Success Rate of Evolution Strategies on the Multimodal Griewank Function

Lisa Schönenberger and Hans-Georg Beyer
Vorarlberg University of Applied Sciences
Research Center Business Informatics
6850 Dornbirn, Austria
{lisa.schoenenberger, hans-georg.beyer}@fhv.at

***Abstract*—The Griewank function is one of the widely used multimodal benchmark functions. The function is known for its counter intuitive behavior of getting simpler to be optimized with increasing dimension, although the number of local minima increases with the problem dimension. A frozen noise model is introduced that is able to partially explain the empirically observed behaviors. The influence of the different Evolution Strategies and their parameters on the success rate are analyzed. Empirical investigations are used to show the limitations of this model. These investigations reveal some unexpected behaviors regarding the influence of the population size on the success rate of the Evolution Strategies that cannot be explained by the current theory.**

***Index Terms*—Evolution Strategies, Griewank function, global optimization, multi-modal objective function, global convergence**

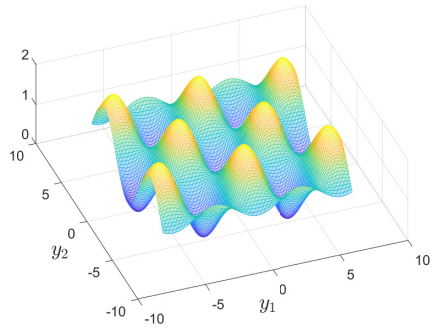

Fig. 1. Griewank function with standard parameters for $N = 2$.

I. Introduction

The multimodal Griewank function was first introduced in [1] and used as a benchmark problem in nonlinear optimization. Although it is meanwhile used in many benchmark suites, the analytical analysis of Griewank is not yet very advanced. This is generally the case for multimodal test functions. Progress rate analysis has only been done for the Rastrigin function [2], [3]. For the Griewank function such analyses are still pending. Due to the complex structure of the Griewank function it is also not clear whether a progress rate analysis is tractable.

There is only a limited amount of literature available for the Griewank function. In [4] it was shown that the number of minima grows exponentially with the number of dimensions, but that the function gets easier to be optimized for larger N considering classical nonlinear optimization strategies based on gradient descent. The number of minima of the Griewank function up to dimension 3 was estimated in [5]. In [6], it was shown for a certain algorithm that the Griewank function first becomes more difficult and then becomes easier to optimize with increasing dimension. In a deeper analysis, the two-scale structure of the Griewank function was investigated, i.e., the decomposition into a global sphere part with a unique global minimum and an oscillatory non-convex function with an infinite number of local minima. Such a decomposition was also done for the Rastrigin function in [7]. By interpreting the oscillatory part as frozen noise, a population sizing model was derived that allows for estimating the population size needed to achieve a given success probability. One of the main questions of this paper is to find out whether this model can be applied to the Griewank function.

The remaining sections are organized as follows: First, a detailed description of the Griewank function is given. After a short introduction of the ES algorithms, their behavior in the Griewank landscape is described. In Section IV, the applicability of the frozen noise model to the Griewank function is investigated. In Sections V and VI, the success probability and the scaling behavior of the population size are experimentally investigated. Finally, a summary of the results and an outlook on future research are given in a concluding section.

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

II. The Griewank Test Function

The Griewank test function F for an N-dimensional search vector $\mathbf{y} = (y_1, \ldots, y_N)$ is defined by

$$F(\mathbf{y}) = \frac{1}{A}\sum_{i=1}^{N} y_i^2 - \prod_{i=1}^{N} \cos(b_i y_i) + 1. \tag{1}$$

The standard values for the parameters A and b_i are $A = 4000$ and $b_i = 1/\sqrt{i}$ for $i = 1, \ldots, N$. These parameters are used for all experiments unless otherwise specified. Beside the standard case, the special case where $b_i = 1$ will also be investigated. Figure 1 shows the Griewank function with standard parameters for $N = 2$.

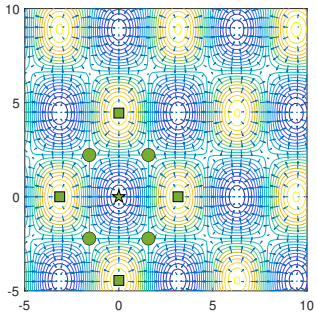

Fig. 2. Global attractor region of the Griewank function for $N = 2$. The star shows the global optimizer, circles the nearest stationary points and squares the nearest stationary points in coordinate direction, cf. (6). Arrows show the negative gradient flow.

The number of local minima of the Griewank function is finite and grows exponentially with the dimension [4]. The global optimizer of the Griewank function is located at $\mathbf{0}$. The region around the global optimizer, where the negative gradients converge toward the optimizer, is defined as the *global attractor*. It is bracketed by stationary points the location of which is obtained by the zeros of the partial derivative of the Griewank function in direction i

$$\frac{\partial F}{\partial y_i} = \frac{2y_i}{A} + \sin(b_i y_i)\, b_i \prod_{j\neq i} \cos(b_j y_j) = 0$$

$$\Leftrightarrow \frac{2y_i}{Ac_i b_i} = -\sin(b_i y_i) \tag{2}$$

where $c_i = \prod_{j\neq i} \cos(b_j y_j)$ depends on the remaining components. Assuming that $2/(Ac_i b_i)$ is sufficiently small, there is an intersection between the linear function $2y_i/(Ac_i b_i)$ and the sine function $-\sin(b_i y_i)$. The intersection can be approximated by the Taylor expansion of $-\sin(b_i y_i)$ at $x_0 = \pi/b_i$. It follows

$$\begin{aligned} -\sin(b_i y_i) &= -\sin(b_i x_0) - b_i \cos(b_i x_0)(y_i - x_0) + \mathcal{O}\left(y_i^2\right) \\ &= -\sin(\pi) - b_i \cos(\pi)\left(y_i - \frac{\pi}{b_i}\right) + \mathcal{O}\left(y_i^2\right) \\ &= b_i y_i - \pi + \mathcal{O}\left(y_i^2\right). \end{aligned} \tag{3}$$

Therefore, one gets

$$\begin{aligned} y_i b_i - \pi &\simeq \frac{2y_i}{Ac_i b_i} \\ \Leftrightarrow \pi &\simeq \left(b_i - \frac{2}{Ac_i b_i}\right) y_i = \frac{Ac_i b_i^2 - 2}{Ac_i b_i} y_i \end{aligned} \tag{4}$$

and finally

$$y_i \simeq \frac{Ac_i b_i \pi}{Ac_i b_i^2 - 2} \quad \text{or} \quad y_i = 0 \quad \text{for all} \quad i = 1, \ldots, N \tag{5}$$

as a condition that $\mathbf{y}$ is a stationary point. Due to the c_i this nonlinear system consisting of N variables can only be solved numerically. Only the stationary points on the coordinate axis can be determined analytically.

