Theoretical Computer Science 978 (2023) 114179

Contents lists available at ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Progress analysis of a multi-recombinative evolution strategy on the highly multimodal Rastrigin function ☆

Amir Omeradzic *, Hans-Georg Beyer

Vorarlberg University of Applied Sciences, Research Center Business Informatics, Hochschulstraße 1, 6850 Dornbirn, Austria

ARTICLE INFO

Article history:
Received 24 November 2022
Received in revised form 4 September 2023
Accepted 4 September 2023
Available online 12 September 2023
Communicated by B. Doerr

Keywords:
Evolution strategy
Progress rate
Global optimization
Rastrigin function

ABSTRACT

A first and second order progress rate analysis was conducted for the intermediate multi-recombinative Evolution Strategy $(\mu/\mu_I, \lambda)$-ES with isotropic scale-invariant mutations on the highly multimodal Rastrigin test function. Closed-form analytic solutions for the progress rates are obtained in the limit of large dimensionality and large populations. The first order results are able to model the one-generation progress including local attraction phenomena. Furthermore, a second order progress rate is derived yielding additional correction terms and further improving the progress model. The obtained results are compared to simulations and show good agreement, even for moderately large populations and dimensionality. The progress rates are applied within a dynamical systems approach, which models the evolution using difference equations. The obtained dynamics are compared to real averaged optimization runs and yield good agreement. The results improve further when dimensionality and population size are increased. Local and global convergence is investigated within given model showing that large mutations are needed to maximize the probability of global convergence, which comes at the expense of efficiency. An outlook regarding future research goals is provided.

1. Introduction

The theoretical analysis of the performance of Evolution Strategies (ES) [8] optimizing functions $f(\mathbf{y})$ in real-valued N-dimensional search spaces $\mathbf{y} \in \mathbb{R}^N$ is a challenge. This is due to the probabilistic nature of these algorithms allowing up to now the dynamic progress analysis only on simple test functions such as the sphere model [2,5], the ridge function class [3,14], and the ellipsoid model [7]. These test functions are simple w.r.t. their optimization landscape (also referred to as fitness landscape) in that they have at most one optimizer (i.e., the location $\mathbf{y}$ of the optimum). Analyzing the dynamical behavior of ES on more complex and multimodal test functions appears to be even more demanding. However, ES and other evolutionary algorithms are especially designated to optimize such problems. There is empirical evidence that ES are able to globally optimize highly multimodal optimization problems [11] with in N exponential number of local optima. The question arises how and when these ES are able to locate the global optimizer. It is the long term goal to find conditions the ES must fulfill to not get trapped in the vast amount of local optimizers. Ideally, a theoretical analysis should provide the answers regarding the success probability P_S (of locating the global optimum) depending on the ES parameters such as

☆ This article belongs to Section C: Theory of natural computing, Edited by Lila Kari.

* Corresponding author.
E-mail addresses: amir.omeradzic@fhv.at (A. Omeradzic), hans-georg.beyer@fhv.at (H.-G. Beyer).

https://doi.org/10.1016/j.tcs.2023.114179

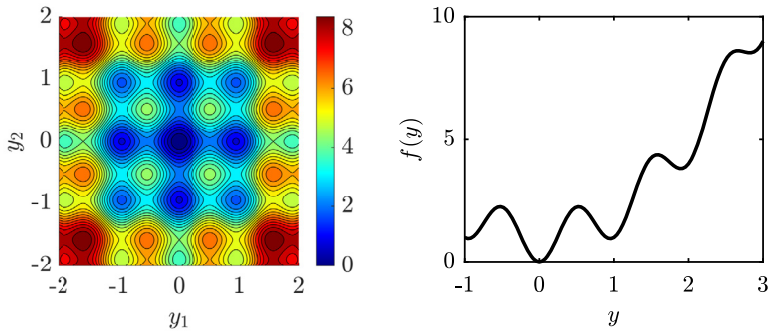

Fig. 1. The heat map shows the optimization landscape for $A = 1$, $\alpha = 2\pi$, and $N = 2$. The global minimizer located at the origin (dark blue) is surrounded by multiple local minima. On the right side the same parameter set is shown for $N = 1$. For increasing y the oscillation contribution is decreasing. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

the population size λ and the test function to be optimized. Furthermore, one is interested in the computational complexity of the optimization process.

One approach successfully applied to the analysis of the ES-performance on simple unimodal test functions mentioned above is the dynamical systems approach [5] which is based on progress rate analysis. The progress rate is a measure of expected positional change in search space between two generations depending on location, strategy and test function parameters. The idea of investigating global search behavior from expected local progress was successfully applied, among others, in [3,7]. It will be shown in this paper that this approach can be extended to the highly multimodal Rastrigin test function

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

where $\mathbf{y} \in \mathbb{R}^N$, with oscillation amplitude A and frequency parameter α. The i-th fitness component in Eq. (1) is defined as

$$f_i(y_i) \coloneqq y_i^2 + A(1 - \cos(\alpha y_i)). \tag{2}$$

Depending on A and α a finite number of local minima M can be observed for each component i. Therefore, the overall number of local minima is scaling as M^N posing a highly multimodal minimization problem with the global optimizer located at $\hat{\mathbf{y}} = \mathbf{0}$. An exemplary optimization landscape of the Rastrigin function is shown in Fig. 1.

The remarkable observation is that ES – unlike classical nonlinear optimization algorithms (e.g. BFGS) – do not follow the local gradient or Hessian ending in one of the $M^N - 1$ local optimizers. That is, ES perform a rather global search. A deeper understanding of this behavior is still missing. Recently, attempts have been made to analyze the problem from the viewpoint of relaxation using kernel smoothing [15]. However, the sampling process needed to transform the original problem into a convex optimization problem is still lacking a link to the ES.

In this paper a simplified and scale-invariant $(\mu/\mu_I, \lambda)$-ES, see Algorithm 1, is analyzed with step-size control defined in Eq. (4). Starting from the so-called parental centroid vector $\mathbf{y}^{(g)}$ a population of λ offspring are generated by adding isotropic Gaussian mutations $\mathbf{x} \sim \sigma \mathcal{N}(\mathbf{0}, \mathbf{1})$ with mutation strength σ in Lines 6 and 7. Thereafter, the fitness is evaluated in Line 8. Selection of the μ best individuals is done in Line 10. It is performed for a given selection (truncation) ratio defined as

$$\vartheta \coloneqq \frac{\mu}{\lambda}, \tag{3}$$

with $\vartheta \in (0, 1)$. It will be an essential quantity for the progress rate results in the limit of large population sizes. Using intermediate recombination with equal weights the best $m = 1, ..., \mu$ individuals are recombined in Line 11 and the new parental centroid $\mathbf{y}^{(g+1)}$ is obtained. In the following, the subscript "$m;\lambda$" can be read as the m-th best solution out of λ candidate solutions. In Line 12 the simplified step-size adaptation is performed. To this end, a constant normalized mutation σ^* using the spherical normalization with $\left\|\mathbf{y}^{(g)}\right\| = R^{(g)}$ is defined as

$$\sigma^* \coloneqq \frac{\sigma^{(g)} N}{\left\|\mathbf{y}^{(g)}\right\|} = \frac{\sigma^{(g)} N}{R^{(g)}}. \tag{4}$$

This property ensures scale invariance and therefore global convergence of the algorithm, as the mutation strength $\sigma^{(g)}$ decreases if and only if the residual distance $R^{(g)}$ decreases. The quantity σ^* is unknown during black-box optimizations, but it is very useful for theoretical investigations to obtain scale-invariant mutations strengths.

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

$g \leftarrow 0$
$\mathbf{y}^{(0)} \leftarrow \mathbf{y}^{(\text{init})}$
$\sigma^{(0)} \leftarrow \sigma^* \left\| \mathbf{y}^{(0)} \right\| / N$
repeat
  for $l \leftarrow 1, ..., \lambda$ **do**
    $\tilde{\mathbf{x}}_l \leftarrow \sigma^{(g)} \mathcal{N}_l(\mathbf{0}, \mathbf{1})$
    $\tilde{\mathbf{y}}_l \leftarrow \mathbf{y}^{(g)} + \tilde{\mathbf{x}}_l$
    $\tilde{f}_l \leftarrow f(\tilde{\mathbf{y}}_l)$
  end for
  $(\tilde{\mathbf{y}}_{1;\lambda}, \ldots, \tilde{\mathbf{y}}_{\mu;\lambda}) \leftarrow \text{sort}\left(\tilde{\mathbf{y}} \text{ w.r.t. ascending } \tilde{f}\right)$
  $\mathbf{y}^{(g+1)} \leftarrow \frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{\mathbf{y}}_{m;\lambda}$
  $\sigma^{(g+1)} \leftarrow \sigma^* \left\| \mathbf{y}^{(g+1)} \right\| / N$
  $g \leftarrow g + 1$
until termination criterion

The remainder of this paper is organized as follows. In the next section the local performance measures will be introduced being the basis for both the progress rate analysis and the dynamical systems approach. Section 3 is devoted to the determination and evaluation of the first order progress rate. Section 4 describes the derivation of the second order progress rate, which will rely on first order progress rate results. Section 5 uses the local performance measures to establish the evolution equations that govern the dynamical behavior of the ES. Experiments will be presented to show the usefulness of the approach. In the final Section 6 conclusions will be drawn and being based on open problems the further research direction will be outlined.

2. Local performance measures and quality gain distribution

The performance of an ES between two generations can be evaluated in both fitness and search space. The quality gain $Q_{\mathbf{y}}(\mathbf{x})$ of fitness f at a position $\mathbf{y}^{(g)}$ due to an isotropic mutation $\mathbf{x} \sim \sigma \mathcal{N}(\mathbf{0}, \mathbf{1})$ is defined as

$$Q_{\mathbf{y}}(\mathbf{x}) \coloneqq f\left(\mathbf{y}^{(g)} + \mathbf{x}\right) - f\left(\mathbf{y}^{(g)}\right), \tag{5}$$

and yields in the case of fitness improvement (minimization considered) a negative value $Q_{\mathbf{y}} < 0$. The definition (5) measures the fitness change before selection and will be needed for the evaluation of the two progress rates (7) and (8). The quality gain components are decomposed using f_i from Eq. (2) as $Q_i \coloneqq f_i(y_i + x_i) - f_i(y_i)$, such that

$$Q_{\mathbf{y}}(\mathbf{x}) = \sum_{i=1}^{N} Q_i(x_i) = \sum_{i=1}^{N} \left[f_i\left(y_i^{(g)} + x_i\right) - f_i\left(y_i^{(g)}\right) \right]. \tag{6}$$

That is, the quality gain corresponds to the difference between fitness values before and after the mutation application. A probabilistic model for the distribution of quality values will be presented below. It will be important for the subsequent progress rate derivations, as selection is based on fitness values.

Analyzing the progress towards the optimizer in search space, the first order progress rate on the Rastrigin function has already been investigated in [17] as a first approach. In this paper, a new approach is presented which significantly improves the prediction quality.

The first order progress rate between two generations for the parental component y_i is defined as

$$\varphi_i \coloneqq \mathrm{E}\left[y_i^{(g)} - y_i^{(g+1)} \,\middle|\, \mathbf{y}^{(g)}, \sigma^{(g)} \right], \tag{7}$$

given parental position $\mathbf{y}^{(g)}$ and mutation strength $\sigma^{(g)}$ at generation g. It is a measure of expected positional difference in search space. Positive expected progress $\varphi_i > 0$ is defined in the case $y_i^{(g)} > \mathrm{E}[y_i^{(g+1)}]$ for $y_i^{(g)} > 0$ and $\mathrm{E}[y_i^{(g+1)}] > 0$. In this case the distance to the optimizer $\hat{y}_i = 0$ is reduced in expectation. This assumption is only valid as long as the sign of $\mathrm{E}[y_i^{(g+1)}]$ does not change, i.e., for small mutations compared to the residual distance. Therefore φ_i has limited applicability when studying the convergence behavior in the vicinity of the optimizer. As has been shown in [7] regarding the performance analysis on the ellipsoid model, a second order progress rate is needed. It is defined as

$$\varphi_i^{\mathrm{II}} \coloneqq \mathrm{E}\left[\left(y_i^{(g)}\right)^2 - \left(y_i^{(g+1)}\right)^2 \,\middle|\, \mathbf{y}^{(g)}, \sigma^{(g)} \right]. \tag{8}$$

Squaring the positions yields $\varphi_i^{\mathrm{II}} > 0$ independent of the sign, if the distance to $\hat{y}_i = 0$ decreases in expectation. Additionally, the derivation will yield expressions containing a progress gain and loss part, which is necessary for a more accurate model of convergence. Both progress rates will be expressed using integral equations for the expected values and approximations will be necessary to find closed-form solutions. In a second step the progress rates can be applied within difference equations to model the expected dynamics over many generations in order to investigate the global convergence behavior.

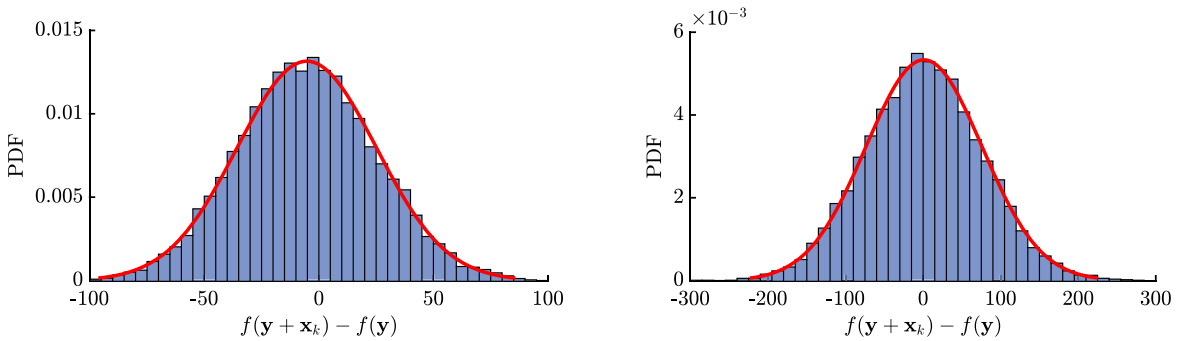

Fig. 2. The histograms show sampled values of $Q_{\mathbf{y}}(\mathbf{x})$ from (5) with fixed $\mathbf{y}$ by applying random mutations $\mathbf{x}_k \sim \sigma\mathcal{N}(\mathbf{0},\mathbf{1})$ ($\sigma = 1$ with $k = 1, ..., 10^4$ samples) at $N = 10$ (left) and $N = 100$ (right) with $A = 10$. The $\mathbf{y}$-values were initialized randomly at $\|\mathbf{y}\| = 10$ where local attraction is significant. The red envelope curves show the respective normal approximation (9) using mean value (30) and variance (31). The p-values of the Anderson-Darling-test for normality are $p = 0.48$ (left) and $p = 0.53$ (right).

The selection of individuals is based on the attained fitness values. The quality gain measures the fitness change before selection according to (5). When the progress rate of an ES is modeled, the cumulative distribution function (CDF) $P_Q(q)$ of the quality gain and its probability density function (PDF) $p_Q(q)$ are needed as a function of $\mathbf{y}$ and σ. Obtaining an exact CDF for $Q_{\mathbf{y}}(\mathbf{x})$ is not feasible at this point. Since $Q_{\mathbf{y}}(\mathbf{x}) = \sum_{i=1}^{N} Q_i(x_i)$ with independent random variables Q_i, the application of the Central Limit Theorem seems appropriate to show that the distribution is asymptotically normal.[1] However, proving its validity rigorously seems hard or even impossible for arbitrary $\mathbf{y}$. Therefore, we resort to normality as an approximation for the quality gain distribution. This is backed up by experimental results in Fig. 2, where sampled $Q_{\mathbf{y}}(\mathbf{x})$-values are compared to the normal approximation. A standard Anderson-Darling test was performed to check whether the sampled data was drawn from a normal distribution with known mean and variance according to (9). The hypothesis test fails to reject the normality assumption at p-values $p = 0.48$ (left) and $p = 0.53$ (right), where rejection is usually defined for $p < 0.05$. Even at relatively small $N = 10$ the results agree well. Good experimental agreement is also observed for the variation of the location $\mathbf{y}$ and mutation strength σ (not shown). Therefore, the normality assumption does not pose a strong restriction on the overall prediction quality of the progress rates in the subsequent sections, such that we approximate

$$Q_{\mathbf{y}}(\mathbf{x}) = \sum_{i=1}^{N} Q_i(x_i) \sim \mathcal{N}(\mathrm{E}\left[Q_{\mathbf{y}}(\mathbf{x})\right], \mathrm{Var}\left[Q_{\mathbf{y}}(\mathbf{x})\right]). \tag{9}$$

Furthermore, the following abbreviations are introduced

$$E_Q := \mathrm{E}\left[Q_{\mathbf{y}}(\mathbf{x})\right] = \sum_{i=1}^{N} \mathrm{E}\left[Q_i\right] \tag{10}$$

$$D_Q^2 := \mathrm{Var}\left[Q_{\mathbf{y}}(\mathbf{x})\right] = \sum_{i=1}^{N} \mathrm{Var}\left[Q_i\right]. \tag{11}$$

At this point an additional assumption for the coordinates $\mathbf{y} = (y_1, ..., y_N)$ has to be made to justify subsequent variance approximations (13) and (14). Given the search vector $\mathbf{y} = (y_1, ..., y_N)$ and residual distance $R^2 = \|\mathbf{y}\|^2$ it is assumed that the components contribute *approximately* equally (in expectation) to the residual distance, i.e., there is no dominating component, such that

$$\frac{y_i^2}{R^2} \approx \frac{1}{N}, \quad \text{for all } i = 1, ..., N. \tag{12}$$

Property (12) will also be referred to as *component equipartition*. The concept was introduced in [6] and proven for the noisy ellipsoid in [12]. Its applicability to the Rastrigin function was shown in [19]. The equipartition assumption is necessary in

[1] For independently distributed quality gain components $Q_i(x_i)$ with finite mean and variance the Central Limit Theorem holds [10], if for some $\delta > 0$ the Lyapunov condition

$$\lim_{N\to\infty} \frac{\sum_{i=1}^{N} \mathrm{E}\left[|Q_i - \mathrm{E}[Q_i]|^{2+\delta}\right]}{D_Q^{2+\delta}} = 0$$

holds. The validation of the condition could be approached using Eqs. (24), (25), and (26) for the respective quantities by evaluating higher order moments.

