On a Population Sizing Model for Evolution Strategies in Multimodal Landscapes

Lisa Schönenberger and Hans-Georg Beyer

***Abstract*—This paper derives a population sizing model for standard Evolution Strategies (ES) in highly multimodal fitness landscapes with exponentially many local optima. The Rastrigin, Bohachevsky, and Ackley test functions are considered. Due to the highly non-convex structure of these functions a detailed analytical description of the behavior of the ES is a challenge. Therefore, a model is derived that simplifies the complex structure of the functions under consideration. The main idea of this model is the interpretation of local landscape oscillations as frozen noise. This allows for an estimation of the success probability of the ES converging to the global optimum and in turn an estimation of the population size required. It is shown that the population size scales usually sublinearly with the search space dimension N. For the Rastrigin and Bohachevsky function, the population size scales with $\mathcal{O}(\sqrt{N}\ln(N))$. As for Ackley, the scaling behavior depends strongly on the initial values. If the algorithm starts in a certain vicinity of the global optimizer, the dependence on the dimension N is rather weak. However, if the initial value exceeds a certain distance R to the optimizer, the population size scales exponentially with R.**

***Index Terms*—Evolution Strategies, global optimization, multimodal objective function, global convergence, population sizing**

I. Introduction

Optimization in highly multimodal, real-valued fitness landscapes is a challenging topic and the theoretical analysis is still in its infancy. Due to their underlying stochastic nature, Evolution Strategies (ES) have proven to be well suited for the optimization of highly multimodal problems. However, the success rate of ES for finding the global optimizer depends strongly on the population size, as already observed in [1] for the CMA-ES [2]. This also holds for $(\mu/\mu_I,\lambda)$-ES using isotropic mutations with σ self-adaptation (σSA) or cumulative step-size adaptation (CSA) for mutation strength control. On the one hand the population size must not be too small; in this case the ES will not find the global optimizer. A too large population size on the other hand would require too many resources. Furthermore, the optimal population size depends on the search space dimensionality N. This is because the number of local minima typically increases with increasing N.

This increase usually disqualifies the application of classical non-linear optimization algorithms. That is, single-run gradient-based strategies are not able to find the global optimizer. Instead, restart strategies must be applied. However, since the number of restarts must increase exponentially with N, such strategies are not a real alternative because this implies an exponentially increasing number of function evaluations. In contrast, there are indications that for ES the population size scales slower than exponentially, at most quadratically or even partially sub-linearly with the problem dimension, thus, keeping the number of function evaluations at a moderate order. In [1] this was experimentally demonstrated for several multimodal test functions, such as Rastrigin, Bohachevsky, Ackley, Schaffer, and Schwefel. In [3] these experimental results were confirmed by an analytical model for the Rastrigin function.

It is the aim of this paper to extend this model to other multimodal test functions that share certain similarities with the Rastrigin function, i.e., the Bohachevsky and the Ackley function. As a result, population sizing equations will be obtained that scale sub-linearly with the search space dimension N. The approach taken and findings made may be regarded as one of the first steps towards an analysis of ES on highly multimodal optimization problems. The population sizing results can be used to evaluate the effectiveness of population sizing rules. This in turn may be a starting point for further algorithm as well as benchmark designs. Last but not least, the analysis also provides a deeper understanding of how ES finds global optima.

The remainder of this paper is organized as follows. After a short introduction of the ES algorithms, the multimodal test functions will be introduced in Section III. In Section IV the frozen noise model will be developed. Equations for the success probability will be derived in Section V. In Section VI the scaling behavior of the population size will be investigated. Finally, a summary of the results and an outlook on future research will be given.

II. ES-Algorithms

The basic $(\mu/\mu_I,\lambda)$-ES algorithms investigated here consist of μ parents and λ offspring with truncation ratio $\vartheta := \mu/\lambda$. The subscript "$m;\lambda$" denotes the selection of the $m = 1,\ldots,\mu$ best individuals out of λ. The control of the strength

Manuscript received February 23, 2024; revised May 14, 2024 and June 28, 2024. This research was funded by the Austrian Science Fund (FWF) under grant P33702-N. For open access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission. The authors gratefully thank Amir Omeradzic for providing valuable feedback.

The authors are with the Research Center Business Informatics, Vorarlberg University of Applied Sciences, 6850 Dornbirn, Austria (e-mail: lisa.schoenenberger@fhv.at, hans-georg.beyer@fhv.at).

This article has supplementary material provided by the authors available at xx.xx.xx.

Digital Object Identifier XX.XXXX/TEVC.XXXX.XXXXXXX

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

σ of the isotropic Gaussian mutations used is done by either σ self-adaptation (σSA), see Alg. 1, or cumulative step-size adaptation (CSA), see Alg. 2. For the σSA-ES the offspring mutation strengths are sampled from a log-normal distribution with learning parameter τ. The standard choice of the learning parameter is $\tau = 1/\sqrt{2N}$, which guarantees optimal performance on the sphere model [4]. A smaller τ yields a slower adaptation. This is an advantage in case of multimodal problems, which is one result of the investigations here. Therefore, a smaller learning rate of $\tau = 1/\sqrt{8N}$ will also be considered. For the CSA-ES, the standard choice for the cumulation time parameter c is $1/N$ and $1/\sqrt{N}$ [5], [6], where the latter results in faster convergence but a lower success probability P_{s}, which is also confirmed in the following sections.

III. Multimodal Test Functions

The multimodal test functions considered have a high number of local optima: Rastrigin, Bohachevsky, and Ackley. The Rastrigin function $F_{\mathcal{R}}$ for an N-dimensional search vector $\mathbf{y} = (y_1, \ldots, y_N)$ is given by

$$F_{\mathcal{R}}(\mathbf{y}) = \sum_{i=1}^{N} \left[y_i^2 + A\left(1 - \cos(\alpha y_i)\right)\right] \tag{1}$$

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

Initialize $\left(\mathbf{y}^{(0)}, \sigma^{(0)}, \sigma_{\mathrm{stop}}, g = 0\right)$
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 $\sigma^{(g)} < \sigma_{\mathrm{stop}}$

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

Initialize $\left(\mathbf{y}^{(0)}, \sigma^{(0)}, \sigma_{\mathrm{stop}}, \mathbf{s} = \mathbf{1}, g = 0\right)$
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, see [6, p.13]
until $\sigma^{(g)} < \sigma_{\mathrm{stop}}$

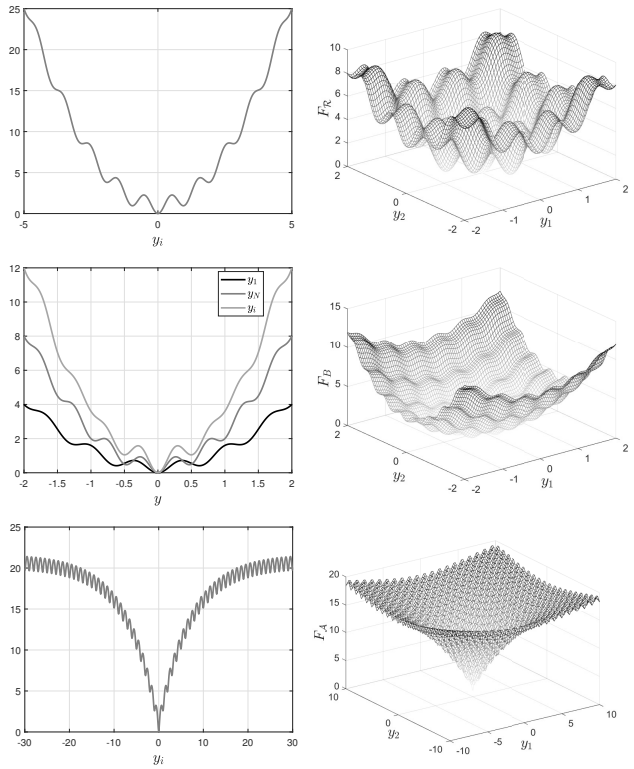

Fig. 1. Multimodal test functions for $N = 1$ and $N = 2$. From top to bottom: Rastrigin (1), Bohachevsky (2), Ackley (3).

where the parameter A denotes the oscillation amplitude and α denotes the frequency. Unless otherwise stated, the parameters A and α have in all experiments the values $A = 1$ and $\alpha = 2\pi$. The global optimizer is located at $\hat{\mathbf{y}} = \mathbf{0}$ and is surrounded by local minima. The number of local minima is $\kappa^N - 1$ where κ increases with A and α (e.g., for $\alpha = 2\pi$, $A = 1$: $\kappa = 7$ and for $\alpha = 2\pi$, $A = 10$: $\kappa = 63$). The top plots of Fig. 1 show the Rastrigin function with standard parameters for $N = 1$ and $N = 2$, respectively.

The Bohachevsky function is given by

$$F_{\mathcal{B}}(\mathbf{y}) = \sum_{i=1}^{N-1} \big[y_i^2 + 2y_{i+1}^2 - B_1 \cos\left(\beta_1 y_i\right) - B_2 \cos\left(\beta_2 y_{i+1}\right) + B_1 + B_2 \big]. \tag{2}$$

The parameters B_1, B_2, β_1, and β_2 control the size and frequency of the oscillations of the function. Larger values of B_1 and B_2 result in larger oscillations. Larger values of β_1 and β_2 increase the frequency and therefore the number of local minima. The standard values of these parameters are $B_1 = 0.3$, $B_2 = 0.4$, $\beta_1 = 3\pi$, and $\beta_2 = 4\pi$. Unless otherwise stated, these standard values will be used for all experiments. Compared to the Rastrigin function, the Bohachevsky function is not equally shaped in each dimension, as visualized in the middle-left plot of Fig. 1. It represents single coordinate contributions of the Bohachevsky function.

The Ackley function is given by

$$F_{\mathcal{A}}(\mathbf{y}) = C_1 - C_1 \mathrm{e}^{-C_2 \sqrt{\frac{1}{N}\sum_{i=1}^{N} y_i^2}} + \mathrm{e} - \mathrm{e}^{\frac{1}{N}\sum_{i=1}^{N} \cos(\gamma y_i)} \tag{3}$$

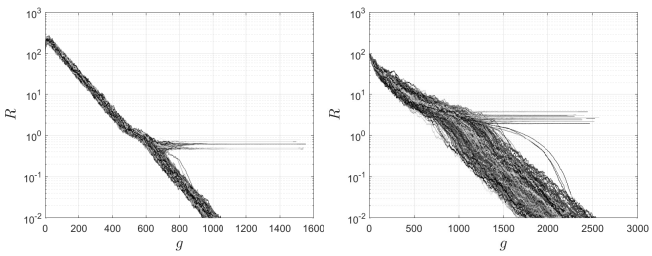

Fig. 2. R-dynamics for 500 $(\mu/\mu_I, 2\mu)$-σSA-ES runs with $\tau = 1/\sqrt{8N}$ for $N = 100$. Left plot: Bohachevsky landscape with $\mu = 50$. Right plot: Ackley landscape with $\mu = 5$. The success probabilities are $P_{\mathrm{s}} = 0.88$ and $P_{\mathrm{s}} = 0.9$, respectively. For each run, the ES was initialized randomly with distance $R^{(0)} = 100$ to the global optimizer.

and is represented by the bottom plots of Fig. 1 for $N = 1$ and $N = 2$. The function describes a funnel-shaped surface with an infinite number of local minima. Larger values of C_1 lead to larger horizontal asymptotes. The larger C_2, the steeper is the slope around the origin. The frequency is determined by γ. The standard values for the Ackley parameters are $C_1 = 20$, $C_2 = 0.2$, and $\gamma = 2\pi$. Unless otherwise stated, these parameters will be used in all subsequent experiments.

Figure 2 shows the dynamics of the distance of the parental centroid to the global optimizer, i.e., $R(g) := \|\mathbf{y}^{(g)}\|$ of 500 independent σSA-ES runs, Alg. 1, in the Bohachevsky landscape (left plot) and in the Ackley landscape (right plot). Most runs converge to the global optimizer, but a certain percentage of runs is getting trapped in local minima. The percentage of runs that approach the global optimizer (i.e., the successful runs) is referred to as *success probability* P_{s}. The mean value dynamics of the successful runs in Fig. 2 are displayed in Fig. 3 and symbolized by angular brackets. The gray solid lines represent the distance R to the optimizer, mutation strength σ and its normalization

$$\sigma^* := \sigma \frac{N}{R}. \tag{4}$$

The values of σ^* are represented on the right y-axis. Having a closer look at the dynamics, one sees that the evolution process can be divided into different phases. At the beginning the initial parental centroid is initialized far enough from the global optimizer. Self-adaptation decreases the normalized σ^* for the Bohachevsky function, where the initial σ was chosen too large. For the Ackley function, on the other hand, the initial σ was (intentionally) chosen too small. Again σ^* reaches a nearly constant level that depends on the general structure of the function, which is further investigated in Section IV. Getting closer to the global optimizer, one observes a certain decrease of σ^* indicating the influence of the *local attractors* rewarding local fitness gain due to smaller mutation steps. That is, the ES reduces the normalized mutation strength σ^*. Finally, in the third phase, the ES is either trapped in a local attractor or it has hit the *global attractor* $\mathcal{A}_{\mathrm{ES}}$, to be introduced in Section V-A. In the case of successful runs σ^* reaches the constant level again. The conditions under which the global attractor is reached and the success probability P_{s} by which an ES approaches the global optimizer will be determined in Section V-B.

The solid black line in Fig. 3, shows the standard deviation of the residual part (to be defined below), which was determined by experiments. The theoretical calculations for the standard deviation of the residual part are represented by the dashed-dotted black lines, with the horizontal lines indicating the asymptotic behavior. This will be further investigated in Section IV-C. Similar graphs can be obtained for the Rastrigin landscape. The CSA-ES, Alg. 2, exhibits similar behaviors, see supplement material.

IV. The Frozen Noise Model

A. The Frozen Noise Model in a Nutshell

A theoretical analysis of the behavior of ES is usually based on progress rate analysis. This approach has been done for the Rastrigin function in [7]–[9]. However, this approach is very involved and up to now restricted to the analysis of the expected value behavior. In this paper we use an alternative approach first proposed for the Rastrigin function in [3].

The evolution of the ES is interpreted as a walk through a frozen noise landscape. The derivation of the frozen noise model comprises two steps.

First, the function $F(\mathbf{y})$ to be optimized is subdivided into a global part $G(\mathbf{y})$ and a residual part $C(\mathbf{y})$ where the latter contains the non-convex nonlinearities of F.

$$F(\mathbf{y}) =: G(\mathbf{y}) + C(\mathbf{y}). \tag{5}$$

In this paper the global part is assumed to be expressible as a sphere model, i.e., $G(\mathbf{y}) = G(\|\mathbf{y}\|) = G(R)$.

Second, the residual part $C(\mathbf{y})$ is replaced by a noise term the distribution of which and its statistical parameters are to be determined. In order to further apply the noisy sphere theory [10] a Gaussian noise distribution will be used as approximation.

Determining the noise distribution is not arbitrary, but must reflect the real dynamics of the ES run. To this end it is important to recall how the ES explores the search space by a *restricted random walk* [11]. The ES mutations in Line 5 of Alg. 1 and Alg. 2 sample the fitness landscape in an unbiased manner. The direction of the evolution, however, is given by the selection. From this point of view one may regard the evolution process as a restricted random walk because of the selection. Furthermore, each mutation and also the changes of the recombined parents (Line 10 in Alg. 1, 2) do have a certain contribution towards the optimizer, but also contributions perpendicular to it. The length of this perpendicular part is on average by a factor of $\sqrt{N}$ larger than the contribution towards the optimizer [11]. Furthermore, after the ES has reached the steady state, each component of $\mathbf{y}$ is normally distributed with mean zero [10]. As a result, the parental centroids circle around the global optimizer (provided that the mutation strength is sufficiently large). This may be regarded as a sampling process, where the distance R to the global optimizer decreases only slightly from one generation to the next. If one neglects this small R decrease, i.e., considers the mutation process under the assumption of $R = \|\tilde{\mathbf{y}}\| = \mathrm{const.}$ then the fitness distribution of the offspring and therefore its standard deviation (below denoted as σ_{ES} to indicate that

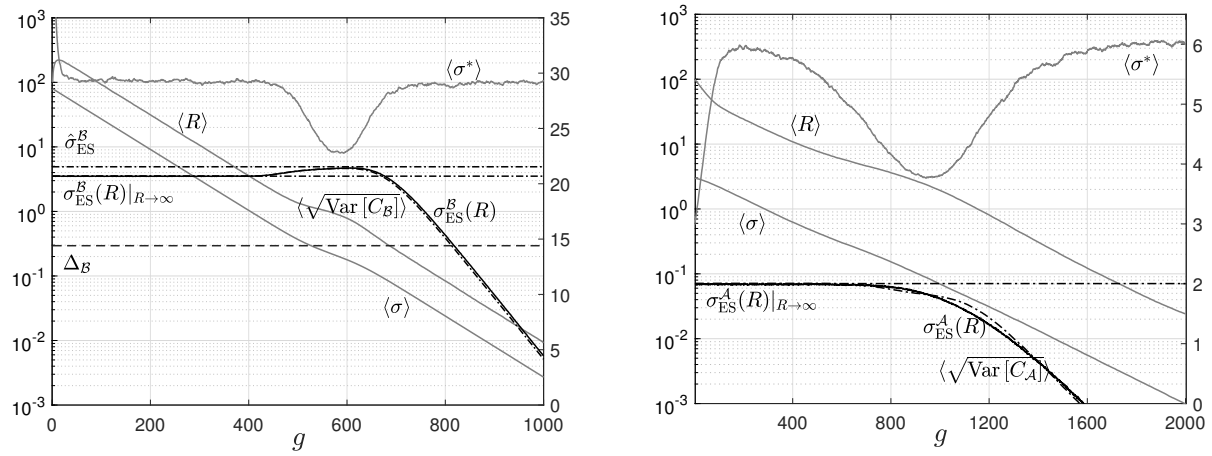

Fig. 3. Mean value dynamics derived from the *successful* runs displayed in Fig. 2. Left plot: Bohachevsky landscape with $\mu = 50$. Right plot: Ackley landscape with $\mu = 5$. In addition to the R dynamics, the mutation strength σ and its normalization σ^* (right y-axis) are displayed. $\sigma_{\mathrm{ES}}(R)$, (15), (16), $\sigma_{\mathrm{ES}}(R)\,|_{R\to\infty}$, (25), (33), $\hat{\sigma}^{\mathcal{B}}_{\mathrm{ES}}$, (26), and $\langle\sqrt{\mathrm{Var}[C]}\rangle$ are discussed in Sect. IV-C. The distance $\Delta_{\mathcal{B}}$ (40) is introduced in Sect. V-A.

the noise is generated by the mutations of the ES) is solely determined by the residual part $C(\tilde{\mathbf{y}})$ of F. Thus, the ES "sees" basically a sphere model disturbed by distance depending noise. This defines the *frozen noise model*

$$\tilde{F}(\mathbf{y}) := G(R) + \sigma_{\mathrm{ES}}(R)\mathcal{N}(0,1), \quad R = \|\mathbf{y}\|, \tag{6}$$

consisting of the global sphere model part $G(\mathbf{y}) = G(R)$ and a normally distributed perturbation part where its conditional variance is defined as

$$\sigma^2_{\mathrm{ES}}(R) := \mathrm{Var}[C(\mathbf{y} + \sigma\mathbf{z}) \,|\, \|\mathbf{y}\| = R], \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{7}$$

The assumption of a normal distribution in (6) must be generally regarded as an approximation. However, there is empirical evidence that this is a reasonable assumption (see, e.g., Fig. 4). In the following $G(R)$ and $\sigma_{\mathrm{ES}}(R)$ will be determined for the different test functions.

B. Building the Models

As for the Rastrigin function (1) the decomposition in terms of (5) reads

$$G_{\mathcal{R}}(\mathbf{y}) := \sum_{i=1}^{N} y_i^2 = R^2, \tag{8}$$

$$C_{\mathcal{R}}(\mathbf{y}) := NA - A\sum_{i=1}^{N} \cos(\alpha y_i), \tag{9}$$

where R is the distance to the global optimizer. The residual $C_{\mathcal{R}}(\mathbf{y})$ describes the oscillations of the Rastrigin function. It holds $C_{\mathcal{R}}(\mathbf{y}) \in [0, 2NA]$.
The decomposition of the Bohachevsky function (2) reads

$$G_{\mathcal{B}}(\mathbf{y}) := y_1^2 + 3\sum_{i=2}^{N-1} y_i^2 + 2y_N^2 = 3R^2 - 2y_1^2 - y_N^2, \tag{10}$$

$$C_{\mathcal{B}}(\mathbf{y}) := (N-1)(B_1 + B_2) - \sum_{i=1}^{N-1} [B_1 \cos(\beta_1 y_i) + B_2 \cos(\beta_2 y_{i+1})]. \tag{11}$$

Note that the global part describes an hyperellipsoid. For $N \to \infty$ one can neglect the distortion of the 1st and last y-component to get a sphere model that will be used as an approximation for smaller N (see also Fig. 5, left plot for an exemplary comparison). For the residual part it holds $C_{\mathcal{B}}(\mathbf{y}) \in [0, 2(N-1)(B_1 + B_2)]$.
The Ackley function (3) yields

$$G_{\mathcal{A}}(\mathbf{y}) := C_1 - C_1 \exp\left(-C_2\sqrt{\frac{1}{N}\sum_{i=1}^{N} y_i^2}\right), \tag{12}$$

$$C_{\mathcal{A}}(\mathbf{y}) := \mathrm{e} - \exp\left(\frac{1}{N}\sum_{i=1}^{N} \cos(\gamma y_i)\right). \tag{13}$$

The global part is a non-quadratic sphere model with a funnel-shaped surface (see Fig. 1). For the residual part one finds $C_{\mathcal{A}}(\mathbf{y}) \in \left[0, \mathrm{e} - \mathrm{e}^{-1}\right]$.

