On a Population Sizing Model for Evolution Strategies Optimizing the Highly Multimodal Rastrigin Function

Lisa Schönenberger
lisa.schoenenberger@fhv.at
Vorarlberg University of Applied Sciences
Research Center Business Informatics
6850 Dornbirn, Austria

Hans-Georg Beyer
hans-georg.beyer@fhv.at
Vorarlberg University of Applied Sciences
Research Center Business Informatics
6850 Dornbirn, Austria

ABSTRACT

A model is presented that allows for the calculation of the success probability by which a vanilla Evolution Strategy converges to the global optimizer of the Rastrigin test function. As a result a population size scaling formula will be derived that allows for an estimation of the population size needed to ensure a high convergence security depending on the search space dimensionality.

CCS CONCEPTS

• **Theory of computation** → *Random search heuristics*; • **Mathematics of computing** → **Bio-inspired optimization**.

KEYWORDS

Evolution Strategies, global optimization, multi-modal objective function, global convergence, population sizing

ACM Reference Format:
Lisa Schönenberger and Hans-Georg Beyer. 2023. On a Population Sizing Model for Evolution Strategies Optimizing the Highly Multimodal Rastrigin Function. In *Genetic and Evolutionary Computation Conference (GECCO '23), July 15–19, 2023, Lisbon, Portugal.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3583131.3590451

1 INTRODUCTION

Finding global optimal solutions in highly multimodal real-valued fitness landscapes by means of Evolution Strategies (ES) [8] depends on the choice of algorithm-specific parameters. Considering highly multimodal test functions such as Rastrigin, Ackley, Fletcher-Powell, and Bohachevsky to name a few, the probability of success of the ES locating the global optimizer is strongly influenced by the choice of the population size. This observation has been made already in [10] regarding the CMA-ES [11], however, this also holds for simple $(\mu/\mu_I, \lambda)$-ES using isotropic mutations in conjunction with σ self-adaptation (σSA) or cumulative stepsize adaptation (CSA) for mutation strength control. Consider the minimization problem $\hat{\mathbf{y}} := \operatorname{argmin}_{\mathbf{y}} F(\mathbf{y}),\ \mathbf{y} \in \mathbb{R}^N$, where $\hat{\mathbf{y}}$ is the global minimizer of F and N is the search space dimensionality. In order to reach $\hat{\mathbf{y}}$ with high probability it seems intuitively clear that sufficiently large population sizes for both the number of parents μ and offspring λ are needed. Furthermore, if the number of local minima increases with the search space dimensionality N it seems plausible that also the population sizes, i.e., μ and λ should increase as well. Since the number of local minima increases exponentially with dimensionality N one could expect that the population size should increase in a similar manner as it would be the case of the number of multi-starts in classical non-linear numerical optimization strategies. Therefore, the empirical findings in [10] came as a big surprise: In most of the cases considered the population sizes did not scale exponentially with N, but seemingly in-between $O(N)$ and $O(N^2)$ (with the Griewank test function as an exception where the population sizes even decreased with N).

What can be learned from these experimental observations? First of all, the ES *does not* perform some kind of gradient following strategy to locate the global optimizer as sometimes claimed [16, p. 75f]. This raises the question how the ES does locate a global optimizer under a huge number of local optima. Furthermore, from the viewpoint of algorithmic efficiency the question of computational complexity would be of interest here. However, this question is intimately connected to the question how to choose the population size since using a population size too small the probability to reach the global optimizer will be very small while choosing the population size too large would be a waste of computing resources. Therefore, this paper is devoted to the derivation of a population sizing equation.

The theoretical analysis of the behavior of ES on highly multimodal test functions is still in its infancy. There are first attempts to extend the progress rate analysis [6] to the Rastrigin function [15]. While this approach is able to take into account many specific details of the ES algorithms and also the influence of the population size parameters μ and λ it is still restricted to the derivation of mean value dynamics. Another approach models the ES mutation process as some kind of convolution. As has been shown in [14] the convolution of Rastrigin like functions with a Gaussian kernel can transform the original non-convex minimization problem into a convex one depending on the kernel parameter (being the mutation strength σ). However, the convolution is an N-fold integration performed only approximately by the ES mutation sampling process. That is, the question of how many samples are needed to get a reliable convex result cannot be easily answered. Therefore, the question still remains how to choose the population size.

It is the goal of this paper to develop a model that describes the convergence behavior of the $(\mu/\mu_I, \lambda)$-ES to the global optimizer of the Rastrigin function. As a result, a population sizing equation will

GECCO '23, July 15–19, 2023, Lisbon, Portugal

ACM ISBN 979-8-4007-0119-1/23/07…$15.00
https://doi.org/10.1145/3583131.3590451

be obtained that scales like $O(\sqrt{N}\ln N)$. That is, for the Rastrigin test function the population size scales even *sublinearly* with the search space dimensionality N. The remainder of this paper is organized as follows. First, the ES algorithms to be considered are briefly reviewed. In Section 3 the Rastrigin test function will be introduced. In Section 4 the convergence model will be developed and the success probability will be derived. Section 5 is devoted to the derivation of the population sizing equation. In the concluding Section 6 a summary will be given and an outlook regarding future research will be presented.

2 ES-ALGORITHMS

It is assumed that the reader is acquainted with the basic $(\mu/\mu_I, \lambda)$-ES algorithms and the order statistics notation "$m;\lambda$" used. The control of the strength σ of the isotropic Gaussian mutations used is done by either σ self-adaptation (σSA), see Alg. 1, or cumulative stepsize adaptation (CSA), see Alg. 2. The performance of the algorithms depends on the choice of the learning parameter τ and the cumulation time constant $1/c$ and D, respectively, where $D = 1/c$ has been chosen. The standard choice of the learning parameter $\tau = 1/\sqrt{2N}$ [12] guarantees optimal performance on the sphere model. As for the choice of c in the CSA-ES, $1/N$ to $1/\sqrt{N}$ defines an admissible range [2, 9] where the latter results in faster convergence rate at the price of lower global success probability P_s on the Rastrigin function (1).