Denote the distance between the global optimizer and the nearest stationary point in coordinate direction i with $\Delta_0(i)$.

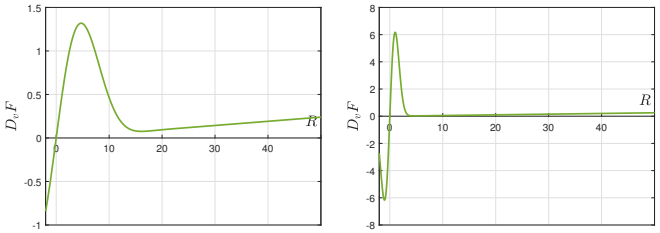

Fig. 3. Directional derivative of the Griewank function along vector $\mathbf{v}$ depending on distance R for $N = 100$, $b_i = 1/\sqrt{i}$ (left plot) and $b_i = 1$ (right plot). All components of $\mathbf{v}$ were chosen randomly.

Since $y_j = 0$ for $j \neq i$ it holds $c_i = \prod_{j\neq i} \cos\left(y_j/\sqrt{j}\right) = 1$. Hence, the nearest stationary point in coordinate direction i is

$$\Delta_0(i) \simeq \frac{Ab_i\pi}{Ab_i^2 - 2}. \tag{6}$$

For the standard Griewank landscape with $b_i = 1/\sqrt{i}$, $\Delta_0(i)$ grows with i. For each additional dimension the distance to the nearest stationary point in coordinate direction becomes larger than that of the previous dimensions. Therefore, the attractor region is not symmetric. If i is very large, then there is no intersection between the linear function and the sine function in (2). In this case the attractor region is unbounded in this coordinate direction. For $A = 4000$, it can be calculated numerically that this is true for dimensions larger than 434. For the special Griewank landscape with $b_i = 1$, the attractor region is symmetric. There are stationary points in each coordinate direction, and the number of local minima is larger than in the standard case.

Figure 2 shows the attractor region for standard parameters and $N = 2$. The nearest stationary points were determined numerically. It is not obvious how the attractor region is structured for larger dimensions. In Section V, an inner bound of the attractor region is estimated using gradient strategies. Fig. 3 shows that the attractor region is unbounded for $N = 100$. For both the standard and the special Griewank landscape there are directions in which the function increases monotonically. It is barely visible in the right plot, but there is no intersection with the R-axis.

III. ES-Algorithms

For the two basic $(\mu/\mu_I, \lambda)$-ES algorithms considered here, the control of the mutation strength σ is done either by σ-self-adaptation (σSA), see Alg. 1, or cumulative step-size adaptation (CSA), see Alg. 2. The population consists of μ parents and λ offspring with truncation ratio $\vartheta = \mu/\lambda$. The subscript "$m;\lambda$" indicates that the $m = 1, \ldots, \mu$ best individuals were selected from λ parents. For the σSA-ES, offspring mutation strengths are chosen from a log-normal distribution with learning parameter τ. The standard choice for the learning parameter, which ensures the best possible performance on the sphere model, is $\tau = 1/\sqrt{2N}$ [8]. A smaller τ results in slower adaptation. In [7] it was shown that a smaller adaptation speed leads to a higher success rate

Algorithm 1 The $(\mu/\mu_I, \lambda)$-σSA Evolution Strategy

Initialize $\left(\mathbf{y}^{(0)}, \sigma^{(0)}, \sigma_{\text{stop}}, g = 0\right)$
repeat
  for $l = 1$ **to** λ **do**
    $\tilde{\sigma}_l = \sigma^{(g)} \mathrm{e}^{\tau \mathcal{N}(0,1)}$ {mutate parental σ}
    $\tilde{\mathbf{y}}_l = \mathbf{y}^{(g)} + \tilde{\sigma}_l \left(\mathcal{N}(0,1), \ldots, \mathcal{N}(0,1)\right)$ {mutate $\mathbf{y}$}
    $\tilde{F}_l = F\left(\tilde{\mathbf{y}}_l\right)$ {evaluate offspring}
  end for
  Sort Individuals $\tilde{\mathbf{y}}$ Ascendingly w.r.t. Fitness $\tilde{F}$
  $g = g + 1$
  $\mathbf{y}^{(g)} = \frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{\mathbf{y}}_{m;\lambda}$ {recombine the μ best $\tilde{\mathbf{y}}$}
  $\sigma^{(g)} = \frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{\sigma}_{m;\lambda}$ {recombine the μ best $\tilde{\sigma}$}
until $\sigma^{(g)} < \sigma_{\text{stop}}$

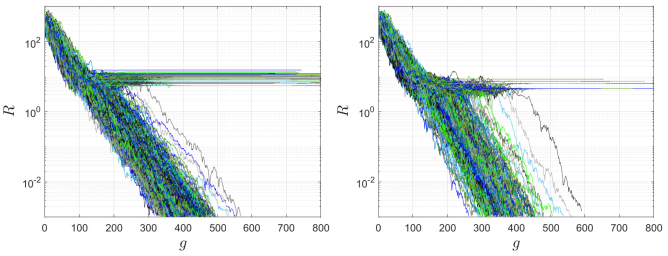

Fig. 4. Residual distance dynamics for 1000 $(3/3_I, 6)$-CSA-ES (Alg. 2) runs with $c = 1/\sqrt{N}$ and $N = 10$. The left plot represents the Griewank landscape with $b_i = 1/\sqrt{i}$ and the right plot represents the Griewank landscape with $b_i = 1$. The success probabilities are $P_{\text{s}} = 0.38$ for the standard case $b_i = 1/\sqrt{i}$ and $P_{\text{s}} = 0.21$ for the special case $b_i = 1$, respectively.

in the Rastrigin landscape. Therefore, a smaller adaptation rate of $\tau = 1/\sqrt{8N}$ is also considered. For the CSA-ES, the standard choice for the cumulation time parameter c is $1/N$ or $1/\sqrt{N}$ and $D = 1/c$ for the damping constant D [9], [10]. The influence of the learning rate and the cumulation time parameter on the success probability in the Griewank landscape is investigated in Section V.