Using (6) together with the global parts (8), (10), and (12), respectively, one obtains the corresponding frozen noise models for the Rastrigin landscape

$$\tilde{F}_{\mathcal{R}}(\mathbf{y}) = R^2 + \sigma^{\mathcal{R}}_{\mathrm{ES}}(R)\mathcal{N}(0,1), \tag{14}$$

for the Bohachevsky landscape

$$\tilde{F}_{\mathcal{B}}(\mathbf{y}) = 3R^2 + \sigma^{\mathcal{B}}_{\mathrm{ES}}(R)\mathcal{N}(0,1), \tag{15}$$

and the Ackley landscape

$$\tilde{F}_{\mathcal{A}}(\mathbf{y}) = C_1 - C_1\mathrm{e}^{-C_2\frac{R}{\sqrt{N}}} + \sigma^{\mathcal{A}}_{\mathrm{ES}}(R)\mathcal{N}(0,1), \tag{16}$$

where $R = \|\mathbf{y}\|$. The calculation of σ_{ES} and the assumption of a normal distribution are examined in the following.

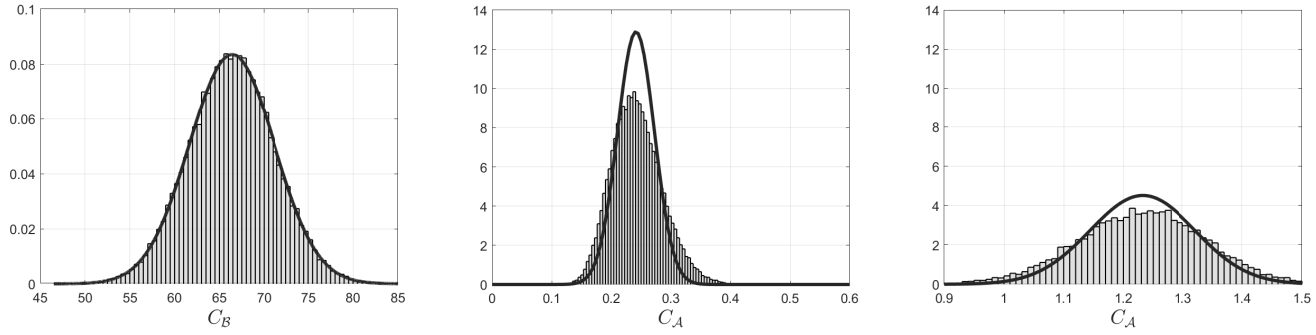

Fig. 4. Histogram of residual parts for 300 $(50/50_I, 100)$-CSA-ES runs with $c = 1/\sqrt{N}$ and $N = 100$ in the Bohachevsky landscape at $R^{\mathcal{B}}_{\rm st} = 1.02$, (36) (left plot) and in the Ackley landscape at $R = 0.22$, (middle plot) and $R = 0.27$ (right plot). Bold solid lines show the pdf of normally distributed variates.

C. Distribution of the Residual Parts

In the noisy sphere models (14) – (16), the assumption of Gaussian noise was made. For the Rastrigin and Bohachevsky function, this can be justified by considering the residual parts $C_{\mathcal{R}}(\mathbf{y})$, (9), and $C_{\mathcal{B}}(\mathbf{y})$, (11) as a sum of independent random variables $\cos(\alpha\tilde{y}_i)$[1] for which the central limit theorem of statistics holds in the limit $N \to \infty$. The assumption of a normal distribution is confirmed by the left plot in Fig. 4 where the histogram of the residual part (11) of the Bohachevsky landscape is shown. The histogram was obtained at the critical distance $R^{\mathcal{B}}_{\rm st}$ (36) where the probability of getting trapped into a local attractor gets larger. The bold line shows the pdf of a normal distribution where the mean and variance were taken from the experimental data.

The calculation of the noise variances needed in (14)–(16) starts from (7) by first determining the variance of $C(\mathbf{y}+\sigma\mathbf{z})$ for a fixed parental centroid state $\mathbf{y}$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$$\sigma^2_{\rm ES}(\mathbf{y}) = {\rm Var}[C(\mathbf{y}+\sigma\mathbf{z})]. \tag{17}$$

It is assumed that each offspring has the same mutation strength σ. This is exactly fulfilled for the CSA-ES (Alg. 2, Line 5), for the σSA-ES it holds asymptotically for $N \to \infty$. In a second step, averaging over the hypersphere $\|\mathbf{y}\| = R$ must be performed

$$\sigma^2_{\rm ES}(R) = \int_{\|\mathbf{y}\|=R} \sigma^2_{\rm ES}(\mathbf{y}) {\rm d}^N y. \tag{18}$$

This can be done only for $N \leq 2$ in closed form, however, for $N \to \infty$ one can express (18) as an expected value of a normally distributed random vector [9]

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{0}, \frac{R^2}{N}\mathbf{I}\right), \tag{19}$$

$$\sigma^2_{\rm ES}(R) = {\rm E}[\sigma^2_{\rm ES}(\mathbf{y})]. \tag{20}$$

The variance and expected value calculations in (17) and (20) are straightforward but lengthy. They are detailed in Supplement Section I. Only the calculation of (17) for the residual (9) of Rastrigin will be sketched. Using (9) in (17), one obtains

$$\begin{aligned}\left(\sigma^{\mathcal{R}}_{\rm ES}(\mathbf{y})\right)^2 &= A^2 {\rm Var}\left[\sum_{i=1}^N \cos(\alpha y_i + \alpha\sigma z_i)\right] \\ &= A^2 \sum_{i=1}^N {\rm Var}[\cos(\alpha y_i + \alpha\sigma z_i)]. \end{aligned} \tag{21}$$

Here, the statistical independences of the $z_i \sim \mathcal{N}(0,1)$ have been taken into account. Since the argument in the cosine can be interpreted as a random variate $x \sim \mathcal{N}(\alpha y_i, (\alpha\sigma)^2)$ one can show (see Supplement)

$$\begin{aligned}&{\rm Var}[\cos(\alpha y_i + \alpha\sigma z_i)] \\ &\quad = \frac{1}{2}\left(1 - {\rm e}^{-(\alpha\sigma)^2}\right)\left(1 - {\rm e}^{-(\alpha\sigma)^2}\cos(2\alpha y_i)\right). \end{aligned} \tag{22}$$

Taking further ${\rm E}[\cos(2\alpha y_i)] = \exp\left(-\frac{2(\alpha R)^2}{N}\right)$ into account and substituting this into (22) and in turn into (21), one finally obtains for (20)

$$\left(\sigma^{\mathcal{R}}_{\rm ES}(R)\right)^2 = A^2\frac{N}{2}\left(1 - {\rm e}^{-(\alpha\sigma)^2}\right)\left(1 - {\rm e}^{-(\alpha\sigma)^2}{\rm e}^{-\frac{2(\alpha R)^2}{N}}\right). \tag{23}$$

This result agrees well with runs of the CSA-ES and shows only small deviations for the σSA-ES (see Fig. 1 in the Supplement). The maximum noise is obtained for $R \to \infty$ since $\sigma = \sigma^* R/N$ one finds using (23)

$$\hat{\sigma}^{\mathcal{R}}_{\rm ES} := \max_R \left[\sigma^{\mathcal{R}}_{\rm ES}(R)\right] = \sigma^{\mathcal{R}}_{\rm ES}(R) \mid_{R\to\infty} = A\sqrt{\frac{N}{2}}. \tag{24}$$

It will serve as an upper bound in subsequent calculations.

For the Bohachevsky function the calculation of $\sigma^{\mathcal{B}}_{\rm ES}(R)$ using (11) is similar, but there are additional covariance terms. The resulting rather large variance expression can be found in the Supplement. It holds

$$\sigma^{\mathcal{B}}_{\rm ES}(R) \mid_{R\to\infty} = \sqrt{\frac{1}{2}(N-1)(B_1^2 + B_2^2)}, \tag{25}$$

and the maximum value of $\sigma^{\mathcal{B}}_{\rm ES}(R)$ can be approximated as

$$\begin{aligned}\hat{\sigma}^{\mathcal{B}}_{\rm ES} &:= \max_R \left[\sigma^{\mathcal{B}}_{\rm ES}(R)\right] \\ &\approx \sqrt{\frac{1}{2}(N-1)(B_1+B_2)^2 - B_1 B_2}. \end{aligned} \tag{26}$$

[1]Here we have used $\tilde{y}_i$ to indicate the offspring components (see Line 5 in Alg. 1 and 2) that are responsible for the landscape sampling of the ES.

(see Supplement Section I-B). These values are represented by the dashed-dotted lines in the left plot of Fig. 3.

In the case of the Ackley function one observes that the cosine sum appears in the exponent of (13). It is shown in Supplement Section I-C that the residual part of the Ackley function becomes asymptotically a log-normally distributed random variate

$$C_{\mathcal{A}}(\mathbf{y}) \approx \mathrm{e} - \mathrm{e}^{\mathcal{N}(\mu_{\mathcal{A}},\sigma_{\mathcal{A}})}. \tag{27}$$

This log-normal distribution can be approximated by a normally distributed random variable with mean and variance

$$\mathrm{E}\left[C_{\mathcal{A}}\right] = \mathrm{e} - \mathrm{e}^{\mu_{\mathcal{A}}+\frac{1}{2}\sigma_{\mathcal{A}}^2} \tag{28}$$

$$\mathrm{Var}\left[C_{\mathcal{A}}\right] = \mathrm{e}^{2\mu_{\mathcal{A}}+\sigma_{\mathcal{A}}^2}\left(\mathrm{e}^{\sigma_{\mathcal{A}}^2}-1\right), \tag{29}$$

where

$$\mu_{\mathcal{A}} = \mathrm{e}^{-\frac{1}{2}(\sigma\gamma)^2}\mathrm{e}^{-\frac{1}{2}\frac{(\gamma R)^2}{N}} \tag{30}$$

$$\sigma_{\mathcal{A}}^2 = \frac{1}{2N}\left(1-\mathrm{e}^{-(\sigma\gamma)^2}\right)\left(1-\mathrm{e}^{-(\sigma\gamma)^2}\mathrm{e}^{-\frac{1}{2}\frac{(\gamma R)^2}{N}}\right). \tag{31}$$

Therefore,

$$\sigma_{\mathrm{ES}}^{\mathcal{A}}(R) = \sqrt{\mathrm{Var}\left[C_{\mathcal{A}}(R)\right]} = \sqrt{\mathrm{e}^{2\mu_{\mathcal{A}}+\sigma_{\mathcal{A}}^2}\left(\mathrm{e}^{\sigma_{\mathcal{A}}^2}-1\right)} \tag{32}$$

(see Supplement Section I-C). The middle and right plot of Fig. 4 depict the experimentally determined distribution of the residual part at two different distances. The bold solid line shows the pdf of a normal distribution with mean and variance predicted by (28) and (29). The values of σ used for the calculation of μ_A and $\sigma_{\mathcal{A}}$ were determined by the experimental values. The pdf of the corresponding log-normal distribution is not displayed since the differences to the normal distribution are nearly not visible. Similar to the other test functions the standard deviation of the residual part $C_{\mathcal{A}}$ approaches a maximum saturation value as $R \to \infty$ as can be seen in the rightmost graph of Fig. 3. Calculating the limit of (30), one sees that $\mu_{\mathcal{A}} \to 0$ and for (31) $\sigma_{\mathcal{A}}^2 \to \frac{1}{2N}$. Thus, one gets from (32) for $N \to \infty$ the saturation value

$$\hat{\sigma}_{\mathrm{ES}}^{\mathcal{A}} := \sigma_{\mathrm{ES}}^{\mathcal{A}}(R)\mid_{R\to\infty} = \frac{1}{\sqrt{2N}} \tag{33}$$

after a short calculation. This saturation value is depicted by the dashed-dotted line in the right plot of Fig. 3.

D. Steady-State in the Noise Model

It is known from the noisy sphere model theory that an ES optimizing an N-dimensional noisy sphere, $F_{\mathcal{S}} = aR^2 + \sigma_{\mathrm{ES}}\mathcal{N}(0,1)$, with constant noise strength σ_{ES} reaches a steady-state R-distribution with $R_{\mathrm{st}} := \mathrm{E}\left[R\right] \neq 0$ where

$$R_{\mathrm{st}} \simeq \sqrt{\frac{\sigma_{\mathrm{ES}}N}{4a\mu c_{\mu/\mu,\lambda}}}, \tag{34}$$

see [4], [12]. It also holds for the noisy sphere [10] that under steady-state conditions, each component of $\mathbf{y}$ is normally distributed with

$$y_i = (\mathbf{y})_i \sim \mathcal{N}\left(0, \frac{R_{\mathrm{st}}^2}{N}\right). \tag{35}$$

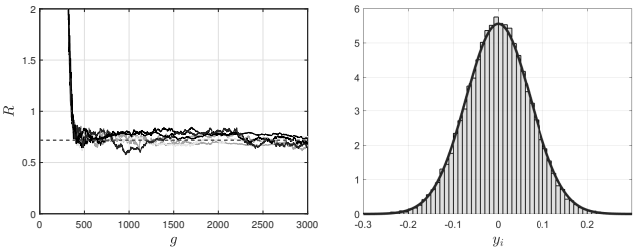

Fig. 5. Left plot: R-dynamics of the noisy hyperellipsoid model (10), (11) (dark lines) and noisy sphere model (15) (light lines) using $\hat{\sigma}_{\mathrm{ES}}^{\mathcal{B}}$ from (26). Black dashed line shows $R_{\mathrm{st}}^{\mathcal{B}}$. Right plot: Histogram of all individual components y_i at distance $R_{\mathrm{st}}^{\mathcal{B}}$. Bold solid line shows the pdf of the $\mathcal{N}\left(0, R_{\mathrm{st}}^2/N\right)$ variate. Experiments executed for the $(100/100_I, 200)$-CSA-ES with $c = 1/\sqrt{N}$, $N = 100$, and $R_{\mathrm{st}}^{\mathcal{B}} = 0.72$, (36).

As already shown in [3], this also holds approximately for the parental distribution of the ES on Rastrigin as long as the ES is not trapped in one of the local attractors. In the case of the Bohachevsky landscape the global part is a hyperellipsoid. As already state, one can assume that for large N the first and last component has only limited influence on the steady-state distribution. Neglecting the first and last component, one gets the noisy sphere model (15). Using (34) the expected value of the steady-state becomes

$$R_{\mathrm{st}}^{\mathcal{B}} \simeq \sqrt{\frac{\hat{\sigma}_{\mathrm{ES}}^{\mathcal{B}}N}{12\mu c_{\mu/\mu,\lambda}}}. \tag{36}$$

This is illustrated in Fig. 5 for dimension $N = 100$, where the left plot shows that the value of $R_{\mathrm{st}}^{\mathcal{B}}$ predicts the residual distance of the experimental data well. Equation (35) also holds approximately for the Bohachevsky function, as illustrated in the right plot of Fig. 5. The histogram shows the distribution of all components lumped together. The variance of the experimental values is 0.0052 as predicted. Note that the distributions of the single first and single last component have larger variances which can be neglected. The influence of this simplification is investigated in supplement Section III-B.

V. The Sucess Probability Model

A. The Global Attractor Region

Using the models (14), (15), and (16), the evolution of the ES in the multimodal landscape can be interpreted as a noisy minimization problem. Recall that the ES reaches the vicinity of the global optimizer up to a distance R_{st} provided that the noise variance does not change. However, if this distance is close enough to the global optimizer the noise strength declines (see Fig. 3) and a large success probability P_{s} can be expected. To specify what is close enough, the following definition is introduced.

The attractor region is defined as the region where the negative gradient flow goes towards the optimizer. In [3] it was already shown for the Rastrigin function that the attractor region is a hypercube

$$\mathcal{A}_{\mathrm{ES}}^{\mathcal{R}} := \left[-\Delta_{\mathcal{R}}, \Delta_{\mathcal{R}}\right]^N, \tag{37}$$

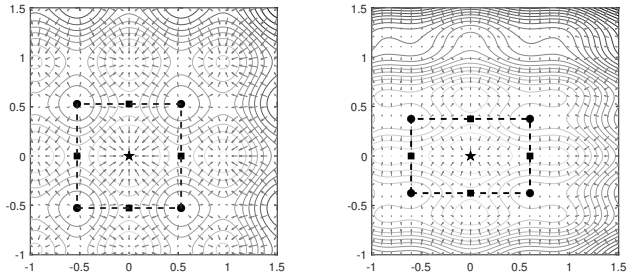

Fig. 6. Global attractor region of the Rastrigin (left) and Bohachevsky (right) function with standard parameters for $N = 2$. Arrows show the negative gradient flow. The star shows the global optimizer, squares the nearest stationary points and circles the farthest stationary points (magnified pictures can be found in supplement Section II-A Fig. 4.)

where $\Delta_{\mathcal{R}}$ is the distance from the global optimizer to the nearest stationary points. It holds

$$\Delta_{\mathcal{R}} \simeq \frac{A\alpha\pi}{A\alpha^2 - 2}. \tag{38}$$

This is visualized for $N = 2$ in the left plot of Fig. 6. For the Bohachevsky function the saddle point $\Delta_{\mathcal{B}}^{(i)}$ next to the global optimizer in coordinate direction $i \notin \{1, N\}$ is given by the 2nd zero of the derivative of (2), i.e.,

$$0 = 6y_i + B_1\beta_1 \sin(\beta_1 y_i) + B_2\beta_2 \sin(\beta_2 y_i) \Leftrightarrow \; y_i = \Delta_{\mathcal{B}}^{(i)}. \tag{39}$$

This nonlinear equation can be approximately solved by expanding the expressions $B\sin(\beta y_i)$ into a Taylor series at $x = \pi/\beta$, i.e., $-\sin(\beta y_i) = -\sin(\beta x) - \beta\cos(\beta x)(y_i - x) + \mathcal{O}(y_i)^2 = \beta y_i - \pi + \mathcal{O}(y_i)^2$ and it follows after a short calculation

$$\Delta_{\mathcal{B}} := \Delta_{\mathcal{B}}^{(i)} = \frac{\pi\left(B_1\beta_1 + B_2\beta_2\right)}{B_1\beta_1^2 + B_2\beta_2^2 - 6}. \tag{40}$$

The estimation of the distances to the nearest stationary point in the first and Nth coordinate directions is similar, i.e.,

$$\Delta_{\mathcal{B}}^{(1)} = \frac{B_1\beta_1\pi}{B_1\beta_1^2 - 2}, \; \Delta_{\mathcal{B}}^{(N)} = \frac{B_2\beta_2\pi}{B_2\beta_2^2 - 4} \tag{41}$$

and the attractor region is a parallelepiped

$$\mathcal{A}_{\mathrm{ES}}^{\mathcal{B}} := \left[-\Delta_{\mathcal{B}}^{(1)}, \Delta_{\mathcal{B}}^{(1)}\right] \times \left[-\Delta_{\mathcal{B}}, \Delta_{\mathcal{B}}\right]^{N-2} \times \left[-\Delta_{\mathcal{B}}^{(N)}, \Delta_{\mathcal{B}}^{(N)}\right]. \tag{42}$$

This is visualized in the right plot of Fig. 6.

For the Ackley function it is not possible to give an exact equation for the attractor region. Only the stationary points that lie on the coordinate axis can be determined analytically. For the stationary points that do not lie on the coordinate axes an analytical calculation is not feasible. In supplement Section II-B, the stationary points on the axes are calculated and it is shown that the attractor region is bounded. The Investigations in Section V-E will show that the frozen noise model is not directly applicable to calculate the success probability. Therefore, a different approach is used which is independent of the attractor size.

