Mathware & Soft Computing 7 (2000) 35-75

Convergence Behavior of the $(1 \overset{+}{,} \lambda)$ Evolution Strategy on the Ridge Functions

A. Irfan Oyman, Hans–Georg Beyer & Hans–Paul Schwefel
University of Dortmund. Department of Computer Science
Systems Analysis Research Group
D-44221 Dortmund
e-mail: {oyman,beyer,schwefel}@LS11.informatik.uni-dortmund.de
Submitted by C. Moraga

Abstract

The convergence behavior of $(1 \overset{+}{,} \lambda)$–ES is investigated at parabolic ridge, sharp ridge, and at the general case of the ridge functions. The progress rate, the distance to the ridge axis, the success rate, and the success probability are used in the analysis.

The strong dependency of the $(1+\lambda)$–ES to the initial conditions is shown using parabolic ridge test function when low distances to the ridge axis are chosen as the start value. The progress rate curve and the success probability curve of the sharp ridge is explained quite exactly using a simple local model. Two members of the corridor model family are compared to some members of the ridge function family, and they do not seem to be the limit case of the ridge function family according to our measures for convergence behavior.

Keywords Evolution Strategy, progress rate, convergence analysis, ridge functions, corridor model

1 Introduction

The convergence behavior of evolutionary algorithms is not as intensely considered as the usage of them for optimization. The analysis of convergence is done on specific landscapes, which are models of the parts of general fitness landscapes. Such models must therefore be typical, they must present a specific property of possible landscapes. Therefore, the results obtained from typical landscapes are assumed to be useful in analyzing the progress behavior of evolutionary algorithms on more complex landscapes. Moreover, this analysis cannot be done generally directly on complex landscapes. As a result, the analysis of simple test functions becomes an important part of the theoretical analysis of evolutionary algorithms.

In this work, the convergence behavior of the $(1 \overset{+}{,} \lambda)$–ES is analyzed on different ridge functions. The whole work mainly consists of simulation results. In several

cases, they are supported by theoretical results. The derivation of the missing parts of the theoretical results is an essential part of our future work.

The underlying work is based on [OBS97], in which the parabolic ridge function is analyzed, which is a member of the ridge family. Here, the analysis continues on the general case. We give the simulation results on the sharp ridge test function. However, most of the underlying work consists of the results obtained for other members of the ridge function family (for $\alpha > 2$).

The algorithm used is described in Section 2. The definitions of the Evolution Strategy, the $(1, \lambda)$–ES, and the $(1+\lambda)$–ES are given in the same section. The ridge function as well as other functions are defined in Section 3. The functions which are used or referred to in this work are explained there (parabolic ridge, sharp ridge, corridor model, rectangular corridor, cylindrical corridor, etc.).

The measures used in the analysis of the convergence behavior are the expected mean distance to the progress axis ($R^{(\infty)}$), the progress rate (φ), the success rate ($sr(\lambda)$), and the success probability ($P_{s\ 1\overset{+}{,}\lambda}$). These measures are used already for the parabolic analysis of the ridge [OBS97]. In this work, they will be used for different members of the ridge family. These convergence measures are defined in Section 4.

In Section 5, the convergence behavior of $(1, \lambda)$–ES is explained at the family of ridge functions ($\alpha \geq 2$). The analysis is repeated in Section 6 for the $(1 + \lambda)$–ES, where the results are also compared for the two selection strategies.

We compare the progress behavior of the $(1 \overset{+}{,} \lambda)$–ES at the parabolic ridge in Section 7, for greater number of offspring ($\lambda = 100$ and $\lambda = 500$). In all the other sections of this work, $\lambda = 10$ is used. Section 8 is dedicated for the worst initial condition of the $(1 + \lambda)$–ES at the ridge functions: The ridge axis. Some simulation results at the parabolic ridge are given for $r^{(0)} = 0$ at $\lambda = 10$, and also for $r^{(0)} < R^{(\infty)}_{1+\lambda}$.

The simulation results obtained for the $(1 \overset{+}{,} 10)$–ES are summarized at the sharp ridge in Section 9; the progress rate formula for $(1, \lambda)$–ES is derived analytically. The two corridor model functions, namely rectangular corridor model and cylindrical corridor model, are investigated in Section 10 and Section 11, respectively. It has been conjectured that these functions might be regarded as the limit case of the ridge functions (as $\alpha \to \infty$). Mainly the progress rates of these corridor model functions are compared with the $\varphi_{1\overset{+}{,}\lambda}$ values of the ridge functions with high α.

Section 12 concludes this work with a short summary of the results and of the research planned for the near future.

2 Algorithm definition

The algorithm used in the simulations is shown in Figure 1. It belongs to the class of Evolutionary Algorithms, which imitate the processes of natural evolution. These processes are (among others) *mutation*, *crossover* (or recombination), and *selection*. The evaluation is done by the environment, in our case the objective function is used for this purpose. A more detailed description of the imitation of

this biological phenomenon is described in [FOW66, Rec73, Rec94, Hol75, Gol89, Sch95, Fog95, Bey96, Rud96].

The Evolution Strategies (ES) is the name of the Evolutionary Algorithm "species" developed in 1960s in the Technical University of Berlin [Sch95]. It imitates the mutation, selection, and recombination processes in the nature, by using (mostly normal) probability distribution functions for mutation, a deterministic mechanism for selection (which chooses just the best set of offspring individuals for the next generation), and a broad repertoire of recombination operators.

The principle in ES is to generate an excess number of offspring from the current generation. With excess we mean that the number of offspring produced is larger than the size of the population in the next generation. In early 1970s, the following notation was introduced for this case [Sch95]: μ represents the number of parents, and λ represents the number of offspring. The parent individual (or individuals for the recombination) are selected arbitrarily from the current generation. The simplest case here is $\mu = 1$, but then one has a parent individual and not a parent population.

The simplest case for the mutation is used in this work, namely isotropic mutation, which has equal properties along all directions (see Equation 1 for the realization of the mutations).

In ES, the normal distribution is used mostly for generating mutations, but this is not a must for the algorithm. The mutation operator is explained in detail further below in this section. One can also use different mutation strengths for each axis. For a more complex model, one can use a correlation *matrix* for the mutations in order to adjust the mutation distribution to the local topology independent of the coordinate axes. In these three cases, the set of *strategy parameters* consists of a single mutation strength σ, of an N–dimensional vector, and of $\frac{N}{2}(N+1)$ (co)variances respectively.

The selection scheme used in ES is different from tournament or proportional selection schemes described in [Gol89]. ES uses a deterministic selection scheme, which selects the best μ individuals as the individuals of the next generation. In the elitist case, these μ individuals are selected from the union set of μ parents and λ offspring. This is called plus strategy, and noted as $(\mu + \lambda)$. If the parent individuals do *not* take part in the selection, then the μ individuals of the next generation are selected from the λ offspring, which is named as comma strategy and notated as (μ, λ). For the plus strategy, one can even choose $\lambda = 1$. However, for the comma strategy, one must fulfill the condition $\lambda > \mu$ to ensure enough descendants for the selection operator. The notation $(\mu \overset{+}{,} \lambda)$ is used to subsume both selection schemes.

Recombination means the usage of the "genetic" information of more than one parent in the generation of each descendant, where one names the complete set of parameters of an individual as its genetic information. For recombination, several schemes can be used [BS93]. Along with the discrete or intermediate (weighted) recombination, where both of which can be considered on global or local scale, other recombination strategies also exist.

The variable setting (or the "genetic information") of an individual consists of its N–dimensional variable vector (which gives the geometrical position of the

individual in the N–dimensional search space), and its strategy parameters.

As a characteristic difference from *Genetic Algorithms* [Hol75, Gol89], the strategy parameters are also considered as part of the evolution process. This means that also the strategy parameters evolve during the process, they adapt to the (local) conditions. This is called the control of the mutation size, and is the first step towards "self-adaptation".

Even in its earliest development stage, ES contained self-adaptation methods for the strategy parameters (such as mutation strength). Self-adaptation is not considered in this work. However, the investigation done here will be the basis for the analysis of self-adaptation processes.

In this work, a simple $(1, \lambda)$–ES and a $(1 + \lambda)$–ES are used. The $(1, \lambda)$–ES is outlined in Figure 1. Since we use just one parent, we do not actually have a parental population. The entity $P^{(g)}$ refers to the parent individual at the generation number g, and $P^{(0)}$ to the starting point. $P'^{(g)}$ refers to the λ individuals generated from the parent $P^{(g)}$.

The $(1, \lambda)$*-ES Algorithm*

$g := 0$
initialize $P^{(0)}$
evaluate $P^{(0)}$
while not terminate do
    $P'^{(g)} :=$ *mutate* $P^{(g)}$
    evaluate $P'^{(g)}$
    $P^{(g+1)} := select_{(1,\lambda)} P'^{(g)}$
    $g := g + 1$
od

Figure 1: *The* $(1, \lambda)$*–ES Algorithm without self-adaptation*

In Figure 1, subroutines of the algorithm are written in *italic*. First, we *initialize* the parent $P^{(0)}$. The starting point can be chosen arbitrarily, however, in order to be able to compare the results with different mutation strengths, all the $P^{(0)}$ are initialized in the same manner. The *evaluate* function calculates the objective function value(s) of the parameter setting represented in the individual(s).

The *mutate* operator generates offspring according to

$$a_j^{'} := a_j + \sigma Z_{\mathcal{N}_j}. \tag{1}$$

This operator uses normally distributed random numbers, $Z_{\mathcal{N}} \sim \mathcal{N}(0, 1)$ (with zero mean and variance one) to generate the λ individuals in $P'^{(g)}$. New random samples are used for each individual and for each variable of all individuals. We denote the λ individuals of the offspring $P'^{(g)}$ as I_i such that $P'^{(g)} = \{I_i | i = 1, \ldots, \lambda\}$. The N variables of an individual I_i are called $a_j^{'}$; $I_i = \{a_j^{'} | j = 0, \ldots, N-1\}$, and of $P^{(g)}$ (the only parent) as $P^{(g)} = \{a_j | j = 0, \ldots, N-1\}$, using the subscript j. The $a_j^{'}$ are therefore $\mathcal{N}(a_j, \sigma^2)$ distributed. The mutation strength σ is provided by the user

at the beginning of the simulation, i.e. it is considered as an exogenous constant in this work.

The simulation lasts for a fixed number of generations given by the user, therefore the *terminate* function is also simple. The *select* operator records the individual having the highest objective function value among the offspring, and selects it as the parent of the next generation.

The parent of the offspring does not take part in the selection if one has the comma strategy. In case of the plus strategy, we denote the selection operator as $select_{(1+\lambda)} P'^{(g)} \cup P^{(g)}$. At the last step, the generation counter is incremented, and the loop restarts for the next generation.

3 Test functions

Several test functions are analyzed in the simulations of this work. In this section, these test functions will be listed, as well as the definitions of some related test functions. In this paper, the *ridge functions* are analyzed, and later the *rectangular* and *cylindrical corridor model*. The hyperplane test function and the sphere model are used in the comparisons made during the analysis. All the test functions given in this section are to be maximized by the $(1 \overset{+}{,} \lambda)$–Evolution Strategy.

The *sphere model* [Rec73, p.115] comprises all functions with spherically distributed quality function values. Thus the concentric hyperspheres are ordered around the optimum, which is chosen to be located in the coordinate origin in order to make the definition of the function simpler. The quality function value increases as the distance to the optimum decreases. In the general case, the sphere model can be expressed as

$$F(\mathbf{x}) = F_0 - f\left(\sqrt{\sum_{i=1}^{N} x_i^2}\right), \tag{2}$$

and $f(.)$ stands for any nondecreasing monotonous function. The general case of the *hyperplane* test function is defined as

$$F(\mathbf{x}) = a_0 + \sum_{i=1}^{N} a_i x_i, \qquad a_0 \in I\!R,\ a_i \in I\!R^+. \tag{3}$$

The coordinate axes can be rotated so that the $F(\mathbf{x})$ in (3) depends only on one variable, e.g. x_1. The ridge function family is defined by

$$F(\mathbf{x}) = x_0 - d\left(\sum_{i=1}^{N-1} x_i^2\right)^{\frac{\alpha}{2}}, \tag{4}$$

where $\alpha \in I\!R$, $d \in I\!R^+$. The positive factor d is used to scale the effect of the nonlinear part on the progress behavior. It influences the "hardness" of the optimization problem. The α parameter defines the ridge function type, for $\alpha = 1$ one

gets the sharp ridge, and for $\alpha = 2$ the parabolic ridge test function. The parabolic ridge was already analyzed in [OBS97]. Parabolic ridge and sharp ridge cases are shown for two variables ($N = 2$) in Figure 2 and Figure 3, respectively. For $\alpha = 0$ one gets the hyperplane.

