Self-Adaptation of Multi-Recombinant Evolution Strategies on the Highly Multimodal Rastrigin Function

Amir Omeradzic, Hans-Georg Beyer
Vorarlberg University of Applied Sciences, Research Center Business Informatics
{amir.omeradzic, hans-georg.beyer}@fhv.at

Abstract—**The self-adaptive, multi-recombinative $(\mu/\mu_I, \lambda)$-ES (Evolution Strategy) is investigated on the highly-multimodal Rastrigin test function by theoretical and experimental means. To this end, the self-adaptation response function is derived in the limit of large populations, which are necessary to achieve high success rates. The established dynamical systems approach is introduced and steady state conditions on Rastrigin are discussed and compared to the sphere function. Then, a characteristic τ is derived to tune the sampling process of the self-adaptive ES. The obtained result is compared to default τ-values. Furthermore, expected runtime experiments are conducted varying τ and population parameters of the ES. Theoretical and experimental results regarding τ are compared in terms of efficiency and robustness showing good agreement.**

Index Terms—**Evolution Strategy, Rastrigin Function, Self-Adaptation, Multimodality**

I. INTRODUCTION

The optimization of real-valued highly multimodal test functions poses a major challenge for conventional gradient-based optimization algorithms. Evolutionary Algorithms, such as Evolution Strategies (ES) [1], have experimentally proven to be capable of localizing the global optimizer among a vast number of local optima provided that the strategy parameters are chosen appropriately [2]. Besides the population size [3], [4], a key ingredient to improve global search is the adaptation of the mutation strength σ. State-of-the-art implementations for ES are self-adaptation (σSA) and cumulative step-size adaptation (CSA). CSA works in a derandomized way by comparing a measured σ-path-length to the path length expected under random selection [5]. Self-adaptation, which is the focus of this paper, samples offspring mutation strengths from a distribution (usually log-normal). Selection by fitness implicitly selects corresponding mutation strengths [6], [7]. σSA-ES have been studied theoretically on the sphere [7], [8], rigde function [8], and the general ellipsoid model [9], which resulted in recommended default parameter settings for the σ-sampling.

This paper extends the analysis of the self-adaptive multi-recombinative $(\mu/\mu_I, \lambda)$-σSA-ES to the Rastrigin test function. Theoretical as well as experimental investigations will be conducted to investigate the dynamics and global convergence properties. In Sec. II the problem is formulated and the ES is introduced. Definitions and the overall model of the $(\mu/\mu_I, \lambda)$-σSA-ES are presented in Secs. III and IV. The self-adaptation response function is derived in Sec. V-A and then tested within a dynamical systems approach in Sec. V-B. Then, a characteristic quantity τ for the σ-sampling on the Rastrigin function is derived. In Sec. VI the efficiency of the σSA-ES is experimentally evaluated. Finally, conclusions are drawn in Sec. VII.

II. TEST FUNCTION AND EVOLUTION STRATEGY

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

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

with oscillation amplitude A and frequency α (default value $\alpha = 2\pi$ is used throughout the paper). Minimization of f is considered with the global minimizer located at $\hat{\mathbf{y}} = \mathbf{0}$. Given A and α the function has a finite number of local attractors in each dimension, such that the overall number of local minima scales exponentially with N. The Rastrigin function has a quadratic global structure and the convergence of the ES on f is evaluated using the residual distance R defined as $R^2 := \sum_i y_i^2$.

The Rastrigin function is investigated using a $(\mu/\mu_I, \lambda)$-ES, see Algorithm 1, with μ parents, λ offspring, intermediate multi-recombination, and generation counter g. The selection pressure (or truncation ratio) is defined as $\vartheta := \mu/\lambda$. Isotropic normally distributed mutations of strength σ are applied and σ is adapted using self-adaptation by sampling log-normally distributed values for each offspring (Lines 5-7), see also density (29). The adaptation is tuned via the learning parameter τ, with smaller values yielding a slower, more robust adaptation and larger values resulting in faster, more efficient, but less robust adaptation. The default value for τ was derived in [8] on the sphere in the limit $N \to \infty$ as $\tau = 1/\sqrt{2N}$. On the Rastrigin function one observes a higher success probability P_S of Alg. 1, if smaller values for τ are chosen. The τ-dependence will be investigated further throughout the paper.

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

1: $g \leftarrow 0$
2: initialize $\left(\mathbf{y}^{(0)}, \sigma^{(0)}\right)$
3: **repeat**
4:     **for** $l = 1, ..., \lambda$ **do**
5:         $\tilde{\sigma}_l \leftarrow \sigma^{(g)} \mathrm{e}^{\tau \mathcal{N}_l(\mathbf{0}, \mathbf{1})}$
6:         $\tilde{\mathbf{x}}_l \leftarrow \tilde{\sigma}_l \mathcal{N}_l(\mathbf{0}, \mathbf{1})$
7:         $\tilde{\mathbf{y}}_l \leftarrow \mathbf{y}^{(g)} + \tilde{\mathbf{x}}_l$
8:         $\tilde{f}_l \leftarrow f(\tilde{\mathbf{y}}_l)$
9:     **end for**
10:    $(\tilde{f}_{1;\lambda}, ..., \tilde{f}_{m;\lambda}, ..., \tilde{f}_{\mu;\lambda}) \leftarrow \mathrm{sort}(\tilde{f}_1, ..., \tilde{f}_\lambda)$
11:    $\mathbf{y}^{(g+1)} \leftarrow \frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{\mathbf{y}}_{m;\lambda}$
12:    $\sigma^{(g+1)} \leftarrow \frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{\sigma}_{m;\lambda}$
13:    $g \leftarrow g + 1$
14: **until** termination criterion

III. DEFINITIONS AND PRELIMINARY RESULTS

In this section the local performance measures will be defined, which will be necessary for the dynamic model presented in Sec. IV. As the focus of this paper is the investigation of the self-adaptation, important preliminary results will also be stated here.

A. Local Performance Measures

The strategy's local performance in search space can be measured using the so-called progress rate (denoted by φ). The first order i-th component progress rate φ_i is defined as the expected value (given position $\mathbf{y}$ and mutation strength σ)

$$\varphi_i := \mathrm{E}\left[y_i^{(g)} - y_i^{(g+1)} \,\big|\, \mathbf{y}^{(g)}, \sigma^{(g)} \right]. \quad (2)$$

Similarly to (2), a second order progress along the i-th component is defined by squaring the values as

$$\varphi_i^{\mathrm{II}} := \mathrm{E}\left[(y_i^{(g)})^2 - (y_i^{(g+1)})^2 \,\big|\, \mathbf{y}^{(g)}, \sigma^{(g)} \right]. \quad (3)$$

φ_i^{II} was first introduced in [9] analyzing self-adaptation on the ellipsoid model. The first order φ_i can be used if comparably small mutation strengths σ are used relative to the residual distance R. For larger mutation strengths the ES performs large steps, such that y_i often changes its sign between generations. In this case φ_i^{II} is necessary for a more accurate model of progress towards the optimizer, as it can be approached independent of the sign of y_i. φ_i^{II} measures the progress of single components, while global convergence is measured (usually) as a function of the residual distance R. One can think of examples where global progress ($R^{(g+1)} < R^{(g)}$) occurs with some components having $\varphi_i^{\mathrm{II}} > 0$ and others showing $\varphi_i^{\mathrm{II}} < 0$, such that demanding $\varphi_i^{\mathrm{II}} > 0$ for all i would be too strict. Therefore, an additional aggregation step is needed to obtain an R-dependent measure. Summing over (3) and using $R^2 = \sum_i y_i^2$ one defines [4]

$$\sum_{i=1}^{N} \varphi_i^{\mathrm{II}} = \mathrm{E}\left[(R^{(g)})^2 - (R^{(g+1)})^2 \big| R^{(g)}, \sigma^{(g)} \right] =: \varphi_R^{\mathrm{II}}. \quad (4)$$

The evaluation of (4) is done in Sec. III-B. It is important to note that all progress rates above will be evaluated asymptotically in the limit $\tau \to 0$. This has two reasons. First, the evaluation of the expected values proves to be very hard, if not impossible, when $\tau > 0$ is chosen. Second, to obtain closed-form solutions of the self-adaptation response (SAR) at all, the limit $\tau \to 0$ is also needed.

The SAR, denoted by the symbol ψ, measures the relative change of the mutation strength σ and is defined as

$$\psi := \mathrm{E}\left[(\sigma^{(g+1)} - \sigma^{(g)})/\sigma^{(g)} \,\big|\, \mathbf{y}^{(g)}, \sigma^{(g)} \right]. \quad (5)$$

The evaluation of all progress rates and the SAR requires to model the selection process of the ES. To this end, one defines the quality gain Q at location $\mathbf{y}$ due to isotropic mutation $\mathbf{x} \sim \sigma \mathcal{N}(\mathbf{0}, \mathbf{1})$ as the local fitness change

$$Q := f(\mathbf{y} + \mathbf{x}) - f(\mathbf{y}). \quad (6)$$

As the quality gain is a random variate, a probabilistic model for the distribution of Q is required. The cumulative distribution function (CDF) of Q is denoted as P_Q. Exact solutions for P_Q cannot be obtained due to the underlying structure of (1), as it contains sums over trigonometric and squared random variates. However, based on the idea of the central limit theorem (CLT), one can apply a normal approximation of the quality gain distribution for comparably large N (see discussion in [10]). While the applicability of the CLT cannot be rigorously proven, experiments do show very good agreement. The normality assumption is therefore not critical for the further analysis. Using expected value $E_Q := \mathrm{E}[Q]$ and variance $D_Q^2 := \mathrm{Var}[Q]$ one approximates P_Q using the standard normal CDF $\Phi(\cdot)$ as

$$\widetilde{P}_Q = \Phi\left(\frac{q - E_Q}{D_Q} \right). \quad (7)$$

Expected value and variance in (7) are evaluated in [10] as functions of the Rastrigin parameters and mutation strength σ. One has [10, (30)]

$$E_Q(\mathbf{y}) = \sum_{i=1}^{N} \left[\sigma^2 + A \cos(\alpha y_i) \left(1 - \mathrm{e}^{-\frac{(\alpha\sigma)^2}{2}} \right) \right], \quad (8)$$

and the variance is evaluated as [10, (31)]

$$\begin{aligned}
D_Q^2(\mathbf{y}) = \sum_{i=1}^{N} \bigg\{ & 4\sigma^2 y_i^2 + 2\sigma^4 \\
& + \frac{A^2}{2}\left(1 - \mathrm{e}^{-(\alpha\sigma)^2} \right)\left(1 - \cos(2\alpha y_i)\mathrm{e}^{-(\alpha\sigma)^2} \right) \\
& + 2A\alpha\sigma^2 \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2}\left(\alpha\sigma^2 \cos(\alpha y_i) + 2y_i \sin(\alpha y_i) \right) \bigg\}. \quad (9)
\end{aligned}$$

Based on the quality gain distribution (7) progress rates φ_i and φ_i^{II} have been derived in [10]. For the further analysis only second order φ_i^{II} is needed. It was derived in [10, p. 24] assuming $\mu, \lambda, N \to \infty$ (constant ϑ) as

$$\varphi_i^{\mathrm{II}} \simeq c_\vartheta \frac{\sigma^2}{D_Q}\left(4y_i^2 + \mathrm{e}^{-\frac{1}{2}(\alpha\sigma)^2} 2\alpha A y_i \sin(\alpha y_i) \right) - \frac{\sigma^2}{\mu}. \quad (10)$$

Note that in (10) the exponential function models the limited range of local attraction. Additionally, the asymptotic generalized progress coefficients $e_\vartheta^{a,b}$ were derived [10, (45)] for $a \geq 1$, $b \geq 0$ as

$$e_\vartheta^{a,b} := \left(e^{-\frac{1}{2}\left[\Phi^{-1}(\vartheta)\right]^2}/(\sqrt{2\pi}\vartheta) \right)^a \left(-\Phi^{-1}(\vartheta)\right)^b, \quad (11)$$

with the special case $c_\vartheta := e_\vartheta^{1,0}$ used in (10). They correspond to the generalized progress coefficients $e_{\mu,\lambda}^{a,b}$ of [7, (5.112)] in the limit of infinitely large population size.