Algorithm 2 The $(\mu/\mu_I, \lambda)$-CSA Evolution Strategy

Initialize $\left(\mathbf{y}^{(0)}, \sigma^{(0)}, \sigma_{\text{stop}}, \mathbf{s} = \mathbf{1}, g = 0\right)$
repeat
  for $l = 1$ **to** λ **do**
    $\tilde{\mathbf{z}}_l = \left(\mathcal{N}(0,1), \ldots, \mathcal{N}(0,1)\right)$ {search direction}
    $\tilde{\mathbf{y}}_l = \mathbf{y}^{(g)} + \sigma^{(g)} \tilde{\mathbf{z}}_l$ {mutate $\mathbf{y}$}
    $\tilde{F}_l = F\left(\tilde{\mathbf{y}}_l\right)$ {evaluate offspring}
  end for
  Sort Individuals $\tilde{\mathbf{y}}$ Ascendingly w.r.t. Fitness $\tilde{F}$
  $g = g + 1$
  $\mathbf{y}^{(g)} = \frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{\mathbf{y}}_{m;\lambda}$ {recombine the μ best $\tilde{\mathbf{y}}$}
  $\mathbf{z}^{(g)} = \frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{\mathbf{z}}_{m;\lambda}$ {recombine the μ best $\tilde{\mathbf{z}}$}
  $\mathbf{s} = (1 - c)\mathbf{s} + \sqrt{\mu c\,(2 - c)}\, \mathbf{z}^{(g)}$ {update s-path}
  $\sigma^{(g)} = \sigma^{(g-1)} \exp\left(\frac{\|\mathbf{s}\|^2 - N}{2DN}\right)$ {update σ, [10, p.13]}
until $\sigma^{(g)} < \sigma_{\text{stop}}$

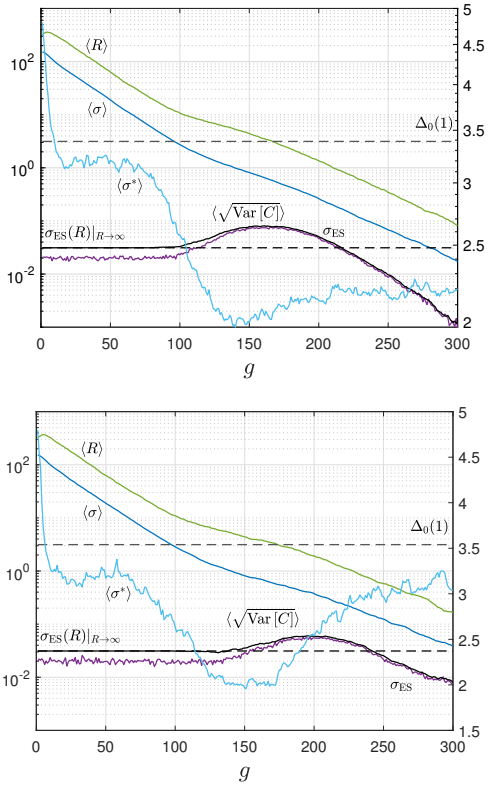

Fig. 5. Mean value dynamics of the $(3/3_I, 6)$-CSA-ES runs derived from the successful runs of Fig. 4. The top plot represents the Griewank landscape with $b_i = 1/\sqrt{i}$ and the bottom plot with $b_i = 1$, respectively. The green line shows the distance to the global optimizer, the blue line the mutation strength σ and the light blue line the normalized mutation strength σ^* (right y-axis). The purple line shows the standard deviation of the cosine part. Its theoretical estimate (13) is represented by the black solid line whereas the lower black dashed line shows its asymptotic behavior, cf. (14). The upper dashed black line represents (6) the distance to the nearest stationary point in coordinate direction 1.

Figure 4 shows the dynamics of the distance $R = \|\mathbf{y}\|$ of the parental centroid to the global optimizer. 1000 independent $(3/3_I, 6)$-CSA-ES runs with $c = 1/\sqrt{N}$ for $N = 10$ were performed to obtain the results. The left plot represents the results for the Griewank landscape with $b_i = 1/\sqrt{i}$ and the right plot for $b_i = 1$, respectively. Fig. 5 shows the mean value dynamics of the *successful runs* from Fig. 4. The light blue line represents the normalized mutation strength $\sigma^* := \sigma N/R$. For better visibility it is displayed on the right y-axis. A closer look at the dynamics shows that the ES goes through three phases. In the first phase, the parental centroid is located far away from the global optimizer and the perturbations caused by the cosine part are relatively small compared to the sphere model part. In this case, the behavior of the ES on Griewank is similar to that of a sphere model. This can be seen in Fig. 5, where the initial mutation strength was chosen intentionally too large to demonstrate, that the CSA is able to adapt σ to the steady-state value. As the ES gets closer to the global optimizer, the influence of the local attractors becomes more dominant. The ES reduces the speed at which the global optimizer is approached. This is visible in the mean value dynamics of

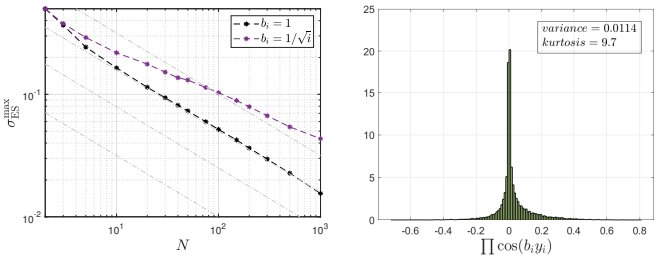

Fig. 6. Left plot: Maximal value of σ_{ES} experimentally determined. Gray dashed-dotted lines show functions $\propto 1/\sqrt{N}$ for comparison. Right plot: Histogram of cosine parts at distance $R = 3.14$ to the global optimizer for the $(10/10_I, 20)$-CSA-ES with $c = 1/\sqrt{N}$ and $N = 100$.

Fig. 5. In the third phase, the ES either terminates in a local minimum and fails or it reaches the global attractor and ends successfully. In this phase, there are differences between the standard and the special Griewank landscape. For the special case with $b_i = 1$, the normalized mutation strength increases and then returns to its steady-state value. For the standard case with $b_i = 1/\sqrt{i}$, the normalized mutation strength remains at the lower level.

These 3 phases have already been observed at the Rastrigin function in [7] where the frozen noise model was proposed to analyze this behavior. This model was used in turn to estimate the population size required to achieve a given success rate. Whether the application of such a model is also possible for Griewank will be analyzed in the following section.

