Convergence Properties of the $(\mu/\mu_I, \lambda)$-ES on the Rastrigin Function

Amir Omeradzic
amir.omeradzic@fhv.at
Vorarlberg University of Applied Sciences
Research Center Business Informatics
Dornbirn, Austria

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

ABSTRACT

The highly multimodal Rastrigin test function is analyzed by deriving a new aggregated progress rate measure. It is derived as a function of the residual distance to the optimizer by assuming normally distributed positional coordinates around the global optimizer. This assumption is justified for successful ES-runs operating with sufficiently slow step-size adaptation. The measure enables the investigation of further convergence properties. For moderately large mutation strengths a characteristic distance-dependent Rastrigin noise floor is derived. For small mutation strengths local attraction is analyzed and an escape condition is established. Both mutation strength regimes combined pose a major challenge optimizing the Rastrigin function, which can be counteracted by increasing the population size. Hence, a population scaling relation to achieve high global convergence rates is derived which shows good agreement with experimental data.

CCS CONCEPTS

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

KEYWORDS

Evolution Strategy, global optimization, progress rate, multimodal function

ACM Reference Format:
Amir Omeradzic and Hans-Georg Beyer. 2023. Convergence Properties of the $(\mu/\mu_I, \lambda)$-ES on the Rastrigin Function. In *Proceedings of the 17th ACM/SIGEVO Conference on Foundations of Genetic Algorithms (FOGA '23), August 30-September 1, 2023, Potsdam, Germany*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3594805.3607126

1 INTRODUCTION

Evolution Strategies (ES) have proven to be well suited for the optimization of highly multimodal real-valued fitness functions due to their underlying stochastic nature. Test functions such as the Rastrigin function contain a huge number of local minima scaling exponentially with the search space dimensionality N. Sampling the search space and applying multiple restarts with standard gradient-based optimization algorithms quickly becomes unfeasible. On the other side, ES achieve high success rates for global convergence on certain multimodal functions, if sufficiently large population sizes are chosen. Experimental investigations in [9] indicate a non-exponential population scaling at most of $O(N^2)$ for the tested multimodal functions with the Rastrigin function scaling sub-linearly in N. However, there is little understanding of how the ES is exploring the fitness landscape to find the global optimizer without getting trapped in one of the local minima. This paper investigates the convergence properties on the Rastrigin function based on progress rate theory results of [11] and [12]. To this end, a new aggregated progress rate is introduced modeling the progress as a function of the residual distance to the optimizer. The obtained results are compared to real ES-runs operating with isotropic mutations and self-adaptation for step-size control.

In Sec. 2 the ES under investigation is introduced. In Sec. 3 the Rastrigin function is defined and averaging methods are discussed. Then, the method is applied to component-wise progress rate equations in Sec. 4 obtaining an aggregated measure. In Sec. 5 a population sizing relation is derived and compared to experimental results. Local attraction is discussed in Sec. 6 and a characteristic "escape" mutation strength is derived. Finally, in Sec. 7 conclusions and an outlook are provided.

2 THE ES-ALGORITHM

The ES under investigation, see Algorithm 1, consists of μ parents and λ offspring with truncation ratio $\vartheta = \mu/\lambda$ $(1 \leq \mu < \lambda)$. Selection of the $m = 1, ..., \mu$ best individuals (out of λ) is denoted by subscript "$m; \lambda$". Normally distributed isotropic mutations of strength σ are applied. Intermediate multi-recombination with equal weights is used to obtain the parental location $\mathbf{y}^{(g)} = [y_1^{(g)}, ..., y_N^{(g)}]$ in the N-dimensional search space for each generation g. For the σ-adaptation (self-adaptation) the offspring mutation strengths are chosen from a log-normal distribution with learning parameter τ. A smaller τ-value therefore yields a slower σ-adaptation, which will be important later. The default choice is $\tau = 1/\sqrt{2N}$, see [10], which ensures optimal performance on the sphere in the limit $N \to \infty$. Recombination also applies to selected σ-values.

3 RASTRIGIN FUNCTION AND AVERAGING

The Rastrigin test function f is defined for a real-valued search vector $\mathbf{y} = [y_1, ..., y_N]$ as

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

FOGA '23, August 30-September 1, 2023, Potsdam, Germany

ACM ISBN 979-8-4007-0202-0/23/08.
https://doi.org/10.1145/3594805.3607126

Algorithm 1 $(\mu/\mu_I, \lambda)$-σSA-ES

$g \leftarrow 0$
initialize $(\mathbf{y}^{(0)}, \sigma^{(0)})$
repeat
  for $l = 1, ..., \lambda$ **do**
    $\tilde{\sigma}_l \leftarrow \sigma^{(g)} \mathrm{e}^{\tau \mathcal{N}_l(\mathbf{0},\mathbf{1})}$
    $\tilde{\mathbf{x}}_l \leftarrow \tilde{\sigma}_l \mathcal{N}_l(\mathbf{0}, \mathbf{1})$
    $\tilde{\mathbf{y}}_l \leftarrow \mathbf{y}^{(g)} + \tilde{\mathbf{x}}_l$
    $\tilde{f}_l \leftarrow f(\tilde{\mathbf{y}}_l)$
  end for
  $(\tilde{f}_{1;\lambda}, ..., \tilde{f}_{m;\lambda}, ..., \tilde{f}_{\mu;\lambda}) \leftarrow \mathrm{sort}(\tilde{f}_1, ..., \tilde{f}_\lambda)$
  $\mathbf{y}^{(g+1)} \leftarrow \frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{\mathbf{y}}_{m;\lambda}$
  $\sigma^{(g+1)} \leftarrow \frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{\sigma}_{m;\lambda}$
  $g \leftarrow g + 1$
until termination criterion

with oscillation amplitude A and frequency α. The function is minimized and the global minimizer is located at $\hat{\mathbf{y}} = \mathbf{0}$. All simulations in this paper will be conducted at default $\alpha = 2\pi$ (unless stated otherwise). Depending on A, the function shows a finite number M of local attractors for each of the N dimensions, such that the function contains $M^N - 1$ local minima scaling exponentially with N. Note that for $|y_i| > \alpha A/2$ the derivative $\partial f/\partial y_i \neq 0$ (for any i), such that no further local minima occur.

As an introductory example, real optimization runs are shown in Fig. 1 using Algorithm 1. The global convergence is investigated by evaluating the dynamics of the parental distance to the global optimizer denoted by $R^{(g)} = \left\|\mathbf{y}^{(g)}\right\|$. All runs are initialized far away from the local attractor landscape. The black line shows the median of all successful runs (the mean can also be used, but the median is more robust w.r.t. outliers). The global convergence probability is denoted as P_S. After the initial phase, the σSA-ES maintains a nearly constant normalized mutation strength $\sigma^* \approx 30$ with a characteristic dip at $g \approx 200$. σ^* is defined as

$$\sigma^* = \frac{\sigma N}{R}. \tag{2}$$

A constant σ^*-level ensures scale-invariance on the sphere function and therefore linear convergence. The observations from Fig. 1 will be investigated in more detail throughout the paper.

The Rastrigin fitness (1) is defined as a function of $\mathbf{y} = [y_1, ..., y_N]$. Convergence however is usually measured as a function of the residual distance R, see Fig. 1. Quantity R is therefore an aggregated measure over all components. In general, the convergence properties can be investigated using progress rate equations (see Sec. 4). A component-wise progress rate however, derived and discussed in [11], has the disadvantage of not being an aggregated measure. As an example, positive progress (convergence) between two generations occurs for decreasing $R^{(g+1)} < R^{(g)}$ even if some components deteriorate and show negative progress. Furthermore, analytic treatment of N equations is unfeasible for large dimensionalities. Hence, the idea will be to express $\mathbf{y}$-dependent functions as average values over all positions satisfying $\|\mathbf{y}\| = R$ to obtain aggregated measures. First, the approach is presented on the Rastrigin function and later transferred to its corresponding progress rate in Sec. 4.

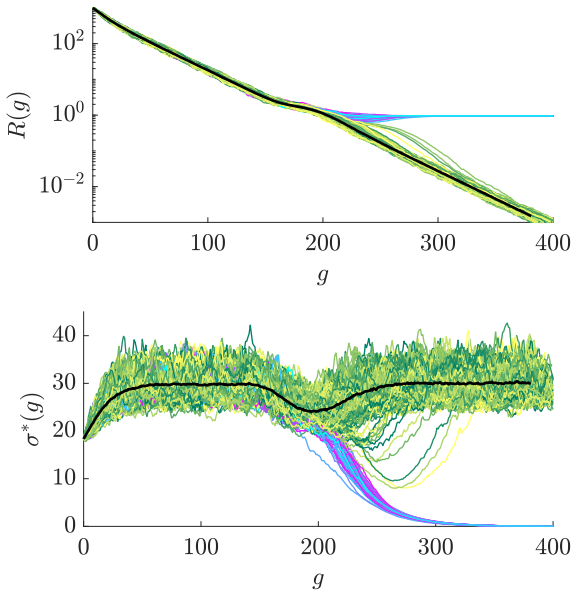

Figure 1: Dynamic runs (500 trials) of Algorithm 1 using $(100/100_I, 200)$-σSA-ES on the Rastrigin function with $N = 100$ and $A = 1$ at learning parameter $\tau = 1/\sqrt{2N}$. The upper plot shows the R-dynamics, while the lower plot shows the normalized mutation strength (2). The green-yellow color palette depicts globally converging runs, while the cyan-magenta colors show locally converging runs. The black line marks the median of all successful runs. The measured success probability $P_S = 0.91$.