order to justify certain approximation steps and to provide a closed-form solution for the progress rate. Furthermore, it will be a reasonable assumption to obtain a model of the algorithm's progress and dynamics in expectation. This assumption also justifies a linear scaling of the variance with dimensionality N provided that the components are contributing equally to the overall variance, such that

$$D_Q^2 = \sum_{i=1}^{N} \mathrm{Var}\left[Q_i\right] = \Theta(N). \tag{13}$$

Additionally, for large N an important approximation will be used for the variance to significantly simplify the obtained lengthy results. If no single i-th component is dominating the sum, i.e., $\mathrm{Var}\left[Q_i\right] / \sum_{j\neq i} \mathrm{Var}\left[Q_j\right] \to 0$ (for any i in the limit $N \to \infty$), the contribution of a single term is negligible for $N \to \infty$. Therefore, the two sums over N and $N-1$ terms, respectively, are asymptotically equal with

$$D_Q^2 = \sum_{i=1}^{N} \mathrm{Var}\left[Q_i\right] \simeq \sum_{j\neq i} \mathrm{Var}\left[Q_j\right] = D_i^2. \tag{14}$$

Note that quantity D_i^2 is formally introduced in (20). Returning to Eq. (9), the expression is rewritten using a standardized random variate Z as

$$Z = \frac{Q_{\mathbf{y}}(\mathbf{x}) - E_Q}{D_Q} \overset{N\to\infty}{\sim} \mathcal{N}(0,1). \tag{15}$$

Approximation 1 *(Quality gain distribution).* The local quality gain at position $\mathbf{y}$ due to random mutation vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{1})$ is approximately normally distributed. Therefore, $P_Q(q)$ and $p_Q(q)$ can be approximated as

$$\widetilde{P}_Q(q) = \Phi\left(\frac{q - E_Q}{D_Q}\right) \tag{16}$$

$$\widetilde{p}_Q(q) = \frac{1}{\sqrt{2\pi}D_Q} \exp\left[-\frac{1}{2}\left(\frac{q - E_Q}{D_Q}\right)^2\right]. \tag{17}$$

Within the normal approximation (16) the inverse $\widetilde{P}_Q^{-1}(p)$ given some probability p can be easily obtained by using the quantile function $\Phi^{-1}(p)$ of the normal distribution. This relation will be used later to obtain a quality gain for some given probability p using

$$q = E_Q + D_Q\,\Phi^{-1}(p). \tag{18}$$

For the derivation of the i-th component progress rate the conditional distribution function $P_Q(q|x_i)$ of the quality gain is needed for a given component x_i. In this case expected value and variance are given by

$$E_{Q|x_i} \coloneqq \mathrm{E}\left[Q_{\mathbf{y}}(\mathbf{x})|x_i\right] = Q_i(x_i) + \sum_{j\neq i} \mathrm{E}\left[Q_j\right] \tag{19}$$

$$D_i^2 \coloneqq \mathrm{Var}\left[Q_{\mathbf{y}}(\mathbf{x})|x_i\right] = \sum_{j\neq i} \mathrm{Var}\left[Q_j\right], \tag{20}$$

where the sum $j \neq i$ is taken for fixed i over the remaining $N-1$ components. Therefore, a normal approximation for the conditional CDF is introduced using (19) and (20).

Approximation 2 *(Quality gain distribution given x_i).* The quality gain distribution at position $\mathbf{y}$ given fixed mutation component x_i and random mutation vector $(\mathbf{x})_{j\neq i} \sim (\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{1}))_{j\neq i}$ is approximately normally distributed. Therefore, $P_Q(q|x_i)$ and $p_Q(q|x_i)$ can be approximated as

$$\widetilde{P}_Q(q|x_i) = \Phi\left(\frac{q - E_{Q|x_i}}{D_i}\right) \tag{21}$$

$$\widetilde{p}_Q(q|x_i) = \frac{1}{\sqrt{2\pi}D_i} \exp\left[-\frac{1}{2}\left(\frac{q - E_{Q|x_i}}{D_i}\right)^2\right]. \tag{22}$$

Having derived approximations of the quality gain distribution functions, the quantities $\mathrm{E}[Q_i]$ and $\mathrm{Var}[Q_i]$ remain to be determined. As the components are independent, it is sufficient to consider a single component and then perform the summation. Starting from definition (6), one can evaluate the quality gain of a single component $Q_i(x_i)$. After applying trigonometric identity $\cos(\alpha(y_i + x_i)) = \cos(\alpha y_i)\cos(\alpha x_i) - \sin(\alpha y_i)\sin(\alpha x_i)$, one gets

$$Q_i(x_i) = f_i(y_i + x_i) - f_i(y_i) \tag{23}$$

$$= x_i^2 + 2y_i x_i + A\cos(\alpha y_i) - A\cos(\alpha y_i)\cos(\alpha x_i) + A\sin(\alpha y_i)\sin(\alpha x_i), \tag{24}$$

of which $\mathrm{E}[Q_i]$ and $\mathrm{Var}[Q_i] = \mathrm{E}\left[Q_i^2\right] - \mathrm{E}[Q_i]^2$ need to be evaluated. The results will be expressed as expected values containing trigonometric functions. As a remark, terms containing moments of $x_i \sim \mathcal{N}(0, \sigma^2)$, i.e., $\mathrm{E}\left[x_i^k\right]$ with $k \geq 1$, are silently evaluated as they are assumed to be widely known. Starting with $\mathrm{E}[Q_i]$ one has

$$\mathrm{E}[Q_i] = \sigma^2 + A\cos(\alpha y_i)(1 - \mathrm{E}[\cos(\alpha x_i)]), \tag{25}$$

where odd powers of $\mathrm{E}\left[x_i^k\right] = 0$, which also yields $\mathrm{E}[\sin(\alpha x_i)] = 0$. Evaluating $\mathrm{Var}[Q_i]$ yields

$$\begin{aligned}\mathrm{Var}[Q_i] &= \mathrm{E}\left[Q_i^2\right] - \mathrm{E}[Q_i]^2 \\ &= 2\sigma^4 + 4y_i^2\sigma^2 + A^2\sin^2(\alpha y_i)\,\mathrm{Var}[\sin(\alpha x_i)] \\ &\quad + A^2\cos^2(\alpha y_i)\,\mathrm{Var}[\cos(\alpha x_i)] - 2A\cos(\alpha y_i)\,\mathrm{E}\left[x_i^2\cos(\alpha x_i)\right] \\ &\quad + 2A\sigma^2\cos(\alpha y_i)\,\mathrm{E}[\cos(\alpha x_i)] + 4Ay_i\sin(\alpha y_i)\,\mathrm{E}[x_i\sin(\alpha x_i)]\,.\end{aligned} \tag{26}$$

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

$$\chi_x(\alpha) = \mathrm{E}\left[\mathrm{e}^{\imath\alpha x}\right] = \mathrm{e}^{\imath\alpha\mu - \frac{1}{2}\alpha^2\sigma^2} = \mathrm{e}^{-\frac{1}{2}\alpha^2\sigma^2}[\cos(\alpha\mu) + \imath\sin(\alpha\mu)], \tag{27}$$

with the imaginary unit denoted by $\imath = \sqrt{-1}$ in (27) and (28). Now the k-th derivatives with respect to α can be applied to both sides

$$\begin{aligned}\frac{\mathrm{d}^k}{\mathrm{d}\alpha^k}\mathrm{E}\left[\mathrm{e}^{\imath\alpha x}\right] = \mathrm{E}\left[\frac{\mathrm{d}^k}{\mathrm{d}\alpha^k}\mathrm{e}^{\imath\alpha x}\right] &= \mathrm{E}\left[\frac{\mathrm{d}^k}{\mathrm{d}\alpha^k}\cos(\alpha x)\right] + \imath\,\mathrm{E}\left[\frac{\mathrm{d}^k}{\mathrm{d}\alpha^k}\sin(\alpha x)\right] \\ &\stackrel{!}{=} \frac{\mathrm{d}^k}{\mathrm{d}\alpha^k}\left[\mathrm{e}^{-\frac{(\alpha\sigma)^2}{2}}[\cos(\alpha\mu) + \imath\sin(\alpha\mu)]\right],\end{aligned} \tag{28}$$

such that corresponding real and imaginary parts can be identified by comparing both sides (denoted by $\stackrel{!}{=}$) of Eq. (28). Given $\mu = 0$ for $k = \{0, 1, 2\}$ the required expectations of trigonometric terms can be derived. Additionally, trigonometric identities $\cos^2(x) = 1/2 + \cos(2x)/2$ and $\sin^2(x) = 1/2 - \cos(2x)/2$ are used. The results are

$$\begin{aligned}\mathrm{E}[\cos(\alpha x)] &= \mathrm{e}^{-\frac{(\alpha\sigma)^2}{2}}, \quad \mathrm{E}\left[\cos^2(\alpha x)\right] = \frac{1}{2} + \frac{1}{2}\mathrm{e}^{-\frac{(2\alpha\sigma)^2}{2}} \\ \mathrm{E}\left[\sin^2(\alpha x)\right] &= \frac{1}{2} - \frac{1}{2}\mathrm{e}^{-\frac{(2\alpha\sigma)^2}{2}}, \quad \mathrm{E}[x\sin(\alpha x)] = \alpha\sigma^2\mathrm{e}^{-\frac{(\alpha\sigma)^2}{2}} \\ \mathrm{E}\left[x^2\cos(\alpha x)\right] &= (\sigma^2 - \alpha^2\sigma^4)\mathrm{e}^{-\frac{(\alpha\sigma)^2}{2}}, \quad \mathrm{Var}[(\cdot)] = \mathrm{E}\left[(\cdot)^2\right] - \mathrm{E}[(\cdot)]^2\,.\end{aligned} \tag{29}$$

Inserting relations (29) into (25) and (26), summing over all N components and collecting the resulting terms one obtains the expected value

$$E_Q = \sum_{i=1}^{N}\left[\sigma^2 + A\cos(\alpha y_i)\left(1 - \mathrm{e}^{-\frac{(\alpha\sigma)^2}{2}}\right)\right]. \tag{30}$$

Analogously, the variance of the Rastrigin quality gain yields

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

The quantities $E_{Q|x_i}$ from (19) and D_i^2 from (20) are given analogously by summing over $N-1$ components. Expressions E_Q and D_Q could be inserted into (16), and $E_{Q|x_i}$ with $Q_i(x_i)$ and D_i into (21). However, it is omitted at this point for better readability.

As an important remark, expression (23) can be linearized w.r.t. mutation x_i to obtain analytically solvable progress rate integrals, see also discussion after Eq. (51). Taylor-expanding f_i around y_i for small x_i gives $f_i(y_i+x_i) = f_i(y_i) + \frac{\partial f_i}{\partial y_i}x_i + O\left(x_i^2\right)$, such that after setting $f_i' \coloneqq \frac{\partial f_i}{\partial y_i}$ and evaluating the derivative one has

$$\begin{aligned} Q_i(x_i) &= f_i(y_i+x_i) - f_i(y_i) = f_i' x_i + O\left(x_i^2\right) \\ &= (2y_i + \alpha A \sin(\alpha y_i))x_i + O\left(x_i^2\right) = (k_i + d_i)x_i + O\left(x_i^2\right), \end{aligned} \tag{32}$$

with following definitions applied to (32)

$$f_i' \coloneqq k_i + d_i, \quad \text{with} \quad k_i \coloneqq 2y_i, \quad \text{and} \quad d_i \coloneqq \alpha A \sin(\alpha y_i). \tag{33}$$

Component k_i is the derivative of the quadratic term y_i^2, cf. Eq. (2), which follows the global quadratic structure of the function. Conversely, derivative d_i follows the local oscillation, such that it will be very important for the model of local attraction during the progress rate derivations in Secs. 3 and 4.

3. First order progress rate

While the first order progress rate (7) does not suffice to completely describe the convergence behavior of the ES on Rastrigin, see Sec. 5, it is a necessary step in the calculation of the second order progress rate in Sec. 4. Given definition (7) and the parental location $\mathbf{y}^{(g)}$, one has to find the expected value over the i-component location $\mathrm{E}\left[y_i^{(g+1)}\right]$. The positional update $\mathbf{y}^{(g)} \to \mathbf{y}^{(g+1)}$ performed by the ES is realized by consecutively applying mutation, selection, and recombination (see Algorithm 1), such that one can write

$$\mathbf{y}^{(g+1)} = \frac{1}{\mu}\sum_{m=1}^{\mu}(\mathbf{y}^{(g)} + \mathbf{x}_{m;\lambda}) = \mathbf{y}^{(g)} + \frac{1}{\mu}\sum_{m=1}^{\mu}\mathbf{x}_{m;\lambda}, \tag{34}$$

where $\mathbf{x}_{m;\lambda}$ denotes the mutation vector of the m-th best offspring after selection. Considering the i-th component of Eq. (34), abbreviating the mutation component as $x_{m;\lambda} \coloneqq (\mathbf{x}_{m;\lambda})_i$, and taking the expected value thereof yields

$$\mathrm{E}\left[y_i^{(g+1)}\middle|\mathbf{y}^{(g)}, \sigma^{(g)}\right] = y_i^{(g)} + \frac{1}{\mu}\sum_{m=1}^{\mu}\mathrm{E}\left[x_{m;\lambda}\middle|\mathbf{y}^{(g)}, \sigma^{(g)}\right]. \tag{35}$$

The progress rate can therefore be evaluated by inserting (35) into (7) giving

$$\varphi_i = -\frac{1}{\mu}\sum_{m=1}^{\mu}\mathrm{E}\left[x_{m;\lambda}\middle|\mathbf{y}^{(g)}, \sigma^{(g)}\right]. \tag{36}$$

Before starting the derivation of (36), the important large population theorem is stated which will be used during the derivation of both first and second order progress rate. Its application also yields the so-called asymptotic generalized progress coefficients presented in Eq. (45).

Theorem 1. *Let $\lambda > \mu + 1$ and $\mu > a$ with $a \geq 1$ and $\vartheta = \mu/\lambda$ with $0 < \vartheta < 1$, such that $t^{\lambda-\mu-1}(1-t)^{\mu-a}$ exhibits its maximum on $(0,1)$ and vanishes at $t \in \{0,1\}$. Let $f_x(t)$ be a function defined for constant $x \in \mathbb{R}$, such that $f_x\colon [0,1] \to [0,1]$ with bounded derivatives on $[0,1]$ and let B denote the beta function. Furthermore, let p_x denote the PDF of a normally distributed variate and let $p_n(x)$ denote a polynomial of degree n in x. For infinitely large $\mu, \lambda \to \infty$ and constant $\vartheta = \mu/\lambda$ the following limit holds*

$$\begin{aligned} &\lim_{\substack{\mu,\lambda\to\infty \\ \vartheta=\text{const.}}} \int_{-\infty}^{\infty} p_n(x)p_x(x)\frac{1}{\mathrm{B}(\lambda-\mu,\mu)}\int_0^1 t^{\lambda-\mu-1}(1-t)^{\mu-a}f_x(t)\,\mathrm{d}t\,\mathrm{d}x \\ &\quad = \frac{1}{\vartheta^{a-1}}\int_{-\infty}^{\infty} p_n(x)p_x(x)f_x(1-\vartheta)\,\mathrm{d}x\,. \end{aligned} \tag{37}$$

Proof. The dominated convergence theorem is applied. First, the following sequence is defined for $\mu = 1, 2, ...$, with $\lambda(\mu) = \mu/\vartheta$ and constant ϑ

$$g_\mu(x) \coloneqq \frac{1}{\mathrm{B}(\lambda-\mu,\mu)} \int_0^1 t^{\lambda-\mu-1}(1-t)^{\mu-a} f_x(t)\,\mathrm{d}t\,. \tag{38}$$

Note that g_μ is measured over the density of the normal distribution. In [18] it was shown that $g_\mu(x)$ converges for any x according to

$$\lim_{\substack{\mu,\lambda\to\infty \\ \vartheta=\text{const.}}} g_\mu(x) = \frac{f_x(1-\vartheta)}{\vartheta^{a-1}}. \tag{39}$$

An upper bound of g_μ can be estimated using $0 \le f_x \le 1$ and the definition of the beta function $\mathrm{B}(z_1,z_2) = \int_0^1 t^{z_1-1}(1-t)^{z_2-1}\,\mathrm{d}t$ as

$$\begin{aligned}
\left|g_\mu(x)\right| &\le \frac{\mathrm{B}(\lambda-\mu,\mu-a+1)}{\mathrm{B}(\lambda-\mu,\mu)} = \frac{(\lambda-\mu-1)!(\mu-a)!}{(\lambda-a)!}\,\frac{(\lambda-1)!}{(\lambda-\mu-1)!(\mu-1)!} \\
&= \frac{(\lambda-1)(\lambda-2)\cdots(\lambda-a+1)(\lambda-a)!}{(\mu-1)(\mu-2)\cdots(\mu-a+1)(\mu-a)!}\,\frac{(\mu-a)!}{(\lambda-a)!} \\
&= \left(\frac{\lambda}{\mu}\right)^{a-1} \frac{(1-1/\lambda)\cdots(1-(a-1)/\lambda)}{(1-1/\mu)\cdots(1-(a-1)/\mu)} \\
&\le \frac{1}{\vartheta^{a-1}}\,\frac{1}{(1-(a-1)/\mu)^{a-1}}.
\end{aligned} \tag{40}$$

A lower bound for the denominator of (40) can be given as

$$\left(1-\frac{a-1}{\mu}\right)^{a-1} \ge \frac{1}{a^{a-1}}. \tag{41}$$

Inequality (41) can be shown easily by setting $\mu = a+k$ with integers $a \ge 1$ and $k \ge 1$ (ensuring $\mu > a$). This yields

$$\begin{aligned}
1-\frac{a-1}{\mu} = \frac{a+k-a+1}{a+k} &\ge \frac{1}{a} \\
ak &\ge k,
\end{aligned} \tag{42}$$

which is fulfilled for any $a \ge 1$ and $k \ge 1$. Using (41) in (40) one gets

$$\left|g_\mu(x)\right| \le \left(\frac{a}{\vartheta}\right)^{a-1}. \tag{43}$$

As there is a constant upper bound of $\left|g_\mu(x)\right|$, it remains to show that

$$\begin{aligned}
\int_{-\infty}^{\infty} |p_n(x)|\,p_x(x)\,\mathrm{d}x &\le \int_{-\infty}^{\infty} \left|\sum_{k=0}^{n} a_k x^k\right| p_x(x)\,\mathrm{d}x \le \int_{-\infty}^{\infty} \sum_{k=0}^{n} |a_k| \left|x^k\right| p_x(x)\,\mathrm{d}x \\
&\le 2\sum_{k=0}^{n} |a_k| \int_0^{\infty} x^k p_x(x)\,\mathrm{d}x < \infty,
\end{aligned} \tag{44}$$

which is finite due to normal density $p_x(x)$. Hence, the limit in Eq. (37) can be exchanged with the integral over x. Using the limit of (39) the desired result is obtained. □

The limit (39) is readily used in [16] to define the so-called asymptotic generalized progress coefficients for integers $a \ge 1$, $b \ge 0$, and truncation ratio $0 < \vartheta < 1$ as

$$e_\vartheta^{a,b} \coloneqq \left[\frac{\mathrm{e}^{-\frac{1}{2}\left[\Phi^{-1}(\vartheta)\right]^2}}{\sqrt{2\pi}\,\vartheta}\right]^a \left[-\Phi^{-1}(\vartheta)\right]^b. \tag{45}$$

These are characteristic coefficients describing the progress in the limit $\mu, \lambda \to \infty$ with constant $\vartheta = \mu/\lambda$, and are related to the generalized progress coefficients [5, Eq. (5.112)]. They will reappear during the derivation of both φ_i and φ_i^{II}. The derivation of φ_i is presented now.