Figure 2: *The parabolic ridge test function (i.e. $\alpha = 2$) for $N = 2$, $d = 0.01$. x_0 values are shown on the horizontal axis, x_1 on the vertical one. In this contour plot, the regions with larger quality function value are indicated in brighter color. Therefore, the optimum is at far right.*

Figure 3: *The sharp ridge (i.e. $\alpha = 1$) for $d = 0.01$ (figure left) and $d = 1$ (figure right), respectively. $N = 2$. See also the legend of Figure 2.*

The neighborhood of a point $\mathbf{x}$ with $F(\mathbf{x})$ consisting of points with $F(\mathbf{x}') \geq F(\mathbf{x})$ is called success region. The success region of the ridge functions is not bounded.

The optimizer of this function is given as

$$\hat{x}_0 \to +\infty; \qquad \hat{x}_i = 0, \qquad 1 \le i \le N-1. \tag{5}$$

The square root of the sum of the squares is named as the distance to the progress axis, i.e.

$$r := \sqrt{\sum_{i=1}^{N-1} x_i^2}. \tag{6}$$

We also analyze the progress behavior of the Evolution Strategy on the *corridor model* test functions. The common property of the corridor model functions is the narrow corridor which is obtained using *restrictions* in all directions except one: the progress direction. Two corridor models are investigated, which are named after the shape of the corridor walls: The *rectangular* corridor model [Sch95, p.134, p.351] and the *cylindrical* corridor model [Sch95, p.361]. In the problem catalog of [Sch95], the restrictions and the objective function are given for the general case of these two models. They will be simplified so that one has the x_0 axis as the progress axis, and the constraint boundaries parallel to the x_0 axis. Therefore, the rectangular corridor model is given as

$$F(\mathbf{x}) = \begin{cases} cx_0 & \text{if all } G_j(\mathbf{x}) \text{ fulfilled} \\ -\infty & \text{otherwise} \end{cases} \tag{7}$$

$$G_j(\mathbf{x}) : |x_j| \le b, \qquad b \in I\!R^+, \text{ for } \quad j = 1, \ldots, N-1. \tag{8}$$

The cylindrical corridor model is defined as

$$F(\mathbf{x}) = \begin{cases} cx_0 & \text{if } G(\mathbf{x}) \text{ is fulfilled} \\ -\infty & \text{otherwise} \end{cases} \tag{9}$$

$$G(\mathbf{x}) : \sqrt{\sum_{i=1}^{N-1} x_i^2} \le b', \qquad b' \in I\!R^+. \tag{10}$$

We selected $c = 1$ for our simulations (in general, $c \in I\!R^+$).

Note, linear transformations of the object variables, i.e. arbitrary rotations of the coordinate system, do not change the nature of the problem for the Evolution Strategy. Since isotropic mutations are used, the progress behavior (and in particular progress rate) is not affected by the rotations of the reference frame. This is a desired property for evolutionary algorithms [Sal96].

4 The progress measures

The behavior of an evolutionary algorithm can be defined as the long term evolution of the population in the time domain. In order to derive its dynamic behavior, the behavior of the ES in a single generation must be analyzed. As a result of this

analysis, one obtains the expected values (first order moments) of respective quantities. If these results can determined for all possible values of the state variables, one obtains an estimate for the overall behavior of the algorithm. The measures concerning the changes in a single generation are called local measures (local in time, *not* necessarily in space). In this work, the measures of interest are φ, $\overline{Q}$, $sr(\lambda)$, $P_{s\,1^{+}_{,}\lambda}$, and r. They are defined below.

The progress rate φ is usually defined as the expected change in the distance of the population to the optimum in a single generation

$$\varphi_{1^{+}_{,}\lambda} := \mathrm{E}\{\|\hat{\mathbf{x}} - \mathbf{x}^{(g)}\| - \|\hat{\mathbf{x}} - \mathbf{x}^{(g+1)}\|\} \ , \tag{11}$$

where $\hat{\mathbf{x}}$ denotes the optimum vector. That is, φ is positive if this distance is expected to decrease. Without reference to (11), the progress rate φ for the hyperplane test function (Eq. (3)) is measured in the direction of the normal vector of the plane. Similarly, for ridge functions (Eq. (4)), φ is simply defined as the expected difference in the x_0 variable of the successive parental generations, since the optimum is at infinity.

In contrast to φ, the other progress measure $\overline{Q}$ is measured in the scalar domain of fitness function values. It gives the expected change in the quality function value of the population in a single generation

$$\overline{Q}_{1^{+}_{,}\lambda} := \mathrm{E}\{F^{(g+1)} - F^{(g)}\} \ . \tag{12}$$

For the corridor models (defined in (7) and (9)), the relation between the two progress measures is as follows: $\overline{Q}_{1,\lambda} = c\varphi$.

The success rate $sr(\lambda)$ is the probability of getting a descendant with a better quality function value than the parent in a *single* trial. The computation of $sr(\lambda)$ and success measures is simpler for the $(1,\lambda)$-ES case, than for the plus case.

The local quality function ($Q_{\mathbf{x}}(\mathbf{z})$) is required for the formal definition of $sr(\lambda)$. $Q_{\mathbf{x}}(\mathbf{z})$ is defined as the difference of the fitness function values of successive generations, i.e. $Q_{\mathbf{x}}(\mathbf{z}) := F(\mathbf{x}_{\mathrm{P}} + \mathbf{z}) - F(\mathbf{x}_{\mathrm{P}})$, $\mathbf{x}_{\mathrm{P}}$ being the position of the parent individual, and $\mathbf{z}$ the mutation vector yielding the best descendant. If the expected value and standard deviation of $Q_{\mathbf{x}}(\mathbf{z})$ are named as $M_Q := \mathrm{E}\{Q_{\mathbf{x}}(\mathbf{z})\}$ and $S_Q := \sqrt{\mathrm{E}\{Q^2_{\mathbf{x}}(\mathbf{z})\} - [\mathrm{E}\{Q_{\mathbf{x}}(\mathbf{z})\}]^2}$, respectively, the cumulative distribution function of being better than the parent can be approximated as (see [Bey94])

$$sr(\lambda) \approx 1 - \Phi\left(\frac{0 - M_Q}{S_Q}\right) = \Phi\left(\frac{M_Q}{S_Q}\right) \ , \tag{13}$$

where "$\Phi(.)$" denotes the cumulative distribution function of the standard normal distribution, $\mathcal{N}(0,1)$.

The success probability $P_{s\,1^{+}_{,}\lambda}$ is the probability that at least one of the λ descendants will have a better objective function value than the parent. $P_{s\,1,\lambda}$ can simply computed as

$$P_{s\,1,\lambda} = 1 - (1 - sr(\lambda))^{\lambda} \approx 1 - \left[\Phi\left(-\frac{M_Q}{S_Q}\right)\right]^{\lambda} \ . \tag{14}$$

The distance r to the progress axis in the ridge function case is given by (6). It gives the distance of the population mean in the N-dimensional space to the progress axis, and is needed in the formulae of φ, $\overline{Q}$, and $sr(\lambda)$.

The value of r fluctuates in a simulation in consecutive generations. If one calls the r value at generation g as $r^{(g)}$, the $r^{(g)}$ values averaged over many generations give the same value in different simulations of the same ES for the same values of α, σ, d, N, and λ after a certain transient period. This mean value for the observed mean r will be denoted as

$$R^{(\infty)} := \mathrm{E}\{r\} \ . \tag{15}$$

5 The $(1,\lambda)$–ES performance

In this section we will give the simulation results for the general case of the ridge functions (for $\alpha \in \{2,3,4,5,8\}$) using $(1,\lambda)$–ES, with $\lambda = 10$, $d = 0.01$, and $N = 100$. The simulation length is chosen as $G = 100,000$ generations; $2,000$ generations are reserved additionally at the beginning of the simulations to avoid influences from the transient phase.

The results for the measures of the performance behavior are reported in separate sections. the simulation results are summarized for the distance to the ridge axis ($R^{(\infty)}$) in Section 5.1. The results on the progress rate (φ^*) are given in Section 5.2. Lastly, in Section 5.3, the success probability ($P_{s\,1,\lambda}$) and success rate ($sr(\lambda)$) values observed are reported.

5.1 The distance to the ridge axis, r

This section gives the simulation results for the $\mathrm{E}\{r\}$ values of some ridge function family members with $\alpha \geq 2$. The r value was defined in (6).

The expected mean of r during the simulation runs (i.e. $R^{(\infty)}$ for $\alpha \geq 2$) is around $D^{(\infty)}$, for sufficiently large mutation strength. The value of $D^{(\infty)}$ is obtained from the sphere model theory [Bey93, p.186], in the following equation it is expressed for $N-1$ dimensions. A comparison of the definitions of the sphere model in (2) and of the ridge functions in (4) indicates the similarity of these two test functions

$$R^{(\infty)} := \mathrm{E}\{r\} \approx D^{(\infty)} = \frac{\sigma(N-1)}{2c_{1,\lambda}} \ . \tag{16}$$

The progress coefficient $c_{1,\lambda}$ for the $(1,\lambda)$-ES is the expectation of the λ-th order statistics of λ samples generated using the standard normal distribution,

$$c_{1,\lambda} = \frac{\lambda}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t\mathrm{e}^{-\frac{1}{2}t^2}[\Phi(t)]^{\lambda-1}\mathrm{d}t \ . \tag{17}$$

As an example, $c_{1,10} \approx 1.53875$.

As already mentioned, the quantity $R^{(\infty)}$ gives the expected distance to the optimum of a $(1,\lambda)$–ES with fixed mutation strength σ as the simulation length

G goes to infinity. The formula for $R^{(\infty)}$ for $\alpha = 2$ (parabolic ridge case) was derived analytically in [OBS97]. In this section, the condition for the validity of approximation $D^{(\infty)}$ for $\mathrm{E}\{r\}$ ($\alpha > 2$) will be shown. One uses the normalizations

$$\sigma^* = d^{\frac{1}{\alpha-1}}(N-1)\sigma \tag{18}$$

$$R^{(\infty)^*} = d^{\frac{1}{\alpha-1}}(N-1)R^{(\infty)} \tag{19}$$

in order to be able to compare the simulation results for different α.

PSfrag replacements
$R^{(\infty)^*}$, [normalized]
σ^*, sigma [normalized]
$D^{(\infty)}$
$\alpha = 2$
$\alpha = 3$
$\alpha = 4$
$\alpha = 5$
$\alpha = 8$

PSfrag replacements
$R^{(\infty)^*}$, [normalized]
σ^*, sigma [normalized]
$D^{(\infty)}$
$\alpha = 2$
$\alpha = 3$
$\alpha = 4$
$\alpha = 5$
$\alpha = 8$

Figure 4: *Normalized distance to the ridge axis versus normalized mutation strength for the $(1, 10)$–ES. Three members of the ridge family are displayed in the figures ($\alpha \geq 2$). The $\mathrm{E}\{r\}$ value of the ridge functions with $\alpha \geq 3$ can be approximated by $D^{(\infty)}$ (Eq. (16)) for $\sigma^* \geq 2c_{1,10} \approx 3$.*

Simulations are made for $\alpha \in \{2, 3, 4, 5, 8\}$; the results for $\alpha = 2, 4, 8$ are given in Figure 4. For $\sigma^* \geq 2c_{1,10} \approx 3$, one can use the approximation $\mathrm{E}\{r\} \approx D^{(\infty)}$. The dependence of the progress rate φ^* to the mutation strength σ^* will be investigated in the next section. $D^{(\infty)}$ will be used in the progress rate formula, and obtain accurate results for $\sigma^* \geq 2c_{1,10}$. For $\sigma^* < 2c_{1,10}$, the simulation results for the mean r should be used instead. Therefore, for smaller values of the mutation strength, one needs a more accurate theoretical formula for $\mathrm{E}\{r\}$.

Finding an accurate analytic formula for $\mathrm{E}\{r\}$ for small values of σ^* is a part of our research ongoing. The $\mathrm{E}\{r\}$ value for $\alpha < 2$ will also be investigated by simulations. For $\alpha = 2$, however, $\mathrm{E}\{r\}$ was already derived in [OBS97, p.29]

$$R^{(\infty)}_{1,\lambda}\Big|_{\alpha=2} = D^{(\infty)} \sqrt{\frac{1}{2}\left[1 + \sqrt{1 + \left(\frac{1}{dD^{(\infty)}}\right)^2}\right] - \frac{2c^2_{1,\lambda}}{N-1}} \tag{20}$$

with $D^{(\infty)}$ given by (16).

5.2 The progress rate φ

In this section,the simulation results will be given for the progress rate φ of some members of the ridge family. An approximation with first order partial derivatives

for the theoretical φ is used to explain the simulation results. This formula is obtained using a local model at the steady state, i.e. when the $r^{(g)}$ has a value around $R^{(\infty)}$.

The φ formula to be derived for in this subsection is very accurate if the $R^{(\infty)}$ values from the simulations are used, if $D^{(\infty)}$ in (16) is used instead, the φ formula is valid for sufficiently large σ^* (see Figure 4 for the reason). The idea of the following derivation has been published in [OBS98]. Its predictive power has been tested for the case $\alpha = 2$ only. Here, the usefulness of this approach will be shown for other α values.