The averaging problem to be solved is

$$\overline{f}(R) \coloneqq \underset{\|\mathbf{y}\|=R}{\mathrm{average}}[f(\mathbf{y})]. \tag{3}$$

An approach to obtain $\overline{f}(R)$ is to integrate the function over the $(N-1)$-dimensional sphere surface $S_N(R)$ with radius $\|\mathbf{y}\| = R$ and normalize by the sphere surface according to

$$\overline{f}(R) = \frac{1}{S_N} \int_{\|\mathbf{y}\|=R} f(\mathbf{y})\, \mathrm{d}s, \tag{4}$$

where $\mathrm{d}s$ denotes the (hyper-)surface element. The sphere surface area for $N \geq 2$ evaluated using the gamma function Γ is given by

$$S_N(R) = \frac{2\pi^{N/2} R^{N-1}}{\Gamma(N/2)}. \tag{5}$$

Applying (4) to (1), the first two terms can be evaluated easily noting that $R^2 = \sum_i y_i^2$. Integrating over a constant yields S_N. Therefore, one gets the intermediate result

$$\overline{f}(R) = R^2 + NA + T(R), \tag{6}$$

with

$$T(R) \coloneqq -\frac{A}{S_N} \int_{\|\mathbf{y}\|=R} \sum_{i=1}^{N} \cos(\alpha y_i)\, \mathrm{d}s\,. \tag{7}$$

Closed-form solutions of (7) can be obtained for $N = 1$ and $N = 2$. Starting with $N = 1$ only two discrete points are relevant (no integration necessary) with two possible solutions $y_1 = \pm R$. Averaging over two points therefore yields

$$T(R) = -\frac{A}{2} \sum_{y_1=\pm R} \cos(\alpha y_1) = -A\cos(\alpha R). \tag{8}$$

For $N = 2$ one can use polar coordinates $(y_1, y_2) = (R\cos\phi, R\sin\phi)$ with derivative vector $\frac{\mathrm{d}(y_1,y_2)}{\mathrm{d}\phi} = (-R\sin\phi, R\cos\phi)$ on $\phi \in [0, 2\pi)$. Additionally, one has $S_2 = 2\pi R$. Therefore, inserting this parametrization into (7) and using path element length $\left\|\frac{\mathrm{d}(y_1,y_2)}{\mathrm{d}\phi}\right\| = R$ one has

$$\begin{aligned} T(R) &= -\frac{A}{2\pi R} \int_0^{2\pi} \sum_{i=1}^{2} \cos\left[\alpha y_i(R,\phi)\right] \left\|\frac{\mathrm{d}(y_1,y_2)}{\mathrm{d}\phi}\right\| \mathrm{d}\phi \\ &= -\frac{A}{2\pi} \int_0^{2\pi} \left[\cos(\alpha R\cos\phi) + \cos(\alpha R\sin\phi)\right] \mathrm{d}\phi\,. \end{aligned} \tag{9}$$

The integrals obtained in (9) can be solved in terms of Bessel functions of the first kind $J_n(x)$ with $n \geq 0$ by applying the integral identity [1, p. 360, 9.1.18]

$$J_0(x) = \frac{1}{\pi}\int_0^{\pi} \cos(x\sin t)\,\mathrm{d}t = \frac{1}{\pi}\int_0^{\pi} \cos(x\cos t)\,\mathrm{d}t\,. \tag{10}$$

Due to the periodicity integrating $\cos t$ and $\sin t$ over $[0, \pi]$ yields the same contribution as the integration over $[\pi, 2\pi]$. Thus, one can extend the integral bounds of (10) as

$$2J_0(x) = \frac{1}{\pi}\int_0^{2\pi} \cos(x\sin t)\,\mathrm{d}t = \frac{1}{\pi}\int_0^{2\pi} \cos(x\cos t)\,\mathrm{d}t\,. \tag{11}$$

Comparing (9) with (11) and setting $x = \alpha R$, the expression (9) is evaluated as

$$T(R) = -\frac{A}{2\pi}\left[2\pi J_0(\alpha R) + 2\pi J_0(\alpha R)\right] = -2AJ_0(\alpha R). \tag{12}$$

The final result for $f(R)$ is summarized as

$$\overline{f}(R) = R^2 + A(1 - \cos(\alpha R)) \quad \text{for } N = 1 \tag{13}$$

$$\overline{f}(R) = R^2 + 2A(1 - J_0(\alpha R)) \quad \text{for } N = 2. \tag{14}$$

Examples of result (13) and (14) are shown in Fig. 2. The analytic equations are compared to sampled results, where for each R random isotropic positions are chosen with $\|\mathbf{y}\| = R$ and averaged over 10^4 trials. Excellent agreement can be observed. Furthermore, one notices a decrease of the oscillation effect when N is increased. This will be useful for the subsequent approach. Unfortunately, integral (4) yields analytically exact results only for the cases $N < 3$. In the context of progress rate theory of Sec. 4 an approach is needed which can be applied to any arbitrary large dimensionality N.

The approach presented now will evaluate the average value of $f(\mathbf{y})$ assuming independent, normally distributed random coordinates y_i with zero mean and variance σ_y^2 according to

$$y_i \sim \sigma_y \mathcal{N}(0,1). \tag{15}$$

The approach can be considered as averaging by stochastic sampling (assuming large N) instead of analytic integration. For the

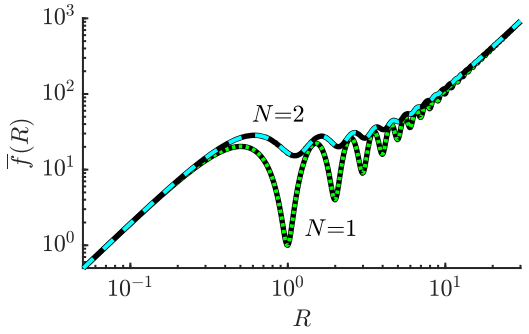

Figure 2: Average Rastrigin function evaluated for $A = 10$ as a function of R. The solid black lines show sampled results using Eq. (1) for $N = 1$ and $N = 2$, respectively. The overlaid dotted green line shows (13) and the dashed cyan line (14).

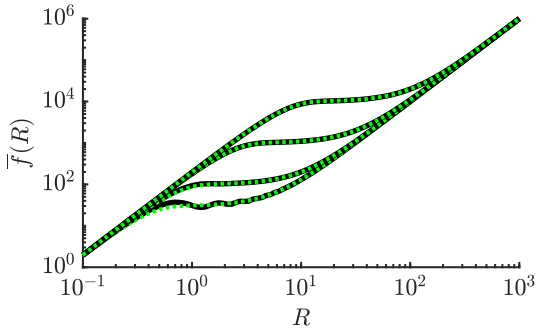

Figure 3: Average Rastrigin function evaluated for $A = 10$ as a function of R. The solid black lines show sampled results of Eq. (1). The overlaid dotted green lines show Eq. (22) for $N = \{3, 10, 100, 1000\}$, from bottom to top.

determination of σ_y from (15) one demands

$$R^2 \stackrel{!}{=} \mathrm{E}\left[\sum_{i=1}^{N} y_i^2\right] = \sigma_y^2\,\mathrm{E}\left[\sum_{i=1}^{N} \mathcal{N}_i^2(0,1)\right] = \sigma_y^2\,\mathrm{E}\left[\chi_N^2\right] = \sigma_y^2 N. \tag{16}$$

It was used that the sum over N independent standard normally distributed variables squared is equal to the chi-squared distributed variable χ_N^2 with $\mathrm{E}\left[\chi_N^2\right] = N$. Solving (16) for σ_y, expression (15) can be rewritten as

$$y_i \sim \frac{R}{\sqrt{N}}\mathcal{N}(0,1). \tag{17}$$

Equation (17) will be useful for averaging sums over trigonometric functions of y_i, where analytic integration is unfeasible. Furthermore, successful ES runs on the Rastrigin function operating under default step-size adaptation also show normally distributed y_i as in (17), see also experiments in Fig. 4. This property is used again in Sec. 4. As y_i is treated as a random variate for the cosine terms, Eq. (1) is now rewritten as

$$f(R, Y) \sim R^2 + NA - AY, \tag{18}$$

with Y containing the sum over the random terms

$$Y \coloneqq \sum_{i=1}^{N} \cos(\alpha y_i). \tag{19}$$

By the Central Limit Theorem in the limit $N \to \infty$, the sum over i.i.d. variates approaches a normal distribution with $Y \sim \mathrm{E}[Y] + \sqrt{\mathrm{Var}[Y]}\mathcal{N}(0,1)$. Additionally, it is shown in Appendix (A.7) that

$$\frac{\sqrt{\mathrm{Var}[Y]}}{\mathrm{E}[Y]} \overset{N\to\infty}{\longrightarrow} 0, \tag{20}$$

with ratio $\sqrt{\mathrm{Var}[Y]}/\mathrm{E}[Y]$ vanishing as $O(1/\sqrt{N})$. In the asymptotic limit the fluctuation term of Y is negligible, which means that the random variate can be replaced by its expected value $\mathrm{E}[Y] = N\mathrm{e}^{-\frac{1}{2}\frac{(\alpha R)^2}{N}}$ evaluated in Appendix (A.1). Equation (18) therefore yields (overline denoting the average in the limit $N \to \infty$)

$$\overline{f}(R) = R^2 + NA - A\,\mathrm{E}[Y] \tag{21}$$

$$= R^2 + NA\left(1 - \mathrm{e}^{-\frac{1}{2}\frac{(\alpha R)^2}{N}}\right). \tag{22}$$

Exemplary evaluations of (22) are shown in Fig 3. The derived results match the sampled results well. Smaller deviations are observed for small values $N = 3$ or $N = 10$, which was expected. In the limit $N \to \infty$ the deviations are smoothed out. The limits $R \to 0$ and $R \to \infty$ yield R^2-dependent functions, i.e., sphere functions. For $R \to \infty$ the exponential vanishes and NA is negligible, while for $R \to 0$ one has $\mathrm{e}^{-\frac{1}{2}\frac{(\alpha R)^2}{N}} = 1 - \frac{1}{2}\frac{(\alpha R)^2}{N} + O(R^4)$ giving $f(R) = (1 + \frac{\alpha^2 A}{2})R^2$.