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

```
Initialize $\left(\mathbf{y}^{(0)}, \sigma^{(0)}, \sigma_\text{stop}, g = 0\right)$
repeat
    for $l = 1$ to $\lambda$ do
        $\tilde{\sigma}_l = \sigma^{(g)} \mathrm{e}^{\tau \mathcal{N}(0,1)}$        ▷ mutate parental $\sigma$
        $\tilde{\mathbf{y}}_l = \mathbf{y}^{(g)} + \tilde{\sigma}_l\,(\mathcal{N}(0,1), \dots, \mathcal{N}(0,1))$        ▷ mutate $\mathbf{y}$
        $\tilde{F}_l = F\left(\tilde{\mathbf{y}}_l\right)$        ▷ evaluate offspring
    end for
    Sort Individuals 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 $\mu$ best $\tilde{\mathbf{y}}$
    $\sigma^{(g)} = \frac{1}{\mu}\sum_{m=1}^{\mu} \tilde{\sigma}_{m;\lambda}$        ▷ recombine the $\mu$ best $\tilde{\sigma}$
until $\sigma^{(g)} < \sigma_\text{stop}$
```

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

```
Initialize $\left(\mathbf{y}^{(0)}, \sigma^{(0)}, \sigma_\text{stop}, \mathbf{s} = \mathbf{1}, g = 0\right)$
repeat
    for $l = 1$ to $\lambda$ do
        $\tilde{\mathbf{z}}_l = (\mathcal{N}(0,1), \dots, \mathcal{N}(0,1))$        ▷ generate 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 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 $\mu$ best $\tilde{\mathbf{y}}$
    $\mathbf{s} = (1-c)\mathbf{s} + \sqrt{\mu c\,(2-c)}\,\frac{1}{\mu}\sum_{m=1}^{\mu} \tilde{\mathbf{z}}_{m;\lambda}$        ▷ update $\mathbf{s}$-path
    $\sigma^{(g)} = \sigma^{(g-1)}\exp\left(\frac{\|\mathbf{s}\|^2 - N}{2DN}\right)$        ▷ update $\sigma$, see [2, p.13]
until $\sigma^{(g)} < \sigma_\text{stop}$
```

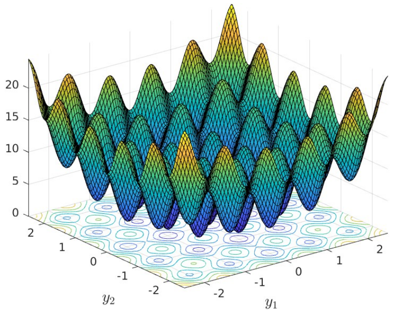

Figure 1: 3D-plot of an $N = 2$ dimensional Rastrigin function (1) with $\alpha = 2\pi$ and $A = 3$.

3 THE RASTRIGIN FUNCTION

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

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

where the parameter $A > 0$ denotes the oscillation amplitude and α denotes the frequency. Unless otherwise stated, the parameters $A = 1$ and $\alpha = 2\pi$ are used in all experiments. The global optimizer located at $\hat{\mathbf{y}} = \mathbf{0}$ is surrounded by $\kappa^N - 1$ local minima (e.g., for $\alpha = 2\pi$, $A = 1$: $\kappa = 7$ and for $\alpha = 2\pi$, $A = 10$: $\kappa = 63$). Figure 1 shows an example 3D-plot of the Rastrigin function. Looking at the contour map in Fig. 2 that includes the global optimizer at $\hat{\mathbf{y}} = \mathbf{0}$ one sees a squared domain (bounded by green lines) in which the negative gradient flow (expressed by small arrows) is directed towards the global minimizer. That is, a gradient strategy initialized

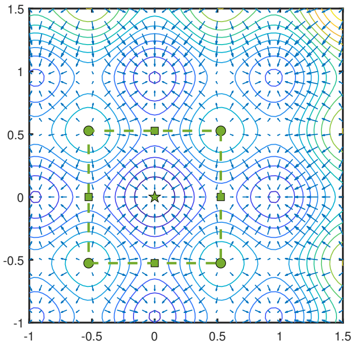

Figure 2: Global attractor region of the Rastrigin function for $N = 2$ (green dashed square), $\alpha = 2\pi$ and $A = 1$. The star shows the global optimizer, squares the nearest stationary points, and circles the farthest stationary points. Arrows show the negative gradient flow.

in this *global attractor region* would converge to the global attractor. The global attractor region is defined by the hypercube

$$\mathcal{A}_0 := [-\Delta_0, \Delta_0]^N, \tag{2}$$

where Δ_0 is the distance from the global optimizer $\mathbf{0}$ (the star) to the nearest stationary point(s) (the small filled squares in Fig. 2). The value of Δ_0 is determined by a non-linear equation. One finds for $A\alpha \gg 2$ asymptotically (see Appendix A)

$$\Delta_0 \simeq \frac{A\alpha\pi}{A\alpha^2 - 2}. \tag{3}$$

Unlike gradient strategies, it cannot be guaranteed that the $(\mu/\mu_I, \lambda)$-ES converges globally if the parental centroid $\mathbf{y}$ is in $\mathcal{A}_0$. Especially, parents $\mathbf{y}$ located in the vicinity of the corners of $\mathcal{A}_0$ will produce better offspring only with a probability of about 2^{-N}, thus, requiring an exponentially large population size for improvements. On the other hand, parents in the vicinity of the stationary points can even be located outside $\mathcal{A}_0$ and still produce better offspring allowing for convergence to the global optimizer. That is, for fixed values of A and α, the global attractor domain of an ES denoted by $\mathcal{A}_{\mathrm{ES}}$, depends on the strategy-specific parameters such as the truncation ratio $\vartheta := \mu/\lambda$, the actual mutation strength σ, the learning parameter τ and the time constant $1/c$, respectively. Simplifications are needed to get a manageable model $\mathcal{A}_{\mathrm{ES}}$. It turns out that

$$\mathcal{A}_{\mathrm{ES}} = [-(\Delta_0 + \varepsilon), \Delta_0 + \varepsilon]^N, \tag{4}$$

can serve as such a model. ε is a small correction term that varies depending on the specific strategy.

In Fig. 3 the dynamics of the distance of the parental centroid to the global optimizer, i.e., $R(g) := \|\mathbf{y}^{(g)}\|$ is displayed for 200 independent runs of the $(100/100_I, 200)$-σSA-ES, Alg. 1, on the Rastrigin function. One observes a certain percentage of runs getting trapped in local minima. The other runs converge to the global optimizer. Similar graphs can be obtained when running the CSA-ES, Alg. 2. Determining the *success probability* P_s by which the ES approaches the global optimizer will be the task of the following section. Having a closer look at the dynamics, one sees that there are basically three phases in the evolution process. The first phase can be observed when the initial parental centroid is initialized far enough from the global optimizer. In that case, the ES "sees" basically a sphere model. The influence of the cosine terms in (1) can be neglected and one observes a linear convergence behavior. Getting closer to the global optimizer, the y_i^2 parts get comparable to the magnitude of the cosine terms, defining the phase II. The influence of the local attractors becomes dominant, slowing down the speed by which the global optimizer is approached. This slowdown can also be seen in the mean value dynamics displayed in Fig. 4. There, the averaged dynamics of the *successful* individual ES runs is displayed, symbolized by angular brackets. At the end of phase II, the ES is either confined in a local attractor or it has hit the global attractor $\mathcal{A}_{\mathrm{ES}}$. This defines the begin of phase III where one observes again increased linear convergence order.