Proposition 1. *Let $\mu, \lambda \in \mathbb{N}$ with $\mu \geq 1$ and $\mu < \lambda$ and let p_x denote the PDF of the random mutation $x \sim \mathcal{N}(0, \sigma^2)$. Let $x_{m;\lambda}$ denote the m-th best value (out of λ) of the i-th mutation component $(\mathbf{x}_{m;\lambda})_i$. Furthermore, let P_Q and P_Q^{-1} denote the quality gain CDF (and its inverse), respectively, with* B *denoting the beta function. Then, the first order component-wise progress rate is given by*

$$\begin{aligned}\varphi_i &= -\frac{1}{\mu}\sum_{m=1}^{\mu} \mathrm{E}\left[x_{m;\lambda}\right] \\ &= -\frac{\lambda}{\mu}\int_{x_i=-\infty}^{x_i=\infty} x_i\, p_x(x_i)\frac{1}{\mathrm{B}(\lambda-\mu,\mu)}\int_{t=0}^{t=1} t^{\lambda-\mu-1}(1-t)^{\mu-1} P_Q\left(P_Q^{-1}(1-t)|x_i\right)\mathrm{d}t\,\mathrm{d}x_i\,.\end{aligned} \tag{46}$$

Proof. From now on the conditional dependency on $\mathbf{y}^{(g)}$ and $\sigma^{(g)}$ will be implicitly assumed as given for better readability of the equations. The expected value of the i-th mutation component $x_{m;\lambda}$ after selection can be expressed as an integral over the order statistic density $p_{m;\lambda}(x_i)$ of the m-th best individual, such that (36) is rewritten as

$$\varphi_i = -\frac{1}{\mu}\sum_{m=1}^{\mu} \mathrm{E}\left[x_{m;\lambda}\right] = -\frac{1}{\mu}\sum_{m=1}^{\mu}\int_{-\infty}^{\infty} x_i\; p_{m;\lambda}(x_i)\,\mathrm{d}x_i\,. \tag{47}$$

The subsequent task will be to derive the density $p_{m;\lambda}$ as a function of mutation and quality gain distributions. Mutations are distributed normally with zero mean and variance σ^2 according to the normal density

$$p_x(x_i) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x_i}{\sigma}\right)^2\right]. \tag{48}$$

Given mutation x_i (and implicitly position $\mathbf{y}$), a random quality gain value Q is distributed according to a conditional probability density $p_Q(q|x_i)$. Given that the m-th best individual attains a quality gain within $[q, q+\mathrm{d}q]$, there must be $m-1$ better individuals having a smaller quality value with probability $[\Pr\{Q \leq q\}]^{m-1} = \left[P_Q(q)\right]^{m-1}$, and $\lambda - m$ individuals having a larger value with $[\Pr\{Q > q\}]^{\lambda-m} = \left[1 - P_Q(q)\right]^{\lambda-m}$. To account for all relevant combinations one has $\frac{\lambda!}{(m-1)!(\lambda-m)!}$, where $1/(m-1)!$ and $1/(\lambda-m)!$ exclude the irrelevant combinations among the two groups of better and worse individuals, respectively. The conditional density for the m-th individual as a function of the quality gain q yields

$$p_{Q;m;\lambda}(q|x_i) = \frac{\lambda!}{(m-1)!(\lambda-m)!} p_Q(q|x_i) P_Q(q)^{m-1}[1-P_Q(q)]^{\lambda-m}. \tag{49}$$

By integrating (49) over all attainable quality gain values $q \in [q_l, q_u]$, one arrives at the density

$$p_{m;\lambda}(x_i) = p_x(x_i)\frac{\lambda!}{(m-1)!(\lambda-m)!}\int_{q_l}^{q_u} p_Q(q|x_i) P_Q(q)^{m-1}[1-P_Q(q)]^{\lambda-m}\,\mathrm{d}q\,. \tag{50}$$

Inserting the order statistic density from (50) into the progress rate (47), one obtains the intermediate result

$$\varphi_i = -\frac{1}{\mu}\sum_{m=1}^{\mu}\frac{\lambda!}{(m-1)!(\lambda-m)!}\int_{-\infty}^{\infty} x_i p_x(x_i)\int_{q_l}^{q_u} p_Q(q|x_i) P_Q(q)^{m-1}[1-P_Q(q)]^{\lambda-m}\,\mathrm{d}q\,\mathrm{d}x_i\,. \tag{51}$$

A few important remarks can be made regarding Eq. (51). A closed-form analytic solution cannot be obtained without applying further approximations. It can be approached in an analogous way to the φ_i-derivation of the Ellipsoid in [13] to obtain a solution in terms of the well-known progress coefficient $c_{\mu/\mu,\lambda}$ [5, p. 216]. However, a closed-form solution with this approach requires a linear relation of Q_i w.r.t. x_i, see relation (32). The effect of a linearized quality gain on the progress rate of the Rastrigin function was already studied in [17] and showed that the progress due to local attraction is not modeled correctly, as the oscillation terms have to be either dropped or linearized for small x_i.

Therefore a different approach is followed here assuming the infinite population limit, an approach which was applied within the analysis of functions with noise-induced multi-modality [9]. The approach will yield correction terms including the effects of the trigonometric terms from (24), in contrast to only taking linearized terms from (32). Starting from Eq. (51) and moving the sum including the m-dependent prefactors into the innermost integral yields

$$\varphi_i = -\frac{\lambda!}{\mu}\int_{-\infty}^{\infty} x_i p_x(x_i)\int_{q_l}^{q_u} p_Q(q|x_i)\sum_{m=1}^{\mu}\frac{P_Q(q)^{m-1}[1-P_Q(q)]^{\lambda-m}}{(m-1)!(\lambda-m)!}\,\mathrm{d}q\,\mathrm{d}x_i\,. \tag{52}$$

Now a transformation can be applied for the sum $\sum_m(\cdot)$ yielding an expression as a function of the regularized incomplete beta function [5, p. 147]. One has

$$\sum_{m=1}^{\mu} \frac{P(q)^{m-1}[1-P(q)]^{\lambda-m}}{(m-1)!(\lambda-m)!} = \frac{1}{(\lambda-\mu-1)!(\mu-1)!} \int_0^{1-P(q)} t^{\lambda-\mu-1}(1-t)^{\mu-1}\,\mathrm{d}t\,. \tag{53}$$

Furthermore, one can rewrite the resulting population-dependent factor as follows

$$\frac{\lambda!}{\mu}\frac{1}{(\lambda-\mu-1)!(\mu-1)!} = \frac{\lambda}{\mu}\frac{(\lambda-1)!}{(\lambda-\mu-1)!(\mu-1)!} = \frac{\lambda}{\mu}\frac{\Gamma(\lambda)}{\Gamma(\lambda-\mu)\,\Gamma(\mu)} = \frac{\lambda}{\mu}\frac{1}{\mathrm{B}(\lambda-\mu,\mu)}, \tag{54}$$

where we have used the property of the gamma function $\Gamma(n) = (n-1)!$ (for any integer $n > 0$) and the known relation between gamma and beta functions $\frac{\Gamma(x)\,\Gamma(y)}{\Gamma(x+y)} = \mathrm{B}(x,y)$. These replacements will be useful later. After replacing the sum and refactoring we arrive at the following progress rate integral

$$\varphi_i = -\frac{\lambda}{\mu}\frac{1}{\mathrm{B}(\lambda-\mu,\mu)} \int_{x_i=-\infty}^{x_i=\infty} x_i\,p_x(x_i) \int_{q=q_l}^{q=q_u} p_Q(q|x_i) \int_{t=0}^{t=1-P_Q(q)} t^{\lambda-\mu-1}(1-t)^{\mu-1}\,\mathrm{d}t\,\mathrm{d}q\,\mathrm{d}x_i\,. \tag{55}$$

Now the integration order of t and q is exchanged. In Eq. (55) one has the bounds

$$q_l \le q \le q_u, \qquad 0 \le t \le 1 - P_Q(q). \tag{56}$$

Defining the inverse transformation $q = P_Q^{-1}(1-t)$ and integrating over t first, one obtains the new ranges

$$0 \le t \le 1, \qquad q_l \le q \le P_Q^{-1}(1-t). \tag{57}$$

The progress rate yields

$$\varphi_i = -\frac{\lambda}{\mu}\frac{1}{\mathrm{B}(\lambda-\mu,\mu)} \int_{x_i=-\infty}^{x_i=\infty} x_i\,p_x(x_i) \int_{t=0}^{t=1} t^{\lambda-\mu-1}(1-t)^{\mu-1} \int_{q=q_l}^{q=P_Q^{-1}(1-t)} p_Q(q|x_i)\,\mathrm{d}q\,\mathrm{d}t\,\mathrm{d}x_i\,. \tag{58}$$

Now the innermost integral can be solved using $p_Q(q|x_i) = \mathrm{d}P_Q(q|x_i)/\mathrm{d}q$

$$\int_{q_l}^{P_Q^{-1}(1-t)} p_Q(q|x_i)\,\mathrm{d}q = P_Q(P_Q^{-1}(1-t)|x_i) - P_Q(q_l|x_i) = P_Q(P_Q^{-1}(1-t)|x_i), \tag{59}$$

where the probability $P_Q(q_l|x_i) = \Pr(Q \le q_l|x_i) = 0$ for any lower bound value q_l. Inserting (59) into (58), we arrive at the progress rate integral (46). □

Unfortunately a closed-form solution of (46) after inserting Approximation 1 and Approximation 2 for the quality gain CDF is not possible due to the underlying structure of the integrand. Hence, asymptotic approximations will be introduced assuming large populations and large dimensionality to successively simplify the integral in a way that closed-form solutions can be provided. First, the large population theorem will be applied and then the quality gain CDF is inserted. Thereafter, the normal CDF is Taylor-expanded with the first two terms yielding analytically solvable results and higher order terms vanishing as $O(1/N)$. The results are further simplified in the end assuming component equipartition (12), which finally gives the progress rate result in (96).

Theorem 2. *Let p_x denote the PDF of the random mutation $x \sim \mathcal{N}(0,\sigma^2)$. Let P_Q denote the quality gain CDF with its quantile function given by P_Q^{-1}. For a truncation ratio $\vartheta = \mu/\lambda$ with $0 < \vartheta < 1$ the component-wise progress rate for large populations yields*

$$\lim_{\substack{\mu,\lambda\to\infty \\ \vartheta=\text{const.}}} \varphi_i = -\frac{1}{\vartheta}\int_{-\infty}^{\infty} x_i\,p_x(x_i)\,P_Q(P_Q^{-1}(\vartheta)|x_i)\,\mathrm{d}x_i\,. \tag{60}$$

Proof. Starting from Eq. (46) and applying the infinite population size limit, the result of Theorem 1 can be applied with $a = 1$, $p_n(x_i) = x_i$, and $f_x(t) = P_Q(P_Q^{-1}(1-t)|x_i)$. Evaluating $f_x(t)$ at $t = 1-\vartheta$ gives

$$f_x(t)\big|_{t=1-\vartheta} = P_Q(P_Q^{-1}(1-t)|x_i)\big|_{t=1-\vartheta} = P_Q(P_Q^{-1}(\vartheta)|x_i), \tag{61}$$

which yields the result (60). □

The next step requires the use of Approximation 1 and Approximation 2 for the quality gain distributions in Eq. (60). To this end, one uses the conditional normal distribution function $\Phi\left(\frac{q-E_{Q|x_i}}{D_i}\right)$, see (21), and the inverse transformation $q = E_Q + D_Q\,\Phi^{-1}(p)$ evaluated at $p = \vartheta$, see (18). One obtains

$$\widetilde{P}_Q\,(\widetilde{P}_Q^{-1}(\vartheta)|x_i) = \Phi\left(\frac{E_Q + D_Q\,\Phi^{-1}(\vartheta) - E_{Q|x_i}}{D_i}\right). \tag{62}$$

Given the normal approximation (62), an expression for $E_{Q|x_i}$ is needed. Using definition (19) with Q_i-result (24) the (conditional) expected value is written as

$$E_{Q|x_i} = Q_i(x_i) + \sum_{j\neq i} \mathrm{E}\left[Q_j\right] = k_i x_i + \delta_i(x_i) + E_i. \tag{63}$$

In (63) the following definitions are introduced as abbreviations

$$\begin{aligned} k_i &:= 2y_i \\ \delta_i(x_i) &:= x_i^2 + A\cos(\alpha y_i)(1-\cos(\alpha x_i)) + A\sin(\alpha y_i)\sin(\alpha x_i) \\ E_i &:= \sum_{j\neq i} \mathrm{E}\left[Q_i\right]. \end{aligned} \tag{64}$$

Given Eq. (63), quantity $\delta(x_i)$ includes all non-linear terms in x_i. This will be important when the normal CDF is expanded and analytically solved. Inserting relation (63) into (62) and the result into (60) yields

$$\varphi_i \simeq -\frac{1}{\vartheta}\int_{-\infty}^{\infty} x_i\,p_x(x_i)\,\Phi\left(\frac{E_Q + D_Q\,\Phi^{-1}(\vartheta) - (k_i x_i + \delta_i(x_i) + E_i)}{D_i}\right)\mathrm{d}x_i\,. \tag{65}$$

A closed-form solution of (65) cannot be obtained with $\Phi(\delta_i(x_i))$ containing non-linear terms in x_i. However, a solution in terms of a Taylor expansion can be provided by introducing the decomposition $\Phi(g(x_i)+h(x_i))$ with $g(x_i)$ being a linear function, and $h(x_i)$ being a small non-linear perturbation according to

$$g(x_i) := -\frac{k_i}{D_i}x_i + \frac{E_{Q_i} + D_Q\,\Phi^{-1}(\vartheta)}{D_i} \tag{66}$$

$$h(x_i) := -\frac{\delta(x_i)}{D_i}. \tag{67}$$

In (66), the abbreviation $E_{Q_i} = E_Q - E_i = \mathrm{E}[Q_i]$, cf. Eq. (10), is used to denote the expected value of the i-th summand of the quality gain (6). Using functions $g(x_i)$ and $h(x_i)$ Eq. (65) becomes

$$\varphi_i \simeq -\frac{1}{\vartheta}\int_{-\infty}^{\infty} x_i\,p_x(x_i)\,\Phi(g(x_i)+h(x_i))\,\mathrm{d}x_i\,. \tag{68}$$

Approximation 3 *(Truncated cumulative distribution function series).* Under the assumption of a normally distributed quality gain, see Approximation 1 and Approximation 2, and a quality gain variance scaling with N according to Eq. (13), the CDF of the normal distribution is expanded at $g(x_i)$ in the limit of $N\to\infty$ as

$$\varphi_i \simeq -\frac{1}{\vartheta}\int_{-\infty}^{\infty} x_i\,p_x(x_i)\left(\Phi(g(x_i)) + \phi(g(x_i))h(x_i) + O\left(\frac{1}{N}\right)\right)\mathrm{d}x_i\,. \tag{69}$$

Relation (69) is derived now. Starting from (68), the Taylor-expansion of $\Phi(\cdot)$ up to first order with the remainder denoted by r yields

$$\Phi(g+h) = \sum_{n=0}^{\infty}\frac{1}{n!}\frac{\mathrm{d}^n\Phi}{\mathrm{d}g^n}h^n = \Phi(g) + \phi(g)h + r(N). \tag{70}$$

Note that all derivatives of the normal distribution exist as $\frac{\mathrm{d}^n\phi(x)}{\mathrm{d}x^n} = (-1)^n\,\mathrm{He}_n(x)\,\phi(x)$ with $\mathrm{He}_n(x)$ denoting the n-th order probabilist's Hermite polynomials. In the following the scaling properties of the remainder as a function of N are investigated. It will be shown that $r = O(1/N)$. To this end, (70) is rewritten as

$$r(N) = \Phi(g+h) - \Phi(g) - \phi(g)h. \tag{71}$$

For the further analysis of $r(N)$ the equipartition of components is assumed as introduced in Eqs. (12), (13), and (14). Hence, the variance D_i can be written as a function of N as

$$D_i = s\sqrt{N}, \tag{72}$$

where the prefactor $s \neq s(N)$ depends on A, α, $\mathbf{y}$, and σ. With these assumptions the functions g and h are written as (using $E \coloneqq E_{Q_i}$, $\Phi_\vartheta^{-1} \coloneqq \Phi^{-1}(\vartheta)$, dropping the subscript i for brevity and using $D_i \simeq D_Q$)

$$g = \frac{E - kx}{s\sqrt{N}} + \Phi_\vartheta^{-1}, \quad h = -\frac{\delta}{s\sqrt{N}}. \tag{73}$$

