using $x_i = \sigma z_i$ and $f_i' = \mathrm{sgn}\,(f_i')\,|f_i'|$, expression (18) is reformulated as

$$Q_{\mathbf{y}}(\mathbf{x}) = -\,\mathrm{sgn}\,(f_i')\,|f_i'|\sigma z_i + E_i + \mathcal{N}(0, D_i^2) \quad (19)$$

$$\frac{Q_{\mathbf{y}}(\mathbf{x}) - E_i}{|f_i'|\sigma} = \mathrm{sgn}\,(-f_i')\,z_i + \mathcal{N}\left(0, \frac{D_i^2}{(f_i'\sigma)^2}\right). \quad (20)$$

The decomposition using sign function and absolute value is needed for correct ordering of selected values w.r.t. z_i in (20).

Given result (20), one can define the linearly transformed quality measure $v_i := (Q_{\mathbf{y}}(\mathbf{x}) - E_i)/|f_i'|\sigma$ and noise variance $\sigma_\epsilon^2 := (D_i/f_i'\sigma)^2$, such that the selection of mutation component $\operatorname{sgn}(-f_i')\, z_i$ is disturbed by a noise term due to the remaining $N-1$ components. A relation of the form (16) is obtained up to the sign function.

In [2] Arnold calculated the expected value of arbitrary sums S_P of products of noisy ordered variates containing ν factors per summand

$$S_P = \sum_{\{n_1,\ldots,n_\nu\}} z_{n_1;\lambda}^{p_1} \cdots z_{n_\nu;\lambda}^{p_\nu}, \tag{21}$$

with random variate z introduced in Eqs. (15) and (16). The vector $P = (p_1, ..., p_\nu)$ denotes the positive exponents and distinct summation indices are denoted by the set $\{n_1, ..., n_\nu\}$. The generic result for the expectation of (21) is provided in [2, p. 142, D.28] and was adapted to account for the sign difference between (16) and (20) resulting in possible exchanged ordering. Performing simple substitutions in Arnold's calculations in [2] and recalling that in our case $\gamma_1 = \gamma_2 = 0$, the expected value yields

$$\mathrm{E}[S_P] = \operatorname{sgn}(-f_i')^{\|P\|_1} \sqrt{\kappa_2}^{\|P\|_1} \frac{\mu!}{(\mu-\nu)!} \sum_{n=0}^{\nu} \sum_{k\geq 0} \zeta_{n,0}^{(P)}(k) h_{\mu,\lambda}^{\nu-n,k}. \tag{22}$$

Note that expression (22) deviates from Arnold's formula only in the sign in front of $\sqrt{\kappa_2}$. The coefficients $\zeta_{n,0}^{(P)}(k)$ are defined in terms of a noise coefficient a according to

$$a = \sqrt{\frac{\kappa_2}{\kappa_2 + \sigma_\epsilon^2}} \quad \text{with} \quad \zeta_{n,0}^{(P)}(k) = \text{Polynomial}(a), \tag{23}$$

for which tabulated results are presented in [2, p. 141]. The coefficients $h_{\mu,\lambda}^{i,k}$ are numerically obtainable solving

$$h_{\mu,\lambda}^{i,k} = \frac{\lambda-\mu}{\sqrt{2\pi}} \binom{\lambda}{\mu} \int_{-\infty}^{\infty} \mathrm{He}_k(x)\, \mathrm{e}^{-\frac{1}{2}x^2} [\phi(x)]^i [\Phi(x)]^{\lambda-\mu-1} [1-\Phi(x)]^{\mu-i}\, \mathrm{d}x\,. \tag{24}$$

Now we are in the position to calculate expectation (13) using (22). Since $z \sim \mathcal{N}(0,1)$, it holds $\kappa_2 = 1$. Identifying $P = (1)$, $\|P\|_1 = 1$ and $\nu = 1$ yields

$$\begin{aligned} \mathrm{E}\left[\sum_{m=1}^{\mu} z_{m;\lambda}\right] &= \operatorname{sgn}(-f_i') \frac{\mu!}{(\mu-1)!} \sum_{n=0}^{1} \sum_{k\geq 0} \zeta_{n,0}^{(1)}(k) h_{\mu,\lambda}^{1-n,k} \\ &= \operatorname{sgn}(-f_i')\, \mu \zeta_{0,0}^{(1)}(0) h_{\mu,\lambda}^{1,0} = -\operatorname{sgn}(f_i')\, \mu a c_{\mu/\mu,\lambda}, \end{aligned} \tag{25}$$