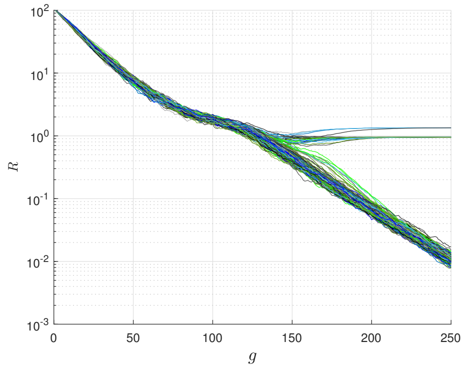

Figure 3: Residual distance dynamics for 200 $(100/100_I, 200)$-σSA-ES runs with $\tau = 1/\sqrt{2N}$ for $N = 100$. For each run, the ES was initialized randomly at an expected residual distance $R(0) = 100$. The success probability is $P_\mathrm{s} = 0.88$.

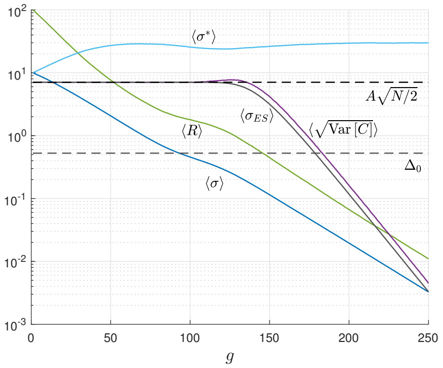

Figure 4: Mean value dynamics of the $(100/100_I, 200)$-σSA-ES with $\tau = 1/\sqrt{2N}$ for $N = 100$ derived from the *successful* runs displayed in Fig. 3. In addition to the R dynamics, the mutation strength σ and its normalization σ^* are displayed. The distance $\Delta_0 \approx 0.527$, cf. Eq. (3), (dashed line) indicates the $R = \|\mathbf{y}\|$ below which any parental component $y_i \in [-\Delta_0, \Delta_0]$. The remaining curves regarding σ_{ES}, cf. Eq. (30), and $\langle\sqrt{\mathrm{Var}[C]}\rangle$ are discussed in Sect. 4.1. The asymptote $A\sqrt{N/2}$ indicates the maximum of the σ_{ES} curve.

In addition to the mean value dynamics of R, Fig. 4 shows also the dynamics of the mutation strength and its normalization

$$\sigma^* := \sigma N/R. \tag{5}$$

As one can see, the initial $\sigma = 10$ was (intentionally) chosen too small. Therefore, self-adaptation increased the normalized σ^* to reach typical sphere model values. Entering the phase II, one observes a certain decrease of σ^*. This reflects the tendency of getting trapped in local attractors. This phase ends at about generation $g = 125$. At $g = 150$ it already holds $R < \Delta_0$ and the ES evolves safely in the global attractor. The central question to be answered in the next section concerns the conditions under which the global attractor is reached with the success probability P_s.

4 THE SUCCESS PROBABILITY MODEL

4.1 The Frozen Noise Model

The Rastrigin function (1) can be divided into two parts. The first is the sphere function $R^2 := \sum_{i=1}^{N} y_i^2$ indicating the squared distance to the global optimizer. The second

$$C(\mathbf{y}) := NA - A\sum_{i=1}^{N}\cos\left(\alpha y_i\right) \tag{6}$$

is called *cosine* part. It describes the oscillations of the Rastrigin function where $C(\mathbf{y}) \in [0, 2NA]$.

In the case where the distance to the global optimizer R is very large, i.e. $R^2 \gg NA$ (being phase I), the perturbations caused by the cosine parts are relatively small compared to the sphere model part R^2. In this case the behavior of the ES on Rastrigin is similar to that of the sphere model. In real runs of the σSA-ES (cf. Fig. 4) and the CSA-ES one observes σ^* values, Eq. (5), that are in an order of magnitude of the asymptotically optimal sphere model value $\sigma^* = \mu c_{\mu/\mu,\lambda}$ [4, 13]. $c_{\mu/\mu,\lambda}$ is the progress coefficient [6] which is in the range of roughly $[0.8, 1.2]$ for truncation ratios $\vartheta \in [1/4, 1/2]$ and sufficiently large λ. Due to (5) it holds $\sigma = \sigma^* R/N = \mu c_{\mu/\mu,\lambda}R/N$. Since $\cos(\alpha y_i)$ is periodic, the minimum distance between two of its maxima is at $2\pi/\alpha$. This determines the extent of the local attractor regions. As long as $\sigma \gtrapprox 2\pi/\alpha$, there will be a high probability to jump over those regions. This yields the condition $\mu c_{\mu/\mu,\lambda}R/N \gtrapprox 2\pi/\alpha$ that will be fulfilled for sufficiently large R, i.e., if one is far apart from the global optimizer.

If the ES is getting closer to the global optimizer (phase II) the influence of the cosine parts becomes more pronounced compared to the R^2-part in (1). The ripples caused by the cosine parts (6) can be interpreted as *frozen noise*. Thus, the evolution process can be modeled as optimizing a noisy sphere model

$$F(\mathbf{y}) = R^2 + NA + \sigma_{\mathrm{ES}}(R)\mathcal{N}(0,1) \quad \text{with} \quad R = \|\mathbf{y}\|, \tag{7}$$

where σ_{ES} is the noise strength depending on the distance R to the global optimizer. This needs further justifications: Under the assumption of a sufficient large mutation strength σ the ES performs a restricted random walk that can be interpreted as *exploration*. As described in [5] the *exploitation step* towards the global optimizer is only of order $1/\sqrt{N}$ compared to the *exploration step*, i.e., the step perpendicular to the optimizer. This is a random sampling process of global kind (i.e., it is not confined in a local attractor region provided that σ is sufficiently large). The assumption of $\mathcal{N}(0,1)$ Gaussian noise in (7) can be justified by considering the cosine part $C(\mathbf{y})$, Eq. (6), as a sum of independent random variables $\cos\left(\alpha y_i\right)$ for which the central limit theorem of statistics holds. The standard deviation $\sigma_{\mathrm{ES}}(R)$ of this noise produced by the offspring $\tilde{\mathbf{y}}$

$$\sigma_{\mathrm{ES}} = \sqrt{\mathrm{Var}\left[C\right]} = A\sqrt{\mathrm{Var}\left[\sum_{i=1}^{N}\cos(\alpha\tilde{y}_i)\right]} \tag{8}$$

will be derived in Appendix B, it is also displayed in Fig. 4 both experimentally as $\langle\sqrt{\mathrm{Var}[C]}\rangle$ and by a theoretical estimate σ_{ES} (for further discussion, see below).