4 PROGRESS RATE

4.1 Derivation

The method introduced in Sec. 3 will now be applied to results for the progress rate on the Rastrigin function. The progress rate (denoted by φ) measures the expected positional change in search space between two generations $g \to g+1$ as a function of fitness and ES parameters. A positive value corresponds to the ES approaching the optimizer and vice versa. The second order component-wise progress rate φ_i^{II} for the parental location $\mathbf{y}^{(g)}$ is defined as [8]

$$\varphi_i^{\mathrm{II}} \coloneqq \mathrm{E}\left[\left(y_i^{(g)}\right)^2 - \left(y_i^{(g+1)}\right)^2 \,\middle|\, \mathbf{y}^{(g)}, \sigma^{(g)}\right]. \tag{23}$$

The second order refers to the square of y_i-values which ensures $\varphi_i^{\mathrm{II}} > 0$ for $(y_i^{(g+1)})^2 < (y_i^{(g)})^2$ independent of the sign of y_i. A second order model is needed for a correct model of convergence involving large mutation strengths. The expected values for the determination of (23) were already evaluated in [11] in the asymptotic limit $N, \mu, \lambda \to \infty$ ($\vartheta = \mu/\lambda = \mathrm{const}$). The result yields

$$\varphi_i^{\mathrm{II}} = c_\vartheta \frac{\sigma^2}{D_Q}\left(4y_i^2 + \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2} 2\alpha A y_i \sin(\alpha y_i)\right) - \frac{\sigma^2}{\mu}. \tag{24}$$

In (24) the asymptotic progress coefficient [11] is given by

$$c_\vartheta = \frac{\mathrm{e}^{-\frac{1}{2}\left[\Phi^{-1}(\vartheta)\right]^2}}{\sqrt{2\pi}\vartheta}, \tag{25}$$

with $\Phi^{-1}(\cdot)$ denoting the quantile function of the standard normal variate. The c_ϑ is related to the progress coefficient $c_{\mu/\mu,\lambda} \simeq c_\vartheta$ (for $\mu, \lambda \to \infty$ with constant $\vartheta = \mu/\lambda$), see also [6, p. 249]. The quality gain variance D_Q^2 at location $\mathbf{y}$ given σ was evaluated in [12] giving

$$\begin{aligned} D_Q^2(\mathbf{y}) = \sum_{i=1}^{N} \Big\{ & 4\sigma^2 y_i^2 + 2\sigma^4 \\ & + \frac{A^2}{2}\left[1 - \mathrm{e}^{-(\alpha\sigma)^2}\right]\left[1 - \cos(2\alpha y_i)\mathrm{e}^{-(\alpha\sigma)^2}\right] \\ & + 2A\alpha\sigma^2 \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}\left[\alpha\sigma^2\cos(\alpha y_i) + 2y_i \sin(\alpha y_i)\right]\Big\}. \end{aligned} \tag{26}$$

The first term of (24) is usually referred to as the gain term, while the second term is the loss term characteristic for intermediate recombination. A distinct property of the Rastrigin function is that the gain term (y_i-dependent) is not necessarily positive as it is the case for unimodal functions. This property will be discussed later.

The first step to obtain an R-dependent aggregation of expression (23) is to sum over all N components

$$\begin{aligned} \sum_{i=1}^{N} \varphi_i^{\mathrm{II}} &= \mathrm{E}\left[\sum_{i=1}^{N}\left(y_i^{(g)}\right)^2 - \sum_{i=1}^{N}\left(y_i^{(g+1)}\right)^2\right] \\ &= \mathrm{E}\left[\left(R^{(g)}\right)^2 - \left(R^{(g+1)}\right)^2\right], \end{aligned} \tag{27}$$

such that one can define the R-dependent progress rate

$$\varphi_R^{\mathrm{II}} \coloneqq \mathrm{E}\left[\left(R^{(g)}\right)^2 - \left(R^{(g+1)}\right)^2 \,\middle|\, R^{(g)}, \sigma^{(g)}\right]. \tag{28}$$

Given the sphere function $f_{\mathrm{sph}}(R) = R^2$, one can relate (28) to the sphere quality gain $\mathrm{E}\left[f_{\mathrm{sph}}(R^{(g+1)}) - f_{\mathrm{sph}}(R^{(g)})\right] = -\varphi_R^{\mathrm{II}}$, such that the quality gain normalization [5, p. 173] is applicable. This yields the normalized R-dependent progress rate (labeled by the asterisk "*")

$$\varphi_R^{\mathrm{II},*} \coloneqq \frac{N}{2R^2}\varphi_R^{\mathrm{II}}. \tag{29}$$

For $N \to \infty$ one has $\varphi_R^{\mathrm{II},*} \simeq \varphi_{\mathrm{sph}}^*$, see [2, p. 16], yielding the normalized sphere progress rate. Expression (29) has two important properties. First, it is an aggregated progress rate measure over all N components, which is new for the Rastrigin function. Second, its relation to the sphere function enables direct comparison of progress rates.

A prerequisite for the further derivation will be the assumption of normally distributed $y_i \sim \mathcal{N}(0, R^2/N)$, see (17). This property is experimentally confirmed in Fig. 4, using the data of 500 trials shown in Fig. 1, displayed at two residual distances. Good agreement is observed between the expected density (red curve) and the histogram. Each component contributes roughly as $y_i^2 \approx R^2/N$ to the overall residual distance R^2. This concept of "equal contribution" is not new and was investigated in [7] for the quality gain on the ellipsoid. Slightly larger deviations occur at $R \approx 1$ (right), where local attraction is more significant, see also later discussion of Fig. 12. At small mutation strengths where local attraction occurs the assumption of course breaks down.

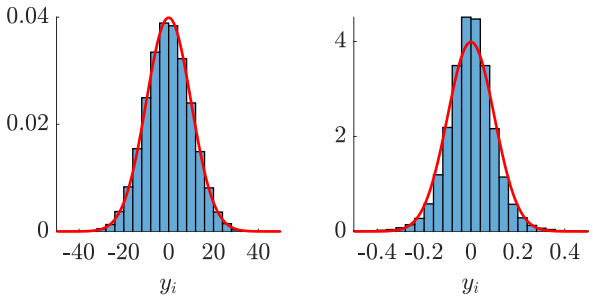

Figure 4: Distribution of realized y_i-values ($i = 1, ..., N$ over 500 trials) for $(100/100_I, 200)$-ES, $N = 100$, $A = 1$, and $\tau = 1/\sqrt{2N}$ at $R \approx 100$ (left) and $R \approx 1$ (right). The red solid curve shows the density of the normal distribution with $y_i \sim \mathcal{N}(0, R^2/N)$.

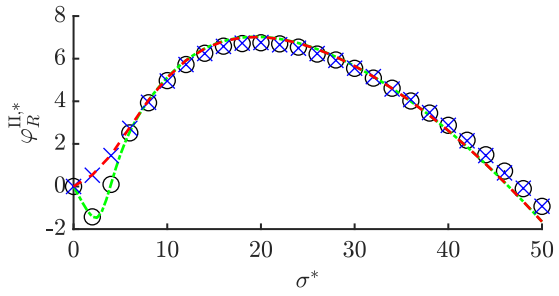

Figure 5: One generation experiments with 10^4 repetitions for $(100/100_I, 200)$-ES, $N = 100$, $A = 1$ at constant $R = 7$. Black circles show experimentally evaluated (23), summed over i, for constant $\mathbf{y} = [0.7, ..., 0.7]$. Blue crosses show (28), where $\mathbf{y}$ is randomly sampled for each trial, such that $\|\mathbf{y}\| = R$. The green dash-dotted line shows (24) (summed over i) and the red dashed line (35). All values are normalized using (29). The error bars are vanishing and not shown.

Now φ_R^{II} is derived starting from (24). Performing the summation one gets

$$\varphi_R^{\mathrm{II}}(R, \mathbf{y}) = c_\vartheta \frac{\sigma^2}{D_Q} \left(4R^2 + \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2} 2\alpha A \sum_{i=1}^{N} y_i \sin(\alpha y_i) \right) - N\frac{\sigma^2}{\mu}. \tag{30}$$

Similarly, the summation of the variance terms in (26) yields

$$\begin{aligned} D_Q^2(R, \mathbf{y}) &= 4\sigma^2 R^2 + 2N\sigma^4 \\ &+ \frac{A^2}{2} \left[1 - \mathrm{e}^{-(\alpha\sigma)^2} \right] \sum_{i=1}^{N} \left[1 - \cos(2\alpha y_i) \mathrm{e}^{-(\alpha\sigma)^2} \right] \\ &+ 2A\alpha\sigma^2 \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2} \left[\alpha\sigma^2 \sum_{i=1}^{N} \cos(\alpha y_i) + 2 \sum_{i=1}^{N} y_i \sin(\alpha y_i) \right]. \end{aligned} \tag{31}$$

Analogous to (19) and (20), the sums over the y_i-dependent trigonometric terms of (30) and (31) will be replaced by their respective expectation values assuming $y_i \sim \frac{R}{\sqrt{N}}\mathcal{N}(0, 1)$ and neglecting fluctuations for $N \to \infty$. The needed expected values are derived in Appendix (A.1), (A.2), and (A.3) giving

$$\mathrm{E}\left[\sum_{i=1}^{N} \cos(\alpha y_i) \right] = N\mathrm{e}^{-\frac{1}{2}\frac{(\alpha R)^2}{N}} \tag{32}$$

$$\mathrm{E}\left[\sum_{i=1}^{N} \cos(2\alpha y_i) \right] = N\mathrm{e}^{-2\frac{(\alpha R)^2}{N}} \tag{33}$$

$$\mathrm{E}\left[\sum_{i=1}^{N} y_i \sin(\alpha y_i) \right] = \alpha R^2 \mathrm{e}^{-\frac{1}{2}\frac{(\alpha R)^2}{N}}. \tag{34}$$

Furthermore, it is shown in Appendix A that $\sqrt{\mathrm{Var}\left[\sum_i(\cdot)\right]}/\mathrm{E}\left[\sum_i(\cdot)\right] \to 0$ for $N \to \infty$ for all three sums. Finally, a fully R-dependent expression can be given for the progress rate

$$\varphi_R^{\mathrm{II}} = c_\vartheta \frac{2R^2\sigma^2}{D_Q(R)} \left(2 + \alpha^2 A \mathrm{e}^{-\frac{\alpha^2}{2}\left(\sigma^2 + \frac{R^2}{N}\right)} \right) - N\frac{\sigma^2}{\mu}. \tag{35}$$

Analogously, the R-dependent quality gain variance yields

$$\begin{aligned} D_Q^2(R) &= 4R^2\sigma^2 + 2N\sigma^4 + \\ &+ \frac{NA^2}{2} \left[1 - \mathrm{e}^{-(\alpha\sigma)^2} \right] \left[1 - \mathrm{e}^{-\alpha^2\left(\sigma^2 + 2\frac{R^2}{N}\right)} \right] \\ &+ 2NA\alpha^2\sigma^2 \mathrm{e}^{-\frac{\alpha^2}{2}\left(\sigma^2 + \frac{R^2}{N}\right)} \left[\sigma^2 + 2\frac{R^2}{N} \right]. \end{aligned} \tag{36}$$

Result (35) is important as it measures the progress on the Rastrigin function in R-space aggregating the individual progress rates φ_i^{II}. Note that the first term of (35), i.e., the gain term, is now strictly positive, which is in contrast to Eq. (24).