5.2.1 On the derivation of the φ^* formula

The analysis starts with substituting r defined in (6) into the ridge function definition in (4), yielding

$$F(x_0, r) = x_0 - dr^\alpha. \tag{21}$$

At ridge functions, the isometric surfaces can locally be approximated by *hyperplanes* for $r \approx R^{(\infty)}$ (see Figure 2). As a result, the progress rate in the direction of the gradient vector $\mathbf{a}$ is as large as the φ of the hyperplane, i.e. $\varphi_{\mathbf{a}} \approx c_{1,\lambda}\sigma\mathbf{a}/\|\mathbf{a}\|$. If the unit vector in x_0 direction is denoted by $\mathbf{e}_0$, the resulting progress rate can be written as $\varphi_{ridge} \approx \mathbf{e}_0^T \cdot \varphi_{\mathbf{a}}$.

By partially differentiating (21) with respect to x_0 and r, the gradient vector $\mathbf{a}$ at the location (x_0, r) is obtained as

$$\mathbf{a} = \begin{pmatrix} \frac{\partial F}{\partial x_0} \\ \\ \frac{\partial F}{\partial r} \end{pmatrix} = \begin{pmatrix} 1 \\ -d\alpha r^{\alpha-1} \end{pmatrix}. \tag{22}$$

Using the quantity $\mathbf{a}$, one gets a first estimate for the progress rate φ

$$\varphi_{ridge} \approx \frac{\sigma c_{1,\lambda}}{\|\mathbf{a}\|} = \frac{\sigma c_{1,\lambda}}{\sqrt{1 + (d\alpha r^{\alpha-1})^2}}, \tag{23}$$

and after substituting r by $D^{(\infty)}$ for sufficiently large σ

$$\varphi_{ridge} \approx \frac{\sigma c_{1,\lambda}}{\sqrt{1 + \left[d\alpha \left(\frac{\sigma(N-1)}{2c_{1,\lambda}}\right)^{\alpha-1}\right]^2}}. \tag{24}$$

The approximation $r \approx D^{(\infty)}$ is valid for $\sigma^* \geq 2c_{1,\lambda}$, as shown in Section 5.1.

Now, one can introduce a normalization scheme which allows us to compare the progress rates of different ridge functions. This scheme is actually a generalization of the scheme introduced for the parabolic ridge test function in [OBS97]

$$\varphi^* = d^{\frac{1}{\alpha-1}}(N-1)\varphi. \tag{25}$$

The normalization of σ has been given already by Eq. (18). Using this scheme, the approximate formula for the progress rate reduces to

$$\varphi^*_{ridge} \approx \frac{\sigma^* c_{1,\lambda}}{\sqrt{1+\alpha^2\left(\frac{\sigma^*}{2c_{1,\lambda}}\right)^{2\alpha-2}}}. \tag{26}$$

5.2.2 Simulation results

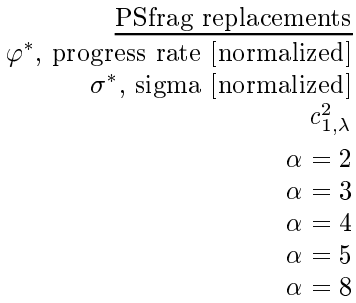


Figure 5: *Normalized progress rate versus normalized mutation strength for the* $(1, 10)$*–ES. Five members of the ridge family are displayed. The theoretical* φ^* *curves for parabolic ridge (i.e.* $\alpha = 2$*, Eq. (28)) and hyperplane function (*$\alpha = 0$*) are also indicated for comparison, as well as the horizontal asymptote* $c^2_{1,\lambda}$*.*

In Figure 5 the normalized progress rate curves are given for normalized mutation strength. Firstly, all of the ridge functions (like other function classes) have progress rate values less than $\varphi^*_{hyperplane}$. This can also be seen from (24), since the denominator is never smaller than one. The value φ^*_{ridge} is always nonnegative, since there is always a component of the mutation vectors in direction toward the optimum. Secondly, the progress rates of all ridge functions for $\alpha > 2$ start to decrease, for sufficiently large σ^* values. For $\alpha \geq 3$, this decrease starts at $\sigma^* \overset{<}{\approx} 2$. The φ^* value for the cases $\alpha = 2 \pm \varepsilon$ ($\varepsilon \ll 1$, i.e. for α values very near to two) are investigated using (26). Simulation runs are also made for $\alpha = 2.1$ and $\alpha = 2.5$, and obtained results supporting this claim. Therefore, the limit value of φ^* using the $(1, \lambda)$–ES for increasing mutation strength is

$$\lim_{\sigma^*\to\infty} \varphi^*_{ridge}\Big|_{\alpha<2} = \infty, \qquad \lim_{\sigma^*\to\infty} \varphi^*_{ridge}\Big|_{\alpha>2} = 0. \tag{27}$$

For $\alpha = 2$, one has $c^2_{1,\lambda}$ as the limit value (see Figure 5 and Section 7 for details). The accurate $\varphi^*_{1,\lambda}$ for $\alpha = 2$ is [OBS97, p.22]

$$\varphi^*_{1,\lambda}\Big|_{\alpha=2} = \frac{c^2_{1,\lambda}}{\sqrt{\frac{c^2_{1,\lambda}}{\sigma^{*2}} + \frac{1}{2}\left(1 + \sqrt{1 + \left(\frac{2c_{1,\lambda}}{\sigma^*}\right)^2}\right)}}. \tag{28}$$

After explaining the progress behavior of the ridge family, it will be shown how good the approximate progress rate formula (26) is for finite σ^*. One considers the two extremes from Figure 5, namely $\alpha = 2$ and $\alpha = 8$. In Figure 6, one sees that the approximate formula in Equation 23 gives accurate results for $\alpha = 2$ using the r values obtained from the simulations.

PSfrag replacements
φ^*, progress rate [normalized]
σ^*, sigma [normalized]
$c^2_{1,\lambda}$
$\alpha = 2$

Figure 6: *Simulation results compared with theory for the* $(1,10)$*–ES with* $\alpha = 2$*. The simple estimate from (24) (with the approximation* $r \approx D^{(\infty)}$*, indicated as theo) differs much from the simulation results* ($\alpha = 2$)*; however, it is asymptotically correct. If one uses the approximation in (23), and insert the mean value of* r *obtained from the simulations, one gets the curve "*`theo |r`*". The analytically derived formula for* φ^* *(Eq. (28)) is indicated as "*`theo2`*".*

The same idea can be tested for $\alpha = 8$. As one sees in Figure 7, one can compute the φ^* value by inserting the mean of the r values from the simulations in Equation (23) (and applying the normalization afterwards) quite accurately. This is quite surprising since this simple estimate for φ^* was obtained by using first order partial derivatives only. One can expect this method to work also for a wide range of α values. The other theoretical curve obtained using (24) gives poor results. Therefore, finding a theoretical (if possible an analytically driven) equation for $R^{(\infty)}$, i.e. the expected mean value of r for a given α remains as a future work.

The φ^*–versus–σ^* plots for $\alpha < 2$ have as their asymptotes curves with a positive slope. The slopes of the asymptotes increase up to $c_{1,\lambda}$ as α is decreased

PSfrag replacements
φ^*, progress rate [normalized]
σ^*, sigma [normalized]
$c^2_{1,\lambda}$
$\alpha = 8$

Figure 7: *Normalized progress rate is plotted versus normalized mutation strength for the* $(1, 10)$*–ES and* $\alpha = 8$*. Again, the simulation results are almost exactly explained by "*`theo |r`*". See also the legend of Figure 6.*

down to zero (remember that $\varphi_{hyperplane} = c_{1,\lambda}\sigma$). For $\alpha < 0$, the simulation results coincide with the theoretical progress rate curve of the hyperplane.

Therefore, one can say that the *parabolic ridge* test function is the *only* member of the ridge function family with a horizontal asymptote ($c^2_{1,\lambda}$), which is somewhat obvious after considering (26).

5.3 The success rate $sr(\lambda)$ and the success probability $P_{s\,1,\lambda}$

The "success" is measured using two measures: The probability of having a descendant which has a better quality function value than its parent is called *success rate* ($sr(\lambda)$). The probability that at least one of the λ offspring will have a better quality function value than its parent is called *success probability* ($P_{s\,1,\lambda}$).

The $sr(\lambda)$ and $P_{s\,1,\lambda}$ values were investigated in [OBS97] for the parabolic ridge case ($\alpha = 2$). The analytical formula for the success rate using the comma selection strategy is [OBS97, p.31f]

$$sr(\lambda) \approx \Phi\left[-\left(\frac{1}{\sigma^{*2}} + \frac{1}{2c^2_{1,\lambda}}\left[1 + \sqrt{1 + \left(\frac{2c_{1,\lambda}}{\sigma^*}\right)^2}\right]\right)^{-\frac{1}{2}}\right] \tag{29}$$

with the theoretical limit given as

$$\lim_{\sigma^* \to \infty} sr(\lambda) = \Phi(-c_{1,\lambda}). \tag{30}$$

The analytical formula for the success probability $P_{s\,1,\lambda}$ is [OBS97, p.31f]

$$P_{s\,1,\lambda} \approx 1 - \left\{ \Phi \left[\left(\frac{1}{\sigma^{*2}} + \frac{1}{2c_{1,\lambda}^2} \left[1 + \sqrt{1 + \left(\frac{2c_{1,\lambda}}{\sigma^*} \right)^2} \right] \right)^{-\frac{1}{2}} \right] \right\}^{\lambda} \tag{31}$$

with the theoretical limit given as

$$\lim_{\sigma^* \to \infty} P_{s\,1,\lambda} = 1 - [\Phi(c_{1,\lambda})]^{\lambda}. \tag{32}$$

In this section, the simulation results are given for $\alpha = 2, 4, 8$. We also made simulation runs for $\alpha = 3$ and $\alpha = 5$. Since the success curves changed gradually as α is increased, these two intermediary curves are removed from the pictures.

PSfrag replacements
$sr(10)$, success rate
σ^*, sigma [normalized]
$\alpha = 2$
$\alpha = 3$
$\alpha = 4$
$\alpha = 5$
$\alpha = 8$
theo $\alpha = 2$

Figure 8: *The success rate curves versus normalized mutation strength for different members of the ridge family,* $\lambda = 10$. *The analytically obtained curve for the parabolic ridge is also indicated in the figure (theo* $\alpha = 2$, *Eq. (29)): The asymptote* $\Phi(-c_{1,10})$ *(Eq. (30)) is reached faster at larger* α *values.*

For $\lambda = 10$, one obtains $sr(10, \alpha) \geq 0.062$ and $P_{s\,1,10} \geq 0.472$. Therefore, the limits for the parabolic ridge seem also to be valid for the other members of the ridge family (see [OBS97, p.31] for the analytical derivation of $sr(\lambda)$ and $P_{s\,1,\lambda}$ for the parabolic ridge). The limit value is attained faster (i.e. at smaller σ^* values) for larger values of α.

The limit value of $P_{s\,1,\lambda}$ in (32) changes very little as λ is increased. For $\lambda = 3$ it is around 0.485, for $\lambda = 10$ it decreases to 0.472. For $\lambda = 1,000$ and $100,000$ one gets the values 0.448 and 0.441, respectively.

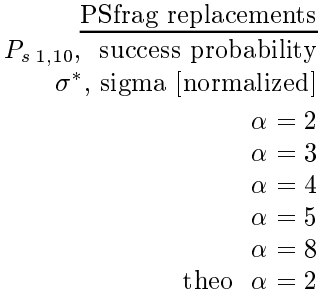


Figure 9: *The success probability curves versus normalized mutation strength for different members of the ridge family, $\lambda = 10$. The analytically obtained curve for the parabolic ridge is also indicated in the figure (theo $\alpha = 2$, Eq. (31)). The asymptote $1 - [\Phi(c_{1,10})]^{\lambda}$ (Eq. (32)) is reached faster at larger α values.*

6 The $(1 + \lambda)$–ES performance

The simulation results obtained using $(1, \lambda)$–ES are reported for $\alpha \geq 2$ in Section 5. We will report here the simulation results for $R^{(\infty)}$ (Section 6.1), $P_{s\,1\overset{+}{,}\lambda}$ (Section 6.2), and φ^* (Section 6.3). In these sections, the $(1, \lambda)$–ES and $(1 + \lambda)$–ES, will be compared and investigate the limit behaviors of these strategies. The simulation runs were done for $\alpha \in \{2, 3, 4, 5, 8, 64\}$, $d = 0.01$, $N = 100$, and $\lambda = 10$. In order to make the diagrams more readable, the results are not displayed for all these α values. Each simulation run lasted $G = 100,000$ generations, and additional $2,000$ generations are reserved for the transient phase.

6.1 The distance to the ridge axis, $R^{(\infty)}_{1\overset{+}{,}\lambda}$

Here the dependence of the $R^{(\infty)}_{1\overset{+}{,}\lambda}$ values on σ^* will be investigated. Furthermore, we will show the dependence of $R^{(\infty)}_{1\overset{+}{,}\lambda}$ on α and the selection strategy. The $R^{(\infty)}_{1\overset{+}{,}\lambda}$ values are normalized according to (19).