To ensure convergence to the global optimizer, the parental centroid should be located in the global attractor region, i.e., $\mathbf{y} \in \mathcal{A}_{\mathrm{ES}}$. Although the attractor regions $\mathcal{A}_{\mathrm{ES}}^{\mathcal{R}}$ (37) or $\mathcal{A}_{\mathrm{ES}}^{\mathcal{B}}$ (42) for the Rastrigin and Bohachevsky function, respectively, can be specified exactly, the following considerations must be taken into account for further investigations: A gradient strategy, starting inside the attractor region is always successful. For an ES this is not necessarily the case. Especially when the parental centroid is located in a corner, the probability to produce a better offspring is about 2^{-N}. On the other hand, when the parental centroid is located outside the attractor but in the vicinity of an edge, the probability of producing an offspring inside the attractor region is approximately[2] $1/2$. This leads to the introduction of a correction term ε that depends on the strategy-specific parameters such as the current mutation strength σ, truncation ratio $\vartheta = \mu/\lambda$, learning parameter τ or cumulation constant c, respectively. The attractor region containing the correction term is defined as

$$\mathcal{A}_{\mathrm{ES}}^{\mathcal{R}}(\varepsilon) := \left[-\Delta_R - \varepsilon, \Delta_R + \varepsilon\right]^N, \tag{43}$$

for the Rastrigin function and

$$\begin{aligned}\mathcal{A}_{\mathrm{ES}}^{\mathcal{B}}(\varepsilon) := &\left[-\Delta_{\mathcal{B}}^{(1)} - \varepsilon, \Delta_{\mathcal{B}}^{(1)} + \varepsilon\right] \\ &\times \left[-\Delta_{\mathcal{B}} - \varepsilon, \Delta_{\mathcal{B}} + \varepsilon\right]^{N-2} \times \left[-\Delta_{\mathcal{B}}^{(N)} - \varepsilon, \Delta_{\mathcal{B}}^{(N)} + \varepsilon\right],\end{aligned} \tag{44}$$

for the Bohachevsky function, respectively. By introducing ε, the attractor model can be used for any ES version.

B. Estimating the Success Probability

As illustrated in the figures of Section III the noise strength σ_{ES} is bounded. Therefore, it can be deduced from (34) and (36) that the probability of reaching the attractor mainly depends on the choice of a sufficiently large population size μ (assuming $\vartheta = \mathrm{const.}$). If the population size is too small, then R_{st} is too large and there is no intersection between the attractor region and the sphere with radius R_{st}. In this case one can assume that global convergence is virtually impossible. On the other hand, if μ is very large such that the R_{st}-sphere lies completely inside the attractor region, every ES reaches the attractor that leads to a success probability near one. For the cases where the attractor region intersects the R_{st}-sphere positive progress is possible with a certain success probability P_{s}.

As has been shown in Section IV-D (35), reaching the steady-state distance R_{st} the single parental components in the Rastrigin and Bohachevsy landscape are approximately normally distributed with zero mean and standard deviation

$$\sigma_{\mathrm{st}} := \frac{R_{\mathrm{st}}}{\sqrt{N}}. \tag{45}$$

For the Rastrigin and Bohachevsky function the attractor region containing the correction term ε can be described by (43) and (44). Taking this into account, one can derive an equation for the success probability, i.e., the probability that the parental centroid is located in the attractor region under the assumption that the single components are normally

[2]Suppose σ is chosen to be sufficiently small to avoid jumping over the attractor region, and sufficiently large to overcome the distance to the attractor region.

distributed with mean zero and standard deviation σ_{st}. Using the independence of the parental centroid components in the steady-state, it follows for the Bohachevsky function

$$\begin{aligned} P_{\mathrm{s}}^{\mathcal{B}} = \Pr\left[\mathbf{y} \in \mathcal{A}_{\mathrm{ES}}^{\mathcal{B}}\right] = \Pr\Bigg[&\left(-\Delta_{\mathcal{B}}^{(1)} - \varepsilon \le y_1 \le \Delta_{\mathcal{B}}^{(1)} + \varepsilon\right) \\ &\wedge \left(-\Delta_{\mathcal{B}} - \varepsilon \le y_2 \le \Delta_{\mathcal{B}} + \varepsilon\right) \wedge \ldots \\ &\wedge \left(-\Delta_{\mathcal{B}}^{(N)} - \varepsilon \le y_N \le \Delta_{\mathcal{B}}^{(N)} + \varepsilon\right)\Bigg] \\ \approx \Pr\left[-\Delta_{\mathcal{B}} - \varepsilon \le y \le \Delta_{\mathcal{B}} + \varepsilon\right]^N . \end{aligned} \tag{46}$$

In the last approximation the influence of the first and last component was neglected, the attractor region is assumed as a hypercube. This simplification has only a small impact on further calculations as it is shown in the supplement Section III-B. Using the fact that the single components are normally distributed, one gets for a single component

$$\begin{aligned} &\Pr\left[-\Delta_{\mathcal{B}} - \varepsilon \le y \le \Delta_{\mathcal{B}} + \varepsilon\right] \\ &\quad = \Pr\left[-\frac{\Delta_{\mathcal{B}} + \varepsilon}{\sigma_{\mathrm{st}}} \le z \le \frac{\Delta_{\mathcal{B}} + \varepsilon}{\sigma_{\mathrm{st}}}\right] \\ &\quad = \Phi\left(\frac{\Delta_{\mathcal{B}} + \varepsilon}{\sigma_{\mathrm{st}}}\right) - \Phi\left(-\frac{\Delta_{\mathcal{B}} + \varepsilon}{\sigma_{\mathrm{st}}}\right) \end{aligned} \tag{47}$$

where $\Phi(z)$ is the cdf of the standard normal variate $\mathcal{N}(0,1)$ and σ_{st} is given by (45). Thus, one gets for the success probability (46)

$$P_{\mathrm{s}}^{\mathcal{B}} = \left[2\Phi\left(\frac{\Delta_{\mathcal{B}} + \varepsilon}{\sigma_{\mathrm{st}}}\right) - 1\right]^N , \tag{48}$$

where σ_{st} depends on R_{st} and therefore on the noise strength $\sigma_{\mathrm{ES}}^{\mathcal{B}}(R_{\mathrm{st}})$. Considering the left plot of Fig. 3, the general course of the curve is that $\sigma_{\mathrm{ES}}^{\mathcal{B}}(R)$ first stays constant for large R followed by a small increase and then starts to drop in the region where the distance R is of the order of $\Delta_{\mathcal{B}}$. That is, even if the ES enters the global attractor region $\mathcal{A}_{\mathrm{ES}}^{\mathcal{B}}$, $\sigma_{\mathrm{ES}}^{\mathcal{B}}(R)$ is still in the vicinity of its maximum value. Therefore, the maximum value $\hat{\sigma}_{\mathrm{ES}}^{\mathcal{B}}$ (26) for the Bohachevsky function is used and it follows from (36) and (45)

$$\sigma_{\mathrm{st}}^{\mathcal{B}} = \sqrt{\frac{\hat{\sigma}_{\mathrm{ES}}^{\mathcal{B}}}{12\mu c_{\mu/\mu,\lambda}}} = \sqrt{\frac{\sqrt{\frac{1}{2}(N-1)(B_1+B_2)^2 - B_1 B_2}}{12\mu c_{\mu/\mu,\lambda}}}. \tag{49}$$

As already mentioned in Section IV, since the global part of the Bohachevsky function is a hyperellipsoid, the standard deviation of the first and last component is larger. However, it can again be assumed that this effect is neglectable for large N, which is demonstrated in more detail in the supplement Section III-B. Inserting this into (48), one finally obtains the success probability in the Bohachevsky case

$$P_{\mathrm{s}}^{\mathcal{B}} = \left[2\Phi\left(\sqrt{\frac{12\mu c_{\mu/\mu,\lambda}}{\hat{\sigma}_{\mathrm{ES}}^{\mathcal{B}}}}\left(\Delta_{\mathcal{B}} + \varepsilon\right)\right) - 1\right]^N , \tag{50}$$

with $\hat{\sigma}_{\mathrm{ES}}^{\mathcal{B}} = \sqrt{\frac{1}{2}(N-1)(B_1+B_2)^2 - B_1 B_2}$. The derivation of an equation for the Rastrigin function is almost analogous and has already been done in [3]. Using (24) and (34) with $a = 1$ the success probability of the Rastrigin function becomes

$$P_{\mathrm{s}}^{\mathcal{R}} = \left[2\Phi\left(\sqrt{\frac{4\sqrt{2}\mu c_{\mu/\mu,\lambda}}{A\sqrt{N}}}\left(\Delta_{\mathcal{R}} + \varepsilon\right)\right) - 1\right]^N . \tag{51}$$

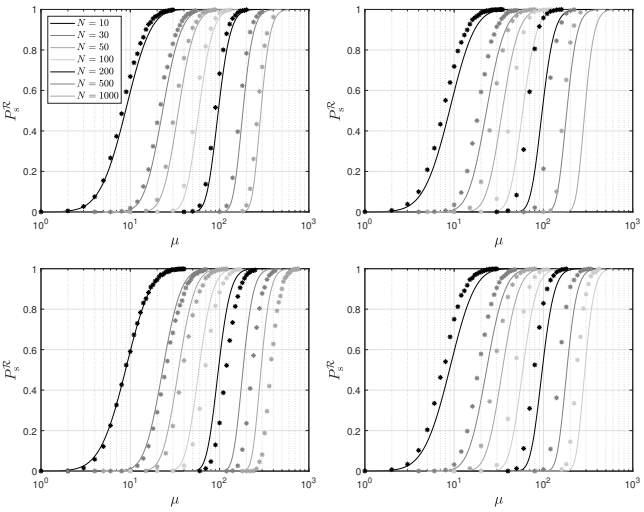

Fig. 7. Success $P_{\mathrm{s}}^{\mathcal{R}}$ depending on population size μ for the Rastrigin landscape. Solid lines show $P_{\mathrm{s}}^{\mathcal{R}}$ predicted by (51) with $\varepsilon = 0$ and $\vartheta = 0.5$. Experimental values are displayed by the stars. From top left to bottom right: CSA-ES with $c = 1/\sqrt{N}$, $c = 1/N$, σSA-ES with $\tau = 1/\sqrt{2N}$ and $\tau = 1/\sqrt{8N}$.

C. Comparison with Experiments

Figures 7 and 8 evaluate the predictive quality of the success probability (51), and (50) using $\Delta_{\mathcal{R}}$ from (38), $\Delta_{\mathcal{B}}$ from (40) and $\varepsilon = 0$. The equations are represented by the solid lines and are compared with experimental values represented by the stars. Each data point was obtained by at least 1000 independent ES runs. The experiments were executed for the σSA-ES and the CSA-ES. As expected, there are differences between the experimental data and the predictions. However, in most cases the general tendencies are well covered by the predictions from (51) and (50). The shape of the curves differs more in the case of the Bohachevsky function. The success probability increases slower than predicted. This is most evident in the case of the σSA-ES with standard choice $\tau = 1/\sqrt{2N}$ where the prediction (50) is not usable.

For both Rastrigin and Bohachevsky, there are considerable differences between the strategies investigated here. The CSA-ES with cumulation constant $c = 1/N$ has a larger success probability than the CSA-ES with the larger cumulation constant $c = 1/\sqrt{N}$. This is also the case for the σSA-ES where the smaller learning parameter $\tau = 1/\sqrt{8N}$ leads to larger success probabilities. For both strategy types, smaller strategy parameters lead to slower adaptation, which is advantageous in multimodal landscapes: The strength of the mutations remains stable longer and sufficiently large so that the probability of being caught in a local minimum is lower. These observations do also hold for other truncation ratios such as $\vartheta = 1/4$ (see supplement Section III-C).

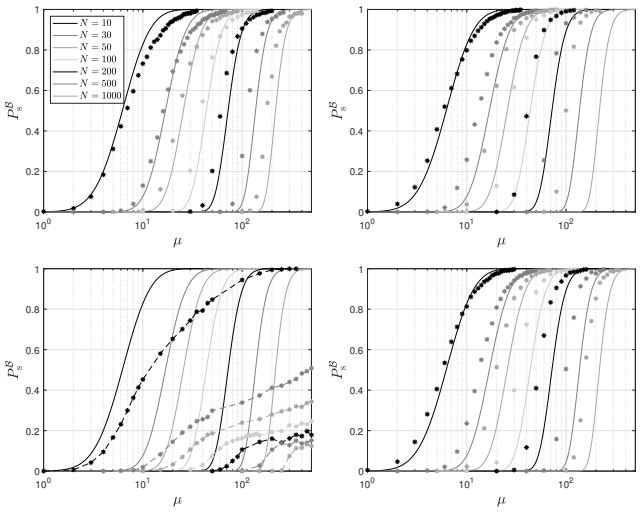

Fig. 8. Success $P_{\mathrm{s}}^{\mathcal{B}}$ depending on population size μ for the Bohachevsky landscape. Solid lines show $P_{\mathrm{s}}^{\mathcal{B}}$ predicted by (50) with $\varepsilon = 0$ and $\vartheta = 0.5$. Experimental values are displayed by the stars. From top left to bottom right: CSA-ES with $c = 1/\sqrt{N}$, $c = 1/N$, σSA-ES with $\tau = 1/\sqrt{2N}$ and $\tau = 1/\sqrt{8N}$.

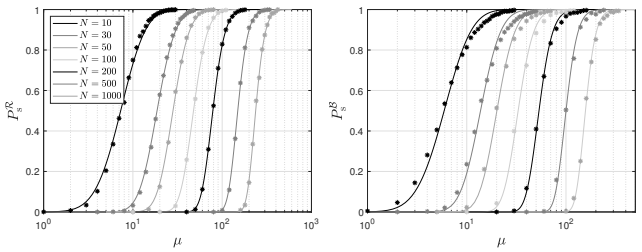

Fig. 9. Success P_{s} predicted by (51) and (50) with ε values according to Fig. 10 for σSA-ES with $\tau = 1/\sqrt{8N}$ and $\vartheta = 0.5$ in the Rastrigin landscape (left plot) and in the Bohachevsky landscape (right plot).

D. Taking Strategy-Specific Properties into Account by ε

Due to the conceptional simplicity of the frozen noise model presented not all ES specific aspects can be modeled. That is, the actual size of the global attractor region depends on strategy-specific aspects such as τ, c, ϑ, etc. To improve the prediction quality, the correction term $\varepsilon \neq 0$ is introduced in (51) and (50). The results for the σSA-ES with $\tau = 1/\sqrt{8N}$ are presented in Fig. 9 where ε was chosen according to Fig. 10. The ε values were determined empirically by minimizing the sum of the squares of the differences between the equations and the experimental values. In the Rastrigin landscape this leads to curves that do well agree with the real ES runs, as visible in the left plot of Fig. 9. For the Bohachevsky function (right plot), there are smaller deviations. Similar results are provided for the other strategy types in the supplement Section III-D.

E. The Ackley Function

Since the size and shape of the global attractor region are not available it is not possible to derive an equation for the success probability as (50). Fig. 11 shows the experimental results for the success probability. Each data point was determined by at least 1000 ES runs. The dependence on N is much weaker than for the Rastrigin or Bohachevsky function. Especially for the CSA-ES with $c = 1/N$, the probabilities become only slightly smaller with increasing N. For the σSA-ES with $\tau = 1/\sqrt{2N}$ and for large N the success probabilities are very low compared to the other strategy types. This is a result of the small initial mutation strength that prevents divergence but leads to

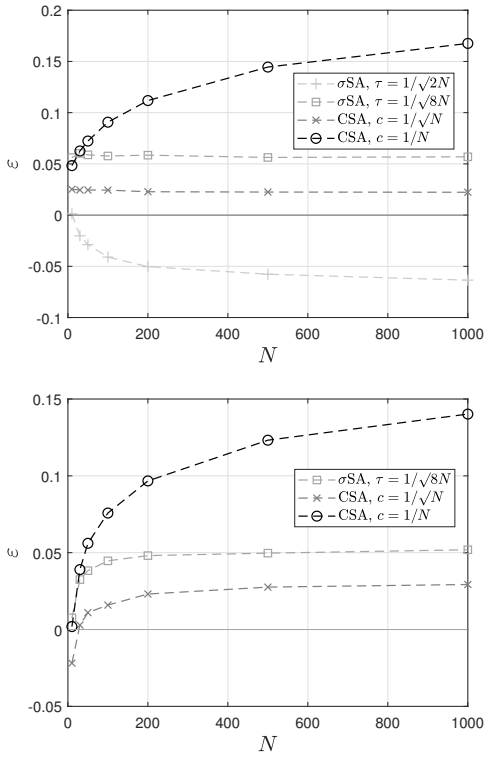

Fig. 10. Optimal ε that minimizes the deviations between (51) and (50) and the experimental values of Figs. 7 and 8. Top plot represents the results for the Rastrigin function and bottom plot for the Boachevsky function, respectively.

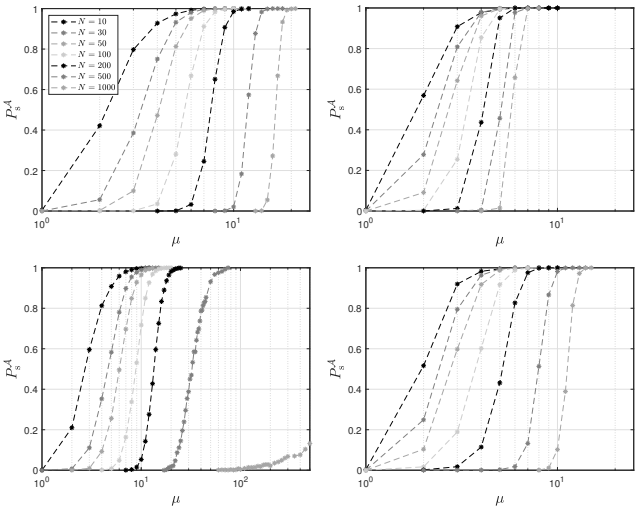

Fig. 11. Success $P_{\mathrm{s}}^{\mathcal{A}}$ depending on population size μ in the Ackley landscape with $\vartheta = 0.5$. Experimental values are displayed by the stars. From top left to bottom right: CSA-ES with $c = 1/\sqrt{N}$, $c = 1/N$, σSA-ES with $\tau = 1/\sqrt{2N}$ and $\tau = 1/\sqrt{8N}$.

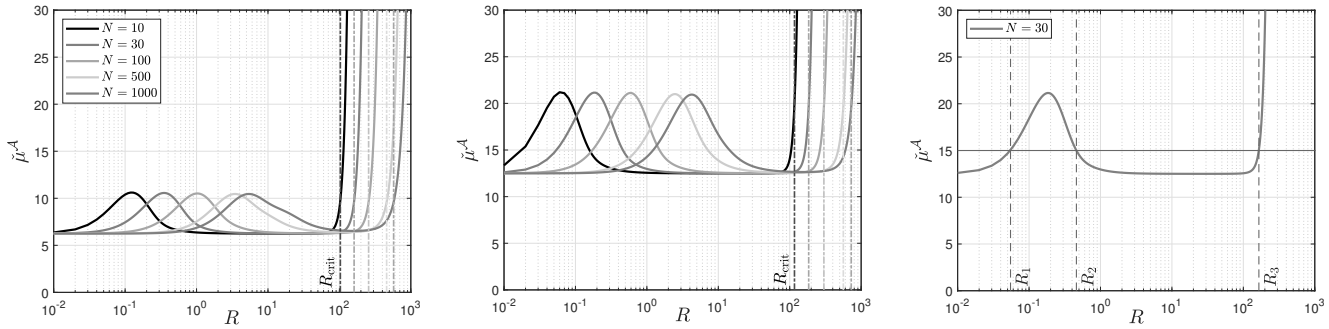

Fig. 12. Solid lines show $\breve{\mu}^{\mathcal{A}}$ from (54) with $\sigma^* = 10$ (left plot) and $\sigma^* = 20$ (middle and right plot). Vertical dashed-dotted lines in left and middle plot show R_{crit} (56).

local convergence in this case. All experiments were executed with moderate initial values $R^{(0)} = 10\sqrt{N}$ and $\sigma^{*(0)} = 3$. Compared to the other multimodal landscapes considered, the initial values in the Ackley landscape have a large influence on the success probability. The initial R distance must not be too large. This can be explained intuitively. Far outside the global part of the Ackley function the landscape is very flat, as visible in Fig. 1. In an almost flat landscape only small progress is possible due to the oscillations resulting in a large noise to signal ratio. Thus, no positive progress is possible and the ES diverges. This is also the case if the normalized mutation strength is too large. For the noisy sphere model this has already been investigated. Therefore, an alternative model approach for the Ackley function can be derived from it, which is described below.

The maximal normalized mutation strength must fulfill the sufficient evolution condition [10]

$$\left(\sigma^*_{\mathrm{ES}}\right)^2 + \left(\sigma^*\right)^2 \leq 4\mu^2 c^2_{\mu/\mu,\lambda}. \tag{52}$$

σ^*_{ES} denotes the normalized noise strength [13]

$$\left(\sigma^{\mathcal{A}}_{\mathrm{ES}}\right)^* = \sigma^{\mathcal{A}}_{\mathrm{ES}}(R)\frac{N}{RG'_{\mathcal{A}}(R)}, \tag{53}$$

where $G'_{\mathcal{A}}$ is the derivative w.r.t. R of the global part (12) of the Ackley function. Resolving the equation in (52) for μ using (53) taking (30), (31), and (32) into account, one gets $\breve{\mu}^{\mathcal{A}}$ above which the ES can converge

$$\breve{\mu}^{\mathcal{A}} = \frac{\sqrt{\left[\left(\sigma^{\mathcal{A}}_{\mathrm{ES}}\right)^*\right]^2 + \left(\sigma^*\right)^2}}{2c_{\mu/\mu,\lambda}}. \tag{54}$$

This minimal parental population size is displayed in Fig. 12. The $\breve{\mu}^{\mathcal{A}}$ curves reveal a very interesting convergence behavior. For example, consider the $N = 30$, $\sigma^* = 20$ curve in the middle plot and draw a horizontal line at $\mu = 15$. One obtains three intersection points, $R_1 < R_2 < R_3$. These points are displayed in the right plot of Fig. 12. If the parental centroid of the ES is at a distance $R < R_1$ then the ES converges to the global optimizer. If R is located between R_1 and R_3, the ES will converge to R_2, i.e., it ends in a local attractor. If, however, $R > R_3$ the ES is expected to diverge. That is, the increase of the necessary population size $\breve{\mu}^{\mathcal{A}}$ beyond the R_3 value is almost exponential w.r.t. the R value: Using (52) with $\sigma^* = 0$ and (53) with (12) and (33), one finds

$$\mu \geq \frac{N}{2\sqrt{2}c_{\mu/\mu,\lambda}RC_1C_2}\mathrm{e}^{\frac{C_2}{\sqrt{N}}R}. \tag{55}$$

Considering the curves for different search space dimensionalities N in Fig. 12 (left and middle), it becomes apparent that the value of the local maximum $\max(\breve{\mu}^{\mathcal{A}})$ of the $\breve{\mu}^{\mathcal{A}}$ curves is rather insensitive w.r.t. N [3]. This means that choosing a $\mu > \max(\breve{\mu}^{\mathcal{A}})$ and an initial R being less than a critical initial R_{crit}, the global convergence is ensured independent of N. This is in contrast to the Rastrigin and Bohachevsky case.