One-generation experiments are conducted in Fig. 5 by performing single optimization steps for given mutation strength and averaging the results of progress rates (23) and (28), respectively, over 10^4 trials. Furthermore, the simulations are compared to analytic expressions (24) and (35). To this end, two configurations (constant $R = 7$) are overlaid with one having $\mathbf{y}$ fixed, and one randomly sampled $\mathbf{y}$-values with $\|\mathbf{y}\| = R$. The values are normalized using (29) and displayed using scale-invariant mutations σ^* of Eq. (2). All results are similar for moderate and large σ^*-values showing good agreement. Differences emerge at small σ^*. The fixed $\mathbf{y} = [0.7, ..., 0.7]$ was chosen to lie within a local attractor. In this case $\sum_i \varphi_i^{\mathrm{II}}(\mathbf{y})$ correctly predicts negative progress for small σ^*, while φ_R^{II} falsely assumes normally distributed y_i-coordinates and predicts positive progress. This error vanishes for large σ^*, i.e., when the ES is searching at larger scales. Therefore, one can conclude that φ_R^{II} is a suitable aggregated measure of component-wise φ_i^{II}, if sufficiently large mutations are applied. Indeed, real (successful) ES-runs, such as in Fig. 1 or later in Fig. 6, tend to maintain high σ^*-levels, such that the normal assumption for y_i stays valid. Local attraction (assuming small σ^*) is investigated further in Sec. 6.

A few important remarks regarding results (35) and (36) are made now. Given expression (36), the variance can be written more

compactly as

$$D_Q^2 = D_{\text{sph}}^2 + D_{\text{Ras}}^2, \tag{37}$$

where $D_{\text{sph}}^2 = 4R^2\sigma^2 + 2N\sigma^4$ corresponds to the quality gain variance of the sphere function [6] . The term $D_{\text{Ras}}^2 = D_Q^2(R) - D_{\text{sph}}^2$ is Rastrigin-specific. In the limit of vanishing exponential functions ($R \to \infty$), see later in Sec. 5, the term will simplify significantly giving the so-called Rastrigin (maximum) noise strength

$$D_{\text{Ras}}^2 \simeq \frac{NA^2}{2} =: \sigma_{\text{Ras}}^2. \tag{38}$$

Having derived (35) and (36), the sphere progress rate φ_{sph}^* can be recovered as a special case. It can be obtained from φ_R^{II} in multiple ways. The technical details are not shown since the calculations are simple and straightforward, only the main steps are explained now. As the normalized progress is constant on the sphere for constant scale-invariant mutations σ^*, Eqs. (35) and (36) need to be rewritten as $\varphi_R^{\text{II}}(\sigma^*)$ and $D_Q(\sigma^*)$ by setting $\sigma = \sigma^* R/N$ via (2). Furthermore, normalization (29) needs to be applied. One way to recover φ_{sph}^* is by setting $A = 0$ or $\alpha = 0$, which removes all Rastrigin-specific terms. Another way is applying the limit $R \to \infty$, which suppresses the exponential terms. Additionally, the constant term $NA^2/2$ is negligible in (36) for $R \to \infty$, see also Appendix (B.5). The third way is the limit $R \to 0$. All exponentials contain arguments being a function $g(R^2)$ after inserting $\sigma = \sigma^* R/N$. Performing a Taylor expansion yields $\mathrm{e}^{-g(R^2)} = 1 - g(R^2) + O\left(R^4\right)$ with negligible higher order terms. After simplification, all three approaches yield

$$\varphi_R^{\text{II},*} = \varphi_{\text{sph}}^* = \frac{c_\vartheta \sigma^*}{\sqrt{1 + \sigma^{*2}/2N}} - \frac{\sigma^{*2}}{2\mu}, \tag{39}$$

and for $N \to \infty$ the well-known asymptotic formula

$$\varphi_{\text{sph}}^* = c_\vartheta \sigma^* - \frac{\sigma^{*2}}{2\mu}. \tag{40}$$

Both (39) and (40) are scale-invariant (R-independent) expressions. As a conclusion, the Rastrigin progress rate yields the sphere progress rate in the limits $R \to \infty$ and $R \to 0$. This result is important and was expected from (1), as y_i^2 is dominating at large scales. For $y_i \to 0$ the global attractor is essentially a quadratic function.

An important property of φ_{sph}^* is that for sufficiently small σ^* one has $\varphi_{\text{sph}}^* > 0$, while for too large σ^*-values the progress rate becomes negative. The second (non-trivial) zero of (39), denoted by $\sigma_{\varphi_0}^*$, is derived in Appendix B by setting $\varphi_{\text{sph}}^* = 0$ and yields in (B.8)

$$\sigma_{\varphi_0}^* = \left[\left(N^2 + 8Nc_\vartheta^2\mu^2\right)^{1/2} - N\right]^{1/2}. \tag{41}$$

Due to the same global (quadratic) structure, result (41) will also be applicable to the Rastrigin function as an upper bound for σ^*.

4.2 Progress Landscape

A more detailed analysis of the progress rate (35) is provided now. Given fitness parameters A, α, and N, the expression $\varphi_R^{\text{II},*}(\sigma^*, R)$ is essentially a function of only two variables. Therefore, the results will be displayed in a two-dimensional σ^*-R-space denoted as the progress landscape. Note that for the sphere function, see Eqs. (39) and (40), the progress rate is constant for all R (given σ^* and N).

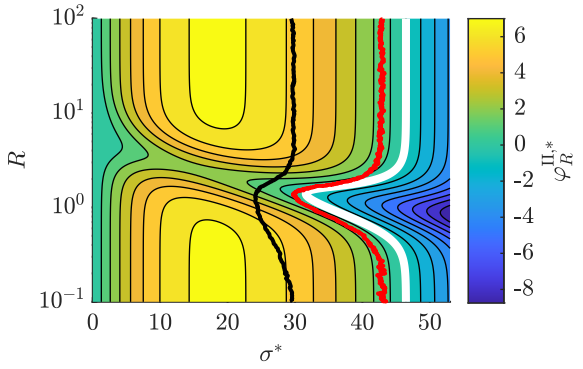

Figure 6: Progress rate $\varphi_R^{\text{II},*}$ for $(100/100_I, 200)$-ES with $N = 100$ and $A = 1$. High progress rate values are shown in yellow and blue values indicate small (negative) progress. The boundary $\varphi_R^{\text{II},*} = 0$ is shown in bold white. The black (left) curve is displaying the median dynamics of Fig. 1 ($\tau = 1/\sqrt{2N}$, $P_S = 0.91$), while the red (right) curve is showing the same ES with $\tau = 1/\sqrt{8N}$ and $P_S = 0.99$.

Figure 6 shows an example progress landscape, evaluated for $\sigma^* \in [0, \sigma_{end}^*]$ and $R \in [10^{-1}, 10^2]$. The value σ_{end}^* is chosen slightly larger than $\sigma_{\varphi_0}^*$, see Eq. (41), as for $\sigma^* > \sigma_{\varphi_0}^*$ the progress rate gets negative. The R-range was chosen large enough to provide good visibility of the relevant characteristics. Thin black lines display regions of equal progress rate level. For $R \to \infty$ and $R \to 0$ the sphere limit is recovered (vertical lines of constant progress).

The median of real runs (black and red curves) show a characteristic σ^*-drop (also visible in Fig. 1), which is directly related to the progress rate zero. The ESs are moving around the progress dip in σ^*-R-space. Interestingly, the σSA-ES with $\tau = 1/\sqrt{2N}$ has a global convergence probability $P_S = 0.91$, while $\tau = 1/\sqrt{8N}$ yields $P_S = 0.99$. Maximizing $\sigma^* \lessapprox \sigma_{\varphi_0}^*$ therefore maximizes P_S, which is associated with a smaller learning parameter τ. This effect can also be observed for the CSA-ES (cumulative step-size adaptation), where a higher P_S is observed for smaller cumulation constant values (due to a slower change of σ). The downside of large mutation strengths is less efficiency optimizing the sphere limits (the sphere-optimal value for the $(100/100_I, 200)$-ES, see Fig. 6, is at $\sigma^* \approx 19$). The respective median R-dynamics reaches the stopping value $R = 10^{-3}$ at $g \approx 400$ ($\tau = 1/\sqrt{2N}$), while $g \approx 1100$ for $\tau = 1/\sqrt{8N}$.

One observes that $\varphi_R^{\text{II},*} > 0$ for sufficiently small σ^*. This means that positive progress is expected at any R for arbitrary small $\sigma^* > 0$, which contradicts experimental observations, see also Fig. 5, as small σ^* significantly increases the local convergence probability. Hence, local attraction is not modeled correctly by $\varphi_R^{\text{II},*}$. Furthermore, the progress dip of Fig. 6 is not related to single local attractors. It is a cumulative effect of oscillations in all N dimensions related to Rastrigin noise term (38). This is investigated in the next section.

5 CONVERGENCE AND POPULATION SIZING

In this section the convergence properties on the Rastrigin function are discussed. Global convergence (in expectation) requires $\varphi_R^{\text{II},*} > 0$ for $R \in (0, \infty)$. The boundary $\varphi_R^{\text{II},*} = 0$, see Fig. 6, is therefore of most interest, especially the progress dip and its location in σ^*-R-space. As it is shown in Appendix B, a closed-form solution can only be obtained under certain (simplified) assumptions. An analytical solution for $R(\sigma^*)$, such that $\varphi_R^{\text{II},*} = 0$, cannot be given due to the non-linearity of the underlying equations.

In the limit of $R \to \infty$, all exponentials of (35) and (36) vanish. The resulting equation for $\varphi_R^{\text{II},*}(\sigma^*, R) = 0$ simplifies significantly with $D_Q^2 = D_{\text{sph}}^2 + NA^2/2$, see (37), such that a fourth order polynomial is obtained in Eq. (B.9) as

$$\sigma^{*4} + 2N\sigma^{*2} + \frac{N^4A^2}{4R^4} - 8Nc_\vartheta^2\mu^2 = 0. \tag{42}$$

Solving (42) for R yields the zero-progress line

$$R_{\varphi_0}(\sigma^*) = \left(\frac{1}{4}\frac{N^4A^2}{8Nc_\vartheta^2\mu^2 - 2N\sigma^{*2} - \sigma^{*4}}\right)^{1/4}, \tag{43}$$

which is visualized in Fig. 7 as a black dashed line. An important relation to the noisy sphere model can be made. In [3] the residual location error R_∞ was derived for the $(\mu/\mu_I, \lambda)$-ES assuming a constant noise strength σ_ϵ in the limit $\sigma^* \to 0$ as

$$R_\infty \simeq \sqrt{\frac{\sigma_\epsilon N}{4c_{\mu/\mu,\lambda}\mu}}. \tag{44}$$

Applying the limit $\sigma^* \to 0$ to Eq. (43), identifying the constant noise strength of the Rastrigin function (for sufficiently large R) as $\sigma_{\text{Ras}}^2 = NA^2/2$ via (38) yields

$$R_{\varphi_0}\big|_{\sigma^*=0} = \left(\frac{N^3A^2}{32c_\vartheta^2\mu^2}\right)^{1/4} = \sqrt{\frac{\sigma_{\text{Ras}}N}{4c_\vartheta\mu}}, \tag{45}$$

which corresponds to result (44) with $c_{\mu/\mu,\lambda} \simeq c_\vartheta$ and $\sigma_\epsilon = \sigma_{\text{Ras}}$.