In Figure 10, the $R^{(\infty)}$ values of plus and comma strategies are compared. For $\alpha = 64$ and $\sigma^* < 2c_{1,10} \approx 3$, this value is roughly the same at both selection strategies. On the figure left, the same comparison is done at $\alpha = 8$, and the curves separate from each other at lower σ^* values than for $\alpha = 64$. In both figures, one can see that the $R^{(\infty)}_{1+\lambda}$ value is smaller than the $R^{(\infty)}_{1,\lambda}$ value. For small σ^* values, the $R^{(\infty)^*}$ value gets larger for larger α as can be seen in Figure 11.

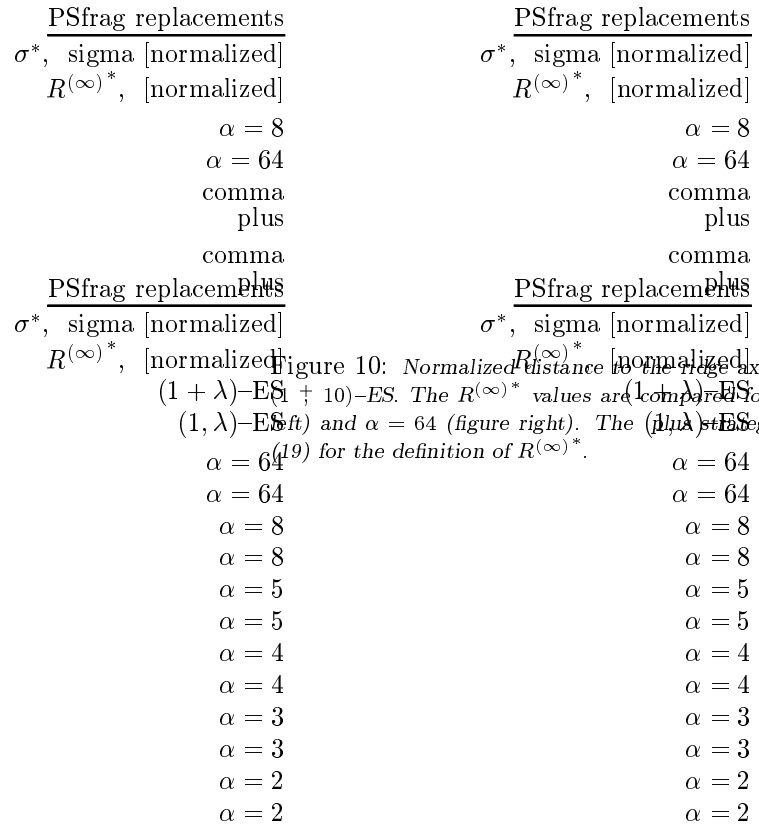


Figure 10: *Normalized distance to the ridge axis versus normalized mutation strength for the* $(1\overset{+}{,}10)$*–ES. The* $R^{(\infty)^*}$ *values are compared for the two selection strategies at* $\alpha = 8$ *(figure left) and* $\alpha = 64$ *(figure right). The* (μ, λ)*–strategy has a lower* $R^{(\infty)^*}$ *value for larger* σ^**. See (19) for the definition of* $R^{(\infty)^*}$*.*

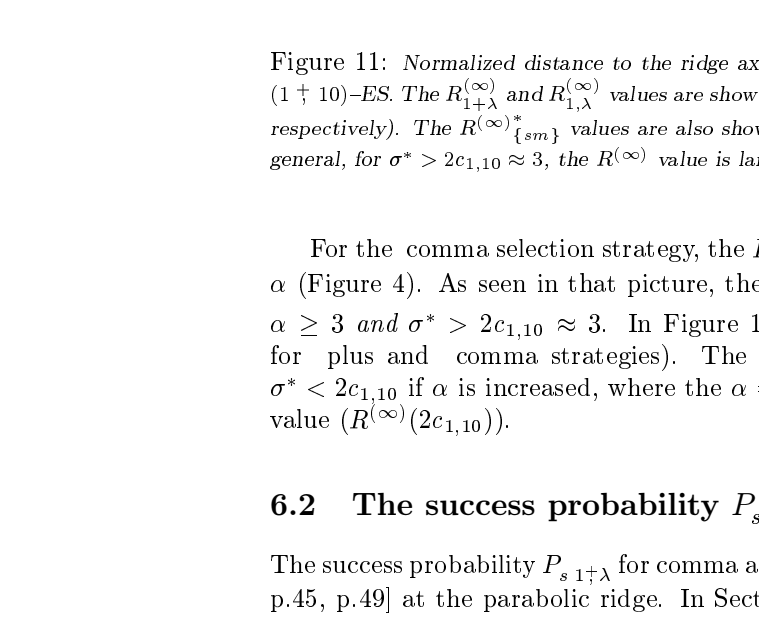


Figure 11: *Normalized distance to the ridge axis versus normalized mutation strength for the* $(1\overset{+}{,}10)$*–ES. The* $R^{(\infty)}_{1+\lambda}$ *and* $R^{(\infty)}_{1,\lambda}$ *values are shown next to each other (figure left and figure right, respectively). The* $R^{(\infty)^*}_{\{sm\}}$ *values are also shown in both pictures to ease the comparison. In general, for* $\sigma^* > 2c_{1,10} \approx 3$*, the* $R^{(\infty)}$ *value is larger for the comma strategy.*

For the comma selection strategy, the $R^{(\infty)^*}$ values are compared for increasing α (Figure 4). As seen in that picture, the $R^{(\infty)}_{1,\lambda}$ values are very near to $D^{(\infty)}$ for $\alpha \geq 3$ *and* $\sigma^* > 2c_{1,10} \approx 3$. In Figure 11, the $R^{(\infty)^*}$ values are compared (i.e. for plus and comma strategies). The $R^{(\infty)}$ value increases in both cases for $\sigma^* < 2c_{1,10}$ if α is increased, where the $\alpha = 64$ curves seem to be very near to the value $(R^{(\infty)}(2c_{1,10}))$.

6.2 The success probability $P_{s\,1\overset{+}{,}\lambda}$

The success probability $P_{s\,1\overset{+}{,}\lambda}$ for comma and plus strategies is investigated in [OBS97, p.45, p.49] at the parabolic ridge. In Section 5.3, the simulation results obtained

using $(1,10)$–ES are given for several α values. Here we will compare the $P_{s\,1^{+}_{,}\lambda}$ results for $\alpha \in \{2,3,4,5,8,64\}$. The results for $\alpha = 3$ and $\alpha = 5$ are not displayed in Figure 13, in order to reduce the number of curves per diagram.

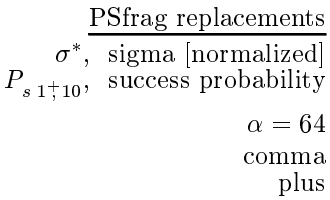


Figure 12: *The success probability curves versus normalized mutation strength for $\alpha = 64$. The limit value for $P_{s\,1,10}$ obtained at the parabolic ridge is also valid for $\alpha = 64$. On the other hand, the $P_{s\,1+10}$ value continues to decrease as σ^* is increased.*

The success probability for $\alpha = 64$ is shown in Figure 12. The curves for both selection strategies differ from each other essentially for $\sigma^* > 2c_{1,10} \approx 3$; and the $P_{s\,1+\lambda}$ continues to decrease although $P_{s\,1,\lambda}$ goes to the limit predicted by (32).

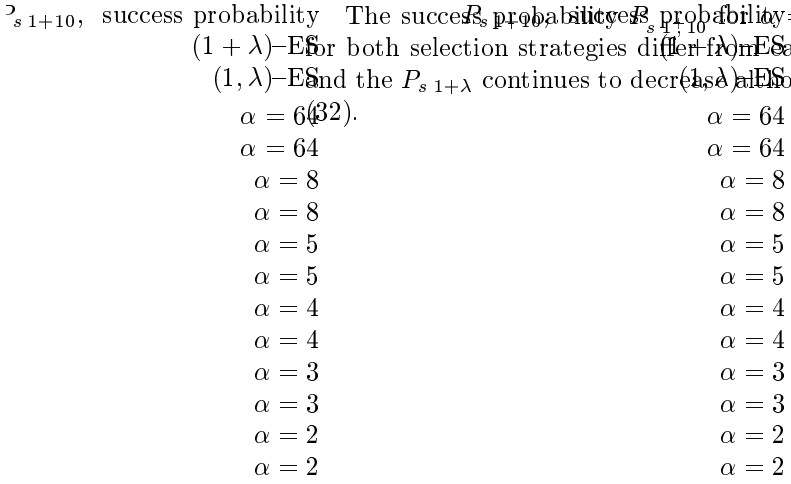


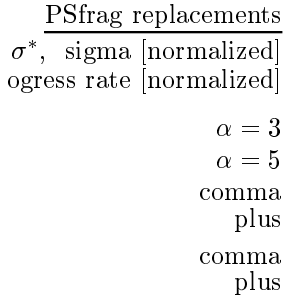
Figure 13: *The success probability curves versus normalized mutation strength for different members of the ridge family. The $P_{s\,1+10}$ curves (figure left) and the $P_{s\,1,10}$ curves (figure right) are plotted next to each other for different values of α. The tendency at the comma strategy is easier to see: for very large values of α, $P_{s\,1,10}$ decreases suddenly near $\sigma^* = 2c_{1,10} \approx 3$ to the asymptotic limit. The $P_{s\,1+10}$ value, however, decreases enormously for increasing σ^*.*

The success probability curves differ enormously for plus and comma selection. In Figure 12, one observes the difference for $\alpha = 64$; in Figure 13, one can see the

difference for several increasing α values. The $P_{s\ 1,\lambda}$ curve has an asymptotic lower limit, which is attained almost at $\sigma^* \approx 2c_{1,10} \approx 3$ for $\alpha = 64$. The $P_{s\ 1+\lambda}$ curve, however, goes to zero for larger σ^*.

6.3 The progress rate $\varphi^*_{1\overset{+}{,}\lambda}$

In this section, the progress rates of the ridge function family (with $\alpha \geq 2$) will be compared for plus and comma selection strategy. This convergence measure was compared for the $\alpha = 2$ case (parabolic ridge) in [OBS97, p.44]. In Section 5.2, it was investigated for $2 \leq \alpha \leq 8$ using $(1, 10)$–ES.



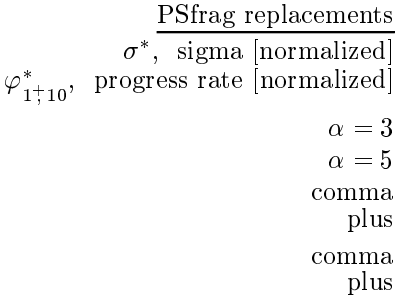


Figure 14: *Normalized progress rate versus normalized mutation strength. The comparison of the ridge functions for* $\alpha = 3$ *and* $\alpha = 5$*. The value* $c^2_{1,10}$ *is also shown in the pictures. The progress rate of the plus strategy surpasses the* $\varphi^*_{1,10}$ *and also the* $\hat{\varphi}^*_{1,10}$ *for* $\alpha = 5$ *(figure right).*

The $\varphi^*_{1\overset{+}{,}\lambda}$ curves are compared for $\alpha = 3$ and $\alpha = 5$ in Figure 14. The comma strategy has better progress rates for any σ^* value at $\alpha = 3$. For $\alpha = 4$, the peak values of these selection strategies are very near to each other; however, the comma strategy shows larger progress rates for any σ^*.

For $\alpha = 5$ (Figure 14 right), the $\hat{\varphi}^*_{1+\lambda}$ value is slightly larger than $\hat{\varphi}^*_{1,\lambda}$. The $\varphi^*_{1+\lambda}$ is larger in the interval $1.6 \overset{<}{\approx} \sigma^* \overset{<}{\approx} 4.5$. The *shape* of the $\varphi^*_{1+\lambda}$ curve does change only a little bit as α is increased from three to five; however, the change at the $\varphi^*_{1,\lambda}$ curve foretells us that the plus strategy will give comparably better progress rates for larger α.

The progress rate of plus and comma strategies for $\alpha = 64$ are compared in Figure 15. For small mutation strengths, both strategies give almost the same results; otherwise (for $\sigma^* \overset{>}{\approx} 1.4$), the $\varphi^*_{1+\lambda}$ value is definitely better.

