On the Performance of $(1, \lambda)$-Evolution Strategies for the Ridge Function Class

Hans-Georg Beyer
University of Dortmund
Department of Computer Science XI
D-44221 Dortmund, Germany
beyer@zappa.cs.uni-dortmund.de

Abstract

This paper presents the N-dependent analysis of the $(1, \lambda)$-evolution strategy (ES) with isotropic mutations for the ridge functions including the special cases of sharp and parabolic ridges. The new approach presented allows for the prediction of the dynamics in ridge direction as well as in radial direction. The central quantities are the corresponding progress rates which are determined in terms of analytical expressions. Its predictive quality is evaluated by ES simulations and the steady-state behavior is discussed in detail.

Keywords

Evolution strategies, induced order statistics, performance measures, progress rates, ridge functions, scalability

I. Introduction

Up until now, the theoretical analysis of the performance of evolution strategies (ES) with Gaussian mutations has been concentrated mainly on the sphere model test functions [1], [2], [3], [4]. Even though the results obtained provide valuable insight into the working of ES algorithms [5], [6], there is still a need for considering further model classes in order to acquire to a deeper understanding of why and how these algorithms really work. Especially when self-adaptation is considered, the sphere model does not cover all essential aspects of the local evolution process. That is, such algorithms can even locally exhibit qualitatively different behavior on test functions which cannot be well approximated by the sphere model. Such a class of *simple* test functions has been investigated empirically by Herdy [7]: the so-called "parabolic ridge" and the "sharp ridge" (see also [8]).

The ridge models may be regarded as extensions of the sphere model breaking its total rotational symmetry in one dimension of the parameter space (for definitions, see below). One might expect that such a small change in the functional structure would not have a severe influence on the performance of the ES. However, the change is of such a kind that each level set of the fitness landscape is an open success domain, and the ridge axis direction appears as a progress direction in which the population can evolve indefinitely.

From the technical point of view, the ridge function class is the one that should be considered *after* the sphere model. In the latter, all N dimensions can be lumped together, thus opening up the possibility for a one-dimensional description of the ES-dynamics. The logical next step is therefore to consider models whose dynamics must be described by two state variables in the parameter space (search space). While this appears logically cogent, a first paper on this topic dates back to 1998 [9]. In that paper, Oyman et al. developed a simple local geometrical model in order to calculate the expected progress, the progress rate φ, in ridge direction for the $(1, \lambda)$-ES given the state of the parent. But this ad hoc approach lacks in some aspects. First, it does not provide any information on the approximation error made. The influence of the parameter space dimension N remained obscure. Second, as a more severe aspect, this model is not able to address the problem of the radial dynamics, i.e., the evolution of the parental distance r to the ridge axis.

The analysis of the dynamics of r^2 for the special case "parabolic ridge" succeeded thereafter in [10], and a thorough analysis of the parabolic ridge for $N \to \infty$ and different ES versions has been done by Oyman [11]. Still there remains the treatment of the radial dynamics for general ridge functions.

This paper provides an approach for calculating the radial as well as the longitudinal (ridge direction) dynamics for $(1, \lambda)$-ES on general ridge functions, thereby closing the remaining gap. The paper is organized as follows.

Section II defines the general rotated ridge function class and its transformation to the normal form. In a third subsection the melioration process on the ridge is discussed and the connection to optimization is established. Subsection II-D gives a short description of the ES algorithm, and in the last subsection the local performance measures are introduced.

Section III is devoted to the progress rate φ_x in ridge direction and Section IV to the progress rate φ_r toward the ridge axis. Both sections are of technical nature. First, the integral representations for φ_x and φ_r, respectively, are derived. Unfortunately, these integrals are not tractable. Therefore, analytical expressions based on normal approximations and linearization techniques must be calculated. The applicability of the approximations used is shown for some examples by comparison with experiments in Section IV-C.

The dynamical aspects and the steady-state behavior of the $(1, \lambda)$-ES are discussed in Section V. In the first part the predictions regarding the dynamics are compared with real ES runs. The most important observation is the appearance of a steady-state behavior keeping the population at a certain (expected) distance r_∞ to the ridge axis. This r_∞ distance is investigated further and the transient time for its appearance is estimated. Finally, the steady-state progress rate is investigated and compared with experiments.

The closing section gives a short outlook at what should be done next.

II. Ridge functions, the $(1, \lambda)$-ES, and progress measures

A. The ridge function family – general definition

The ridge functions $F_R(\mathbf{y})$ have been introduced in order to evaluate the optimization performance of self-adaptive ESs on the problem

$$F_R(\mathbf{y}) \to \mathrm{Max}, \qquad F_R \in \mathbb{R}^1, \ \ \mathbf{y} \in \mathbb{R}^N \tag{1}$$

in N-dimensional real-valued search spaces. One feasible definition of F_R is given by

$$F_R(\mathbf{y}) = \mathbf{v}^{\mathrm{T}}\mathbf{y} - d\left[\sqrt{[(\mathbf{v}^{\mathrm{T}}\mathbf{y})\mathbf{v} - \mathbf{y}]^2}\right]^\alpha, \tag{2}$$

with

$$\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1. \tag{3}$$

Here, $\mathbf{y}$ is the object parameter vector $\mathbf{y} \in \mathbb{R}^N$ and $\mathbf{v}$ defines the so-called ridge direction in which the population is expected to move. The exponent α determines the degree of the ridge function. The case $\alpha = 2$ leads to the so-called *parabolic* ridge, whereas $\alpha = 1$ is known as the *sharp* ridge. The parameter $d \geq 0$ influences the problem difficulty given fixed α and N. Larger d values usually lead to worse performance (see below). For the limit case $d \to 0$, F_R degenerates to a linear fitness landscape, also known as hyperplane, whereas the opposite limit $d \to \infty$ yields something reminiscent of an $(N-1)$-dimensional sphere model. Figure 1 shows two examples of (2) displayed as isofitness plots in two dimensions.

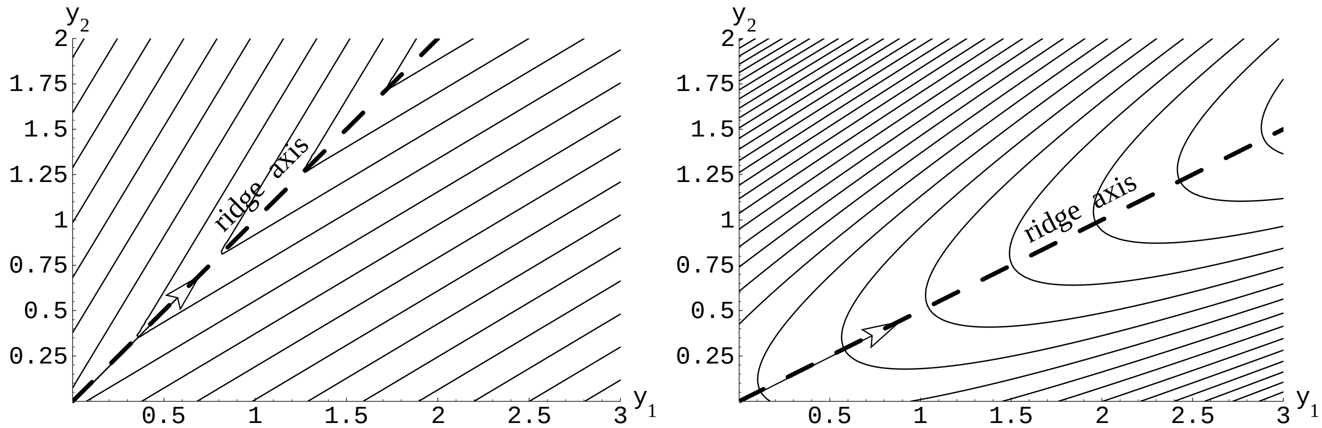

Fig. 1. Curves of constant fitness values (isofitness lines) of the F_R function (2). The ridge axes are indicated by the bold, dashed lines and the $\mathbf{v}$ vectors by arrows. Left picture: sharp ridge ($\alpha = 1$) with $N = 2$, $\mathbf{v} = (1/\sqrt{2}, 1/\sqrt{2})^{\mathrm{T}}$, and $d = 4$. Right picture: parabolic ridge ($\alpha = 2$) with $N = 2$, $\mathbf{v} = (2/\sqrt{5}, 1/\sqrt{5})^{\mathrm{T}}$, and $d = 4$.

B. Transformation to the normal form

Generally, the performance of the ES depends on the direction of $\mathbf{v}$. However, in this paper the investigation will be restricted to ES variants that fulfill the isotropy condition. That is, their performance does not change under linear coordinate transformations. Therefore, the analysis to be presented can be performed on the "normal form" of ridge functions. The normal form is obtained by a coordinate rotation turning the ridge axis $\mathbf{v}$ into the direction of a coordinate axis x_1. This can be accomplished by the linear orthonormal transformation

$$\mathbf{y} = x_1\mathbf{v} + \sum_{i=2}^{N} x_i\mathbf{w}_i = (\mathbf{v}, \mathbf{w}_2, \ldots, \mathbf{w}_N) \cdot (x_1, x_2, \ldots, x_N)^{\mathrm{T}} \tag{4}$$

with

$$\forall i \geq 2: \quad \mathbf{v}^{\mathrm{T}}\mathbf{w}_i = 0 \tag{5}$$

and

$$\forall i \geq 2: \quad \mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j = \delta_{ij}. \tag{6}$$

This can be proven easily by inserting (4) into (2) taking (3), (5), and (6) into account

$$\begin{aligned}
F_R(\mathbf{y}) &= F_R(\mathbf{y}(\mathbf{x})) = \tilde{F}_R(\mathbf{x}) \\
\tilde{F}_R(\mathbf{x}) &= \mathbf{v}^{\mathrm{T}}x_1\mathbf{v} + \sum_{i=2}^{N} x_i\mathbf{v}^{\mathrm{T}}\mathbf{w}_i - d\left[\sqrt{\left[\left(\mathbf{v}^{\mathrm{T}}x_1\mathbf{v} + \sum_{i=2}^{N} x_i\mathbf{v}^{\mathrm{T}}\mathbf{w}_i\right)\mathbf{v} - x_1\mathbf{v} - \sum_{i=2}^{N} x_i\mathbf{w}_i\right]^2}\,\right]^{\alpha} \\
&= x_1 - d\left[\sqrt{\left[(x_1+0)\mathbf{v} - x_1\mathbf{v} - \sum_{i=2}^{N} x_i\mathbf{w}_i\right]^2}\,\right]^{\alpha} \\
&= x_1 - d\left[\sqrt{\left(\sum_{i=2}^{N} x_i\mathbf{w}_i\right)^2}\,\right]^{\alpha} \\
&= x_1 - d\left[\sqrt{\sum_{i=2}^{N}\sum_{j=2}^{N} x_i x_j \mathbf{w}_i\mathbf{w}_j}\,\right]^{\alpha} \\
&= x_1 - d\left[\sqrt{\sum_{i=2}^{N} x_i^2}\,\right]^{\alpha}
\end{aligned} \tag{7}$$

Thus, the ridge function takes the simple form

$$\tilde{F}_R = x_1 - dr^{\alpha} \tag{8}$$

with

$$r = \sqrt{\sum_{i=2}^{N} x_i^2}. \tag{9}$$

By considering the transformation (4) it becomes clear that x_1 measures the coordinate value on the ridge axis $\mathbf{v}$ and the x_i $(2 \leq i \leq N)$ are perpendicular components. Therefore, r, as given by (9), measures the *distance* to the ridge axis.

C. Melioration on the ridge

Formula (8) allows for an interesting interpretation of (1) as noticed by Oyman [11, p.33]: Enlarging $\tilde{F}_R(\mathbf{x})$ comprises two subgoals:
a) Minimizing the distance r to the ridge axis
b) Enlarging x_1
As we will see later, both subgoals are somewhat conflicting in ESs using isotropic mutations. As a result one observes a performance limit for $\alpha \geq 2$, even though the success domain is an unbounded subset of $\mathbb{R}^N$.[1]

One word of caution should be added here concerning the optimization goal (1). Due to (8), the "optimum" of $\tilde{F}_R$ lies at infinity. Therefore, the goal (1) is somewhat ill-posed. There are two possibilities to resolve this "problem:"
(a) The fitness functions (2) and (8) may be regarded as models describing situations far away from the optimum. The goal is to improve the fitness locally (melioration). Therefore, there is no need to refer to the optimum.
(b) The ill-posed problem is resolved by regularization.
While (a) is the standard way of thinking about performance in ES theory, (b) considers (8) as the result of a limit process using a regularization parameter c

$$\tilde{F}_R(\mathbf{x}) = \lim_{c \to 0} F_{\diamond}(\mathbf{x}, c). \tag{10}$$

[1] There exists an additional performance limit caused by the self-adaptation not considered here.

One possible choice of $F_\diamond$ is

$$F_\diamond(\mathbf{x}, c) := x_1 - cx_1^2 - d\left(\sum_{i=2}^{N} x_i^2\right)^{\alpha/2}, \qquad c \ge 0. \tag{11}$$

Its maximum is at

$$x_1 = \hat{x}_1 = \frac{1}{2c} \qquad x_i = \hat{x}_i = 0 \quad (i = 2, \ldots, N). \tag{12}$$

Thus, one has $\forall i = 2, \ldots, N: \ \hat{x}_i = 0$ and as $c \to 0 \Rightarrow \hat{x}_1 \to \infty$. We will return to this when considering the performance definitions.