Accepting the noise model (7), converging to the global optimizer of Rastrigin is equivalent to optimize a noisy sphere model. It is important to note that in the case of constant noise strength σ_{ES}, an ES optimizing a noisy sphere reaches a steady state R-distribution with $R_{\mathrm{st}} := \mathrm{E}\left[R\right] \neq 0$ and

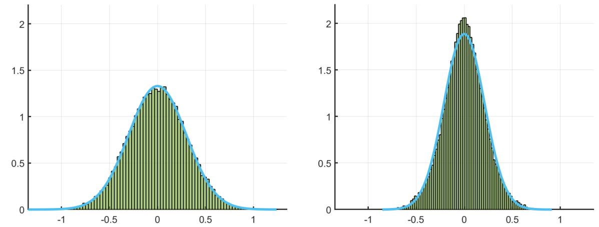

Figure 5: Histogram of all individual components y_i at distance R to the optimum for $R = 3$ (left graph) and $R = 2.1$ (right graph) for σSA-ES runs $N = 100,\ \mu = 50,\ \vartheta = 0.5$. The blue line is the pdf of the $\mathcal{N}\left(0, \frac{R^2}{N}\right)$ variate.

$$R_{\mathrm{st}} \simeq \sqrt{\frac{\sigma_{\mathrm{ES}}N}{4\mu c_{\mu/\mu,\lambda}}}, \tag{9}$$

see [3, 12]. That is, the parental centroid $\mathbf{y}$ calculated in Line 10 of Alg. 1 and 2 has the expected distance R_{st} to the global optimizer. Furthermore, each component of $\mathbf{y}$ is normally distributed [7]

$$y_i = (\mathbf{y})_i \sim \mathcal{N}(0, R_{\mathrm{st}}^2/N). \tag{10}$$

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. Figure 5 shows two examples of the distribution of a single parent component at two different distances R to the global optimizer. The histogram on the rhs has been obtained for an R in the critical range where the ES has a higher probability getting trapped into one of the local attractors.

With model (7), the evolution of the ES on Rastrigin can be analyzed as a noisy minimization problem where the ES reaches the vicinity of the global optimizer up to a distance R_{st}. If this distance is sufficiently small, the ES has reached the global attractor region $\mathcal{A}_{\mathrm{ES}}$ and can converge successfully (phase III). Since σ_{ES} is bounded (see Fig. 4) one can infer from Eq. (9) that global convergence mainly depends on the choice of a sufficiently large μ (assuming $\vartheta = \text{const.}$).

4.2 Estimating the Success Probability

In order to have convergence to the global optimizer, the parental centroid has to be in the global attractor region, i.e., $\mathbf{y} \in \mathcal{A}_{\mathrm{ES}}$, Eq. (4). Using (4), the success probability P_{s} is therefore

$$\begin{aligned} P_{\mathrm{s}} &= \Pr[\mathbf{y} \in \mathcal{A}_{\mathrm{ES}}] \\ &= \Pr\left[(-\Delta_0 - \varepsilon \le y_1 \le \Delta_0 + \varepsilon) \wedge \cdots\right. \\ &\qquad \left.\cdots \wedge (-\Delta_0 - \varepsilon \le y_N \le \Delta_0 + \varepsilon)\right] \\ &= \Pr\left[-\Delta_0 - \varepsilon \le y \le \Delta_0 + \varepsilon\right]^N. \end{aligned} \tag{11}$$

Here, the independence of the parental centroid components in the steady state has been used and y is distributed according to (10). Using

$$\sigma_{\mathrm{st}} := \frac{R_{\mathrm{st}}}{\sqrt{N}} \overset{(9)}{=} \sqrt{\frac{\sigma_{\mathrm{ES}}}{4\mu c_{\mu/\mu,\lambda}}}, \tag{12}$$

for the standard deviation in (10), one gets for a single component

$$\Pr\left[-\Delta_0 - \varepsilon \le y \le \Delta_0 + \varepsilon\right] = \Pr\left[-\frac{\Delta_0+\varepsilon}{\sigma_{\mathrm{st}}} \le z \le \frac{\Delta_0+\varepsilon}{\sigma_{\mathrm{st}}}\right]$$

$$= \Phi\left(\frac{\Delta_0+\varepsilon}{\sigma_{\mathrm{st}}}\right) - \Phi\left(-\frac{\Delta_0+\varepsilon}{\sigma_{\mathrm{st}}}\right) \qquad (13)$$

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

$$P_{\mathrm{s}} = \left[2\Phi\left(\frac{\Delta_0+\varepsilon}{\sigma_{\mathrm{st}}}\right) - 1\right]^N. \qquad (14)$$

Due to Eq. (12), σ_{st} depends on σ_{ES} of the offspring generated frozen noise. This standard deviation which depends on the parental R will be derived for the CSA-ES in Appendix B and is displayed in Fig. 4 together with an experimentally obtained curve for the σSA-ES labeled as $\langle\sqrt{\mathrm{Var}[C]}\rangle$. Apart from small deviations caused by the different offspring $\tilde{\sigma}_l$ values in σSA-ES which do not exist for CSA-ES the general curve tendency is the same: The frozen noise strength σ_{ES} stays constant and only starts to drop if the distance R is of the order of Δ_0. That is, even if the ES enters the global attractor region $\mathcal{A}_{\mathrm{ES}}$, σ_{ES} is still in the vicinity of its maximum value. Therefore, one can replace σ_{ES} by its maximum value $\sigma_{\mathrm{ES}} = A\sqrt{N/2}$. Plugging this into (12) yields

$$\sigma_{\mathrm{st}} = \sqrt{\frac{A\sqrt{N}}{4\sqrt{2}\mu c_{\mu/\mu,\lambda}}}. \qquad (15)$$

If inserted into (14), one finally obtains the success probability formula

$$P_{\mathrm{s}} = \left[2\Phi\left(\sqrt{\frac{4\sqrt{2}\mu c_{\mu/\mu,\lambda}}{A\sqrt{N}}}(\Delta_0+\varepsilon)\right) - 1\right]^N. \qquad (16)$$

4.3 Comparison with Experiments

The predictive quality of the success probability formula (16) with (3) is evaluated for the σSA-ES and the CSA-ES in Fig. 6 using $\varepsilon = 0$. As for the σSA-ES the learning parameter $\tau = 1/\sqrt{2N}$ was used and for the CSA-ES $c = 1/\sqrt{N}$ was chosen. Each data point was obtained by at least 500 independent runs of the ES. As expected, there are differences between experimental data and the predictions. However, the general tendencies are well covered by (16). One can obtain better predictions in the case of the σSA-ES if one chooses $\tau = 1/\sqrt{4N}$ (not shown in this paper).