The drastic difference at $\sigma^* = 2c_{1,10} \approx 3.08$, where the plus strategy is more than *five* times better than the comma selection strategy, will be examined more thoroughly. The $P_{s\ 1\overset{+}{,}\lambda}$ values are almost the same at this σ^* value ($P_{s\ 1,\lambda} \approx 0.51$, $P_{s\ 1+\lambda} \approx 0.53$, Figure 12). The $R^{(\infty)}_{1\overset{+}{,}\lambda}$ values are also almost the same (the $R^{(\infty)}_{1,\lambda}$ is about five per cent greater than $R^{(\infty)}_{1+\lambda}$, Figure 10). Neither the success probability

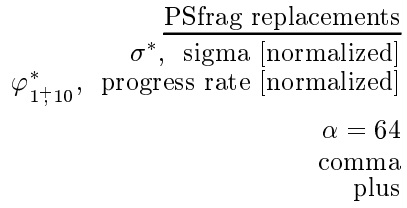


Figure 15: *Normalized progress rate versus normalized mutation strength for the ridge function with $\alpha = 64$. The value $c^2_{1,10}$ is also shown in the picture. For $\sigma^* \lessapprox 1.4$, both selection strategies give similar results; otherwise, the plus strategy is better.*

figures nor the $R^{(\infty)}$ figures can explain the progress rate difference between these two strategies. For this purpose, Figure 16 can be used.

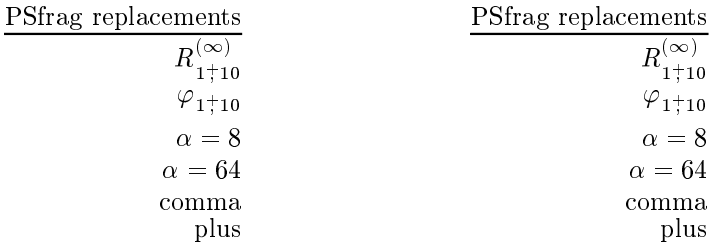


Figure 16: *The progress rate versus distance to the ridge axis. Both axes are unnormalized. The φ values are compared for the selection strategies at $\alpha = 8$ (figure left) and $\alpha = 64$ (figure right). The φ_{1+10} value is greater than $\varphi_{1,10}$ on the figure right, for $R^{(\infty)} > 1$.*

The φ–versus–$R^{(\infty)}$ plots for $\alpha = 8$ and $\alpha = 64$ are given in Figure 16. Both axes are unnormalized. For $\alpha = 64$, the progress rate values decrease after $R^{(\infty)} \approx 1$, and one sees that $\varphi_{1,\lambda} < \varphi_{1+\lambda}$ for the same $R^{(\infty)}$ value (please note that same $R^{(\infty)}$ does not necessarily mean same σ). For $\alpha = 8$, $\hat{\varphi}_{1^+_,\lambda}$ is attained at a larger $R^{(\infty)}$ value, and the plus strategy shows smaller progress rates than the comma strategy for larger $R^{(\infty)}$ values. As a conclusion, one sees that the $\hat{\varphi}_{1^+_,\lambda}$ value depends on the $R^{(\infty)}$ value (to be more precisely, it depends on the σ value at which the critical $R^{(\infty)}$ magnitude is reached). The α value indicates how large the effect of $R^{(\infty)}$ is

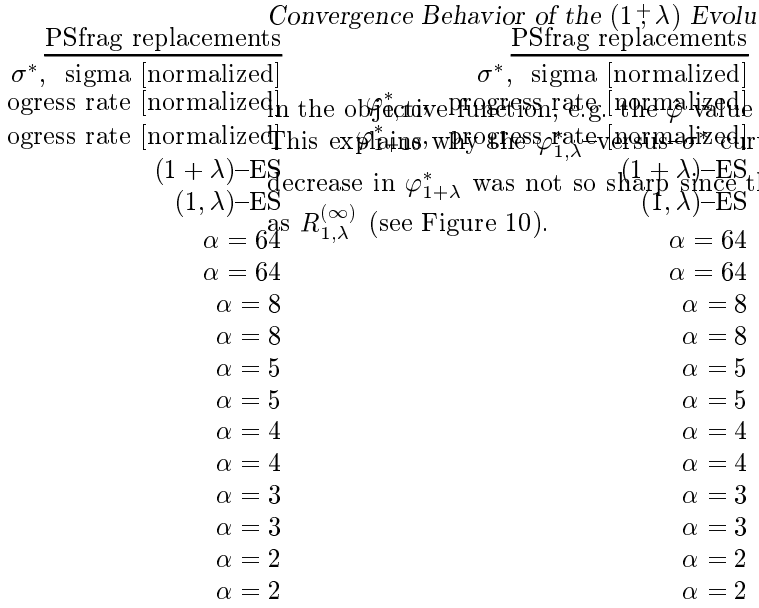


Figure 17: *Normalized progress rate versus normalized mutation strength, for different values of* α. *The* φ^*_{1+10} *values are shown on the figure left. The* $\varphi^*_{1,10}$ *values on the figure right decrease to zero suddenly after* $\sigma^* \gtrapprox 2c_{1,10} \approx 3$.

The normalized progress rates $\varphi^*_{1+\lambda}$ and $\varphi^*_{1,\lambda}$ are plotted next to each other in Figure 17 for different values of α. The case $\alpha = 2$ is discussed in depth in [OBS97, p.48f]. The results for comma strategy are given already in Section 5.2. In Figure 17, one can compare the shapes and the asymptotics of the progress rate figures for the plus and comma strategy. For $\alpha \geq 5$, the plus strategy gives better results (since $\alpha \in I\!R$, one has to note here that the exact value for the change where $\hat{\varphi}^*_{1+\lambda}$ becomes larger than $\hat{\varphi}^*_{1,\lambda}$ should be between four and five). The negative r^α term in the objective function makes the progress in the x_0 direction for the comma strategy very hard (see (22)), since $R^{(\infty)}_{1,\lambda}$ increases faster than $R^{(\infty)}_{1+\lambda}$. The plus strategy profits from the lower $R^{(\infty)}$ values, and $\varphi^*_{1+\lambda}$ decreases more slowly.

7 On the influence of the number of offspring

The theoretical formulae describing the convergence behavior of the $(1, \lambda)$-ES at the parabolic ridge test function will be verified here using larger values for the number of descendants ($\lambda \gg 10$). The formulae for $\varphi^*_{1,\lambda}$, $R^{(\infty)}_{1,\lambda}$, and $P_{s\,1,\lambda}$ are derived analytically in [OBS97]. The derivations do not emphasize a specific value for λ; however, since it is assumed that N (the number of dimensions) is sufficiently large, we doubt the formula to be valid for higher values of λ if N is kept constant.

In this section, simulation results obtained for three different λ values (10, 100, and 500) at the $(1, \lambda)$–ES will be compared with the theoretical formulae, and with the results obtained using $(1 + \lambda)$–ES. One keeps d and N constant ($d = 0.01$ and $N = 100$). The simulation length is $G = 100,000$ for the $(1,10)$–ES and

$G = 50,000$ for the others, reserving additional $2,000$ generations before collecting data. One does not have any theoretical results for the plus strategy, since it is not possible to evaluate the respective integrals analytically.

The limit value for the progress rate at the parabolic ridge test function is

$$\lim_{\sigma^* \to \infty} \varphi^*_{1,\lambda} = c^2_{1,\lambda}. \tag{33}$$

Therefore, in order to be able to compare the results for different λ, the normalized progress rates are divided by $c^2_{1,\lambda}$. Therefore, the horizontal asymptote becomes simply 1. This horizontal line is also indicated in the figure for the plus strategy. The σ and φ are normalized according to (18) and (25), respectively.

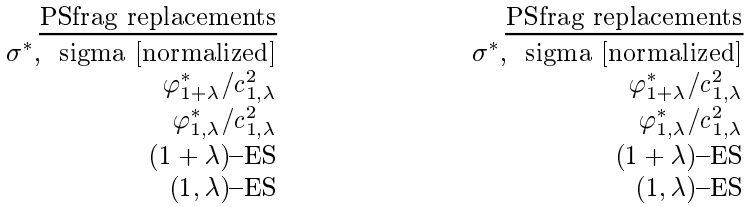


Figure 18: *The normalized progress rates divided by $c^2_{1,\lambda}$ are plotted versus the normalized mutation strength. The values used for λ are 10, 100, and 500 ($d = 0.01$, $N = 100$, $(1 \overset{+}{,} \lambda)$–ES). The simulation results for the comma strategy are shown in the figure left, and for the plus strategy on the figure right, respectively. The curves obtained using the theoretical $\varphi_{1,\lambda}$ formula (Eq. (28)) are also indicated for the $(1,\lambda)$–ES case, the simulation results deviate from these values for $\lambda = 100$ and $\lambda = 500$. The $\varphi^*_{1,\lambda}$ values exceed the $c^2_{1,\lambda}$ value for larger values of λ. The peak value of the progress rate for the plus strategy ($\hat{\varphi}^*_{1+\lambda}/c^2_{1,\lambda}$) increases as λ is increased.*

In Figure 18, the progress rate values for both selection methods are given next to each other. We have chosen $d = 0.01$ and $N = 100$, and for λ the values 10, 100, and 500. The simulation results for the comma strategy are in the figure left, for the plus strategy in the figure right. The φ^* values for $\lambda = 10$ are almost equal for both strategies for approximately $\sigma^* < 1$. If λ is increased further to 100 and 500, this range increases further to $\sigma^* < 3.5$ and $\sigma^* < 6$, respectively.

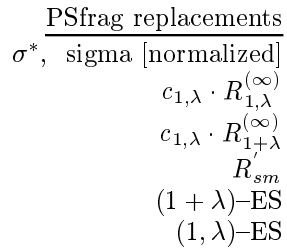
The $\varphi^*_{1,\lambda}/c^2_{1,\lambda}$ value obtained at large σ^* values increases as λ is increased, i.e. the horizontal asymptote is not valid for these large λ values ($N = 100$). The progress rate value goes to a greater limit. An explanation of this limit behavior for large λ will be investigated in the future. However, the progress rate curve obtained from the simulations with large λ does not differ much from the N-independent theoretical formula for larger values of N. For instance, in a simulation run with $N = 1,000$ and $\lambda = 500$, one observes that the horizontal asymptote $c^2_{1,\lambda}$ is again valid. This indicates that a more accurate $\varphi_{1,\lambda}$ formula must be developed which incorporates an N-dependence. Also the values for $\varphi^*_{1+\lambda}$ decreased for this case.

The simulations for the plus strategy (Figure 18 right) showed us that the peak value for the progress rate $\hat{\varphi}^*_{1+\lambda}$ increases as λ is increased. The mutation strength at which this optimum progress rate is attained ($\hat{\sigma}^*$) also increases. However, the $\hat{\varphi}^*_{1+\lambda}$ value does not seem to surpass the $c^2_{1,\lambda}$ value. Some values for comparing these two strategies are given in Table 1. This table summarizes Figure 18 and Figure 20.

λ	$c_{1,\lambda}$	$c^2_{1,\lambda}$	$\hat{\varphi}^*_{1,\lambda}$	$\hat{\sigma}^*_{1+\lambda}$	$sr(\hat{\sigma}^*_{1+\lambda},\lambda)$	$\hat{\varphi}^*_{1+\lambda}$	$\hat{\varphi}^*_{1+\lambda}/c^2_{1,\lambda}$	$\hat{\varphi}^*_{1+\lambda}/\hat{\varphi}^*_{1,\lambda}$
10	1.539	2.368	2.368	2.5	0.09	1.59	0.67	0.67
100	2.508	6.288	6.54	4.4	0.02	4.82	0.77	0.74
500	3.037	9.222	9.90	6.2	0.01	7.73	0.84	0.78

Table 1: *Comparing the* $(1\overset{+}{,}\lambda)$*–ES for large* λ *(*$\alpha = 2$*). The value of* $\hat{\sigma}^*_{1,\lambda}$ *is infinity. The table was generated using Figures 18 and 20. The values for* $\hat{\varphi}^*_{1,\lambda}$ *(*$\lambda = 100$ *and* $\lambda = 500$*) are obtained by averaging the simulation results for* $\sigma^* > 24$*. The table entries which have one or two digits after the decimal point are obtained empirically.*

Several valuable informations are repeated in Table 1. The entries with one or two digits after the decimal point are approximate, and are obtained from the simulations. In summary, one can see that $\hat{\sigma}^*_{1+\lambda}$ increases as λ is increased. The ratio $\hat{\varphi}^*_{1+\lambda}/\hat{\varphi}^*_{1,\lambda}$ increases at the same time. The success rate at $\hat{\sigma}^*_{1+\lambda}$ decreases as λ is increased.