To calculate the thresholds R_{crit} and σ^*_{crit} beyond which convergence becomes unattainable, insert the normalized noise strength (53) with (33) into the evolution condition (52). The resulting nonlinear equation can be solved approximately, yielding

$$R_{\mathrm{crit}} = \frac{x_{\mathrm{crit}}\sqrt{N}}{2C_2} \tag{56}$$

$$\sigma^*_{\mathrm{crit}} = \sqrt{4\mu^2 c^2_{\mu/\mu,\lambda} - \frac{1}{2}\left(\frac{N}{RC_1C_2}\right)^2 \mathrm{e}^{2C_2\frac{R}{\sqrt{N}}}}, \tag{57}$$

where $x_{\mathrm{crit}} \approx x^{(k)}$ is obtained recursively with

$$x^{(0)} = \ln\left(\frac{C_1\left(4\mu^2 c^2_{\mu/\mu,\lambda} - \left(\sigma^*\right)^2\right)}{2N}\right) \tag{58}$$

$$x^{(k)} = x^{(0)} + 2\ln(x^{(k-1)}). \tag{59}$$

Comprehensive calculations and experimental results are represented in supplement Section IV. Equation (56) is represented by the dashed lines in Fig. 12 (left and middle). It lies in the region where $\breve{\mu}^{\mathcal{A}}$ increases strongly.

The severe influence of the initial distance to the optimizer on the convergence behavior gives rise to the question how this value is usually chosen in test beds for the empirical evaluation of evolutionary algorithms. The start value for the CMA-ES in [1] is randomly initialized within the interval $[1, 30]^N$. The initial mutation strength σ is set to half of the initialization

[3]This N-independence also occurs when varying the Ackley parameters or ϑ.

interval. Numerous other papers have adopted these starting values. R_{crit} (56) can fall within this initialization interval. For example, for $N = 100$, $\mu = 10$, and $\vartheta = 0.5$, it holds $R_{\mathrm{crit}} \approx 270 < 300$, where 300 is the maximum distance in the initialization interval. Using Jensen's inequality, it follows for the norm of the vector with uniformly distributed components in $[a, b]$, that

$$\mathrm{E}\left[\sqrt{\sum_{i=1}^{N} X_i^2}\right] \leq \sqrt{\mathrm{E}\left[\sum_{i=1}^{N} X_i^2\right]} = \sqrt{\sum_{i=1}^{N} \mathrm{E}\left[X_i^2\right]} = \sqrt{N \frac{b^3 - a^3}{3(b-a)}} = \sqrt{\frac{N}{3}\left(a^2 + ab + b^2\right)}. \quad (60)$$

The last step uses the second moment of the uniform distribution. Using the above mentioned values, the expected start distance is less than 176. For these parameters and for the CSA-ES with $c = 1/\sqrt{N}$, the runs already diverge from a starting distance of 150.

Swarm optimization algorithms often use a limited search space of $[-32.768, 32.768]^N$ and an initialization interval of $[16.384, 32.768]^N$ [14], $[-32.768, 32.768]^N$ [15], or $[-5, 5]^N$ [16]. Considering these values one might speculate that these initializations have been chosen within or at the edge of global convergence. That is, these choices are not suited to show the limitations of the algorithms. Therefore, test beds should be adopted to also show the limitations.

VI. Population Sizing

A. Derivation of Parent Population Size

A main question of this paper concerns the choice of μ and λ that guarantees convergence of the ES towards the global optimizer. As already mentioned in Section V-B the success probability mainly depends on the choice of the population size μ. A large μ leads to a small R_{st} and the R_{st}-sphere intersects the attractor region. Given a fixed truncation ratio ϑ, it suffices to derive a formula that predicts $\mu(P_{\mathrm{s}})$. It was shown in Section V-D that (50), (51) in conjunction with the correction term ε predict the experimental data well. Therefore, the equations can be used to derive a population sizing formula and to evaluate its scaling behavior. Solving (51) for μ under the assumption $c_{\mu/\mu,\lambda} \simeq f(\vartheta)$ [13, p. 249] yields for the Rastrigin function

$$\mu^{\mathcal{R}} \simeq \frac{\hat{\sigma}^{\mathcal{R}}_{\mathrm{ES}}}{(\Delta_{\mathcal{R}} + \varepsilon)^2 4 c_{\mu/\mu,\lambda}} \left[\Phi^{-1}\left(\frac{1}{2} + \frac{1}{2} P_{\mathrm{s}}^{\frac{1}{N}}\right)\right]^2, \quad (61)$$

where Φ^{-1} is the quantile function of the standard normal distribution. For the Bohachevsky function it follows from (50)

$$\mu^{\mathcal{B}} \simeq \frac{\hat{\sigma}^{\mathcal{B}}_{\mathrm{ES}}}{(\Delta_{\mathcal{B}} + \varepsilon)^2 12 c_{\mu/\mu,\lambda}} \left[\Phi^{-1}\left(\frac{1}{2} + \frac{1}{2} P_{\mathrm{s}}^{\frac{1}{N}}\right)\right]^2. \quad (62)$$

(61) and (62) with $\varepsilon = 0$ behave asymptotically like [3]

$$\mu = \mathcal{O}\left(\sqrt{N}\ln(N)\right). \quad (63)$$

As there is no analytical characterization of the success domain for the Ackley function, it is not possible to provide a μ-dependent equation for the success probability. However, (54) describes the minimal value of μ above which an ES can converge. Fig. 12 illustrates that there is a local maximum for $R < R_{\mathrm{crit}}$. It follows that

$$\mu^{\mathcal{A}} = \max_R \left[\breve{\mu}^{\mathcal{A}}(R) \mid R < R_{\mathrm{crit}}\right] \quad (64)$$

is an upper bound for the population size required to achieve a success rate of 1. Fig. 12 also shows that $\mu^{\mathcal{A}}$ strongly depends on σ^*. On the other hand, the value of the steady-state σ^* depends on μ and on strategy-specific parameters. Therefore, within the current model it is not possible to provide a general equation for $\mu^{\mathcal{A}}$. A value for σ^* must be assumed in order to determine $\mu^{\mathcal{A}}$.

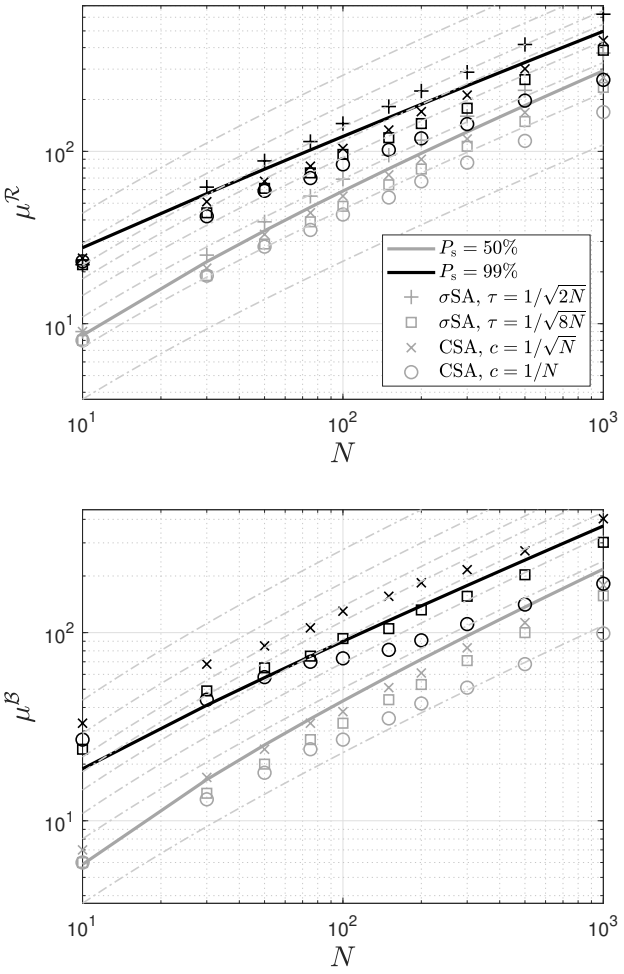

Fig. 13. Population size for $P_{\mathrm{s}} = 50\%$ (lighter curve and markers) and $P_{\mathrm{s}} = 99\%$ (darker curve and markers). Upper plot shows the results for Rastrigin and lower plot for Bohachevsky, both with $\vartheta = 0.5$. Bold lines show (61) and (62) and symbols represent the experimental results. Gray dashed-dotted lines show functions $\propto \sqrt{N}\ln(N)$ for comparison.

B. Comparison with Experiments

Figure 13 compares the predictions of the population size equations depending on N with experiments. At least 300 runs were executed to obtain the experimental data displayed by the markers. The solid lines represent the population sizing equations (61) and (62) with $\varepsilon = 0$ for $P_{\mathrm{s}} = 50\%$ and

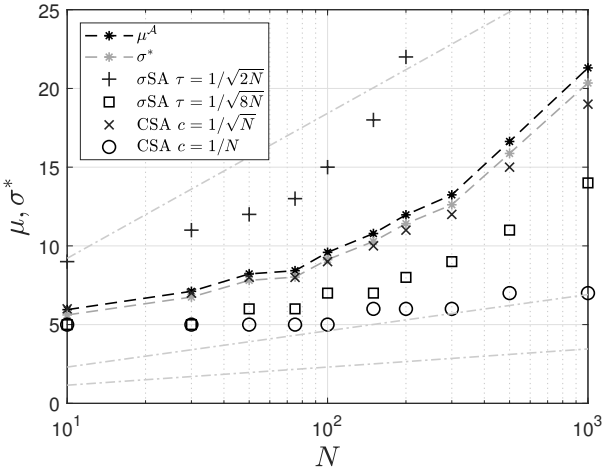

Fig. 14. Population size μ_{99} to reach $P_\mathrm{s} = 99\%$ in the Ackley landscape with $\vartheta = 0.5$ and initialization values $R^{(0)} = 10\sqrt{N}$ and $\sigma^{*(0)} = 3$. Grey stars with dashed lines show σ^* experimentally determined for the CSA-ES with $c = 1/\sqrt{N}$ and $\mu = \mu_{99}$. Black stars show $\mu^{\mathcal{A}}$ (64) using this σ^*. Gray dashed-dotted lines show functions $\propto \ln(N)$ for comparison.

for $P_\mathrm{s} = 99\%$. While the theoretical predictions differ from the experimental values, (61) and (62) predict the general functional tendency well.

The deviations are due to the different values of ε, see Fig. 10. For the CSA-ES with $c = 1/N$, ε is positive and increases as N increases. This results in an overestimation of the population size. The scaling behavior is slower than $\sqrt{N}\ln(N)$, which is consistent with the increasing ε values. For the σSA-ES with $\tau = 1/\sqrt{2N}$ in the Rastrigin landscape the population size is underestimated in accordance with the negative values of ε. The growth rate is slightly above $\sqrt{N}\ln(N)$ for $P_\mathrm{s} = 0.5$ as predicted from the decreasing values of ε. The growth rate for $P_\mathrm{s} = 0.99$ differs slightly from the predicted behavior. For both the CSA-ES with $c = 1/\sqrt{N}$ and the σSA-ES with $\tau = 1/\sqrt{8N}$ the values of ε are positive and approximately constant with N. The growth rate is approximately $\sqrt{N}\ln(N)$. The predicted values from the equations are an upper bound for sufficiently large N, except for the Bohachevsky landscape for $P_\mathrm{s} = 0.99$. In this case the scaling behavior differs. For Bohachevsky and $P_\mathrm{s} = 0.99$, the experimental values display fluctuations, whereas the values for $P_\mathrm{s} = 0.5$ follow a rather smooth course. This discrepancy can be explained by the fact that the increase of P_s within the vicinity of $P_\mathrm{s} \approx 0.99$ becomes weaker, see Fig. 8, and that the data was acquired from a small number of experiments.

In the case of the Ackley function, only an upper bound for the population size can be given. This means that the success probability is expected to approach 1 when μ is larger as predicted by (64). The dashed line with black stars in Fig. 14 shows (64), where σ^* was determined experimentally for the CSA-ES with $c = 1/\sqrt{N}$ and $\mu = \mu_{99}$. The values of σ^*, $\mu^{\mathcal{A}}$, and μ_{99} are almost identical and show the same scaling behavior. When using other strategy types or varying ϑ or the Ackley parameters, differences may be larger. In all cases, the scaling behaviors of σ^* and $\mu^{\mathcal{A}}$ is approximately but slightly larger than μ_{99}.

Except for the σSA-ES with $\tau = 1/\sqrt{2N}$, the experimental values increase slower than linearly. For the σSA-ES with $\tau = 1/\sqrt{2N}$, the values of μ are significantly larger and increase faster than linearly. As demonstrated in Section V-E, the choice of suitable initial values depends on the strategy-specific parameters. By choosing different initial values, this anomalous behavior can be avoided and a sublinear scaling behavior can also be achieved.

VII. Conclusions and Outlook

Each of the multimodal functions considered in this work can be divided into a global part, where the global optimizer is the only minimum, and a residual part with several local minima. The latter part can be interpreted as frozen noise. By applying the noisy sphere theory, a model has been developed for the calculation of the success probability P_s to reach the global optimizer of the multimodal function. P_s depends on the population size parameters μ and λ. For a noisy sphere with a fixed noise strength, an ES with a fixed population size cannot reach the optimizer arbitrarily close, but its parents fluctuate about the optimizer with an expected distance R_st. These fluctuations can be interpreted as a global search behavior. If R_st is sufficiently small the parents hit the global attractor region. In this region the frozen noise wanes and the ES converges to the global optimizer. This convergence model is independent of specific ES strategy parameters and therefore applicable to different types of ES such as σSA or CSA-ES. The model was applied to the Rastrigin and Bohachevsky function, with the result that the parental population size needed to achieve a satisfactory convergence probability scales approximately with $\mathcal{O}(\sqrt{N}\ln(N))$.

Even though the Bohachevsky test function formula appears more complex than that of Rastrigin, the analysis in this paper revealed strong similarities w.r.t. the population sizing behavior. Therefore, from viewpoint of testbed design for empirical performance evaluations it suffices to include only one of these test functions. This should be taken into account when proposing new benchmark competitions. Similar considerations do also hold for the Ackley function.

As for the Ackley function, the frozen noise model does not allow for the determination of the success probability as there is no analytical description for the global attractor domain. Therefore, a different approach was used.

An upper bound for the population size μ was derived that mainly depends on σ^*. Since σ^* increases with N, the upper bound for the population size also depends on N, but only weakly. This is in contrast to the the Rastrigin or Bohachevsky case. However, convergence depends strongly on the initialization. If the start population is located too far from the global optimizer the ES will only converge locally or diverge. This dependency from the initial conditions must be taken more into account when running empirical performance comparisons of evolutionary algorithms. The standard initializations found in literature appear to ensure convergence with a reasonable success probability for small population sizes (see Fig. 11). From this perspective the standard Ackley function is not

a hard test function. However, it could be made harder by decreasing the C_2 parameter in (3). This would elevate the local $\check{\mu}$ maximum in Fig. 12.

In addition to the multimodal functions considered, there is Griewank's function that can also be divided into a global part and a residual part. However, the Griewank function is an exception in the class of these multimodal functions because the success probability increases with N. The main reason for this counterintuitive behavior is the fast vanishing of the noise variance with N. That is, the problem gets easier with increasing N. From the viewpoint of analysis, a second peculiarity seems to complicate the application of the model additionally. The noise cannot be well approximated by a Gaussian distribution. Both peculiarities make the analysis in the interesting low-dimensional N cases appear questionable: The model developed in this paper is based on $N \to \infty$ asymptotics. Thus, yielding only an approximation for small N that does not perform well for N of the order of 10 or below. Therefore, the Griewank case needs perhaps another approach and remains as a topic for another paper.

The investigations presented in this paper do also contribute to a better understanding of how ES perform search and locate the optimizers in highly multimodal fitness landscapes. Contrary to popular belief, the ES does not follow a gradient-like path. The global optimizer is approached from all sides. That is, if correctly designed the ES performs a global search. However, this also means that there must be a global structure of the fitness landscape that guides the search of the ES in a highly noisy environment.

The analysis of ES in highly multimodal fitness landscapes is still in its very first beginning. One may question whether the results presented and methods developed are useful for real-world problems. There is no definite answer, yet. This question defines a research program on its own. Do the multimodal test functions to be found in the BBOB and CEC competitions really reflect real-world problems? Furthermore, progress in a theory-driven analysis will be always gradual. A possible direction could regard multimodal test functions where the global part cannot be well approximated by a sphere model. Also the performance of the ES on dynamically moving optimizer problems could define a research direction. And on the algorithm engineering side, the question of controlling the population size efficiently deserves attention.

REFERENCES

[1] N. Hansen and S. Kern, "Evaluating the CMA Evolution Strategy on Multimodal Test Functions," *Parallel Problem Solving from Nature*, vol. 8, pp. 282–291, 2004.

[2] N. Hansen, S. Müller, and P. Koumoutsakos, "Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES)," *Evolutionary Computation 11*, no. 1, p. 1–18, 2003.

[3] 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.

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

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

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

[7] A. Omeradzic and H.-G. Beyer, "Progress Rate Analysis of Evolution Strategies on the Rastrigin Function: First Results," in *Parallel Problem Solving from Nature XXVII*, H. Aguirre et al., Ed. Berlin: Springer, 2022.

[8] ——, "Progress analysis of a multi-recombinative evolution strategy on the highly multimodal Rastrigin function," *Theoretical Computer Science*, vol. 978, p. 114179, 2023.

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

[10] H.-G. Beyer, D. Arnold, and S. Meyer-Nieberg, "A New Approach for Predicting the Final Outcome of Evolution Strategy Optimization under Noise," *Genetic Programming and Evolvable Machines*, vol. 6, no. 1, pp. 7–24, 2005.

[11] H.-G. Beyer, "On the "Explorative Power" of ES/EP-like Algorithms," in *Evolutionary Programming VII: Proceedings of the Seventh Annual Conference on Evolutionary Programming*, V. Porto, N. Saravanan, D. Waagen, and A. Eiben, Eds. Heidelberg: Springer-Verlag, 1998, pp. 323–334, DOI: 10.1007/BFB0040785.

[12] 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.

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

[14] P. J. Angeline, "Evolutionary optimization versus particle swarm optimization: Philosophy and performance differences," in *International conference on evolutionary programming*. Springer, 1998, pp. 601–610.

[15] G. Zhu and S. Kwong, "Gbest-guided artificial bee colony algorithm for numerical function optimization," *Applied mathematics and computation*, vol. 217, no. 7, pp. 3166–3173, 2010.

[16] B. Niu and H. Wang, "Bacterial colony optimization," *Discrete Dynamics in Nature and Society*, vol. 2012, 2012.

Lisa Schönenberger received the Diploma degree in Mathematics from the University of Vienna, Austria, in 2010 and the Bachelor's degree in Computer Science from the FernUniversität Hagen, Germany, in 2020. Since 2021, she is employed as a research assistant in the field of evolutionary algorithms at the Vorarlberg University of Applied Sciences, Austria. Her current research interests include evolutionary algorithms, multimodal optimization, and restart strategies.

Hans-Georg Beyer is best known for his theoretical analyses and the design of Evolution Strategies based on the stochastic dynamical systems approach. Dr. Beyer received the Diploma degree in Theoretical Electrical Engineering from the Ilmenau Technical University, Germany, in 1982 and the Ph.D. in physics from Bauhaus-University Weimar, Weimar, Germany, in 1989, and the Habilitation degree in computer science from the University of Dortmund, Dortmund, Germany, in 1997. Since 2004 he has been professor with the Vorarlberg University of Applied Sciences, Dornbirn, Austria. He authored the book "The Theory of Evolution Strategies" (Heidelberg: Springer-Verlag, 2001) and authored/coauthored numerous papers in that field. Dr. Beyer was the Editor-in-Chief of the MIT Press Journal "Evolutionary Computation" and served as an Associate Editor for the IEEE "Transactions on Evolutionary Computation" from 1997 to 2021.