In order to improve the predictions one needs the correction term $\varepsilon \neq 0$ in (16). The results are presented in Fig. 7 where ε was chosen according to Fig. 8. The ε values were determined experimentally by minimizing the sum of the squares from the differences between (16) and the experimental values. As one can see, the model of a global success domain $\mathcal{A}_{\mathrm{ES}}$ in terms of (4) provides success curves that do well agree with the real ES runs. Therefore, Eq. (16) can be used to derive a population sizing formula and to evaluate its scaling behavior.

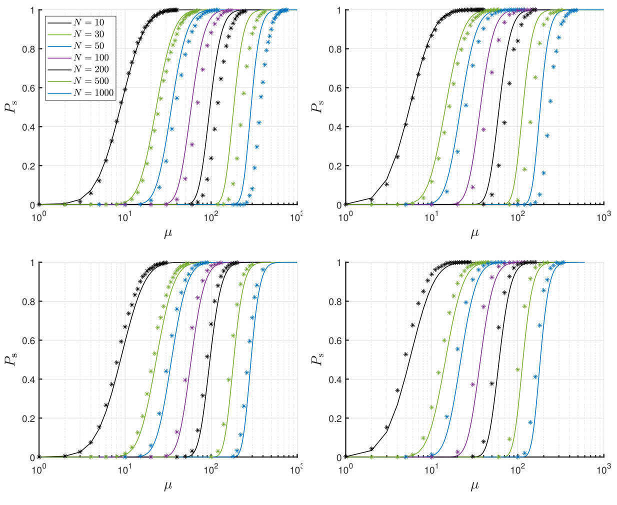

Figure 6: Success P_{s} vs. population size μ predicted by Eq. (16) with $\varepsilon = 0$. Experimental results are displayed by the stars. σSA-ES in top row with $\vartheta = 1/2$ (left) and $\vartheta = 1/4$ (right). CSA-ES in bottom row with $\vartheta = 1/2$ (left) and $\vartheta = 1/4$ (right).

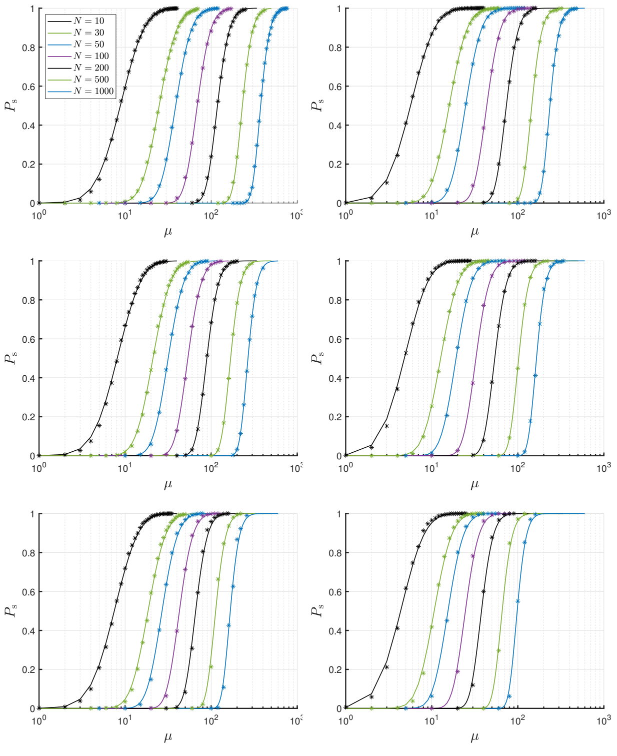

Figure 7: P_{s} predicted by Eq. (16) with ε values according to Fig. 8 for $\vartheta = 1/2$ (left column) and $\vartheta = 1/4$ (right column). Upper row displays the σSA-ES with $\tau = 1/\sqrt{2N}$. Middle row displays the CSA-ES with $c = 1/\sqrt{N}$. Bottom row displays the CSA-ES with $c = 1/N$.

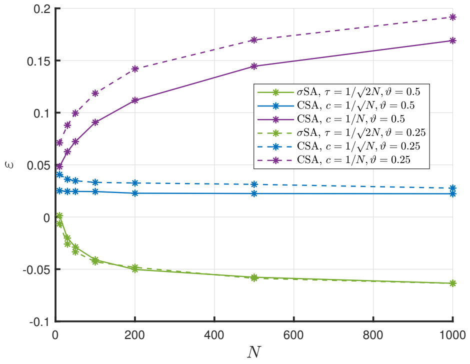

Figure 8: The dependence of ε on N such that the deviations between Eq. (16) and the experimental values are minimal in Fig. 7.

5 POPULATION SIZING

5.1 Derivation of Parent Population Size

The central question of this paper regards the choice of μ and λ that guarantees convergence of the ES towards the global optimizer. Given a fixed truncation ratio ϑ, it suffices to derive a formula that predicts $\mu(P_s)$. Solving Eq. (16) for μ under the assumption $c_{\mu/\mu,\lambda} \simeq f(\vartheta)$ [6, p.249] yields after a simple calculation

$$\mu \simeq \frac{A}{\sqrt{2}c_{\mu/\mu,\lambda}} \frac{\sqrt{N}}{4(\Delta_0+\varepsilon)^2} \left[\Phi^{-1}\left(\frac{1}{2}+\frac{1}{2}P_s^{\frac{1}{N}}\right)\right]^2, \qquad (17)$$

where Φ^{-1} is the quantile function of the standard normal distribution.