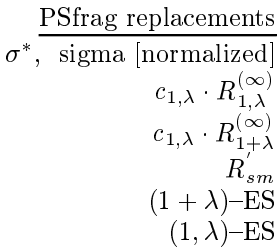


Figure 19: *The distances to the ridge axis multiplied by* $c^2_{1,\lambda}$ *are plotted versus the normalized mutation strength. The values used for* λ *are* 10, 100, *and* 500 *(*$d = 0.01$, $N = 100$, $(1\overset{+}{,}\lambda)$*–ES). The line labeled by "*R'_{sm}*" stands for* $c_{1,\lambda} D^{(\infty)} = \sigma^*(N-1)/2$*.* $D^{(\infty)}$ *is defined by Eq. (16). The simulation results for the comma strategy are shown in the figure left and for the plus strategy on the figure right, respectively.*
The $R^{(\infty)}_{1,100}$ *value does not differ much from the* $R^{(\infty)}_{1+100}$ *for* $\sigma^* \leq 6$ *(approximately), although the* φ^* *values for both selection methods differ remarkably even for* $\sigma^* = 5$ *(see Figure 18). This peculiarity becomes more evident for* $\lambda = 500$*:* $R^{(\infty)}_{1,500} \approx R^{(\infty)}_{1+500}$ *for* $\sigma^* \leq 15$*, and* $\varphi^*_{1,500} \gg \varphi^*_{1+500}$ *for* $\sigma^* \geq 10$*. The progress rate values do not seem to be related to* $R^{(\infty)}$ *values for higher values of* λ*.*

The distance to the progress axis ($R^{(\infty)}$) is another quantity which is used in describing the progress behavior on the ridge functions. In Figure 19, this value for the higher values of λ is investigated for both selection methods. The simulation results are multiplied by $c_{1,\lambda}$ in order to be able to compare the results for different λ in the same picture.

As the number of descendants increased, the plus strategy attained the same $R^{(\infty)}$ values up to larger values of σ^* as the comma strategy. To give an example, one has $R^{(\infty)}_{1,500} \approx R^{(\infty)}_{1+500}$ for $\sigma^* \leq 15$, whereas $R^{(\infty)}_{1,10} \approx R^{(\infty)}_{1+10}$ for $\sigma^* \leq 1$. The strange observation is $\varphi^*_{1,500} > \varphi^*_{1+500}$ for $\sigma^* \geq 6$, i.e. also when $R^{(\infty)}_{1,\lambda} \approx R^{(\infty)}_{1+\lambda}$.

As a result, one can say that the φ^* values are *not* necessarily related to the $R^{(\infty)}$ values. The success probability $P_{s\ 1,\lambda}$ gives us the necessary explanation for the difference between the progress rate values of these two selection methods.

PSfrag replacements
σ^*, sigma [normalized]
$sr(\lambda)$, success rate
$(1+\lambda)$–ES
$(1,\lambda)$–ES

PSfrag replacements
σ^*, sigma [normalized]
$sr(\lambda)$, success rate
$(1+\lambda)$–ES
$(1,\lambda)$–ES

Figure 20: *The success rate values are plotted versus the normalized mutation strength. The values used for λ are 10, 100, and 500 ($d = 0.01$, $N = 100$, $(1\overset{+}{,}\lambda)$–ES). The simulation results for the comma strategy are shown in the figure left, and for the plus strategy on the figure right, respectively.*
The theoretical $sr(\lambda)$ curve (Eq. (29)) and the theoretical asymptote (Eq. (30)) are displayed for the comma strategy. The success rate values for the plus strategy case continue to decrease as σ^* increases, and they are always smaller than the values obtained by the comma selection method.

Success rate value is a measure on the local curvature, i.e. it gives us the ratio of the success volume to the total local volume. The success volume consists of points having a better quality function value than the parent, and which are within the distance of the normalized mutation strength σ^* to the parent individual.

The success rate curves for high values of λ (10, 100, and 500) are shown in Figure 20. For the curves obtained using $(1,\lambda)$–ES, one sees that the asymptotic value is valid for $\lambda = 10$; however, for $\lambda = 100$ and $\lambda = 500$, it does not seem to be a sharp lower limit. The theoretical curve for $sr(500)$ is also not correct, either. However, it describes the tendency correctly.

For the plus strategy (figure right), the curves for these three λ values do not deviate much from each other. For $\sigma^* \leq 8$, one can say that the larger λ the smaller $sr(\lambda)$. This result is somewhat surprising, but one should remember that the success rate is also a function of λ, and the equilibrium condition changes if λ

is changed.

The $sr(\lambda)$ values are the same for both selection strategies at small σ^* values, to be more precise, for $\sigma^* \leq 1$, $\sigma^* \leq 4$, and $\sigma^* \leq 5$ for $\lambda = 10$, $\lambda = 100$, and $\lambda = 500$, respectively.

An interesting result obtained from Figure 18 and Figure 20 is the following: The progress rate of the plus strategy *cannot* be greater than the one of the comma strategy at the parabolic ridge test function. The success rate is always less than $\frac{1}{\lambda}$ for the $\hat{\sigma}^*$ value where $\hat{\varphi}^*_{1,\lambda}$ is reached (for any λ). The plus strategy has a success rate values which are less than or equal to the ones of the comma strategy. For σ^* values less than $\hat{\sigma}^*_{1+\lambda}$, the plus strategy cannot surpass the comma strategy. As the number of offspring goes to infinity, the plus strategy will attain progress rate values closer to $\hat{\varphi}^*_{1,\lambda}$; however, it cannot surpass.

PSfrag replacements

σ^*, sigma [normalized]
$_{1,\lambda}$, success probability
$_{+\lambda}$, success probability
$(1+\lambda)$–ES
$(1,\lambda)$–ES

PSfrag replacements

σ^*, sigma [normalized]
$P_{s\,1,\lambda}$, success probability
$P_{s\,1+\lambda}$, success probability
$(1+\lambda)$–ES
$(1,\lambda)$–ES

Figure 21: *The success probability values are plotted versus the normalized mutation strength. The values used for* λ *are* 10, 100, *and* 500 *(*$d = 0.01$, $N = 100$, $(1\dot{+}\lambda)$*–ES). The simulation results for the comma strategy are shown in the figure left, and for the plus strategy on the figure right, respectively. The figure on the left side has a linear scale on the vertical axis, whereas a logarithmic scale was necessary on the figure right.*
The theoretical $P_{s\,1,\lambda}$ *curve (Eq. (31)) and the theoretical asymptote (Eq. (32)) are plotted in the figure left. It differs from the simulation results for larger* λ *values. The* $P_{s\,1,\lambda}$ *values have the same asymptote for increasing* σ^**. The success probability values for the plus strategy case (*$P_{s\,1+\lambda}$*) continue to decrease as* σ^* *increases, and they are always smaller than the* $P_{s\,1,\lambda}$ *values.*

The success probability $P_{s\,1\dot{+}\lambda}$ is defined as the probability to have at least one descendant out of λ offspring which has a better fitness value than the parent individual. The success probability curves for different values of λ (10, 100, and 500) are shown in Figure 21. The theoretical curve obtained for $P_{s\,1,\lambda}$ deviates for $\lambda = 100$ and $\lambda = 500$ from the simulations. However, all curves seem to have almost equal asymptotic values, which is a remarkable difference to the plus strategy (figure right). Moreover, the $P_{s\,1,\lambda}$ values are much larger than $P_{s\,1+\lambda}$ for larger values of σ^*.

8 The parabolic ridge with (1+10)–ES for $r^{(0)} = 0$

In this section, the effect of the (special) initial position $r^{(0)} = 0$ in the search space to the progress behavior of the plus strategy will be investigated. The $(1+10)$–ES algorithm is used for this purpose. The parabolic ridge is used as the test function ($d = 0.01$, $N = 100$).

In [OBS97, pp.48], the two selection strategies ($(1 \dot{+} \lambda)$–ES) are compared at the parabolic ridge, and explained the progress deficiency of the plus strategy. It has been concluded that the low $R^{(\infty)}_{1+\lambda}$ and $P_{s\,1+\lambda}$ values caused the low $\varphi^*_{1+\lambda}$ rates (for $\sigma^* > 1$). In Section 7, one has seen that $P_{s\,1+\lambda}$ gives more information on the progress behavior. In the simulations done, $r^{(0)} = D^{(\infty)}$ is used, which is a good estimate for the comma strategy, and $R^{(\infty)}_{1+\lambda} < R^{(\infty)}_{1,\lambda}$ is observed at the end of the simulation. A period of $2,000$ generations is reserved before starting the statistics, and observed later that this was a sufficient length for reaching the steady state for $r^{(g)}$ for both selection strategies. The aim here was to start collecting data at the steady state for both algorithms. The question remaining unanswered was on the influence of a much smaller $r^{(0)}$ for the $(1 + \lambda)$-ES.

In order to have a fair comparison, the initial position in the search space should *not* favor either of the algorithms compared [Bey96, p.213]. It is the intention to show here how the initial conditions can affect the comparison.

The first test aims to verify the existence of $R^{(\infty)}_{1+\lambda}$ for different values of σ; i.e. different values for $r^{(0)}$ are chosen. For very large values (up to $r^{(0)} = 200D^{(\infty)}$), the same $R^{(\infty)}_{1+\lambda}$ value is obtained, however, it took just a longer time in generations to reach the steady state, e.g. for $\sigma = 20$ and $r^{(0)} = 10D^{(\infty)}$, $1,000$ generations were sufficient to reach $R^{(\infty)}_{1+\lambda}$.

However, for $r^{(0)} < R^{(\infty)}_{1+\lambda}$, the results are a bit different. The $r^{(g)}$ values fluctuate around the same $R^{(\infty)}_{1+\lambda}$ value; however, an $r^{(0)}$ value chosen smaller than the lower end of this fluctuation interval results in a very long adaptation time. Several generations pass by without a single movement in the search space if $r^{(0)}$ is chosen too small.

For $\sigma = 20$, some experiments are made for $r^{(0)} < R^{(\infty)}_{1+\lambda}$ ($d = 0.01$, $N = 100$, $G = 5 \cdot 10^6$), which has $R^{(\infty)}_{1+\lambda} \approx 250$. If $r^{(0)} \approx 181$ is chosen, the first move in the search space (g_{M1} in short) was the $2,342,082$th, and after 13 moves (at $g = 2,956,830$) the vicinity of the stationary state was reached ($r \approx 240$). In short, more than two million generations were necessary for the first move in the search space, and thereafter, about $6 \cdot 10^5$ generations to reach the vicinity of the steady state. If we started at $r^{(0)} \approx 198$, "only" $176,365$ generations were necessary to reach $r \approx 242$, i.e. at least an order of magnitude faster. On the other hand, for $r^{(0)} \approx 165$, more than four million generations were necessary for the first move.

Therefore, it is recommended to choose a large start value $r^{(0)}$ (such as $D^{(\infty)}$) as long as the $R^{(\infty)}_{1+\lambda}$ is unknown. Otherwise, the $(1 + 10)$–ES needs too many generations to reach the stationary state for r.

The worst value for $r^{(0)}$ is zero, since the $P_{s\,1+\lambda}$ has its smallest value in this

case. Some simulation runs are made for different values of σ between 1 and 8, and recorded g, $x_0^{(g)}$, $r^{(g)}$, and $F(\mathbf{x})^{(g)}$ values for the first 400 moves in the search space. The simulation length was further limited to $G = 5 \cdot 10^6$. The results are summarized in Table 2.

The $P_{s\,1\overset{+}{,}\lambda}$ value gives an estimate for the probability to get at least one descendant better than the parent individual in the next generation. The formula [OBS97, p.31f]

$$P_{s\,1,\lambda} \approx 1 - \left[\Phi\left(-\frac{M_Q}{S_Q}\right)\right]^\lambda = 1 - \left[\Phi\left(\frac{(N-1)d\sigma}{\sqrt{1+(2dr)^2+2\sigma^2d^2(N-1)}}\right)\right]^\lambda \tag{34}$$

is also applicable to the plus strategy, as long as the correct $R^{(\infty)}$ value is inserted. We insert $r^{(0)} = 0$, and get

$$P_{s\,1+\lambda} \approx 1 - \left[\Phi\left(\left(\sigma^{*-2} + 2(N-1)^{-1}\right)^{-\frac{1}{2}}\right)\right]^\lambda. \tag{35}$$

The decrease in the $P_{s\,1+\lambda}$ values explains the increase in g_{M1}, the number of generations necessary for the first move.

σ	g_{M1}	g_{M2}	$r^{(g_{M1})}$	$r^{(g_{M2})}$	$R^{(\infty)}_{1+\lambda}$	$R^{(\infty)}_{1,\lambda}$	$P_{s\,1+\lambda}$
1	2	3	8.801	13.259	44.62	46.412	0.827
2	3	6	19.293	27.910	67.29	75.547	0.241
3	22	23	28.780	41.067	86.14	104.655	0.028
4	23	56	35.143	49.423	100.02	134.318	$2.49 \cdot 10^{-3}$
5	24,198	24,472	44.308	57.191	114.47	164.446	$2.23 \cdot 10^{-4}$
6	222,510	225,254	51.225	66.783	130.94	194.907	$2.38 \cdot 10^{-5}$
7	629,625	644,050	56.117	70.056	139.16	225.599	$3.27 \cdot 10^{-6}$

Table 2: *The results of the simulation run with* $(1 + 10)$*–ES with* $r^{(0)} = 0$ *for* $1 \le \sigma \le 7$ $(N = 100,\ d = 0.01,$ *and* $G = 5 \cdot 10^6)$*. The simulation for* $\sigma = 8$ *could not generate any movement in the search space, i.e. all of the* 50 *million descendants generated had a smaller quality function value than the parent individual* $P^{(0)}$*. The symbols* g_{M1} *and* $r^{(g_{M1})}$ *stand for the generation count at which the first move in the search space is made and the distance to the ridge axis attained hereby, respectively. For larger values of* σ *one sees that* $g_{M2} - g_{M1} \ll g_{M1}$*. The movements thereafter occur even more frequently. The* $P_{s\,1+\lambda}$ *values are added to the table to show why the* g_{M1} *values increase so rapidly. One also sees that* $R^{(\infty)}_{1+\lambda} < R^{(\infty)}_{1,\lambda}$*.*

From Table 2 one sees that the $R^{(\infty)}_{1+\lambda}$ is smaller than $R^{(\infty)}_{1,\lambda}$; however, the plus strategy requires much more time than the comma strategy to reach that mean value when started at a low $r^{(0)}$ value.