As $h \to 0$ for $N \to \infty$, the remainder (71) vanishes accordingly. Therefore, in order to show $r(N) = O(1/N)$, $\lim_{N\to\infty} r(N)N$ is investigated applying l'Hôpital's rule

$$\lim_{N\to\infty} r(N)N = \lim_{N\to\infty} \frac{r(N)}{1/N} = \lim_{N\to\infty} \frac{\frac{\partial r(N)}{\partial N}}{\frac{\partial (1/N)}{\partial N}} = -\lim_{N\to\infty} N^2 \frac{\partial r(N)}{\partial N}. \tag{74}$$

To evaluate (74) the derivative of r from (71) w.r.t. N is evaluated as

$$\begin{aligned}\frac{\partial r}{\partial N} &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(g+h)^2} \left(\frac{\partial g}{\partial N} + \frac{\partial h}{\partial N}\right) - \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}g^2} \frac{\partial g}{\partial N} + \frac{gh}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}g^2} \frac{\partial g}{\partial N} - \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}g^2} \frac{\partial h}{\partial N} \\ &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}g^2} \left[\left(\mathrm{e}^{-gh - \frac{1}{2}h^2} - 1 \right) \left(\frac{\partial g}{\partial N} + \frac{\partial h}{\partial N} \right) + gh \frac{\partial g}{\partial N} \right]. \end{aligned} \tag{75}$$

The term $(\mathrm{e}^{-gh-\frac{1}{2}h^2} - 1)$ of (75) is expanded up to first order discarding higher orders $O((gh + \frac{1}{2}h^2)^2)$

$$\begin{aligned}\frac{\partial r}{\partial N} &\simeq \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}g^2} \left[\left(-gh - \frac{1}{2}h^2 \right) \left(\frac{\partial g}{\partial N} + \frac{\partial h}{\partial N} \right) + gh \frac{\partial g}{\partial N} \right] \\ &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}g^2} \left[-\frac{1}{2}h^2 \left(\frac{\partial g}{\partial N} + \frac{\partial h}{\partial N} \right) - gh \frac{\partial h}{\partial N} \right]. \end{aligned} \tag{76}$$

The derivatives of g and h from Eq. (73) are

$$\frac{\partial g}{\partial N} = -\frac{E - kx}{2sN^{3/2}}, \quad \frac{\partial h}{\partial N} = \frac{\delta}{2sN^{3/2}}. \tag{77}$$

Inserting (73) and (77) into (76) yields after refactoring

$$\begin{aligned}\frac{\partial r}{\partial N} &\simeq \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{E-kx}{s\sqrt{N}} + \Phi_\vartheta^{-1}\right)^2} \left[-\frac{\delta^2}{2s^2 N} \left(-\frac{E - kx}{2sN^{3/2}} + \frac{\delta}{2sN^{3/2}} \right) + \left(\frac{E - kx}{s\sqrt{N}} + \Phi_\vartheta^{-1} \right) \frac{\delta^2}{2s^2 N^2} \right] \\ &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{E-kx}{s\sqrt{N}} + \Phi_\vartheta^{-1}\right)^2} \left(-\frac{\delta^2}{2s^2 N^2} \right) \left[\frac{\delta}{2s\sqrt{N}} - \frac{3}{2} \frac{E - kx}{s\sqrt{N}} - \Phi_\vartheta^{-1} \right]. \end{aligned} \tag{78}$$

Taking the limit (74) of (78) therefore yields

$$\begin{aligned}\lim_{N\to\infty} r(N)N &= -\lim_{N\to\infty} N^2 \frac{\partial r(N)}{\partial N} = \lim_{N\to\infty} \left\{ \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{E-kx}{s\sqrt{N}} + \Phi_\vartheta^{-1}\right)^2} \frac{\delta^2}{2s^2} \left[\frac{\delta}{2s\sqrt{N}} - \frac{3}{2} \frac{E - kx}{s\sqrt{N}} - \Phi_\vartheta^{-1} \right] \right\} \\ &= -\frac{\delta^2 \Phi_\vartheta^{-1}}{2\sqrt{2\pi} s^2} \mathrm{e}^{-\frac{1}{2}\left(\Phi_\vartheta^{-1}\right)^2}, \end{aligned} \tag{79}$$

such that the remainder $r(N)$ can be given as

$$r(N) \simeq -\frac{\delta^2 \Phi_\vartheta^{-1}}{2\sqrt{2\pi} s^2} \mathrm{e}^{-\frac{1}{2}\left(\Phi_\vartheta^{-1}\right)^2} \frac{1}{N} = O\left(\frac{1}{N}\right), \tag{80}$$

which concludes the derivation of (69).

Both integrals of (69) are analytically solvable.[2] The zeroth order term yields a closed form solution due to $g(x_i)$ being linear w.r.t. x_i and gives progress contributions due to the sphere function, i.e., the linear part of the quality gain (63). The

[2] Actually, using the result from (80) one could even calculate a closed-form second-order approximation for (69). However, the resulting formula would be rather complex.

first order term can be solved by applying quadratic completion to the Gaussian product $p_x(x_i)\phi(g(x_i))$ yielding an expected value over a normal density. The expected value over $h(x_i)$ can be regarded as a perturbation of the sphere containing A and α dependencies.

The determination of φ_i via (69) was done in [18] by evaluating both integrals. As the derivation and the final result for φ_i are very lengthy and therefore not practical for further analytic treatment, the obtained expression for φ_i was simplified as a last step assuming large dimensionality N. However, the same result as in [18] can be obtained in a quicker way by simplifying the integrands of (69) under the same assumptions before the integration, instead of simplifying the result afterwards. This will enable a more concise derivation of the final progress rate result.

First the functions g and h from (66) and (67), respectively, are simplified. For large N, the quality gain variance $D_i \simeq D_Q$ using (14). As E_{Q_i} is just the quality gain expectation of a single component, it can be neglected compared to D_Q scaling as $\sqrt{N}$ using (13). Hence, one has

$$g(x_i) \simeq -\frac{k_i x_i}{D_Q} + \Phi^{-1}(\vartheta) \tag{81}$$

$$h(x_i) \simeq -\frac{\delta(x_i)}{D_Q}. \tag{82}$$

Another approximation is introduced regarding the density $p_x(x_i)\phi(g(x_i))$ for the second term of (69). By completing the square one can derive a resulting normal density with mean m and variance ς^2 by demanding

$$p_x(x_i)\phi(g(x_i)) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\frac{x_i^2}{\sigma^2}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}g(x_i)^2} \stackrel{!}{=} C\,\mathrm{e}^{-\frac{1}{2}\frac{(x_i-m)^2}{\varsigma^2}}. \tag{83}$$

Simple calculations yield

$$m = \Phi^{-1}(\vartheta)\frac{D_Q k_i \sigma^2}{D_Q^2 + k_i^2\sigma^2}, \quad \varsigma^2 = \frac{1}{1/\sigma^2 + (k_i^2/D_Q^2)}, \quad C = \frac{\mathrm{e}^{-\frac{1}{2}\left[\Phi^{-1}(\vartheta)\right]^2}}{2\pi\sigma}. \tag{84}$$

Noting that $D_Q^2 = \Theta(N)$ and neglecting contributions of single components for $N \to \infty$, i.e., $k_i^2 \ll D_Q^2$, $(k_i\sigma)^2 \ll D_Q^2$, the quantities m and ς^2 from (84) yield the asymptotic results

$$m \simeq 0, \quad \varsigma^2 \simeq \sigma^2, \tag{85}$$

such that the density of the first order term yields

$$p_x(x_i)\phi(g(x_i)) \simeq \frac{\mathrm{e}^{-\frac{1}{2}\left[\Phi^{-1}(\vartheta)\right]^2}}{\sqrt{2\pi}} p_x(x_i). \tag{86}$$

Using the results from Eqs. (81), (82), and (86), the progress rate integral (69) is further simplified. The prefactors of the resulting integral yield the asymptotic progress coefficient (45)

$$c_\vartheta := e_\vartheta^{1,0} = \frac{1}{\sqrt{2\pi}\vartheta} \mathrm{e}^{-\frac{1}{2}\left[\Phi^{-1}(\vartheta)\right]^2}. \tag{87}$$

Approximation 4 *(Progress rate integral for large dimensionality).* Based on the result of Approximation 3 only the first two terms are considered. Furthermore, the integrands of (69) are approximated and simplified assuming large dimensionality using Eqs. (81), (82), (86), and (87). Hence, one obtains

$$\varphi_i \simeq I_i^0 + I_i^1, \quad \text{with} \tag{88}$$

$$I_i^0 := -\frac{1}{\vartheta} \int_{-\infty}^{\infty} x_i p_x(x_i) \Phi\left(-\frac{k_i x_i}{D_Q} + \Phi^{-1}(\vartheta)\right) \mathrm{d}x_i\,, \quad \text{and} \tag{89}$$

$$I_i^1 := \frac{c_\vartheta}{D_Q} \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x_i \delta(x_i) p_x(x_i)\,\mathrm{d}x_i\,. \tag{90}$$

Calculating I_i^0 from (89) by inserting mutation density $p_x(x_i)$ from (48) and applying the substitution $z = x_i/\sigma$, one gets

$$I_i^0 = -\frac{\sigma}{\sqrt{2\pi}\vartheta} \int_{-\infty}^{\infty} z\mathrm{e}^{-\frac{1}{2}z^2} \Phi\left(-\frac{k_i\sigma}{D_Q} z + \Phi^{-1}(\vartheta)\right) \mathrm{d}z\,. \tag{91}$$

The following integral identity [5, Eq. (A.12)] can be applied

$$\int_{-\infty}^{\infty} t\mathrm{e}^{-\frac{1}{2}t^2}\Phi(at+b)\,\mathrm{d}t = \frac{a}{\sqrt{1+a^2}}\exp\left[-\frac{1}{2}\frac{b^2}{1+a^2}\right]. \tag{92}$$

Evaluating (92) with $a=-k_i\sigma/D_Q$ and $b=\Phi^{-1}(\vartheta)$ yields for the right-hand side of (92)

$$\frac{a}{\sqrt{1+a^2}}\exp\left[-\frac{1}{2}\frac{b^2}{1+a^2}\right] = -\frac{k_i\sigma}{D_Q}\frac{1}{\sqrt{1+(k_i\sigma/D_Q)^2}}\exp\left[-\frac{1}{2}\frac{\left[\Phi^{-1}(\vartheta)\right]^2}{1+(k_i\sigma/D_Q)^2}\right]. \tag{93}$$

Again assuming $(k_i\sigma)^2 \ll D_Q^2$, expression (93) simplifies and the result for (89) is obtained with (87) as

$$I_i^0 \simeq \frac{\mathrm{e}^{-\frac{1}{2}[\Phi^{-1}(\vartheta)]^2}}{\sqrt{2\pi}\vartheta}\frac{k_i\sigma^2}{D_Q} = c_\vartheta\frac{k_i\sigma^2}{D_Q}. \tag{94}$$

Now I_i^1 is solved. One notices that $x_i\delta(x_i) = x_i(x_i^2 + A\cos(\alpha y_i)(1-\cos(\alpha x_i)) + A\sin(\alpha y_i)\sin(\alpha x_i))$, see (64), is integrated over density p_x with zero mean. Therefore, all odd functions of x_i yield no contribution and only the term $x_i\sin(\alpha x_i)$ needs to be evaluated. One gets

$$\begin{aligned} I_i^1 &\simeq c_\vartheta\frac{A\sin(\alpha y_i)}{D_Q}\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} x_i\sin(\alpha x_i)\mathrm{e}^{-\frac{1}{2}\left(\frac{x_i}{\sigma}\right)^2}\,\mathrm{d}x_i \\ &= c_\vartheta\frac{A\sin(\alpha y_i)}{D_Q}\,\mathrm{E}\left[x_i\sin(\alpha x_i)\right] \\ &= c_\vartheta\frac{A\sin(\alpha y_i)}{D_Q}\alpha\sigma^2\mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2} \\ &= c_\vartheta\frac{d_i\sigma^2}{D_Q}\mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}. \end{aligned} \tag{95}$$

In the second line of (95) the expected value definition is used. From second to third line the expected value of $x_i\sin(\alpha x_i)$ is evaluated using (29). In the last line the derivative $d_i = \alpha A\sin(\alpha y_i)$ from (33) is recovered. Using the results from (94) and (95) the first order progress rate approximation for large N and μ can finally be given.

First order progress rate *The first order component-wise progress rate on the Rastrigin function in the asymptotic limits of infinitely large population size μ (constant $\vartheta=\mu/\lambda$) and infinitely large dimensionality N yields*

$$\varphi_i \simeq c_\vartheta\frac{\sigma^2}{D_Q}\left(k_i + \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}d_i\right) = c_\vartheta\frac{\sigma^2}{D_Q}\left(2y_i + \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}\alpha A\sin(\alpha y_i)\right). \tag{96}$$

The expressions for $c_\vartheta = e_\vartheta^{1,0}$ from (45) and D_Q from (31) were not inserted to improve readability. Result (96) shows very interesting properties compared to [17, Eq. (26)], where a linearized quality gain approximation resulted in

$$\varphi_{i,\mathrm{lin}} \simeq c_{\mu/\mu,\lambda}\frac{\sigma^2}{\sqrt{(f_i'\sigma)^2 + D_i^2}}f_i'. \tag{97}$$

First note that the progress coefficient was replaced by its asymptotic form $c_{\mu/\mu,\lambda} \simeq c_\vartheta$. The difference for the variance terms in the denominators of (96) and (97) is negligible for large N with $D_Q^2 \approx D_i^2 + (f_i'\sigma)^2$, see also (14). However, the most notable difference lies between the derivative term $f_i' = k_i + d_i$, see definition (33), and the newly obtained term $k_i + \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}d_i$. It contains an unchanged sphere-dependent term k_i and an exponentially decaying Rastrigin-specific term d_i. This characteristic form will be discussed in the subsequent part. The result (96) will be essential for the determination of the second order φ_i^{II}.

At this point one-generation experiments can be performed and compared to the progress rate (96) to investigate its accuracy. To this end, a random position vector $\mathbf{y}$ is initialized isotropically with $\|\mathbf{y}\| = R$ given some residual distance R. Then, repeated simulations are performed and quantity (7) is averaged over 10^6 trials. The issue with the choice of R is that the "interesting" region with high density of local minima scales with N, such that a relation $R(N)$ is needed. The following argumentation can be given. Assuming w.l.o.g. $\mathbf{y} > \mathbf{0}$ and that all components of the parental position are at some given local minimum denoted by $\hat{y}^{(j)}$. Index j identifies the local attractor along the half-axis, e.g. $j \in \{1,2,3\}$ in Fig. 1 on the right side. For $N=1$ one has $\mathbf{y} = [\hat{y}^{(j)}]$ and therefore $R^2 = (\hat{y}^{(j)})^2$. Having N components at the same j-th local minimum

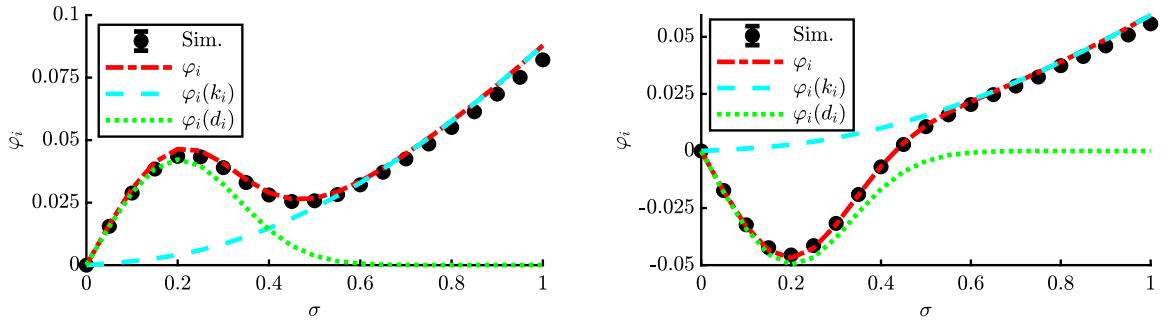

Fig. 3. One-generation experiments with $(10/10, 40)$-ES for $N = 20$, $A = 10$, $\alpha = 2\pi$ at randomly chosen $\|\mathbf{y}\| = R = \sqrt{N}$. The results for φ_i of Eq. (96) are shown for the exemplary components $i = 2$ with $y_i = 1.16$ (left) and $i = 12$ with $y_i = 0.78$ (right) to illustrate the effect of local attraction on the progress rate. The plots show additionally Eq. (96) with $\varphi_i(k_i) = \varphi_i(d_i, k_i)|_{d_i=0}$ [cyan, dashed] and $\varphi_i(d_i) = \varphi_i(d_i, k_i)|_{k_i=0}$ [green, dotted], respectively.

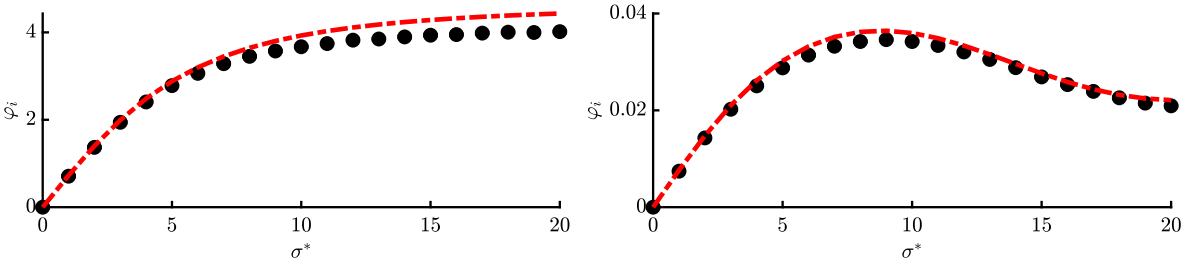

Fig. 4. Progress rate φ_i as a function of the normalized mutation σ^* for $(10/10, 40)$-ES with $N = 20$, $A = 1$, $\alpha = 2\pi$, at two residual distances $R = 10\sqrt{N}$ with $y_i = 11.6$ (left) and $R = 0.1\sqrt{N}$ with $y_i = 0.116$ (right). As in Fig. 3, black dots depict the simulation, while the red dash-dotted line shows result (96). The error bars are very small and therefore not visible.

yields $\mathbf{y} = [\hat{y}^{(j)}, \hat{y}^{(j)}, ..., \hat{y}^{(j)}]$, such that $R^2 = N(\hat{y}^{(j)})^2$. A scaling $R = O(\sqrt{N})$ is therefore needed to stay within a certain region of local attractors when N is increased.