5.2 Comparison with Experiments

Figure. 9 compares the prediction of the population size Eq. (17) depending on N and P_s for $\varepsilon = 0$ with experiments. 300 runs were executed to obtain the experimental data displayed by the markers (+, ×, and ∘). While the theoretical predictions of (17) with $\varepsilon = 0$ differ from the experimental values, Eq. (17) predicts the general functional tendency well. The deviations are due to the different sizes of $\mathcal{A}_{\text{ES}}$ encoded in ε. As for the σSA-ES the population size is underestimated in accordance with the negative values of ε in Fig. 8. In contrast to that, $\varepsilon = 0$ results in an overestimation of the population size for the CSA-ES in both cases $c = 1/N$ and $c = 1/\sqrt{N}$.

As will be shown in Appendix C, (17) with $\varepsilon = 0$ behaves asymptotically like

$$\mu = O\left(\sqrt{N}\ln(N)\right). \qquad (18)$$

Respective curves proportional to $\sqrt{N}\ln(N)$ are displayed by gray dashed-dotted curves in Fig. 9. As one can infer from the data in Fig. 9 and 8, the population size scaling of $O(\sqrt{N}\ln(N))$ can serve as an upper bound for the CSA-ES versions considered. As for the σSA-ES with $\tau = 1/\sqrt{2N}$ the growth rate is slightly above $\sqrt{N}\ln(N)$ for $P_s = 0.5$. However, corrections to the scaling law $\sqrt{N}\ln(N)$ cannot be obtained indicating the limits of the model used that does not take into account the influence of τ and c, respectively, on the σ adaptation.

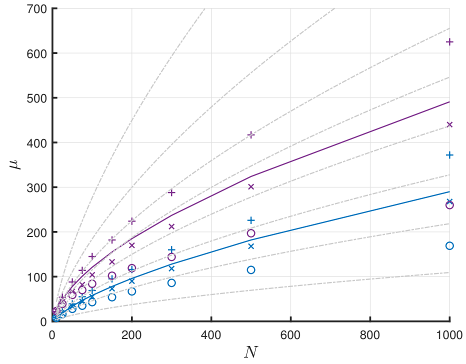

Figure 9: Population size for $P_s = 50\%$ (blue curve and markers) and $P_s = 99\%$ (purple curve and markers). The data points obtained by ES runs represent the σSA-ES (+) with $\tau = 1/\sqrt{2N}$, the CSA-ES with $c = 1/\sqrt{N}$ (×), and with $c = 1/N$ (∘). Gray dashed-dotted lines show functions $\propto \sqrt{N}\ln(N)$ for comparison.

Besides the N-scaling the influence of the Rastrigin parameters A and α on the population sizing is of interest. To this end, a closer look at (3) reveals that for $A \to \infty \quad \Rightarrow \quad \Delta_0 \to \pi/\alpha$. Thus, the influence of A in (17) becomes linear for sufficiently large A. This can be verified by the experiments presented in the left column of Fig. 10 for both the σSA-ES and the CSA-ES. ε in (17) was calculated by minimizing the difference between the slope of the experimental values and those of Eq. (17) for values larger than $A = 4$.

In order to derive the scaling behavior w.r.t. α it is important, to realize, that $\alpha \to \infty \;\Rightarrow\; \Delta_0 \to \pi/\alpha$. That is, the extension of $\mathcal{A}_{\text{ES}}$ shrinks with increasing α. Therefore, ε must shrink analogously $\varepsilon \to \varepsilon/\alpha$. As a result the term in (17) yields $(\Delta_0+\varepsilon)^{-2} \simeq \alpha^2/(\pi+\varepsilon)^2$. Therefore, the population size must grow quadratically with α. This is experimentally confirmed on the rhs of Fig. 10. ε in (17) was determined experimentally for values of α larger than 2.5π by minimizing the sum of the squares from the differences between the experimental values and Eq. (17) using $(\Delta_0+\varepsilon)^{-2} = \alpha^2/(\pi+\varepsilon)^2$.

6 CONCLUSIONS

Reaching the global minimum of Rastrigin, a function that has a huge number of local minima, is a hopeless endeavor when tackled by gradient based nonlinear optimization techniques. Yet, Evolution Strategies are able to find the global optimizer provided that the population size has been chosen sufficiently large. The question answered in this paper is how large is "sufficiently large?" To this end, a model has been developed that allows for the calculation of the probability P_s of reaching the global optimizer of the Rastrigin function depending on the population size parameters μ and λ. The basic idea was the separation of Rastrigin into a Sphere model part and a noise term and to apply the theory of ES performance on noisy Sphere models. Given a fixed noise strength, an ES with fixed μ, λ cannot reach the optimizer of the Sphere model arbitrarily close. Instead, its parents fluctuate about the optimizer with an expected

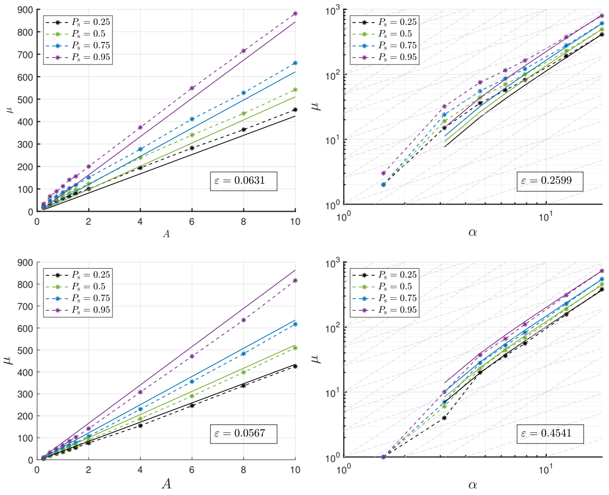

Figure 10: Scaling behavior of μ, Eq. (17), depending on the Rastrigin parameters A (left column) and α (right column). Top row represents the σSA-ES and the bottom row the CSA-ES with $c = 1/N$ and $\vartheta = 1/2$ optimizing the $N = 100$ case. Markers with dashed lines represent the experiments, where each data point was obtained by 500 independent runs. The grey dashed-dotted straight lines are linear and quadratic growth curves for comparison purpose.

distance R_{st} to the optimizer. However, if this R_{st} is sufficiently small such that the parents hit the global attractor region, the ES will converge to the global minimum. The model abstracts from the details of the ES used, i.e., whether σSA or CSA-ES are used. Therefore, the influence of those strategy specific parameters as τ and c are not incorporated in the model. Yet, the model yields remarkable predictions. Basically the parental population size scales like $O(\sqrt{N}\ln(N))$ in order to get reliable convergence to the global optimizer. That is, the growth is sublinear and slightly above $\sqrt{N}$. This is in contrast to gradient-based restart strategies that need an exponential number of restarts. Besides the influence of the search space N, the influence of the Rastrigin parameters on the necessary population size came out of the analysis. While for sufficiently large A the population size scales linearly the influence of the spacial frequency α is quadratic.