Another observation obtained from Table 2 is that one did not need to wait for the second move in the search space so long as for the first one. The reason for it

is that the $r^{(g_{M1})}$ value is remarkably larger than $r^{(0)} = 0$, and therefore the $P_{s\ 1+\lambda}$ value is also much larger.

For $\sigma = 8$ (i.e. $\sigma^* = 7.92$), one could not register any moves at all in five million generations. This would mean a progress rate $\varphi^*_{1+\lambda} = 0$. The actual value at the appropriate $R^{(\infty)}_{1+\lambda}$ is a bit smaller than 0.5 [OBS97, p.48], obtained by averaging the results of only 100,000 generations. Therefore, the careful selection of initial conditions is necessary to obtain correct observations.

9 The sharp ridge

A special case of the ridge functions is the sharp ridge test function to be obtained by using the value $\alpha = 1$ in (4). After substituting r from (6) it reads

$$F(x_0, r) = x_0 - dr. \tag{36}$$

The sharp ridge has interesting progress properties. The progress rate φ_{sharp} is proportional to the mutation strength σ, and the proportionality constant depends only on d and λ. The $R^{(\infty)}$ value for it also increases linearly with σ, and it does *not* have a nonlinear region for small σ values (compare here the $R^{(\infty)}$ values for $\alpha \geq 2$ in Figure 4). Therefore, these early results can be summarized as

$$\varphi_{sharp} \propto \sigma \qquad\qquad R^{(\infty)} \propto \sigma. \tag{37}$$

PSfrag replacements

d
φ, progress rate
$c_{1,10}/\sqrt{1+d^2}$

PSfrag replacements

d
φ, progress rate
$c_{1,10}/\sqrt{1+d^2}$

Figure 22: *Progress rate versus d at the sharp ridge test function on the $(1, 10)$–ES with $\sigma = 1$, $N = 100$. The simulation results are compared with the analytical formula obtained from the local model, Equation (38). For $d \leq 0.1$ one has $\varphi \lessapprox \sigma c_{1,10} \approx \varphi_{hyperplane}$, and for the other extreme, $d \geq 100$, one gets $\varphi \gtrapprox 0$, since the influence of the $N - 1$ variables other than x_0 dominate the objective function. For the d values in between and also for these two extremes the formula $c_{1,10}/\sqrt{1+d^2}$ produces quite exact results.*

The proportionality constant for φ_{sharp} can be calculated using d and $c_{1,\lambda}$. The calculation of $R^{(\infty)}_{sharp}$ remains still to be done. Therefore, we will show here only

how the progress rate for the sharp ridge can be calculated simply from the local model mentioned in Section 5.2.

The normalization scheme for σ^* and φ^* suggested for the ridge functions in (18) and (25), respectively, *cannot* be used for the sharp ridge. Anyway, the linear response to the change in σ makes the φ–versus–σ and $R^{(\infty)}$–versus–σ plots unnecessary. Therefore, the mutation strength is kept constant ($\sigma = 1$) and used different values of d in our simulations ($10^{-5} \leq d \leq 1,000$).

The values $\lambda = 10$ and $N = 100$ are used for the $(1, \lambda)$–ES simulations. The simulation length was $G = 200,000$; additional $2,000$ generations were reserved at the beginning of each simulation before collecting the statistics. A different seed value was used for the pseudo-random number generator in each simulation run.

The φ–versus–d curve of the $(1, 10)$–ES at the sharp ridge is given in Figure 22. The analytical formula for the progress rate at the sharp ridge (also shown in this figure) is obtained from (23) for $\alpha = 1$

$$\varphi_{sharp} \approx \frac{\sigma c_{1,\lambda}}{\|\mathbf{a}\|} = \frac{\sigma c_{1,\lambda}}{\sqrt{1 + d^2}}. \tag{38}$$

This progress rate formula uses just the first order derivatives at the local model. It reflects the linear dependency of φ_{sharp} on the mutation strength and gives surprisingly good results.

PSfrag replacements

d

$R^{(\infty)}_{sharp}$

$D^{(\infty)}$

Figure 23: *The distance to the ridge axis versus d at the sharp ridge test function.* $(1, 10)$*–ES,* $\sigma = 1$*,* $N = 100$*. The simulation results are compared to the* $D^{(\infty)}$ *value (Eq. (16)). The relative difference between the simulation results and this value is less than one per cent for* $d \geq 6$*.*

In Figure 23, the simulation results for $R^{(\infty)}_{sharp}$ for different d values are plotted. The definition of $D^{(\infty)}$ is given by (16), which is obtained from the sphere model theory. The $R^{(\infty)}_{sharp}$ value is very near to this value for $d \geq 6$, in other words, for $\varphi < 0.26$, when the progress rate is very low.

The $R^{(\infty)}_{sharp}$ value goes to infinity as d goes to zero. This phenomenon was also observed for the parabolic ridge: A small d value reduces the effect of r in the

objective function, and the equilibrium for r is established at much larger values.

The performance of the $(1+\lambda)$–ES. In this part, the results obtained using the $(1+10)$–ES will be compared with the ones using the $(1,10)$–ES for the sharp ridge. The simulation uses the same parameters, i.e. $\lambda = 10$, $\sigma = 1$, $G = 200,000$, and $N = 100$. The $r^{(0)}$ value was again chosen large ($10D^{(\infty)}$ see Section 8 for the effect of the too small $r^{(0)}$ value using the `plus` strategy); and we started collecting the statistics after the first $2,000$ generations.

PSfrag replacements

d

$P_{s\ 1^{+}_{,}10}$, `success probability`

`comma`

`plus`

Figure 24: *The success probability versus d at the sharp ridge test function for the two selection strategies, with $\lambda = 10$, $\sigma = 1$, $N = 100$. For $d \leq 1$, both the $(1,10)$–ES and the $(1+10)$–ES have the same success probability. However, for $d > 1$ the $P_{s\ 1+10}$ decreases enormously, whereas the $P_{s\ 1,10}$ goes to the same limit which one observed at other ridge functions.*

The success probability values for both strategies are indicated in Figure 24. For $d \leq 1$, both strategies have the same success probability curve. For larger d values, the `plus` strategy has much smaller success probability values.

In Figure 25, the progress rate curves for the $(1+10)$–ES and the $(1,10)$–ES are displayed. For $d < 1$, both strategies yield almost equal performance. Actually, one almost has hyperplane conditions for such low d values, and for $\lambda = 10$, the progress rate of the `plus` strategy at the hyperplane is slightly larger than the one of the `comma` strategy. If one increases the d value, the `plus` strategy gives much worse results than the comma strategy: For $d = 4.5$ about three order of magnitudes. Such low progress rates for the `plus` strategy are explained by low success probabilities, which are caused by low $R^{(\infty)}_{1+10}$ values (see Figure 26).

The $R^{(\infty)}$ values of the $(1,10)$–ES and of the $(1+10)$–ES are compared in Figure 26. The `plus` strategy has smaller $R^{(\infty)}$ values for $d > 1$. For $d < 1$, both selection strategies attain nearly the same $R^{(\infty)}$ values.

To summarize these three figures, one can say that the behavior of the two selection strategies differ for $d > 1$. For these d values, the `plus` strategy attains smaller values for $R^{(\infty)}$ and the success probability, as a result, it has a lower

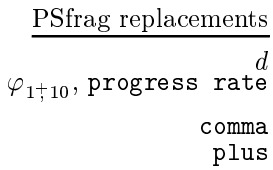


Figure 25: *Progress rate versus* d *at the sharp ridge test function of the two selection strategies, with* $\lambda = 10$, $\sigma = 1$, $N = 100$. *For* $d < 1$, *the simulation results are almost equal to each other* ($\varphi \lessapprox \sigma c_{1,10}$), *whereas the* $(1 + 10)$*–ES exhibits worse results for larger* d. *For* $d > 4.5$, *the simulation results for the* `plus` *strategy were not reliable since the* $P_{s\ 1+10}$ *values decreased enormously (see Figure 24).*

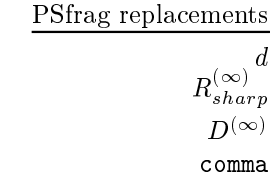


Figure 26: *The distance to the ridge axis versus* d *at the sharp ridge test function, for the two selection strategies, the* $(1, 10)$*–ES and* $(1 + 10)$*–ES,* $\sigma = 1$, $N = 100$. *The* $R^{(\infty)}_{1,10}$ *value goes approximately to the limit* $D^{(\infty)}$ *(see Figure 23 for the larger* d *values), whereas the* `plus` *strategy attains smaller* $R^{(\infty)}$ *values. For* $d < 1$, *the* $R^{(\infty)}$ *values are almost equal.*

progress rate performance.

10 The rectangular corridor model

The rectangular corridor model (Equation (7)) is analyzed in this section. We investigate whether it can be used as the limit case of the ridge functions or not. The measure φ^* is used in this comparison, it will also be mentioned the comparison results for $R^{(\infty)}$ and for success measures. The simulation length was $50,000$ generations, plus $2,000$ generations at the beginning before collecting statistics. The values $\lambda = 10$ and $N = 100$ are used in the simulations.

In order to compare the simulation results with the normalized values for the ridge functions, the normalization in [Sch95, p.138]

$$\sigma^* = \frac{N-1}{b}\sigma \ , \tag{39}$$

$$\varphi^* = \frac{N-1}{b}\varphi \tag{40}$$

has been used. Early simulations showed that the shape of the φ^*-versus-σ^* curve for the rectangular corridor model seems to be similar to the curve for the ridge functions with $\alpha > 2$. However, there remained some questions on the comparability of these two functions. The corridor models contain constraints in their definition. What should be done if all the λ descendants of the $(1, \lambda)$-ES are infeasible, i.e. outside the lethal restriction boundaries? One can use the idea given in [Sch95, p.137], and select the parent individual as the parent individual of the next generation. This requires a small but important modification in the algorithm: The new selection strategy is *not* a comma strategy anymore, since the parent individual can live more than one generation.

Another alternative is to continue generating descendants until one has λ feasible, offspring (i.e. generating an unknown number of offspring until this condition is satisfied), and applying the selection operator thereafter. However, this requires a more radical change in the algorithm, since the $(1 \overset{+}{,} \lambda)$–ES algorithms are compared based on the number of descendants actually generated.

As has been mentioned at the beginning of this section, we want to compare the behavior of the rectangular corridor model with ridge functions having larger values for α, in order to see whether it can be a limit case for the ridge functions or not. The parameters of the ridge functions chosen for the comparison are $N = 100$, $d = 0.01$ and $\alpha \in \{8, 64\}$. The normalization scheme given in (18) and (25), respectively is used. For the rectangular corridor, the parameters $c = 1$ and $b \in \{15, 150, 450\}$ are used, and the normalization given above in this section.

The normalized progress rates of the rectangular corridor model and of two members of the ridge family are compared in Figure 27. If one considers the change in the φ^* figure from $\alpha = 8$ to $\alpha = 64$, one can infer that the rectangular ridge *cannot* be the limit case for the ridge functions. For both of the normalization schemes used in this figure for the rectangular corridor model, the $\hat{\varphi}^*$ value and the shape of the progress rate differ for both function classes.

The change in the $R^{(\infty)}$ values with respect to σ^* shows the difference between these function families from another point of view. The $R^{(\infty)}$ value is the result of a different dynamics. For the rectangular corridor case, $R^{(\infty)}$ is dictated solely

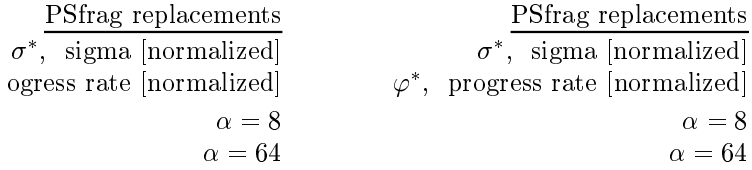


Figure 27: *Normalized progress rate versus normalized mutation strength. The simulation results for the rectangular corridor model are compared to the ridge function with* $\alpha = 8$ *and* $\alpha = 64$*. The figure right is obtained using* $(N-1)/2b$ *as the normalization constant for the rectangular corridor, the left one using the regular scheme given above in this section.*

by the restrictions. The observed value for this distance is not affected by the objective function at all. It increases much faster than in the ridge case, and it stays at a value governed by the restrictions for larger σ^* values.