Supplementary Document of "On a Population Sizing Model for Evolution Strategies in Multimodal Landscapes"

Lisa Schönenberger and Hans-Georg Beyer

This is the supplementary material to the paper "On a Population Sizing Model for Evolution Strategies in Multimodal Landscapes", published in IEEE Transactions on Evolutionary Computation. This material provides all the detailed calculations. It also contains additional plots and experiments.

I. Distribution of the Residual Parts

The standard deviation of the residual part is denoted with σ_{ES}, i.e.,

$$\sigma_{\mathrm{ES}} = \sqrt{\mathrm{Var}\left[C\right]} \tag{1}$$

where C denotes the residual part of a multimodal function. The first step in the derivation of σ_{ES} is the calculation of the variance of $\cos\left(x\right)$ where $x \sim \mathcal{N}\left(m, s^2\right)$. According to [1, p.406], the characteristic function of x using Euler's formula is

$$\begin{aligned}\mathrm{E}[\mathrm{e}^{\imath t x}] &= \exp\left(\imath m t - \frac{1}{2}s^2t^2\right)\\ &= \exp\left(-\frac{1}{2}s^2t^2\right)\left(\cos(mt) + \imath\sin(mt)\right)\\ &= \mathrm{E}[\cos(tx)] + \imath\mathrm{E}[\sin(tx)].\end{aligned} \tag{2}$$

Comparing the real parts it follows for $t = 1$

$$\mathrm{E}[\cos(x)] = \exp\left(-\frac{1}{2}s^2\right)\cos(m). \tag{3}$$

Using the identity $\cos(x)^2 = \frac{1}{2} + \frac{1}{2}\cos(2x)$, one gets

$$\begin{aligned}\mathrm{E}[\cos(x)^2] &= \frac{1}{2} + \frac{1}{2}\mathrm{E}[\cos(2x)]\\ &= \frac{1}{2} + \frac{1}{2}\exp\left(-\frac{1}{2}(2s)^2\right)\cos(2m).\end{aligned} \tag{4}$$

Taking the identity $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$ into account and use (3) and (4), one obtains

$$\begin{aligned}\mathrm{Var}[\cos(x)] =& \frac{1}{2} + \frac{1}{2}\exp\left(-\frac{1}{2}(2s)^2\right)\cos(2m)\\ &- \exp\left(-s^2\right)\left(\frac{1}{2} + \frac{1}{2}\cos(2m)\right)\\ =& \frac{1}{2}\left(1 - \mathrm{e}^{-s^2}\right)\left(1 - \mathrm{e}^{-s^2}\cos(2m)\right)\end{aligned} \tag{5}$$

Each offspring individual $\tilde{\mathbf{y}}_l$ in the CSA-ES is generated from the parent $\mathbf{y}$ with the same mutation strength σ (see Alg. 2 from main paper). It holds $\tilde{y}_i = y_i + \sigma z_i$, where $z_i \sim \mathcal{N}(0,1)$ and therefore $\alpha\tilde{y}_i \sim \mathcal{N}(\alpha y_i, (\alpha\sigma)^2)$. Inserting this into (5) it follows

$$\begin{aligned}\mathrm{Var}[\cos(\alpha\tilde{y}_i)] = &\frac{1}{2}\left(1 - \mathrm{e}^{-(\alpha\sigma)^2}\right)\\ &\times\left(1 - \mathrm{e}^{-(\alpha\sigma)^2}\cos(2\alpha y_i)\right).\end{aligned} \tag{6}$$

A. Rastrigin Function

The residual part of the Rastrigin function is

$$C_{\mathcal{R}}(\mathbf{y}) := NA - A\sum_{i=1}^{N}\cos\left(\alpha y_i\right), \tag{7}$$

and due to the stochastic independence of the z_i, it follows

$$\begin{aligned}\mathrm{Var}\left[C_{\mathcal{R}}\right] &= \mathrm{Var}\left[A\sum_{i=1}^{N}\cos(\alpha\tilde{y}_i)\right]\\ &= A^2\sum_{i=1}^{N}\mathrm{Var}\left[\cos(\alpha y_i + \alpha\sigma z_i)\right]\\ &= A^2\frac{N}{2}\left(1 - \mathrm{e}^{-(\alpha\sigma)^2}\right)\\ &\times\left(1 - \mathrm{e}^{-(\alpha\sigma)^2}\frac{1}{N}\sum_{i=1}^{N}\cos(2\alpha y_i)\right).\end{aligned} \tag{8}$$

This equation depends on the current location of the parent $\mathbf{y}$. The components y_i in the variance expression (8) can be removed by an expected value operation. As has been explained in Section IV-C of the main paper, it holds under steady-state conditions that the single parental components are normally distributed with mean 0 and variance R^2/N. Therefore, it follows from (3) that

$$\mathrm{E}[\cos(2\alpha y_i)] = \mathrm{e}^{-\frac{1}{2}\frac{(2\alpha R)^2}{N}}\cos(0). \tag{9}$$

Inserting this in (8) it follows,

$$\begin{aligned}\mathrm{E}\left[\mathrm{Var}\left[C_{\mathcal{R}}\right]\right] &= A^2\frac{N}{2}\left(1 - \mathrm{e}^{-(\alpha\sigma)^2}\right)\\ &\times\left(1 - \mathrm{e}^{-(\alpha\sigma)^2}\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left[\cos(2\alpha y_i)\right]\right)\\ &= A^2\frac{N}{2}\left(1 - \mathrm{e}^{-(\alpha\sigma)^2}\right)\\ &\times\left(1 - \mathrm{e}^{-(\alpha\sigma)^2}\frac{1}{N}\sum_{i=1}^{N}\mathrm{e}^{-2\frac{(\alpha R)^2}{N}}\right)\end{aligned} \tag{10}$$

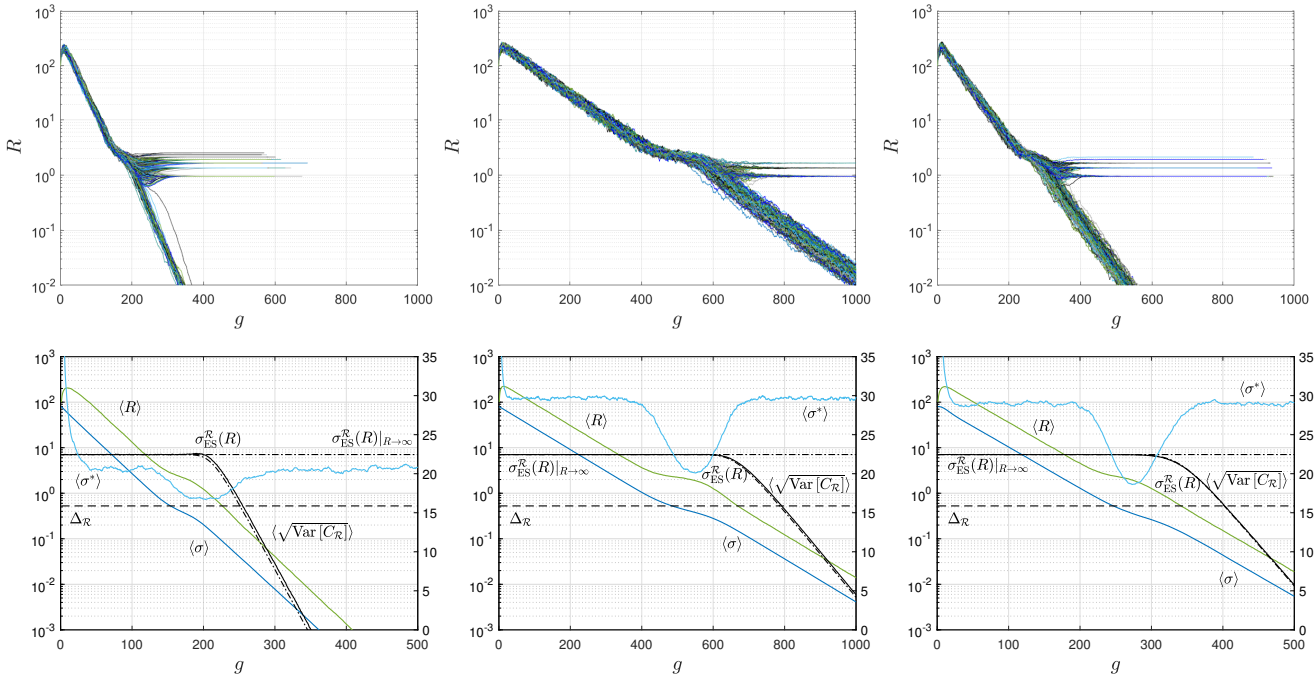

Fig. 1. Residual distance dynamics and mean value dynamics in the Rastrigin landscape for $N = 100$. Left plots: $(50/50_I, 100)$-σSA-ES with $\tau = 1/\sqrt{2N}$ and $P_{\mathrm{s}} = 0.12$, middle plots: $(50/50_I, 100)$-σSA-ES with $\tau = 1/\sqrt{8N}$ and $P_{\mathrm{s}} = 0.6$, right plots: $(50/50_I, 100)$-CSA-ES with $c = 1/\sqrt{N}$ and $P_{\mathrm{s}} = 0.38$. The green line shows the residual distance R, the blue line shows the mutation strength σ and the light blue line its normalization σ^*. The black solid line shows the experimentally determined values for the standard deviation of the residual part. The dashed-dotted curve shows (11) the theoretical calculation for the standard deviation of the residual part $\sigma^{\mathcal{R}}_{\mathrm{ES}}(R)$ while the dashed-dotted straight line shows its asymptotic behavior $\sigma_{\mathrm{ES}}(R)\mid_{R\to\infty}$ (12). σ^* is represented by the right y-axis, all other values are represented by the left y-axis.

and finally one gets an equation for $\sigma_{\mathrm{ES}}(R)$ that only depends on the distance R to the global optimizer, i.e.,

$$\begin{aligned}\sigma^{\mathcal{R}}_{\mathrm{ES}}(R) &= \sqrt{\mathrm{Var}\,[C_{\mathcal{R}}]} \\ &= A\sqrt{\frac{N}{2}}\sqrt{\left(1-\mathrm{e}^{-(\alpha\sigma)^2}\right)\left(1-\mathrm{e}^{-(\alpha\sigma)^2}\mathrm{e}^{-2\frac{(\alpha R)^2}{N}}\right)}. \end{aligned} \quad (11)$$

If the distance to the global optimizer R is very large, then the mutation strength σ is also very large, and the exponential terms in (11) vanish. Therefore, the asymptotic behavior of σ_{ES} is

$$\sigma_{\mathrm{ES}}(R)\mid_{R\to\infty} = A\sqrt{\frac{N}{2}}. \quad (12)$$

For the CSA-ES this is also the maximum value. For the σSA-ES each offspring $\tilde{\mathbf{y}}_l$ is generated by an individual $\tilde{\sigma}_l$ (see Alg. 1 from main paper). The variation of σ results in covariances, which leads to a larger standard deviation and complicates the calculations not considered here. (11) is represented in Fig. 1 by the black dashed dotted curves. The values for σ were determined experimentally. The straight, dashed-dotted lines show (12). The black solid line represents the experimental results. In the bottom plot for the CSA-ES, the differences between (11) and the experimental results are nearly not visible. In the left plot for the σSA-ES with $\tau = 1/\sqrt{2N}$, there are small deviations between the theoretical and experimental values. There is a region where the standard deviation is slightly larger than the theoretical maximum from (12). Slight deviations can also be seen for the σSA-ES with $\tau = 1/\sqrt{8N}$ in the middle plot, but these are much smaller than those for $\tau = 1/\sqrt{2N}$. The extent of the influence of this deviation on further calculations is examined in Section III.

B. Bohachevsky Function

The residual part of the Bohachevsky function is

$$\begin{aligned}C_{\mathcal{B}}(\mathbf{y}) := \sum_{i=1}^{N-1}&[B_1 - B_1\cos(\beta_1 y_i) \\ &+ B_2 - B_2\cos(\beta_2 y_{i+1})]. \end{aligned} \quad (13)$$

The calculation of $\sigma^{\mathcal{B}}_{\mathrm{ES}}(R)$ is similar to Rastrigin, but there are additional covariance terms, i.e., although the individual mutations z_i are independent of each other, they occur twice for components $i = 2, \ldots, N-1$ in the sum of (13). Using the identity

$$\begin{aligned}\mathrm{Var}\,[aX + bY] &= a^2\mathrm{Var}\,[X] \\ &+ b^2\mathrm{Var}\,[Y] + 2ab\mathrm{Cov}\,[X, Y] \end{aligned} \quad (14)$$

for any random variables X and Y it follows for a parent $\mathbf{y}$

$$\begin{aligned}\mathrm{Var}\,[C_{\mathcal{B}}(\mathbf{y})] = B_1^2\sum_{i=1}^{N-1}&\mathrm{Var}\,[\cos(\beta_1\tilde{y}_i)] \\ &+ B_2^2\sum_{i=2}^{N}\mathrm{Var}\,[\cos(\beta_2\tilde{y}_i)] \\ &+ 2B_1B_2\sum_{i=2}^{N-1}\left(\mathrm{Cov}\,[\cos(\beta_1\tilde{y}_i), \cos(\beta_2\tilde{y}_i)]\right). \end{aligned} \quad (15)$$

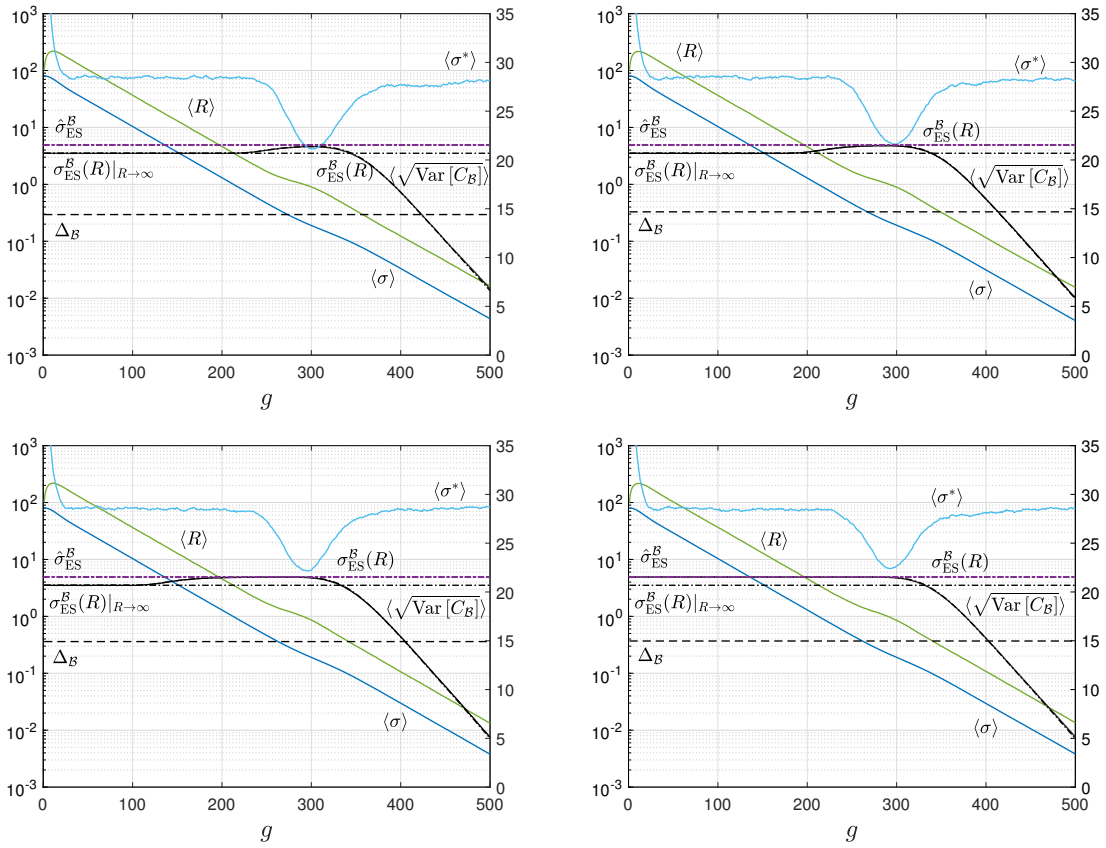

Fig. 2. Mean value dynamics for the $(50/50_I, 100)$-CSA-ES with $c = 1/\sqrt{N}$ in the Bohachevsky landscape with $B_1 = 0.3,\ B_2 = 0.4,\ \beta_1 = 3\pi$, and $N = 100$. From top left to bottom right: $\beta_2 = 4\pi,\ P_{\mathrm{s}} = 0.77$, $\beta_2 = 3.5\pi,\ P_{\mathrm{s}} = 0.79$, $\beta_2 = 3.1\pi,\ P_{\mathrm{s}} = 0.81$, and $\beta_2 = 3\pi,\ P_{\mathrm{s}} = 0.85$, respectively. The green line shows the residual distance R, the blue line shows the mutation strength σ and the light blue line its normalization σ^*. The black solid line shows the experimentally determined values for the standard deviation of the residual part. The dashed-dotted curve shows $\sigma^{\mathcal{B}}_{\mathrm{ES}}(R)$ the theoretical calculation for the standard deviation of the residual parts given by the square root of (22) and (25) for the bottom right plot, respectively. The black dashed-dotted straight line shows $\sigma^{\mathcal{B}}_{\mathrm{ES}}(R)\,|_{R\to\infty}$ (23) and the upper dashed-dotted purple line shows $\hat{\sigma}^{\mathcal{B}}_{\mathrm{ES}}$ (26). $\Delta_{\mathcal{B}}$ shows (53). σ^* is represented by the right y-axis, all other values are represented by the left y-axis.

where $\tilde{y}_i = y_i + \sigma z_i$ with $z_i \sim \mathcal{N}(0,1)$ is the single component of an offspring individual. The calculation of the first two terms is analogous to Rastrigin. Using (6) and (9) it follows for a parent $\mathbf{y}$, which is located on the R-sphere

$$B_1^2 \sum_{i=1}^{N-1} \mathrm{Var}\left[\cos\left(\beta_1 \tilde{y}_i\right)\right] = \frac{B_1^2(N-1)}{2}\left(1 - \mathrm{e}^{-\beta_1^2\sigma^2}\right) \times \left(1 - \mathrm{e}^{-\beta_1^2\sigma^2}\mathrm{e}^{-2\beta_1^2\frac{R^2}{N}}\right). \quad (16)$$

Taking the identity $\mathrm{Cov}\,[XY] = \mathrm{E}\,[XY] - \mathrm{E}\,[X]\,\mathrm{E}\,[Y]$ into account one gets for the covariance part in (15)

$$\mathrm{Cov}\left[\cos\left(\beta_1\tilde{y}_i\right), \cos\left(\beta_2\tilde{y}_i\right)\right] = \mathrm{E}\left[\cos\left(\beta_1\tilde{y}_i\right)\cos\left(\beta_2\tilde{y}_i\right)\right] - \mathrm{E}\left[\cos\left(\beta_1\tilde{y}_i\right)\right]\mathrm{E}\left[\cos\left(\beta_2\tilde{y}_i\right)\right]. \quad (17)$$

The second term of (17) can be calculated with (3), i.e.,

$$\begin{aligned}&\mathrm{E}\left[\cos\left(\beta_1\tilde{y}_i\right)\right]\mathrm{E}\left[\cos\left(\beta_2\tilde{y}_i\right)\right] \\ &= \mathrm{e}^{-\frac{\beta_1^2\sigma^2}{2}}\cos\left(\beta_1 y_i\right)\mathrm{e}^{-\frac{\beta_2^2\sigma^2}{2}}\cos\left(\beta_2 y_i\right) \quad (18)\\ &= \frac{1}{2}\mathrm{e}^{-\frac{\left(\beta_1^2+\beta_2^2\right)\sigma^2}{2}}\left(\cos\left((\beta_1-\beta_2)y_i\right) + \cos\left((\beta_1+\beta_2)y_i\right)\right).\end{aligned}$$

The identity $\cos(a)\cos(b) = 1/2\left(\cos(a-b) + \cos(a+b)\right)$ was used for the last step. For the calculation of the first term, use the above identity again and it follows with (3)

$$\begin{aligned}&\mathrm{E}\left[\cos\left(\beta_1\tilde{y}_i\right)\cos\left(\beta_2\tilde{y}_i\right)\right] \\ &= \frac{1}{2}\left(\mathrm{E}\left[\cos\left((\beta_1-\beta_2)\tilde{y}_i\right)\right] + \mathrm{E}\left[\cos\left((\beta_1+\beta_2)\tilde{y}_i\right)\right]\right) \\ &= \frac{1}{2}\mathrm{e}^{-\frac{(\beta_1-\beta_2)^2\sigma^2}{2}}\cos\left((\beta_1-\beta_2)y_i\right) \\ &+ \frac{1}{2}\mathrm{e}^{-\frac{(\beta_1+\beta_2)^2\sigma^2}{2}}\cos\left((\beta_1+\beta_2)y_i\right) \quad (19)\end{aligned}$$