with $\zeta_{1,0}^{(1)}(k) = 0$ for any k, and $\zeta_{0,0}^{(1)}(k) \neq 0$ only for $k = 0$ yielding a. The expression $h_{\mu,\lambda}^{1,0}$ is equivalent to the progress coefficient definition $c_{\mu/\mu,\lambda}$ [3, p. 216]. Inserting (25) back into (13), using $a = \sqrt{1/(1+(D_i/f_i'\sigma)^2)} = |f_i'|\sigma/\sqrt{(f_i'\sigma)^2 + D_i^2}$

with the requirement $a > 0$, and noting that $f_i' = \mathrm{sgn}\,(f_i')\,|f_i'|$ one finally obtains for the i-th component first order progress rate

$$\varphi_i(\sigma, \mathbf{y}) = c_{\mu/\mu,\lambda} \frac{f_i'(y_i)\sigma^2}{\sqrt{(f_i'(y_i)\sigma)^2 + D_i^2(\sigma, (\mathbf{y})_{j\neq i})}}. \tag{26}$$

The population dependency is given by progress coefficient $c_{\mu/\mu,\lambda}$. The fitness dependent parameters are contained in f_i', see (7), and in D_i^2 calculated in (33). For better readability the derivative f_i' and variance D_i^2 are not inserted into (26). An exemplary evaluation of D_i^2 as a function of the residual distance R using normalization (10) is shown in the supplementary material in Fig. 4.

4.3 Comparison of Simulation and Approximation

Figure 1 shows an experimentally obtained progress rate compared to the result of (26). Due to large N one exemplary φ_i-graph is shown on the left, and corresponding $i = 1, ..., N$ errors are shown on the right.

The left plot shows the progress rate over a σ-range of $[0, 1]$. This magnitude was chosen in order to study the oscillation, as the frequency $\alpha = 2\pi$. The initial position was chosen randomly to be on the sphere surface $R = 10$.

The red dashed curve uses f_i' as linearization, while the blue dash-dotted curve assumes $f_i' = k_i$ (with $d_i = 0$), see also (7). As f_i' approximates the quality change locally, agreement for the progress is given only for very small mutations σ. For larger σ very large deviation may occur, depending on the local derivative.

The blue curve $\varphi_i(k_i)$ neglects the oscillation ($d_i = 0$) and therefore follows the progress of the quadratic function $f(\mathbf{y}) = \sum_i y_i^2$ for large σ with very good agreement. Due to a *linearized* form of $Q_i(x_i)$ in (6) neither approximation can reproduce the oscillation for moderately large σ.

To verify the approximation quality, the error between (26) and simulation is displayed on the right side of Fig. 1 for all $i = 1, ..., N$. It was done for small $\sigma = 0.1$ and large $\sigma = 1$. The deviations are very similar in magnitude for all i, given randomly chosen y_i. Note that for $\sigma = 1$ the red points show very large errors compared to blue, which was expected.

Figure 2 shows the progress rate φ_i over σ^*, for $i = 2$ as in Fig. 1, with $\mathbf{y}$ randomly on the surface radii $R = \{100, 10, 1, 0.1\}$. Using σ^* the mutation σ is normalized by the residual distance R with spherical normalization (10). Far from the origin with $R = \{100, 10\}$ the quadratic terms are dominating giving better results using $\varphi_i(k_i)$. Reaching $R = 1$ local minima are more relevant and mixed results are obtained with $\varphi_i(f_i')$ better for smaller σ^* and $\varphi_i(k_i)$ for larger σ^*. Within the global attractor $R = 0.1$ the local structure dominates and $\varphi_i(f_i')$ yields better results. These observations will be relevant analyzing the dynamics in Fig. 3 where both approximations show strengths and weaknesses.