Results (43) and (45) explain the σ^*-decrease observed in Fig. 1 and Fig. 6, occurring for the ES approaching the Rastrigin noise floor. The red curve ($\tau = 1/\sqrt{8N}$) decreases σ^* to have positive progress at all, while the black curve ($\tau = 1/\sqrt{2N}$) exhibits smaller σ^*-values keeping a larger distance to the $\varphi_R^{\text{II},*} = 0$ boundary. Thus, the latter realizes a larger local progress. This is the result of the faster adaptation of σ (due to the larger τ). As a result, one has smaller mutations which are in turn more prone to be trapped in a local attractor. This is reflected by a lower success probability P_S. The smaller τ, however, yields a larger P_S-value by keeping a higher σ^*-level.

In the limit of $R \to 0$, σ^* increases again, as the ES reaches the global attractor optimizing a sphere function with constant σ^* (same level as for $R \to \infty$). Since there is no closed-form solution of the progress dip location, a different approach is needed to model the transition point. Recalling that R_{φ_0} of (43) was derived in the limit of vanishing exponential terms, a natural extension of this model is to parametrize the point at which the terms are vanishing. This can also be motivated by looking at Eq. (22) and Fig. 3, where the exponential term models the transition between the sphere limits. Hence, a transition relation $R_{\text{tr}}(\sigma^*)$ is introduced. It can

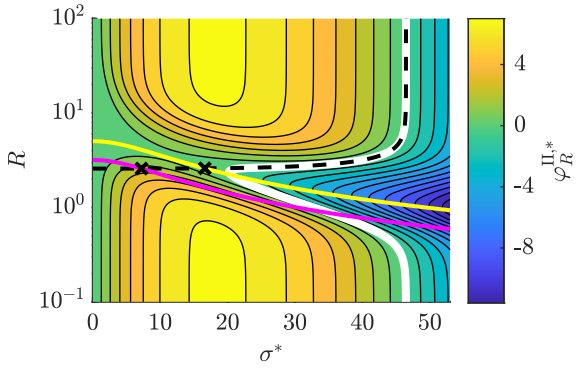

Figure 7: Progress rate $\varphi_R^{\text{II},*}$ for $(100/100_I, 200)$-ES with $N = 100$ and $A = 3$. The black dashed line shows Eq. (43). Two lines show Eq. (46) with $\delta = 5$ (yellow, top) and $\delta = 1$ (magenta, bottom), respectively. Crosses indicate the intersection points obtained by Eq. (49). Note that the progress dip at $A = 3$ is significantly larger compared to $A = 1$ from Fig. 6.

be obtained by investigating the characteristic exponential term of φ_R^{II} in (35), which also occurs in variance (36). Introducing an attenuation factor $\delta > 0$ and setting $\sigma = \sigma^* R_{\text{tr}}/N$, one can demand

$$\begin{aligned} &e^{-\delta} \overset{!}{=} e^{-\frac{(\alpha R_{\text{tr}})^2}{2}\left[\left(\frac{\sigma^*}{N}\right)^2 + \frac{1}{N}\right]}, \quad \text{such that} \\ &R_{\text{tr}}(\sigma^*) = \frac{\sqrt{2\delta N}}{\alpha}\frac{1}{\sqrt{1 + \sigma^{*2}/N}}. \end{aligned} \tag{46}$$

It is assumed that δ is independent of the fitness and strategy parameters. Figure 7 shows R_{φ_0} from (43) and R_{tr} from (46) with two exemplary evaluations $\delta = 1$ and $\delta = 5$. One observes that R_{φ_0} (black dashed line) follows the zero-progress line up until the dip minimum is reached. The dip location along the R-axis is well approximated by the constant noise limit (45) at $\sigma^* = 0$. The R_{tr}-curves (magenta and yellow, respectively) follow a characteristic path depending on the chosen attenuation factor δ. The intersection point σ_{sec}^* of both curves, namely

$$R_{\varphi_0}(\sigma_{\text{sec}}^*) \overset{!}{=} R_{\text{tr}}(\sigma_{\text{sec}}^*), \tag{47}$$

is parametrizing the dip location and will give insight on the population scaling $\mu(N, \alpha, A)$. Setting $R_{\varphi_0} = R_{\text{tr}}$, one obtains a fourth order polynomial in σ^* as

$$\sigma^{*4} + 2N\sigma^{*2} + \frac{N^2\left(\alpha^4A^2 - 128\delta^2c_\vartheta^2\mu^2/N\right)}{\alpha^4A^2 + 16\delta^2} = 0. \tag{48}$$

The real non-negative solution of (48) yields after simplification the intersection point

$$\sigma_{\text{sec}}^* = \left[N\frac{\left(1 + \frac{8c_\vartheta^2\mu^2}{N}\right)^{1/2}}{\left(1 + \frac{\alpha^4A^2}{16\delta^2}\right)^{1/2}} - N\right]^{1/2}, \tag{49}$$

which is visualized in Fig. 7. Convergence on the sphere requires

$$0 < \sigma^*_{\text{sec}} < \sigma^*_{\varphi_0}, \tag{50}$$

with the sphere-zero $\sigma^*_{\varphi_0}$ given in Eq. (41). The relation $\sigma^*_{\text{sec}} < \sigma^*_{\varphi_0}$ follows immediately for any $A, \alpha, \delta > 0$ and setting $A = 0$ or $\alpha = 0$ yields $\sigma^*_{\text{sec}} = \sigma^*_{\varphi_0}$. Demanding $\sigma^*_{\text{sec}} > 0$ in (49) it must hold

$$\frac{8c_\vartheta^2\mu^2}{N} > \frac{\alpha^4 A^2}{16\delta^2}. \tag{51}$$

Solving (51) for μ one arrives at the important population sizing result

$$\mu > \sqrt{\frac{N}{2}}\frac{\alpha^2 A}{8c_\vartheta\delta}. \tag{52}$$

Expression (52) relates the fitness-dependent parameters to the population size μ. For the subsequent experiments we will investigate the scaling properties of (52) without considering the potential prefactors of Eq. (52). To this end, repeated experiments of Algorithm 1 are performed and the success probability P_S is measured. Then, the necessary population size μ is evaluated to achieve a high success rate of $P_S \geq 0.99$. The results of Figs. 8, 9, and 10 show good agreement with the parameter scaling predicted in Eq. (52). The $\mu(N)$-scaling from experimental results is clearly sub-linear (as already observed in [9]) and indicates a scaling slightly larger than $N^{1/2}$. Some fluctuations can be observed which is practically inevitable, as very large N and μ are tested posing limits on the available CPU resources. Furthermore, certain deviations of the experiments to prediction (52) are expected to occur as the underlying model is based on an expected value, see (23), without considering possible higher order moments of the y_i-distribution causing fluctuations.

6 LOCAL ATTRACTION

In this section the limitations of the R-dependent progress rate φ^{II}_R are discussed by investigating local attraction effects. In case of local convergence one has $\sigma \to 0$ (equivalently $\sigma^* \to 0$) while R stagnates. In this case the local structure of the fitness landscape is dominating. Hence, the assumption $y_i \sim \frac{R}{\sqrt{N}}\mathcal{N}(0,1)$ being normally distributed around the optimizer cannot hold. While the progress landscapes show positive progress for small σ^*-values, this does not imply global convergence of real ES runs, see e.g. Fig. 12. It should be intuitively clear that for too small mutation strengths local convergence occurs. This issue was also observed in one-generation experiments in Fig. 5, where negative progress rates are obtained at certain $\mathbf{y}$, if local attraction is present. As the aggregated (R-dependent) formula (35) is not able to model local attraction, a different approach is needed based on the y_i-dependent formula (24). The goal is to derive a σ-condition avoiding local attraction (in expectation). To this end, a characteristic "escape" mutation strength σ_{esc} is derived. It can serve as an additional stability criterion for the ES.

Starting with φ^{II}_i of Eq. (24), the gain function G is defined as

$$G(y_i, \sigma) := 4y_i^2 + \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2} 2\alpha A y_i \sin(\alpha y_i). \tag{53}$$

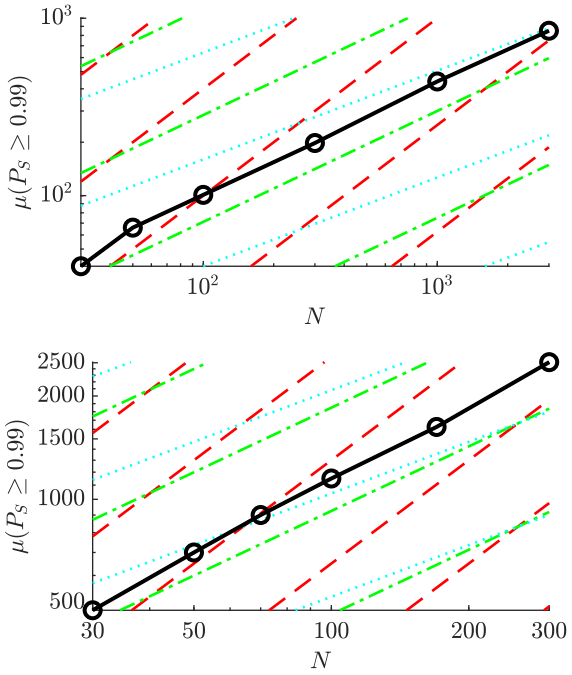

Figure 8: Population sizing $\mu(N)$ using σSA-ES with $\tau = 1/\sqrt{2N}$ and $\alpha = 2\pi$. The top plot shows $\vartheta = 1/4$ with $A = 1$, while the bottom plot shows $\vartheta = 1/2$ with $A = 10$. The dotted cyan lines depict $\mu \propto \sqrt{N}$, dash-dotted green lines $\mu \propto N^{5/8}$, and dashed red lines $\mu \propto N$. The number of evaluated trials, for increasing N, is 2000, 2000, 1000, 700, 500, 400 **(top) and** 3000, 3000, 1500, 1000, 700, 600 **(bottom).**