D. The $(1, \lambda)$-ES with isotropic mutations

In this paper it is assumed that the reader is familiar with the standard ES notations and algorithms. For an overview of this issue, see [12]. The ES algorithm to be considered here can also be found in [4]. It is displayed in Figure 2 in its self-adaptive version. However, in the analysis and the ES experiments, the learning parameter τ is fixed at $\tau = 0$ (the analysis for $\tau \neq 0$ will still remain as a challenge for the future). With $\tau = 0$, the mutation strengths $\tilde{\sigma}_l =: \sigma$ stay

Procedure $(1, \lambda)$-ES;	line #
Begin	1
$g := 0$;	2
`initialize`$(\mathbf{y}^{(g)}, \sigma^{(g)})$	3
Repeat	4
For $l := 1$ **To** λ **Do Begin**	5
$\tilde{\sigma}_l := \sigma^{(g)} \cdot \exp(\tau \mathcal{N}(0,1))$;	6
$\mathbf{z}_l := \tilde{\sigma}_l \cdot (\mathcal{N}(0,1), \ldots, \mathcal{N}(0,1))^T$;	7
$\tilde{\mathbf{y}}_l := \mathbf{y}^{(g)} + \mathbf{z}_l$;	8
$F_l := F(\tilde{\mathbf{y}}_l)$	9
End;	10
$l_p :=$ `selection`$_{1;\lambda}(F_1, F_2, \ldots, F_\lambda)$;	11
$\sigma^{(g+1)} := \tilde{\sigma}_{l_p}$;	12
$\mathbf{y}^{(g+1)} := \tilde{\mathbf{y}}_{l_p}$;	13
$g := g + 1$;	14
Until stop-criterion	15
End	16

Fig. 2. Algorithm of the $(1, \lambda)$-ES with self-adaptation. The learning parameter τ is usually choosen as $\tau \propto 1/\sqrt{N}$ (e.g. $\tau = c_{1,\lambda}/\sqrt{N}$, see [4]). However, in this paper self-adaptation is switched off by setting $\tau = 0$.
Note, each $\mathcal{N}(0,1)$ represents an independent sample from an $\mathcal{N}(0,1)$ normally distributed random number generator. The `selection` operator in line 11 returns the index of the $\tilde{\mathbf{y}}$ individual with the greatest fitness value.

constant throughout the ES run. That is, the mutations $\mathbf{z}$ produced in line 7 of Figure 2 have components which are $\mathcal{N}(0, \sigma^2)$ distributed. Therefore, the density function of a single component of the mutation vector reads

$$p_z(z) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2}. \tag{13}$$

E. How to evaluate the performance – local progress measures

E.1 General remarks

Generally, one can differentiate between global and local performance measures. While the global measures evaluate the performance of the algorithm for a long time period, e.g., the number of function evaluations needed to get into a certain vicinity of the optimum, the local measures evaluate the change of the state from generation g to $g + 1$. Both measures are more or less strongly connected to each other, however, in a one-way direction. The local measures are very often the microscopic basis on which the evolutionary dynamics are established. Thus, global performance is a byproduct of the evolutionary dynamics.

Local performance can be measured in the fitness space as well as in the search space. The expected (average population) fitness change per generation is called the *quality gain* $\overline{Q}$. If performance is evaluated in the search space, one speaks of *progress rates* φ. The emphasis is here on *progress*; this notion is much more general than "convergence rates." Here, only progress rates will be considered. These rates build the microscopic basis of the performance analysis. Note, the *quality gain* $\overline{Q}$ cannot be used for such an analysis because in general the underlying evolutionary dynamics cannot be reconstructed from the one-dimensional $\overline{Q}$ information:

Considering the evolution of a parent $\mathbf{y}_P$ in $(1,\lambda)$-ES requires its description in the N-dimensional state space, i.e., $\mathbf{y}_P \in \mathbb{R}^N$. Depending on symmetries in the fitness function $F(\mathbf{y})$, components of the state vector $\mathbf{y}_P$ can be lumped together. In the extreme case of the sphere model, one has fitness functions F_s of the type $F_s(\mathbf{y}) = f_s(\|\hat{\mathbf{y}} - \mathbf{y}\|)$, i.e., f_s depends only on the Euclidean distance R of the state vector $\mathbf{y}$ to the optimum. Provided that the ES obeys the symmetry condition,[2] the state of the $(1,\lambda)$-ES is fully determined by the parental R, as long as one is interested in the evolution dynamics of the residual distance to the optimum. As to the ridge function class, due to (8) and (9), the state description reduces to a two-dimensional vector $(x_1, r)^{\mathrm{T}}$. Thus, one has to deal with two dynamical equations, one for the progress in x_1 direction and one for the residual dynamics of r. That means that there are two progress measures φ_x and φ_r, the first measuring the progress in ridge direction and the latter the approach to the ridge axis. Unlike the sphere model where φ can be reconstructed from the quality gain $\overline{Q}$ under certain conditions, the one-dimensional $\overline{Q}$ cannot be used to reconstruct the two-dimensional progress vector $(\varphi_x, \varphi_r)^{\mathrm{T}}$. This is the deeper reason why progress rate vectors must be considered.

E.2 Measuring the progress φ_x in ridge direction

In order to define the longitudinal progress, assume that the $(1,\lambda)$-ES is in the (parental) state $(x^{(g)}, r^{(g)})$ (writing x instead of x_1). Its transition to the next (parental) state $(x^{(g+1)}, r^{(g+1)})$ is described by the conditional transition density $p(x, r|x^{(g)}, r^{(g)})$. The progress in x-direction, i.e., the ridge direction, is then defined as the (conditional) expected distance change in this direction

$$\varphi_x := \mathrm{E}\left[x^{(g+1)} - x^{(g)} | x^{(g)}, r^{(g)}\right]. \tag{14}$$

Thus, φ_x can be expressed as

$$\varphi_x = \int_{-\infty}^{\infty}\int_0^{\infty} (x - x^{(g)}) p(x, r|x^{(g)}, r^{(g)}) \mathrm{d}r \mathrm{d}x. \tag{15}$$

Introducing the marginal density $p_{1;\lambda}$

$$p_{1;\lambda}(x|x^{(g)}, r^{(g)}) = \int_0^{\infty} p(x, r|x^{(g)}, r^{(g)}) \mathrm{d}r \tag{16}$$

and considering the translation invariance of $x^{(g+1)} - x^{(g)}$ in the ridge function (8) one ends up with[3]

$$\varphi_x = \int_{-\infty}^{\infty} z p_{1;\lambda}(z|r^{(g)}) \mathrm{d}z. \tag{17}$$

The "$|r^{(g)}$" in the density $p_{1;\lambda}$ indicates that – apart from the strategy parameters λ and σ (not displayed here) – φ_x still depends on the distance of the parent to the ridge axis.

Equation (14) may be regarded as the definition of the progress in x_1 direction. However, alternatively it can be obtained from the general progress rate definition which measures the parental (Euclidean) distance change to the optimum $\hat{\mathbf{x}}$

$$\varphi := \mathrm{E}\left[\, \|\hat{\mathbf{x}} - \mathbf{x}^{(g)}\| - \|\hat{\mathbf{x}} - \mathbf{x}^{(g+1)}\| \,\Big|\, \mathbf{x}^{(g)}\right] \tag{18}$$

To this end $\tilde{F}_R(\mathbf{x})$ must be regarded as a limiting function of $F_\diamond(\mathbf{x}, c)$ given by (11). With (12) and (9) one obtains (substituting $x = x_1$)

$$\varphi(c) = \mathrm{E}\left[\sqrt{(\hat{x} - x^{(g)})^2 + (r^{(g)})^2} - \sqrt{(\hat{x} - x^{(g+1)})^2 + (r^{(g+1)})^2}\right] \tag{19}$$

[2]Symmetry of the ES algorithm means that there is no preference on search directions neither explicitly nor implicitly built into the algorithm. For example, the (μ,λ)-ES and the $(\mu/\mu_I,\lambda)$-ES (with intermediate multiparent recombination) are symmetrical, whereas the $(\mu/\mu_D,\lambda)$-ES (with dominant recombination) is not symmetrical.

[3]Here, $z := x - x^{(g)}$ has been written in order to simplify the notation.

Since we are interested in the case $c \to 0$, it follow $\hat{x} \to \infty$, $\hat{r} = 0$ and, provided that $x^{(g)} \ll \infty$, $r^{(g)} \ll \infty$, the square roots can be expanded into Taylor series leading to

$$\begin{aligned}
\varphi(c) &= \mathrm{E}\left[(\hat{x}-x^{(g)})\sqrt{1+\frac{\left(r^{(g)}\right)^2}{\left(\hat{x}-x^{(g)}\right)^2}} - (\hat{x}-x^{(g+1)})\sqrt{1+\frac{\left(r^{(g+1)}\right)^2}{\left(\hat{x}-x^{(g+1)}\right)^2}}\right] \\
\varphi(c) &= \mathrm{E}\left[(\hat{x}-x^{(g)})\left(1+\frac{1}{2}\frac{\left(r^{(g)}\right)^2}{\left(\hat{x}-x^{(g)}\right)^2}+\ldots\right) - (\hat{x}-x^{(g+1)})\left(1+\frac{1}{2}\frac{\left(r^{(g+1)}\right)^2}{\left(\hat{x}-x^{(g+1)}\right)^2}+\ldots\right)\right] \\
\varphi(c) &= \mathrm{E}\left[x^{(g+1)}-x^{(g)}+\frac{1}{2}\left(\frac{\left(r^{(g)}\right)^2}{\hat{x}-x^{(g)}}-\frac{\left(r^{(g+1)}\right)^2}{\hat{x}-x^{(g+1)}}\right)+\ldots\right]
\end{aligned} \tag{20}$$

Provided that the expected value exists, then taking the limit $c \to 0$, one finally obtains with (12) $\varphi(c) \stackrel{c\to 0}{\longrightarrow} \varphi_x$, i.e., Equation (14). As one can see, from this point of view, φ_x is also the progress rate in the classical sense.

E.3 Measuring the progress φ_r toward the ridge axis

This progress rate (the radial progress) is similar to that defined for the sphere model. It measures the expected distance change to the ridge axis

$$\varphi_r := \mathrm{E}\left[r^{(g)} - r^{(g+1)} | x^{(g)}, r^{(g)}\right]. \tag{21}$$

Given the state $(x^{(g)}, r^{(g)})$ and the transition density $p(x, r|x^{(g)}, r^{(g)})$ already introduced in Section II-E.2, φ_r can be expressed as

$$\varphi_r = \int_{-\infty}^{\infty}\int_0^{\infty} (r^{(g)} - r) p(x, r|x^{(g)}, r^{(g)}) \mathrm{d}r \mathrm{d}x. \tag{22}$$

Similar to (16) the marginal density is introduced as

$$p_{1;\lambda}(r|x^{(g)}, r^{(g)}) = \int_{-\infty}^{\infty} p(x, r|x^{(g)}, r^{(g)}) \mathrm{d}x \tag{23}$$

Due to the x_1 translation property of $\tilde{F}_R$, $p_{1;\lambda}$ does not depend on $x^{(g)}$, therefore one obtains

$$\varphi_r = r^{(g)} - \int_0^{\infty} r p_{1;\lambda}(r|r^{(g)}) \mathrm{d}r. \tag{24}$$

Rewriting (21) as

$$\mathrm{E}\left[r^{(g+1)} | r^{(g)}\right] = r^{(g)} - \varphi_r(r^{(g)}) = \int_0^{\infty} r p_{1;\lambda}(r|r^{(g)}) \mathrm{d}r, \tag{25}$$

one sees that the r-dynamics do *not* depend on x, whereas from (14) and (17) it follows that

$$E\left[x^{(g+1)} | x^{(g)}, r^{(g)}\right] = x^{(g)} + \varphi_x(r^{(g)}) = x^{(g)} + \int_{-\infty}^{\infty} x p_{1;\lambda}(x|r^{(g)}) \mathrm{d}r, \tag{26}$$

i.e., the x-dynamics is *controlled* by the r-dynamics.

III. On the calculation of the longitudinal progress φ_x

This section is organized as follows. First, the general φ_x-integral for the $(1,\lambda)$-ES is derived. In this integral the conditional density $p(Q|z)$ and the cumulative distribution function $P(Q)$ of the mutation-induced quality change Q are needed. These functions are derived as normal approximations allowing for an analytical treatment of the progress rate integral.

A. *Local quality function and the φ_x integral*

For $(1,\lambda)$-strategies it suffices to consider one generation step. Let us assume for notational simplicity that the ES is in the state $(x^{(g)}, R)$ with $R = r^{(g)}$. The fitness of this state is given by (8) and (9). After application of the mutations $\mathbf{z}_l$ on $\mathbf{x}$ one obtains λ new states $\tilde{\mathbf{x}}_l = \mathbf{x}^{(g)} + \mathbf{z}_l$ resulting in λ new x values and λ new r values. The fitness of those states is given by (8)

$$F(\mathbf{x}^{(g)} + \mathbf{z}_l) = x^{(g)} + z_l - d \cdot (r_l)^{\alpha}. \tag{27}$$

Here, $x^{(g)} := x_1^{(g)}$ and $z := z_1$ has been written to simplify the notations. Note, z_l refers to the z_1 coordinate of the lth mutation vector $\mathbf{z}_l$. The local quality change

$$Q := F(\mathbf{x}^{(g)} + \mathbf{z}) - F(\mathbf{x}^{(g)}) \tag{28}$$

for an arbitrary mutation becomes

$$Q = z - d(r^{\alpha} - R^{\alpha}), \tag{29}$$

i.e., it is independent of the parental x value. An offspring state $\tilde{\mathbf{x}}$ survives the $(1,\lambda)$ selection if its $Q(\tilde{\mathbf{y}})$ value is the best one, i.e., the largest (maximization considered) out of the λ offspring Q-values. The state r and the mutation component z leading to this best offspring will be denoted by $r_{1;\lambda}$ and $z_{1;\lambda}$, respectively. They are random variates whose density functions $p_{1;\lambda}(z)$ and $p_{1;\lambda}(r)$ have already been introduced in the integrals (17) and (25), respectively. It is important to realize that these random variates are *not* usual order statistics known from the literature (e.g. [13]) which are expressed by the symbol "$Q_{m:\lambda}$" with the meaning

$$Q_{1:\lambda} \leq Q_{2:\lambda} \leq \cdots \leq Q_{\lambda:\lambda}. \tag{30}$$

Looking at (29) it becomes clear that (30) does *not* imply $Q_{m:\lambda} \Rightarrow z_{m:\lambda}$. Therefore the standard techniques of order statistics cannot be applied to determine $p_{1;\lambda}(z)$. Instead the technique of *induced order statistics*, coined in [14], [6], must be applied. The derivation of $p_{1;\lambda}(z)$ will be explained in detail now.