IV. The frozen noise model

The Griewank function can be divided into two parts. The first part $1/A \cdot \sum_{i=1}^{N} y_i^2$ describes a simple sphere model with a unique global minimum. The second part

$$C(\mathbf{y}) = 1 - \prod_{i=1}^{N} \cos(b_i y_i) \tag{7}$$

is called the *cosine part*. It consists of an infinite number of local minima and its codomain is in the interval $[0, 2]$. The perturbations caused by the cosine part when sampled by the ES mutations can be interpreted as frozen noise, resulting in a noisy sphere model of the form

$$F(\mathbf{y}) = \frac{R^2}{A} + \sigma_{\mathrm{ES}}(R)\mathcal{X} \quad \text{with} \quad R = \|\mathbf{y}\|, \tag{8}$$

where σ_{ES} is the noise strength and $\mathcal{X}$ describes the distribution of the noise. The noise strength $\sigma_{\mathrm{ES}}(R)$ is given by the standard deviation of the cosine part of the offspring population before selection. For a given parent $\mathbf{y} = (y_1, \ldots, y_N)^T$ and mutation strength σ, the standard deviation of the cosine parts for the CSA-ES is $\sigma_{\mathrm{ES}} = \sqrt{\mathrm{Var}\left[C(\mathbf{y}, \sigma)\right]}$

$$\sigma_{\mathrm{ES}} = \sqrt{\mathrm{Var}\left[\prod_{i=1}^{N} \cos\left(b_i \left(y_i + \sigma z_i\right)\right)\right]}, \tag{9}$$

where z_i are independent $\mathcal{N}(0, 1)$ distributed random variables. It was shown in [7] that

$$\mathrm{E}\left[\cos\left(b_i \tilde{y}_i\right)^2\right] = \frac{1}{2} + \frac{1}{2}\mathrm{e}^{-2b_i^2\sigma^2}\cos\left(2b_i y_i\right) \tag{10}$$

$$\mathrm{E}\left[\cos\left(b_i \tilde{y}_i\right)\right]^2 = \frac{1}{2}\mathrm{e}^{-b_i^2\sigma^2} + \frac{1}{2}\mathrm{e}^{-b_i^2\sigma^2}\cos\left(2b_i y_i\right). \tag{11}$$

where $\tilde{y}_i = y_i + \sigma z_i$. Using the identity $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$ and the independence of the components z_i, it follows

$$\begin{aligned}\mathrm{Var}\left[C(\mathbf{y}, \sigma)\right] =& \mathrm{E}\left[\prod_{i=1}^{N}\cos\left(b_i\tilde{y}_i\right)^2\right] - \mathrm{E}\left[\prod_{i=1}^{N}\cos\left(b_i\tilde{y}_i\right)\right]^2 \\ =& \prod_{i=1}^{N}\mathrm{E}\left[\cos\left(b_i\tilde{y}_i\right)^2\right] - \prod_{i=1}^{N}\mathrm{E}\left[\cos\left(b_i\tilde{y}_i\right)\right]^2 \end{aligned} \tag{12}$$

and finally, with Eqs. (10) and (11)

$$\begin{aligned}\mathrm{Var}\left[C(\mathbf{y}, \sigma)\right] =& \frac{1}{2^N}\prod_{i=1}^{N}\left(1 + \mathrm{e}^{-2b_i^2\sigma^2}\cos\left(2b_i y_i\right)\right) \\ &- \frac{1}{2^N}\prod_{i=1}^{N}\mathrm{e}^{-b_i^2\sigma^2}\left(1 + \cos\left(2b_i y_i\right)\right). \end{aligned} \tag{13}$$

The standard deviation $\sigma_{\mathrm{ES}}(\mathbf{y}, \sigma) = \sqrt{\mathrm{Var}\left[C(\mathbf{y}, \sigma)\right]}$ is shown in Fig. 5 both experimentally (purple line) and by the theoretical estimate (black line). Eq. (13) depends on the current location of $\mathbf{y}$. An equation for $\sigma_{\mathrm{ES}}(R)$ that depends only on the distance R to the global optimizer can be derived by taking the expected value w.r.t. y_i. It will be demonstrated at the end of this section that in general the distributions of the individual y_i-components differ (Fig. 7). This complicates the derivation of an R-dependent equation. Nevertheless, it is possible to derive an equation for the asymptotic behavior of $\sigma_{\mathrm{ES}}(R)$. When an ES has reached its steady-state behavior, the normalized mutation strength σ^* is roughly constant, and a large value of R results in a large value of σ. The expression in the first parenthesis of (13) approaches 1 for large σ and the second product vanishes. Therefore, for a large residual distance R and sufficiently large σ, it holds

$$\sigma_{\mathrm{ES}}(R)|_{R\to\infty} = \sqrt{\frac{1}{2^N}}. \tag{14}$$

This is represented by the lower black dashed line in Fig. 5. The standard deviation of the cosine part decreases with increasing dimension. Therefore, the influence of the cosine parts is expected to decrease with increasing dimension. Thus, the success rate is expected to increase with increasing N. This already provides an explanation for the (at first glance) astonishing observation that minimizing Griewank becomes simpler with increasing search space dimensionality. Note that (14) does not represent the maximum value of σ_{ES}. This is visible in Fig. 5. It is also visible that the maximum standard deviation occurs in the region where the ES reaches the distance $\Delta_0(i)$, cf. (6). This explains why the ES can locally converge due to the high noise level experienced. The maximum values of σ_{ES} are shown in the left plot of Fig. 6.

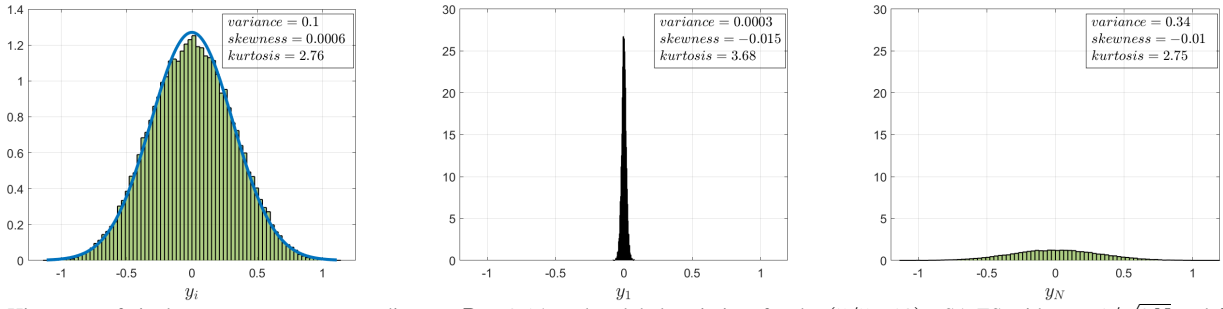

Fig. 7. Histogram of single parent components at distance $R = 3.14$ to the global optimizer for the $(5/5_I, 10)$-σSA-ES with $\tau = 1/\sqrt{2N}$ and $N = 100$. Left plot shows the histogram for all components with $b_i = 1$, middle plot for the first component with $b_i = 1/\sqrt{i}$, and the right plot for the Nth component with $b_i = 1/\sqrt{i}$. The blue line in the left figure is the pdf of the $\mathcal{N}(0, R^2/N)$ variate.