B. Component Aggregation

As already discussed above Eq. (4) the aggregation of all component-wise φ_i^{II} is needed for a global (R^2-dependent) progress rate. This method will be useful to model the dynamics of the ES as a function of the residual distance R, see Eq. (23). Furthermore, some important relations to the sphere function can be established easily.

Starting from (4), the evaluation of $\sum_i \varphi_i^{\mathrm{II}}(\mathbf{y})$ given some value for $R^2 = \sum_i y_i^2$ is not trivial, as terms containing trigonometric functions in y_i appear. This issue was discussed in [4] and a stochastic averaging method was introduced, as analytic averaging is intractable for $N \geq 3$. Stochastic averaging assumes that the components y_i are random i.i.d. variates distributed normally around the global optimizer $\hat{\mathbf{y}} = 0$ for any g, such that

$$y_i \sim \mathcal{N}(0, R^2/N). \quad (12)$$

Note that the expectation of the sum over squared components yields $\mathrm{E}\left[\sum_i y_i^2\right] = R^2$. By assumption (12) each component contributes roughly as $y_i^2 \approx R^2/N$ to the residual distance R^2. Using random variates (12) the sums over trigonometric functions can be approximated by their corresponding expected values (applying the CLT for $N \to \infty$), such that the aggregated progress rate is derived as [4, (35)]

$$\varphi_R^{\mathrm{II}} = c_\vartheta \frac{2R^2\sigma^2}{D_Q(R)} \left(2 + \alpha^2 A e^{-\frac{\alpha^2}{2}\left(\sigma^2 + \frac{R^2}{N}\right)} \right) - N\frac{\sigma^2}{\mu}. \quad (13)$$

Result (13) is now R^2-dependent and contains the (aggregated) information of N progress rates. Using the same approach, the aggregated quality gain variance $D_Q^2(R)$ is derived from (9) as [4, (36)]

$$D_Q^2(R) = 4R^2\sigma^2 + 2N\sigma^4 +$$
$$+ \frac{NA^2}{2}\left(1 - e^{-(\alpha\sigma)^2}\right)\left(1 - e^{-\alpha^2\left(\sigma^2 + 2\frac{R^2}{N}\right)}\right)$$
$$+ 2NA\alpha^2\sigma^2 e^{-\frac{\alpha^2}{2}\left(\sigma^2 + \frac{R^2}{N}\right)}\left(\sigma^2 + 2\frac{R^2}{N}\right). \quad (14)$$

Similarly, the expectation value (8) yields [11, (1.74)]

$$E_Q(R) = N\sigma^2 + NA e^{-\frac{1}{2}\frac{(\alpha R)^2}{N}}\left(1 - e^{-\frac{(\alpha\sigma)^2}{2}}\right). \quad (15)$$

The results (13) and (14) can be readily used to obtain the sphere quality gain $\mathrm{E}\left[(R^{(g)})^2 - (R^{(g+1)})^2\right]$ as a special case by setting $A = 0$ (or $\alpha = 0$). The sphere is considered here as it provides similarities in the dynamical behavior compared to the Rastrigin function. The sphere progress rate is defined as $\varphi := \mathrm{E}\left[R^{(g)} - R^{(g+1)}\right]$. By introducing normalizations

(denoted by $*$) as $\varphi^* = \varphi N/R$ and $\varphi_R^{\mathrm{II},*} = \varphi_R^{\mathrm{II}} N/(2R^2)$, see [12, p. 16], the asymptotic equality ($N \to \infty$) holds

$$\varphi^* \simeq \varphi_R^{\mathrm{II},*}, \quad (16)$$

which yields the sphere progress rate (setting $A = 0$) as

$$\varphi_{\mathrm{sph}}^* = \frac{c_\vartheta \sigma^*}{\sqrt{1 + \sigma^{*2}/2N}} - \frac{\sigma^{*2}}{2\mu}. \quad (17)$$

Equation (17) is a function of scale-invariant mutation strength σ^*, which is obtained by normalizing σ with dimension N and residual distance R according to

$$\sigma^* = \sigma N/R. \quad (18)$$

On the sphere one has $\varphi_{\mathrm{sph}}^* > 0$ for any $\sigma^* \in (0, \sigma_{\varphi_0}^*)$, where $\sigma_{\varphi_0}^*$ denotes the (second) zero of φ_{sph}^*, which was evaluated in [4, (41)] as

$$\sigma_{\varphi_0}^* = \left[\left(N^2 + 8Nc_\vartheta^2\mu^2\right)^{1/2} - N\right]^{1/2}. \quad (19)$$

Depending on the choice of learning parameter τ a constant (in expectation) steady state σ^*-level is attained on the sphere, which results in a constant scale-invariant progress rate (17) for given N, μ, and ϑ, see e.g. Fig. 2. On the Rastrigin function the condition $\sigma^* \in (0, \sigma_{\varphi_0}^*)$ also holds due to its quadratic global structure. Furthermore, small τ raise the steady state σ^*-level closer to $\sigma_{\varphi_0}^*$, which increases the success rate at the expense of having a lower efficiency. This will be important for the derivation of τ in Sec. V-B.

IV. DYNAMICS AND STEADY STATE

The progress rate and SAR defined in Sec. III model the expected change of $\mathbf{y}$ and σ between two generations. We are interested in modeling and predicting the dynamics of the σSA-ES over many generations to investigate its convergence properties. An established method is the dynamical systems approach introduced in [7, Ch. 7]. It enables to test the analytically obtained formulas against actual optimization runs over many generations. Furthermore, steady state conditions of the dynamic system can be derived. This section should outline the overall goal, such the derivation of required quantities follows in the subsequent sections.

Starting with the second order progress rate (3) one can rearrange the equation in terms of an iterative mapping $g \to g + 1$ for the i-th component as

$$(y_i^{(g+1)})^2 = (y_i^{(g)})^2 - \varphi_i^{\mathrm{II}}(\sigma^{(g)}, \mathbf{y}^{(g)}) + \epsilon_y(\sigma^{(g)}, \mathbf{y}^{(g)}), \quad (20)$$

with the random variate ϵ_y modeling fluctuations in the search space. Analogously, the definition (5) of the SAR-function is rearranged giving an iterative mapping

$$\sigma^{(g+1)} = \sigma^{(g)}(1 + \psi(\sigma^{(g)}, \mathbf{y}^{(g)})) + \epsilon_\sigma(\sigma^{(g)}, \mathbf{y}^{(g)}), \quad (21)$$

with ϵ_σ modeling the fluctuations of the σ-adaptation. As φ_i^{II} and ψ are expected values of their respective quantities, it holds $\mathrm{E}\left[\epsilon_y\right] = \mathrm{E}\left[\epsilon_\sigma\right] = 0$. For the subsequent investigations the fluctuation terms will be neglected by setting $\epsilon_y = \epsilon_\sigma = 0$. This assumption is of course a simplification resulting in a deterministic iteration. However, it will be compared with

averaged dynamics of real optimization runs. One obtains the evolution equations of the **y**-dependent iteration as

$$(y_i^{(g+1)})^2 = (y_i^{(g)})^2 - \varphi_i^{\mathrm{II}}(\sigma^{(g)}, \mathbf{y}^{(g)})$$
$$\sigma^{(g+1)} = \sigma^{(g)}(1 + \psi(\sigma^{(g)}, \mathbf{y}^{(g)})), \qquad (22)$$

with $i = 1, ..., N$, which yields a set of $N+1$ equations to be iterated. It will be abbreviated as "**y**-iteration". Note that both $\varphi_i^{\mathrm{II}}(\mathbf{y})$ and $\psi(\mathbf{y})$ are even functions of y_i, such that taking the square-root of $(y_i^{(g+1)})^2$ does not pose an issue.

Using the same argumentation as for Eq. (20), one can define R^2-dependent evolution equations using (4) and $\psi(\sigma, R)$

$$(R^{(g+1)})^2 = (R^{(g)})^2 - \varphi_R^{\mathrm{II}}(\sigma^{(g)}, R^{(g)})$$
$$\sigma^{(g+1)} = \sigma^{(g)}(1 + \psi(\sigma^{(g)}, R^{(g)})), \qquad (23)$$

which reduces the number of equations to two compared to (22). It will be abbreviated as "R-iteration" and further investigated in Sec. V-B.

A goal of the analysis is to derive a steady state condition for the evolution of scale-invariant σ^*, see (18), which rescales σ by current residual distance R (and N). For **y**-iteration (22) there are N positional coordinates to be evaluated. The steady state analysis of the $N+1$ dimensional system is not tractable, such that the focus is set on R-iteration (23). Recalling that $\varphi = \mathrm{E}\left[R^{(g)} - R^{(g+1)}\right]$, $\varphi^* = \varphi N/R$, and $\varphi^* \simeq \varphi_R^{\mathrm{II},*}$, R-iteration (23) is re-formulated using (18) as

$$R^{(g+1)} = R^{(g)}\left(1 - \varphi_R^{\mathrm{II},*}/N\right) \qquad (24)$$

$$\sigma^{*,(g+1)} R^{(g+1)}/N = (\sigma^{*,(g)} R^{(g)}/N)(1 + \psi). \qquad (25)$$

Inserting (24) into (25) yields the σ^*-evolution

$$\sigma^{*,(g+1)} = \sigma^{*,(g)}(1 + \psi)/(1 - \varphi_R^{\mathrm{II},*}/N). \qquad (26)$$

Imposing the steady state condition $\sigma_{ss}^* = \sigma^{*,(g+1)} = \sigma^{*,(g)}$ and noting that in the general case $\varphi_R^{\mathrm{II},*} = \varphi_R^{\mathrm{II},*}(\sigma_{ss}^*, R)$ and $\psi = \psi(\sigma_{ss}^*, R, \tau)$, one has

$$\varphi_R^{\mathrm{II},*}(\sigma_{ss}^*, R) = -N\psi(\sigma_{ss}^*, R, \tau). \qquad (27)$$

On the sphere function one has scale-invariant progress $\varphi_R^{\mathrm{II},*}(\sigma^*)$ and self-adaptation $\psi(\sigma^*, \tau)$, such that (27) yields the simplified condition (analogous to [7, (7.162)])

$$\varphi_R^{\mathrm{II},*}(\sigma_{ss}^*) = -N\psi(\sigma_{ss}^*, \tau). \qquad (28)$$

In principle, one is interested in solving Eq. (27) to obtain the strategy parameter τ. As discussed later in Sec. V-B there is no closed-form solution of (27) for Rastrigin. Furthermore, a resulting τ should be scale-invariant according to Eq. (28) to guarantee good performance on the sphere, which is the global structure of Rastrigin. Conditions for a steady state σ_{ss}^* will be discussed in Sec. V-B showing that scale-invariance cannot be achieved on Rastrigin. Instead, a "critical" σ_{crit}^* is derived, which characterizes the Rastrigin-specific dynamics well in terms of the sphere steady state. It will enable to solve (28) for learning parameter τ, which in turn enables to tune the step-size adaptation of the ES.

V. SELF-ADAPTATION

The progress rates defined in Sec. III evaluate the expected positional change given $\sigma^{(g)}$. A full description of the self-adaptive σSA-ES requires the evaluation of the σ-change using the SAR function defined in (5), which is done now.