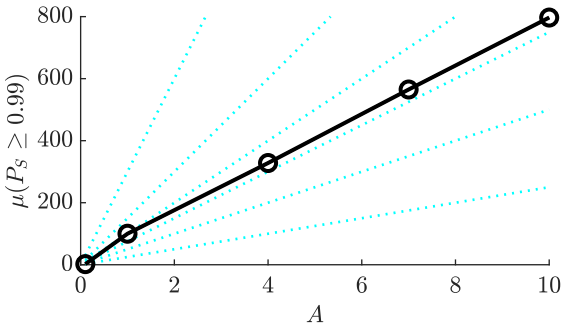

Figure 9: Population sizing $\mu(A)$ using σSA-ES with $\vartheta = 1/4$, $N = 100$, $\tau = 1/\sqrt{2N}$, and $\alpha = 2\pi$. The dotted cyan lines show $\mu \propto A$. For each data point 2000 **trials were evaluated.**

Requiring positive progress $\varphi^{\text{II}}_i > 0$, Eq. (24) yields

$$\frac{c_\vartheta}{D_Q} G(y_i, \sigma) > \frac{1}{\mu}. \tag{54}$$

At this point the infinite population limit $\mu \to \infty$ is assumed in order to obtain closed-form solutions. As $1/\mu \to 0$ it suffices to show that $G > 0$ for $\varphi^{\text{II}}_i > 0$ to hold. The function G is plotted

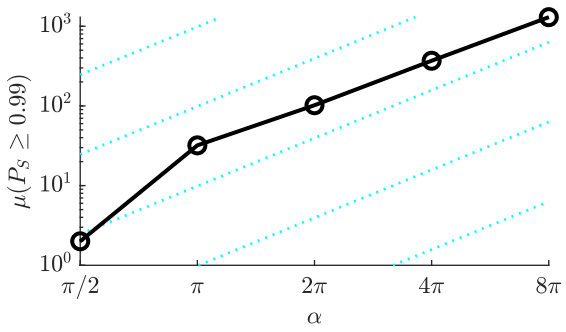

Figure 10: Population sizing $\mu(\alpha)$ using σSA-ES with $\vartheta = 1/4$, $N = 100$, $\tau = 1/\sqrt{2N}$, and $A = 1$. The dotted cyan lines show $\mu \propto \alpha^2$. For each data point 2000 trials were evaluated.

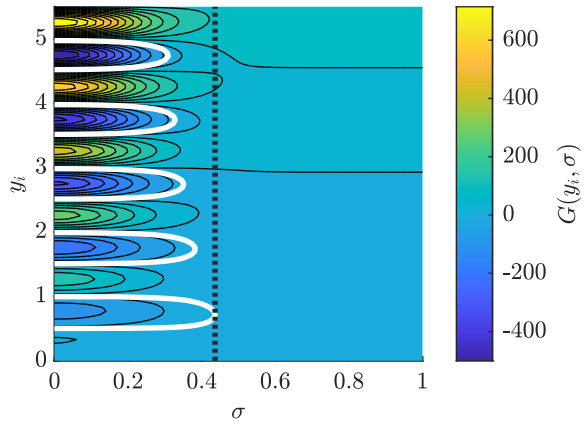

Figure 11: Gain function (53) visualized for $A = 10$ and $\alpha = 2\pi$. The boundary $G = 0$ is shown in bold white, enclosing regions of negative progress. σ_{esc} from (64) is shown as black dotted. Only the first five local attractors are shown (out of 31).

in Fig. 11 as a function of σ and y_i. One observes local attraction regions for small mutations located at $y_0 \approx \{1, 2, 3, 4, 5\}$. For small σ each of the attractors is a "stable" point, as $y_0 + \epsilon$ (with $\epsilon > 0$) yields positive gain (decreasing y_i in expectation), while $y_0 - \epsilon$ yields negative gain (increasing y_i). For sufficiently large $\sigma > \sigma_{\text{esc}}$ (black dotted line) positive progress can be ensured. The threshold σ_{esc} is derived now. Starting with (53), requiring $G \stackrel{!}{=} 0$ and assuming $y_i \neq 0$, G is refactored as

$$G = 2y_i\widetilde{G}, \tag{55}$$

with $\widetilde{G}$ defined as

$$\widetilde{G} := 2y_i + \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}\alpha A \sin(\alpha y_i) \stackrel{!}{=} 0, \tag{56}$$

yielding a first condition. The second condition $\frac{\partial G}{\partial y_i} = 0$ (at $\sigma = \sigma_{\text{esc}}$) can be inferred from Fig. 11, which yields for (55)

$$\frac{\partial G}{\partial y_i} = 2\widetilde{G} + 2y_i\frac{\partial \widetilde{G}}{\partial y_i} \stackrel{!}{=} 0. \tag{57}$$

As $\widetilde{G} = 0$ and $y_i \neq 0$, $\frac{\partial G}{\partial y_i} = 0$ is equivalent to $\frac{\partial \widetilde{G}}{\partial y_i} = 0$. Therefore, one has

$$\frac{\partial \widetilde{G}}{\partial y_i} = 2 + \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}\alpha^2 A\cos(\alpha y_i) = 0, \tag{58}$$

such that the following condition is obtained

$$\mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}\alpha A = -\frac{2}{\alpha\cos(\alpha y_i)}. \tag{59}$$

Inserting condition (59) into (56), it follows

$$2y_i - \frac{2}{\alpha}\frac{\sin(\alpha y_i)}{\cos(\alpha y_i)} = 0. \tag{60}$$

Introducing the substitution $x = \alpha y_i$ and applying $\sin x/\cos x = \tan x$ yields

$$\frac{2}{\alpha}(x - \tan x) = 0. \tag{61}$$

The first non-trivial solution of (61) is the most interesting, as it corresponds to $G = 0$ of the first local attractor at $y_i \approx 0.75$, see Fig. 11. Furthermore, negative gain contributions are due to the sine term in (53). For small $|y_i| < 1$ one has $y_i^2 < |y_i|$, such that the first local attractor corresponds to the worst case requiring the largest σ to obtain $G = 0$. Numerical solving yields the zero of (61) as

$$x_0 \approx 4.493. \tag{62}$$

Multiplying (56) by α, identifying $x_0 = \alpha y_i$ and $\sigma = \sigma_{\text{esc}}$ (point of vanishing gain) results in

$$\begin{aligned} 2x_0 + \mathrm{e}^{-\frac{1}{2}(\alpha\sigma_{\text{esc}})^2}\alpha^2 A\sin x_0 &= 0 \\ \mathrm{e}^{\frac{1}{2}(\alpha\sigma_{\text{esc}})^2} &= -\frac{\alpha^2 A\sin x_0}{2x_0}. \end{aligned} \tag{63}$$

Resolving (63) for σ_{esc} yields the final result

$$\begin{aligned} \sigma_{\text{esc}} &= \frac{1}{\alpha}\sqrt{2\ln\left(-\frac{\alpha^2 A\sin x_0}{2x_0}\right)} \\ &\approx \frac{1}{\alpha}\sqrt{2\ln\left(0.1086\alpha^2 A\right)}. \end{aligned} \tag{64}$$

Figure 12 shows experiments of the $(400/400_I, 800)$-σSA-ES with $\alpha = 2\pi$, and relatively large $A = 10$, such that one has $\sigma_{\text{esc}} \approx 0.436$ from result (64). A constant σ translates to $R(\sigma^*) = \sigma N/\sigma^*$, see normalization (2), showing a $1/\sigma^*$ characteristics (red dashed line) in the progress landscape. The median of real *unsuccessful* runs is shown for different learning parameters τ. The sharp decrease $\sigma^* \to 0$ indicates local convergence, which agrees well with the σ_{esc}-line. However, dropping below the threshold $\sigma < \sigma_{\text{esc}}$ does not imply that local convergence must occur (see success rate $P_S > 0$ for all τ). Conversely, it is a stability criterion that maintains positive component-wise progress in expectation, if $\sigma > \sigma_{\text{esc}}$ is kept large enough. Of course, this can only hold up to the global attractor, at which $\sigma \to 0$ must be ensured to have convergence.

Figure 13 shows numerically evaluated progress dip locations, see e.g. the dip at $\sigma^* \approx 25$ and $R \approx 2$ in Fig. 12, for increasing μ-values while keeping ϑ and the fitness parameters constant. It shows how increasing μ shifts the dip location to larger σ^*-values and smaller residual distances R. Using larger populations enables the ES to operate at larger mutation strengths and approach the optimizer

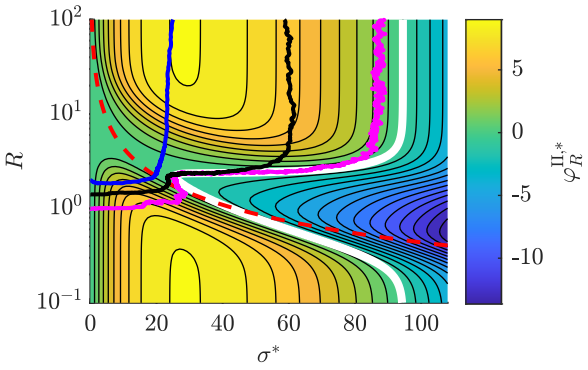

Figure 12: Median dynamics of unsuccessful runs (out of 100 trials) for $(400/400_I, 800)$-σSA-ES, $N = 100$, $A = 10$, with $\sigma_{\text{esc}} \approx 0.436$ (red dashed line). The learning parameter was set to $\tau = 1/\sqrt{N}$ (blue, left, $P_S = 0.01$), $\tau = 1/\sqrt{2N}$ (black, center, $P_S = 0.08$), and $\tau = 1/\sqrt{8N}$ (magenta, right, $P_S = 0.29$).