The progress rates of two exemplary components for a single experiment are shown in Fig. 3. For both plots $\sigma \in [0, 1]$ was chosen in order to investigate the effects of the oscillation as $\alpha = 2\pi$. On the left, one observes enhanced progress for moderate σ-values due to local attraction, as both local and global attractor are aligned along the same direction. On the right, there is negative progress for moderate σ, as the local attractor is driving the ES away from the global attractor. For larger σ, the overall spherical shape is dominating and both exhibit positive progress. A decomposition of the progress rate in terms of $\varphi_i = \varphi_i(d_i, k_i)|_{k_i=0} + \varphi_i(d_i, k_i)|_{d_i=0}$ is displayed in Fig. 3. It shows the large-scale behavior of the k_i-term, dashed cyan, and limited range of the d_i-term, dotted green. As $k_i = \partial(y_i^2)/\partial y_i$, its progress term models the global quadratic structure of Rastrigin, see derivative definitions (33). The second term $\mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2} d_i$ models the Rastrigin-specific local oscillation having limited range depending on the mutation strength σ (or α). By defining scale-invariant mutations using (4) with $\sigma = \sigma^* R/N$, the oscillations vanish via $\mathrm{e}^{-\frac{1}{2}(\alpha\sigma^* R/N)^2}$ for large residual distance R, where the sphere function is recovered. This model significantly improves the progress rate formula (97) from [17].

As a note, changing one of the fitness parameters A or α directly affects Fig. 3. The change of amplitude A rescales both the (local) peak and dip heights accordingly, increasing the effects of local attraction for larger A. Increasing frequency α has mostly short-range effects as the overall range is reduced due to suppression via $\mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}$ of (96). In the subsequent parts, the progress rate is investigated for $A = 1$ and $\alpha = 2\pi$ as an example.

In Figs. 4 and 5 the progress rate is evaluated over scale-invariant σ^* for two different N-values and population sizes. One can see that the approximation quality improves for larger N and μ, as expected from the applied approximations. The overall agreement between simulation and approximation is good for larger and smaller residual distances R, see left and right plots, respectively. The σ^*-range was chosen large enough, such that the progress rate of the corresponding sphere function [5, Eq. (6.54)] reaches negative values due to mutations being too large. This boundary directly translates to Rastrigin, as the global structure is the same. However, due to φ_i being first order, no negative progress occurs even for large σ^*. Therefore the second order progress rate φ_i^{II} needs to be derived in Sec. 4, where loss terms will provide additional correction terms.

4. Second order progress rate

The second order progress rate (8) requires the evaluation of $\mathrm{E}\left[(y_i^{(g+1)})^2\right]$. Starting with intermediate result (34) and referring to the i-th component, the expression yields after squaring

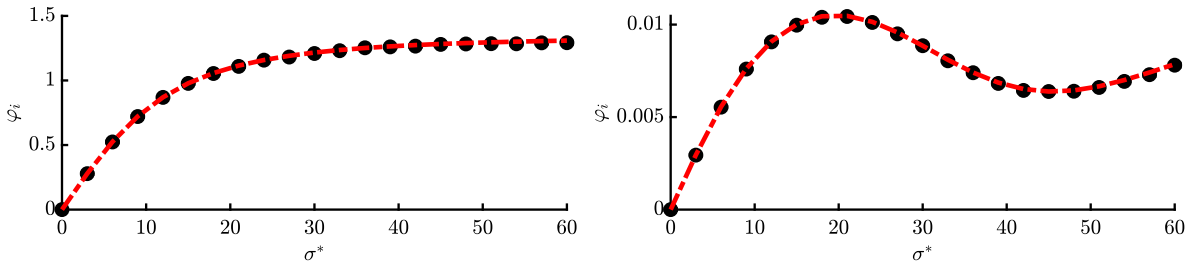

Fig. 5. Progress rate φ_i as a function of the normalized mutation σ^* for $(100/100, 200)$-ES with $N = 100$, $A = 1$, $\alpha = 2\pi$, at two residual distances $R = 10\sqrt{N}$ with $y_i = 11.9$ (left) and $R = 0.1\sqrt{N}$ with $y_i = 0.119$ (right). The approximation quality improves compared to Fig. 4 and shows very good agreement.

$$\begin{aligned}\left(y_i^{(g+1)}\right)^2 &= \left(y_i^{(g)} + \frac{1}{\mu}\sum_{m=1}^{\mu} x_{m;\lambda}\right)^2 \\ &= \left(y_i^{(g)}\right)^2 + 2y_i^{(g)}\frac{1}{\mu}\sum_{m=1}^{\mu} x_{m;\lambda} + \frac{1}{\mu^2}\left(\sum_{m=1}^{\mu} x_{m;\lambda}\right)^2.\end{aligned} \tag{98}$$

Squaring the last term can be evaluated by separating the sum into equal and unequal indices

$$\begin{aligned}\left(\sum_{m=1}^{\mu} x_{m;\lambda}\right)^2 &= \left(\sum_{k=1}^{\mu} x_{k;\lambda}\right)\left(\sum_{l=1}^{\mu} x_{l;\lambda}\right) = \sum_{m=1}^{\mu}\left(x_{m;\lambda}\right)^2 + \sum_{k\neq l} x_{k;\lambda}x_{l;\lambda} \\ &= \sum_{m=1}^{\mu}\left(x_{m;\lambda}\right)^2 + 2\sum_{l=2}^{\mu}\sum_{k=1}^{l-1} x_{k;\lambda}x_{l;\lambda}.\end{aligned} \tag{99}$$

Inserting (99) into (98) and taking the expected value (conditional variables $\mathbf{y}^{(g)}$ and $\sigma^{(g)}$ are implicitly assumed to be given) yields

$$\mathrm{E}\left[\left(y_i^{(g+1)}\right)^2\right] = \left(y_i^{(g)}\right)^2 + 2y_i^{(g)}\frac{1}{\mu}\sum_{m=1}^{\mu}\mathrm{E}\left[x_{m;\lambda}\right] + \frac{1}{\mu^2}\sum_{m=1}^{\mu}\mathrm{E}\left[\left(x_{m;\lambda}\right)^2\right] + \frac{2}{\mu^2}\sum_{l=2}^{\mu}\sum_{k=1}^{l-1}\mathrm{E}\left[x_{k;\lambda}x_{l;\lambda}\right]. \tag{100}$$

Noting that $\varphi_i = -\frac{1}{\mu}\sum_{m=1}^{\mu}\mathrm{E}\left[x_{m;\lambda}\right]$, see Eq. (36), and using (100) in φ_i^{II}-definition (8) yields the second order i-th component progress rate

$$\varphi_i^{\mathrm{II}} = 2y_i^{(g)}\varphi_i - \frac{1}{\mu^2}E^{(2)} - \frac{2}{\mu^2}E^{(1,1)}, \tag{101}$$

for which the two following expected values need to be determined

$$\frac{1}{\mu^2}E^{(2)} \coloneqq \frac{1}{\mu^2}\sum_{m=1}^{\mu}\mathrm{E}\left[\left(x_{m;\lambda}\right)^2\right] \tag{102}$$

$$\frac{1}{\mu^2}E^{(1,1)} \coloneqq \frac{1}{\mu^2}\sum_{l=2}^{\mu}\sum_{k=1}^{l-1}\mathrm{E}\left[x_{k;\lambda}x_{l;\lambda}\right]. \tag{103}$$

In the subsequent parts the solutions to Eqs. (102) and (103) will be derived. Starting with (102), the solution requires order statistic density (50) for the m-th individual, large population identity (37), and the expansion of the normal CDF (69) up to first order. The resulting two integrals can then be solved analytically for large N and the results will simplify significantly.

Proposition 2. *Let $\mu, \lambda \in \mathbb{N}$ with $\mu \geq 1$ and $\mu < \lambda$ and let p_x denote the PDF of the random mutation $x \sim \mathcal{N}(0, \sigma^2)$. Let $x_{m;\lambda}$ denote the m-th best value (out of λ) of the i-th mutation component $(\mathbf{x}_{m;\lambda})_i$. Furthermore, let P_Q and P_Q^{-1} denote the quality gain CDF (and its inverse), respectively, with* B *denoting the beta function. Then, the second order expected value reads*

$$\frac{1}{\mu}\sum_{m=1}^{\mu}\mathrm{E}\left[\left(x_{m;\lambda}\right)^2\right] = \frac{\lambda}{\mu}\int_{x_i=-\infty}^{x_i=\infty} x_i^2\, p_x(x_i)\frac{1}{\mathrm{B}(\lambda-\mu,\mu)}\int_{t=0}^{t=1} t^{\lambda-\mu-1}(1-t)^{\mu-1}P_Q\left(P_Q^{-1}(1-t)|x_i\right)\mathrm{d}t\,\mathrm{d}x_i\,. \tag{104}$$

Proof. Starting from (102) and rewriting the expected value as an integral over order statistic density $p_{m;\lambda}(x_i)$ yields

$$\frac{1}{\mu}\sum_{m=1}^{\mu}\mathrm{E}\left[\left(x_{m;\lambda}\right)^2\right]=\frac{1}{\mu}\sum_{m=1}^{\mu}\int_{-\infty}^{\infty}x_i^2\,p_{m;\lambda}(x_i)\,\mathrm{d}x_i\,. \tag{105}$$

Both (47) and (105) have the same structure after inserting $p_{m;\lambda}(x_i)$ from (50) and the integration over the squared mutation component is performed as the last step. The same steps as presented in the proof of Proposition 1 can therefore be applied with squared quantity x_i^2, which directly gives the result (104). □

Analogously to the derivation of the first order progress rate in Sec. 3, a closed-form solution for (104) can only be provided by first applying the limit of large populations and then introducing approximations assuming large dimensionality N.

Theorem 3. *Let p_x denote the PDF of the random mutation $x\sim\mathcal{N}(0,\sigma^2)$ and let $x_{m;\lambda}$ denote the m-th best value (out of λ) of the i-th mutation component $(\mathbf{x}_{m;\lambda})_i$. Let P_Q denote the quality gain CDF with its quantile function given by P_Q^{-1}. For a truncation ratio ϑ the limit of the second order expected value reads*

$$\lim_{\substack{\mu,\lambda\to\infty\\ \vartheta=\text{const.}}}\frac{1}{\mu}\sum_{m=1}^{\mu}\mathrm{E}\left[\left(x_{m;\lambda}\right)^2\right]=\frac{1}{\vartheta}\int_{-\infty}^{\infty}x_i^2\,p_x(x_i)\,P_Q\left(P_Q^{-1}(\vartheta)|x_i\right)\mathrm{d}x_i\,. \tag{106}$$

Proof. Starting from Eq. (104) and applying the infinite population size limit, the result of Theorem 1 can be applied with $a=1$, $p_n(x_i)=x_i^2$, and $f_x(t)\big|_{t=1-\vartheta}=P_Q\left(P_Q^{-1}(\vartheta)|x_i\right)$, which yields the result (106). □

Given result (106), approximations are again applied to provide closed-form solutions. Inserting quality gain Approximation 1 and Approximation 2 via Eq. (62) into (106) leads (again) to an analytically not solvable integral due to non-linear terms in x_i within $\Phi(\cdot)$. Therefore, the CDF is expanded using Approximation 3 neglecting higher order terms $O(1/N)$. Finally, the integrands are simplified assuming large dimensionality using Approximation 4. The result is therefore given after inserting $g(x_i)$ and $h(x_i)$ from (81) and (82) as

$$\frac{1}{\mu^2}E^{(2)}\simeq I_i^0+I_i^1,\quad\text{with} \tag{107}$$

$$I_i^0:=\frac{1}{\mu\vartheta}\int_{-\infty}^{\infty}x_i^2\,p_x(x_i)\Phi\left(-\frac{k_i x_i}{D_Q}+\Phi^{-1}(\vartheta)\right)\mathrm{d}x_i\,,\quad\text{and} \tag{108}$$

$$I_i^1:=-\frac{c_\vartheta}{\mu D_Q}\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}x_i^2\,\delta(x_i)\,p_x(x_i)\,\mathrm{d}x_i\,. \tag{109}$$

The two integrals abbreviated as I_i^0 and I_i^1 are evaluated now. For I_i^0, the substitution $z=x_i/\sigma$ is introduced

$$I_i^0=\frac{\sigma^2}{\sqrt{2\pi}\mu\vartheta}\int_{-\infty}^{\infty}z^2\mathrm{e}^{-\frac{1}{2}z^2}\Phi\left(-\frac{k_i\sigma z}{D_Q}+\Phi^{-1}(\vartheta)\right)\mathrm{d}z\,. \tag{110}$$

The following integral identity [16] is applied for real parameters a and b

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}t^2\mathrm{e}^{-\frac{1}{2}t^2}\Phi(at+b)\,\mathrm{d}t=\Phi\left(\frac{b}{(1+a^2)^{1/2}}\right)-\frac{1}{\sqrt{2\pi}}\frac{a^2b}{(1+a^2)^{3/2}}\mathrm{e}^{-\frac{1}{2}\frac{b^2}{1+a^2}}. \tag{111}$$

Evaluating (111) with $a=-k_i\sigma/D_Q$, $b=\Phi^{-1}(\vartheta)$ from (108) yields for the right-hand side of (111)

$$\begin{aligned}&\Phi\left(\frac{b}{(1+a^2)^{1/2}}\right)-\frac{1}{\sqrt{2\pi}}\frac{a^2b}{(1+a^2)^{3/2}}\mathrm{e}^{-\frac{1}{2}\frac{b^2}{1+a^2}}\\&=\Phi\left(\frac{\Phi^{-1}(\vartheta)}{(1+(k_i\sigma)^2/D_Q^2)^{1/2}}\right)-\frac{1}{\sqrt{2\pi}}\frac{(k_i\sigma)^2\Phi^{-1}(\vartheta)}{D_Q^2(1+(k_i\sigma)^2/D_Q^2)^{3/2}}\mathrm{e}^{-\frac{1}{2}\frac{\left[\Phi^{-1}(\vartheta)\right]^2}{1+(k_i\sigma)^2/D_Q^2}}.\end{aligned} \tag{112}$$

Assuming $(k_i\sigma)^2 \ll D_Q^2$ for large N further simplifies (112) and one obtains the result

$$I_i^0 \simeq \frac{\sigma^2}{\mu}\left[1 - \Phi^{-1}(\vartheta)\left[\frac{\mathrm{e}^{-\frac{1}{2}\left[\Phi^{-1}(\vartheta)\right]^2}}{\sqrt{2\pi}\vartheta}\right]\frac{(k_i\sigma)^2}{D_Q^2}\right]. \tag{113}$$

For (113) the asymptotic generalized progress coefficient definition $e_\vartheta^{1,1}$ from (45) can be applied with parameters $a = 1$ and $b = 1$

$$e_\vartheta^{1,1} = -\Phi^{-1}(\vartheta)\left[\frac{\mathrm{e}^{-\frac{1}{2}\left[\Phi^{-1}(\vartheta)\right]^2}}{\sqrt{2\pi}\vartheta}\right]. \tag{114}$$

This leads to following result for the first integral I_i^0

$$I_i^0 \simeq \frac{\sigma^2}{\mu}\left[1 + e_\vartheta^{1,1}\frac{(k_i\sigma)^2}{D_Q^2}\right]. \tag{115}$$

Second integral I_i^1 from (109) is expressed using expected values over the normal density p_x of the terms given by $x_i^2\delta(x_i)$. With $\delta(x_i)$ given in Eq. (64) one gets

$$I_i^1 \simeq -\frac{c_\vartheta}{\mu D_Q}\left(\mathrm{E}\left[x_i^4\right] + A\sin(\alpha y_i)\,\mathrm{E}\left[x_i^2\sin(\alpha x_i)\right] + A\cos(\alpha y_i)\,\mathrm{E}\left[x_i^2\right] - A\cos(\alpha y_i)\,\mathrm{E}\left[x_i^2\cos(\alpha x_i)\right]\right). \tag{116}$$

One has $\mathrm{E}\left[x_i^4\right] = 3\sigma^4$ and $\mathrm{E}\left[x_i^2\right] = \sigma^2$. Using results from (29) the remaining expected values read

$$\mathrm{E}\left[x_i^2\sin(\alpha x_i)\right] = 0, \quad \mathrm{E}\left[x_i^2\cos(\alpha x_i)\right] = (\sigma^2 - \alpha^2\sigma^4)\mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}. \tag{117}$$

Therefore, one gets

$$I_i^1 \simeq -\frac{c_\vartheta\sigma^2}{\mu D_Q}\left[3\sigma^2 + A\cos(\alpha y_i)\left(1 - \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2} + \alpha^2\sigma^2\mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}\right)\right]. \tag{118}$$

Collecting the results (115) and (118) with $k_i = 2y_i$ and inserting them back into (107) the expected value finally reads

$$\frac{1}{\mu^2}E^{(2)} \simeq \frac{\sigma^2}{\mu}\left\{1 + e_\vartheta^{1,1}\frac{(2y_i)^2\sigma^2}{D_Q^2} - \frac{c_\vartheta}{D_Q}\left[3\sigma^2 + A\cos(\alpha y_i)\left(1 - \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2} + \alpha^2\sigma^2\mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}\right)\right]\right\}. \tag{119}$$

The solution of the second expected value $\frac{1}{\mu^2}E^{(1,1)}$ from (103) is presented now. First an exact integral is derived. Then, approximations are applied to give closed-form solutions.