A. Self-Adaptation Response (SAR)

For the $(\mu/\mu_I, \lambda)$-σSA-ES, $\sigma^{(g)}$ denotes the recombined mutation strength of the $m = 1, ..., \mu$ selected values (see Alg. 1, Line 12). From this point the conditional variables $\mathbf{y}^{(g)}$ and $\sigma^{(g)}$ are implicitly assumed to be given and $\sigma = \sigma^{(g)}$ is used for brevity. Random mutations $\tilde{\sigma}$ are sampled according to the probability density of the log-normal mutation operator $p_\sigma(s)$, see Line 5. It is defined in terms of learning parameter τ given mutation strength σ for $s \in (0, \infty)$ as

$$p_\sigma(s) := \frac{1}{\sqrt{2\pi}\tau s} \exp\left[-\frac{1}{2}\left(\frac{\ln(s/\sigma)}{\tau}\right)^2\right]. \qquad (29)$$

First, expected value (5) after selection and recombination can be rewritten in terms of an order statistic density $p_{m;\lambda}$ for the m-th best individual as

$$\psi = \frac{1}{\mu} \sum_{m=1}^{\mu} \int_0^\infty \left(\frac{s - \sigma}{\sigma}\right) p_{m;\lambda}(s|\sigma)\, ds. \qquad (30)$$

Equation (30) was evaluated on the sphere in [8] in the asymptotic limits $N \to \infty$ ($\mu \ll N$) and $\tau \to 0$. In this paper a slightly different approach is taken. As the progress rate (10) was derived for large populations $\mu, \lambda \to \infty$ (constant $\vartheta = \mu/\lambda$), the goal is to obtain a corresponding SAR for infinitely large populations. The details of the derivation are presented in Appendix A. To this end, a large population identity is applied and the limit $\tau \to 0$ is assumed to provide a closed-form solution. The final result (A.20) yields

$$\psi \simeq \tau^2\left(\frac{1}{2} - c_\vartheta \sigma \frac{E_Q'(\sigma)}{D_Q(\sigma)} + e_\vartheta^{1,1} \sigma \frac{D_Q'(\sigma)}{D_Q(\sigma)}\right). \qquad (31)$$

Result (31) has interesting properties. It is a generic expression for which the positional dependency and the test function parameters are contained within E_Q, D_Q, and its corresponding derivatives. The population dependence is contained solely in c_ϑ and $e_\vartheta^{1,1}$. Furthermore, it can be mapped directly to result [8, (C.23)], where the derivation of ψ was performed for finite μ and λ giving $\psi \simeq \tau^2(\frac{1}{2} - c_{\mu/\mu,\lambda}\sigma E_Q'/D_Q + e_{\mu,\lambda}^{1,1}\sigma D_Q'/D_Q)$. Note that the same functional dependencies are obtained with the large population limit only affecting the coefficients giving $e_{\mu,\lambda}^{a,b} \to e_\vartheta^{a,b}$. This transformation of progress coefficients into their asymptotic form was also observed for the progress rates in [10].

The evaluation of ψ on the Rastrigin function requires to insert E_Q and D_Q^2 (with respective derivatives) into (31). As the results are very lengthy this step is omitted for better readability. However, the expressions will be evaluated during subsequent experiments. The SAR as a function of **y** and R, respectively, yields

$$\psi(\mathbf{y}) \simeq \tau^2\left(\frac{1}{2} - c_\vartheta \sigma \frac{E_Q'(\mathbf{y})}{D_Q(\mathbf{y})} + e_\vartheta^{1,1} \sigma \frac{D_Q'(\mathbf{y})}{D_Q(\mathbf{y})}\right) \qquad (32)$$

$$\psi(R) \simeq \tau^2 \left(\frac{1}{2} - c_\vartheta \sigma \frac{E'_Q(R)}{D_Q(R)} + e_\vartheta^{1,1} \sigma \frac{D'_Q(R)}{D_Q(R)} \right), \quad (33)$$

with $E_Q(\mathbf{y})$ (8), $D_Q^2(\mathbf{y})$ (9), $E_Q(R)$ (15), and $D_Q^2(R)$ (14). Note that in (32) and (33) the σ-dependence is implicitly assumed as given and the derivative is taken w.r.t. σ.

At this point one-generation experiments are conducted to investigate the theoretical prediction for ψ. To this end, single optimization steps are performed and (5) is averaged over 10^5 trials. The results are then compared to closed-form results (32) and (33).

Figure 1a shows ψ for two different random realizations of $\mathbf{y}$ illustrating the effects of local attraction. Very good agreement between simulation and $\psi(\mathbf{y})$ can be observed, while $\psi(R)$ shows larger deviations for smaller σ. Given very small σ ψ is always positive. In this case the explored region is very small and any test function can be approximated by a hyperplane. For moderate $\sigma < 0.5$ a larger region of the function is explored and local attraction (or repulsion in the case of a local maximum) effects are prominent. On the left side $\psi(\mathbf{y})$ increases from its initial value indicating that selected individuals have larger sampled $\tilde{\sigma}$. This occurs for example when the ES is close to a local maximum. On the right side $\psi(\mathbf{y})$ drops significantly for moderate σ, which indicates local attraction effects and smaller sampled $\tilde{\sigma}$. $\psi(R)$ does not model local attraction correctly, as it evaluates an average response at $\|\mathbf{y}\| = R$ and no specific location $\mathbf{y}$ (fixed in the experiment). In dynamic optimization runs (at small σ) one observes that component aggregation (12) is violated and the y_i are not isotropically distributed due to local attraction. For large $\sigma > 0.5$, however, $\psi(\mathbf{y})$ and $\psi(R)$ are very similar, as the global structure of the Rastrigin function, i.e., the sphere, becomes dominating and (12) holds (global search of the ES). In this case $\psi < 0$ is desired as $\sigma^{(g+1)} < \sigma^{(g)}$ with $R^{(g+1)} < R^{(g)}$ when minimizing a quadratic function. This observation suggests displaying $\psi(\sigma^*)$ as a function of the scale-invariant mutation strength (18).

Figure 1b shows $\psi(\sigma^*)$ for two values of τ. The range of σ^* was chosen large enough to cover the range of positive progress on the sphere defined by $\sigma_{\varphi_0}^*$ in (19). One observes larger deviations of the derived result if τ is chosen larger (left plot), which was expected. Most of the deviations are due to Taylor expansion (A.15) and neglecting higher orders of $O(\tau^4)$. As $\psi(R)$ represents an aggregated measure over all y_i-components with $\|\mathbf{y}\| = R$, it can be regarded as an average value. For moderate and large σ^* the deviations between $\psi(\mathbf{y})$ and $\psi(R)$ vanish and local attraction is negligible (analogous to Fig. 1a). One may conclude that $\psi(R)$ yields good agreement with experiments for sufficiently large σ^*, where the spherical structure of Rastrigin is dominating.

The conducted experiments show that $\psi(\mathbf{y})$ yields very good agreement provided that τ is relatively small. It also models the SAR in the presence of local attraction. This also holds well for moderately large N and μ. On the other hand, $\psi(R)$ significantly simplifies the positional dependence (from N components to one). Experiments show its limitations when predicting local attraction effects for very small σ^*, but it yields good results for larger σ^*-values.
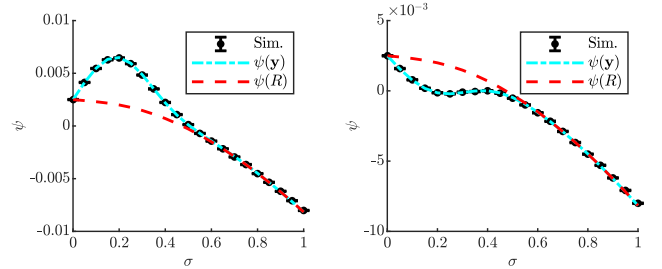
(a) $\psi(\sigma)$ with left and right plots showing different random $\mathbf{y}$-realizations given the same R.
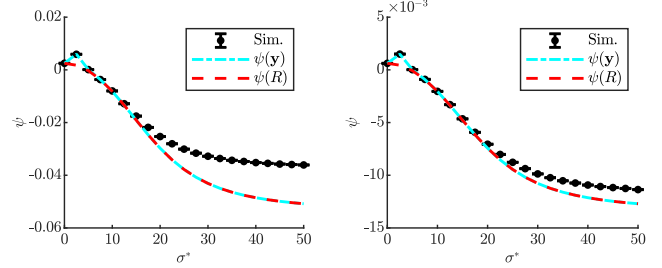
(b) $\psi(\sigma^*)$ as a function of $\sigma^* = \sigma N/R$, see (18), for $\tau = 1/\sqrt{2N}$ (left) and $\tau = 1/\sqrt{8N}$ (right).

Fig. 1: One-generation experiments for ψ with $(100/100_I, 200)$-ES, $A = 10$, $N = 100$, initialized at random $\mathbf{y}$ with $\|\mathbf{y}\| = R = \sqrt{N}$, where local attraction is present. Simulations of (5) are displayed as black dots (vanishing error bars, 10^5 trials) Closed-form result (32) is shown as dash-dotted cyan and (33) as dashed red.

B. Dynamics and τ-Derivation

The one-generation experiments of the last section have shown good agreement with simulations. Now the results for $\psi(\mathbf{y})$ and $\psi(R)$ are tested over many generations using iterations (22) and (23). The experiments in Fig. 2 show the $R(g)$- and $\sigma^*(g)$-dynamics of real optimization runs (Alg. 1) compared to iterated dynamics. The median is taken as a measure of central tendency, as it is more stable w.r.t. outliers compared to the mean. Depending on the chosen A one observes globally (Fig. 2a) and locally converging (Fig. 2b) runs with measured success rate P_S. A characteristic drop (and rise) of $\sigma^*(g)$ is observed. Both iterations show good agreement w.r.t. the measured median dynamics. Despite R-iteration being aggregated over all N components, it shows similar results to $\mathbf{y}$-iteration. Larger deviations occur in Fig. 2b during the phase of local convergence at $g \geq 1000$, where $\sigma^* \to 0$. This would be critical if P_S was evaluated for the iterations. In our case, the focus is set on the investigation of the characteristic σ^*-decrease.

Figure 3 shows the progress landscape $\varphi_R^{\mathrm{II},*}(\sigma^*, R)$, see (13) with $\varphi_R^{\mathrm{II},*} = \varphi_R^{\mathrm{II}} N/(2R^2)$, for the same configurations as shown in Fig. 2 with overlaid dynamics (g is implicitly given). The σ^*-bound was chosen slightly larger than $\sigma_{\varphi_0}^*$ (range of positive progress), see (19), and $R \in [10^{-1}, 10^2]$ to provide good visibility of the relevant characteristics. White regions correspond to high $\varphi_R^{\mathrm{II},*}$ and dark regions to low $\varphi_R^{\mathrm{II},*}$ (zero-progress-line shown as bold white). Thin black lines
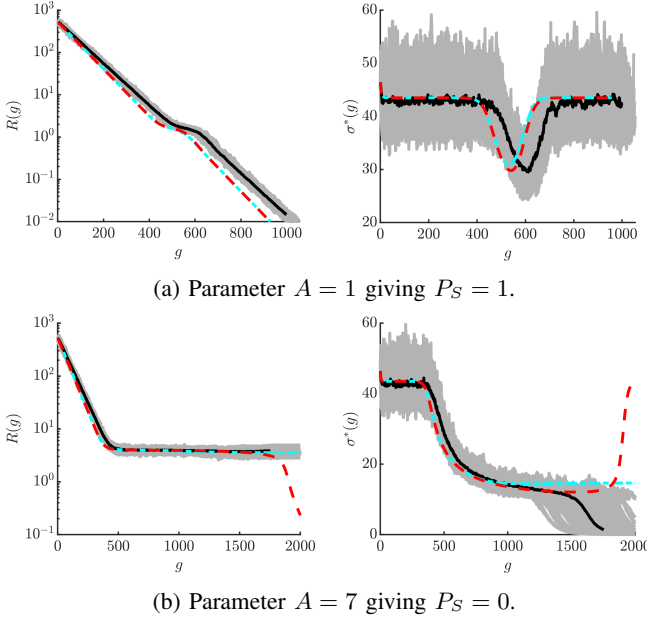
(a) Parameter $A = 1$ giving $P_S = 1$.

(b) Parameter $A = 7$ giving $P_S = 0$.