The values have been experimentally determined. The values decrease slower than $\mathcal{O}(1/\sqrt{2N})$, but rather with $\mathcal{O}(1/\sqrt{N})$.

In the Rastrigin model [7], the noise is normally distributed. This is not the case for the Griewank noise model (8), where the cosine part is a product of independent random variables. The right plot in Fig. 6 shows a histogram of the cosine part at the residual distance $R = \Delta_0(1) = 3.14$, cf. (6). Also in [7], the individual y_i-components were $\mathcal{N}(0, R^2/N)$ distributed. This is also true for the special case with $b_i = 1$, as shown in Fig. 7. For the standard case with $b_i = 1/\sqrt{i}$, the distribution is different for each component. The variance becomes larger for smaller b_i.

In summary, the noise is not normally distributed and for $b_i = 1/\sqrt{i}$ the variances of the single y_i-components are different. In addition, the attractor region is unbounded and cannot be described in terms of closed analytical expressions. Therefore, the frozen noise model is not suited for quantitatively estimating the population size of the Griewank function. Only the size of σ_{ES} can be estimated well. σ_{ES} becomes smaller with increasing N. Therefore, it is expected that the success probability P_{s} increases with increasing N. In the following, the success probability and the population size will be estimated experimentally.

V. Success Probability

In Fig. 8, the success probability for the Griewank function with $b_i = 1$ was determined experimentally with at least 1000 runs. For small $N \leq 5$ and large μ, data for the σSA-ES are not available. For these values, R oscillates while σ remains approximately constant [1]. For the available data, there are no significant differences between the different strategies. Fig. 9 represents the results for the Griewank function with $b_i = 1/\sqrt{i}$. For a better overview, only the values for the CSA-ES are shown. Again, the differences between the different strategies are much smaller than in other multimodal landscapes, such as the Rastrigin landscape [7].

For $b_i = 1$ and $N \geq 5$ the success probability increases with increasing N, and the success probability is constant 1 for $N \geq 100$. For $b_i = 1/\sqrt{i}$, the behavior is slightly different. For

[1]This behavior is due to the use of log-normal mutations in Line 4 of Alg. 1 that introduce a bias in the σ-mutations.

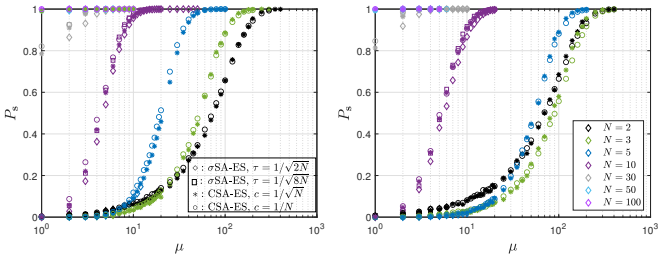

Fig. 8. Experimental values of the success probability P_{s} for ES in the Griewank landscape with $b_i = 1$, $\vartheta = 0.5$ (left figure) and $\vartheta = 0.25$ (right figure), respectively.

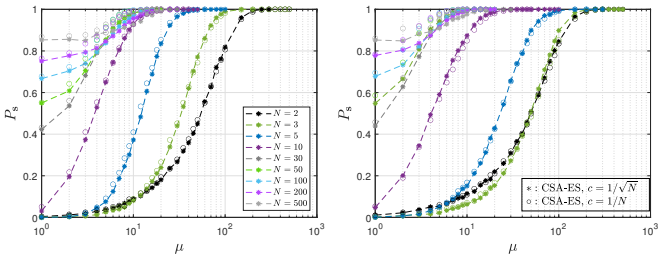

Fig. 9. Experimental values of the success probability P_{s} for ES in the Griewank landscape with $b_i = 1/\sqrt{i}$, $\vartheta = 0.5$ (left figure) and $\vartheta = 0.25$ (right figure), respectively.

$N \geq 3$ the probabilities first become significantly larger as N increases. This is true only up to a value of $N = 30$. Between $N = 30$ and $N = 50$, the success probability increases only slightly. Contrary to expectation, the probabilities decrease slightly for $N \geq 100$ and large values of μ. This effect will be analyzed in more detail in Section VI.

A comparison of the left and right plots in Figs. 8 and 9 shows that the success probabilities tend to be higher for $\vartheta = 0.5$ than for $\vartheta = 0.25$. However, it is expected that a larger offspring population size λ for a constant population size μ will lead to higher success rates because the parental population can be selected from a larger sample. The influence of the offspring population size λ on the success probability is examined in Fig. 10. This figure shows the success probability as a function of the offspring population size λ, where the population size μ is constant. The experimental results were obtained with at least 1000 CSA-ES runs with $c = 1/\sqrt{N}$. The

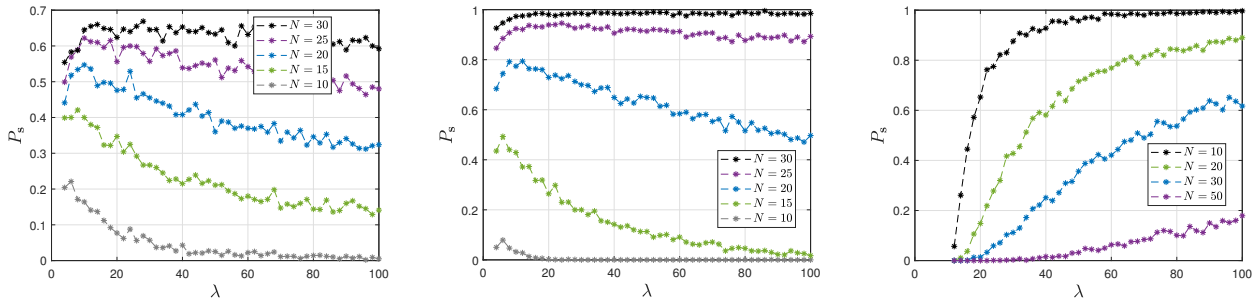

Fig. 10. Success probability depending on offspring population size λ for the CSA-ES with $c = 1/\sqrt{N}$. From left to right: $\mu = 2$ in the Griewank landscape with $b_i = 1/\sqrt{i}$, $\mu = 2$ in the Griewank landscape with $b_i = 1$, and as a contrast $\mu = 10$ in the Rastrigin landscape.