In order to determine $p_{1;\lambda}(z)$, recall that according to (29) the best Q value, i.e., $Q_{\lambda:\lambda}$, is obtained by a mutation $\mathbf{z}$ whose x_1 component z is a sample from the $\mathcal{N}(0, \sigma^2)$ distribution with density $p_z(z)$ given by Eq. (13). If one considers a single trial, say the first one (out of λ), the probability of having this trial in an infinitesimal interval $\mathrm{d}z$ around z *and* accept it as the best trial is $p_z(z)\mathrm{d}z \cdot P_{\mathrm{a}}(z)$. Here, $P_{\mathrm{a}}(z)$ is the acceptance probability. Since there are λ independent trials, there are λ independent and mutually exclusive possibilities of being the best. Therefore, the probability $p_{1;\lambda}(z)\mathrm{d}z$ becomes

$$p_{1;\lambda}(z)\mathrm{d}z = \lambda p_z(z) P_{\mathrm{a}}(z)\mathrm{d}z. \tag{31}$$

The acceptance probability $P_{\mathrm{a}}(z)$ is determined as follows. Given a fixed state z, the local quality Q (29) depends also on the random variate r. That is, Q is a random variate conditional on z. The probability of a single trial having quality values in an interval $\mathrm{d}Q$ around Q is given by $p(Q|z)\mathrm{d}Q$. In order to have Q accepted as the best trial out of λ trials, the remaining $(\lambda-1)$ Q values must be smaller (maximization considered here). For a single trial this occurs with probability $P(Q)$, where $P(Q)$ is the cumulative distribution function of the (nonconditional) random variate Q. Since there are $(\lambda-1)$ independent trials, the probability of being smaller than the Q generated by the first trial becomes $[P(Q)]^{\lambda-1}$. Thus one gets

$$p(Q|z)\mathrm{d}Q[P(Q)]^{\lambda-1} \tag{32}$$

for the acceptance probability of a z state producing a quality value in an interval $\mathrm{d}Q$ around Q. The random variate Q itself has the support $Q \in [-\infty, z + dR^{\alpha}]$, because of (29) and $r \in [0, \infty]$. Therefore, one obtains the acceptance probability $P_{\mathrm{a}}(z)$ by integration of (32)

$$P_{\mathrm{a}}(z) = \int_{Q=-\infty}^{Q=z+dR^{\alpha}} p(Q|z)[P(Q)]^{\lambda-1}\mathrm{d}Q, \tag{33}$$

and with (31), it follows

$$p_{1;\lambda}(z|R) = \lambda p_z(z) \int_{Q=-\infty}^{Q=z+dR^{\alpha}} p(Q|z)[P(Q)]^{\lambda-1}\mathrm{d}Q. \tag{34}$$

Here, it has been indicated that $p_{1;\lambda}(z)$ still depends on the parental state $R := r^{(g)}$.

The progress rate φ_x can be expressed now by the integral (17)

$$\varphi_x = \lambda \int_{z=-\infty}^{z=\infty} z p_z(z) \int_{Q=-\infty}^{Q=z+dR^\alpha} p(Q|z)[P(Q)]^{\lambda-1} \mathrm{d}Q \mathrm{d}z. \tag{35}$$

After changing the order of integration, one alternatively obtains

$$\varphi_x = \lambda \int_{Q=-\infty}^{Q=\infty} [P(Q)]^{\lambda-1} \int_{z=Q-d\cdot R^\alpha}^{z=\infty} z p_z(z) p(Q|z) \mathrm{d}z \mathrm{d}Q. \tag{36}$$

This integral will be used for the φ_x calculation. However, while $p_z(z)$ is given by (13), the conditional density $p(Q|z)$ and the cumulative distribution $P(Q)$ remain still to be determined.

B. Normal approximation of $P(Q)$

In order to keep (36) tractable, $P(Q)$ must be approximated by a normal distribution. Such an approximation should be performed in such a way that:
(a) the first moments of the random variate Q are preserved and
(b) the approximation is asymptotically, i.e., for $N \to \infty$, exact.
According to (29), the random variate Q depends on the two random variates z and r. For z one has $p_z(z)$, given by Eq.(13), however, for r there is no analytical $p(r)$. The asymptotically exact normal approximation already derived in [2, p.387, Eq.(22)] will be used

$$p(r) = \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \exp\left[-\frac{1}{2}\left(\frac{r - \sqrt{R^2 + \sigma^2(N-1)}}{\tilde{\sigma}}\right)^2\right] \quad \text{with} \quad \tilde{\sigma} = \sigma\sqrt{\frac{R^2 + \sigma^2(N-1)/2}{R^2 + \sigma^2(N-1)}}. \tag{37}$$

In order to get a normal approximation for $p(Q)$, the dependency of r in the Q expression (29) must be linear. Apart from the case $\alpha = 1$, which is linear by definition, the $\alpha \neq 1$ cases are to be handled by Taylor expansion breaking off after the linear r term

$$\begin{aligned} Q(r) &= Q(R + (r - R)) \\ &= z - d[R + (r - R)]^\alpha + dR^\alpha \\ &= z - dR^\alpha - d\alpha R^{\alpha-1}(r - R) - \cdots + dR^\alpha \\ &= z - d\alpha R^{\alpha-1} r + d\alpha R^\alpha - \dots . \end{aligned} \tag{38}$$

This expansion is still exact for the case $\alpha = 1$. Its application to cases $\alpha \neq 1$ relies on the smallness assumption of $(r - R)$ compared to R. This may not always be fulfilled, however, for the steady-state case it is (see below).

Due to the linearity of (38), Q appears normally distributed as sum of two normally distributed random variates. Considering (37) and (13), its expectation $\overline{Q} := \mathrm{E}[Q]$ is therefore

$$\overline{Q} = -d\alpha R^{\alpha-1}\mathrm{E}[r] + d\alpha R^\alpha = d\alpha R^{\alpha-1}\left(R - \sqrt{R^2 + \sigma^2(N-1)}\right) \tag{39}$$

and its standard deviation $\sigma_Q := \mathrm{D}[Q] = \sqrt{\mathrm{D}^2[z] + \mathrm{D}^2[d\alpha R^{\alpha-1} r]}$

$$\sigma_Q = \sqrt{\sigma^2 + (d\alpha R^{\alpha-1})^2 \tilde{\sigma}^2}. \tag{40}$$

Thus, one obtains for the cumulative distribution $P(Q)$

$$P(Q) = \Phi\left(\frac{Q - \overline{Q}}{\sigma_Q}\right) = \Phi\left(\frac{Q + d\alpha R^{\alpha-1}(\sqrt{R^2 + \sigma^2(N-1)} - R)}{\sqrt{\sigma^2 + (d\alpha R^{\alpha-1})^2\tilde{\sigma}^2}}\right) \tag{41}$$

with $\Phi(x)$ as the distribution function of the standard normal distribution

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{1}{2}t^2} \mathrm{d}t. \tag{42}$$

C. Normal approximation of $p(Q|z)$

The determination of the conditional density $p(Q|z)$ is accomplished by the same technique as has been used for $P(Q)$. The only difference is that one has now to deal with conditional expectations. Starting from the linear approximation (38) keeping z fixed (because this is the condition), one easily finds with (37)

$$\mathrm{E}[Q|z] = z - d\alpha R^{\alpha-1}\mathrm{E}[r] + d\alpha R^{\alpha} \tag{43}$$

and with (39) and $\overline{Q}_{|z} := \mathrm{E}[Q|z]$

$$\overline{Q}_{|z} = z + \overline{Q} = z + d\alpha R^{\alpha-1}(R - \sqrt{R^2 + \sigma^2(N-1)}). \tag{44}$$

For the conditional standard deviation $\sigma_{Q|z} := \mathrm{D}[Q|z]$ one obtains from (38), (37), and (40), taking $\mathrm{D}[z] = 0$ into account,

$$\sigma_{Q|z} = d\alpha R^{\alpha-1}\tilde{\sigma} = \sqrt{\sigma_Q^2 - \sigma^2}. \tag{45}$$

Therefore, the normal approximation of $p(Q|z)$ reads

$$p(Q|z) = \frac{1}{\sqrt{2\pi}\sigma_{Q|z}} \exp\left[-\frac{1}{2}\left(\frac{Q - \overline{Q}_{|z}}{\sigma_{Q|z}}\right)^2\right] = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_Q^2 - \sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{Q - \overline{Q} - z}{\sqrt{\sigma_Q^2 - \sigma^2}}\right)^2\right], \tag{46}$$

$$p(Q|z) = \frac{1}{\sqrt{2\pi}d\alpha R^{\alpha-1}\tilde{\sigma}} \exp\left[-\frac{1}{2}\left(\frac{Q + d\alpha R^{\alpha-1}(\sqrt{R^2 + \sigma^2(N-1)} - R) - z}{d\alpha R^{\alpha-1}\tilde{\sigma}}\right)^2\right]. \tag{47}$$

D. Calculation of the φ_x-integral

The starting point is Eq.(36). In order to simplify the integration, the lower integration limit $z = Q - dR^{\alpha}$ in (36) is extended to $z = -\infty$. The inner integral then reads

$$I_x(Q) = \int_{-\infty}^{\infty} z p_z(z) p(Q|z) \mathrm{d}z \tag{48}$$

and φ_x becomes

$$\varphi_x = \lambda \int_{Q=-\infty}^{\infty} [P(Q)]^{\lambda-1} I_x(Q) \mathrm{d}Q. \tag{49}$$

The extension of the lower integration limit in (48) is justified by the fact that most of the probability mass of the random variable z is concentrated in the interval $z \in [-3\sigma, 3\sigma]$ (Gaussian distribution (13)). The location of the maximum of $p(Q|z)$ with respect to z (consider the exponent (47)), denoted by $\hat{z}$, fulfills the condition $\hat{z} > Q - dR^{\alpha}$. Furthermore, almost the entire area defined by the integrand in (48) is located for $z > Q - dR^{\alpha}$. Therefore, the error made by extending the lower limit to $z = -\infty$ is of small order and can be neglected. Actually, it vanishes for $N \to \infty$. Estimating this error for small N is difficult and depends also on the normal approximation used in order to get (47) and (41). In any case, results obtained by (48), (49) should be compared with simulation experiments (see below).

The calculation of $I_x(Q)$ starts with (46) and (13) inserted in (48)

$$I_x(Q) = \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{\sqrt{2\pi}\sigma_{Q|z}} \int_{-\infty}^{\infty} z \mathrm{e}^{-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2} \exp\left[-\frac{1}{2}\left(\frac{Q - \overline{Q} - z}{\sigma_{Q|z}}\right)^2\right] \mathrm{d}z. \tag{50}$$

After the substitution $t = z/\sigma$, $\mathrm{d}z = \sigma \mathrm{d}t$, one gets

$$I_x(Q) = \frac{\sigma}{\sqrt{2\pi}\sigma_{Q|z}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t \mathrm{e}^{-\frac{1}{2}t^2} \exp\left[-\frac{1}{2}\left(-\frac{\sigma}{\sigma_{Q|z}}t + \frac{Q - \overline{Q}}{\sigma_{Q|z}}\right)^2\right] \mathrm{d}t. \tag{51}$$

With the integral formula (for its derivation, see e.g. [14, p.322])

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t \mathrm{e}^{-\frac{1}{2}t^2} \mathrm{e}^{-\frac{1}{2}(at+b)^2} \mathrm{d}t = \frac{-ab}{(\sqrt{1+a^2})^3} \exp\left[-\frac{1}{2}\left(\frac{b^2}{1+a^2}\right)\right], \tag{52}$$

one obtains with (45)

$$I_x(Q) = \frac{1}{\sqrt{2\pi}} \frac{\sigma^2}{\sigma_Q^3} (Q - \overline{Q}) \exp\left[-\frac{1}{2} \left(\frac{Q - \overline{Q}}{\sigma_Q} \right)^2 \right]. \tag{53}$$

After inserting (53) into (49) and taking (41) into account, one gets

$$\varphi_x = \frac{\lambda}{\sqrt{2\pi}} \frac{\sigma^2}{\sigma_Q^3} \int_{-\infty}^{\infty} (Q - \overline{Q}) \exp\left[-\frac{1}{2} \left(\frac{Q - \overline{Q}}{\sigma_Q} \right)^2 \right] \left[\Phi \left(\frac{Q - \overline{Q}}{\sigma_Q} \right) \right]^{\lambda - 1} \mathrm{d}Q, \tag{54}$$

and with the substitution $x = (Q - \overline{Q})/\sigma_Q$, $\mathrm{d}x = \mathrm{d}Q/\sigma_Q$, it follows

$$\varphi_x = \frac{\sigma^2}{\sigma^Q} \frac{\lambda}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \mathrm{e}^{-\frac{1}{2}x^2} [\Phi(x)]^{\lambda-1} \mathrm{d}x. \tag{55}$$

Since the integral in (55) is the well known progress coefficient $c_{1,\lambda}$ (see e.g. [2, p.385])

$$c_{1,\lambda} := \frac{\lambda}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \mathrm{e}^{-\frac{1}{2}x^2} [\Phi(x)]^{\lambda-1} \mathrm{d}x, \tag{56}$$

the progress rate along the ridge axis becomes

$$\varphi_x = \frac{\sigma^2 c_{1,\lambda}}{\sigma_Q} \tag{57}$$

and by back-substitution, taking (40) and (37) into account, the final φ_x formula reads

$$\boxed{\varphi_x = \frac{\sigma c_{1,\lambda}}{\sqrt{1 + (d\alpha R^{\alpha-1})^2 \frac{R^2 + \sigma^2 (N-1)/2}{R^2 + \sigma^2 (N-1)}}}.} \tag{58}$$

This formula predicts the expected progress in ridge direction depending on α, d, N, R, and σ. Its basic properties are discussed next.

E. Discussion of general properties of φ_x

Given fixed values of $\alpha, d, \sigma < \infty$ and N, the asymptotic behavior of φ_x can be evaluated. Depending on α one obtains different lower and upper progress rate limits for the $(1, \lambda)$-ES:

$$\alpha > 1: \qquad \begin{array}{lcl} R \to 0 & \Rightarrow & \varphi_x = \sigma c_{1,\lambda} \\ R \to \infty & \Rightarrow & \varphi_x = 0, \end{array} \tag{59}$$

for the sharp ridge

$$\alpha = 1: \qquad \begin{array}{lcl} R \to 0 & \Rightarrow & \varphi_x = \dfrac{\sigma c_{1,\lambda}}{\sqrt{1 + d^2/2}} \\ R \to \infty & \Rightarrow & \varphi_x = \dfrac{\sigma c_{1,\lambda}}{\sqrt{1 + d^2}}, \end{array} \tag{60}$$

and for the remaining cases, excluding $\alpha = 0$

$$\alpha < 1: \qquad \begin{array}{lcl} R \to 0 & \Rightarrow & \varphi_x = 0 \\ R \to \infty & \Rightarrow & \varphi_x = \sigma c_{1,\lambda}. \end{array} \tag{61}$$

The special case $\alpha = 0$ corresponds to the hyperplane yielding $\varphi_x = \sigma c_{1,\lambda}$ independent of the R value chosen. The progress rate of the hyperplane appears also in (59) and (61) as a limit case for $R \to 0$ and $R \to \infty$, respectively. Whereas the progress behavior of the sharp ridge is nearly independent of R given fixed σ. However, apart from this special case, the progress along the ridge axis depends strongly on the parental distance R to the ridge axis. Therefore, knowing the $R = r^{(g)}$ dynamics is necessary for the calculation of φ_x. This is done in Section IV. The approximation quality of the asymptotically exact progress rate formula (58) is investigated together with the φ_r results in Section V.