Proposition 3. *Let $\mu, \lambda \in \mathbb{N}$ with $\mu \geq 1$ and $\mu < \lambda$ and let p_x denote the PDF of the random mutation $x \sim \mathcal{N}(0, \sigma^2)$. Let $x_{k;\lambda}$ denote the k-th best value (out of λ) of the i-th mutation component $(\mathbf{x}_{k;\lambda})_i$. Furthermore, let P_Q and P_Q^{-1} denote the quality gain CDF (and its inverse), respectively, with* B *denoting the beta function. Then, the second order expected value reads*

$$\begin{aligned}\frac{1}{\mu^2}\sum_{l=2}^{\mu}\sum_{k=1}^{l-1}\mathrm{E}\left[x_{k;\lambda}x_{l;\lambda}\right] = \frac{1}{2}\frac{\lambda}{\mu}\frac{\mu-1}{\mu}\int_{-\infty}^{\infty}x_1 p_x(x_1)\int_{-\infty}^{\infty}x_2 p_x(x_2)\\ \times\left(\frac{1}{\mathrm{B}(\lambda-\mu,\mu)}\int_0^1 t^{\lambda-\mu-1}(1-t)^{\mu-2}P_Q\left(P_Q^{-1}(1-t)|x_1\right)P_Q\left(P_Q^{-1}(1-t)|x_2\right)\mathrm{d}t\right)\mathrm{d}x_2\,\mathrm{d}x_1\,.\end{aligned} \tag{120}$$

Proof. First, a joint order statistic density has to be derived for the expected value. Then, the double sum is converted into a single integral using a known identity. The resulting five-fold integration is restructured by exchanging bounds and then successively solved.

Starting with (103), the double sum includes mixed contributions from the k-th and l-th best elements of the i-th mutation component. To avoid confusion with the summation indices k and l, the integration variables associated with k-th element will be denoted as x_1 (mutation) and q_1 (quality), while the l-th element is integrated over x_2 and q_2. The ordering $1 \leq k < l \leq \lambda$ is assumed with k yielding a smaller (better) quality value $q_1 < q_2$. Additionally, the joint probability density $p_{k,l;\lambda}(x_1, x_2)$ is needed, such that the expected value can be formulated as

$$\frac{1}{\mu^2}E^{(1,1)} = \frac{1}{\mu^2}\sum_{l=2}^{\mu}\sum_{k=1}^{l-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2\, p_{k,l;\lambda}(x_1, x_2)\,\mathrm{d}x_2\,\mathrm{d}x_1\,. \tag{121}$$

The mutation densities are independent and denoted by $p_x(x_1)$ and $p_x(x_2)$, respectively. Given mutation components x_1 and x_2, the conditional density obtaining the quality values q_1 and q_2 is $p_Q(q_1|x_1)$ and $p_Q(q_2|x_2)$, respectively. Given q_1 and q_2, one has $k-1$ values smaller than q_1, $l-k-1$ values between q_1 and q_2 and $\lambda - l$ values larger than q_2 with probabilities

$$\begin{aligned}\Pr\{Q \le q_1\}^{k-1} &= P_Q(q_1)^{k-1}\\ \Pr\{q_1 \le Q \le q_2\}^{l-k-1} &= [P_Q(q_2) - P_Q(q_1)]^{l-k-1}\\ \Pr\{Q > q_2\}^{\lambda-l} &= [1 - P_Q(q_2)]^{\lambda-l},\end{aligned} \tag{122}$$

and $P_Q(q)$ denoting the quality gain CDF. The joint probability density can therefore be written as

$$\begin{aligned}p_{k,l;\lambda}(x_1, x_2) = p_x(x_1)p_x(x_2)\int_{q_{\min}}^{\infty} p_Q(q_1|x_1)\int_{q_1}^{\infty} p_Q(q_2|x_2)\\ \times \lambda!\frac{P_Q(q_1)^{k-1}[P_Q(q_2) - P_Q(q_1)]^{l-k-1}[1 - P_Q(q_2)]^{\lambda-l}}{(k-1)!(l-k-1)!(\lambda-l)!}\,\mathrm{d}q_2\,\mathrm{d}q_1\,,\end{aligned} \tag{123}$$

with integration ranges $q_{\min} \le q_1 < \infty$ and $q_1 < q_2 < \infty$ as $k < l$. Lower bound $q_{\min}$ denotes the smallest possible quality value, which is resolved later. The factorials exclude the irrelevant combinations among the three groups given in (122). Plugging (123) into (121) and moving the sum into the innermost integral gives

$$\begin{aligned}\frac{1}{\mu^2}E^{(1,1)} = \frac{\lambda!}{\mu^2}\int_{-\infty}^{\infty} x_1 p_x(x_1)\int_{-\infty}^{\infty} x_2 p_x(x_2)\int_{q_{\min}}^{\infty} p_Q(q_1|x_1)\int_{q_1}^{\infty} p_Q(q_2|x_2)\\ \times\sum_{l=2}^{\mu}\sum_{k=1}^{l-1}\frac{P_Q(q_1)^{k-1}[P_Q(q_2) - P_Q(q_1)]^{l-k-1}[1 - P_Q(q_2)]^{\lambda-l}}{(k-1)!(l-k-1)!(\lambda-l)!}\,\mathrm{d}q_2\,\mathrm{d}q_1\,\mathrm{d}x_2\,\mathrm{d}x_1\,.\end{aligned} \tag{124}$$

The double sum of (124) over the P_Q-values will be expressed by an integral. This can be done using an identity from [4, p. 113]. Setting $\nu = 2$ and identifying the indices as $i_1 = l$ and $i_2 = k$, the identity yields

$$\sum_{l=2}^{\mu}\sum_{k=1}^{l-1}\frac{Q_1^{\lambda-l}[Q_2 - Q_1]^{l-k-1}[1 - Q_2]^{k-1}}{(\lambda-l)!(l-k-1)!(k-1)!} = \frac{1}{(\lambda-\mu-1)!(\mu-2)!}\int_0^{Q_1} t^{\lambda-\mu-1}(1-t)^{\mu-2}\,\mathrm{d}t\,, \tag{125}$$

for real values Q_1 and Q_2, with integers $\nu \le \mu < \lambda$. Now the substitution $Q_1 = 1 - P_Q(q_2)$, $Q_2 = 1 - P_Q(q_1)$ can be performed and the double sum of (124) can be recognized by comparing with (125). Applying the identity therefore yields

$$\begin{aligned}&\sum_{l=2}^{\mu}\sum_{k=1}^{l-1}\frac{[1 - P_Q(q_2)]^{\lambda-l}[P_Q(q_2) - P_Q(q_1)]^{l-k-1}[P_Q(q_1)]^{k-1}}{(\lambda-l)!(l-k-1)!(k-1)!}\\ &\quad= \frac{1}{(\lambda-\mu-1)!(\mu-2)!}\int_0^{1-P_Q(q_2)} t^{\lambda-\mu-1}(1-t)^{\mu-2}\,\mathrm{d}t\,.\end{aligned} \tag{126}$$

Hence, Eq. (124) is expressed as

$$\begin{aligned}\frac{1}{\mu^2}E^{(1,1)} = \frac{\lambda!}{\mu^2}\frac{1}{(\lambda-\mu-1)!(\mu-2)!}\int_{-\infty}^{\infty} x_1 p_x(x_1)\int_{-\infty}^{\infty} x_2 p_x(x_2)\\ \times\int_{q_{\min}}^{\infty} p_Q(q_1|x_1)\int_{q_1}^{\infty} p_Q(q_2|x_2)\int_0^{1-P_Q(q_2)} t^{\lambda-\mu-1}(1-t)^{\mu-2}\,\mathrm{d}t\,\mathrm{d}q_2\,\mathrm{d}q_1\,\mathrm{d}x_2\,\mathrm{d}x_1\,.\end{aligned} \tag{127}$$

The prefactor of Eq. (127) can be evaluated as

$$\frac{\lambda!}{\mu^2}\frac{1}{(\lambda-\mu-1)!(\mu-2)!}=\frac{\lambda(\lambda-1)!(\mu-1)}{\mu^2(\lambda-\mu-1)!(\mu-1)!}=\frac{1}{\vartheta}\frac{\mu-1}{\mu}\frac{1}{\mathrm{B}(\lambda-\mu,\mu)}. \tag{128}$$

Now the integration order will be exchanged twice in (127). First the order between t and q_2 is exchanged. Then the order between t and q_1 is exchanged, such that both q-integrations are performed before the t-integration enabling the application of the large population identity (37). Starting with integration bounds

$$q_1 \le q_2 < \infty, \quad 0 \le t \le 1 - P_Q(q_2), \tag{129}$$

and using the inverse function P_Q^{-1} with $q_2 = P_Q^{-1}(1-t)$ the exchanged bounds between t and q_2 are

$$0 \le t \le 1 - P_Q(q_1), \quad q_1 \le q_2 \le P_Q^{-1}(1-t). \tag{130}$$

Using factor (128) and exchanged bounds (130), the expression (127) is reformulated as

$$\begin{aligned}\frac{1}{\mu^2}E^{(1,1)} = \frac{1}{\vartheta}\frac{\mu-1}{\mu}\frac{1}{\mathrm{B}(\lambda-\mu,\mu)}\int_{-\infty}^{\infty} x_1 p_x(x_1)\int_{-\infty}^{\infty} x_2 p_x(x_2)\\ \times\int_{q_{\min}}^{\infty} p_Q(q_1|x_1)\int_0^{1-P_Q(q_1)} t^{\lambda-\mu-1}(1-t)^{\mu-2}\int_{q_1}^{P_Q^{-1}(1-t)} p_Q(q_2|x_2)\,\mathrm{d}q_2\,\mathrm{d}t\,\mathrm{d}q_1\,\mathrm{d}x_2\,\mathrm{d}x_1\,.\end{aligned} \tag{131}$$

Now the integration order between t and q_1 is exchanged starting from

$$q_{\min} \le q_1 < \infty, \quad 0 \le t \le 1 - P_Q(q_1), \tag{132}$$

yielding exchanged bounds

$$0 \le t \le 1, \quad q_{\min} \le q_1 \le P_Q^{-1}(1-t). \tag{133}$$

Therefore, one arrives at the following integral to be solved (beta function has been moved inside as it will be evaluated during the t-integration)

$$\begin{aligned}\frac{1}{\mu^2}E^{(1,1)} = \frac{1}{\vartheta}\frac{\mu-1}{\mu}\int_{-\infty}^{\infty} x_1 p_x(x_1)\int_{-\infty}^{\infty} x_2 p_x(x_2)\\ \times\left(\frac{1}{\mathrm{B}(\lambda-\mu,\mu)}\int_0^1 t^{\lambda-\mu-1}(1-t)^{\mu-2}\right.\\ \left.\times\left[\int_{q_{\min}}^{P_Q^{-1}(1-t)} p_Q(q_1|x_1)\left\{\int_{q_1}^{P_Q^{-1}(1-t)} p_Q(q_2|x_2)\,\mathrm{d}q_2\right\}\mathrm{d}q_1\right]\mathrm{d}t\right)\mathrm{d}x_2\,\mathrm{d}x_1\,.\end{aligned} \tag{134}$$

Now the integrals in (134) will be successively solved. Starting with integral $\{\cdot\}$ over q_2 one has

$$\int_{q_1}^{P_Q^{-1}(1-t)} p_Q(q_2|x_2)\,\mathrm{d}q_2 = \Big[P_Q(q_2|x_2)\Big]_{q_1}^{P_Q^{-1}(1-t)} = P_Q(P_Q^{-1}(1-t)|x_2) - P_Q(q_1|x_2). \tag{135}$$

The q_1-integration within $[\cdot]$ using (135) yields

$$\int_{q_{\min}}^{P_Q^{-1}(1-t)} p_Q(q_1|x_1)\Big(P_Q(P_Q^{-1}(1-t)|x_2) - P_Q(q_1|x_2)\Big)\,\mathrm{d}q_1 \tag{136}$$

$$= P_Q(P_Q^{-1}(1-t)|x_2)\int_{q_{\min}}^{P_Q^{-1}(1-t)} p_Q(q_1|x_1)\,\mathrm{d}q_1 \tag{137}$$

$$-\int_{q_{\min}}^{P_Q^{-1}(1-t)} p_Q(q_1|x_1)P_Q(q_1|x_2)\,\mathrm{d}q_1\,. \tag{138}$$

First integral (137) is easily evaluated, as the conditional density is integrated over its support giving

$$P_Q(P_Q^{-1}(1-t)|x_2)\int_{q_{\min}}^{P_Q^{-1}(1-t)} p_Q(q_1|x_1)\,\mathrm{d}q_1 = P_Q(P_Q^{-1}(1-t)|x_2)\Big[P_Q(q_1|x_1)\Big]_{q_{\min}}^{P_Q^{-1}(1-t)} \\ = P_Q(P_Q^{-1}(1-t)|x_2)P_Q(P_Q^{-1}(1-t)|x_1), \tag{139}$$

with $P_Q(q_{\min}|x_1) = \Pr\{Q \le q_{\min}|x_1\} = 0$. Note that the resulting factors are equal up to the conditional variables x_1 and x_2.

The second integral (138) will be simplified using integration by parts. Thereafter, one can exchange the x_1 and x_2 variables to find a simpler expression for the original integral. Integration by parts yields

$$\int_{q_{\min}}^{P_Q^{-1}(1-t)} p_Q(q_1|x_1)P_Q(q_1|x_2)\,\mathrm{d}q_1 \\ = P_Q(P_Q^{-1}(1-t)|x_1)P_Q(P_Q^{-1}(1-t)|x_2) - \int_{q_{\min}}^{P_Q^{-1}(1-t)} P_Q(q_1|x_1)p_Q(q_1|x_2)\,\mathrm{d}q_1\,. \tag{140}$$

Equation (140) inserted into (134) has to be integrated over x_1 and x_2, of which the order can be exchanged. For the following step the t-integration and the prefactors of (134) have no influence, such that they are dropped for better readability. Integrating both sides of (140) yields

$$\int_{-\infty}^{\infty} x_1 p_x(x_1) \int_{-\infty}^{\infty} x_2 p_x(x_2) \int_{q_{\min}}^{P_Q^{-1}(1-t)} p_Q(q_1|x_1)P_Q(q_1|x_2)\,\mathrm{d}q_1\,\mathrm{d}x_2\,\mathrm{d}x_1 \\ = \int_{-\infty}^{\infty} x_1 p_x(x_1) \int_{-\infty}^{\infty} x_2 p_x(x_2) P_Q(P_Q^{-1}(1-t)|x_1)P_Q(P_Q^{-1}(1-t)|x_2)\,\mathrm{d}x_2\,\mathrm{d}x_1 \\ - \int_{-\infty}^{\infty} x_2 p_x(x_2) \int_{-\infty}^{\infty} x_1 p_x(x_1) \int_{q_{\min}}^{P_Q^{-1}(1-t)} P_Q(q_1|x_2)p_Q(q_1|x_1)\,\mathrm{d}q_1\,\mathrm{d}x_1\,\mathrm{d}x_2\,, \tag{141}$$

where in the last line the integration order of x_1 and x_2 was exchanged, such that an expression equivalent to the left-hand side of (141) is obtained with given arguments for p_Q and P_Q. Collecting the terms, Eq. (141) can be formulated as

$$\int_{-\infty}^{\infty} x_1 p_x(x_1) \int_{-\infty}^{\infty} x_2 p_x(x_2) \int_{q_{\min}}^{P_Q^{-1}(1-t)} p_Q(q_1|x_1)P_Q(q_1|x_2)\,\mathrm{d}q_1\,\mathrm{d}x_2\,\mathrm{d}x_1 \\ = \frac{1}{2}\int_{-\infty}^{\infty} x_1 p_x(x_1) \int_{-\infty}^{\infty} x_2 p_x(x_2) P_Q(P_Q^{-1}(1-t)|x_1)P_Q(P_Q^{-1}(1-t)|x_2)\,\mathrm{d}x_2\,\mathrm{d}x_1\,. \tag{142}$$

Noting that the right-hand side of result (142) is one half of the first integration result (139) after x-integration and noting the minus sign in (138), one gets for (136) the expression

$$\int_{-\infty}^{\infty} x_1 p_x(x_1) \int_{-\infty}^{\infty} x_2 p_x(x_2) \int_{q_{\min}}^{P_Q^{-1}(1-t)} p_Q(q_1|x_1)\Big(P_Q(P_Q^{-1}(1-t)|x_2) - P_Q(q_1|x_2)\Big)\,\mathrm{d}q_1\,\mathrm{d}x_2\,\mathrm{d}x_1 \\ = \int_{-\infty}^{\infty} x_1 p_x(x_1) \int_{-\infty}^{\infty} x_2 p_x(x_2)\left(1-\frac{1}{2}\right) P_Q(P_Q^{-1}(1-t)|x_1)P_Q(P_Q^{-1}(1-t)|x_2)\,\mathrm{d}x_2\,\mathrm{d}x_1\,. \tag{143}$$

Inserting the results of (143) back into $[\cdot]$ of (134) and including all prefactors, the five-fold integral simplifies providing the desired result of Eq. (120). □

Theorem 4. *Let p_x denote the density of the i-th component mutation $x \sim \mathcal{N}(0, \sigma^2)$ and let $x_{k;\lambda}$ denote the k-th best value (out of λ) of the i-th mutation component $(\mathbf{x}_{k;\lambda})_i$. Let P_Q denote the quality gain CDF with its quantile function given by P_Q^{-1}. For a truncation ratio ϑ the limit of the second order expected value reads*

$$\lim_{\substack{\mu,\lambda\to\infty \\ \vartheta=\text{const.}}} \frac{1}{\mu(\mu-1)} \sum_{l=2}^{\mu} \sum_{k=1}^{l-1} \mathrm{E}\left[x_{k;\lambda} x_{l;\lambda}\right] = \frac{1}{2}\left[\frac{1}{\vartheta} \int_{-\infty}^{\infty} x_i p_x(x_i) P_Q\left(P_Q^{-1}(\vartheta)|x_i\right) \mathrm{d}x_i\right]^2 . \tag{144}$$