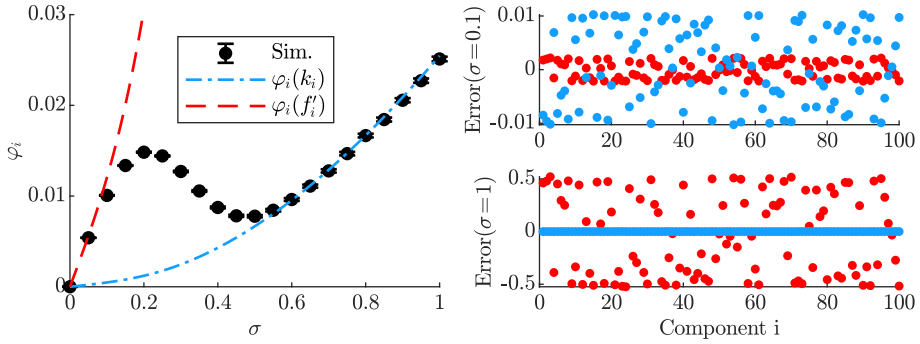

Fig. 1. One-generation experiments with $(150/150, 300)$-ES, $N = 100$, $A = 10$ are performed and quantity (11) is measured averaging over 10^5 runs. Left: φ_i over σ for $i = 2$ at position $y_2 \approx 1.19$, where $\mathbf{y}$ was chosen randomly such that $\|\mathbf{y}\| = R = 10$. Right: error measure $\varphi_i - \varphi_{i,\mathrm{sim}}$ between (26) and simulation for $i = 1, ..., N$ evaluated at $\sigma = \{0.1, 1\}$. The colors are set according to the legend.

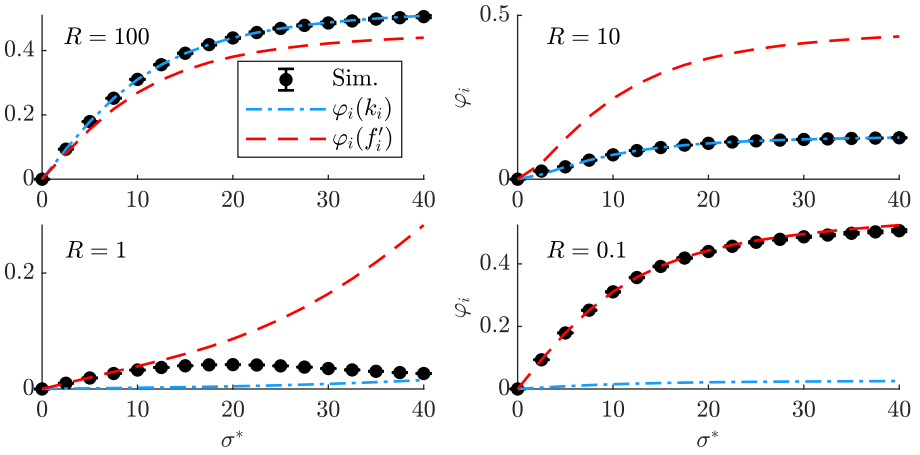

Fig. 2. One-generation progress φ_i $(i = 2)$ over normalized mutation σ^* for $(150/150, 300)$-ES, $N = 100$, $A = 1$ and $R = \{100, 10, 1, 0.1\}$. Simulations are averaged over 10^5 runs. These experiments are preliminary investigations related to the dynamics shown in Fig. 3 with $\sigma^* = 30$. Given a constant σ^* the approximation quality varies over different magnitudes of R.

5 Evolution Dynamics

As we are interested in the dynamical behavior of the ES, averaged real optimization runs from Algorithm 1 will be compared to the iterated dynamics using progress result (26) by applying the dynamical systems approach [3]. Neglecting fluctuations, i.e., $y_i^{(g+1)} = \mathrm{E}\left[y_i^{(g+1)} \middle| \sigma^{(g)}, \mathbf{y}^{(g)}\right]$ the mean value dynamics for the mapping $y_i^{(g)} \to y_i^{(g+1)}$ immediately follows from (11) giving

$$y_i^{(g+1)} = y_i^{(g)} - \varphi_i(\sigma^{(g)}, \mathbf{y}^{(g)}). \tag{27}$$

The control scheme of $\sigma^{(g)}$ was introduced in Eq. (10) and yields simply

$$\sigma^{(g)} = \sigma^* \left\| \mathbf{y}^{(g)} \right\| / N. \tag{28}$$

Equations (27) and (28) describe a deterministic iteration in search space and rescaling of mutations according to the residual distance. For a convergence analysis, we are interested in the dynamics of $R^{(g)} = \left\| \mathbf{y}^{(g)} \right\|$ rather than the actual position values $\mathbf{y}^{(g)}$. Hence in Fig. 3 the $R^{(g)}$-dynamics of the conducted experiments is shown.