Plugging (18) and (19) into (17) one gets for the covariance term

$$\begin{aligned}&\mathrm{Cov}\left[\cos\left(\beta_1\tilde{y}_i\right), \cos\left(\beta_2\tilde{y}_i\right)\right] \\ &= \frac{1}{2}\cos\left((\beta_1-\beta_2)y_i\right)\left(\mathrm{e}^{-\frac{(\beta_1-\beta_2)^2\sigma^2}{2}} - \mathrm{e}^{-\frac{(\beta_1^2+\beta_2^2)\sigma^2}{2}}\right) \\ &+ \frac{1}{2}\cos\left((\beta_1+\beta_2)y_i\right)\left(\mathrm{e}^{-\frac{(\beta_1+\beta_2)^2\sigma^2}{2}} - \mathrm{e}^{-\frac{(\beta_1^2+\beta_2^2)\sigma^2}{2}}\right). \quad (20)\end{aligned}$$

Applying (9) to (20), one gets an R-dependent equation for the covariance term, i.e.,

$$\begin{aligned}
&\mathrm{E}\left[\mathrm{Cov}\left[\cos\left(\beta_1\tilde{y}_i\right),\cos\left(\beta_2\tilde{y}_i\right)\right]\right]\\
&=\frac{1}{2}\mathrm{E}\left[\cos\left((\beta_1-\beta_2)y_i\right)\right]\left(\mathrm{e}^{-\frac{(\beta_1-\beta_2)^2\sigma^2}{2}}-\mathrm{e}^{-\frac{(\beta_1^2+\beta_2^2)\sigma^2}{2}}\right)\\
&+\frac{1}{2}\mathrm{E}\left[\cos\left((\beta_1+\beta_2)y_i\right)\right]\left(\mathrm{e}^{-\frac{(\beta_1+\beta_2)^2\sigma^2}{2}}-\mathrm{e}^{-\frac{(\beta_1^2+\beta_2^2)\sigma^2}{2}}\right)\\
&=\frac{1}{2}\mathrm{e}^{-(\beta_1-\beta_2)^2\frac{R^2}{2N}}\left(\mathrm{e}^{-\frac{(\beta_1-\beta_2)^2\sigma^2}{2}}-\mathrm{e}^{-\frac{(\beta_1^2+\beta_2^2)\sigma^2}{2}}\right)\\
&+\frac{1}{2}\mathrm{e}^{-(\beta_1+\beta_2)^2\frac{R^2}{2N}}\left(\mathrm{e}^{-\frac{(\beta_1+\beta_2)^2\sigma^2}{2}}-\mathrm{e}^{-\frac{(\beta_1^2+\beta_2^2)\sigma^2}{2}}\right). \qquad (21)
\end{aligned}$$

Inserting (21) and (16) into (15) it follows after some straightforward calculations that

$$\begin{aligned}
&\mathrm{Var}\left[C_{\mathcal{B}}(R)\right]\\
&=\frac{B_1^2(N-1)}{2}\left(1-\mathrm{e}^{-\beta_1^2\sigma^2}\right)\left(1-\mathrm{e}^{-\beta_1^2\sigma^2}\mathrm{e}^{-2\beta_1^2\frac{R^2}{N}}\right)\\
&+\frac{B_2^2(N-1)}{2}\left(1-\mathrm{e}^{-\beta_2^2\sigma^2}\right)\left(1-\mathrm{e}^{-\beta_2^2\sigma^2}\mathrm{e}^{-2\beta_2^2\frac{R^2}{N}}\right)\\
&+B_1B_2(N-2)\left(\mathrm{e}^{-\frac{(\beta_1-\beta_2)^2\sigma^2}{2}}-\mathrm{e}^{-\frac{(\beta_1^2+\beta_2^2)\sigma^2}{2}}\right)\\
&\times\mathrm{e}^{-(\beta_1-\beta_2)^2\frac{R^2}{2N}}\\
&+B_1B_2(N-2)\left(\mathrm{e}^{-\frac{(\beta_1+\beta_2)^2\sigma^2}{2}}-\mathrm{e}^{-\frac{(\beta_1^2+\beta_2^2)\sigma^2}{2}}\right)\\
&\times\mathrm{e}^{-(\beta_1+\beta_2)^2\frac{R^2}{2N}} \qquad (22)
\end{aligned}$$

and it holds $\sigma_{\mathrm{ES}}^{\mathcal{B}}(R)=\sqrt{\mathrm{Var}\left[C_{\mathcal{B}}(R)\right]}$. This is visualized in Fig. 2 by the black dashed-dotted curve for the CSA-ES. The solid black line shows the experimental values for comparison. Note that the differences are nearly invisible. As with the Rastrigin function, slight deviations are expected for the σSA-ES, as there are additional covariances.

Under steady-state conditions, it holds $\sigma^*\approx\mathrm{const.}$ and for $R\to\infty$, the mutation strength σ increases. Thus, the terms $\mathrm{e}^{-\beta^2\sigma^2}$ vanish and the covariance terms in (22) vanish. It follows

$$\sigma_{\mathrm{ES}}^{\mathcal{B}}(R)\mid_{R\to\infty}=\sqrt{(N-1)\frac{B_1^2+B_2^2}{2}}. \qquad (23)$$

This is represented in Fig. 2 by the straight black dashed-dotted line. It can also be seen that (23) is not the maximum value and that there is an area with an elevation. For large R, the values agree with the experimental values and (22), except for the right plot. This plot represents the special case with $\beta_1=\beta_2$. In this case the covariance term in (20) is an additional variance term. The shape of the curve differs, and the asymptotic value is larger. Fig. 2 shows that the range of this increase depends on the difference between β_1 and β_2. The smaller the difference, the longer the range extends over generations.

When $\beta_1=\beta_2$, the covariance term in (15) gets $\mathrm{Cov}\left[\cos\left(\beta_1\tilde{y}_i\right),\cos\left(\beta_2\tilde{y}_i\right)\right]=\mathrm{Var}\left[\cos\left(\beta_1\tilde{y}_i\right)\right]$ and it follows

$$\begin{aligned}
\mathrm{Var}\left[C_{\mathcal{B}}(\mathbf{y})\right]&=B_1^2\sum_{i=1}^{N-1}\mathrm{Var}\left[\cos\left(\beta_1\tilde{y}_i\right)\right]\\
&+B_2^2\sum_{i=2}^{N}\mathrm{Var}\left[\cos\left(\beta_1\tilde{y}_i\right)\right]\\
&+2B_1B_2\sum_{i=2}^{N-1}\mathrm{Var}\left[\cos\left(\beta_1\tilde{y}_i\right)\right]. \qquad (24)
\end{aligned}$$

By applying (16) it follows

$$\begin{aligned}
\mathrm{Var}\left[C_{\mathcal{B}}(R)\right]&=\left(1-\mathrm{e}^{-\beta_1^2\sigma^2}\right)\\
&\times\left(1-\mathrm{e}^{-\beta_1^2\sigma^2}\mathrm{e}^{-2\beta_1^2\frac{R^2}{N}}\right)\\
&\times\left((N-1)\frac{(B_1^2+B_2^2)}{2}+(N-2)B_1B_2\right). \qquad (25)
\end{aligned}$$

$\sigma_{\mathrm{ES}}^{\mathcal{B}}(R):=\sqrt{\mathrm{Var}\left[C_{\mathcal{B}}(R)\right]}$ for the case $\beta_1=\beta_2$ is shown in the bottom right plot of Fig. 2.

For a large distance to the optimum, the terms $\mathrm{e}^{-\beta^2\sigma^2}$ in (25) vanish and it holds

$$\begin{aligned}
\hat{\sigma}_{\mathrm{ES}}^{\mathcal{B}}&:=\sqrt{\mathrm{Var}\left[C_{\mathcal{B}}(R)\right]}\mid_{R\to\infty}\\
&=\sqrt{(N-1)\frac{(B_1^2+B_2^2)}{2}+(N-2)B_1B_2}\\
&=\sqrt{\frac{1}{2}(N-1)(B_1+B_2)^2-B_1B_2}. \qquad (26)
\end{aligned}$$

This is also the maximal value of (22) which is visualized in Fig. 2 by the dashed-dotted purple line.

C. Ackley Function

The residual part of the Ackley function is

$$C_{\mathcal{A}}(\mathbf{y}):=\mathrm{e}-\mathrm{e}^{\frac{1}{N}\sum_{i=1}^{N}\cos(\gamma y_i)}. \qquad (27)$$

For an offspring individual $\tilde{\mathbf{y}}$ the exponent in (27) is a sum of iid random variables $\cos(\gamma\tilde{y}_i)$. Using the Central Limit Theorem, it follows that the random variable $\sum_{i=1}^{N}\cos\left(\gamma\tilde{y}_i\right)$ is asymptotically normal distributed with mean

$$\begin{aligned}
\mathrm{E}\left[\sum_{i=1}^{N}\cos\left(\gamma\tilde{y}_i\right)\right]&=\sum_{i=1}^{N}\mathrm{E}\left[\cos\left(\gamma\tilde{y}_i\right)\right]\\
&=\sum_{i=1}^{N}\mathrm{e}^{-\frac{1}{2}(\sigma\gamma)^2}\cos\left(\gamma y_i\right) \qquad (28)
\end{aligned}$$

and variance

$$\begin{aligned}
&\mathrm{Var}\left[\sum_{i=1}^{N}\cos\left(\gamma\tilde{y}_i\right)\right]=\sum_{i=1}^{N}\mathrm{Var}\left[\cos\left(\gamma\tilde{y}_i\right)\right]\\
&=\sum_{i=1}^{N}\frac{1}{2}\left(1-\mathrm{e}^{-(\sigma\gamma)^2}\right)\left(1-\mathrm{e}^{-(\sigma\gamma)^2}\cos(2\gamma y_i)\right), \qquad (29)
\end{aligned}$$

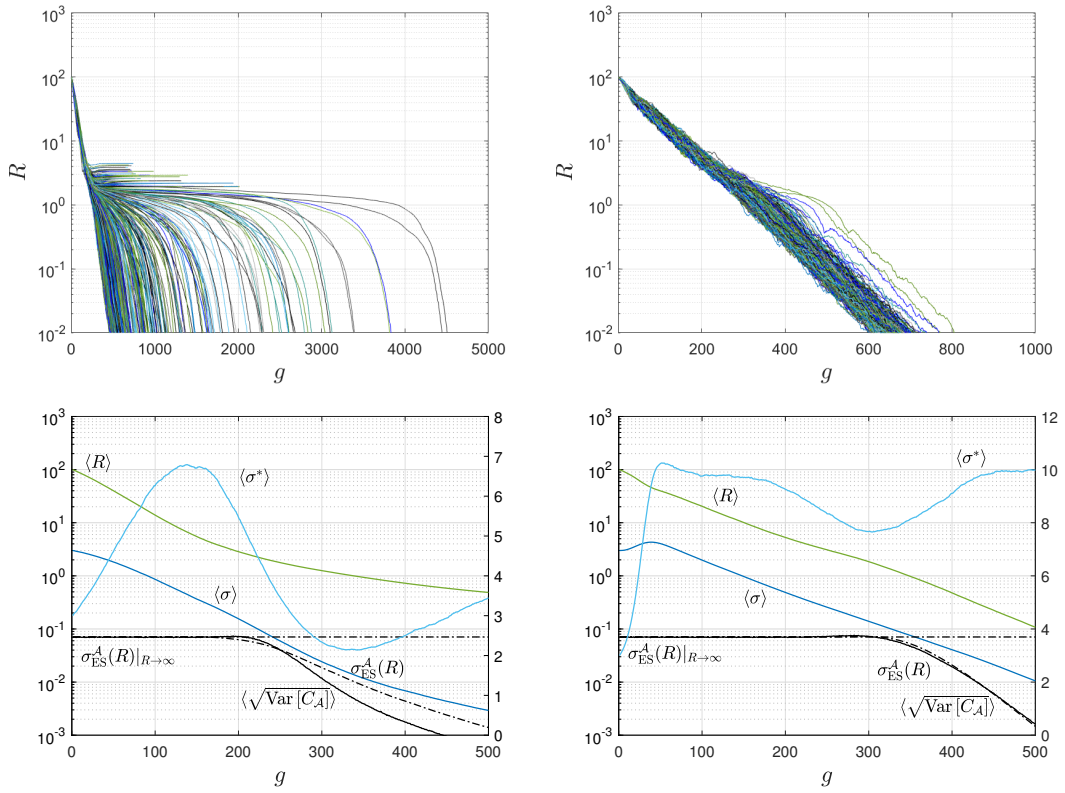

Fig. 3. Residual distance dynamics and mean value dynamics in the Ackley landscape with $N = 100$. Left plots: $(10/10_I, 20)$-σSA-ES with $\tau = 1/\sqrt{2N}$ and $P_\mathrm{s} = 0.75$, and right plots: $(10/10_I, 20)$-CSA-ES with $c = 1/\sqrt{N}$ and $P_\mathrm{s} = 1$. The green line shows the residual distance R, the blue line shows the mutation strength σ and the light blue line its normalization σ^*. The black solid line shows the experimentally determined values for the standard deviation of the residual part. The dashed-dotted curve shows $\sigma^{\mathcal{A}}_{\mathrm{ES}}(R)$ given by (37) while the dashed-dotted straight line shows its asymptotic behavior $\sigma^{\mathcal{A}}_{\mathrm{ES}}(R)\,|_{R\to\infty}$ (39). σ^* is represented by the right y-axis, all other values are represented by the left y-axis.

where (3) and (5) were applied. By applying (9) it follows for a parent $\mathbf{y}$ located on the R-sphere that the exponent in (27) is normally distributed with mean and variance

$$\mu_{\mathcal{A}} := \mathrm{e}^{-\frac{1}{2}(\sigma\gamma)^2}\mathrm{e}^{-\frac{1}{2}\frac{(\gamma R)^2}{N}} \tag{30}$$

$$\sigma^2_{\mathcal{A}} := \frac{1}{2N}\left(1-\mathrm{e}^{-(\sigma\gamma)^2}\right)\left(1-\mathrm{e}^{-(\sigma\gamma)^2}\mathrm{e}^{-\frac{1}{2}\frac{(\gamma R)^2}{N}}\right). \tag{31}$$

Therefore, the residual part of the Ackley function can be approximated by a log-normally distributed random variable, i.e.,

$$C_{\mathcal{A}}(\mathbf{y}) \approx \mathrm{e} - \mathrm{e}^{\mathcal{N}\left(\mu_{\mathcal{A}},\sigma^2_{\mathcal{A}}\right)}. \tag{32}$$

For the expected value and the variance of a log-normally distributed random variable, it holds [2]

$$\mathrm{E}\left[\mathrm{e}^{\mathcal{N}\left(\mu,\sigma^2\right)}\right] = \mathrm{e}^{\mu+\frac{1}{2}\sigma^2} \tag{33}$$

$$\mathrm{Var}\left[\mathrm{e}^{\mathcal{N}\left(\mu,\sigma^2\right)}\right] = \mathrm{e}^{2\mu+\sigma^2}\left(\mathrm{e}^{\sigma^2}-1\right) \tag{34}$$

and it follows

$$\mathrm{E}\left[C_{\mathcal{A}}\right] = \mathrm{e} - \mathrm{e}^{\mu_{\mathcal{A}}+\frac{1}{2}\sigma^2_{\mathcal{A}}} \tag{35}$$

$$\mathrm{Var}\left[C_{\mathcal{A}}\right] = \mathrm{e}^{2\mu_{\mathcal{A}}+\sigma^2_{\mathcal{A}}}\left(\mathrm{e}^{\sigma^2_{\mathcal{A}}}-1\right). \tag{36}$$

Approximating the log-normal random variable $C_{\mathcal{A}}(\mathbf{y})$ by a normally distributed random variable with variance $\mathrm{Var}\left[C_{\mathcal{A}}\right]$ it follows

$$\sigma^{\mathcal{A}}_{\mathrm{ES}}(R) = \sqrt{\mathrm{Var}\left[C_{\mathcal{A}}(R)\right]} = \sqrt{\mathrm{e}^{2\mu_{\mathcal{A}}+\sigma^2_{\mathcal{A}}}\left(\mathrm{e}^{\sigma^2_{\mathcal{A}}}-1\right)} \tag{37}$$

where $\mu_{\mathcal{A}}$ and $\sigma^2_{\mathcal{A}}$ are given by (30) and (31). Eq. (37) is represented in Fig. 3 by the black dashed-dotted curve. The solid line shows the experimental values for comparison. As expected there are some deviations between the experimental values and its predictions. These deviations are small for the CSA-ES and larger for the σSA-ES.

For increasing residual distance R and under steady-state conditions, the mutation strength σ increases, causing the terms $\mathrm{e}^{-(\gamma\sigma)^2}$ to vanish. It follows

$$\mu_{\mathcal{A}}(R)\,|_{R\to\infty} = 0,\ \sigma^2_{\mathcal{A}}(R)\,|_{R\to\infty} = \frac{1}{2N} \tag{38}$$

and therefore

$$\begin{aligned}\sigma^{\mathcal{A}}_{\mathrm{ES}}(R)|_{R\to\infty} &= \sqrt{\mathrm{e}^{\frac{1}{2N}}\left(\mathrm{e}^{\frac{1}{2N}}-1\right)} \\ &= \sqrt{\mathrm{e}^{\frac{1}{N}}-\mathrm{e}^{\frac{1}{2N}}} \\ &\simeq \sqrt{1+\frac{1}{N}-1-\frac{1}{2N}} \\ &= \frac{1}{\sqrt{2N}}.\end{aligned} \tag{39}$$

where the Taylor approximation $e^x \approx 1 + x$ was used. (39) is represented in Fig. 3 by the straight dashed-dotted line.

II. The Global Attractor Region

A. Plots of the Attractor Region

Figure 4 shows the global Attractor regions for $N = 2$.

B. Attractor Region of the Ackley Function

To determine the nearest stationary points on the coordinate axes, consider the derivative of the Ackley function, i.e.,

$$\frac{\partial F}{\partial y_i} = \frac{C_1 C_2}{\sqrt{N \sum_{j=1}^{N} y_j^2}} y_i \mathrm{e}^{-C_2 \sqrt{\frac{1}{N}\sum_{j=1}^{N} y_j^2}} + \frac{\gamma}{N} \sin(\gamma y_i)\, \mathrm{e}^{\frac{1}{N}\sum_{j=1}^{N}\cos(\gamma y_j)}. \tag{40}$$

Since $\partial F / \partial y_i = 0$ for $y_i = 0$ the distance $\Delta_{\mathcal{A}}$ from the stationary points on the axis to the origin is the solution of

$$\operatorname{sgn}(y) \frac{C_1 C_2}{\sqrt{N}} \mathrm{e}^{-\frac{C_2}{\sqrt{N}}|y|} + \frac{\gamma}{N} \quad \sin(\gamma y)\, \mathrm{e}^{\frac{1}{N}\cos(\gamma y) + \frac{N-1}{N}} = 0. \tag{41}$$

It follows for $y > 0$

$$\mathrm{e}^{-\frac{C_2}{\sqrt{N}} y} = -\frac{\gamma}{C_1 C_2 \sqrt{N}} \sin(\gamma y)\, \mathrm{e}^{\frac{1}{N}\cos(\gamma y) + \frac{N-1}{N}}. \tag{42}$$

Due to the symmetry of the Ackley function, the attractor region is also symmetric and the case $y < 0$ follows analogously. $\Delta_{\mathcal{A}}$ is given by the intersection between the two functions

$$f_{exp} := \mathrm{e}^{-\frac{C_2}{\sqrt{N}} y} \tag{43}$$

$$f_{osc} := -\frac{\gamma}{C_1 C_2 \sqrt{N}} \sin(\gamma y)\, \mathrm{e}^{\frac{1}{N}\cos(\gamma y) + \frac{N-1}{N}}. \tag{44}$$

The exponential function f_{exp} is monotonically decreasing, whereas f_{osc} oscillates and is bounded. This is visualized in the top plot of Fig. 5. For large N, the fluctuations of the function $\mathrm{e}^{\frac{1}{N}\cos(\gamma y)}$ are minimal compared to the fluctuations of $\sin(\gamma y)$. The maximum value of f_{osc} is approximately reached at the point where $\sin(\gamma y)$ is minimal, i.e., at $3\pi/2\gamma$, and therefore

$$\max_y f_{osc}(y) \approx \frac{\gamma}{C_1 C_2 \sqrt{N}} \mathrm{e}^{\frac{N-1}{N}}. \tag{45}$$

The value where f_{exp} intersects this maximum serves as a lower bound for $\Delta_{\mathcal{A}}$, i.e., the y-value with

$$\begin{aligned} \mathrm{e}^{-\frac{C_2}{\sqrt{N}} y} &= \frac{\gamma}{C_1 C_2 \sqrt{N}} \mathrm{e}^{\frac{N-1}{N}} \\ \Leftrightarrow -\frac{C_2}{\sqrt{N}} y &= \ln\left(\frac{\gamma}{C_1 C_2 \sqrt{N}}\right) + \frac{N-1}{N} \\ \Leftrightarrow y &= -\ln\left(\frac{\gamma}{C_1 C_2 \sqrt{N}}\right) \frac{\sqrt{N}}{C_2} - \frac{N-1}{C_2 \sqrt{N}} \end{aligned} \tag{46}$$

is a lower bound for $\Delta_{\mathcal{A}}$. Because the slope of f_{exp} becomes flatter with increasing N, the inequality approaches equality for sufficiently large N. Therefore, the distance from the

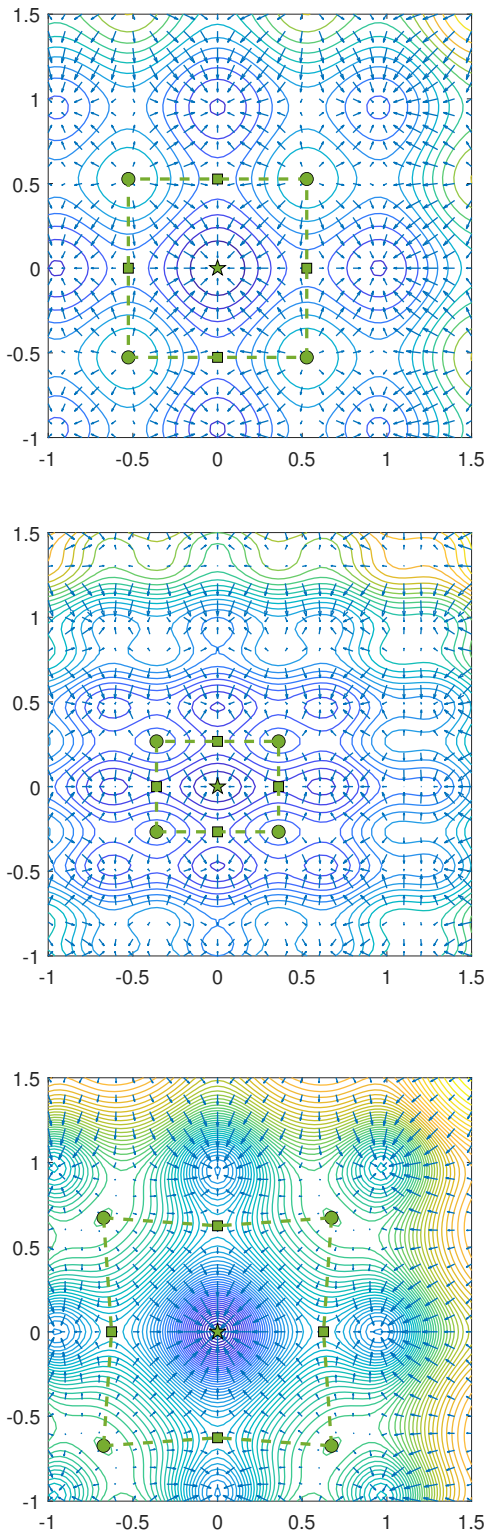

Fig. 4. Global attractor region of the Rastrigin (top), Bohachevsky (middle), and Ackley (bottom) function with standard parameters for $N = 2$. Arrows show the negative gradient flow. The star shows the global optimizer, squares the nearest stationary points, and circles the farthest stationary points.