Proof. Starting from Eq. (120) the μ-dependent prefactor was rearranged in a way that the factor $(\mu-1)/\mu$ in (120) is retained in the final result. Formally one could include $(\mu-1)/\mu$ in the sequence (38) and take the limit. However, it is desirable to keep the factor in the progress rate as a correction for finite μ-values. As a next step, one can define $f_x(t) = P_Q\left(P_Q^{-1}(1-t)|x_1\right) P_Q\left(P_Q^{-1}(1-t)|x_2\right)$. As $0 \le f_x(t) \le 1$ the same bound estimation as in (43) holds. Furthermore, both mutation integrals over density p_x are finite, see also (44). Therefore, the limit is evaluated with $f_x(t)\big|_{t=1-\vartheta} = P_Q\left(P_Q^{-1}(\vartheta)|x_1\right) P_Q\left(P_Q^{-1}(\vartheta)|x_2\right)$ and $a = 2$ as

$$\begin{aligned}
\lim_{\substack{\mu,\lambda\to\infty \\ \vartheta=\text{const.}}} \frac{1}{\mu(\mu-1)} E^{(1,1)} &= \frac{1}{2}\frac{1}{\vartheta^2} \int_{-\infty}^{\infty} x_1 p_x(x_1) \int_{-\infty}^{\infty} x_2 p_x(x_2) P_Q\left(P_Q^{-1}(\vartheta)|x_1\right) P_Q\left(P_Q^{-1}(\vartheta)|x_2\right) \mathrm{d}x_2\, \mathrm{d}x_1 \\
&= \frac{1}{2}\frac{1}{\vartheta^2} \int_{-\infty}^{\infty} x_1 p_x(x_1) P_Q\left(P_Q^{-1}(\vartheta)|x_1\right) \mathrm{d}x_1 \int_{-\infty}^{\infty} x_2 p_x(x_2) P_Q\left(P_Q^{-1}(\vartheta)|x_2\right) \mathrm{d}x_2 \\
&= \frac{1}{2}\left[\frac{1}{\vartheta} \int_{-\infty}^{\infty} x_i p_x(x_i) P_Q\left(P_Q^{-1}(\vartheta)|x_i\right) \mathrm{d}x_i\right]^2 ,
\end{aligned} \tag{145}$$

with x_i re-introduced in the last line to denote the i-th mutation component, which gives Eq. (144). □

In $[\cdot]$ of result (144), one can identify the first order progress rate $-\varphi_i$ within the large population limit derived in Eq. (60). Refactoring (144) to obtain $\frac{1}{\mu^2} E^{(1,1)}$, one can insert the φ_i-approximation from (96). Noting that $c_\vartheta^2 = e_\vartheta^{2,0}$ via (45), one gets

$$\begin{aligned}
\frac{1}{\mu^2} E^{(1,1)} &\simeq \frac{1}{2}\frac{\mu-1}{\mu}\varphi_i^2 \\
&\simeq \frac{1}{2}\frac{\mu-1}{\mu} e_\vartheta^{2,0} \frac{\sigma^4}{D_Q^2}\left(2y_i + \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2} \alpha A \sin(\alpha y_i)\right)^2 .
\end{aligned} \tag{146}$$

Finally, inserting the results from (119) and (146) into (101), one obtains the second order progress rate

$$\begin{aligned}
\varphi_i^{\mathrm{II}} \simeq\ & c_\vartheta \frac{\sigma^2}{D_Q}\left(4y_i^2 + \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2} 2\alpha A y_i \sin(\alpha y_i)\right) \\
& - \frac{\sigma^2}{\mu}\left\{1 + e_\vartheta^{1,1}\frac{(2y_i)^2\sigma^2}{D_Q^2} - \frac{c_\vartheta}{D_Q}\left[3\sigma^2 + A\cos(\alpha y_i)\left(1 - \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2} + \alpha^2\sigma^2 \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}\right)\right]\right. \\
& \left. + (\mu-1) e_\vartheta^{2,0}\frac{\sigma^2}{D_Q^2}\left(2y_i + \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}\alpha A \sin(\alpha y_i)\right)^2\right\},
\end{aligned} \tag{147}$$

which serves as an approximation in the asymptotic limit of infinitely large dimensionality and population size. However, experimental investigations will also show good agreement for finite N, μ, and λ.

For future investigations of the convergence and step-size adaptation properties of the $(\mu/\mu_I, \lambda)$-ES, a simpler expression than (147) is needed. To this end, the N-dependency of the terms within $\{\cdot\}$ of (147) is investigated. It will be shown that for $N \to \infty$ and $\mu = o(N)$ only the term $-\sigma^2/\mu$ yields relevant contributions. The relevant terms in $\{\cdot\}$ of Eq. (147) are abbreviated according to their respective factors as $e_\vartheta^{1,1}$, c_ϑ/D_Q and $e_\vartheta^{2,0}$. In order to maximize the absolute value of the individual terms a lower bound for D_Q^2 is needed. Given the form of D_Q^2 from Eq. (31), no useful lower bound for the variance could be established satisfying $D_Q^2 > 0$ for any y_i due to the trigonometric terms. Therefore, we will restrict the analysis to the sphere limit case $A \to 0$. This assumption might seem crude. However, the most important characteristics are already contained in the first φ_i-dependent term of (147) referred to as the *gain* term in sphere model theory [5]. On the other hand, the *loss* terms in $\{\cdot\}$ are mostly dominated by the first term $-\sigma^2/\mu$. Experiments will affirm this assumption.

As the φ_i^{II}-approximation shall be valid for a constant σ^* given any R-value, the mutation strength is re-normalized using (4)

$$\sigma = \frac{\sigma^* R}{N}. \tag{148}$$

Setting $A = 0$, $\sigma = \sigma^* R/N$, and $\sum_i y_i^2 = R^2$ in (31), one obtains the sphere variance for constant normalized mutation strength as

$$\begin{aligned} D_{\text{Q},\text{sph}}^2 &= \sum_{i=1}^{N} [4\sigma^2 y_i^2 + 2\sigma^4] = 4\sigma^2 R^2 + 2N\sigma^4 = 4R^4 \left(\frac{\sigma^*}{N}\right)^2 + 2N \left(\frac{\sigma^* R}{N}\right)^4 \\ &= 4R^4 \left(\frac{\sigma^*}{N}\right)^2 \left(1 + \frac{\sigma^{*2}}{2N}\right). \end{aligned} \tag{149}$$

In the limit $N \to \infty$ the second term of (149) is negligible for constant σ^* giving

$$D_{\text{Q},\text{sph}}^2 \simeq 4R^4 \left(\frac{\sigma^*}{N}\right)^2. \tag{150}$$

Having obtained the sphere variance asymptotic in (150), the terms within $\{\cdot\}$ of (147) are evaluated. The term with prefactor $e_\vartheta^{1,1}$ yields with $\sigma = \sigma^* R/N$ and using (150)

$$e_\vartheta^{1,1} \frac{\sigma^2}{D_{\text{Q}}^2} (2y_i)^2 = e_\vartheta^{1,1} \frac{\left(\sigma^* \frac{R}{N}\right)^2}{4R^4 (\sigma^*/N)^2} (2y_i)^2 = e_\vartheta^{1,1} \frac{y_i^2}{R^2} = O\left(\frac{1}{N}\right). \tag{151}$$

It was used in (151) that a single component y_i^2 contributes in expectation $1/N$ to the residual distance $R^2 = \sum_{j=1}^{N} y_j^2$, see also (12). The second term with prefactor $c_\vartheta / D_{\text{Q}}$ using $D_{\text{Q}} \simeq 2R^2 \sigma^*/N$ with $A = 0$ as

$$\frac{3c_\vartheta \left(\sigma^* \frac{R}{N}\right)^2}{D_{\text{Q}}} = \frac{3c_\vartheta \left(\sigma^* \frac{R}{N}\right)^2}{2R^2 \sigma^*/N} = O\left(\frac{1}{N}\right). \tag{152}$$

The last term with prefactor $e_\vartheta^{2,0}$ yields with $A = 0$ and using (150)

$$\begin{aligned} (\mu - 1) e_\vartheta^{2,0} \frac{\sigma^2}{D_{\text{Q}}^2} (2y_i)^2 &= (\mu - 1) e_\vartheta^{2,0} \frac{\left(\sigma^* \frac{R}{N}\right)^2}{4R^4 (\sigma^*/N)^2} (2y_i)^2 \\ &= (\mu - 1) e_\vartheta^{2,0} \frac{y_i^2}{R^2} = \begin{cases} O\left(\frac{1}{N}\right) & \text{if } \mu(N) = \text{const.} \\ O\left(\frac{\mu(N)}{N}\right) & \text{else.} \end{cases} \end{aligned} \tag{153}$$

In (153) the notation $\mu(N)$ was introduced to emphasize that the population size is usually chosen depending on the dimensionality of the search space. Finally, inserting the results of the loss term investigation for the three terms (151), (152), and (153) back into progress rate (147), one gets for the loss term in $\{\cdot\}$ of (147)

$$-\frac{\sigma^2}{\mu} \left\{ 1 + O\left(\frac{1}{N}\right) + O\left(\frac{\mu(N)}{N}\right) \right\}. \tag{154}$$

Provided that the population size $\mu = o(N)$, i.e., increasing sub-linearly with N, all terms except "1" in $\{\cdot\}$ can be neglected for $N \to \infty$. Theoretical results concerning population sizing, i.e., choosing the necessary $\mu(N)$ to achieve high global convergence probability (success probability), are not available at this point. It is one of the main future goals of the current research project. Note that treating μ as a constant is also not satisfactory, since for large N an increase of μ is necessary to maintain a high success rate on a highly multimodal problem. However, experimental investigations on the Rastrigin function including step-size adaptation suggest a sub-linear relation, which validates the approximation. Finally, the lengthy result (147) is simplified using the loss term asymptotic of (154) and the second order progress rate approximation is obtained.

Second order progress rate *The second order component-wise progress rate on the Rastrigin function in the asymptotic limits of infinitely large population size μ (constant $\vartheta = \mu/\lambda$) and infinitely large dimensionality N with $\mu = o(N)$ yields*

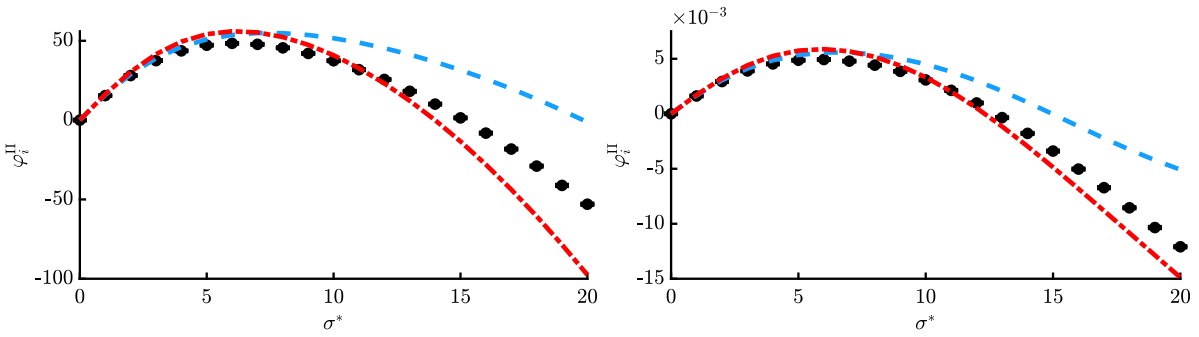

Fig. 6. Second order progress rate φ_i^{II} as a function of σ^* for $(10/10, 40)$-ES with $N = 20$, $A = 1$, $\alpha = 2\pi$, at two residual distances $R = 10\sqrt{N}$ with $y_i = 11.6$ (left) and $R = 0.1\sqrt{N}$ with $y_i = 0.116$ (right). The dashed blue curves show Eq. (147) and the dash-dotted red curves Eq. (156).

$$\varphi_i^{\mathrm{II}} \simeq 2y_i\varphi_i - \frac{\sigma^2}{\mu} \tag{155}$$

$$\simeq c_\vartheta \frac{\sigma^2}{D_Q}\left(4y_i^2 + \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2} 2\alpha A y_i \sin(\alpha y_i)\right) - \frac{\sigma^2}{\mu}. \tag{156}$$

The expressions for $c_\vartheta = e_\vartheta^{1,0}$ from (45) and D_Q from (31) were not inserted to improve readability. The first line (155) emphasizes the dependence of $\varphi_i^{\mathrm{II}}(\varphi_i)$ and can be thought of as a more general formula provided that φ_i is known and the loss term behaves similarly to the sphere function loss term $-\sigma^2/\mu$. The second line (156) shows the explicit results for the Rastrigin function. The results (155) and (156) can be mapped to the Evolutionary Progress Principle [5] as the expressions contain a progress gain and loss term, respectively. Here, the gain part scales with c_ϑ and it is a y_i-dependent expression. Hence, depending on the sign of $y_i \sin(\alpha y_i)$ it may also yield negative contributions due to local attraction moving the ES away from the global optimizer, cf. Fig. 3. The loss term $-\sigma^2/\mu$ is characteristic for intermediate recombination. It introduces significant loss for large σ, but can be decreased using a larger μ due to recombination effects.

Results of one-generation experiments are presented in Figs. 6 and 7 by evaluating (8) over 10^6 trials (black dots with vanishing error bars) and comparing with the obtained approximations. The red dash-dotted line is showing simplified result (156), while the blue dashed line is showing (147). The positions $\mathbf{y}$ were initialized randomly (given R) and kept constant over all repetitions. Fig. 6 shows a smaller dimensionality $N = 20$ and truncation ratio $\vartheta = 1/4$, while Fig. 7 shows larger values $N = 100$ with $\vartheta = 1/2$. This was done to exemplarily investigate the results at different parameter sets.

First thing to note is that the loss term allows negative progress for large σ^*, which was not the case for φ_i. The approximation quality is good for different R-values (see left and right plots, respectively) and improves for larger N and μ in Fig. 7, which was expected. Simplified expression φ_i^{II} from (156) [red, dash-dotted] yields good results compared to (147) [blue, dashed], with (147) giving slightly better results for smaller σ^* and (156) better results at larger σ^*. This indicates that additional terms of the Taylor expansion (70) would be needed to further improve the results of (147). However, this would make the expression more involved, which is not desired. Furthermore, the results of Fig. 6 are relatively good considering that a rather small population $(10/10, 40)$-ES was used at low dimensionality $N = 20$. One can conclude that (156) yields very good results considering its "simplicity". It will therefore be used in Sec. 5 to investigate the dynamical behavior of the ES. It should be noted that at this point there is no aggregated progress measure over all N components, such as the R-dependent sphere progress rate. Given some $\mathbf{y}^{(g)}$ one can evaluate all $i = 1, \ldots, N$ values for φ_i^{II} and obtain a progress vector, but the overall effect on $R^{(g)} \to R^{(g+1)}$ is not known. This will be part of future research. However, the cumulative effect of all N progress rates can be evaluated within a dynamical systems model to be shown in the next chapter.

5. Evolution equations

In the previous sections one-generation experiments were conducted and compared against progress rate results (96), (147), and (156). In order to have an aggregated measure over all components and many generations, φ_i and φ_i^{II} will be used within the evolution equations and compared to real optimization runs of Algorithm 1. Using this method the (mean) global convergence behavior can be investigated.

Given definitions for first and second order progress (7) and (8), the expressions can be reformulated as stochastic iterative mappings between two generations $g \to g + 1$ according to

$$y_i^{(g+1)} = y_i^{(g)} - \varphi_i(\sigma^{(g)}, \mathbf{y}^{(g)}) + \epsilon^{(1)}(\sigma^{(g)}, \mathbf{y}^{(g)}) \tag{157}$$

$$\left(y_i^{(g+1)}\right)^2 = \left(y_i^{(g)}\right)^2 - \varphi_i^{\mathrm{II}}(\sigma^{(g)}, \mathbf{y}^{(g)}) + \epsilon^{(2)}(\sigma^{(g)}, \mathbf{y}^{(g)}). \tag{158}$$

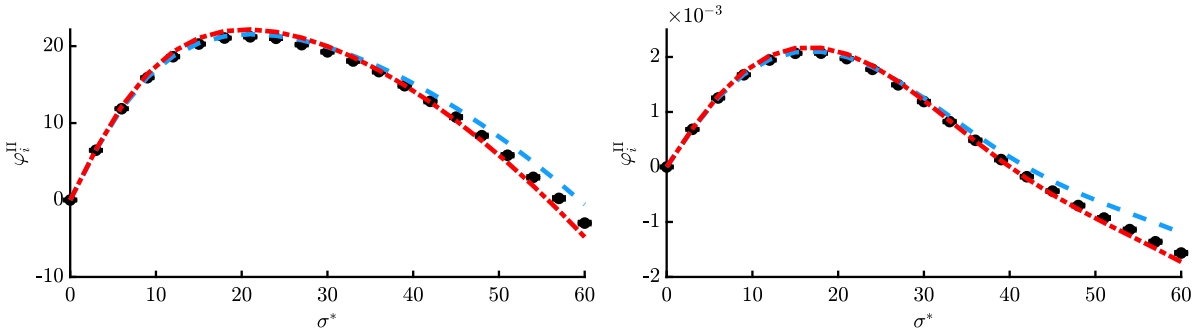

Fig. 7. Second order progress rate φ_i^{II} as a function of σ^* for $(100/100, 200)$-ES with $N = 100$, $A = 1$, $\alpha = 2\pi$, at two residual distances $R = 10\sqrt{N}$ with $y_i = 11.9$ (left) and $R = 0.1\sqrt{N}$ with $y_i = 0.119$ (right). The dashed blue curves show Eq. (147) and the dash-dotted red curves Eq. (156).

The two terms $\epsilon^{(1)}$ and $\epsilon^{(2)}$ can be interpreted as fluctuations w.r.t. the expected values (provided by φ_i and φ_i^{II}). Thus, it holds $\mathrm{E}\left[\epsilon^{(1)}\right] = 0 = \mathrm{E}\left[\epsilon^{(2)}\right]$. However, the exact transition densities for $g \to g+1$ are not known at this point. In principle, they could be approximated using a finite number of higher order moments (or cumulants) to model the fluctuations [5, Ch. 7]. However, for a first study of the progress rate results on the dynamics, the fluctuations are neglected by setting $\epsilon^{(1)} = 0 = \epsilon^{(2)}$. Therefore, one arrives at the (deterministic) equations describing the mean-value dynamics of the parental position coordinates

$$y_i^{(g+1)} = y_i^{(g)} - \varphi_i(\sigma^{(g)}, \mathbf{y}^{(g)}) \tag{159}$$

$$\left(y_i^{(g+1)}\right)^2 = \left(y_i^{(g)}\right)^2 - \varphi_i^{\mathrm{II}}(\sigma^{(g)}, \mathbf{y}^{(g)}), \tag{160}$$

with constant normalized mutation strength σ^* from Eq. (4) giving

$$\sigma^{(g)} = \sigma^* \left\|\mathbf{y}^{(g)}\right\| / N. \tag{161}$$

Two important issues need to be discussed. Firstly, the positional iterations are defined for a single component i. For large N however, it is not feasible to display each component individually. While the components will be iterated separately, the dynamics will be presented as a function of the residual distance $R = \left\|\mathbf{y}^{(g)}\right\|$. Secondly, for the evaluation of φ_i^{II} being a function of $\mathbf{y}^{(g)}$, the square root of the components $(y_i^{(g)})^2$ has to be taken after iteration giving two solutions $\pm y_i^{(g)}$. As the corresponding terms of φ_i^{II} and $D_Q^2(\mathbf{y})$ are even in $y_i^{(g)}$, both solutions are equivalent.