Fig. 2: Dynamics of R and σ^* (100 trials of Alg. 1) for $(100/100_I, 200)$-ES, $N = 100$, $\tau = 1/\sqrt{8N}$ and varying A (median: solid black; individual trials: solid gray). The cyan dash-dotted line shows (22), and the red dashed line (23). The runs are initialized outside the local attraction region.
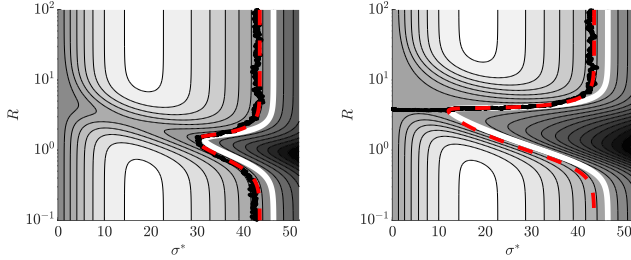

Fig. 3: Progress rate $\varphi_R^{\mathrm{II},*}(\sigma^*, R)$, see Eq. (13), for configuration of Fig. 2 with $A = 1$ (left) and $A = 7$ (right) with overlaid dynamics. The median is displayed as solid black. The iterated dynamics of (23) is shown as dashed red. $\varphi_R^{\mathrm{II},*}(\sigma^*, R)$ and $\psi(\sigma^*, R)$ are displayed in color for $A = 1$ in the supplementary material B-A.

are levels of equal progress. For $R \to \infty$ and $R \to 0$ the sphere function is recovered (see vertical lines indicating scale-invariance). On the left, the ES moves around a characteristic "progress dip" (see also [4]), which leads to the σ^*-reduction. On the right, local convergence occurs as $\sigma^* \to 0$ ($R > 0$). The observed dip is an effect of the presence of many local attractors acting similar to noise during the selection. It slows down the strategy's progress significantly.

On the Rastrigin function there is no σ_{ss}^* independent of R, such that (28) could be applied. There is a high σ^*-change and a steady state σ_{ss}^* only exists in the sphere limits for large and small R, respectively. Choosing slow adaptation $\tau = 1/\sqrt{8N}$ yields a relatively high σ_{ss}^*-level close to $\sigma_{\varphi_0}^*$, which improves the success rate P_S (see also Fig. 5). However, this has the

downside of requiring many generation to decrease σ^* from σ_{ss}^* to reach the critical point of the progress dip location. The critical point is defined as

$$\sigma_{\mathrm{crit}}^* := \min_{0 < \sigma^* < \sigma_{\varphi_0}^*} \sigma^*, \quad \text{such that} \quad \varphi_R^{\mathrm{II},*} = 0. \quad (34)$$

Note that increasing A (see Fig. 3) extends the progress dip, such that σ_{crit}^* decreases further. This has the effect of requiring more generations for the adaptation. Instead of choosing τ arbitrarily small (to maximize P_S), the goal is to derive $\tau = \tau(N, \alpha, A)$ by including information about σ_{crit}^*. To this end, an approximation for σ_{crit}^* will be derived first. Then, one may set $\sigma_{ss}^* = \sigma_{\mathrm{crit}}^*$ in Eq. (28) demanding a steady state on the sphere, i.e., on the global structure of Rastrigin. Finally, Eq. (28) can be solved obtaining a characteristic τ. Choosing $\tau(N, \alpha, A)$ the ES should (approximately) approach the region of the progress dip with $\sigma_{ss}^* \approx \sigma_{\mathrm{crit}}^*$. This should lead to a more efficient search compared to choosing τ arbitrarily small.

First, condition (28) is evaluated for the sphere using progress rate (17) and SAR (33). The derivation is straightforward and can be found in the supplementary material B-B. It relies on setting $A = 0$ in (14) and (15), and evaluating expected derivatives. The obtained expressions are further simplified assuming large populations with $\sqrt{N} \ll \mu$, such that the steady state condition on the sphere yields

$$c_\vartheta \sqrt{2N} - \frac{\sigma_{ss}^{*2}}{2\mu} = -N\tau^2 \left(\frac{1}{2} - c_\vartheta \sqrt{2N} + 2e_\vartheta^{1,1} \right). \quad (35)$$

A closed-form solution for σ_{crit}^* can only be obtained under certain simplifications. The residual distance R at which σ^* decreases (see Fig. 4, right) can be approximated using the Rastrigin maximum noise strength $\sigma_\epsilon^2 := NA^2/2$ as [3]

$$R_\infty^2 = \sigma_\epsilon N/(4c_\vartheta\mu) = N^{3/2}A/(\sqrt{32}c_\vartheta\mu). \quad (36)$$

R_∞ is obtained in [4] using a simplified variance (14) by setting all exponential terms to zero, which gives $\widetilde{D}_Q^2(R) = 4R^2\sigma^2 + 2N\sigma^4 + NA^2/2$. This corresponds to the variance of a sphere under constant noise $\sigma_\epsilon^2 = NA^2/2$. Intuitively speaking, as the ES approaches a large number of local attractors, its selection is similarly disturbed as on the noisy sphere. In this case the variance can be approximated accordingly.

Inserting $\widetilde{D}_Q(R)$ into progress rate (13) and SAR (33) the simplified model is tested in Fig 4 (left). It approximates the σ^*-dip reasonably well. Deriving σ_{crit}^* one therefore demands $\varphi_R^{\mathrm{II},*} = 0$ at $R = R_\infty$. Starting from (13), one sets $\varphi_R^{\mathrm{II},*} = \varphi_R^{\mathrm{II}} N/(2R^2)$, $\sigma = \sigma^* R/N$, uses $\widetilde{D}_Q(R)$, such that after rearranging one obtains the expression

$$1 + \frac{\alpha^2 A}{2} e^{-\frac{\alpha^2 A N^{3/2}}{\sqrt{128}c_\vartheta\mu}\left(\frac{\sigma^{*2}}{N^2} + \frac{1}{N}\right)} = \sqrt{1 + \frac{\sigma^{*2}}{4c_\vartheta^2\mu^2} + \frac{\sigma^{*4}}{8Nc_\vartheta^2\mu^2}}. \quad (37)$$

Solving (37) for σ^* requires further simplifications. As $0 < \sigma^* < \sigma_{\varphi_0}^*$ and $\sigma_{\varphi_0}^* \simeq (8N)^{1/4}(c_\vartheta\mu)^{1/2}$ (see supplementary material B-C), one may assume $\sigma^* = \gamma\sigma_{\varphi_0}^*$ ($0 < \gamma < 1$). Then, one has $\sigma^{*2}/(4c_\vartheta^2\mu^2) = O(\gamma^2\sqrt{N}/\mu)$ and $\sigma^{*4}/(8Nc_\vartheta^2\mu^2) = O(\gamma^4)$. Therefore, the second order term is neglected compared to the fourth order term in (37) for large $\mu \gg \sqrt{N}$. Furthermore, the square-root is expanded

using $\sqrt{1+x} = 1 + x/2 + O(x^2)$ and the term $1/N$ in the exponential function can be dropped compared to σ^{*2}/N^2. Finally, one obtains the simplified expression

$$8\alpha^2 A N c_\vartheta^2 \mu^2 e^{-\frac{\alpha^2 A \sigma^{*2}}{\sqrt{128N}c_\vartheta\mu}} = \sigma^{*4}. \tag{38}$$

Equation (38) can be solved in terms of the Lambert W-function (or product-logarithm). For complex numbers y, z and some integer k it holds

$$z = ye^y, \quad \text{with inverse function} \quad W_k(z) = y. \tag{39}$$

Introducing $a := 8\alpha^2 A N c_\vartheta^2 \mu^2$ and $b := \frac{\alpha^2 A}{\sqrt{128N}c_\vartheta\mu}$ in (38), simple manipulations yield $z = a^{1/2}b/2$ (real-valued $z > 0$), and $y = b\sigma^{*2}/2$, such that with $k = 0$ one solves (38) as

$$\sigma^*_{\text{crit}} = \frac{(512N)^{1/4}(c_\vartheta\mu)^{1/2}}{\alpha A^{1/2}} \sqrt{W_0\left(\frac{\alpha^3 A^{3/2}}{8}\right)}. \tag{40}$$

An exemplary evaluation of (40) is shown in Fig. 4 (right, marked as $*$). Additional parameter variations can be found in the supplementary material B-C. The experiments show good agreement of numerically obtained σ^*_{crit} compared to (40), especially for increasing μ given constant N and A. Setting $\sigma^*_{ss} = \sigma^*_{\text{crit}}$ and inserting (40) into (35) gives

$$\tau^2 = \frac{1}{N} \frac{\frac{8c_\vartheta\sqrt{2N}}{\alpha^2 A}W_0\left(\frac{\alpha^3 A^{3/2}}{8}\right) - c_\vartheta\sqrt{2N}}{1/2 - c_\vartheta\sqrt{2N} + 2e_\vartheta^{1,1}}$$

$$= \frac{1}{N} \frac{1 - \frac{8}{\alpha^2 A}W_0\left(\frac{\alpha^3 A^{3/2}}{8}\right)}{1 - 1/(2c_\vartheta\sqrt{2N}) - 2e_\vartheta^{1,1}/(c_\vartheta\sqrt{2N})}, \tag{41}$$

such that after neglecting $O(1/\sqrt{N})$ in the denominator of (41) one obtains

$$\tau(N, \alpha, A) \simeq \sqrt{\frac{1}{N}\left(1 - \frac{8}{\alpha^2 A}W_0\left(\frac{\alpha^3 A^{3/2}}{8}\right)\right)}. \tag{42}$$

$\tau(N, \alpha, A)$ is real-valued for $\alpha\sqrt{A}/2 > 1.47$ and the limit $\alpha^2 A \to \infty$ of (42) can be easily evaluated as $\tau \simeq 1/\sqrt{N}$ (see supplementary material B-D). This result is interesting, as it suggests choosing a larger τ compared to default $\tau = 1/\sqrt{2N}$ in the limit of high multimodality. Choosing a larger τ corresponds to a less robust but faster search showing (usually) smaller success rates P_S, which is problematic. For a more complete picture the efficiency of the ES has to be investigated.

In Figure 5 result (42) is compared to asymptotic $\tau = 1/\sqrt{N}$ and very small $\tau = 1/\sqrt{8N}$ for two different configurations. The sphere steady state σ^*_{ss} ($R \to \infty, R \to 0$) realized by choosing τ via (42) indeed approaches the region of σ^*_{crit} for both configurations. Deviations are expected to occur to some degree as many approximation steps were necessary. The ES moves around the critical point to increase its progress. As expected, $\tau = 1/\sqrt{8N}$ yields very high P_S with a large mean number of generations $\overline{g}$ to reach the stopping criterion. $\tau(N, \alpha, A)$ yields high P_S with good efficiency in terms of $\overline{g}$. Value $\tau = 1/\sqrt{N}$ is less stable giving $P_S = 0$ left, and high P_S right. However, these experiments were done at fixed population sizes. In the next section, the efficiency is investigated by varying learning parameter τ and population parameters λ and ϑ.
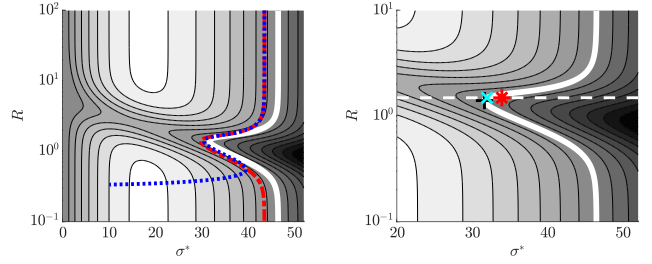

Fig. 4: Plots show configuration of Fig. 2 ($A = 1$). Left plot shows R-iteration using regular variance (14) (dash-dotted red) and simplified variance by setting exponentials to zero (dotted blue). The right plot shows a zoomed-in view with R_∞ (dashed white) and σ^*_{crit} marked (numeric: $+$; $\varphi_R^{\text{II},*}(R_\infty) = 0$: $\times$; result (40): $*$).
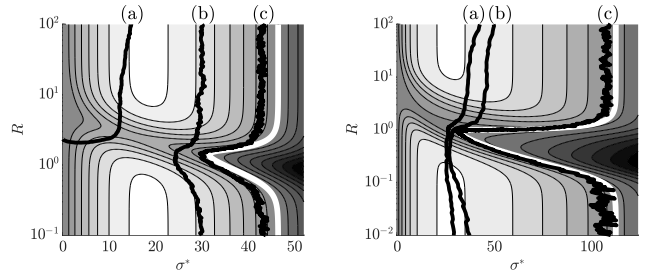

Fig. 5: Left side shows $(100/100_I, 200)$-ES, $N = 100$, $A = 1$, and right side $(500/500_I, 2000)$-ES, $N = 50$, $A = 10$. Both plots show the median over 100 ES-runs with $\tau = 1/\sqrt{N}$ (a), $\tau(N, \alpha, A)$ (b), and $1/\sqrt{8N}$ (c). Success rate P_S and mean number of generations $\overline{g}$ to reach $R = 10^{-4}$ are left ($P_S = 0, 0.92, 0.99$; $\overline{g} = -, 490, 1491$), and right ($P_S = 0.94, 0.97, 1$; $\overline{g} = 175, 185, 1041$). The median is taken over all successful runs (if existing), otherwise over all unsuccessful runs.