It seems that the analysis method presented can be extended to other highly multimodal test functions provided that a reasonable noisy Sphere model can be constructed. This will be a future road of research.

ACKNOWLEDGMENTS

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

REFERENCES

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

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

[3] D.V. Arnold and H.-G. Beyer. 2002. Performance Analysis of Evolution Strategies with Multi-Recombination in High-Dimensional $\mathbb{R}^N$-Search Spaces Disturbed by Noise. *Theoretical Computer Science* 289 (2002), 629–647.

[4] D.V. Arnold and H.-G. Beyer. 2004. Performance Analysis of Evolutionary Optimization With Cumulative Step Length Adaptation. *IEEE Trans. Automat. Control* 49, 4 (2004), 617–622.

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

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

[7] H.-G. Beyer, D.V. Arnold, and S. Meyer-Nieberg. 2005. A New Approach for Predicting the Final Outcome of Evolution Strategy Optimization under Noise. *Genetic Programming and Evolvable Machines* 6, 1 (2005), 7–24.

[8] H.-G. Beyer and H.-P. Schwefel. 2002. Evolution Strategies: A Comprehensive Introduction. *Natural Computing* 1, 1 (2002), 3–52.

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

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

[11] N. Hansen, S.D. Müller, and P. Koumoutsakos. 2003. Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES). *Evolutionary Computation* 11, 1 (2003), 1–18.

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

[13] S. Meyer-Nieberg and H.-G. Beyer. 2005. On the Analysis of Self-Adaptive Recombination Strategies: First Results. In *Proceedings of the CEC'05 Conference.* IEEE, Piscataway, NJ, 2341–2348.

[14] N. Müller and T. Glasmachers. 2021. Non-local optimization: imposing structure on optimization problems by relaxation. In *Foundations of Genetic Algorithms, 16.* ACM, 1–10. https://doi.org/10.1145/3450218.3477307

[15] A. Omeradzic and H.-G. Beyer. 2022. Progress Rate Analysis of Evolution Strategies on the Rastrigin Function: First Results. In *Parallel Problem Solving from Nature XVII*, H. Aguirre et al. (Ed.). Springer, Berlin, 499–511. DOI: 10.1007/978-3-031-14721-0_35.

[16] I. Rechenberg. 1994. *Evolutionsstrategie '94.* Frommann-Holzboog Verlag, Stuttgart.

A DERIVATION OF Δ_0

The saddle point Δ_0 nearest to the global optimizer $\hat{\mathbf{y}} = \mathbf{0}$ is defined by the 2nd zero of the derivative of (1) w.r.t. y_i, i.e.,

$$2y_i + A\alpha\sin(\alpha y_i) = 0 \quad\Leftrightarrow\quad y_i = \Delta_0. \tag{19}$$

This is a non-linear equation that must be solved numerically, however, it can be asymptotically approximated. For sufficiently large $A\alpha \gg 2$, Δ_0 is given by the 2nd zero of the sinus function $\alpha\Delta_0 = \pi$ leading to $\Delta_0 = \pi/\alpha$. This holds exactly for $A\alpha \to \infty$. For $A\alpha < \infty$ one can expand (19) in a Taylor series at $y_i = \pi/\alpha$ such that $\Delta_0 = \pi/\alpha + h$

$$\begin{aligned} 0 &= 2\left(\frac{\pi}{\alpha} + h\right) + A\alpha\sin\left(\alpha\left(\frac{\pi}{\alpha} + h\right)\right) \\ &= 2\frac{\pi}{\alpha} + 2h + A\alpha^2\cos(\pi)h + O(h^2). \end{aligned} \tag{20}$$

Neglecting higher order terms, one gets $h = \frac{2\pi}{\alpha(A\alpha^2-2)}$ and finally

$$\Delta_0 = \frac{\pi}{\alpha} + h = \frac{\pi}{\alpha} + \frac{2\pi}{\alpha(A\alpha^2 - 2)} = \frac{A\alpha\pi}{A\alpha^2 - 2}. \tag{21}$$

B DERIVATION OF σ_{ES}

The derivation of σ_{ES} defined by Eq. (8) differs for the σSA-ES and the CSA-ES. While in the CSA-ES each offspring individual $\tilde{\mathbf{y}}_l$ is generated from the parent $\mathbf{y}$ by the same mutation strength σ in Line 5 of Alg. 2, the σSA-ES produces each offspring $\tilde{\mathbf{y}}_l$ by an individual $\tilde{\sigma}_l$ in Lines 4 and 5 in Alg. 1. This may increase the variance due to the variation of σ (see the slight superelevation of

σ_{ES} in Fig. 4) and it complicates the calculations. While σ_{ES} can also be derived for the σSA-ES, it will be presented for the CSA-ES for brevity.

The offspring components in CSA-ES are generated according to $\tilde{y}_i = y_i + \sigma z_i$ where the z_i are iid $z_i \sim \mathcal{N}(0,1)$. Due to the stochastic independence of the z_i the sum in the variance expression in (8) can be taken out of the variance

$$\text{Var}\left[\sum_{i=1}^{N} cos(\alpha\tilde{y}_i)\right] = \sum_{i=1}^{N} \text{Var}\left[cos(\alpha y_i + \alpha\sigma z_i)\right]. \tag{22}$$

In the next step, the variance of a single component of the sum in (22) will be calculated using the variance formula

$$\text{Var}[cos(w_i)] = \text{E}[cos(w_i)^2] - \text{E}[cos(w_i)]^2. \tag{23}$$

where $w_i = \alpha\tilde{y}_i$ was substituted. Therefore, the first two moments of $cos(w_i)$ are needed where $w_i \sim \mathcal{N}(\alpha y_i, (\alpha\sigma)^2)$. According to [1, p.406], the characteristic function of w_i reads using Euler's formula

$$\begin{aligned}\text{E}[\text{e}^{\imath t w_i}] &= \exp\left(\imath\alpha y_i t - \frac{1}{2}(\alpha\sigma)^2 t^2\right)\\ &= \exp\left(-\frac{1}{2}(\alpha\sigma)^2 t^2\right)(\cos(\alpha y_i t) + \imath\sin(\alpha y_i t))\\ &= \text{E}[\cos(t w_i)] + \imath\text{E}[\sin(t w_i)]. \end{aligned}\tag{24}$$