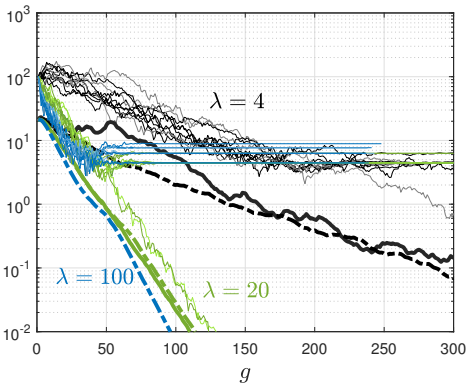

Fig. 11. R and σ dynamics of the CSA-ES with $c = 1/\sqrt{N}$, $\mu = 2$, $N = 10$, $b_i = 1/\sqrt{i}$. The bold lines represent the corresponding mean values of σ for successful runs (solid lines) and unsuccessful runs (dashed lines), respectively.

left plot shows the results for $b_i = 1/\sqrt{i}$ and the middle plot for $b_i = 1$, respectively. First, the success probability increases with increasing population size. This is true for all dimensions. Then, the success probability decreases. This effect is stronger for smaller values of N. This behavior is not as expected. It is expected that a larger offspring population leads to a larger success. This is the case in the right plot of Fig. 10, where the same experiment was executed in the Rastrigin landscape. There, the success probability increases with increasing λ for all dimensions.

Figure 11 shows the R and σ dynamics for the Griewank function with $b_i = 1/\sqrt{i}$. The black lines show the results for $\lambda = 4$. In this case σ decreases more slowly than for $\lambda = 20$ (green), or $\lambda = 100$ (blue). When the mutation strength becomes small very quickly, the probability of getting trapped by a local minimum is higher.

Comparing Fig. 8 with Fig. 9, it also becomes apparent that $b_i = 1$ yields larger success probabilities than $b_i = 1/\sqrt{i}$ for $N \geq 30$. This is not obvious, because for $b_i = 1$ the number of local minima is larger than for $b_i = 1/\sqrt{i}$. This effect becomes stronger for larger N. However, looking at Fig. 12, it can be seen that for the gradient strategy described in Alg. 3, the results are similar.

A possible reason for the unexpected behavior is that for large dimensions and $b_i = 1$, the local attractors are very small. Thus, the probability of hitting them by randomly

Algorithm 3 Gradient Strategy

Initialize $(R, R_{\text{stop}}, \eta, g_{\max}, g = 0, succ = 0)$
$\mathbf{z} = (\mathcal{N}(0,1), \ldots, \mathcal{N}(0,1))^T$
$\mathbf{y} = \frac{\mathbf{z}}{\|\mathbf{z}\|} R$
repeat
  $\nabla F(\mathbf{y}) = \left(\frac{\partial F}{\partial y_1}(\mathbf{y}), \ldots, \frac{\partial F}{\partial y_N}(\mathbf{y}) \right)^T$
  $\mathbf{y} = \mathbf{y} - \eta \nabla F(\mathbf{y})$
  $g = g + 1$
until $g > g_{\max}$
if $\|\mathbf{y}\| < R_{\text{stop}}$ **then**
  $succ = 1$
end if

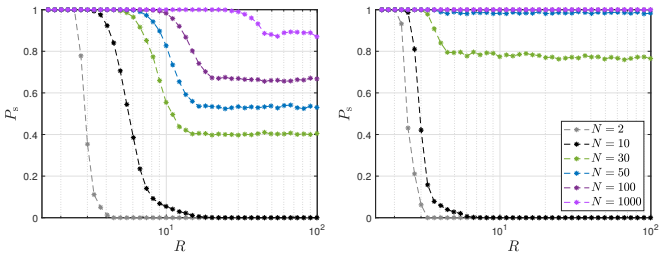

Fig. 12. Success probability of the gradient strategy, Alg. 3, with $\eta = 0.1$, $R_{\text{stop}} = 0.1$, and $g_{\max} = 2 \cdot 10^6$ in the Griewank landscape with $b_i = 1/\sqrt{i}$ (left plot) and $b_i = 1$ (right plot), respectively. Start values were chosen randomly on the R-sphere.

choosing a starting point on the R-sphere is approximately zero for large N. If the starting point is not located in one of the local attractors, it is located in the unbounded global attractor region and the ES converges globally. For $b_i = 1/\sqrt{i}$, the local attractors are larger and therefore the probability of choosing a start value inside one of these attractors is larger. This is visualized in Fig 13. The figures show the Griewank function in a 2-dimensional subspace of $\mathbb{R}^N$ for $N = 100$. In the left plots, the components y_i have been set to zero for $i = 3, \ldots, N$. There are several local attractors and the attractor region is bounded for this intersection. In the remaining plots, the components $y_3, \ldots, y_N$ were randomly chosen from a normal distribution. The farther the plot is to the right, the farther out the intersection point was chosen. The chosen distances d differ between $b_i = 1/\sqrt{i}$ and $b_i = 1$.