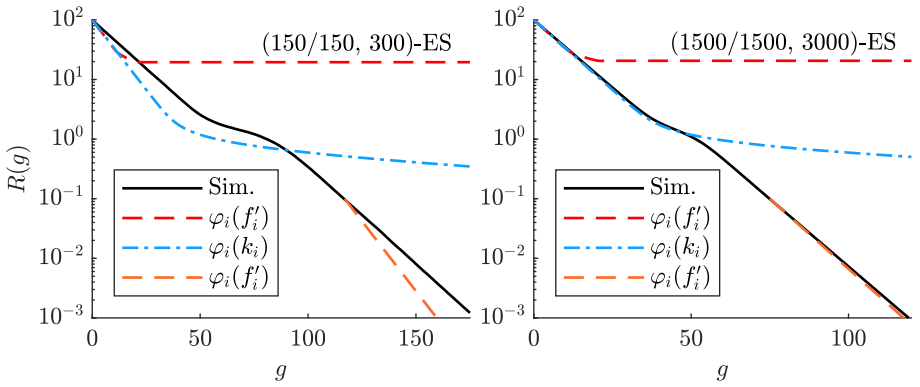

Fig. 3. Comparing average of 100 optimization runs of Algorithm 1 (black, solid) with iterated dynamics from Eq. (27) under constant $\sigma^* = 30$ for $A = 1$ and $N = 100$. Large populations sizes are chosen to ensure global convergence (left: $\mu = 150$; right: $\mu = 1500$; constant $\mu/\lambda = 0.5$). Iteration using progress (26) is performed for both $f_i' = k_i + d_i$ (red/orange dashed) and $f_i'(d_i = 0) = k_i$ (blue dash-dotted) using Equations (27) and (28). The orange dashed iteration was initialized with $R^{(0)} = 0.1$ and translated to the corresponding position of the simulation for easier comparison. The evaluation of quality variance $D_i^2(R)$ is shown in Fig. 4 in the Appendix.

In Fig. 3, all runs of Algorithm 1 exhibit global convergence with the black line showing the average. The left and right plots differ by population size.

Iteration $\varphi_i(k_i)$, blue dash-dotted curve, also converges globally, though very slowly and therefore not shown entirely. The convergence behavior of iteration $\varphi_i(f_i')$, red and orange dashed curves, strongly depends on the initialization and is discussed below.

Three phases can be observed for the simulation. It shows linear convergence at first being followed by a slow-down due to local attractors. Reaching the global attractor the convergence speed increases again. Iteration $\varphi_i(k_i)$ is able to model the first two phases to some degree. Within the global attractor the slope information d_i is missing such that the progress is largely underestimated.

Iteration $\varphi_i(f_i')$ converges first, but yields a stationary state with $R^{st} \approx 20$ when the progress φ_i becomes dominated by derivative term d_i. Starting from $R^{(0)} = 10^2$ the stationary y_i^{st} are either fixed or alternating between coordinates depending on σ, D_i, k_i, and d_i. This effect is due to attraction of local minima and due to the deterministic iteration disregarding fluctuations. It occurs also with varying initial positions. Initialized at $R^{(0)} = 10^{-1}$ orange iteration $\varphi_i(f_i')$ is globally converging.

It turns out that the splitting point of the two approximations in Fig. 3 occurs at a distance R to the global optimizer where the ES approaches the attractor region of the "first" local minima. For the model parameters considered in the experiment this is at about $R \approx 28.2$ – the distance of the farest local minimizer to the global optimizer (obtained by numerical analysis).

Plots in Fig. 3 differ by population size. The convergence speed, i.e. the slopes, show better agreement for large populations, which can be attributed to the fluctuations neglected in (27). Investigations on unimodal funtions Sphere [3] and Ellipsoid [5] have shown that progress is decreased by fluctuations due to a loss-term scaling with $1/\mu$, which agrees with Fig. 3. On the left the iterated progress is faster due to neglected but present fluctuations, while on the right better agreement is observed due to insignificant fluctuations. These observations will be investigated in future research.