IV. On the calculation of the radial progress φ_r

Most of the considerations to be made are similar to those of Section III. First, the progress rate integral φ_r for the $(1,\lambda)$-ES is derived and a normal approximation for the conditional density $p(Q|r)$ is provided. Second, the integrations are performed yielding φ_r. In the third subsection, the results are compared with experiments.

A. The φ_r integral

According to (21), (24), and (25) the conditional expectation $\mathrm{E}[r^{(g+1)}|r^{(g)}]$ must be calculated in order to obtain φ_r. This conditional expectation, defined by the integral (25), requires determining the density $p_{1;\lambda}(r)$. This function is the density of the random variate $r = r_{1;\lambda}$, i.e., the r-variate belonging to the trial (out of λ trials) that produced the largest Q value by Eq. (29). Again, we have to consider an *induced order statistics.* That is, the same considerations made in Subsection III-A in order to derive $p_{1;\lambda}(z)$ can be applied to $p_{1;\lambda}(r)$. Therefore, the derivation will be sketched by reference to the respective equations in Subsection III-A.

The probability $p_{1;\lambda}(r)\mathrm{d}r$ reads, similarly to (31),

$$p_{1;\lambda}(r)\mathrm{d}r = \lambda p(r) P_\mathrm{a}(r)\mathrm{d}r. \tag{62}$$

Here, $p(r)$ is given by (37) and $P_\mathrm{a}(r)$ is the acceptance probability that a trial producing this r value appears as the best, i.e., its Q value is larger than the Q values of the remaining $(\lambda-1)$ trials. Given a (fixed) r, the acceptance probability for Q values in an infinitesimal interval $\mathrm{d}Q$ around Q reads similarly to (32): $p(Q|r)\mathrm{d}Q[P(Q)]^{\lambda-1}$. The acceptance probability is obtained by integration over all possible Q values. Unlike the $P_\mathrm{a}(z)$ case, where $Q \in [-\infty, z + dR^\alpha]$ was found, here one has $Q \in [-\infty, \infty]$. This is so because of Eq. (29), given a fixed r (condition!), Q depends on the random variate z which has as an $\mathcal{N}(0,\sigma^2)$ normal variate and the support $z \in [-\infty, \infty]$. Thus, one obtains instead of (33)

$$P_\mathrm{a}(r) = \int_{Q=-\infty}^{Q=\infty} p(Q|r)[P(Q)]^{\lambda-1}\mathrm{d}Q \tag{63}$$

and with (62)

$$p_{1;\lambda}(r) = \lambda p(r) \int_{Q=-\infty}^{Q=\infty} p(Q|r)[P(Q)]^{\lambda-1}\mathrm{d}Q. \tag{64}$$

Inserting (64) into (24) and exchanging the order of integrations, the φ_r integral reads (writing $R = r^{(g)}$)

$$\varphi_r = R - \lambda \int_{Q=-\infty}^{Q=\infty} [P(Q)]^{\lambda-1} \int_{r=0}^{r=\infty} r p(r) p(Q|r)\mathrm{d}r\mathrm{d}Q. \tag{65}$$

The calculation of (65) requires the determination of the conditional density $p(Q|r)$ ($P(Q)$ and $p(r)$ have already been determined). Looking at (29), one sees that given a fixed r (condition), Q depends on the random variable z which is according to (13) $\mathcal{N}(0,\sigma^2)$ distributed. Thus, Q is $\mathcal{N}(-d(r^\alpha - R^\alpha), \sigma^2)$ distributed. Unfortunately, using this (exact) distribution would lead to an intractable r-integral in (65). Therefore, the linear approximation (38) must be used resulting in

$$p(Q|r) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{Q-\overline{Q}_{|r}}{\sigma}\right)^2\right] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{Q + d\alpha R^{\alpha-1}(r-R)}{\sigma}\right)^2\right] \tag{66}$$

with the conditional expectation

$$\overline{Q}_{|r} = d\alpha R^{\alpha-1}(R-r). \tag{67}$$

B. Calculation of the φ_r integral

The first step in calculating (65) concerns the inner integral denoted by $I_r(Q)$

$$I_r(Q) := \int_{r=0}^{r=\infty} r p(r) p(Q|r)\mathrm{d}r. \tag{68}$$

With (37) and (66), one gets

$$I_r(Q) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{2\pi}\tilde{\sigma}} \int_{r=0}^{\infty} r \exp\left[-\frac{1}{2}\left(\frac{r-\sqrt{R^2+\sigma^2(N-1)}}{\tilde{\sigma}}\right)^2\right] \exp\left[-\frac{1}{2}\left(\frac{Q + d\alpha R^{\alpha-1}(r-R)}{\sigma}\right)^2\right] \mathrm{d}r. \tag{69}$$

By the substitution $t = (r - \sqrt{R^2 + \sigma^2(N-1)})/\tilde{\sigma}$, i.e., $\mathrm{d}t = \mathrm{d}r/\tilde{\sigma}$ and $r = \tilde{\sigma}t + \sqrt{R^2 + \sigma^2(N-1)}$, one obtains

$$\begin{aligned} I_r(Q) &= \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{2\pi}}\int_{t=-\sqrt{R^2+\sigma^2(N-1)}/\tilde{\sigma}}^{t=\infty} \left(\tilde{\sigma}t + \sqrt{R^2+\sigma^2(N-1)}\right) \mathrm{e}^{-\frac{1}{2}t^2} \\ &\qquad \times \exp\left[-\frac{1}{2}\left(\frac{\tilde{\sigma} d\alpha R^{\alpha-1}}{\sigma}t + \frac{Q + d\alpha R^{\alpha-1}(\sqrt{R^2+\sigma^2(N-1)} - R)}{\sigma}\right)^2\right] \mathrm{d}t. \end{aligned} \tag{70}$$

Since $\exp(-t^2/2)$ is concentrated around $t = 0$ and the lower integration limit $t_l = -\sqrt{R^2 + \sigma^2(N-1)}/\tilde{\sigma} \leq -\sqrt{N-1}$, this limit can be extended to $-\infty$ as $N \to \infty$. With (39) one obtains

$$I_r(Q) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{2\pi}}\int_{t=-\infty}^{t=\infty} \left(\tilde{\sigma}t + \sqrt{R^2+\sigma^2(N-1)}\right) \mathrm{e}^{-\frac{1}{2}t^2} \exp\left[-\frac{1}{2}\left(\frac{\tilde{\sigma} d\alpha R^{\alpha-1}}{\sigma}t + \frac{Q-\overline{Q}}{\sigma}\right)^2\right] \mathrm{d}t. \tag{71}$$

With the help of the integral formula (52) and

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}t^2}\mathrm{e}^{-\frac{1}{2}(at+b)^2}\mathrm{d}t = \frac{1}{\sqrt{1+a^2}}\exp\left[-\frac{1}{2}\left(\frac{b^2}{1+a^2}\right)\right], \tag{72}$$

which can be easily proven by completing the square in the exponent, one obtains

$$I_r(Q) = -\frac{\tilde{\sigma}^2 d\alpha R^{\alpha-1}}{\sqrt{2\pi}\sigma_Q^3}(Q-\overline{Q})\exp\left[-\frac{1}{2}\left(\frac{Q-\overline{Q}}{\sigma_Q}\right)^2\right] + \frac{R^2+\sigma^2(N-1)}{\sqrt{2\pi}\sigma_Q}\exp\left[-\frac{1}{2}\left(\frac{Q-\overline{Q}}{\sigma_Q}\right)^2\right]. \tag{73}$$

Now, the solution (73) to the inner integral (68) of (65) can be inserted in (65) taking (41) into account

$$\begin{aligned} \varphi_r &= R - \frac{\lambda}{\sqrt{2\pi}\sigma_Q}\int_{Q=-\infty}^{Q=\infty}\left[\Phi\left(\frac{Q-\overline{Q}}{\sigma_Q}\right)\right]^{\lambda-1}\sqrt{R^2+\sigma^2(N-1)}\exp\left[-\frac{1}{2}\left(\frac{Q-\overline{Q}}{\sigma_Q}\right)^2\right]\mathrm{d}Q \\ &\quad + \frac{\tilde{\sigma}^2 d\alpha R^{\alpha-1}}{\sigma_Q^3}\frac{\lambda}{\sqrt{2\pi}}\int_{Q=-\infty}^{Q=\infty}(Q-\overline{Q})\exp\left[-\frac{1}{2}\left(\frac{Q-\overline{Q}}{\sigma_Q}\right)^2\right]\left[\Phi\left(\frac{Q-\overline{Q}}{\sigma_Q}\right)\right]^{\lambda-1}\mathrm{d}Q. \end{aligned} \tag{74}$$

As a final step, the substitution $x := (Q - \overline{Q})/\sigma_Q$ is performed

$$\varphi_r = R - \sqrt{R^2+\sigma^2(N-1)}\frac{\lambda}{\sqrt{2\pi}}\int_{x=-\infty}^{x=\infty}\mathrm{e}^{-\frac{1}{2}x^2}[\Phi(x)]^{\lambda-1}\mathrm{d}x + \frac{\tilde{\sigma}^2 d\alpha R^{\alpha-1}}{\sigma_Q}\frac{\lambda}{\sqrt{2\pi}}\int_{x=-\infty}^{x=\infty}x\mathrm{e}^{-\frac{1}{2}x^2}[\Phi(x)]^{\lambda-1}\mathrm{d}x. \tag{75}$$

The integrand in the first integral of (75) can be simplified, because of (42) $\frac{\mathrm{d}}{\mathrm{d}x}[\Phi(x)]^\lambda = \frac{\lambda}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}x^2}[\Phi(x)]^{\lambda-1}$. The second integral is matched by (56). Thus, one gets

$$\varphi_r = R - \sqrt{R^2+\sigma^2(N-1)} + \frac{\tilde{\sigma}^2 d\alpha R^{\alpha-1}}{\sigma_Q}c_{1,\lambda}. \tag{76}$$

Taking the $\tilde{\sigma}$ definition (37) and the σ_Q definition (40) into account, one finally obtains

$$\boxed{\varphi_r = R - \sqrt{R^2+\sigma^2(N-1)} + d\alpha R^{\alpha-1}\frac{R^2+\sigma^2(N-1)/2}{R^2+\sigma^2(N-1)}\frac{\sigma c_{1,\lambda}}{\sqrt{1+(d\alpha R^{\alpha-1})^2\frac{R^2+\sigma^2(N-1)/2}{R^2+\sigma^2(N-1)}}}.} \tag{77}$$

Interestingly, (77) can be simplified by considering (58), φ_r becomes

$$\varphi_r = R - \sqrt{R^2+\sigma^2(N-1)} + d\alpha R^{\alpha-1}\frac{R^2+\sigma^2(N-1)/2}{R^2+\sigma^2(N-1)}\varphi_x(R). \tag{78}$$

C. Comparison with experiments

For comparison purposes similarly to the sphere model a normalization is introduced

$$\sigma^* = \sigma\frac{N-1}{R}, \qquad \varphi^* = \varphi\frac{N-1}{R}, \qquad q = d\alpha R^{\alpha-1}. \tag{79}$$

The normalizations for σ and φ are taken from the sphere model theory (see [1], [2]), the only difference is in the additional parameter q. One obtains from (77)

$$\varphi_r^* = (N-1)\left[1-\sqrt{1+\frac{\sigma^{*2}}{N-1}}\right] + q\frac{1+\frac{\sigma^{*2}}{2(N-1)}}{1+\frac{\sigma^{*2}}{N-1}}\frac{\sigma^* c_{1,\lambda}}{\sqrt{1+q^2\frac{1+\sigma^{*2}/2(N-1)}{1+\sigma^{*2}/(N-1)}}}. \tag{80}$$

Using the normalization (79), the $(N-1)$-dimensional sphere model [2, p.385] is obtained as a limit case

$$(N-1)\text{-D sphere case:} \qquad q \to \infty. \tag{81}$$

By choosing $q = d\alpha R^{\alpha-1}$, the dependency of φ_r on α, d, and R is reduced to a single parameter q. However, this q contains R implicitly. Therefore, q must be taken into account when $\varphi^*(\sigma^*)$-plots are generated. That is, from (79) one finds

$$R = \sqrt[\alpha-1]{\frac{q}{d\alpha}} \qquad \Rightarrow \qquad \sigma = \frac{\sigma^*}{N-1}\sqrt[\alpha-1]{\frac{q}{d\alpha}} \qquad \text{for} \qquad \alpha \neq 1. \tag{82}$$

For the $\alpha = 1$ case, (79) simplifies to $q = d$ and R can be chosen arbitrarily, leading to $\sigma = \sigma^* R/(N-1)$.

For the experimental verification of the predictions made by (58) and (77), so-called "one-generation-experiments" have been performed. That is, given a fixed initial (parental) state, the $(1,\lambda)$-ES is performed for only one generation. The resulting state change in x-direction and the radial state change are recorded and then averaged over G independent "one-generation-experiments."