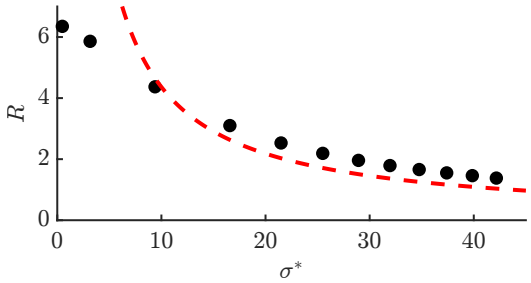

Figure 13: Numerically evaluated progress dip locations (black dots) for $\vartheta = 1/2$, $N = 100$, and $A = 10$ with increasing values $\mu = \{10, 50, 100, 200, ..., 1000\}$ from left to right. The red dashed curve shows σ_{esc} (displayed as $R = \sigma_{\text{esc}} N/\sigma^*$).

more closely resulting in higher global convergence probabilities. The σ_{esc}-line (red dashed) remains constant as it was derived for $\mu \to \infty$. A progress dip located below the σ_{esc}-line is critical, as both noise floor and local attraction effects overlap yielding effectively zero success rates.

The results obtained from Fig. 12 suggest a synthetic explicit σ-control rule for understanding the meaning of σ_{esc}. This rule uses a constant mutation strength σ for a sufficiently high number of generations until the global attractor is reached, and then decreases $\sigma \to 0$. This is realized by defining a $\sigma^{(g)}$-schedule being constant for the first $g < 9000$ generations. For $9000 \leq g \leq 10^4$ it is decreased multiplicatively as $\sigma^{(g+1)} = c\sigma^{(g)}$ $(0 < c < 1)$, such that the stopping criterion $\sigma < 10^{-6}$ is reached at the last generation. The corresponding experiments are conducted in Fig. 14. The single-trial dynamics show that only the run at σ_{esc} converges globally (repeated experiments shown in Fig. 15). ES-runs with $\sigma < \sigma_{\text{esc}}$ tend to converge locally at large R due to the ES getting stuck in the local minima landscape. ES-runs operating at $\sigma > \sigma_{\text{esc}}$ are less prone to local attraction and they reach the Rastrigin noise floor at moderately large R (see intersection of red and white line in Fig. 12).

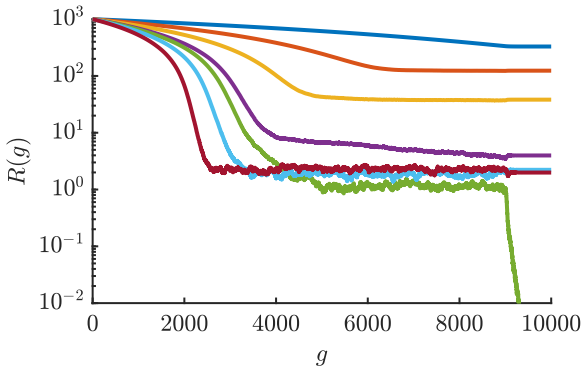

Figure 14: Single runs using constant σ for $(400/400_I, 800)$-ES, $N = 100$, and $A = 10$. A schedule for $\sigma^{(g)}$ was defined being constant during the first 9000 generations and converging exponentially within the last 1000 generations. One has $\sigma = \{0.1, 0.2, 0.3, 0.4, \sigma_{\text{esc}}, 0.5, 0.6\}$, from top to bottom (see ordering at $g = 2000$).

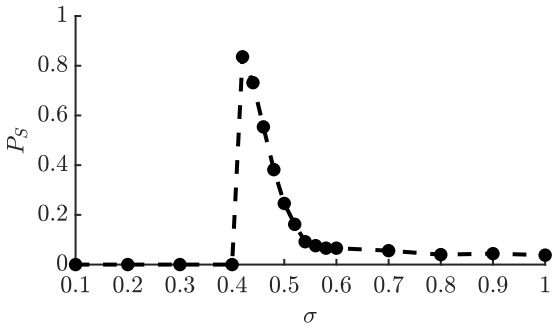

Figure 15: The success probability P_S is evaluated using parameters of Fig. 14 and 500 repetitions for each σ. The peak occurs around $\sigma \approx \sigma_{\text{esc}} = 0.436$.

In Fig. 15 different σ-values are tested and the success rate P_S is evaluated over 500 repetitions. One observes that P_S is maximized at $\sigma \approx \sigma_{\text{esc}}$. As expected, values $\sigma < \sigma_{\text{esc}}$ are less successful due to local attraction. For $\sigma > \sigma_{\text{esc}}$ local attraction is avoided, but the ES fluctuates at a larger residual distance before $\sigma \to 0$, such that it is more likely to miss the global attractor.

7 CONCLUSIONS AND OUTLOOK

In this paper results from progress rate theory were applied and extended to investigate the convergence properties on the Rastrigin function. An aggregated residual distance dependent progress rate was obtained assuming normally distributed y_i-locations around the optimizer. The progress rate yields useful insights on the search behavior of the ES, which can be illustrated by recalling Fig. 1. Far away from local attraction the ES is optimizing the sphere keeping a constant scale-invariant mutation strength. Approaching the local attractor landscape leads to a significant reduction of the (normalized) mutation strength compared to the initial level.

As the mutation strength σ decreases together with σ^*, it may fall below the σ_{esc}-threshold (see Fig. 12). Having $\sigma \gg \sigma_{\text{esc}}$ (at σ^*-levels comparable to sphere-optimal values) the ES is performing a global search. It is not significantly influenced by single local attractors. For $\sigma \lessapprox \sigma_{\text{esc}}$ the search can be regarded as rather local and individual attractors gain importance, such that local convergence occurs with higher probability. Within the global attractor the sphere function is optimized again. Considering the ES performance a two-fold positive effect of large populations on the success rate can be identified. First, large μ-values decrease the expected residual distance (45) to the global optimizer (similar to optimizing the sphere under constant noise). Second, intermediate recombination reduces the magnitude of the loss term $-\sigma^{*2}/(2\mu)$ in (39). Large μ and recombination therefore allow the ES to operate at larger σ-levels keeping $\sigma > \sigma_{\text{esc}}$ and enabling a global search.

Furthermore, the progress rate analysis enabled the derivation of the population scaling result in (52), which could be experimentally verified. The result can serve to some extent as a guidance for the investigation of other highly multimodal test functions, provided that a global (spherical) structure exists with local perturbations.

There are multiple issues requiring further research. While it is now clear why large populations and mutation strengths are beneficial optimizing Rastrigin, a detailed analysis of the full σSA-ES or CSA-ES including the step-size adaptation is still pending. Additionally, the ES-efficiency in terms of fitness evaluations as a function of population size, truncation ratio, and learning parameter was not yet investigated. As the population size is a crucial parameter, the idea of using dynamic population control methods seems natural, see e.g. [4]. Actually, the theoretical analysis of population size control strategies is an uncharted research field. Furthermore, a probabilistic model would be useful to predict the success rate P_S as a function of fitness and ES parameters. Whether the obtained results can be transferred to other multimodal functions also remains part of future research.

ACKNOWLEDGMENTS

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

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] A. Auger and N. Hansen. 2005. A Restart CMA Evolution Strategy with Increasing Population Size. In *Congress on Evolutionary Computation, CEC'05*, Vol. 2. IEEE, 1769–1776.
[5] H.-G. Beyer. 1993. Toward a Theory of Evolution Strategies: Some Asymptotical Results from the $(1,^+\lambda)$-Theory. *Evolutionary Computation* 1, 2 (1993), 165–188.
[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 A. Melkozerov. 2014. The Dynamics of Self-Adaptive Multi-Recombinant Evolution Strategies on the General Ellipsoid Model. *IEEE Transactions on Evolutionary Computation* 18, 5 (2014), 764–778. DOI: 10.1109/TEVC.2013.2283968.
[9] 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.
[10] S. Meyer-Nieberg. 2007. *Self-Adaptation in Evolution Strategies.* Dissertation. Universität Dortmund, Dortmund, Germany.
[11] A. Omeradzic and H.-G. Beyer. 2022. *Progress Analysis of a Multi-Recombinative Evolution Strategy on the Highly Multimodal Rastrigin Function.* Report. Vorarlberg University of Applied Sciences. https://opus.fhv.at/frontdoor/index/index/docId/4722.
[12] 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 – PPSN XVII*, G. Rudolph, A. V. Kononova, H. Aguirre, P. Kerschke, G. Ochoa, and T. Tušar (Eds.). Springer International Publishing, 499–511.

A EXPECTED VALUES

In Secs. 3 and 4 the expected values and variances over the sums of $i = 1, ..., N$ trigonometric terms with random variable $y_i \sim \sigma_y \mathcal{N}(0,1)$, $\sigma_y = \frac{R}{\sqrt{N}}$, are needed assuming i.i.d. components. The expected values and variances are taken over sums of the terms $\cos(\alpha y_i)$, $\cos(2\alpha y_i)$, and $y \sin(\alpha y_i)$, with $i = 1, ..., N$, respectively. They were already derived in [12], such that one has

$$\mathrm{E}\left[\sum_{i=1}^{N} \cos(\alpha y_i)\right] = N \mathrm{e}^{-\frac{1}{2}\frac{(\alpha R)^2}{N}} \tag{A.1}$$

$$\mathrm{E}\left[\sum_{i=1}^{N} \cos(2\alpha y_i)\right] = N \mathrm{e}^{-2\frac{(\alpha R)^2}{N}} \tag{A.2}$$

$$\mathrm{E}\left[\sum_{i=1}^{N} y_i \sin(\alpha y_i)\right] = \alpha R^2 \mathrm{e}^{-\frac{1}{2}\frac{(\alpha R)^2}{N}}. \tag{A.3}$$