6 Summary and Outlook

A first order progress rate φ_i was derived for the $(\mu/\mu_I, \lambda)$-ES by means of noisy order statistics in (26) on the Rastrigin function (1). To this end, the mutation induced variance of the quality change D_i^2 is needed. Starting from (4) a derivation yielding D_i^2 in (33) has been presented in the Appendix. Furthermore, the approximation quality of φ_i was investigated using Rastrigin and quadratic derivatives f_i' and k_i, respectively, by comparing with one-generation experiments.

Linearization f_i' shows good agreement for small-scale mutations, but very large deviations for large mutations. Conversely, linearization k_i yields significantly better results for large mutations as the quadratic fitness term dominates. A progress rate modeling the transition between the regimes is yet to be determined. First numerical investigations of (14) including all terms of (4) indicate

that nonlinear terms are needed for a better progress rate model, which is an open challenge and part of future research.

The obtained progress rate was used to investigate the dynamics by iterating (27) using (28) and comparing with ES runs. Iteration via f_i' only converges globally if initialized close to the optimizer, since local attraction is strongly dominating. Dynamics via k_i converges globally independent of initialization, but the observed rate matches only for the initial phase and for very large populations. This confirms the need for a higher order progress rate modeling the effect of fluctuations, especially when function evaluations are expensive and small populations must be used. Additionally, an advanced progress rate formula is needed combining effects of global and local attraction to model all three phases of the dynamics correctly.

The investigations done so far are a first step towards a full dynamical analysis of the ES on the multimodal Rastrigin function. Future investigations must also include the complete dynamical modeling of the mutation strength control. One aim is the tuning of mutation control parameters such that the global convergence probability is increased while still maintaining search efficiency. Our final goal will be the theoretical analysis of the full evolutionary process yielding also recommendations regarding the choice of the minimal population size needed to converge to the global optimizer with high probability.

Acknowledgments This work was supported by the Austrian Science Fund (FWF) under grant P33702-N. Special thanks goes to Lisa Schönenberger for providing valuable feedback and helpful discussions.

Appendix

In the Appendix, the derivation of the quality variance $D_i^2 = \sum_{j\neq i} \mathrm{Var}\big[Q_j(x_j)\big]$ introduced in Eq. (18) is explained for random mutations $x \sim \mathcal{N}(0,\sigma^2)$. Only the main steps are presented in a concise manner.

A single component j can be evaluated and summed over due to mutual independence. Noting that $\mathrm{E}\left[x^2\right] = \sigma^2$, $\mathrm{E}\left[x^4\right] = 3\sigma^4$, $\mathrm{E}\left[x^k\right] = 0$ for odd k, and $\mathrm{Var}\left[(\cdot)\right] = \mathrm{E}\left[(\cdot)^2\right] - \mathrm{E}\left[(\cdot)\right]^2$, the variance of Q_j using Eq. (4) is evaluated as

$$\begin{aligned}\mathrm{Var}\left[Q_j\right] &= \mathrm{E}\left[Q_j^2\right] - \mathrm{E}\left[Q_j\right]^2 \\ &= 2\sigma^4 + 4y_j^2\sigma^2 + A^2\sin^2(\alpha y_j)\,\mathrm{Var}\left[\sin(\alpha x_j)\right] \\ &\quad + A^2\cos^2(\alpha y_j)\,\mathrm{Var}\left[\cos(\alpha x_j)\right] - 2A\cos(\alpha y_j)\,\mathrm{E}\left[x^2\cos(\alpha x_j)\right] \\ &\quad + 2A\sigma^2\cos(\alpha y_j)\,\mathrm{E}\left[\cos(\alpha x_j)\right] + 4Ay_j\sin(\alpha y_j)\,\mathrm{E}\left[x\sin(\alpha x_j)\right].\end{aligned} \tag{29}$$

Obtaining (29) it was used that for $x \sim \mathcal{N}(0, \sigma^2)$ we have $\mathrm{E}\left[x^k \sin(\alpha x)\right] = 0$ for even k and $\mathrm{E}\left[x^k \cos(\alpha x)\right] = 0$ for odd k, which is due to odd sine and even cosine function, respectively.