Figure 3 shows simulation results of the $(1,10)$-ES for $q = 0.2$. The experiments were done using the sharp ridge, i.e.,

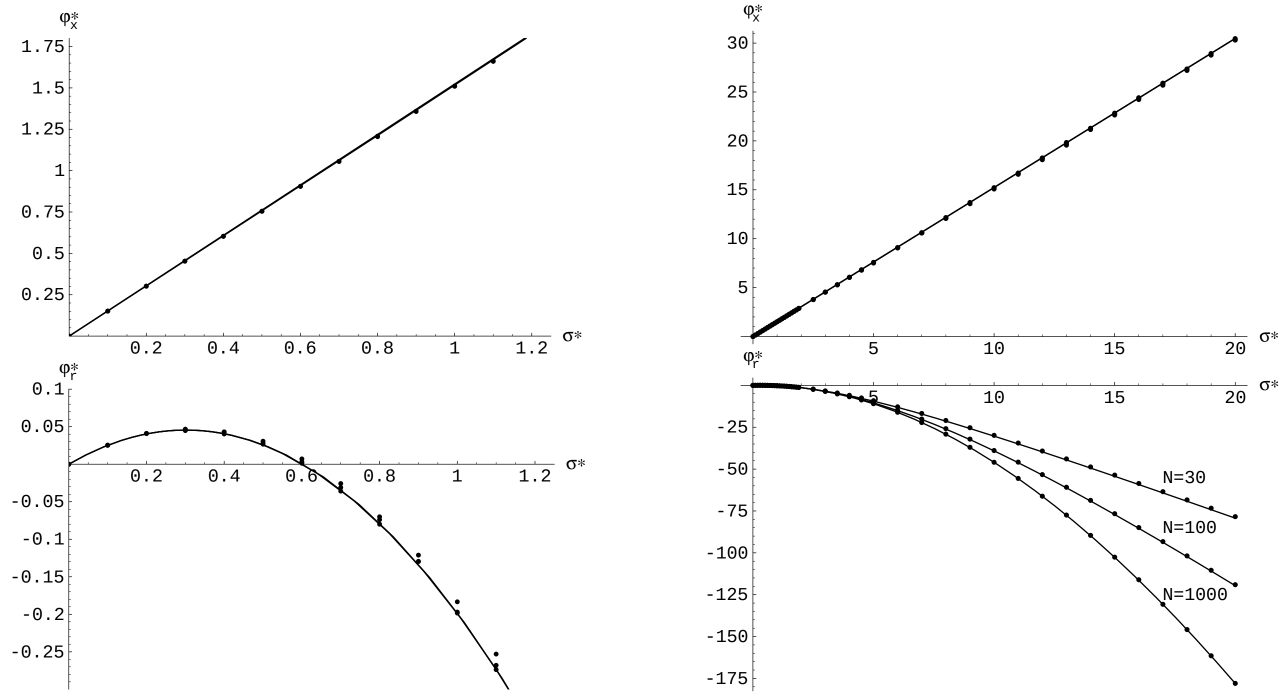

Fig. 3. Comparison between theory (solid curves) and experiment (dots) for φ_x^* and φ_r^* on the sharp ridge ($\alpha = 1$) with $d = 0.2$ for different parameter space dimensions $N = 30$, 100, and 1,000. The left pictures are magnifications of the right ones emphasizing the strategy behavior at small σ^* values. The simulation results for the longitudinal progress rate φ_x^* are almost identical for different N, whereas for the radial progress φ_r^*, there are differences predicted well by the theory.

$\alpha = 1$, with $d = 0.2$ (leading to $q = 0.2$), and an initial distance $r^{(0)}$ to the ridge axis of $R = 5$ has been chosen. Different parameter space dimensions $N = 30$, 100, and 1,000 have been tested. The number of one-generation-experiments were $G = 500{,}000$ (for $N = 30$), $G = 300{,}000$ (for $N = 100$), and $G = 200{,}000$ (for $N = 1{,}000$). Apart from some small deviations for φ_r^* at small N and σ^*, the theory predicts the outcome of the experiments well.

Results of similar experiments for the parabolic ridge, i.e., $\alpha = 2$, with $d = 0.05$ and $q = 2$ (leading to $R = 20$) are presented in Figure 4 and for the cubic ridge, i.e., $\alpha = 3$, with $d = 0.01$ and $q = 3$ (leading to $R = 10$) in Figure 5. In both cases the experimental conditions were $G = 100{,}000$ (for $N = 30$), $G = 50{,}000$ (for $N = 100$), and $G = 30{,}000$

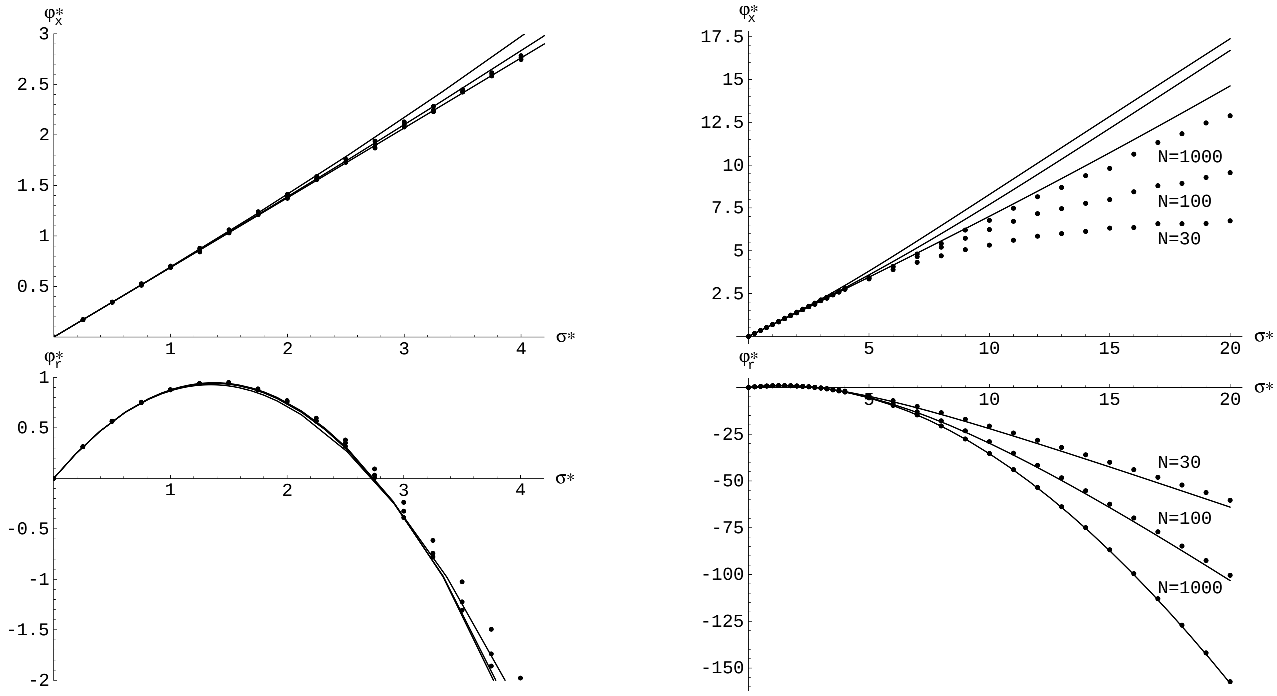

Fig. 4. Comparison between theory (solid curves) and experiment (dots) for φ_x^* and φ_r^* on the parabolic ridge ($\alpha = 2$) with $d = 0.05$ and $q = 2$ for different parameter space dimensions N. The left pictures are magnifications of the right ones emphasizing the strategy behavior at small σ^* values. While there is moderate predictive quality of the φ_r^* formula, the φ_x^* predictions are acceptable for sufficiently small σ^* only. However, in the vicinity of the second zero of φ_r^*, which corresponds to the steady-state behavior (see Section V-B, especially the remarks below Eq. (129)), the predictive quality of the φ_x^* formula is moderate.

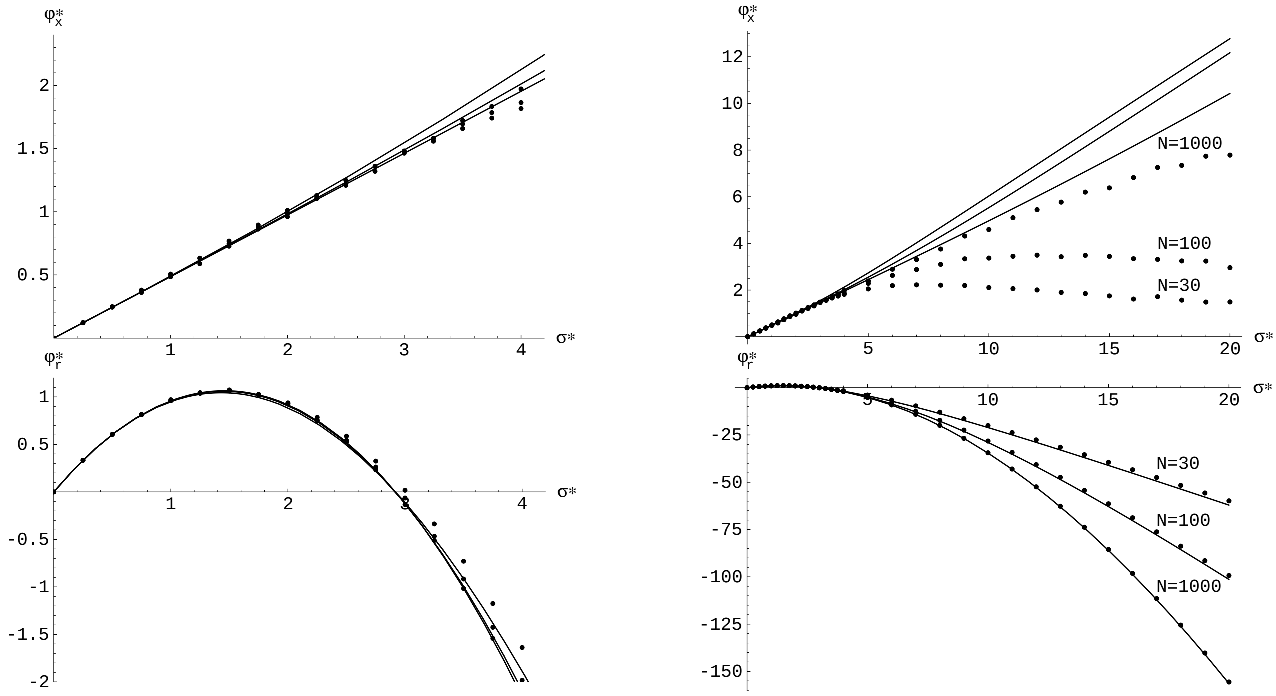

Fig. 5. Comparison between theory (solid curves) and experiment (dots) for φ_x^* and φ_r^* on the cubic ridge ($\alpha = 3$) with $d = 0.01$ and $q = 3$ for different parameter space dimensions N. The left pictures are magnifications of the right ones emphasizing the strategy behavior at small σ^* values. While there is moderate predictive quality of the φ_r^* formula, the φ_x^* predictions are acceptable for sufficiently small σ^* only (see also caption of Figure 4).

(for $N = 1{,}000$). The prediction quality of φ_r^* and especially of φ_x^* is not so good as for the sharp ridge case in Figure 3. This does not come as a big surprise, because the theory was developed mainly for the $\alpha = 1$ case. The cases $\alpha \neq 1$ were treated by the linear Taylor approximation (38). Therefore, one has to expect deviations in the case of $\sigma\sqrt{N-1}$ values which are comparable with the parental R value. By definition (79) this corresponds to large σ^* values. Conversely, one can expect good approximation quality for sufficiently small σ^*. Indeed, this is observed in the left pictures of Figure 4 and 5. Furthermore, one notices that at the (second) root of φ_r^* the φ_x^* value is well predicted. This is an observation of certain importance, because this point corresponds to the steady-state behavior of the ES to be discussed in Section V-B.

V. Mean value dynamics and steady-state behavior

This section discusses some dynamical aspects of the evolution process of $(1,\lambda)$-ES on the ridge function class. For this purpose, recall the equations that describe the mean value dynamics of the r and x value evolution. After that, the steady-state behavior of these equations is investigated. The resulting formulae are compared with simulations and some aspects of the long term dynamics are discussed.

A. Mean value dynamics

As to the r-dynamics, the mean value evolution from g to $g+1$ is given by Eq. (25) and (77) (substituting R back to $r^{(g)}$

$$\mathrm{E}[r^{(g+1)}|r^{(g)}] = \sqrt{\left(r^{(g)}\right)^2 + \sigma^2(N-1)} - d\alpha\left(r^{(g)}\right)^{\alpha-1}\varepsilon(r^{(g)})\frac{\sigma c_{1,\lambda}}{\sqrt{1+\left(d\alpha\left(r^{(g)}\right)^{\alpha-1}\right)^2\varepsilon(r^{(g)})}} \tag{83}$$

with

$$\varepsilon(r^{(g)}) := \frac{\left(r^{(g)}\right)^2 + \sigma^2(N-1)/2}{\left(r^{(g)}\right)^2 + \sigma^2(N-1)}, \qquad \varepsilon(r^{(g)}) \in [\frac{1}{2}, 1]. \tag{84}$$

The x dynamics is obtained from (26) and (58)

$$\mathrm{E}[x^{(g+1)}|x^{(g)}, r^{(g)}] = x^{(g)} + \frac{\sigma c_{1,\lambda}}{\sqrt{1+\left(d\alpha\left(r^{(g)}\right)^{\alpha-1}\right)^2\varepsilon(r^{(g)})}}. \tag{85}$$

It is important to realize that the expectations in (83) and (85) are conditional on the state (x, r) at time g. Since $x^{(g)}$ and $r^{(g)}$ themselves are random variates, the expectations are also random variates. In order to obtain the mean value dynamics, one has to calculate the expectation with respect to the state $(x^{(g)}, r^{(g)})$. Given the density functions at g, i.e., $p^{(g)}(x^{(g)})$ and $p^{(g)}(r^{(g)})$, the mean values, symbolized by a bar over the variables, read

$$\overline{r^{(g)}} := \int_{r=0}^{r=\infty} r p^{(g)}(r)\mathrm{d}r \tag{86}$$

and

$$\overline{x^{(g)}} := \int_{x=-\infty}^{x=\infty} x p^{(g)}(x)\mathrm{d}x. \tag{87}$$

In order to obtain the mean value at time $g+1$, the conditional expectations (83) and (85), respectively, are to be averaged over the state densities at time g

$$\overline{r^{(g+1)}} = \int_{r^{(g)}=0}^{r^{(g)}=\infty} \mathrm{E}[r^{(g+1)}|r^{(g)}]p^{(g)}(r^{(g)})\mathrm{d}r^{(g)} \tag{88}$$

and similarly

$$\overline{x^{(g+1)}} = \int_{x^{(g)}=-\infty}^{x^{(g)}=\infty}\int_{r^{(g)}=0}^{r^{(g)}=\infty} \mathrm{E}[x^{(g+1)}|x^{(g)}, r^{(g)}]p^{(g)}(r^{(g)})p^{(g)}(x^{(g)})\mathrm{d}r^{(g)}\mathrm{d}x^{(g)}. \tag{89}$$