VI. EXPECTED RUNTIME EXPERIMENTS

In the previous sections the σSA-ES was analyzed by means of evaluating the self-adaptation response together with the progress rate modeling the ES-dynamics. Theoretical investigations certainly have limitations as $\tau \to 0$ is required in general. Furthermore, multiple approximation steps were necessary to provide closed-form solutions for $\tau(N, \alpha, A)$ at all. This section should complement the previous results by experimentally evaluating the performance of Alg. 1. To this end, the test function parameters A, α, and N will be fixed and the strategy parameters λ, τ, and ϑ will be varied over a larger range. Furthermore, restarts of Alg. 1 will be allowed and the efficiency is measured in terms of necessary functions evaluations to reach global convergence. To this end, the expected runtime is evaluated. It was derived in [13] and yields

$$E_r = \left(\frac{1}{P_S} - 1\right)\frac{F_u}{n_u} + \frac{F_s}{n_s} = \frac{F_u + F_s}{n_s}, \tag{43}$$

with the number of total successful and unsuccessful function evaluations denoted as F_s, F_u, respectively, and the number of successful and unsuccessful trials as n_s and n_u, respectively.

For the subsequent experiments $A = 10$, $\alpha = 2\pi$, and $N = 50$ were chosen with moderately large test function parameters. The varied strategy parameters are $\lambda = 200, 400, ..., 2000$, $\tau = 0.05, 0.07, ..., 0.17$ and $\vartheta = 0.1, 0.2, ..., 0.7$ giving 490 configurations. Each configuration is evaluated using 200 trials, such that P_S and E_r are obtained. An overall budget of function evaluations was set to 10^7 (based on $\lambda = 2000$, $\vartheta = 0.5$, $\tau = 0.05$ requiring $E_r \approx 3.3 \cdot 10^6$ with added headroom).

The initialization and termination of each run was chosen by following argumentation. One wants to assess the efficiency on the Rastrigin problem covering the complete landscape of local attraction, but it is desirable to reduce the effects of optimizing the sphere. Therefore, one sets $\mathbf{y}^{(0)} = 30 \cdot \mathbf{1}$ with $\sigma^{(0)} = \sigma_{\varphi_0}^* \|\mathbf{y}\|/N$, which initializes the ES at the beginning of the local attraction region (for $A = 10$) with comparably large $\sigma^{(0)}$, which reduces the influence of the initial $\mathbf{y}^{(0)}$. For the termination of successful runs one sets $f_{\mathrm{stop}} = 10^{-1}$ for the ES being in the global attractor, as its closest neighboring local attractor corresponds to $f \approx 1$. This prevents overvaluation of the ES optimizing the sphere limits. Unsuccessful runs terminate below $\sigma_{\mathrm{stop}} = 10^{-3}$ (local convergence) or by reaching the maximum number of function evaluations.

Exemplary evaluations of the experiments are shown in Fig. 6. The left plots, respectively, show the success rate $P_S \in [0, 1]$ with contour lines showing levels of equal P_S (white: $P_S > 0.9$, black: $P_S < 0.1$), see legend at the top. The right plots show $\log_{10}(E_r) \in [5, 8]$ displaying the order of magnitude. White regions have no E_r-value as $P_S = 0$. Note that the contour plots use interpolation between the discrete data points, which improves the visual interpretability.

In general, the highest P_S-values are obtained for large λ and small τ, which was expected. However, these values do not necessarily correspond to the lowest E_r. As an example, in Fig. 6a (top) one obtains the lowest E_r at $\tau \in [0.11, 0.15]$. The notable τ-values are $1/\sqrt{N} \approx 0.14$, $\tau(N, \alpha, A) \approx 0.13$, $1/\sqrt{2N} = 0.1$, and $1/\sqrt{8N} = 0.05$. Choosing $\tau(N, \alpha, A)$ agrees well with the region of lowest E_r. The ES does benefit from approaching the region of σ_{crit}^* resulting in a better efficiency. One observes that E_r deteriorates as τ decreases. For very large τ-values the σ-adaptation becomes faster and more unstable, such that for some data points no E_r is obtained. Note that a minimum population $\lambda = 400$ is needed to obtain results at all, while for larger $\lambda \geq 800$ E_r does not change much for given τ.

Figures 6b shows the effect of the truncation ratio ϑ given τ. Ratios $\vartheta \in [0.2, 0.5]$ yield satisfying results, which agrees well with standard values $\vartheta = 0.25$ (close to sphere optimal [7, Sec. 6.3]) and $\vartheta = 0.5$ (e.g. used for the CMA-ES [14]). Values $\vartheta = \{0.1, 0.6\}$ do work, but they are usually less efficient. $\vartheta = 0.7$ does not yield E_r-results within the given budget of function evaluations. Note that configuration $\vartheta = 0.5$, $\tau = 0.05$ is relatively inefficient, but yields high P_S. For a single optimization run it can be recommended due to its robustness. However, its E_r-performance can be improved by either decreasing ϑ or increasing τ. One may conclude that τ and ϑ should not both be chosen too conservatively, as small τ and large ϑ both significantly reduce the overall convergence
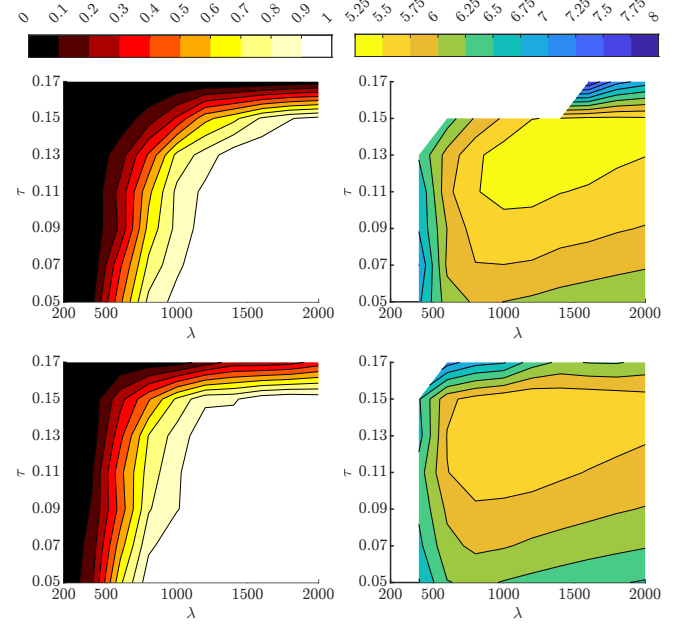
(a) Variation of λ (x-axis) and τ (y-axis) for $\vartheta = 0.3$ (top) and $\vartheta = 0.5$ (bottom).
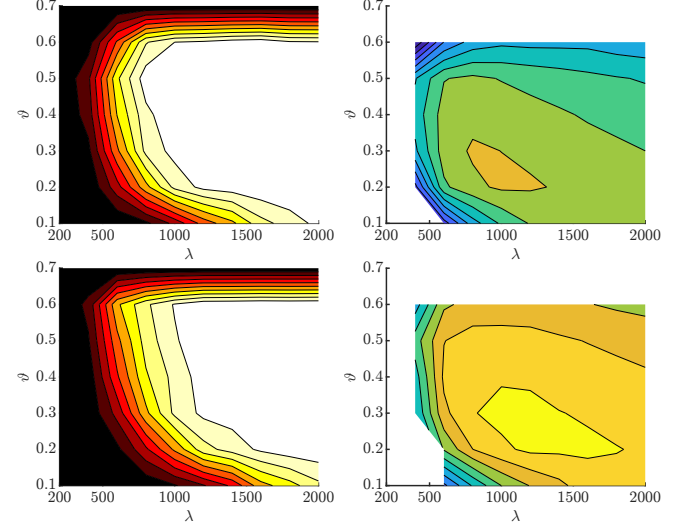
(b) Variation of λ (x-axis) and ϑ (y-axis) for $\tau = 0.05$ (top) and $\tau = 0.11$ (bottom).
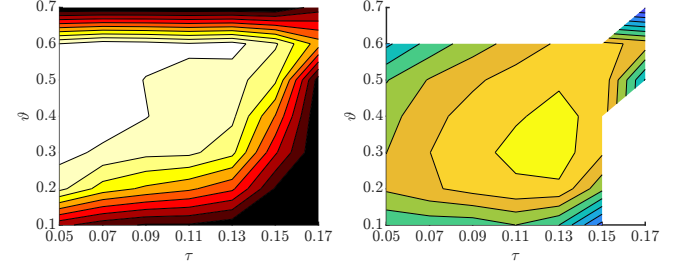
(c) Variation of τ (x-axis) and ϑ (y-axis) for $\lambda = 1000$.

Fig. 6: Variation of λ, τ, and ϑ at $A = 10$ and $N = 50$.

rate requiring more function evaluations.

Figure 6c shows the variation of τ and ϑ for fixed λ. As mentioned before $\tau = 0.05$ is less efficient. The default value $\tau = 1\sqrt{2N} = 0.10$ shows better results. However, it should be

noted that $\tau = 1\sqrt{2N}$ was derived in [8] assuming $N \to \infty$ and $\mu \ll N$, which is clearly violated in this experiment. $\tau(N, \alpha, A) \approx 0.13$ yields very good results in terms of E_r. One observes that even larger $1/\sqrt{N} \approx 0.14$ yields good results (provided that λ is large enough).

The results of a second experiment at $N = 200$, $A = 1$, and $\alpha = 2\pi$ are comparable to Fig. 6 and are given in the supplementary material B-E.

VII. Conclusion and Outlook

In this paper an asymptotically exact self-adaptation response function was derived in (31) for large populations and evaluated on the Rastrigin function. Within a dynamic model of the σSA-ES, progress rate and SAR results were used to investigate the convergence and steady state properties of the ES as a function of the learning parameter τ. As there is no R-independent steady state (as on the sphere), the characteristic progress dip was studied and a model for a critical σ^*_{crit} was derived. Imposing σ^*_{crit} for the (sphere) steady state a relation $\tau(N, \alpha, A)$ could be derived for the Rastrigin function.

Expected runtime experiments show that $\tau(N, \alpha, A)$ does not maximize the success rate P_S compared to smaller choices of τ. Instead, it shows a higher efficiency of the search in terms of function evaluations. Therefore, comparably large $\tau \in [1/\sqrt{2N}, 1/\sqrt{N}]$ may be beneficial on multimodal functions, if restarts are allowed and the population size is sufficiently large. Whether this observation holds for other multimodal test functions with adequate global structure is an open question. On the other hand, the highest robustness (at the expense of efficiency) is achieved for small τ, higher ϑ, and large populations, which was expected. Whether and how the results could be transferred to cumulative step-size adaptation (CSA) is left for future research.

A general important observation is that minimal population sizes are needed to obtain global convergence at all (see also [3]). In black-box optimization there is no knowledge about the degree of multimodality, i.e., parameters α and A, to tune the ES appropriately. Furthermore, allowing algorithm restarts opens up the question of how to re-initialize the ES-parameters. Therefore, adaptive population control appears to be a natural direction for future research.