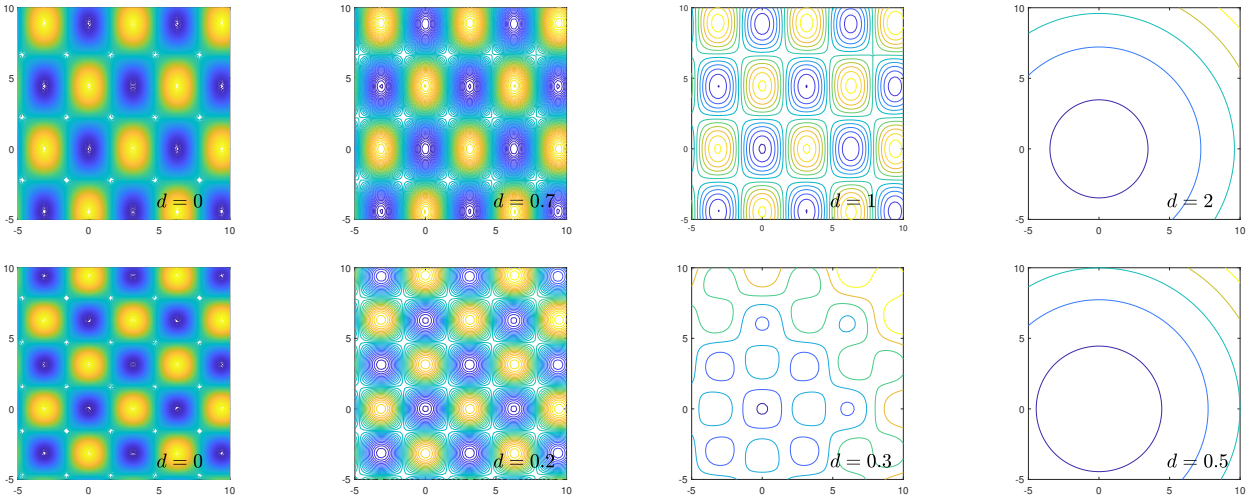

Fig. 13. Griewank landscapes in a 2-dimensional subspace of $\mathbb{R}^{100}$ with $b_i = 1/\sqrt{i}$ (top plots) and $b_i = 1$ (bottom plots), respectively. Components y_i for $i = 3, \ldots, 100$ where set to $d\mathcal{N}(0, 1)$.

The reason for this is that for $b_i = 1$, differently shaped landscapes are generated even for very small distances d, but not for $b_i = 1/\sqrt{i}$. All plots are typical, i.e., repeating this experiment with other random numbers leads to similar plots in almost all cases. These plots visualize that the probability of hitting a specific local attractor in the N-dimensional space is smaller for $b_i = 1$. Additional experiments with a smaller step size of $\eta = 0.001$ show no significant differences in the results of Fig. 12.

Figure 12 also shows that for $N \geq 30$ the success probability remains positive even if the starting point is located far outside. This is consistent with the observations of Section II where it was depicted that the attractor region is unbounded. This figure also allows to estimate an inner bound of the attractor region of about 2.5 for the standard case with $b_i = 1/\sqrt{i}$ and 2 for the special case with $b_i = 1$, respectively.

VI. Estimating the Population Size

A central question concerns the choice of the population size that guarantees the convergence of the ES to the global optimizer. An analytical derivation of a scaling behavior is not possible within the scope of these investigations. However, it can be expected that the population size decreases with increasing N, since the influence of the cosine component becomes less pronounced with larger dimensions. This was shown in Section IV. However, in the experiments in Section V it was also observed that the success rate can slightly drop for large N and μ in the Griewank landscape with standard parameters (Fig. 9).

Figure 14 shows the population size required to achieve a given success probability. The experimental values were determined with at least 500 CSA-ES runs with $c = 1/\sqrt{N}$ and $\vartheta = 0.5$. For $b_i = 1$ and for $b_i = 1/\sqrt{i}$ and $P_\mathrm{s} = 50\%$, the population size decreases with increasing N. For $b_i = 1/\sqrt{i}$ and sufficiently large N the population size grows with N,

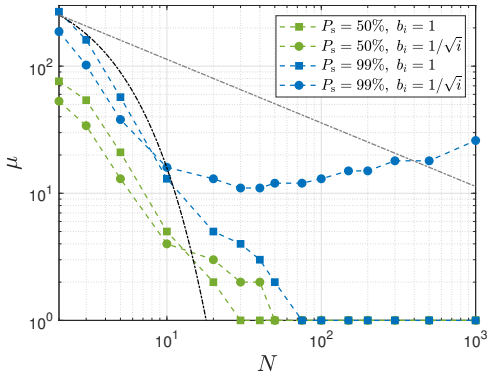

Fig. 14. Population size μ needed to achieve P_s depending on N. Symbols represent the experimental values for the CSA-ES with $c = 1/\sqrt{N}$ and $\vartheta = 0.5$. Gray dashed-dotted line shows (17) with $c_\mathcal{X} = 10$ and $\sigma_\mathrm{ES} = 1/\sqrt{2N}$ and black dashed-dotted line with $\sigma_\mathrm{ES} = 1/\sqrt{2^N}$, respectively.

although both σ_ES^{max} and $\sigma_\mathrm{ES}(R)\,|_{R\to\infty}$ become smaller as N increases. This was demonstrated in Section IV using (14) and the left plot in Fig. 6. This strange behavior will be investigated now.

The relative difference between σ_ES^{max} and $\sigma_\mathrm{ES}(R)\,|_{R\to\infty}$ becomes larger as N increases. σ_ES^{max} scales with $\mathcal{O}(1/\sqrt{N})$ and $\sigma_\mathrm{ES}(R)\,|_{R\to\infty}$ scales with $\mathcal{O}(1/\sqrt{2^N})$. Therefore, there is a region where the noise becomes relatively large compared to the previous noise level. This effect becomes stronger with increasing N. For $b_i = 1/\sqrt{i}$, the maximum standard deviation decreases more slowly than for $b_i = 1$. This can be seen in the left plot of Fig. 6. Therefore, the effect of the noise becoming relatively large compared to the noise strength at large R is also stronger for $b_i = 1/\sqrt{i}$.

If $b_i = 1$, the region where $\sigma_\mathrm{ES} > \sigma_\mathrm{ES}(R)\,|_{R\to\infty}$ is closer to the global optimizer. This can be seen by comparing the top and bottom plots of Fig. 5. This effect becomes stronger with

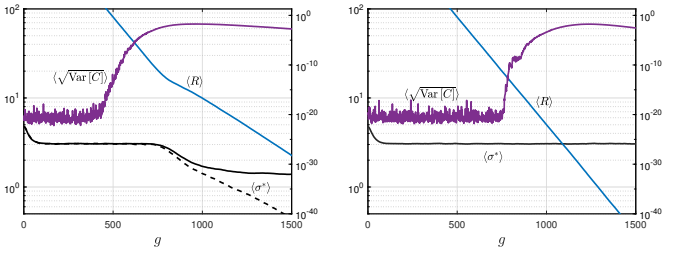

Fig. 15. Mean value dynamics of σ^* (black lines and left y-axis, solid for successful and dashed for unsuccessful runs), R (blue lines and left y-axis) and experimentally determined σ_{ES} (purple lines and right y-axis) for the $(3/3_I, 6)$-CSA-ES with $c = 1/\sqrt{N}$ and $N = 100$. Left plot: $b_i = 1/\sqrt{i}$, right plot: $b_i = 1$.

larger N, as visualized in Fig. 15 for $N = 100$. The purple line shows the mean value of σ_{ES} at generation g. The values were determined experimentally. For $b_i = 1$, $\sigma_{\mathrm{ES}}^{max}$ is reached when the ES has already reached the global attractor region (see Fig. 12) and the ES is converging to the global optimizer. This is a possible reason why the rising population size only occurs for $b_i = 1/\sqrt{i}$.