In the following, the deterministic iterations (159) and (160) using mutation strength rescaling (161) are compared to real optimization runs. For the initialization, $\mathbf{y}^{(0)}$ is chosen randomly such that $\left\|\mathbf{y}^{(0)}\right\| = R^{(0)}$ for a given $R^{(0)}$. The starting position is kept constant for consecutive runs of the same experiment. For the magnitude of $R^{(0)}$ it is ensured that the strategy starts far enough away from the local minima landscape. Given Fig. 1 with $A = 1$, the farthermost local minimizer is at $y_i \approx 3$ with resulting $R \approx 3\sqrt{N}$ for N-components, such that $R^{(0)} = 20\sqrt{N} > 3\sqrt{N}$ is chosen.

Considering the choice of σ^* one observes in experiments that larger mutation strengths (compared to a sphere-optimal σ^*) increase the success probability P_S of individual trials to converge to the global optimizer. This is due to the fact that large steps tend to overcome local attraction more easily. However, this comes at the expense of efficiency, since large steps are often overshooting the global optimizer. Therefore in Fig. 8, σ^* is chosen larger than the sphere-optimal value $\hat{\sigma}^*_{\mathrm{sph}}$, which can be obtained numerically from [5, Eq. (6.54)], but small enough to prevent negative progress. The aim was to obtain $P_S \approx 1$.

In order to aggregate the $R^{(g)}$-data of multiple dynamic experiments, the median has shown to be a suitable measure of central tendency. The main issue is that due to fluctuations the $R^{(g)}$-values of distinct ES-runs may differ by orders of magnitude, such that the mean yields biased results due to a skewed distribution. The median is more suitable in this case and a more stable measure.

In Fig. 8 one can observe three phases within the dynamics. First, linear convergence is observed for large $R^{(g)}$-values, where the sphere function dominates. Then, a slow down is observed due to increasing effects of local attraction. For small $R^{(g)}$-values, the ES descends into the global attractor basin and linear convergence can be observed again. One can see that the φ_i-iteration (blue) shows by far too much progress compared to φ_i^{II}-iteration. This is due to the first order model, which does not include loss terms and overestimates the progress significantly, see also discussion of result (96). Iteration via φ_i^{II} (red) shows good results compared to the median curve, especially for larger μ and N (right plot). Better agreement for large populations is also due to reduced fluctuation effects, which were neglected at the beginning of Sec. 5.

In Fig. 9 the effect of reduced σ^* is investigated, which increases the probability of local convergence. The left plot shows $\sigma^* = 5$ with no globally converging runs, as the mutation strength is too low. Technically, for constant σ^* there is no local convergence as the algorithm never stops if R is not decreasing. Still, the experiments are stopped after some

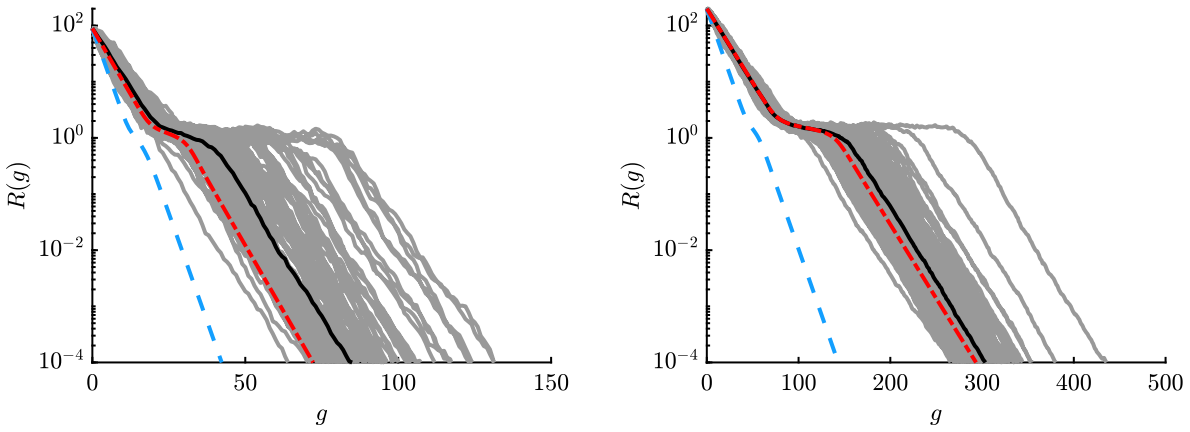

Fig. 8. Comparison of real optimization runs with mean value dynamics using progress rates φ_i via (157) [dashed blue] and φ_i^{II} via (158) [dash-dotted red]. Gray lines show all 100 successful runs of Algorithm 1 and the black line shows the median thereof. The left plot shows $(10/10, 40)$-ES for $N = 20$ with $\sigma^* = 7$ ($\hat{\sigma}^*_{\text{sph}} = 5.7$) and the right one $(100/100, 200)$-ES for $N = 100$ with $\sigma^* = 30$ ($\hat{\sigma}^*_{\text{sph}} = 18.3$). For both experiments $A = 1$, and $\alpha = 2\pi$ are chosen. The resulting success probability $P_S = 1$.

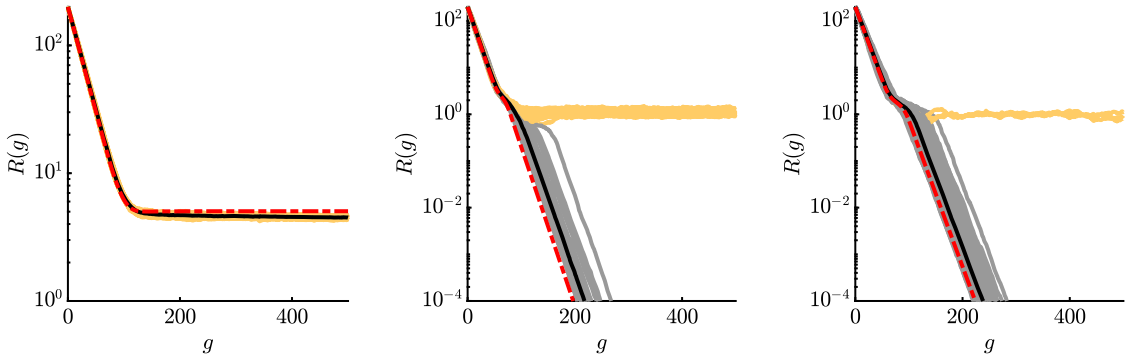

Fig. 9. Variation of σ^* for $(100/100, 200)$-ES for $N = 100$, $A = 1$, and $\alpha = 2\pi$. From left to right $\sigma^* = \{5, 18.3, 25\}$, with $\hat{\sigma}^*_{\text{sph}} = 18.3$, and success rate $P_S = \{0, 0.45, 0.97\}$. The experiment with $\sigma^* = 30$ ($P_S = 1$) was already shown in Fig. 8. Globally converging trials are shown in gray, and non-converging runs in light-orange. The median is taken over the globally converging runs, except for the left plot where none exist, in which the median over all unsuccessful runs is taken.

g-threshold is reached. The stagnating behavior of the ES around some $R^{(g)}$ can be illustrated using Fig. 3. For $\sigma = 0.2$ one has $\sigma^* = \sigma N/R \approx 0.9$, which is small compared to $\hat{\sigma}^*_{\text{sph}} \approx 5.7$. Both left and right progress components of Fig. 3 are significantly influenced by the local attraction region at $\sigma = 0.2$. While some components may be improved (positive value left), others are worsened (negative value right) resulting in a cumulative effect of $R^{(g)}$-stagnation. One way out can be increasing σ (or equivalently σ^*). However, the local minima landscape changes with changing R and arbitrary σ^*-increase is not possible. Stagnation may appear at different σ^* and $R^{(g)}$-values depending on fitness and strategy parameters. For an active step-size adaptation, changing σ appropriately – without converging locally – poses a major challenge.

In the central plot of Fig. 9 roughly half of the runs are globally converging at increased $\sigma^* = \hat{\sigma}^*_{\text{sph}}$. In this case the deterministic iteration follows a single converging path, as no fluctuations are modeled. The residual distance of the locally converging runs is reduced compared to ES-runs with $\sigma^* = 5$. Note that the convergence speed is faster (steeper negative slope) for the globally converging runs compared to $\sigma^* = 30$ of Fig. 8 due to sphere-optimal $\hat{\sigma}^*_{\text{sph}}$. However, this comes with the disadvantage of a lower P_S, as more trials are converging locally. The right plot with $\sigma^* = 25$ is similar to $\sigma^* = 30$ of Fig. 8, but with several non-converging runs. Again, the ES convergence speed is faster, if σ^* is chosen closer to $\hat{\sigma}^*_{\text{sph}}$, but shows a slightly reduced P_S-value. The overall prediction quality of the iterative mapping (160) is good and the results affirm the expectation, that relatively large mutations are favorable to maximize P_S on the Rastrigin function.

To confirm the expectation that the approximation quality increases further for larger μ and N, experiments are shown in Fig. 10. First thing to notice is that positional fluctuations of the ES trials decrease further, such that nearly all runs show a similar R-dynamics. This is related to the intermediate recombination, see Eq. (34), as position $\mathbf{y}^{(g+1)}$ is obtained by averaging over a large number of individuals. One can see good agreement, but for the left plot there is still some room for improvement. This is related to truncation ratio $\vartheta = 1/4$, such that the Taylor expansion point in Eq. (70) via function $g(x_i)$ is shifted by $\Phi^{-1}(\vartheta)$. For $\vartheta = 1/2$ and even larger N and μ (right plot), very good agreement is observed.

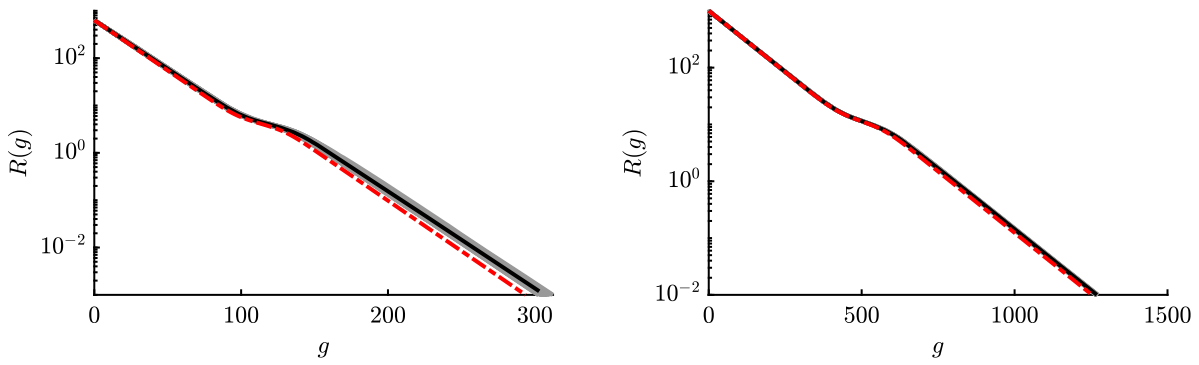

Fig. 10. The left plot shows $(1000/1000, 4000)$-ES with $\sigma^* = 110$ for $N = 1000$, $A = 1$, and $\alpha = 2\pi$. The right plot shows $(10000/10000, 20000)$-ES with $\sigma^* = 400$ for $N = 10000$ (same α and A), evaluated for 50 trials due to CPU resource restrictions.

6. Conclusion and outlook

In this paper the full first and second order progress rate analysis of the $(\mu/\mu_I, \lambda)$-ES has been presented. In order to obtain closed-form expressions for φ_i and φ_i^{II} it was necessary to consider the large dimensionality and large population assumption. While the latter does not present a serious issue because large populations are needed to ensure global convergence, it was the key prerequisite to solve and simplify the expected value integrals. As the experiments have shown, the approximation quality of the progress rate expressions is rather good even for N as small as 20 and comparably small populations of $\mu = 10$. For larger N and μ the approximation quality improves further, as expected. The first order progress rate result is able to model the local attraction effects on the Rastrigin function. This is a very important step, as all subsequent investigations in this paper are based on φ_i-results. The second order progress rate derivation was needed to obtain additional loss terms completing the progress model, which was especially needed for larger mutation strengths and close to the global optimizer.

Using the progress rate expressions, the dynamics of the evolution process have been investigated. There is a good agreement between the iterations and real ES-runs using median aggregation of the residual distance R to the global optimizer. As has been shown, depending on the choice of the normalized mutation strength, one can model global as well as local convergence behavior. Additionally, one observes a trade-off between efficiency and success rate, as relatively large mutations have to be chosen to maximize the success probability.

The conducted experiments assume scale-invariance, i.e., the mutation strength is controlled by the residual distance R. This is in contrast to the full self-adaptive ES where σ evolves during the ES run either by mutative self-adaptation (SA), cumulative step-size adaptation (CSA), or Meta-ES. The incorporation of the self-adaptation process will be the next step completing the analysis of the $(\mu/\mu_I, \lambda)$-ES on Rastrigin. To this end, the self-adaptation response (SAR) function must be derived. Combining N progress rates with the SAR function yields $N + 1$ evolution equations. In order to get manageable expressions that allow for analytic population sizing and expected runtime investigations, additional aggregation is needed. One possible approach would be the aggregation of individual parental y_i components into the parental distance R modeling the expected progress as a function of the residual distance. This would reduce the number of evolution equations to two and making further analytic treatment more accessible. A first step in this direction has been done in [19].

Finally, the presented approach to model the ES-dynamics is based on mean value considerations. That is, fluctuations are not considered so far. Whether the approach presented can be extended to allow for the calculation of the global attractor convergence probability as a function of strategy and fitness parameters remains an open question.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

This work was supported by the Austrian Science Fund (FWF) under grant P33702-N.

References

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

[2] A. Agapie, O. Solomon, L. Bădin, Theory of (1+1) ES on SPHERE revisited, IEEE Trans. Evol. Comput. (2022) 938–948.
[3] A. Agapie, O. Solomon, M. Giuclea, Theory of (1 + 1) ES on the RIDGE, IEEE Trans. Evol. Comput. 26 (3) (2022) 501–511.
[4] D.V. Arnold, Noisy Optimization with Evolution Strategies, Kluwer Academic Publishers, Dordrecht, 2002.
[5] H.-G. Beyer, The Theory of Evolution Strategies, Natural Computing Series, Springer, Heidelberg, 2001.
[6] H.-G. Beyer, D.V. Arnold, S. Meyer-Nieberg, A new approach for predicting the final outcome of evolution strategy optimization under noise, Genet. Program. Evol. Mach. 6 (1) (2005) 7–24.
[7] H.-G. Beyer, A. Melkozerov, The dynamics of self-adaptive multi-recombinant evolution strategies on the general ellipsoid model, IEEE Trans. Evol. Comput. 18 (5) (2014) 764–778, https://doi.org/10.1109/TEVC.2013.2283968.
[8] H.-G. Beyer, H.-P. Schwefel, Evolution strategies: a comprehensive introduction, Nat. Comput. 1 (1) (2002) 3–52.
[9] H.-G. Beyer, B. Sendhoff, Toward a steady-state analysis of an evolution strategy on a robust optimization problem with noise-induced multi-modality, IEEE Trans. Evol. Comput. 21 (4) (2017) 629–643, https://doi.org/10.1109/TEVC.2017.2668068.
[10] P. Billingsley, Probability and Measure, Wiley Series in Probability and Statistics, Wiley, 1995.
[11] N. Hansen, S. Kern, Evaluating the CMA evolution strategy on multimodal test functions, in: X. Yao, et al. (Eds.), Parallel Problem Solving from Nature 8, Springer, Berlin, 2004, pp. 282–291.
[12] M. Hellwig, H.-G. Beyer, On the steady state analysis of covariance matrix self-adaptation evolution strategies on the noisy ellipsoid model, Theor. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2018.05.016.
[13] A. Melkozerov, H.-G. Beyer, On the analysis of self-adaptive evolution strategies on elliptic model: first results, in: J. Branke, et al. (Eds.), GECCO'10: Proceedings of the Genetic and Evolutionary Computation Conference, ACM, New York, 2010, pp. 369–376.
[14] S. Meyer-Nieberg, Self-Adaptation in Evolution Strategies, PhD thesis, University of Dortmund, CS Department, Dortmund, Germany, 2007.
[15] N. Müller, T. Glasmachers, Non-local optimization: imposing structure on optimization problems by relaxation, in: Foundations of Genetic Algorithms, 16, ACM, 2021, pp. 1–10.
[16] A. Omeradzic, H.-G. Beyer, Progress Analysis of a Multi-Recombinative Evolution Strategy on the Highly Multimodal Rastrigin Function, Report, Vorarlberg University of Applied Sciences, 2022, https://opus.fhv.at/frontdoor/index/index/docId/4722.
[17] A. Omeradzic, H.-G. Beyer, Progress rate analysis of evolution strategies on the Rastrigin function: first results, in: G. Rudolph, A.V. Kononova, H. Aguirre, P. Kerschke, G. Ochoa, T. Tušar (Eds.), Parallel Problem Solving from Nature – PPSN XVII, Springer International Publishing, 2022, pp. 499–511.
[18] A. Omeradzic, H.-G. Beyer, Rastrigin Function: Quality Gain and Progress Rate for $(\mu/\mu_I, \lambda)$-ES, Report, Vorarlberg University of Applied Sciences, 2023, https://opus.fhv.at/frontdoor/index/index/docId/5151.
[19] A. Omeradzic, H.-G. Beyer, 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, Association for Computing Machinery, New York, NY, USA, 2023, pp. 117–128.