In the general case, expectations of the form $\mathrm{E}\left[x^k \cos \alpha x\right]$ and $\mathrm{E}\left[x^k \sin \alpha x\right]$ for $k \geq 0$ can be obtained by using the definition of the characteristic function χ of $x \sim \mathcal{N}(\mu, \sigma^2)$ and its known result [1]

$$\chi_x(\alpha) = \mathrm{E}\left[\mathrm{e}^{i\alpha x}\right] = \mathrm{e}^{i\alpha\mu - \frac{1}{2}\alpha^2\sigma^2} = \mathrm{e}^{-\frac{1}{2}\alpha^2\sigma^2}[\cos(\alpha\mu) + i\sin(\alpha\mu)]. \tag{30}$$

Then, the k-th derivatives with respect to α can be applied to both sides

$$\begin{aligned} \frac{\mathrm{d}^k}{\mathrm{d}\alpha^k}\mathrm{E}\left[\mathrm{e}^{i\alpha x}\right] = \mathrm{E}\left[\frac{\mathrm{d}^k}{\mathrm{d}\alpha^k}\mathrm{e}^{i\alpha x}\right] &= \mathrm{E}\left[\frac{\mathrm{d}^k}{\mathrm{d}\alpha^k}\cos(\alpha x)\right] + i\,\mathrm{E}\left[\frac{\mathrm{d}^k}{\mathrm{d}\alpha^k}\sin(\alpha x)\right] \\ &\stackrel{!}{=} \frac{\mathrm{d}^k}{\mathrm{d}\alpha^k}\left[\mathrm{e}^{-\frac{(\alpha\sigma)^2}{2}}[\cos(\alpha\mu) + i\sin(\alpha\mu)]\right], \end{aligned} \tag{31}$$

such that corresponding real and imaginary parts can be identified. Given $\mu = 0$ for $k = \{0, 1, 2\}$ the required quantities of (29) can be derived. Additionally, trigonometric identities $\cos^2(x) = 1/2 + \cos(2x)/2$ and $\sin^2(x) = 1/2 - \cos(2x)/2$ are used. The results are

$$\begin{aligned} \mathrm{E}\left[\cos(\alpha x)\right] &= \mathrm{e}^{-\frac{(\alpha\sigma)^2}{2}}, \quad \mathrm{E}\left[\cos^2(\alpha x)\right] = \frac{1}{2} + \frac{1}{2}\mathrm{e}^{-\frac{(2\alpha\sigma)^2}{2}} \\ \mathrm{E}\left[\sin^2(\alpha x)\right] &= \frac{1}{2} - \frac{1}{2}\mathrm{e}^{-\frac{(2\alpha\sigma)^2}{2}}, \quad \mathrm{E}\left[x\sin(\alpha x)\right] = \alpha\sigma^2\mathrm{e}^{-\frac{(\alpha\sigma)^2}{2}} \\ \mathrm{E}\left[x^2\cos(\alpha x)\right] &= (\sigma^2 - \alpha^2\sigma^4)\mathrm{e}^{-\frac{(\alpha\sigma)^2}{2}}, \quad \mathrm{Var}\left[(\cdot)\right] = \mathrm{E}\left[(\cdot)^2\right] - \mathrm{E}\left[(\cdot)\right]^2. \end{aligned} \tag{32}$$

Inserting relations (32) into (29), summing over $N-1$ components and collecting the resulting terms the Rastrigin quality variance is obtained

$$\begin{aligned} D_i^2 = \sum_{j\neq i} 2\sigma^4 + 4y_j^2\sigma^2 &+ \frac{A^2}{2}\left[1 - \mathrm{e}^{-(\alpha\sigma)^2}\right] \\ &+ \frac{A^2}{2}\mathrm{e}^{-(\alpha\sigma)^2}\cos(2\alpha y_j)\left[\mathrm{e}^{-(\alpha\sigma)^2} - 1\right] \\ &+ 2A\alpha\sigma^2\mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}\Big[\alpha\sigma^2\cos(\alpha y_j) + 2y_j\sin(\alpha y_j)\Big]. \end{aligned} \tag{33}$$