Inserting (83) into (88), one sees that an exact calculation of the mean values will almost always be excluded. Even when one uses normal approximations for $p^{(g)}(r)$ and $p^{(g)}(x)$ which can be calculated from (64) and (34), respectively, the integrals over the root terms in (83) and (85) are not tractable. The usual way to overcome this situation is to expand the $r^{(g)}$ expressions in (83) and (85) in a Taylor series at the parental mean value $\overline{r^{(g)}}$. Thus, (88) and (89) become series over the central moments $\left(r^{(g)} - \overline{r^{(g)}}\right)^k$ which can be calculated similarly to the first-order moment (compare Eq. (21), (22), and (65))

$$\overline{\left(r^{(g)} - \overline{r^{(g)}}\right)^k} = \int_{r=0}^{\infty} (r - \overline{r})^k p_{1;\lambda}(r) \mathrm{d}r. \tag{90}$$

Even though this approach is tractable, it appears that for most of the questions of interest, series expansion up to the linear term (i.e., $k = 1$) suffices. Therefore, only the mean value dynamics in linear approximation will be considered. That is, (88) and (89) are obtained through replacement of $r^{(g)}$ by $\overline{r^{(g)}}$ (note, the linear term vanishes because $k = 1$ in (90).) By this procedure, one gets from (88) with (83)

$$\overline{r^{(g+1)}} = \sqrt{\left(\overline{r^{(g)}}\right)^2 + \sigma^2(N-1)} - d\alpha\left(\overline{r^{(g)}}\right)^{\alpha-1} \varepsilon(\overline{r^{(g)}}) \frac{\sigma c_{1,\lambda}}{\sqrt{1 + \left(d\alpha\left(\overline{r^{(g)}}\right)^{\alpha-1}\right)^2 \varepsilon(\overline{r^{(g)}})}} + \dots \tag{91}$$

and from (89) with (85)

$$\overline{x^{(g+1)}} = \overline{x^{(g)}} + \frac{\sigma c_{1,\lambda}}{\sqrt{1 + \left(d\alpha\overline{(r^{(g)})}^{\alpha-1}\right)^2 \varepsilon(\overline{r^{(g)}})}} + \dots . \tag{92}$$

The system (91), (92) describes the mean value evolution of the $(1, \lambda)$-ES on the ridge functions. A closer look at these equations reveals that the x dynamics, i.e., the longitudinal progress, is controlled by the r dynamics, i.e., the evolution of the residual distance of the parent to the ridge axis. However, this dependency is one-way: the r evolution does not depend on x. Even though this simplifies the dynamical equations, a closed analytical solution is excluded. Only special cases can be treated analytically (see below). However, a numerical treatment of the system (91), (92) can easily be done and the results can be compared with real ES runs.

Figures 6, 7, and 8 show example runs of $(1, 10)$-ES with fixed mutation strength and different values of α, d, and N. All experiments were initialized at the state $\mathbf{x}^{(0)} = \mathbf{0}$, i.e., one has $\overline{r^{(0)}} = 0$ and $\overline{x^{(0)}} = 0$. The plots are the result of

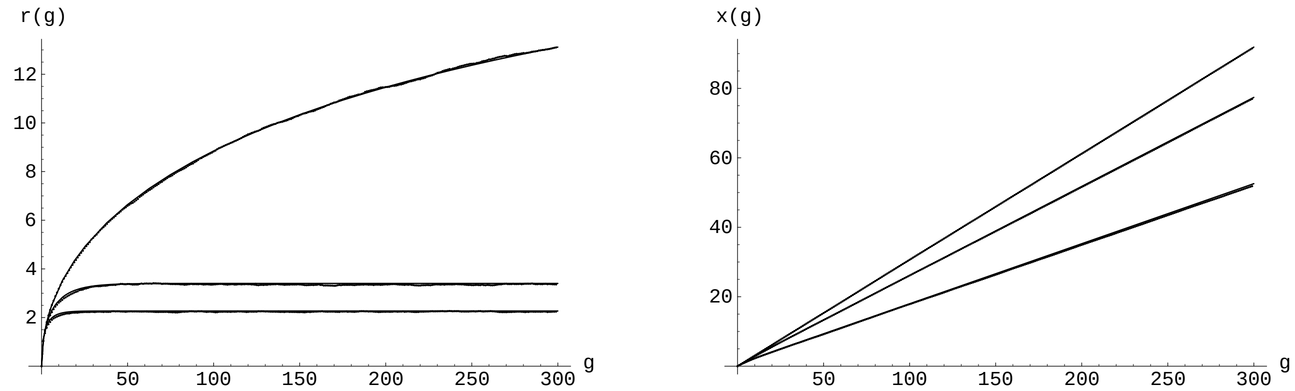

Fig. 6. On the mean value dynamics of the residual distance to the ridge axis, left picture, denoted by $r(g)$, and the distance traveled in longitudinal (ridge) direction, right picture, denoted by $x(g)$. The theoretical predictions obtained from (91), (92) by numerical iteration and displayed as a continuous curve are almost completely covered by the experimental results displayed as small dots. The ridge parameters are $N = 30$, $d = 0.1$ and the $(1, 10)$-ES used $\sigma = 0.2$ as mutation strength. The top curves are for the sharp ridge, $\alpha = 1$, the middle for the parabolic and the lower curves are for the $\alpha = 3$ ridge.

500 independent ES runs averaged in order to smooth the data points. The left pictures display the r dynamics and the right ones the x dynamics. As the most striking characteristic of the r dynamics, one observes "saturation" behavior: The expected value of the distance to the ridge axis approaches a steady-state value r_∞, i.e.,

$$r_\infty := \lim_{g\to\infty} \overline{r^{(g)}}. \tag{93}$$

The consequences for the x dynamics can be read immediately from Eq. (92): The expected value of x experiences a constant change from g to $g + 1$

$$\overline{x^{(g+1)}} = \overline{x^{(g)}} + \varphi_x(r_\infty) \tag{94}$$

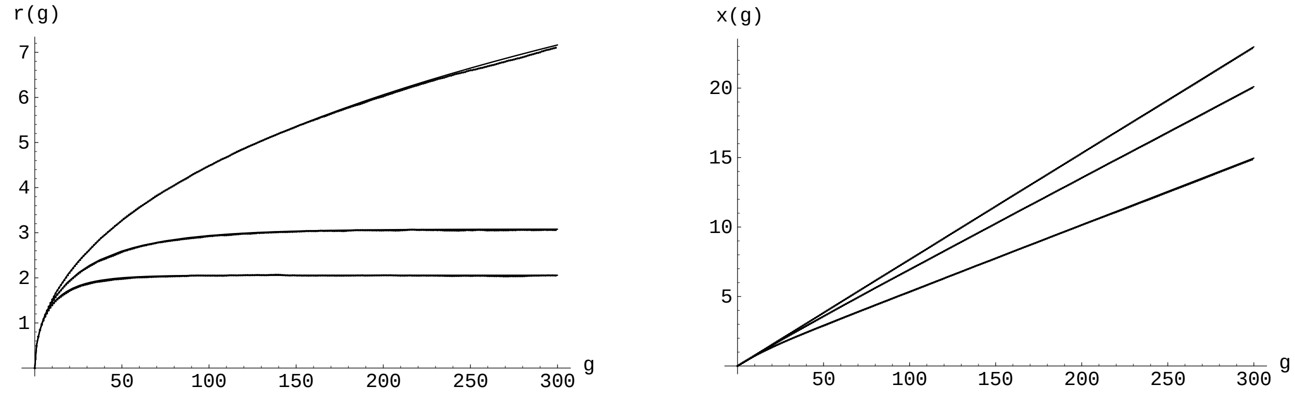

Fig. 7. Mean value dynamics of a $(1, 10)$-ES with $\sigma = 0.05$ on ridge functions with $N = 100$ and $d = 0.1$. The upper curves are for $\alpha = 1$ (sharp ridge), the middle curves are for $\alpha = 2$ (parabolic ridge), and the lower curves are for $\alpha = 3$. For further explanations, see Figure 6.

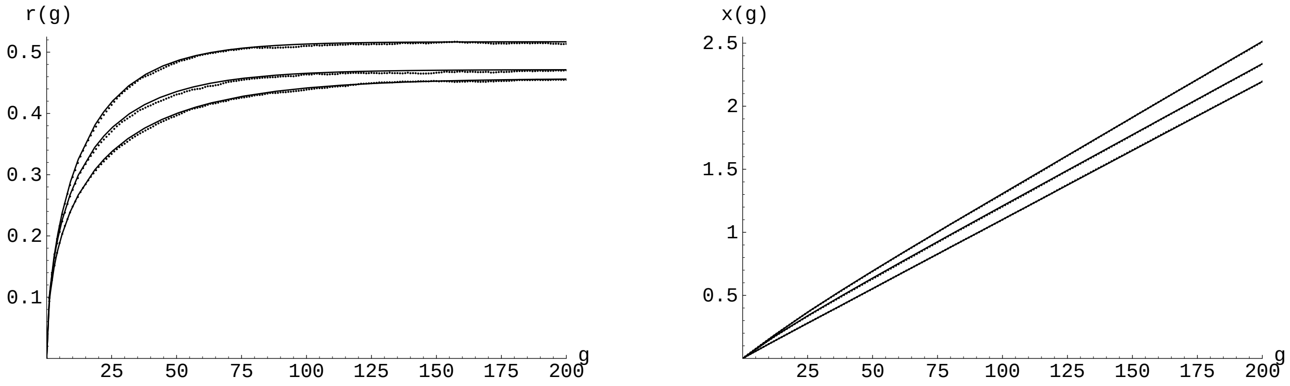

Fig. 8. Mean value dynamics of a $(1, 10)$-ES with $\sigma = 0.01$ on ridge functions with $N = 100$ and $d = 1.0$. The upper curves are for $\alpha = 3$, the middle curves are for $\alpha = 2$ (parabolic ridge), and the lower curves are for $\alpha = 1$ (sharp ridge). For further explanations, see Figure 6.

leading to a linear increase in x direction

$$\text{steady-state:} \qquad \overline{x^{(g)}} = \overline{x^{(g_0)}} + (g - g_0) \cdot \varphi_x(r_\infty). \tag{95}$$

While this dynamical behavior is exact for $g_0 \to \infty$, the curves in Figures 6–8 show that the vicinity of the steady-state is usually reached after a relatively small number of generations g_t (so-called transient time). Unfortunately, there is no easy way to derive analytical expressions for the transient time g_t for arbitrary values of α. However, as seen in the next subsection, this time scales with N.

B. Steady state behavior

B.1 On the expected distance to the ridge axis

The steady-state behavior occurs after a certain transient time g_t in the case of ES operating with constant mutation strength σ. Its main characteristic is the appearance of a constant expected value r_∞ of the distance of the population (and its parent, respectively) to the ridge axis. That is, one has for the

$$\text{steady-state:} \quad \overline{r^{g+1}} = \overline{r^g} = r_\infty \quad \Leftrightarrow \quad \varphi_r = 0, \tag{96}$$

leading with (91) to the nonlinear equation

$$0 = \sqrt{r_\infty^2 + \sigma^2(N-1)} - r_\infty - d\alpha r_\infty^{\alpha-1}\varepsilon(r_\infty)\frac{\sigma c_{1,\lambda}}{\sqrt{1 + (d\alpha r_\infty^{\alpha-1})^2 \varepsilon(r_\infty)}}. \tag{97}$$

Figure 9 shows r_∞ plots depending on d for different σ, N, and α obtained by numerical techniques (analytical expressions can only be derived in some special cases, see below). From these double-logarithmic plots one can infer that there are two different r_∞ asymptotics. The $d \to \infty$ case results in a constant (d-independent) r_∞ value, whereas the $d \to 0$ case can be described by a power law. In the following, the r_∞ asymptotics will be derived.

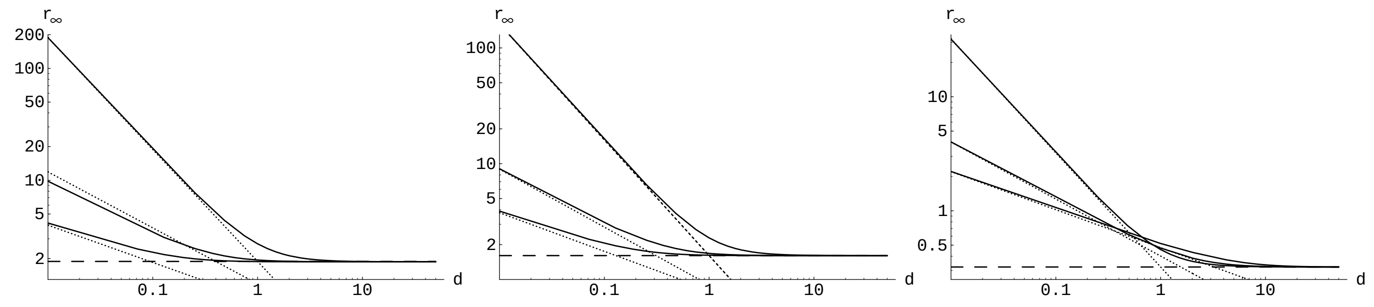

Fig. 9. On the dependence of the steady-state distance to the ridge axis on the ridge parameter d for different values of α, σ, and N. The curves have been obtained by numerically solving Eq. (97) for r_∞. The upper curves are for $\alpha = 1$, the middle ones for $\alpha = 2$, and the lower one for $\alpha = 3$. The dotted lines are the corresponding $d \to 0$ asymptotics and the dashed lines indicate the $d \to \infty$ asymptotics. Left picture: $(1, 10)$-ES with $N = 30$, $\sigma = 0.2$; middle picture: $N = 100$, $\sigma = 0.05$; right picture: $N = 100$, $\sigma = 0.01$.