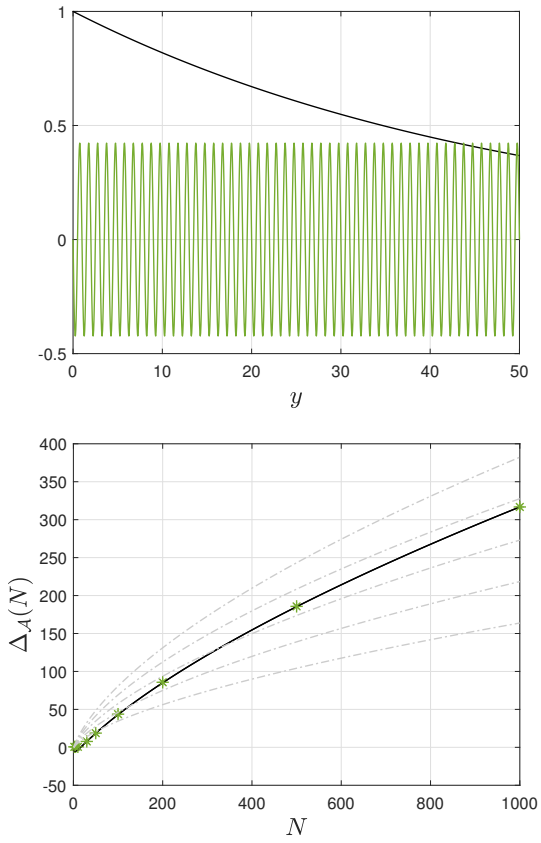

Fig. 5. Top plot: f_{exp} (43) (black line) and f_{osz} (44) (green line) for $N = 100$. Bottom plot: (47) (black line) compared with numerical solutions of (41) (green stars).

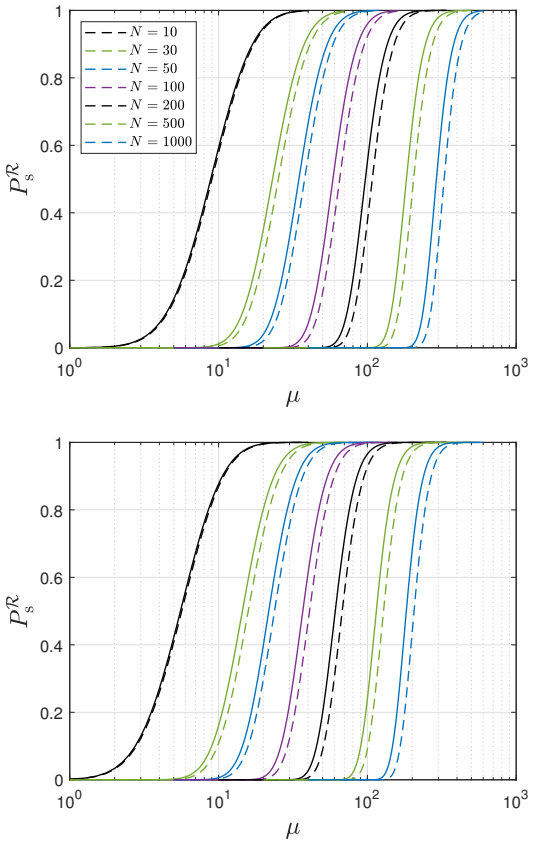

Fig. 6. Success probability $P_{\mathrm{s}}^{\mathcal{R}}$ (51) depending on μ for the Rastrigin landscape with $\vartheta = 0.5$ (top plot) and $\vartheta = 0.25$ (bottom plot). Solid lines use $\hat{\sigma}_{\mathrm{ES}}^{\mathcal{R}} = A\sqrt{N/2}$, dashes lines use the maximal value of $\sigma_{\mathrm{ES}}^{\mathcal{R}}$ determined experimentally for the σSA-ES with $\tau = 1/\sqrt{2N}$.

optimizer to the nearest stationary point on the coordinate axis is

$$\Delta_{\mathcal{A}}(N) \approx -\ln\left(\frac{\gamma}{C_1 C_2 \sqrt{N}}\right)\frac{\sqrt{N}}{C_2} - \frac{N-1}{C_2\sqrt{N}}. \tag{47}$$

These points lie farther outside for larger N. (47) is represented by the black line in the bottom plot of Fig. 5 and is compared with the numerical solution of (41).

For the stationary points that do not lie on the coordinate axis, no analytical calculation is possible. The derivative of the Ackley function in direction i also depends on the other directions j. There is no general formula that allows for the determination of stationary points that do not lie on the coordinate axes.

It is not yet clear whether the attractor region is bounded. This will be investigated in the following. The Ackley function can be expressed as a function of the distance to the optimizer R, i.e.,

$$F_{\mathcal{A}}(R) = C_1 - C_1 \mathrm{e}^{-\frac{C_2}{\sqrt{N}}R} + \mathrm{e} - \mathrm{e}^{\frac{1}{N}\sum_{i=1}^{N}\cos\left(\gamma\frac{z_i}{\|\mathbf{z}\|}R\right)} \tag{48}$$

where $\mathbf{z} = \mathbf{z}/\|\mathbf{z}\|R$ is represented as a directional vector multiplied with the distance R. The directional derivatives of the Ackley function are

$$\frac{\partial F_{\mathcal{A}}}{\partial R}(R) = \frac{C_1 C_2}{\sqrt{N}}\mathrm{e}^{-\frac{C_2}{\sqrt{N}}R} + \frac{\gamma}{N}\sum_{i=1}^{N}\frac{z_i}{\|\mathbf{z}\|}\sin\left(\gamma\frac{z_i}{\|\mathbf{z}\|}R\right) \times \mathrm{e}^{\frac{1}{N}\sum_{i=1}^{N}\cos\left(\gamma\frac{z_i}{\|\mathbf{z}\|}R\right)}. \tag{49}$$

Set the directional derivative to zero. It follows

$$\frac{C_1 C_2}{\gamma}\mathrm{e}^{-\frac{C_2}{\sqrt{N}}R} = -\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{z_i}{\|\mathbf{z}\|}\sin\left(\gamma\frac{z_i}{\|\mathbf{z}\|}R\right) \times \mathrm{e}^{\frac{1}{N}\sum_{i=1}^{N}\cos\left(\gamma\frac{z_i}{\|\mathbf{z}\|}R\right)}. \tag{50}$$

The function on the lhs decreases to zero while the second expression on the rhs lies between e^{-1} and e. The first expression on the rhs is a sum of sine curves $a_i \sin(b_i R)$. This sum is a continuous function with at least one zero, at $R = 0$. Therefore the sum contains positive and negative values. The sum of periodic functions is almost periodic [3]. From the definition of an almost periodic function it follows that when the function contains positive and negative values, the function has an infinite number of zeros. Therefore, the first expression of the rhs has an infinite number of zeros. As the lhs approaches zero, there are infinitely many stationary points in each direction.

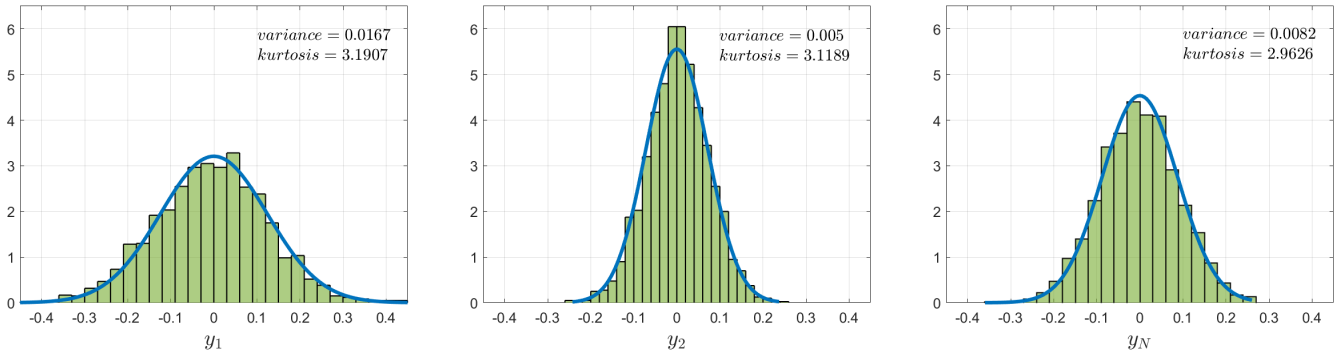

Fig. 7. Histogram of parent components y_i at distance $R_{\mathrm{st}}^{\mathcal{B}} = 0.72$ (see (36) from the main paper) for the $(100/100_I, 200)$-σSA-ES with $\tau = 1/\sqrt{8N}$ and $N = 100$. Blue lines show the probability density functions of an $\mathcal{N}\left(0, \sigma_{\mathrm{st}}(i)^2\right)$-distributed random variable with $\sigma_{\mathrm{st}}(i)$ from (57).

III. The Success Probability Model

A. Influence of Larger σ_{ES} on P_{s} for the σSA-ES

The success probability for the Rastrigin function was determined with

$$P_{\mathrm{s}}^{\mathcal{R}} = \left[2\Phi\left(\sqrt{\frac{4\mu c_{\mu/\mu,\lambda}}{\hat{\sigma}_{\mathrm{ES}}^{\mathcal{R}}}}\left(\Delta_{\mathcal{R}} + \varepsilon\right)\right) - 1\right]^N. \tag{51}$$

In Section I, it was shown that the standard deviation of the cosine part for the σSA-ES is slightly larger than predicted from $\sigma_{\mathrm{ES}}(R)$, (11). In the derivation of (51) the maximum value of $\sigma_{\mathrm{ES}}(R)$ for the CSA-ES was used. For the σSA-ES, the value of $\sigma_{\mathrm{ES}}(R)$ can be slightly larger, as shown in Fig. 1. In the bottom left plot, which represents the σSA-ES with $\tau = 1/\sqrt{2N}$, there are small deviations between the black solid and the black dashed-dotted curve. In the bottom middle plot, which represents the σSA-ES with $\tau = 1/\sqrt{8N}$, there are also deviations, although they are barely noticeable.

How large the influence of this simplification on (51) is represents Fig. 6. The values of $\max \sigma_{\mathrm{ES}}(R)$ were determined experimentally for the σSA-ES with $\tau = 1/\sqrt{2N}$ in the Rastrigin landscape. The differences are visible and increase slightly with increasing N. However, the differences are small enough to be neglected in this simple model.

B. Influence of Neglected Anisotropy Aspects of the Bohachevsky Function on $P_{\mathrm{s}}^{\mathcal{B}}$

In the derivation of the success probability equation for the Bohachevsky function

$$P_{\mathrm{s}}^{\mathcal{B}} = \left[2\Phi\left(\sqrt{\frac{12\mu c_{\mu/\mu,\lambda}}{\hat{\sigma}_{\mathrm{ES}}^{\mathcal{B}}}}\left(\Delta_{\mathcal{B}} + \varepsilon\right)\right) - 1\right]^N, \tag{52}$$

simplifications were made. The global part of the Bohachevsky function is an ellipsoid. This fact was neglected in the derivation of (52). On the one hand, the attractor region was assumed to be a hypercube, although it is actually a parallelepiped. The lengths of the edges from the attractor region are (see (40) and (41) in the main paper)

$$\Delta_{\mathcal{B}} := \Delta_{\mathcal{B}}^{(i)} = \frac{\pi\left(B_1\beta_1 + B_2\beta_2\right)}{B_1\beta_1^2 + B_2\beta_2^2 - 6} \tag{53}$$

$$\Delta_{\mathcal{B}}^{(1)} = \frac{B_1\beta_1\pi}{B_1\beta_1^2 - 2} \tag{54}$$

$$\Delta_{\mathcal{B}}^{(N)} = \frac{B_2\beta_2\pi}{B_2\beta_2^2 - 4}. \tag{55}$$

In addition, the distribution of the first and last parental component differs from the distribution of the other components. In general, it holds that in the noisy ellipsoid model

$$\tilde{F}(\mathbf{y}) = \sum_{i=1}^{N} q_i y_i^2 + \sigma_{\mathrm{ES}}\mathcal{N}(0,1), \tag{56}$$

the variance of the single components under steady-state conditions is [4]

$$\sigma_{\mathrm{st}}(i)^2 := \mathrm{Var}\left[y_i\right] = \frac{\sigma_{\mathrm{ES}}}{4\mu c_{\mu/\mu,\lambda} q_i}. \tag{57}$$

In the Bohachevsky landscape, it holds $q_1 = 1$, $q_N = 2$, and $q_i = 3$ for $i = 2, \ldots, N-1$. Therefore, the variances of the first and last components are larger than the variances of the remaining components. This is also visualized in Fig. 7, where histograms for single components are shown. The values of variance and kurtosis given in the respective plot were determined from the experimental data. In all cases, the variance measured from the experimental data is consistent with (57).

In the following, these two effects are taken into account in the derivation of (52). It follows

$$\begin{aligned} P_{\mathrm{s}}^{\mathcal{B}} &= \Pr\left[\mathbf{y} \in \mathcal{A}_{\mathrm{ES}}^{\mathcal{B}}\right] \\ &= \Pr\Big[\left(-\Delta_{\mathcal{B}}^{(1)} - \varepsilon \le y_1 \le \Delta_{\mathcal{B}}^{(1)} + \varepsilon\right) \\ &\quad \wedge \left(-\Delta_{\mathcal{B}} - \varepsilon \le y_2 \le \Delta_{\mathcal{B}} + \varepsilon\right)^{(N-2)} \\ &\quad \wedge \left(-\Delta_{\mathcal{B}}^{(N)} - \varepsilon \le y_N \le \Delta_{\mathcal{B}}^{(N)} + \varepsilon\right)\Big]. \end{aligned} \tag{58}$$

Using the fact that the single components are normally distributed with variance from (57), one gets for a single component y_i

$$\begin{aligned}
&\Pr\left[-\Delta_{\mathcal{B}}^{(i)} - \varepsilon \leq y_i \leq \Delta_{\mathcal{B}}^{(i)} + \varepsilon\right] \\
=&\Pr\left[-\frac{\Delta_{\mathcal{B}}^{(i)} + \varepsilon}{\sigma_{\mathrm{st}}(i)} \leq z_i \leq \frac{\Delta_{\mathcal{B}}^{(i)} + \varepsilon}{\sigma_{\mathrm{st}}(i)}\right] \\
=&\Phi\left(\frac{\Delta_{\mathcal{B}}^{(i)} + \varepsilon}{\sigma_{\mathrm{st}}(i)}\right) - \Phi\left(-\frac{\Delta_{\mathcal{B}}^{(i)} + \varepsilon}{\sigma_{\mathrm{st}}(i)}\right) \\
=&2\Phi\left(\frac{\Delta_{\mathcal{B}}^{(i)} + \varepsilon}{\sigma_{\mathrm{st}}(i)}\right) - 1 \qquad (59)
\end{aligned}$$

where $\Phi(z)$ is the cdf of the standard normal variate $\mathcal{N}(0,1)$. Thus, one gets for the success probability

$$\begin{aligned}
P_{\mathrm{s}}^{\mathcal{B}} = &\left[2\Phi\left(\frac{\Delta_{\mathcal{B}}^{(1)} + \varepsilon}{\sigma_{\mathrm{st}}(1)}\right) - 1\right] \qquad (60) \\
&\times \left[2\Phi\left(\frac{\Delta_{\mathcal{B}} + \varepsilon}{\sigma_{\mathrm{st}}(2)}\right) - 1\right]^{N-2} \\
&\times \left[2\Phi\left(\frac{\Delta_{\mathcal{B}}^{(N)} + \varepsilon}{\sigma_{\mathrm{st}}(N)}\right) - 1\right].
\end{aligned}$$

The differences between (52) and (60) are visualized in Fig. 8. The larger values for $\sigma_{\mathrm{st}}(1)$ and $\sigma_{\mathrm{st}}(N)$ lead to lower success probabilities, but already for $N \geq 30$ the differences are relatively small.

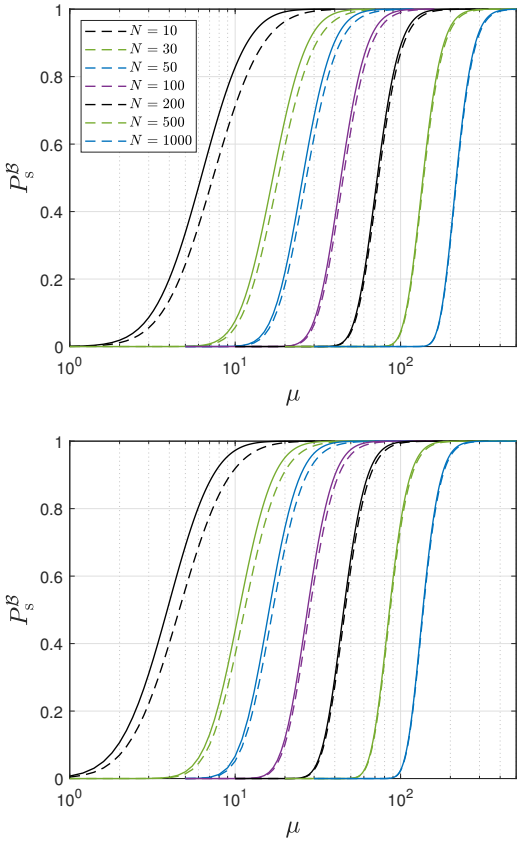

Fig. 8. Success probability $P_{\mathrm{s}}^{\mathcal{B}}$ depending on μ for the Bohachevsky landscape with $\vartheta = 0.5$ (top plot) and $\vartheta = 0.25$ (bottom plot). Solid lines represent (52) and dashed lines (60), respectively.

C. Results for ϑ = 0.25

The following figures show the success probability depending on μ for a truncation ratio of $\vartheta = 0.25$. For the termination condition, a σ_{stop} between 10^{-4} and 10^{-7} was used. An additional termination condition for the Ackley function is that the ES must not move too far away from the starting point, contrary to the optimum. Fig. 9 represents the results for the Rastrigin function, Fig. 10 for the Bohachevsky function, and Fig. 11 for the Ackley function, respectively.