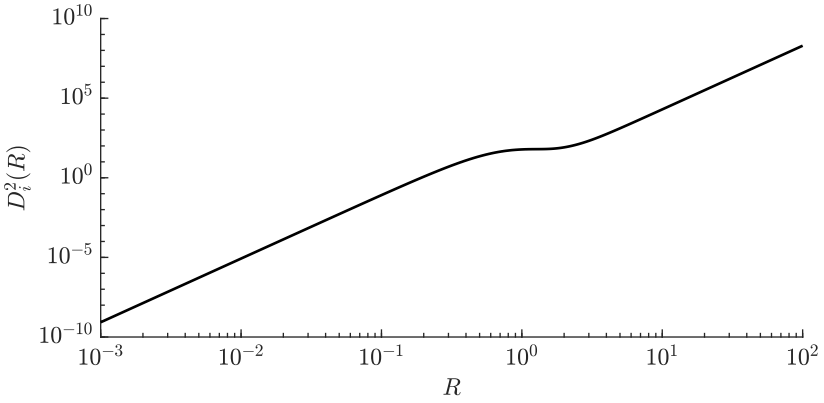

Fig. 4. Quality variance of (33) as a function of R for $A = 1$, $\sigma^* = 30$ and $N = 100$ (excluding the i-th component) with σ obtained using normalization (10). The parameters were chosen according to the dynamic experiments in Fig. 3 and $D_i^2(R)$ shows the variance over different magnitudes of R. Positions $\mathbf{y}$ were randomly initialized with $\|\mathbf{y}\| = R$. Variance fluctuations due to different cartesian realizations of the same R are negligible for large N. Similar to Fig. 3, a transition region can be observed.

References

1. Abramowitz, M., Stegun, I.A.: Pocketbook of Mathematical Functions. Verlag Harri Deutsch, Thun (1984)
2. Arnold, D.: Noisy Optimization with Evolution Strategies. Kluwer Academic Publishers, Dordrecht (2002)
3. Beyer, H.G.: The Theory of Evolution Strategies. Natural Computing Series, Springer, Heidelberg (2001), DOI: 10.1007/978-3-662-04378-3
4. Beyer, H.G.: Convergence Analysis of Evolutionary Algorithms That are Based on the Paradigm of Information Geometry. Evolutionary Computation **22**(4), 679–709 (2014), DOI: 10.1162/EVCO_a_00132
5. Beyer, H.G., Melkozerov, A.: The Dynamics of Self-Adaptive Multi-Recombinant Evolution Strategies on the General Ellipsoid Model. IEEE Transactions on Evolutionary Computation **18**(5), 764–778 (2014), DOI: 10.1109/TEVC.2013.2283968
6. Beyer, H.G., Sendhoff, B.: Simplify Your Covariance Matrix Adaptation Evolution Strategy. IEEE Transactions on Evolutionary Computation **21**(5), 746–759 (2017), DOI: 10.1109/TEVC.2017.2680320
7. Glasmachers, T., Schaul, T., Sun, Y., Wierstra, D., Schmidhuber, J.: Exponential Natural Evolution Strategies. In: Branke et al., J. (ed.) GECCO'10: Proceedings of the Genetic and Evolutionary Computation Conference. pp. 393–400. ACM, New York (2010)
8. Hansen, N., Kern, S.: Evaluating the CMA Evolution Strategy on Multimodal Test Functions. In: Yao et al., X. (ed.) Parallel Problem Solving from Nature 8. pp. 282–291. Springer, Berlin (2004)
9. Hansen, N., Müller, S., Koumoutsakos, P.: Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES). Evolutionary Computation **11**(1), 1–18 (2003)

10. Mobahi, H., Fisher, J.: A Theoretical Analysis of Optimization by Gaussian Continuation. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. pp. 1205–1211. AAAI Press (2015)
11. Müller, N., Glasmachers, T.: Non-local optimization: imposing structure on optimization problems by relaxation. In: Foundations of Genetic Algorithms, 16. pp. 1–10. ACM (2021). https://doi.org/10.1145/3450218.3477307, https://doi.org/10.1145/3450218.3477307
12. Ollivier, Y., Arnold, L., Auger, A., Hansen, N.: Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. J. Mach. Learn. Res. **18**(18), 1–65 (2017)
13. Rechenberg, I.: Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution. Frommann-Holzboog Verlag, Stuttgart (1973)
14. Schwefel, H.P.: Numerical Optimization of Computer Models. Wiley, Chichester (1981)
15. Zhang, J., Bi, S., Zhang, G.: A directional Gaussian smoothing optimization method for computational inverse design in nanophotonics. Materials & Design **197**, 109213 (2021). https://doi.org/10.1016/j.matdes.2020.109213