Acknowledgments

References

[1] H.-G. Beyer and H.-P. Schwefel, "Evolution Strategies: A Comprehensive Introduction," *Natural Computing*, vol. 1, no. 1, pp. 3–52, 2002.

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

[3] L. Schönenberger and H.-G. Beyer, "On a Population Sizing Model for Evolution Strategies Optimizing the Highly Multimodal Rastrigin Function," in *Proceedings of the Genetic and Evolutionary Computation Conference*, ser. GECCO '23, New York, NY, USA, 2023, p. 848–855. [Online]. Available: https://doi.org/10.1145/3583131.3590451

[4] A. Omeradzic and 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*, ser. FOGA '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 117–128. [Online]. Available: https://doi.org/10.1145/3594805.3607126

[5] N. Hansen and A. Ostermeier, "Completely Derandomized Self-Adaptation in Evolution Strategies," *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, 2001, https://doi.org/10.1162/106365601750190398.

[6] I. Rechenberg, *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Stuttgart: Frommann-Holzboog Verlag, 1973.

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

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

[9] H.-G. Beyer and A. Melkozerov, "The Dynamics of Self-Adaptive Multi-Recombinant Evolution Strategies on the General Ellipsoid Model," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 5, pp. 764–778, 2014, DOI: 10.1109/TEVC.2013.2283968.

[10] A. Omeradzic and H.-G. Beyer, "Progress Analysis of a Multi-Recombinative Evolution Strategy on the Highly Multimodal Rastrigin Function," *Theoretical Computer Science*, vol. 978, 2023. [Online]. Available: https://doi.org/10.1016/j.tcs.2023.114179

[11] ——, "Rastrigin Function: Quality Gain and Progress Rate for $(\mu/\mu_I, \lambda)$-ES," Vorarlberg University of Applied Sciences, Report, 2023. [Online]. Available: https://opus.fhv.at/frontdoor/index/index/docId/5151

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

[13] A. Auger and N. Hansen, "Performance Evaluation of an Advanced Local Search Evolutionary Algorithm," in *Congress on Evolutionary Computation, CEC'05*, vol. 2. IEEE, 2005, pp. 1777–1784.

[14] N. Hansen, "The CMA Evolution Strategy: A Tutorial," 2023.

[15] Hoorfar, A. and Hassani, M., "Inequalities on the lambert function and hyperpower function," 2008. [Online]. Available: http://eudml.org/doc/130024

Appendix A
Self-Adaptation Response

Starting from (30) the order statistic density $p_{m;\lambda}(s|\sigma)$ needs to be derived. An offspring samples its mutation strength according to density $p_\sigma(s)$ in (29). Furthermore, a quality gain q is observed with conditional density $p_Q(q|s)$ and CDF $P_Q(q|s)$. The conditional CDF of the quality gain given σ for an individual sampling mutations from $p_\sigma(s)$ is obtained by

$$P_{Q,s}(q|\sigma) = \int_0^\infty P_Q(q|s) p_\sigma(s)\, \mathrm{d}s. \qquad (A.1)$$

For the m-th best individual having quality gain q, there are $m - 1$ better and $\lambda - m$ worse individuals, respectively. The order statistic density for the m-th best individual given σ is obtained by integration over all quality gain values using (A.1)

$$p_{m;\lambda}(s|\sigma) = p_\sigma(s) \frac{\lambda!}{(m-1)!(\lambda - m)!} \int_{q_l}^{q_u} p_Q(q|s)$$
$$\times P_{Q,s}(q|\sigma)^{m-1} (1 - P_{Q,s}(q|\sigma))^{\lambda - m}\, \mathrm{d}q. \quad (A.2)$$

For log-normal mutations (29) and a normally distributed quality gain (7), a closed-form solution of Eq. (A.1) cannot be provided. However, for small learning parameter τ it will be shown by the same argumentation as in Eqs. (A.15) and (A.16) that

$$P_{Q,s}(q|\sigma) = P_Q(q|\sigma) + O\left(\tau^2\right). \qquad (A.3)$$

Inserting (A.3) into (A.2), and the result into (30), one gets by neglecting higher orders of τ and exchanging the sum and integral

$$\psi = \int_0^\infty \left(\frac{s-\sigma}{\sigma}\right) p_\sigma(s) \frac{\lambda!}{\mu} \int_{q_l}^{q_u} p_Q(q|s)$$
$$\times \sum_{m=1}^\mu \frac{P_Q(q|\sigma)^{m-1}(1-P_Q(q|\sigma))^{\lambda-m}}{(m-1)!(\lambda-m)!}\, dq\, ds. \quad \text{(A.4)}$$

Given (A.4) the following identity from [7, (5.14)] is applied

$$\sum_{m=1}^\mu \frac{P(q)^{m-1}[1-P(q)]^{\lambda-m}}{(m-1)!(\lambda-m)!}$$
$$= \frac{1}{(\lambda-\mu-1)!(\mu-1)!}\int_0^{1-P(q)} t^{\lambda-\mu-1}(1-t)^{\mu-1}\, dt, \quad \text{(A.5)}$$

such that (A.4) yields by using (A.5)

$$\psi = \int_0^\infty \left(\frac{s-\sigma}{\sigma}\right) p_\sigma(s) \frac{\lambda!}{(\lambda-\mu-1)!\mu!} \int_{q_l}^{q_u} p_Q(q|s)$$
$$\times \int_0^{1-P_Q(q|\sigma)} t^{\lambda-\mu-1}(1-t)^{\mu-1}\, dt\, dq\, ds. \quad \text{(A.6)}$$

The population dependent factor in (A.6) is rewritten in terms of the beta function B and truncation ratio $\vartheta = \mu/\lambda$ as $\frac{\lambda!}{(\lambda-\mu-1)!\mu!} = \frac{1}{\vartheta} \frac{1}{B(\lambda-\mu,\mu)}$. Furthermore, one can exchange the integral bounds $q_l \le q \le q_u$, and $0 \le t \le 1-P_Q(q|\sigma)$ by applying the inverse CDF P_Q^{-1} to $q = P_Q^{-1}(1-t)$ giving $0 \le t \le 1$, and $q_l \le q \le P_Q^{-1}(1-t|\sigma)$. Refactoring and exchanging the bounds in (A.6) yields

$$\psi = \frac{1}{\vartheta} \int_0^\infty \left(\frac{s-\sigma}{\sigma}\right) \frac{p_\sigma(s)}{B(\lambda-\mu,\mu)} \int_0^1 t^{\lambda-\mu-1}(1-t)^{\mu-1}$$
$$\times \int_{q_l}^{P_Q^{-1}(1-t|\sigma)} p_Q(q|s)\, dq\, dt\, ds. \quad \text{(A.7)}$$

The innermost integration over q can be easily evaluated as

$$\int_{q_l}^{P_Q^{-1}(1-t|\sigma)} p_Q(q|s)\, dq = P_Q\left(P_Q^{-1}(1-t|\sigma)\Big|s\right), \quad \text{(A.8)}$$

where the probability $P_Q(q_l|s) = \Pr(Q \le q_l|s) = 0$ for any lower bound value q_l. Inserting (A.8) into (A.7) yields

$$\psi = \frac{1}{\vartheta} \int_0^\infty \left(\frac{s-\sigma}{\sigma}\right) \frac{p_\sigma(s)}{B(\lambda-\mu,\mu)} \int_0^1 t^{\lambda-\mu-1}(1-t)^{\mu-1}$$
$$\times P_Q\left(P_Q^{-1}(1-t|\sigma)\Big|s\right)\, dt\, ds. \quad \text{(A.9)}$$

The t-integral in (A.9) is now evaluated in the limit $\mu,\lambda \to \infty$ ($\vartheta = \mu/\lambda$) by applying the same method as in the proof of [10, Theorem 1] using the dominated convergence theorem. One can define a sequence $g_\mu := \frac{1}{B(\lambda-\mu,\mu)}\int_0^1 t^{\lambda-\mu-1}(1-t)^{\mu-a} f_{\sigma,s}(t)\, dt$ for $a = 1$, $\mu > a$ and $\lambda(\mu) = \mu/\vartheta$ with $f_{\sigma,s}(t)$ defined for constant values of $\sigma, s \ge 0$ and $0 \le f_{\sigma,s}(t) \le 1$. It can be shown that $|g_\mu(x)| \le 1$ using [10, Eq. (43)]. Furthermore, the s-integral in (A.9) is finite with $\mathrm{E}\left[\frac{s-\sigma}{\sigma}\right] = \tau^2/2$ using (A.16). Therefore, an upper bound of (A.9) is given by $\tau^2/2\vartheta$, such that limit and s-integral can be exchanged

in (A.9). The limit of g_μ yields [10, Eq. (39)] with $f_{\sigma,s} = P_Q(P_Q^{-1}(1-t|\sigma)|s)$

$$\lim_{\substack{\mu,\lambda\to\infty \\ \vartheta=\text{const.}}} g_\mu = f_{\sigma,s}(t)\big|_{t=1-\vartheta} = P_Q\left(P_Q^{-1}(\vartheta|\sigma)\Big|s\right), \quad \text{(A.10)}$$

such that the asymptotic equality for the SAR is obtained

$$\psi \simeq \frac{1}{\vartheta} \int_0^\infty \left(\frac{s-\sigma}{\sigma}\right) p_\sigma(s) P_Q\left(P_Q^{-1}(\vartheta|\sigma)\Big|s\right)\, ds. \quad \text{(A.11)}$$

Now the actual distribution for P_Q and its inverse for P_Q^{-1} are required to further evaluate (A.11). Using the normal approximation (7) one has

$$P_Q(q|s) \simeq \Phi\left(\frac{q - E_Q(s)}{D_Q(s)}\right) = p, \quad \text{(A.12)}$$

yielding a probability value p. The inverse P_Q^{-1} is obtained using the inverse of the normal distribution $\Phi^{-1}(p)$ given probability p and mutations strength σ

$$q = E_Q(\sigma) + D_Q(\sigma)\Phi^{-1}(p). \quad \text{(A.13)}$$

From now on we use the abbreviations $E(s) := E_Q(s)$, $D(s) := D_Q(s)$ and $\Phi_\vartheta^{-1} := \Phi^{-1}(\vartheta)$ for brevity. Inserting (A.13) into (A.12) and the result into (A.11) yields

$$\psi \simeq \frac{1}{\vartheta} \int_0^\infty \left(\frac{s-\sigma}{\sigma}\right) p_\sigma(s)$$
$$\times \Phi\left(\frac{E(\sigma) + D(\sigma)\Phi_\vartheta^{-1} - E(s)}{D(s)}\right) ds. \quad \text{(A.14)}$$

A closed-form solution of (A.14) cannot be given as the functional dependencies of $E(s)$, see (8), and $D(s)$, see (9), are too complex. However, for $\tau \to 0$ a solution can be given in terms of a Taylor expansion around σ by neglecting higher order terms of $\left(\frac{s-\sigma}{\sigma}\right)^k$. In [8, Appendix C.2] it was shown that for $k \ge 3$ one has $\mathrm{E}\left[\left(\frac{s-\sigma}{\sigma}\right)^k\right] = O(\tau^4)$. Therefore, the expected value of a function f w.r.t. $p_\sigma(s)$ is expanded in terms of factor $\Delta_s^k := (s-\sigma)^k/\sigma^k$ and truncated for $k \ge 3$. Using abbreviation $\partial^k f/\partial s^k |_{s=\sigma} =: f_\sigma^{(k)}$ one gets

$$\mathrm{E}[f(s)] = \int_0^\infty f(s) p_\sigma(s)\, ds$$
$$= \int_0^\infty \sum_{k=0}^\infty \frac{\sigma^k}{k!} \frac{\partial^k f}{\partial s^k}\bigg|_{s=\sigma} \Delta_s^k p_\sigma(s)\, ds$$
$$= \int_0^\infty \left[f_\sigma^{(0)} + \sigma f_\sigma^{(1)}\Delta_s + \frac{\sigma^2 f_\sigma^{(2)}}{2}\Delta_s^2 + O(\Delta_s^3)\right] p_\sigma(s)\, ds$$
$$\simeq f_\sigma^{(0)} + \sigma f_\sigma^{(1)}\,\mathrm{E}[\Delta_s] + \frac{\sigma^2 f_\sigma^{(2)}}{2}\,\mathrm{E}[\Delta_s^2] + O(\tau^4). \quad \text{(A.15)}$$