D. Additional plots for P_{s} with $\varepsilon \neq 0$

In the following figures the success probability equations (51) and (52) with $\varepsilon \neq 0$ are compared with experimental values. Fig. 12 shows the results for the Rastrigin landscape and Fig. 13 for the Bohachevsky landscape, respectively. The corresponding ε values are represented in Fig. 14. For the σSA-ES with $\tau = 1/\sqrt{2N}$ in the Bohachevsky landscape, no approximation with ε is possible. In this case the curves are much flatter than predicted from (52). The ε-value only shifts the curve to the left or right but has no influence on the increase.

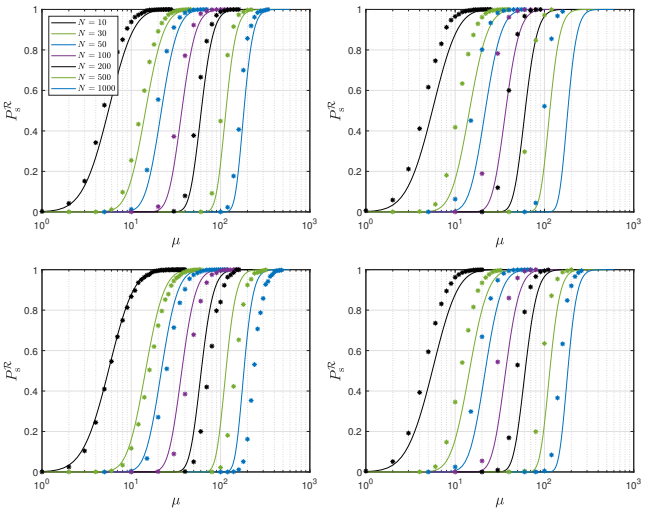

Fig. 9. Success $P_{\mathrm{s}}^{\mathcal{R}}$ depending on population size μ in the Rastrigin landscape. Solid lines show $P_{\mathrm{s}}^{\mathcal{R}}$ predicted by (51) with $\varepsilon = 0$ and $\vartheta = 0.25$. Experimental values are displayed by the stars. From top left to bottom right: CSA-ES with $c = 1/\sqrt{N}$, $c = 1/N$, σSA-ES with $\tau = 1/\sqrt{2N}$ and $\tau = 1/\sqrt{8N}$.

IV. Initial Values of the Ackley Function

To describe the threshold R_{crit} beyond which convergence becomes unattainable, start with the evolution condition

$$\left(\sigma_{\mathrm{ES}}^*\right)^2 \leq 4\mu^2 c_{\mu/\mu,\lambda}^2 - \left(\sigma^*\right)^2 \qquad (61)$$

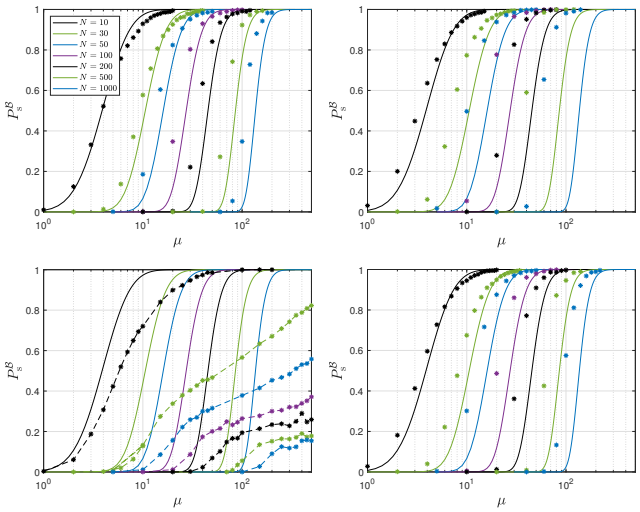

Fig. 10. Success $P_s^{\mathcal{B}}$ depending on population size μ in the Bohachevsky landscape. Solid lines show $P_s^{\mathcal{B}}$ predicted by (52) with $\varepsilon = 0$ and $\vartheta = 0.25$. Experimental values are displayed by the stars. From top left to bottom right: CSA-ES with $c = 1/\sqrt{N}$, $c = 1/N$, σSA-ES with $\tau = 1/\sqrt{2N}$ and $\tau = 1/\sqrt{8N}$.

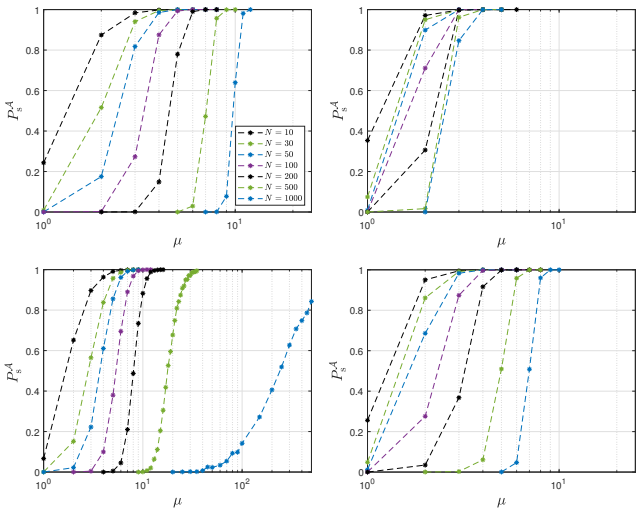

Fig. 11. Success $P_s^{\mathcal{A}}$ depending on population size μ in the Ackley landscape with $\vartheta = 0.25$. Experimental values are displayed by the stars. From top left to bottom right: CSA-ES with $c = 1/\sqrt{N}$, $c = 1/N$, σSA-ES with $\tau = 1/\sqrt{2N}$ and $\tau = 1/\sqrt{8N}$.

which necessarily implies $\sigma^* \leq 2\mu c_{\mu/\mu,\lambda}$. Using the normalization of the noise strength

$$\begin{aligned}\left(\sigma_{\mathrm{ES}}^{\mathcal{A}}\right)^* &= \sigma_{\mathrm{ES}}^{\mathcal{A}}(R)\frac{N}{RG'_{\mathcal{A}}(R)} \\ &= \sigma_{\mathrm{ES}}^{\mathcal{A}}(R)\frac{N\sqrt{N}}{RC_1C_2}\mathrm{e}^{C_2\frac{R}{\sqrt{N}}}\end{aligned} \tag{62}$$

where $G_{\mathcal{A}}$ is the global part of the Ackley function, leads to

$$\left(\sigma_{\mathrm{ES}}^{\mathcal{A}}\right)^2 \frac{N^3}{R^2C_1^2C_2^2}\mathrm{e}^{2C_2\frac{R}{\sqrt{N}}} \leq 4\mu^2c_{\mu/\mu,\lambda}^2 - \left(\sigma^*\right)^2. \tag{63}$$

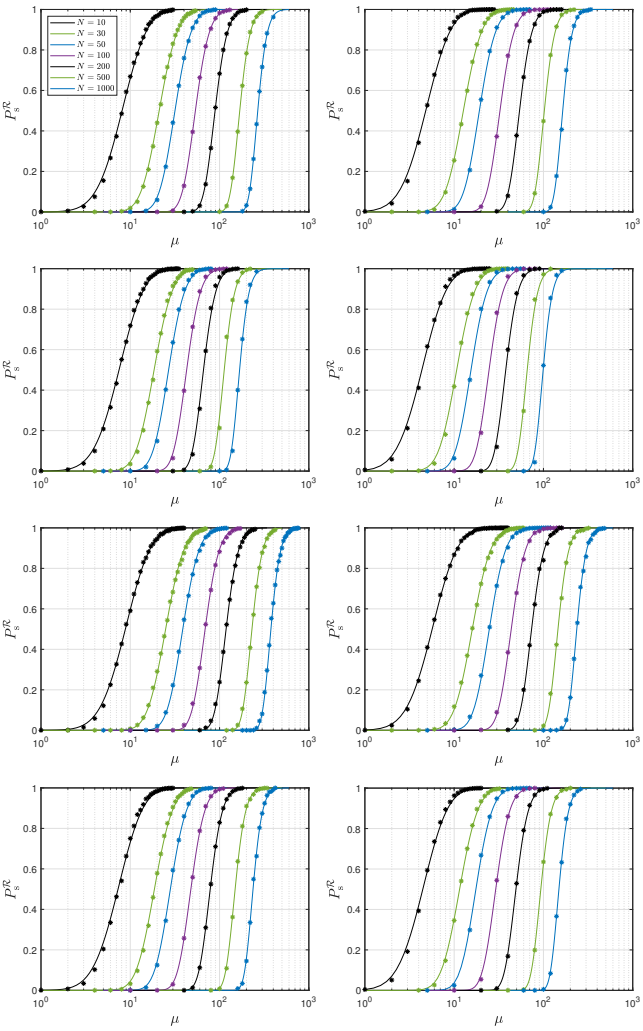

Fig. 12. Success $P_s^{\mathcal{R}}$ predicted by (51) with ε values according to Fig. 14 in the Rastrigin landscape with $\vartheta = 0.5$ (left plots) and $\vartheta = 0.25$ (right plots), respectively. From top to bottom: CSA-ES with $c = 1/\sqrt{N}$, $c = 1/N$, σSA-ES with $\tau = 1/\sqrt{2N}$, and $\tau = 1/\sqrt{8N}$.

By substituting

$$R = \frac{x\sqrt{N}}{2C_2} \Leftrightarrow x = \frac{2RC_2}{\sqrt{N}}, \tag{64}$$

and inserting $\hat{\sigma}_{\mathrm{ES}}^{\mathcal{A}} = \frac{1}{\sqrt{2N}}$ (39), it follows

$$\begin{aligned}&\frac{2N}{x^2C_1^2}\mathrm{e}^x \leq 4\mu^2c_{\mu/\mu,\lambda}^2 - \left(\sigma^*\right)^2 \\ &\Leftrightarrow \mathrm{e}^x \leq x^2\frac{C_1^2\left(4\mu^2c_{\mu/\mu,\lambda}^2 - \left(\sigma^*\right)^2\right)}{2N} \\ &\Leftrightarrow x \leq 2\ln(x) + \ln\left(\frac{C_1^2\left(4\mu^2c_{\mu/\mu,\lambda}^2 - \left(\sigma^*\right)^2\right)}{2N}\right).\end{aligned} \tag{65}$$

Consider equality and use a fixed-point iteration to determine the solution x. Because $x \gg 2\ln(x)$ for large x use

$$x^{(0)} = \ln\left(\frac{C_1^2\left(4\mu^2c_{\mu/\mu,\lambda}^2 - \left(\sigma^*\right)^2\right)}{2N}\right) \tag{66}$$

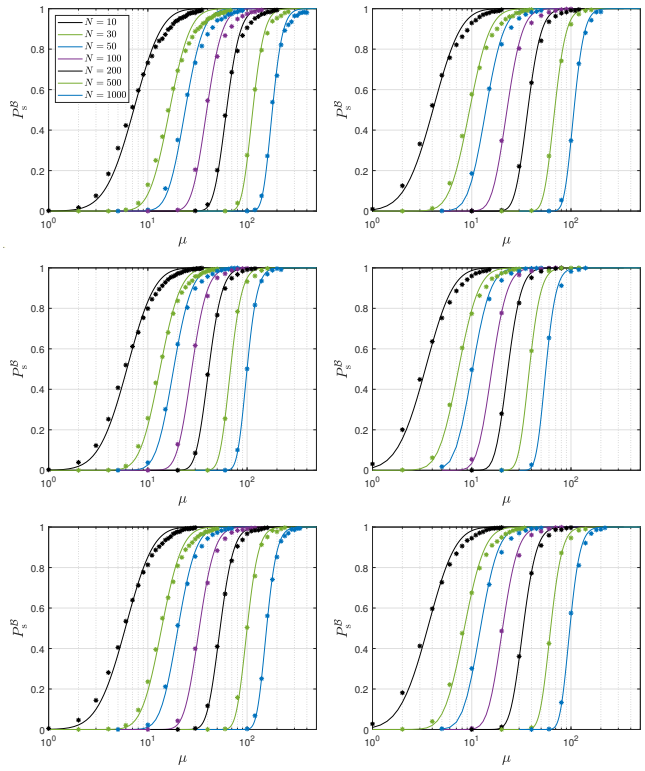

Fig. 13. Success $P_s^{\mathcal{B}}$ predicted by (52) with ε values according to Fig. 14 in the Bohachevsky landscape with $\vartheta = 0.5$ (left plots) and $\vartheta = 0.25$ (right plots), respectively. From top to bottom: CSA-ES with $c = 1/\sqrt{N}$, $c = 1/N$, and σSA-ES with $\tau = 1/\sqrt{8N}$.

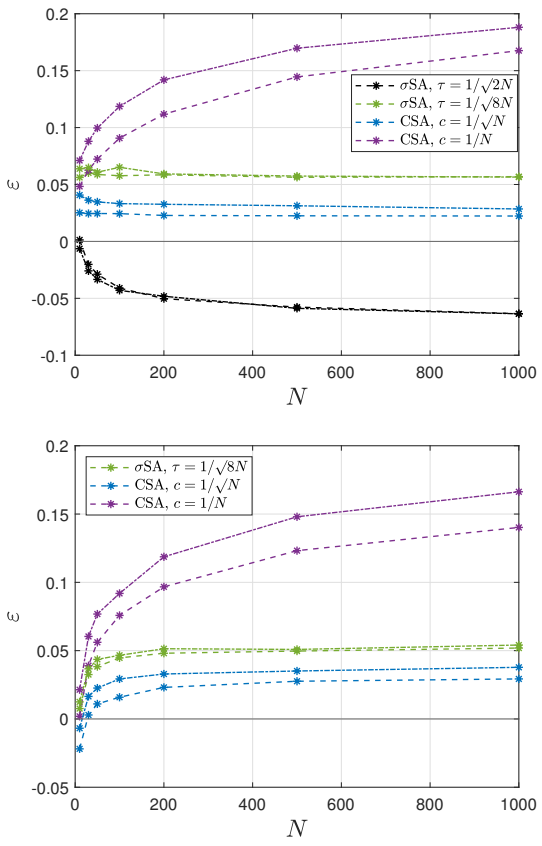

Fig. 14. Optimal ε that minimizes the deviations between (51) and (52) and the experimental values of Figs. 9 and 10 for the Rastrigin function (top plot) and Bohachevsky function (bottom plot). Dashed lines represent $\vartheta = 0.5$ and dashed-dotted lines represent $\vartheta = 0.25$, respectively.

as start value of the fix-point iteration to determine the second zero. The subsequent iteration step yields

$$x^{(1)} = \ln\left(\frac{C_1^2\left(4\mu^2 c_{\mu/\mu,\lambda}^2 - (\sigma^*)^2\right)}{2N}\right) + 2\ln\left(\ln\left(\frac{C_1^2\left(4\mu^2 c_{\mu/\mu,\lambda}^2 - (\sigma^*)^2\right)}{2N}\right)\right). \quad (67)$$

Higher order approximations are given by

$$x^{(k)} = \ln\left(\frac{C_1^2\left(4\mu^2 c_{\mu/\mu,\lambda}^2 - (\sigma^*)^2\right)}{2N}\right) + 2\ln\left(x^{(k-1)}\right). \quad (68)$$

Using the substitution (64), the distance R_{crit} can be numerically calculated with

$$R_{\text{crit}} = \frac{x^{(k)}\sqrt{N}}{2C_2}. \quad (69)$$

The maximal value σ_{crit} necessary for convergence can be calculated with (61) and (62). It follows

$$(\sigma^*)^2 \leq 4\mu^2 c_{\mu/\mu,\lambda}^2 - \left(\sigma_{\text{ES}}^{\mathcal{A}}\frac{N\sqrt{N}}{RC_1C_2}\mathrm{e}^{C_2\frac{R}{\sqrt{N}}}\right)^2 \quad (70)$$

and finally with $\hat{\sigma}_{\text{ES}}^{\mathcal{A}} = \frac{1}{\sqrt{2N}}$

$$\sigma_{\text{crit}}^* = \sqrt{4\mu^2 c_{\mu/\mu,\lambda}^2 - \frac{1}{2}\left(\frac{N}{RC_1C_2}\right)^2 \mathrm{e}^{2C_2\frac{R}{\sqrt{N}}}}. \quad (71)$$

Equations (69) and (71) are examined in Fig. 15, where the probability of divergence (P_{div}) depending on the initial values $R^{(0)}$ and $\sigma^{(0)}$ is illustrated. The right plot shows the R-dynamics of CSA-ES runs. The starting values were selected to cause a significant proportion of the runs to diverge. In the left plot, the probability of divergence depending on $R^{(0)}$ is examined. The probability of divergence is smaller for larger μ, resulting in a larger R_{crit}. The probability of divergence is larger for larger $\sigma^{*(0)}$, leading to a smaller R_{crit}. These observations agree with the predictions of (69). For $\sigma^{*(0)} = 3$, the value of R_{crit} falls within the range where P_{div} lies between 0 and 1. For $\sigma^{*(0)} = 6$ the probability of divergence is larger and all runs diverge even before reaching R_{crit} indicating the limitations of the model used.

The middle plot shows P_{div} depending on $\sigma^{*(0)}$ with vertical dashed lines representing (71). The experimental values are smaller for larger μ, resulting in a larger σ_{crit}. This observation agrees with (71). For larger $R^{(0)}$, the experimental values are considerably larger compared to (71) where σ_{crit}^* is nearly constant for small values of R. The differences between

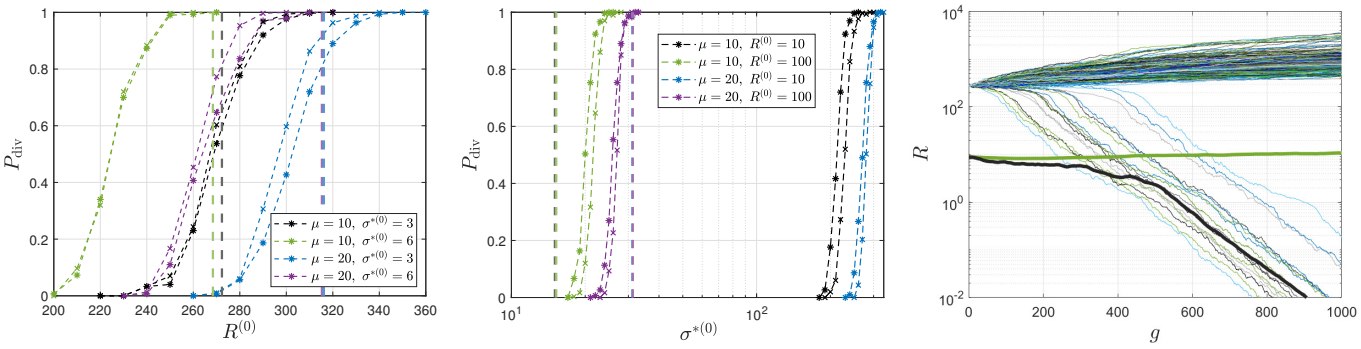

Fig. 15. Probability that an ES in the Ackley landscape diverges depending on initial distance $R^{(0)}$ (left plot) and initial mutation strength $\sigma^{(0)}$ (middle plot). Experimental values were obtained from 300 independent ES runs with $N = 100$ and $\vartheta = 0.5$ for the CSA-ES with $c = 1/\sqrt{N}$ ($*$) and the σSA-ES with $\tau = 1/\sqrt{2N}$ ($\times$). Vertical dashed lines show (69) (left plot) with $k = 8$ and (71) (middle plot). Right plot shows the residual distance dynamics for the CSA-ES with $c = 1/\sqrt{N},\ N = 100,\ \mu = 10$ and $\vartheta = 0.5$. Each run was initialized randomly at residual distance of $R^{(0)} = 280$ and started with a mutation strength $\sigma^{(0)} = 9$. Black bold line shows the mean σ-dynamics of the convergent runs and green bold line of the divergent runs, respectively.

$R^{(0)} = 10$ and $R^{(0)} = 100$ of (71) are nearly not visible in the plot. The predicted values from (71) are significantly smaller than the theoretical ones for $R^{(0)} = 10$.

For small μ no solution of (69) and (71) exists. This means that there are no initial values $R^{(0)}$ and $\sigma^{(0)}$ that guarantee convergence. For instance, for $N = 500$ and $\mu = 2$ no solutions exist for R_{crit} and σ_{crit}. Experiments show that even for very low start values, almost all runs diverge, such as for the σSA-ES with $\tau = 1/\sqrt{8N},\ R^{(0)} = 100$, and $\sigma^{(0)} = 3$.

REFERENCES

[1] I. Abramowitz, M. Stegun, *Pocketbook of Mathematical Functions*. Harri Deutsch, Thun, 1984.

[2] N. Johnson and S. Kotz, *Continuous Univariate Distributions*. J. Wiley, 1970.

[3] H. Bohr, "Zur Theorie der Fast Periodischen Funktionen: I. Eine Verallgemeinerung der Theorie der Fourierreihen," *Acta Mathematica*, vol. 45, no. 1, pp. 29–127, 1925.

[4] H.-G. Beyer and D. Arnold, "The Steady State Behavior of $(\mu/\mu_I, \lambda)$-ES on Ellipsoidal Fitness Models Disturbed by Noise," in *GECCO'03: Proceedings of the Genetic and Evolutionary Computation Conference*, E. Cantú-Paz et al., Ed. Berlin, Germany: Springer, 2003, pp. 525–536.