Let us consider the $d \to \infty$ case first. If one considers the ridge functions (2), (8), the influence of the d parameter becomes clear. It controls the nonlinearity of the ridge. Large d values make the ridge more sphere-like (given fixed σ, α, and N). That is, one can expect a behavior similar to an $(N-1)$-dimensional sphere. Recall, given a fixed mutation strength σ, one expects for the $(1, \lambda)$-ES a residual radius R_∞ [1, p.186].

$$R_\infty = \frac{\sigma(N-1)}{2c_{1,\lambda}}. \tag{98}$$

This result can also be obtained for r_∞ on the ridge functions. By multiplying the third expression in (97) with $\frac{1/d}{1/d}$ and taking the limit $d \to \infty$, one obtains for (97)

$$d \to \infty: \quad 0 = \sqrt{r_\infty^2 + \sigma^2(N-1)} - r_\infty - \sigma c_{1,\lambda}\sqrt{\varepsilon(r_\infty)}. \tag{99}$$

Under the condition $r_\infty \gg \sigma\sqrt{N-1}$, i.e., $\sigma\sqrt{N-1}/r_\infty \to 0$, one finds for (84)

$$\sigma\sqrt{N-1}/r_\infty \to 0 \quad \Rightarrow \quad \varepsilon(r_\infty) \to 1 \tag{100}$$

and the square root can be expanded in a Taylor series leading to

$$\sqrt{r_\infty^2 + \sigma^2(N-1)} - r_\infty = r_\infty\sqrt{1 + \frac{\sigma^2(N-1)}{r_\infty^2}} - r_\infty = \frac{\sigma^2(N-1)}{2r_\infty} + \dots. \tag{101}$$

Substituting the results (101) in (99) and resolving for r_∞ gives finally (considering (98))

$$d \to \infty: \quad r_\infty = \frac{\sigma(N-1)}{2c_{1,\lambda}} = R_\infty. \tag{102}$$

The asymptotic r_∞ value does not depend on d. This is in agreement with the numerical results displayed in Figure 9 as horizontal dashed lines.

The $d \to 0$ asymptotic concerns the case where the nonlinear term in the ridge function (2), (8) appears as a “disturbance” on a hyperplane fitness landscape. Therefore, one can expect a more “hyperplane-like” behavior, e.g., $r_\infty \gg R_\infty \Rightarrow \frac{R_\infty}{r_\infty} \to 0$. That is, again one can assume (101) and the validity of $\varepsilon(r_\infty) \to 1$, leading with (97) to

$$\frac{1}{r_\infty}\frac{\sigma(N-1)}{2c_{1,\lambda}} = \frac{R_\infty}{r_\infty} = \frac{d\alpha r_\infty^{\alpha-1}}{\sqrt{1 + (d\alpha r_\infty^{\alpha-1})^2}}. \tag{103}$$

The right-hand side can only take values between 0 and 1. From (102) we know that 1 corresponds to $d \to \infty$. Therefore, $d \to 0$ might correspond to zero. Furthermore, assuming $d\alpha r_\infty^{\alpha-1} \ll 1$ leads to $R_\infty/r_\infty \to d\alpha r_\infty^{\alpha-1}$ in (103), i.e.,

$$d \to 0 \ \wedge \ d\alpha r_\infty^{\alpha-1} \ll 1: \quad \frac{\sigma(N-1)}{2c_{1,\lambda}} = d\alpha r_\infty^{\alpha}, \tag{104}$$

and consequently $r_\infty = \sqrt[\alpha]{R_\infty/d\alpha}$ with $R_\infty = \sigma(N-1)/2c_{1,\lambda}$. It is important to realize here that this asymptotic behavior for $d \to 0$ still depends on σ, because $d\alpha r_\infty^{\alpha-1} \ll 1$ must be fulfilled. Inserting the r_∞ formula in the condition

$d\alpha r_\infty^{\alpha-1} \ll 1$ leads to $R_\infty \ll (d\alpha)^{-1/(\alpha-1)}$ or alternatively $d \ll 1/\alpha R_\infty^{\alpha-1}$. Thus, one finally gets

$$\left.\begin{array}{l} d \ll \frac{1}{\alpha R_\infty^{\alpha-1}} \\ \text{or} \\ \sigma \ll \frac{2c_{1,\lambda}}{(N-1)\sqrt[\alpha-1]{d\alpha}} \end{array}\right\}: \quad r_\infty = \sqrt[\alpha]{\frac{R_\infty}{d\alpha}}, \qquad \text{with} \quad R_\infty = \frac{\sigma(N-1)}{2c_{1,\lambda}}. \tag{105}$$

As one can see, this asymptotic holds for sufficiently small d (given a fixed σ) and for sufficiently small mutation strengths σ (given fixed d), respectively.[4] Taking the logarithm in (105), one obtains $\log r_\infty = \frac{1}{\alpha}\log\frac{R_\infty}{\alpha} - \frac{1}{\alpha}\log d$. That is, the logarithm of r_∞ is a linear decreasing function of the logarithm of d. Its slope is $1/\alpha$. In Figure 9, these asymptotic curves are indicated by dotted lines.

While there is no analytical solution to (97) in general, the case $\alpha = 1$ can be treated under the condition $\varepsilon(r_\infty) \to 1$ and the case $\alpha = 2$ by additionally taking (101) into account. For the sharp ridge one gets from (97) with $\varepsilon(r_\infty) = 1$

$$\alpha = 1: \quad \sqrt{r_\infty^2 + \sigma^2(N-1)} = r_\infty + \frac{d\sigma c_{1,\lambda}}{\sqrt{1+d^2}}. \tag{106}$$

After squaring this equation and resolving for r_∞ one obtains using the definition of R_∞ (98)

$$\boxed{\alpha = 1: \quad r_\infty = R_\infty\sqrt{1+\frac{1}{d^2}}\left(1 - \frac{d^2}{1+d^2}\frac{c_{1,\lambda}^2}{N-1}\right).} \tag{107}$$

Since usually $c_{1,\lambda}^2 \ll N$ does hold, the second term in the parentheses can be neglected.

As to the parabolic ridge ($\alpha = 2$), one starts from (97) with $\varepsilon(r_\infty) = 1$ and (101) and substitutes (98) yielding

$$\alpha = 2: \quad 2dR_\infty = \frac{(2dr_\infty)^2}{\sqrt{1+(2dr_\infty)^2}}. \tag{108}$$

Using the substitutions $y := (2dr_\infty)^2$ and $a := 2dR_\infty$ leads to the quadratic equation $y^2 - a^2 y - a^2 = 0$. Its formal (positive) root reads

$$y = a^2\left(\frac{1}{2} + \frac{1}{2}\sqrt{1+\frac{4}{a^2}}\right). \tag{109}$$

After back-substitution one gets

$$\boxed{\alpha = 2: \quad r_\infty = R_\infty\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{1+\frac{1}{(dR_\infty)^2}}}.} \tag{110}$$

The case $\alpha = 3$ leads to a third-order equation; its analytical solution is omitted here. Generally, the cases $\alpha \notin \{1, 2, 3\}$ must be handled by numerical root finding techniques.

B.2 On the transient time behavior

The dynamical process of reaching the r_∞ value is described by Eq. (91). Again, its analytical solution is excluded. However, for the $\alpha = 1$ case one can approximately calculate the number of generations g needed to reach a certain vicinity κr_∞ of the steady-state. Furthermore, it is possible to provide a lower bound for the general case.

In order to calculate the number of generations needed to reach a certain $r = \kappa r_\infty$, Eq. (91) is approximated by a differential equation such that $\overline{r^{(g+1)}} - \overline{r^{(g)}} = \frac{\mathrm{d}r}{\mathrm{d}g} + \ldots$. Assuming $\varepsilon(r) = 1$ and using the approximation (101), one thus obtains (neglecting higher order terms)

$$\frac{\mathrm{d}r}{\mathrm{d}g} = \frac{\sigma^2(N-1)}{2r} - \frac{d\sigma c_{1,\lambda}}{\sqrt{1+d^2}} =: \frac{a}{r} - b. \tag{111}$$

Its integration is easily carried out by separation

$$\int_{r_0}^{r_g} \frac{r}{a-br}\mathrm{d}r = \int_{g_0}^{g_r} g\mathrm{d}g. \tag{112}$$

[4] Conversely, Eq. (102) is also valid for *constant* $d < \infty$ and sufficiently large σ, as can be easily verified by considering the asymptotic behavior of (103) for $d = \text{const.}$, $\sigma \to \infty$.

Introducing the relative deviation of $r_g = r(g)$ and $r_0 = r(0)$ from the steady-state r_∞ by κ_g and κ_0, respectively

$$r_g = \kappa_g r_\infty \quad \text{and} \quad r_0 = \kappa_0 r_\infty, \tag{113}$$

one gets from (112)

$$g_r - g_0 = \left[-\frac{r}{b} - \frac{a}{b^2} \ln(a - br) \right] \Big|_{\kappa_0 r_\infty}^{\kappa_g r_\infty}. \tag{114}$$

Substituting a and b by the expressions from (111) and taking (107) in terms of $r_\infty = \frac{\sigma(N-1)}{2c_{1,\lambda}}\sqrt{1+\frac{1}{d^2}}$ into account (second term in (107) is neglected, because $c_{1,\lambda}^2 \ll (N-1)$ can be assumed), one finally obtains

$$\alpha = 1: \quad g_r - g_0 = \frac{N-1}{2c_{1,\lambda}^2}\left(1 + \frac{1}{d^2}\right)\left[-(\kappa_g - \kappa_0) - \ln\left(\frac{1-\kappa_g}{1-\kappa_0}\right)\right]. \tag{115}$$

This formula predicts the number of generations well, as one can verify by explicit numerical calculation for the experimental settings in Figures 6–8 ($\alpha = 1$, $\kappa_0 = 0$, $c_{1,10} = 1.53388$). Since κ_g and κ_0 are constants with respect to N, the number of generations needed to reach a certain vicinity of r_∞ scales linearly in N. That is, the transient time g_t is of $\mathcal{O}(N)$.

As to the cases $\alpha \neq 1$, no satisfactory approximation formula has been found up until now. However, one can easily show that there is a lower bound on the transient time g_t being of order N. To this end, we consider the difference equation (91) again. Writing r instead of $\overline{r^{(g)}}$, one successively obtains

$$\begin{aligned} r^{(g+1)} - r \quad &= \quad \sqrt{r^2 + \sigma^2(N-1)} - r - d\alpha r^{\alpha-1}\varepsilon(r)\frac{\sigma c_{1,\lambda}}{\sqrt{1 + (d\alpha r^{\alpha-1})^2\varepsilon(r)}} \\ &\leq \sqrt{r^2 + \sigma^2(N-1)} - r \leq \frac{\sigma^2(N-1)}{2r} \end{aligned} \tag{116}$$

$$r^{(g+1)} - r^{(g)} \leq \frac{\sigma^2(N-1)}{2r^{(g)}}. \tag{117}$$

This difference inequality can be satisfied by

$$r^{(g)} = \sigma\sqrt{N-1}\sqrt{g}. \tag{118}$$

To see this, insert (118) in $r^{(g+1)} - r^{(g)}$

$$r^{(g+1)} - r^{(g)} = \sigma\sqrt{N-1}(\sqrt{g+1} - \sqrt{g}) \tag{119}$$

and take $\sqrt{g+1} - \sqrt{g} = \sqrt{g}\left(\sqrt{1+\frac{1}{g}} - 1\right) \leq \sqrt{g}\left(1 + \frac{1}{2g} - 1\right) = \frac{1}{2\sqrt{g}}$ into account

$$r^{(g+1)} - r^{(g)} \leq \frac{\sigma\sqrt{N-1}}{2}\frac{1}{\sqrt{g}} = \frac{\sigma^2(N-1)}{2r^{(g)}}. \tag{120}$$

Actually, (118) is also the continuous time solution to the differential equation corresponding to (117). Therefore, Eq. (118) represents an upper bound on $\overline{r^{(g)}}$ for initial $r = r_0 < r_\infty$. This solution increases with a $\sqrt{g}$ law. Since $r_\infty < \infty$ and (117), the graph of (118) must be above the real r dynamics. That is, equating (118) with κr_∞ yields the number of generations g_t to reach the κr_∞ value by the $\sqrt{g}$ dynamics starting from $r^{(0)} = 0$

$$g_t = \frac{\kappa^2 r_\infty^2}{\sigma^2(N-1)} \geq \frac{\kappa^2 R_\infty^2}{\sigma^2(N-1)} = \frac{N-1}{4c_{1,\lambda}^2}\kappa^2. \tag{121}$$

Since the (118) dynamics represents an upper bound on r, (121) represents a lower bound on the transient time g_t. As one can see, this lower bound on the transient time scales with N.

Unfortunately, up to now no α and d dependent upper bound has been derived on the transient time g_t. Looking at Figures 6–8, one can conjecture that the $\alpha = 1$ case might serve as a bound for the $\alpha > 1$ case. If this were correct, Eq. (115) could be applied to securely estimate the transient time g_t.

B.3 The steady-state progress

Reaching the vicinity of the steady-state after g_0 generations, the longitudinal progress (in the ridge direction) is governed by Eq. (95). Analytical expressions for the progress rate $\varphi_x(r_\infty)$ are therefore of certain interest.

Because $r_\infty \geq R_\infty$ (which follows immediately from (103)), the $\varepsilon(r)$ in the $\varphi_x(r)$ expression (58) can be assumed to be 1 (for $N \gg 1$). This leads to the steady-state progress rate formula

$$\varphi_{\text{ss}} = \frac{\sigma c_{1,\lambda}}{\sqrt{1 + (d\alpha r_\infty^{\alpha-1})^2}}, \tag{122}$$

a very simple formula already obtained by Oyman et al. [9] using a different approach based on a local geometrical model (see also [15]). Because $r_\infty \geq R_\infty$, inserting $r_\infty = R_\infty$ in (122) provides an upper bound on the steady-state progress rate

$$\varphi_{\text{ss}} \leq \frac{\sigma c_{1,\lambda}}{\sqrt{1 + \left[d\alpha \left(\frac{\sigma(N-1)}{2c_{1,\lambda}}\right)^{\alpha-1}\right]^2}}. \tag{123}$$

One might expect that this formula can also be used as an analytical expression for the real φ_{ss} value. Unfortunately, a comparison with (122) using r_∞ obtained by numerical solution of (103) reveals considerable deviations (up to 50%). However, (123) can be used as a guide to find a φ-σ-normalization. As first noticed in [9], the normalization

$$\left.\begin{array}{lll} \alpha > 1: & \varphi^\circ := d^{\frac{1}{\alpha-1}}(N-1)\varphi, & \sigma^\circ := d^{\frac{1}{\alpha-1}}(N-1)\sigma \\ \alpha = 1: & \varphi^\circ := \varphi, & \sigma^\circ := \sigma \end{array}\right\} \tag{124}$$

simplifies (123) to

$$\alpha > 1: \quad \varphi^\circ \leq \frac{\sigma^\circ c_{1,\lambda}}{\sqrt{1 + \alpha^2 \left(\frac{\sigma^\circ}{2c_{1,\lambda}}\right)^{2(\alpha-1)}}}. \tag{125}$$

That is, one might conjecture that by the application of (124) the influence of d and N on the normalized progress rate could be removed ($\alpha \neq 1$). This would reduce the numerical calculation of $\varphi(\sigma)$ curves to just one curve for a given $\alpha \neq 1$ and λ.[5] The benefit of such a normalization becomes clear when discussing the x dynamics described by (95). Applying (124) to (95) gives

$$\overline{x^{(g)}} = \overline{x^{(g_0)}} + \frac{\varphi^\circ(\sigma^\circ)}{d^{\frac{1}{\alpha-1}}(N-1)}(g - g_0). \tag{126}$$

Conversely, one can calculate the number of generations $G := g - g_0$ needed to travel a distance $D := x^{(g)} - x^{(g_0)}$ along the ridge direction. Resolving (126) for $G := g - g_0$, one obtains

$$G = \frac{d^{\frac{1}{\alpha-1}}(N-1)}{\varphi^\circ(\sigma^\circ)} D. \tag{127}$$

This formula allows for an assessment of the time complexity of the steady-state evolution process.