Using $\mathrm{E}[s^k] = \sigma^k e^{\frac{1}{2}k^2\tau^2}$ for a log-normal variate and expansion $e^{\frac{1}{2}\tau^2} = 1 + \frac{1}{2}\tau^2 + O(\tau^4)$, one has

$$\mathrm{E}[\Delta_s] = \tau^2/2 + O(\tau^4), \quad \mathrm{E}[\Delta_s^2] = \tau^2 + O(\tau^4). \quad \text{(A.16)}$$

In (A.3) only the zeroth order term of (A.15) is used to simplify the CDF neglecting $O(\tau^2)$. Expression (A.14) already contains Δ_s, such that for $O(\tau^4)$ one needs to evaluate $f_\sigma^{(0)}$

and $f_\sigma^{(1)}$ using for $f(s) = \Phi(\dots)$. One obtains by denoting the derivative w.r.t. σ with prime symbol $'$

$$f_\sigma^{(0)} = \Phi\left(\frac{E(\sigma) + D(\sigma)\Phi_\vartheta^{-1} - E(\sigma)}{D(\sigma)}\right) = \vartheta$$

$$f_\sigma^{(1)} = \left.\frac{\partial f}{\partial s}\right|_{s=\sigma} = -\frac{e^{-\frac{1}{2}[\Phi_\vartheta^{-1}]^2}}{\sqrt{2\pi}}\frac{E'(\sigma) + \Phi_\vartheta^{-1}D'(\sigma)}{D(\sigma)}.$$

(A.17)

Using (A.16) and (A.17) in (A.15), and dropping $O\left(\tau^4\right)$-terms, the SAR yields

$$\psi \simeq \frac{1}{\vartheta}\left(\vartheta\frac{1}{2}\tau^2 - \sigma\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}[\Phi_\vartheta^{-1}]^2}\frac{E'(\sigma) + \Phi_\vartheta^{-1}D'(\sigma)}{D(\sigma)}\tau^2\right)$$

$$= \tau^2\left[\frac{1}{2} - \sigma\frac{e^{-\frac{1}{2}[\Phi_\vartheta^{-1}]^2}}{\sqrt{2\pi}\vartheta}\frac{E'(\sigma)}{D(\sigma)} - \sigma\frac{\Phi_\vartheta^{-1}}{\sqrt{2\pi}\vartheta}e^{-\frac{1}{2}[\Phi_\vartheta^{-1}]^2}\frac{D'(\sigma)}{D(\sigma)}\right].$$

(A.18)

In (A.18) the asymptotic generalized progress coefficient $e_\vartheta^{a,b}$ can be recognized from (11) as

$$e_\vartheta^{1,0} = c_\vartheta = \frac{e^{-\frac{1}{2}[\Phi_\vartheta^{-1}]^2}}{\sqrt{2\pi}\vartheta}, \quad e_\vartheta^{1,1} = -\frac{\Phi_\vartheta^{-1}e^{-\frac{1}{2}[\Phi_\vartheta^{-1}]^2}}{\sqrt{2\pi}\vartheta}.$$

(A.19)

We finally arrive at the SAR in the limit of $N \to \infty$, $\tau \to 0$, and $(\mu, \lambda) \to \infty$ (constant $\vartheta = \mu/\lambda$) as

$$\psi \simeq \tau^2\left(\frac{1}{2} - c_\vartheta\sigma\frac{E'_Q(\sigma)}{D_Q(\sigma)} + e_\vartheta^{1,1}\sigma\frac{D'_Q(\sigma)}{D_Q(\sigma)}\right).$$

(A.20)

APPENDIX B
SUPPLEMENTARY MATERIAL

A. Progress Rate and SAR Landscapes

Figure 7 shows the result of Fig. 3 (left plot, $A = 1$) in color. Furthermore, the corresponding SAR is displayed. Bold white lines denote the respective zeros of $\varphi_R^{\mathrm{II},*}$ and ψ. Regions of high progress are shown in yellow, while regions of low (negative) progress are dark blue. As τ is relatively small, the median dynamics of the ES achieves relatively high σ^*-levels. In the sphere limits ($R \to \infty, R \to 0$) one observes σ^* close to $\sigma_{\varphi_0}^* \approx 47$. During the transitional phase σ^* stays close to the progress dip. In the sphere limits one observes vertical lines indicating scale-invariant progress rate and SAR. In this limit the sphere steady state condition (28) is valid and convergence occurs for $\varphi_R^{\mathrm{II},*}(\sigma^*, R) > 0$ and $\psi < 0$ (satisfied in the plot). Along the progress dip ψ changes significantly as σ^* is reduced to achieve positive progress.
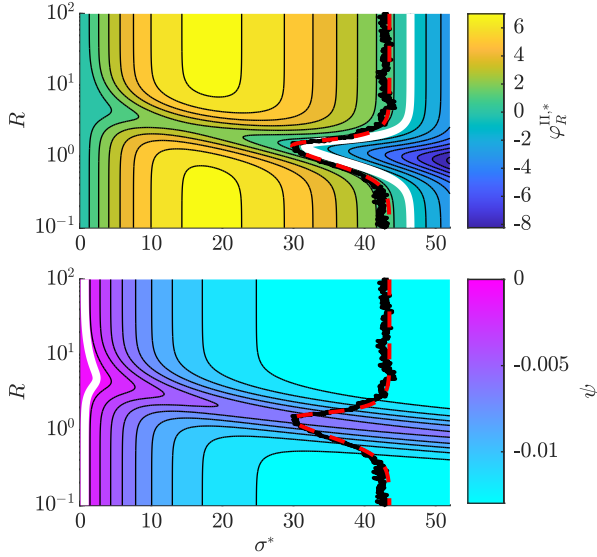

Fig. 7: The top plot shows the progress rate $\varphi_R^{\mathrm{II},*}(\sigma^*, R)$, see Eq. (13), and the bottom plot the SAR function $\psi(\sigma^*, R)$, see (33), for the $(100/100_I, 200)$-ES, $N = 100$, $\tau = 1/\sqrt{8N}$, with overlaid median dynamics of Alg. 1 (black) and iterated dynamics of Eq. (23) (dashed red).

B. Steady State Condition

The steady state condition (35) for the sphere starting from (28) is derived. The progress rate is already given in (17). The SAR on the sphere requires setting $A = 0$ in $E_Q(R)$, see (15), and $D_Q^2(R)$, see (14), and evaluating respective derivatives. One gets for the sphere after using $\sigma = \sigma^* R / N$

$$E_Q = N\sigma^2$$
$$E_Q' = \frac{dE_Q}{d\sigma} = 2N\sigma = 2N\frac{\sigma^* R}{N}, \tag{B.1}$$

and for the variance terms

$$D_Q^2 = 4R^2\sigma^2 + 2N\sigma^4 = 4R^2\left(1 + \frac{\sigma^{*2}}{2N}\right)$$
$$D_Q' = \frac{d(D_Q^2)}{d\sigma}\frac{1}{2D_Q} = \frac{8R^2\sigma + 8N\sigma^3}{2(4\sigma^2 R^2 + 2N\sigma^4)^{1/2}}$$
$$= \frac{2R\left(1 + \frac{\sigma^{*2}}{N}\right)}{\sqrt{1 + \frac{\sigma^{*2}}{2N}}}. \tag{B.2}$$

Using (B.1) and (B.2) in SAR-function (31) yields

$$\psi_{\mathrm{sph}}(\sigma^*) = \tau^2\left(\frac{1}{2} - c_\vartheta \frac{\sigma^*}{\sqrt{1 + \frac{\sigma^{*2}}{2N}}} + e_\vartheta^{1,1}\frac{1 + \frac{\sigma^{*2}}{N}}{1 + \frac{\sigma^{*2}}{2N}}\right). \tag{B.3}$$

Now one inserts (17) and (B.3) into (28) and sets $\sigma^* = \sigma_{ss}^*$ (steady state), which gives

$$\frac{c_\vartheta \sigma_{ss}^*}{\sqrt{1 + \frac{\sigma_{ss}^{*2}}{2N}}} - \frac{\sigma_{ss}^{*2}}{2\mu}$$
$$= -N\tau^2\left(\frac{1}{2} - \frac{c_\vartheta \sigma_{ss}^*}{\sqrt{1 + \frac{\sigma_{ss}^{*2}}{2N}}} + e_\vartheta^{1,1}\frac{1 + \frac{\sigma_{ss}^{*2}}{N}}{1 + \frac{\sigma_{ss}^{*2}}{2N}}\right). \tag{B.4}$$

Condition (B.4) has several important properties [1]. In the limit $N \to \infty$ (small σ^* as $\mu \ll N$) the terms simplify significantly by dropping $O\left(\sigma^{*2}/N\right)$ giving $c_\vartheta \sigma_{ss}^* - \frac{\sigma_{ss}^{*2}}{2\mu} = -N\tau^2(\frac{1}{2} - c_\vartheta \sigma_{ss}^* + e_\vartheta^{1,1})$, for which (after replacing c_ϑ by $c_{\mu/\mu,\lambda}$) the default $\tau = 1/\sqrt{2N}$ was derived in [8, Sec. 4.1]. In the case of the Rastrigin function the assumption of $\mu \ll N$ does not hold, as μ must be chosen relatively large to obtain high success rates. Additionally, choosing large μ rescales the steady state σ^*, see scaling of second zero (B.7) w.r.t. μ. Therefore, by assuming $\mu \to \infty$, constant N, and the scaling (B.7), one can simplify the last term of ψ by evaluating $\lim_{\sigma^* \to \infty}(1 + \frac{\sigma_{ss}^{*2}}{N})/(1 + \frac{\sigma_{ss}^{*2}}{2N}) = 2$. Furthermore, one has $\sqrt{1 + \sigma_{ss}^{*2}/2N} \simeq \sigma_{ss}/\sqrt{2N}$. Under these assumptions (B.4) simplifies giving

$$c_\vartheta \sqrt{2N} - \frac{\sigma_{ss}^{*2}}{2\mu} = -N\tau^2\left(\frac{1}{2} - c_\vartheta \sqrt{2N} + 2e_\vartheta^{1,1}\right). \tag{B.5}$$

---

[1] For condition (B.4) it should be noted that a closed-form solution for σ_{ss}^* is in principle possible, but due to its very lengthy result practically not usable. The same observation holds for the optimal $\hat{\sigma}_{ss}^*$ maximizing the progress rate in (17), which (technically) can be obtained by evaluating $\frac{d\varphi^*}{d\sigma^*} = 0$, but yields very lengthy results.

C. Deriving and Evaluating σ^*_{crit}

First, the second zero of the progress rate is simplified. Starting from (19), the second zero for $\mu^2 \gg N$ yields

$$
\begin{aligned}
\sigma^*_{\varphi_0} &= \left[N \left(1 + \frac{8c_\vartheta^2 \mu^2}{N} \right)^{1/2} - N \right]^{1/2} \\
&\simeq \left[N \left(\frac{8c_\vartheta^2 \mu^2}{N} \right)^{1/2} - N \right]^{1/2} \\
&= \left[(8N)^{1/2} c_\vartheta \mu - N \right]^{1/2} \qquad \text{(B.6)} \\
&= (8N)^{1/4} (c_\vartheta \mu)^{1/2} \left[1 - \frac{N}{(8N)^{1/2} c_\vartheta \mu} \right]^{1/2} \\
&= (8N)^{1/4} (c_\vartheta \mu)^{1/2} \left[1 - O\left(\frac{\sqrt{N}}{\mu} \right) \right],
\end{aligned}
$$

which yields after neglecting $O(\sqrt{N}/\mu)$

$$
\sigma^*_{\varphi_0} \simeq (8N)^{1/4} (c_\vartheta \mu)^{1/2}. \qquad \text{(B.7)}
$$

Equation (B.7) is then used to neglect terms in Eq. (37) and provide a closed-form solution of σ^*_{crit}.