Comparing the real parts in the 2nd and 3rd line of (24), one gets for $t = 1$

$$\text{E}[\cos(w_i)] = \exp\left(-\frac{1}{2}(\alpha\sigma)^2\right)\cos(\alpha y_i). \tag{25}$$

Taking the identity $\cos(w_i)^2 = \frac{1}{2} + \frac{1}{2}\cos(2w_i)$ into account, one gets with (25)

$$\begin{aligned}\text{E}[\cos(w_i)^2] &= \frac{1}{2} + \frac{1}{2}\text{E}[\cos(2w_i)]\\ &= \frac{1}{2} + \frac{1}{2}\exp\left(-\frac{1}{2}(2\alpha\sigma)^2\right)\cos(2\alpha y_i).\end{aligned}\tag{26}$$

Plugging (25) and (26) into (23), one obtains

$$\begin{aligned}\text{Var}[cos(\alpha\tilde{y}_i)] =& \frac{1}{2} + \frac{1}{2}\exp\left(-\frac{1}{2}(2\alpha\sigma)^2\right)\cos(2\alpha y_i)\\ &- \exp\left(-(\alpha\sigma)^2\right)\left(\frac{1}{2} + \frac{1}{2}\cos(2\alpha y_i)\right)\\ =& \frac{1}{2}\left(1 - \text{e}^{-(\alpha\sigma)^2}\right)\left(1 - \text{e}^{-(\alpha\sigma)^2}\cos(2\alpha y_i)\right)\end{aligned}\tag{27}$$

and finally for (22)

$$\begin{aligned}&\text{Var}\left[\sum_{i=1}^{N} cos(\alpha y_i + \alpha\sigma z_i)\right]\\ &= \frac{N}{2}\left(1 - \text{e}^{-(\alpha\sigma)^2}\right)\left(1 - \text{e}^{-(\alpha\sigma)^2}\frac{1}{N}\sum_{i=1}^{N}\cos(2\alpha y_i)\right).\end{aligned}\tag{28}$$

This result still depends on the actual location of the parent $\mathbf{y}$ in the search space. In order to get an aggregated measure of $\sigma_{\text{ES}}(R)$ that only depends on the distance R to the global minimizer, the condition in the variance expression (28) must be removed by taking the expected value w.r.t. the y_i. As has been explained in Sect. 4.1, under steady state conditions, Eq. (10) holds, i.e., $y_i \sim \mathcal{N}(0, R^2/N)$. Using the helper variable $w_i := 2\alpha y_i$, it holds $w_i \sim \mathcal{N}(0, (2\alpha R)^2/N)$ and (25) can be used yielding *mutatis mutandis* the expected value

$$\text{E}[\cos(2\alpha y_i)] = \exp\left(-\frac{1}{2}\frac{(2\alpha R)^2}{N}\right)\cos(0). \tag{29}$$

Inserting this in (28) one finally obtains for σ_{ES} in (8)

$$\sigma_{\text{ES}}(R) = A\sqrt{\frac{N}{2}}\sqrt{\left(1 - \text{e}^{-(\alpha\sigma)^2}\right)\left(1 - \text{e}^{-(\alpha\sigma)^2}\text{e}^{-2\frac{(\alpha R)^2}{N}}\right)}. \tag{30}$$

This result was already displayed in Fig. 4. The deviations to the experimental values $\langle\sqrt{\text{Var}[C]}\rangle$ are σSA-ES specific and are not observed in CSA-ES runs. As one can easily infer from (30), it holds

$$\sigma_{\text{ES}}(R) \le A\sqrt{\frac{N}{2}}. \tag{31}$$

That is, for the CSA-ES $\sigma_{\text{ES}}(R)$ takes its maximum of $A\sqrt{N/2}$. In the case of the σSA-ES slightly larger values are expected. This is due to the additional variance caused by the variance of the mutation strength within a single generation.

C ASYMPTOTIC POPULATION SIZE

In order to derive the N-asymptotics of (17) it is noted that apart from the $\sqrt{N}$, only

$$f(N) := \Phi^{-1}\left(\frac{1}{2} + \frac{1}{2}P_{\text{s}}^{\frac{1}{N}}\right) \tag{32}$$

needs closer scrutiny. Taking $\Phi(\cdot)$ on both sides yields

$$\Phi(f(N)) = \frac{1}{2} + \frac{1}{2}P_{\text{s}}^{\frac{1}{N}}. \tag{33}$$

The lhs is expressed by the first term of an asymptotic expansion [1, 26.2.12, p.408] $\Phi(x) \simeq 1 - \exp(-\frac{1}{2}x^2)/\sqrt{2\pi}x$ and the rhs is expanded in a Taylor series neglecting terms of $O(1/N^2)$

$$P_{\text{s}}^{\frac{1}{N}} = \exp\left(\frac{1}{N}\ln(P_{\text{s}})\right) = 1 + \frac{1}{N}\ln(P_{\text{s}}) + O(1/N^2). \tag{34}$$

Thus, (33) becomes

$$-\frac{\exp\left(-\frac{1}{2}f(N)^2\right)}{\sqrt{2\pi}f(N)} \simeq \frac{1}{2N}\ln(P_{\text{s}}). \tag{35}$$

Multiplying by $-\sqrt{2\pi}f(N)$ and taking the logarithm on both sides yields

$$\begin{aligned}-\frac{1}{2}f(N)^2 &\simeq \ln\left(\sqrt{2\pi}f(N)\frac{1}{2N}\ln(P_{\text{s}}^{-1})\right)\\ &= \ln(\sqrt{\pi/2}) + \ln(f(N)) - \ln(N) + \ln\ln(P_{\text{s}}^{-1}).\end{aligned}\tag{36}$$

Multiplying by -2, one gets

$$f(N)^2 \simeq 2\ln(N) - 2\ln(f(N)) - 2\ln\ln(P_{\text{s}}^{-1}) - \ln(\pi/2). \tag{37}$$

This equation can be used to generate successively better $f(N)^2$ approximations. However, for the purpose of this paper it suffices to note that for sufficiently large N the term $\ln(f(N))$ can be neglected compared to $\ln(N)$. Thus, from viewpoint of order notations one gets $f(N)^2 = O(\ln(N))$. Recalling that the square of (32), i.e., $f(N)^2$ appears in (17), one finally gets $\mu = O\left(\sqrt{N}\ln(N)\right)$.