The influence of d and N can be directly assessed in (127), provided that $\varphi^\circ(\sigma^\circ) \neq f(d, N)$. The validity of this conjecture is shown in the Appendix. At this point, the special cases $\alpha = 1$ and $\alpha = 2$ is presented. The $\alpha = 1$ case (sharp ridge) is a special one, because in the steady-state, the progress rate does not depend on r_∞ (NB, this is asymptotically correct, i.e., for $N \gg 1$, see Eq. (58) and take $\varepsilon(r_\infty) \to 1$ into account)

$$\alpha = 1: \quad \varphi^\circ = \frac{\sigma^\circ c_{1,\lambda}}{\sqrt{1 + d^2}}. \tag{128}$$

For the parabolic ridge, i.e., $\alpha = 2$, one finds with (122), (110), (98), and (127)

$$\alpha = 2: \quad \varphi^\circ = \frac{\sigma^\circ c_{1,\lambda}}{\sqrt{1 + \frac{(\sigma^\circ)^2}{2c_{1,\lambda}^2}\left(1 + \sqrt{1 + \frac{4c_{1,\lambda}^2}{(\sigma^\circ)^2}}\right)}}. \tag{129}$$

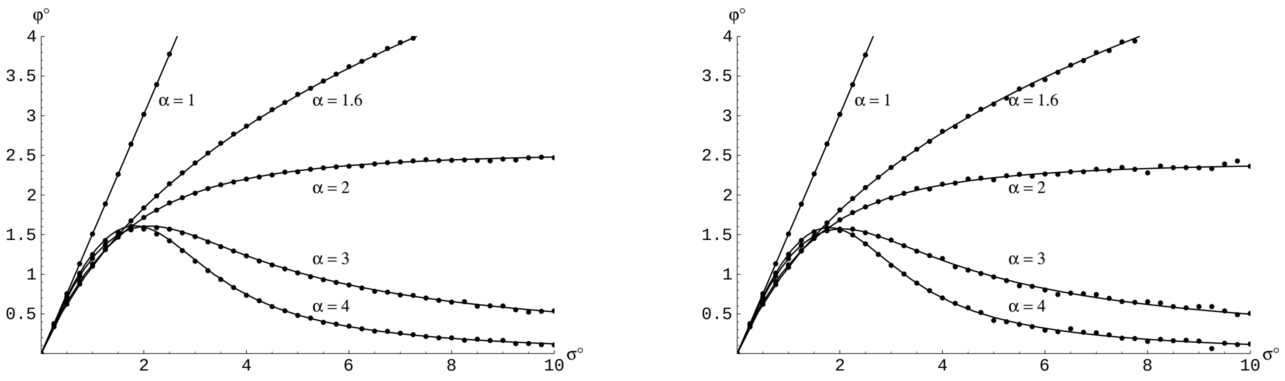

Fig. 10. On the steady-state progress rate (longitudinal progress) of a $(1, 10)$-ES working with isotropic Gaussian mutations with strength σ which is kept constant during the ES run. The normalization (124) has been used. The curves are the prediction from the theory using (58) with an R obtained from (77) by numerically solving $\varphi_r(R) = 0$. The dots are from real ES runs. Each dot corresponds to one ES experiment with a fixed mutation strength σ. The ES was initialized at $\mathbf{x}^{(0)} = \mathbf{O}$. After 5,000 generations, progress data were collected and averaged over a number of G generations (left picture: $N = 30$, $G = 500,000$; right picture: $N = 100$, $G = 100,000$). The d parameters chosen are $d = 0.2$ (for $\alpha = 1$ and $\alpha = 1.6$), $d = 0.05$ (for $\alpha = 2$), $d = 0.01$ (for $\alpha = 3$), and $d = 0.1$ (for $\alpha = 4$).

Cases other than $\alpha = 1$ and $\alpha = 2$ can be obtained numerically. Figure 10 shows such $\varphi^\circ(\sigma^\circ)$ curves for $\alpha = 1$, 1.6, 2, 3, and 4. The dots represent data points obtained by real ES runs using a $(1, 10)$-ES. Even for the $N = 30$ case and $\alpha \neq 1$ (left picture), there is a very good agreement between theory and experiment. This can be explained by means of Figures 3–5: The steady-state corresponds to the φ_r^* values for which $\varphi_r^*(\sigma_0^*) = 0$, where σ_0^* is the second zero of $\varphi_r^*(\sigma^*)$. Looking at the related φ_x^* values reveals that the operating point is in a region of the φ_x^* values for which the approximations developed are of good predictive quality. If σ_0^* were significantly larger, as one would expect for recombinative ES for example, $\varphi^\circ(\sigma^\circ)$ would not have such a high predictive quality.

Figure 10 also shows the influence of α on the shape of the $\varphi^\circ(\sigma^\circ)$ curves. While the $\alpha = 1$ case shows a linearly increasing function of σ°, the parabolic ridge case ($\alpha = 2$) exhibits a saturation behavior for $\sigma^\circ \to \infty$, first discovered in [16]. From (129) one easily obtains $\varphi^\circ \xrightarrow{\sigma^\circ \to \infty} c_{1,\lambda}^2$. Obviously, the cases $1 < \alpha < 2$ lie in between. As to the mutation strength, in order to have high performance, σ can be chosen arbitrarily large. However, this does not hold for $\alpha > 2$. In those cases there exists an optimal $\sigma^\circ = \hat{\sigma}^\circ$ which maximizes the steady-state progress rate. The existence of such a performance maximum can be easily inferred from (125) by showing that for $\sigma^\circ \to \infty$ the right-hand side of (125) goes to zero, provided that $\alpha > 2$. Since $\varphi^\circ(0) = 0$ and $\varphi^\circ(\sigma^\circ) > 0$ $(0 < \sigma^\circ < \infty)$, there must be a φ maximum (at least one). Unfortunately, an analytical expression for the optimal σ° or φ° value cannot be derived.

VI. Outlook

This paper has provided an N-dependent analysis of the $(1, \lambda)$-ES with constant mutation strength (no self-adaptation) on the ridge functions. Unlike the analysis in [15], [10], a simplified but even more accurate approach has been found that also allows for a treatment of the radial dynamics for different α. The predictions of this approach as to the steady-state behavior of the ES, which is reached after $\mathcal{O}(N)$ ES generations, are very accurate. Therefore, they can serve as performance reference when comparing this $(1, \lambda)$-ES with algorithms using self-adaptation.[6] Self-adaptation is necessary in real-world applications and even in the case of the ridge family when α is unknown. Actually, the ridge function class is a good test bed for investigations and evaluations of self-adaptation. As noticed by Herdy already in 1992, standard self-adaptation can fail on the sharp ridge: The mutation strength goes down to zero even though it should increase or at least stay constant.

As for the ridge functions, especially for the sharp ridge, it seems that the self-adaptation mechanism rewards the short-term goal (reduction of the residual distance r) rather than the long-term goal (increasing x along the ridge axis). As long as d is not small enough, self-adaptation is deceived by the short-term goal. While this behavior is not fully understood up to now, because of lack of an appropriate self-adaptation theory, this paper makes a first step toward a deeper understanding of those evolutionary processes, and it predicts the σ values that should be expected if self-adaptation worked optimally.

Besides the analysis of $(1, \lambda)$-σ-self-adaptation on the ridge function class, which should be a research task for the future, the analysis of (μ, λ)-ES and recombination, e.g., the $(\mu/\mu, \lambda)$-ES, must be considered further.

[5] One could even go a step further and define the normalization $\varphi^\bullet := \varphi^\circ / c_{1,\lambda}^2$, $\sigma^\bullet := \sigma^\circ / c_{1,\lambda}$. By this, the influence of λ would be removed.

[6] Note, self-adaptation in the algorithm of Figure 2 can be switched on by choosing $\tau \neq 0$.

Acknowledgment

The author thanks Irfan Oyman for his help and Dirk V. Arnold for his comments. The suggestions and the help of the anonymous reviewers are gratefully acknowledged. This paper is a result of the author's research as a Heisenberg Fellow of the DFG under grant Be 1578/4-1.

Appendix: On the independence of $\varphi^\circ(\sigma^\circ)$ on d and N

In order to prove the correctness of the assumption that for $\alpha > 1$ the normalization (124) yields a function $\varphi^\circ(\sigma^\circ) \neq f(d, N)$ (within the framework of approximations used), one has to show that by applying (124) the square root in (122) as a function of σ° does not depend on d and N. To this end, the right-hand side of (103) is multiplied by $R_\infty^{\alpha-1}/R_\infty^{\alpha-1}$ and the substitution

$$y = r_\infty^2 / R_\infty^2 \tag{130}$$

is introduced. This leads to

$$\frac{1}{\sqrt{y}} = \frac{d\alpha(\sqrt{y})^{\alpha-1}}{\sqrt{R_\infty^{-2(\alpha-1)} + (d\alpha)^2 y^{\alpha-1}}}. \tag{131}$$

By multiplication one gets the nonlinear equation

$$0 = y^\alpha - y^{\alpha-1} - \frac{1}{(d\alpha R_\infty^{\alpha-1})^2}. \tag{132}$$

Back-substitution of $R_\infty = \sigma(N-1)/2c_{1,\lambda}$ and using $\sigma = \sigma^\circ / d^{\frac{1}{\alpha-1}}(N-1)$ yields

$$0 = y^\alpha - y^{\alpha-1} - \frac{(2c_{1,\lambda})^{2(\alpha-1)}}{\alpha^2(\sigma^\circ)^{2(\alpha-1)}} \neq f(d, N) \quad \Rightarrow \quad y \neq f(d, N). \tag{133}$$

Therefore, one has $1/\sqrt{y} = R_\infty / r_\infty \neq f(d, N)$ and with (103)

$$\frac{d\alpha r_\infty^{\alpha-1}}{\sqrt{1 + (d\alpha r_\infty^{\alpha-1})^2}} \neq f(d, N) \quad \Rightarrow \quad d\alpha r_\infty^{\alpha-1} \neq f(d, N). \tag{134}$$

Since $d\alpha r_\infty^{\alpha-1}$ – as a function of σ° – does not depend on d and N, the φ formula obtained from (122) by the normalization (124) must also have this property

$$\varphi^\circ(\sigma^\circ) \neq f(d, N). \tag{135}$$

References

[1] H.-G. Beyer, "Toward a Theory of Evolution Strategies: Some Asymptotical Results from the $(1 \stackrel{+}{,} \lambda)$-Theory," *Evolutionary Computation*, vol. 1, no. 2, pp. 165–188, 1993.
[2] H.-G. Beyer, "Toward a Theory of Evolution Strategies: The (μ, λ)-Theory," *Evolutionary Computation*, vol. 2, no. 4, pp. 381–407, 1995.
[3] H.-G. Beyer, "Toward a Theory of Evolution Strategies: On the Benefit of Sex – the $(\mu/\mu, \lambda)$-Theory," *Evolutionary Computation*, vol. 3, no. 1, 81–111, 1995.
[4] H.-G. Beyer, "Toward a Theory of Evolution Strategies: Self-Adaptation," *Evolutionary Computation*, vol. 3, no. 3, 311–347, 1996.
[5] H.-G. Beyer, "An Alternative Explanation for the Manner in which Genetic Algorithms Operate," *BioSystems*, vol. 41, pp. 1–15, 1997.
[6] H.-G. Beyer, *The Theory of Evolution Strategies.* Natural Computing Series. Heidelberg, Germany: Springer-Verlag, 2001, ISBN 3-540-67297-4.
[7] M. Herdy, "Reproductive Isolation as Strategy Parameter in Hierarchically Organized Evolution Strategies," in *Parallel Problem Solving from Nature, 2*, R. Männer and B. Manderick, Eds., Amsterdam, Netherlands: Elsevier, 1992, pp. 207–217.
[8] I. Rechenberg, *Evolutionsstrategie '94.* Stuttgart, Germany: Frommann–Holzboog Verlag, 1994.
[9] A. I. Oyman, H.-G. Beyer, and H.-P. Schwefel, "Convergence Behavior of the $(1 \stackrel{+}{,} \lambda)$ Evolution Strategy on the Ridge Functions," Tech. Rep. SyS–1/98, Department of Computer Science XI, University of Dortmund, Germany, Feb. 1998.
[10] A. I. Oyman, H.-G. Beyer, and H.-P. Schwefel. "Analysis of a Simple ES on the "Parabolic Ridge"," *Evolutionary Computation*, 8(3): 249–265, 2000.
[11] A. I. Oyman, "Convergence Behavior of Evolution Strategies on Ridge Functions," Ph.D. dissertation, Department of Computer Science, University of Dortmund, Germany, 1999.
[12] T. Bäck, U. Hammel, and H.-P Schwefel, "Evolutionary Computation: Comments on the History and Current State," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 3–17, 1997.
[13] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *A First Course in Order Statistics.* New York: Wiley, 1992.
[14] H.-G. Beyer, *Zur Analyse der Evolutionsstrategien.* Habilitation, Department of Computer Science, University of Dortmund, Germany, 1996.
[15] A. I. Oyman, H.-G. Beyer, and H.-P. Schwefel, "Where Elitists Start Limping: Evolution Strategies at Ridge Functions," in *Parallel Problem Solving from Nature, 2*, A. E. Eiben, T. Bäck, M. Schoenauer, and H.-P. Schwefel, Eds., Heidelberg, Germany: Springer-Verlag, 1998, pp. 34–43.
[16] A. I. Oyman, H.-G. Beyer, and H.-P. Schwefel, Analysis of a Simple ES on the "Parabolic Ridge," Tech. Rep. SyS–2/97, Department of Computer Science XI, University of Dortmund, Germany, Aug. 1997.