In Fig. 8, experiments regarding σ^*_{crit} from Eq. (40) are conducted. To this end, μ, N, and A are varied, respectively. The numeric solution serves as a reference. Approximation (40) yields comparably good results considering the simplifications necessary to obtain a closed-form solution. Especially good results are obtained for large μ (top plot) due to suppressed terms in the square-root of (37). Increasing N at fixed μ deteriorates the result to some extent, as additional correction terms (the second order σ^{*2}-term) would be required to improve the result. The additional term, however, prevents a closed-form solution in terms of the W-function, see (38). Larger deviations occur for varying A, which is not surprising. Parameter A directly affects σ^*_{crit}, see Eq. (40) and Fig. 3, where for larger A one observes a significant decrease of σ^*_{crit} (larger extent of the progress dip). Furthermore, A is contained in the variance terms (14), which were mostly neglected. From the iteration shown in Fig. 4 one expects that the analytic solution should overestimate the numeric solution (see dip minima locations). Indeed, the analytic solution (magenta) lies consistently above the numeric solution in the bottom plot of Fig. 8. However, the overall scaling of approximation (40) w.r.t. A can be regarded as satisfactory. The polynomial solution (expanding Eq. (38) including terms $O(\sigma^{*2})$ and solving) significantly underestimates the expected results in all plots. Expanding the exponential function including $O(\sigma^{*4})$ did not improve the results, but gave mostly complex-valued solutions. This indicates that the exponential function is crucial to obtain acceptable results.

D. Properties of $\tau(N, \alpha, A)$

Result (42) has important properties, which are discussed now. To this end, the substitution $x = \alpha\sqrt{A}/2$ is introduced, such that one has

$$
\tau = \sqrt{\frac{1}{N} \left(1 - \frac{2}{x^2} W_0(x^3) \right)}. \qquad \text{(B.8)}
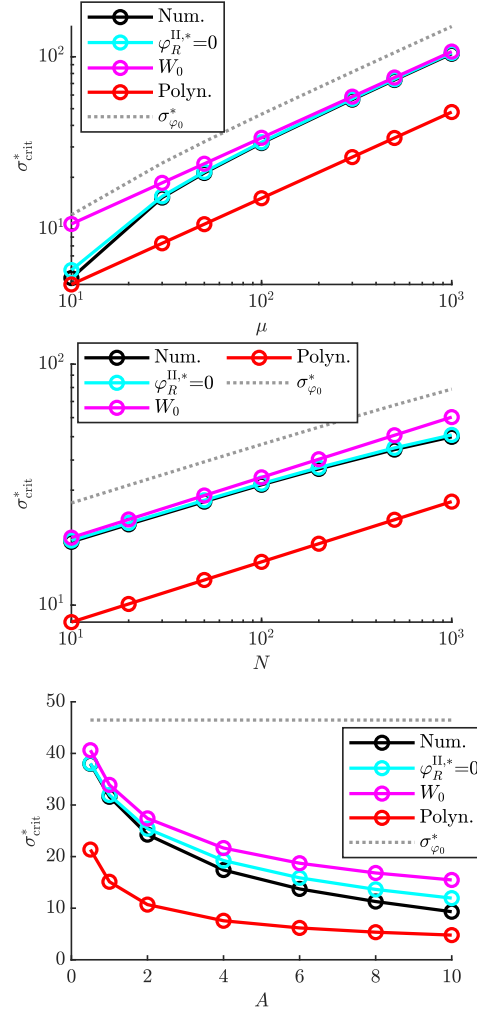$$

Fig. 8: Result for σ^*_{crit} evaluated for top ($N = 100$, $A = 1$, μ varied with $\vartheta = 0.5$), center ($\mu = 100$, $\lambda = 200$, $A = 1$, N varied), and bottom ($\mu = 100$, $\lambda = 200$, $N = 100$, A varied). The numeric evaluation using $\varphi_R^{II,*}(\sigma^*, R)$ (right plot of Fig. 4) is shown in black and serves as a reference. The cyan curve shows the numeric solution of $\varphi_R^{II,*}|_{R=R_\infty} = 0$ including all terms. The magenta curve shows result (40), and the red curve shows the solution of fourth order polynomial from (38) by expanding $e^{-b\sigma^{*2}} = 1 - b\sigma^{*2} + O(\sigma^{*4})$. Note that the upper two plots use logarithmic axes.

Demanding $1 > \frac{2}{x^2} W_0(x^3)$ in (B.8) for a real-valued solution, one obtains by numeric solving

$$
x_0 \approx 1.47, \quad \text{such that} \quad 1.47 < \frac{\alpha\sqrt{A}}{2}. \qquad \text{(B.9)}
$$

This condition is fulfilled for the standard (and larger) choices of α and A. Figure 9 shows $2W_0(x^3)/x^2$ as a function of x. The vertical line indicates x above which there exists a real solution for τ. Note that the expression vanishes for large x. The limit $x \to \infty$ is evaluated now. An upper bound of $W_0(x)$ for $x \geq e$ is given by $W_0(x) \leq \ln(x) - \ln(\ln(x)) + \frac{e}{e-1} \frac{\ln(\ln(x))}{\ln(x)}$
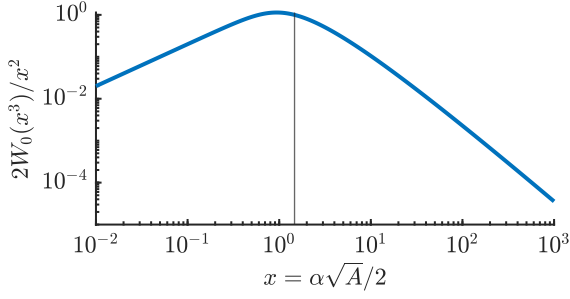
Fig. 9: W_0-dependence of τ from (B.8). The vertical line marks condition (B.9).

[15], such that

$$W_0(x^3) \leq \ln(x^3) - \ln(\ln(x^3)) + \frac{e}{e-1}\frac{\ln(\ln(x^3))}{\ln(x^3)}$$

$$= 3\ln(x) - \ln(3\ln(x)) + \frac{e}{e-1}\frac{\ln(3\ln(x))}{3\ln(x)}. \tag{B.10}$$

The limit w.r.t. the upper bound vanishes as

$$\lim_{x\to\infty} \frac{2}{x^2}\left(3\ln(x) - \ln(3\ln(x)) + \frac{e}{e-1}\frac{\ln(3\ln(x))}{3\ln(x)}\right) = 0. \tag{B.11}$$

In the asymptotic limit $x \to \infty$, Eq. (B.8) therefore yields

$$\tau \simeq \sqrt{\frac{1}{N}}. \tag{B.12}$$

The limit $x \to 0$ is not useful from a modeling perspective, as for $\alpha\sqrt{A} \to 0$ the Rastrigin function becomes the sphere and σ_{crit}^* does not exists. Additionally, the residual distance R_∞ (36) vanishes for the sphere.

E. Additional Expected Runtime Experiment

For the experiment shown in Fig. 10 parameters $A = 1$, $\alpha = 2\pi$, and $N = 200$ were chosen. The varied strategy parameters are $\lambda = 100, 200, ..., 1000$, $\tau = 0.02, 0.03, ..., 0.08$, and $\vartheta = 0.1, 0.2, ..., 0.6$ giving 420 configurations. Each configuration is evaluated using 200 trials, such that P_S and E_r are obtained. An overall budget of function evaluations was set to $5\cdot 10^6$. The legend is shown at the top with P_S on the left and $\log_{10}(E_r)$ on the right. The initialization was set to $\mathbf{y}^{(0)} = 3 \cdot \mathbf{1}$, $\sigma^{(0)} = \sigma_{\varphi_0}^* \|\mathbf{y}\|/N$ and the termination criteria are the same as in Sec. VI.

The overall results are comparable to the experiment in Sec. VI. The lowest E_r is obtained for moderate $\tau \approx 0.5$, which agrees well with Eq. (42) giving $\tau(N, \alpha, A) \approx 0.0495$. The default $\tau = 1/\sqrt{2N} = 0.05$ is also very close to $\tau(N, \alpha, A)$ for this configuration. Large values of τ tend to be more unstable, while small values are less efficient. Again, the highest success rates are obtained for larger ϑ and small τ, which is not surprising. Furthermore, sphere-optimal truncation ratios around $\vartheta = 0.25$ tend to be more efficient.
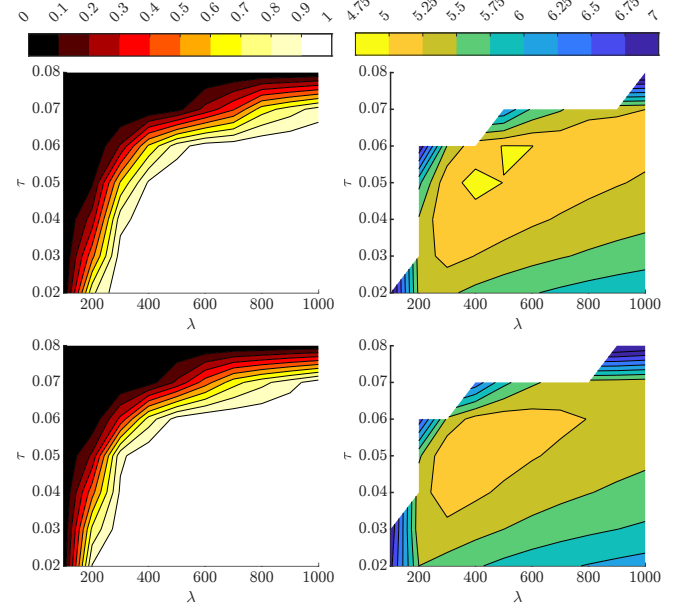
(a) Variation of λ (x-axis) and τ (y-axis) for $\vartheta = 0.3$ (top) and $\vartheta = 0.5$ (bottom).
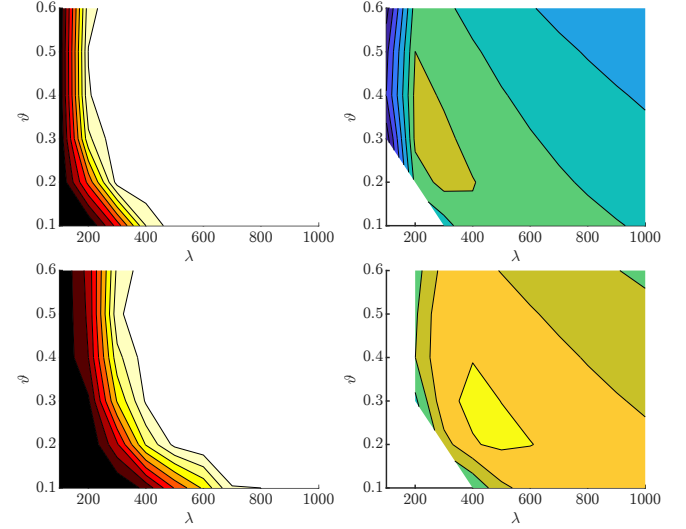
(b) Variation of λ (x-axis) and ϑ (y-axis) for $\tau = 0.02$ (top) and $\tau = 0.05$ (bottom).
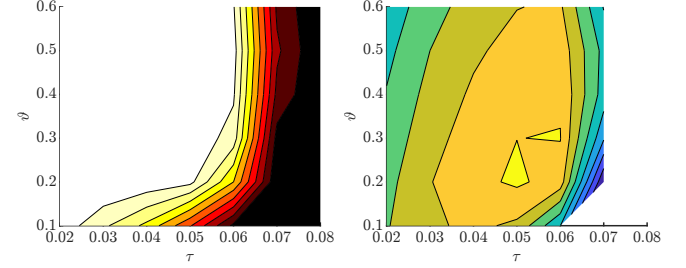
(c) Variation of τ (x-axis) and ϑ (y-axis) for $\lambda = 500$.

Fig. 10: Variation of λ, τ, and ϑ at $A = 1$ and $N = 200$.