The scaling behavior of μ that is examined in the following can be partially explained by the frozen noise model. Fig. 12 shows that there is a distance R_{conv} below which a gradient strategy always ends successfully. A closer scrutiny of the data in Fig. 12 shows for $b_i = 1/\sqrt{i}$ that $R_{\mathrm{conv}} \propto \sqrt{N}$ approximately holds. For $b_i = 1$, it holds $R_{\mathrm{conv}} \to \infty$ for $N \geq 50$. An ES optimizing an N-dimensional noisy sphere $aR^2 + \sigma_{\mathrm{ES}}\mathcal{N}(0,1)$ reaches a steady-state R-distribution with $R_{\mathrm{st}} := \mathrm{E}[R]$ and

$$R_{\mathrm{st}} \simeq \sqrt{\frac{\sigma_{\mathrm{ES}} N}{2a\mu c_{\mu/\mu,\lambda}}}, \tag{15}$$

see [11], [8]. If the R_{st}-sphere is located inside the global attractor region, one can assume a success probability of 1. In the noisy Griewank model the cosine part is not normally distributed and the real distribution $\mathcal{X}$ is unknown. Yet, the general structure of (15) does also hold for the Griewank noise, however with a different constant $c_{\mathcal{X}}$ and a σ_{ES} to be taken from experiments

$$R_{\mathrm{st}} \simeq \sqrt{\frac{\sigma_{\mathrm{ES}} N A}{c_{\mathcal{X}}\mu c_{\mu/\mu,\lambda}}}. \tag{16}$$

If R_{st} from (16) is smaller than R_{conv} a success probability of 1 is expected, i.e., if

$$R_{\mathrm{conv}} \geq R_{\mathrm{st}} \Leftrightarrow \mu \geq \frac{\sigma_{\mathrm{ES}} N A}{c_{\mathcal{X}} c_{\mu/\mu,\lambda} R_{\mathrm{conv}}^2}. \tag{17}$$

Considering the left plot of Fig. 6 and (14), it follows that σ_{ES} scales between $\mathcal{O}(1/\sqrt{2^N})$ and $\mathcal{O}(1/\sqrt{N})$. It holds $\sigma_{\mathrm{ES}}^{\max} \approx 1/\sqrt{2N}$. Inserting this and $R_{\mathrm{conv}} \propto \sqrt{N}$ into (17), it follows that μ scales between $\mathcal{O}(1/\sqrt{2^N})$ and $\mathcal{O}(1/\sqrt{N})$. This is confirmed by the experimental data in Fig. 14 except for $b_i = 1/\sqrt{i}$ and $P_{\mathrm{s}} = 99\%$. The dashed-dotted curves show (17) for $\sigma_{\mathrm{ES}} = 1/\sqrt{2N}$ (grey) and $1/\sqrt{2^N}$ (black), respectively. A value of 10 has been chosen for $c_{\mathcal{X}}$, which results from experiments. For $b_i = 1$ it holds $R_{\mathrm{conv}} = \infty$ for $N \geq 50$, which leads to the conclusion that for $b_i = 1$ and $N \geq 50$, the success probability is 1 for all μ. In Fig. 14, this holds for $N \geq 75$.

VII. Conclusion

The Griewank function is known as an example that is easier to optimize in higher dimensions. As has been shown, this property can qualitatively be well explained by a frozen noise model considering the ES optimization process as acting in a quadratic landscape with noise produced by sampling the Griewank landscape by the ES mutations. While it was possible to calculate the noise variance analytically, the non-Gaussian distribution of the noise pervents a detailed quantitative analysis and an analytical derivation of a population sizing law. The empirical investigations reveal, however, some unexpected behaviors that are partially not well understood. While the model explains qualitatively the rather weak influence of the ES strategy parameters on the success rate in contrast to other multimodal test functions such as Rastrigin, it comes as a surprise that under constant μ the success rate exhibits a maximum for relatively small λ value and decreases monotonically when λ is increased further. Another puzzling observation regards the influence of the number of local minima on the success rate. The simpler structure with $b_i = 1$ leads to higher success rates compared to the $b_i = 1/\sqrt{i}$ case although the former landscape has many more local minima.

Understanding this behavior needs further investigations. Therefore, an analysis of the progress rate remains as a future task.

References

[1] A. O. Griewank, "Generalized descent for global optimization," *Journal of optimization theory and applications*, vol. 34, pp. 11–39, 1981.

[2] 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, p. 114179, 2023.

[3] ——, "Convergence Properties of the $(\mu/\mu_I, \lambda)$-ES on the Rastrigin Function," in *FOGA XVII*, H. Aguirre et al., Ed. ACM, 2023.

[4] M. Locatelli, "A Note on the Griewank Test Function," *J. of Global Optimization*, vol. 25, no. 2, p. 169–174, 2003.

[5] H. Cho, F. Olivera, and S. Guikema, "A derivation of the number of minima of the Griewank function," *Applied Mathematics and Computation*, vol. 204, no. 2, pp. 694–701, 2008.

[6] Y. Huang, J. Li, and P. Wang, "Unusual phenomenon of optimizing the Griewank function with the increase of dimension," *Frontiers of Information Technology & Electronic Engineering*, vol. 20, no. 10, pp. 1344–1360, 2019.

[7] L. Schönenberger and H.-G. Beyer, "On a Population Sizing Model for Evolution Strategies Optimizing the Highly Multimodal Rastrigin Function," ser. GECCO '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 848–855.

[8] S. Meyer-Nieberg, *Self-Adaptation in Evolution Strategies*. Ph. D. Dissertation. University of Dortmund, CS Department, 2007.

[9] N. Hansen, *Verallgemeinerte individuelle Schrittweitenregelung in der Evolutionsstrategie*. Doctoral thesis. Technical University of Berlin, 1998.

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

[11] D. Arnold and H.-G. Beyer, "Performance Analysis of Evolution Strategies with Multi-Recombination in High-Dimensional R^N-Search Spaces Disturbed by Noise," *Theoretical Computer Science*, vol. 289, pp. 629–647, 2002.