Now the corresponding variances are established. Applying $\mathrm{Var}\left[\sum_i (\cdot)\right] = \sum_i \mathrm{E}\left[(\cdot)^2\right] - \mathrm{E}\left[(\cdot)\right]^2$ and using previously obtained results (supplementary material of [12]) with $\sigma_y = R/\sqrt{N}$ yields for the variances

$$\begin{aligned}\mathrm{Var}\left[\sum_{i=1}^{N} \cos(\alpha y_i)\right] &= \sum_{i=1}^{N} \mathrm{E}\left[\cos^2(\alpha y_i)\right] - \mathrm{E}\left[\cos(\alpha y_i)\right]^2 \\ &= N\left(\frac{1}{2} + \frac{1}{2}\mathrm{e}^{-\frac{1}{2}\frac{(2\alpha R)^2}{N}} - \mathrm{e}^{-\frac{(\alpha R)^2}{N}}\right)\end{aligned} \tag{A.4}$$

$$\begin{aligned}\mathrm{Var}\left[\sum_{i=1}^{N} \cos(2\alpha y_i)\right] &= \sum_{i=1}^{N} \mathrm{E}\left[\cos^2(2\alpha y_i)\right] - \mathrm{E}\left[\cos(2\alpha y_i)\right]^2 \\ &= N\left(\frac{1}{2} + \frac{1}{2}\mathrm{e}^{-\frac{1}{2}\frac{(4\alpha R)^2}{N}} - \mathrm{e}^{-\frac{(2\alpha R)^2}{N}}\right)\end{aligned} \tag{A.5}$$

$$\begin{aligned}&\mathrm{Var}\left[\sum_{i=1}^{N} y_i \sin(\alpha y_i)\right] = \sum_{i=1}^{N} \mathrm{E}\left[y_i^2 \sin^2(\alpha y_i)\right] - \mathrm{E}\left[y_i \sin(\alpha y_i)\right]^2 \\ &= N\left(\frac{1}{2}\frac{R^2}{N} - \frac{1}{2}\left[\frac{R^2}{N} - (2\alpha)^2\frac{R^4}{N^2}\right]\mathrm{e}^{-\frac{1}{2}\frac{(2\alpha R)^2}{N}} - \alpha^2\frac{R^4}{N^2}\mathrm{e}^{-\frac{(\alpha R)^2}{N}}\right) \\ &= R^2\left(\frac{1}{2} - \frac{1}{2}\mathrm{e}^{-\frac{1}{2}\frac{(2\alpha R)^2}{N}} + 2\alpha^2\frac{R^2}{N}\mathrm{e}^{-\frac{1}{2}\frac{(2\alpha R)^2}{N}} - \alpha^2\frac{R^2}{N}\mathrm{e}^{-\frac{(\alpha R)^2}{N}}\right).\end{aligned} \tag{A.6}$$

All required expected values for (A.4), (A.5), and (A.6) can be evaluated using the equations in the supplementary material of [12]. Given aforementioned results, ratio $\sqrt{\mathrm{Var}\left[Y\right]}/\mathrm{E}\left[Y\right]$ can be evaluated in the limit $N \to \infty$. As $Y \sim \mathrm{E}\left[Y\right] + \sqrt{\mathrm{Var}\left[Y\right]}\mathcal{N}(0,1)$, this will give a reasoning to neglect the fluctuation term $\sqrt{\mathrm{Var}\left[Y\right]}$ relative to

$\mathrm{E}[Y]$ in the limit of infinite dimensionality. The ratios are evaluated as

$$\frac{\mathrm{Var}\left[\sum_{i=1}^{N}\cos(\alpha y_i)\right]^{\frac{1}{2}}}{\mathrm{E}\left[\sum_{i=1}^{N}\cos(\alpha y_i)\right]} = \frac{\left(\frac{1}{2}+\frac{1}{2}\mathrm{e}^{-\frac{1}{2}\frac{(2\alpha R)^2}{N}}-\mathrm{e}^{-\frac{(\alpha R)^2}{N}}\right)^{\frac{1}{2}}}{\sqrt{N}\mathrm{e}^{-\frac{1}{2}\frac{(\alpha R)^2}{N}}} \overset{N\to\infty}{\longrightarrow} 0 \tag{A.7}$$

$$\frac{\mathrm{Var}\left[\sum_{i=1}^{N}\cos(2\alpha y_i)\right]^{\frac{1}{2}}}{\mathrm{E}\left[\sum_{i=1}^{N}\cos(2\alpha y_i)\right]} = \frac{\left(\frac{1}{2}+\frac{1}{2}\mathrm{e}^{-\frac{1}{2}\frac{(4\alpha R)^2}{N}}-\mathrm{e}^{-\frac{(2\alpha R)^2}{N}}\right)^{\frac{1}{2}}}{\sqrt{N}\mathrm{e}^{-2\frac{(\alpha R)^2}{N}}} \overset{N\to\infty}{\longrightarrow} 0 \tag{A.8}$$

$$\begin{aligned}&\frac{\mathrm{Var}\left[\sum_{i=1}^{N} y_i\sin(\alpha y_i)\right]^{\frac{1}{2}}}{\mathrm{E}\left[\sum_{i=1}^{N} y_i\sin(\alpha y_i)\right]}\\ &= \frac{\left(\frac{1}{2}-\frac{1}{2}\mathrm{e}^{-\frac{1}{2}\frac{(2\alpha R)^2}{N}}+2\alpha^2\frac{R^2}{N}\mathrm{e}^{-\frac{1}{2}\frac{(2\alpha R)^2}{N}}-\alpha^2\frac{R^2}{N}\mathrm{e}^{-\frac{(\alpha R)^2}{N}}\right)^{\frac{1}{2}}}{\alpha R\mathrm{e}^{-\frac{1}{2}\frac{(\alpha R)^2}{N}}} \overset{N\to\infty}{\longrightarrow} 0.\end{aligned} \tag{A.9}$$

Note that the limit considerations hold for constant R and for a scaling relation $R^2 = N$, see (46). This scaling can also be motivated by investigating Fig. 3, where the increase of N shifts the central region (i.e. the region between the sphere limits) to larger R-values. For constant R the exponential factors yield "1" in the limit $N \to \infty$, such that the numerators of (A.7), (A.8), and (A.9) vanish. The denominators of (A.7) and (A.8) are also suppressing the ratio with $O\left(1/\sqrt{N}\right)$, while the denominator of (A.9) remains constant.

B PROGRESS RATE

In Sec. 5 the convergence properties are analyzed. To derive conditions satisfying $\varphi_R^{\mathrm{II}} = 0$ (or equivalently $\varphi_R^{\mathrm{II},*} = 0$), the relevant quantities are transformed to be functions of the normalized mutation σ^*. This ensures scale-invariance on the sphere function.

Setting $\sigma = \sigma^* R/N$ in (36), the variance yields

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

with h defined as

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

The function h was introduced to later solve $\varphi_R^{\mathrm{II}} = 0$ on both the sphere ($h = 0$) and the Rastrigin function.

The progress rate (35) yields after normalization $\varphi_R^{\mathrm{II},*} = \varphi_R^{\mathrm{II}}\frac{N}{2R^2}$ and setting $\sigma = \sigma^* R/N$

$$\varphi_R^{\mathrm{II},*}(\sigma^*,R) = c_\vartheta\frac{2R^2(\sigma^*R/N)^2\,[2+g]}{2R^2\frac{\sigma^*}{N}\sqrt{1+\frac{\sigma^{*2}}{2N}+h}}\frac{N}{2R^2}-\frac{N(\sigma^*R/N)^2}{\mu}\frac{N}{2R^2}, \tag{B.3}$$

with g defined as

$$g(\sigma^*,R) := \alpha^2 A\mathrm{e}^{-\frac{(\alpha R)^2}{2}\left(\frac{\sigma^{*2}}{N^2}+\frac{1}{N}\right)}. \tag{B.4}$$

Analogous to h, function g was introduced to discern between the sphere ($g = 0$) and the Rastrigin function. After simplification, (B.3) yields

$$\varphi_R^{\mathrm{II},*} = \frac{c_\vartheta\sigma^*}{2}\frac{2+g}{\sqrt{1+\frac{\sigma^{*2}}{2N}+h}}-\frac{\sigma^{*2}}{2\mu}. \tag{B.5}$$

Setting $g = h = 0$ in (B.5) recovers the sphere progress rate (39). Requiring $\varphi_R^{\mathrm{II},*} = 0$ in (B.5), the expression can be reformulated as

$$\sigma^{*4}+2N\sigma^{*2}(1+h(\sigma^*,R))-2N(c_\vartheta\mu)^2\left(2+g(\sigma^*,R)\right)^2 = 0. \tag{B.6}$$

The functional dependencies of g and h are explicitly written in (B.6) to illustrate the problem of solvability. We want to solve for $\sigma^*(R)$, which would give a relation yielding $\varphi_R^{\mathrm{II},*} = 0$. One can immediately see that with $g(\sigma^*, R)$ and $h(\sigma^*, R)$ containing exponential functions of σ^* and R, no closed form solution of (B.6) can be given. To obtain the zero of the sphere function $\sigma^*_{\varphi_0}$, one sets $g = h = 0$ in (B.6), such that

$$\sigma^{*4}+2N\sigma^{*2}-8Nc_\vartheta^2\mu^2 = 0. \tag{B.7}$$

The only positive non-complex solution of the fourth order Eq. (B.7) is

$$\sigma^*_{\varphi_0} = \left[\left[N^2+8Nc_\vartheta^2\mu^2\right]^{1/2}-N\right]^{1/2}, \tag{B.8}$$

which is the maximum normalized mutation strength giving positive progress on the sphere function (in the limit $N \to \infty$).

The limit of vanishing exponentials corresponds to $g = 0$, see (B.4), and $h = \frac{N^3A^2}{8R^4\sigma^{*2}}$, see (B.2), such that (B.6) yields the polynomial

$$\sigma^{*4}+2N\sigma^{*2}+\frac{N^4A^2}{4R^4}-8Nc_\vartheta^2\mu^2 = 0. \tag{B.9}$$

Solving (B.9) for $R = R_{\varphi_0}(\sigma^*)$, i.e., the residual distance of zero progress, yields

$$R_{\varphi_0}(\sigma^*) = \left(\frac{1}{4}\frac{N^4A^2}{8Nc_\vartheta^2\mu^2-2N\sigma^{*2}-\sigma^{*4}}\right)^{1/4}. \tag{B.10}$$

The smallest attainable value $R_{\varphi_0}(\sigma^*)$ is given at $\sigma^* = 0$ yielding the constant noise limit (45) for $\sigma^2_{\mathrm{Ras}} = NA^2/2$ according to

$$R_{\varphi_0}\big|_{\sigma^*=0} = \left(\frac{N^3A^2}{32c_\vartheta^2\mu^2}\right)^{1/4} = \left(\frac{NA^2}{2}\frac{N^2}{16c_\vartheta^2\mu^2}\right)^{1/4} = \sqrt{\frac{\sigma_{\mathrm{Ras}}N}{4c_\vartheta\mu}}. \tag{B.11}$$

The maximum σ^*-value yielding a real solution of (B.10) can be found by setting the denominator to zero, which can be identified as the (negative) polynomial of Eq. (B.7). Hence, $R_{\varphi_0}(\sigma^*)$ is defined on $\sigma^* \in [0, \sigma^*_{\varphi_0})$.

It can be easily shown, that the progress rate $\varphi_R^{\mathrm{II},*} > 0$ for any $R > R_{\varphi_0}$. This can be done by setting $g = 0$ and $h = \frac{N^3A^2}{8R^4\sigma^{*2}}$ in (B.5), and solving for R by demanding $\varphi_R^{\mathrm{II},*} > 0$. The calculations are straightforward and therefore not shown here.