The $sr(\lambda)$ and $P_{s\,1,\lambda}$ curves with respect to σ^* also differ for both strategies. There exists a lower limit for the ridge functions with $\alpha \geq 1$ (i.e. $P_{s\,1,\lambda} \geq 1 - [\Phi(c_{1,\lambda})]^{\lambda} > 0$); however, the success probability for the rectangular corridor goes to *zero* as σ^* is increased. The shape of the curves also differ for smaller σ^*.

The rectangular corridor model *cannot* be the limit case for the ridge functions: One observes characteristic differences in the behavior. In Section 11, we will examine whether the *cylindrical* corridor can be the limit case. The distance to the progress axis occurs explicitly in the definition of this function.

11 The cylindrical corridor model

In Section 10, the convergence behavior of the rectangular corridor model is compared with the ridge functions. The rectangular corridor has rectangular edges at the boundaries described by the $N-1$ restrictions. The cylindrical corridor, however, has a single restriction based on the distance to the corridor axis, i.e. its boundary has a cylindrical shape. In this section, it will be shown how well the cylindrical corridor may correspond to the limit case of the ridge functions with $(1, \lambda)$–ES. The values $\lambda = 10$, $b' = 450$, and $N = 100$ are used in this comparison. The simulation length is $G = 200,000$ generations, with additional $2,000$ generations before starting to collect statistics.

11.1 Ignoring infeasible descendants

The first possibility to handle constraints is to ignore infeasible descendants and to generate only λ descendants per generation. This case will be considered in this

subsection.

The cylindrical corridor is defined by (9), and the restriction applied to it by (10). The results are normalized similar to (39) and (40) for the rectangular corridor as follows:

$$\sigma^* = \frac{N-1}{b'}\sigma \tag{41}$$

$$\varphi^* = \frac{N-1}{b'}\varphi \tag{42}$$

Just as in the rectangular corridor model case, one selects the parent as the parent individual of the next generation if all λ offspring were beyond restriction boundaries. This produces a progress picture similar to the plus strategy.

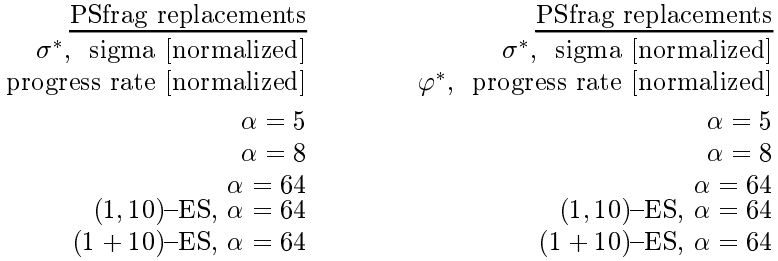


Figure 28: *Normalized progress rate versus normalized mutation strength. The simulation results for the cylindrical corridor model are compared to the ridge function's with* $\alpha \in \{5, 8, 64\}$. *The figure on the left shows the comparison of the cylindrical corridor with the members of the ridge family for the comma selection strategy* ($d = 0.01$, $N = 100$). *On the figure right, it is compared with the progress rate figures of* $(1\,\dagger\,10)$*–ES at the ridge function with* $\alpha = 64$.
The figure left shows that the cylindrical corridor does not correspond to the tendency of the φ^*_{ridge} *with large* α. *The figure right shows that it is similar to the curve of the plus strategy; however, the* $\hat{\varphi}^*$ *and* $\hat{\sigma}^*$ *values do not seem to reflect the values obtained for the ridge functions.*

We compared the normalized progress rate obtained at the cylindrical corridor with the ridge family, for large values of α and for both selection strategies. The results are displayed in Figure 28. On the figure left, one can see (after comparing the curves for different α) that the cylindrical corridor model does not reflect the tendency of the ridge functions. The magnitudes of $\hat{\varphi}^*$ are not comparable, and the shape of the curve for φ^*_{ridge} differs more from the cylindrical corridor model as α is increased.

On the figure right, one sees that the cylindrical corridor model has a progress behavior which is more similar to the (1+10)–ES than to the $(1, 10)$–ES, for $\alpha = 64$; however, the $\hat{\varphi}^*$ is smaller for the cylindrical corridor. Since one knows that $\hat{\varphi}^*$ increases for the ridge functions as α is increased (for $\alpha \geq 3$), the cylindrical corridor *cannot* be regarded as the limit case also for the plus selection strategy.

The $R^{(\infty)}$, $P_{s\,1\dagger\lambda}$, and $sr(\lambda)$ figures are also compared for both function families. The $R^{(\infty)}$ figures are much more different from each other, since the dynamics

governing the distance to the progress axis are different in both function families. The ridge family *has* the r value explicitly in the objective function, whereas the cylindrical corridor model has to obey the restriction $r \leq b'$. The success probability figure (and hereby the $sr(\lambda)$ figure) of cylindrical corridor is very similar to the one of the (1+10)–ES at the ridge function with $\alpha = 64$ although comma selection strategy is used at the cylindrical corridor. To summarize the results in this section and in Section 10, one can say that the corridor models are *not* the limit cases of the ridge functions.

11.2 Considering λ feasible offspring per generation

The algorithm investigated in Section 11.1 works usually with less than λ individuals per generation since λ feasible individuals were not guaranteed per generation. The algorithm considered now will be upgraded such that one continues generating individuals until really λ feasible descendants have been produced [Sch95, Appendix B3].

Therefore, the statistics collected for the progress rate should be interpreted as follows: Since the progress is measured in x_0 direction, one can denote the progress towards optimum in one generation as Δx_0. One has to multiply this number with λ/λ_{actual}, where λ_{actual} is the number of offspring generated to obtain λ feasible descendants. The *progress rate* $\varphi_{cylinder}$ is obtained as usual by averaging these measurements over G generations. If one omits this adjustment, one obtains a progress rate curve which is similar to the one at the hyperplane, since the effect of the cylinder-shaped restriction is canceled out by the excess number of descendants generated.

Once one has λ feasible offspring, one can apply the $(1, 10)$–ES and compare the observed progress behavior with those of the ridge functions and those obtained in the previous section. The other parameters of the simulation were selected as follows: $G = 50,000$ (with maximal 10^7 fitness evaluations), $b' = 450$, $N = 100$. One started at $r^{(0)} = 0$, and started collecting data from the first generation on. The results did not change significantly when collecting data started after the first $2,000$ generations in another simulation set. Since λ_{actual} is much larger for larger σ^*, the simulation length may be shorter than G. The number of offspring generated was limited in a simulation run by 10^7 in order to limit the simulation length in time. In other words, if more than 20 offspring should be generated on average for a nonlethal descendant, the simulation length G is decreased accordingly. The $R^{(\infty)}$, $P_{s\,1,10}$, and φ values are measured for this algorithm.

The progress rate curve obtained is compared in Figure 29 with the one of the previous algorithm in the Section 11 above; and to the one of the ridge function with $\alpha = 64$. Obviously, also the algorithm described in this subsection has a different progress behavior than the one of the ridge functions.

We also compared the $R^{(\infty)}$–versus–σ^* curves of both function families. The members of the ridge function family (with $\alpha \geq 1$) have $R^{(\infty)}_{1,\lambda}$ values which are proportional to the σ^* used, and the proportionality constant $(N-1)/2c_{1,\lambda}$ can be observed in all the simulations (see e.g. Figure 10 for $\sigma^* > 2c_{1,10}$). The $R^{(\infty)}_{1+\lambda}$ values

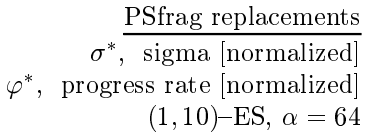


Figure 29: *Normalized progress rate versus normalized mutation strength. The simulation results for the cylindrical corridor model are compared to the ridge function's ($\alpha = 64$). The $\varphi_{hyperplane}$ line is also plotted. The simulation for the cylindrical corridor with effective $\lambda = 10$ is labeled as cylinder new, the other one as cylinder. Both algorithms for the cylindrical corridor differ remarkably from the ridge functions with high α value. The $\hat{\varphi}^*$ values obtained are also smaller than the one obtained for the ridge function.*

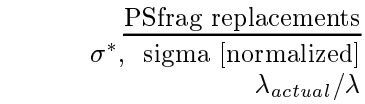


Figure 30: *The simulation length G and the ratio of the λ_{actual}/λ versus σ^*, for $\lambda = 10$. The $(1,10)$–ES is applied to the cylindrical corridor model ($N = 100$, $b' = 450$). For $\sigma^* \geq 4$, the simulation length is less than $50,000$ generations since the total number of function evaluations was restricted by $20 \cdot 10 \cdot 50,000 = 10^7$ generations. The λ_{actual}/λ rate increases exponentially in σ^*: For $\sigma^* = 2$ it is around 3.2, and e.g. for $\sigma^* = 4$ around 22.*

also increase proportionally in σ^*. Moreover, the $R^{(\infty)}$ values are not restricted for the ridge functions. The cylindrical corridor shows another $R^{(\infty)}$ profile for increasing σ^*. It attains values very near to b' already for very small σ^* values ($\sigma^* \approx 0.5$), and changes very little for larger σ^*.

The actual simulation length G and the ratio λ_{actual}/λ are shown in Figure 30. The number of offspring which should be generated to obtain λ nonlethal descendants increases exponentially. This increase causes a reduction in the actual simulation length for $\sigma^* > 4$. The results for $\sigma^* > 6$ are therefore not accurate.

The success probability figures are similar to the $P_{s\,1+10}$ figures of the ridge functions, although the $(1,10)$–ES was used here for the cylindrical corridor. For the ridge functions, the $P_{s\,1,10}$ curve is very different from the $P_{s\,1+10}$ curve (Figure 13). Therefore, one can say that the $P_{s\,1,10}$ curve for the cylindrical corridor differs very much from the $P_{s\,1,10}$ curve for the ridge functions (for any α).

After considering the $P_{s\,1,10}$, $R^{(\infty)}$, and φ^* figures, one can say that the behavior of the cylindrical corridor differs much from that of the ridge functions. Therefore, the cylindrical corridor model does *not* seem to be the limit case of the ridge functions for $\alpha \to \infty$.

11.3 Comparing both corridor models without normalization

In this subsection, the progress rate figures of cylindrical corridor will be compared with the rectangular corridor. These test functions (Equations 7 and 9, respectively) do only differ from each other in the applied constraints. In the comparison, one will use the same size for the cross section in $N-1$ dimensions other than x_0. The size of the cross section of the cylinder is the volume of the $N-1$–dimensional hypersphere. This volume is given in [Smi77, p.326-327].

The cross section volume for the rectangular ridge is given as

$$v_{rectangular} = (2b)^{N-1}; \tag{43}$$

and for the cylindrical corridor as

$$v_{cylinder} = c_{N-1} b^{'\,N-1}, \tag{44}$$

where the c_{N-1} value is

$$c_{N-1} = \frac{\pi^{\frac{N}{2}}}{\Gamma\left(\frac{N}{2}+1\right)}, \qquad c_{99} \approx 9.4714 \cdot 10^{-40}. \tag{45}$$

From the equality $v_{cylinder} \overset{!}{=} v_{rectangular}$, one gets for $b^{'} = 450$ the corresponding b value as approximately 90.783. The simulations for the rectangular ridge are made using this value, and using the double and half of it to show the effect. We also generated $\lambda_{actual} \gg \lambda$ offspring to have $\lambda = 10$ nonlethal individuals for the rectangular ridge as well. The results are shown in Figure 31.

All the curves depicted in the figure are unnormalized. One can conclude that the $(1,\lambda)$–ES has higher progress rates at the cylindrical corridor than at the rectangular ridge for the same cross section size and σ. Moreover, $\varphi_{cylinder}$ is still greater if one doubles the width of the rectangular corridor.

The geometry with sharp angled constraint edges makes the progress very hard. This can be seen easier on the success rate curve (Figure 32). The success rate

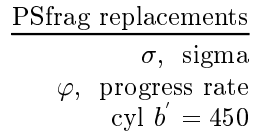


Figure 31: *Progress rate versus mutation strength. Both axes are unnormalized. The simulation results obtained using* $(1, 10)$*-ES by rectangular and cylindrical corridor models are compared. The simulation results at the cylindrical corridor are obtained for* $b^{'} = 450$*, displayed as "cyl* $b^{'} = 450$*". For the rectangular corridor, the results of three different* b *values are displayed for* 181.566*,* 90.783*, and* 45.39*. One obtains smaller progress rates for the rectangular corridor if one uses the same cross section of the cylindrical corridor.*

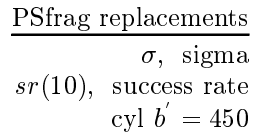


Figure 32: *Success rate versus mutation strength. The* $(1, 10)$*-ES attains at the cylindrical ridge test function larger success rate values than at the rectangular ridge with the same cross section (*$b = 90.783$*) for the same mutation strength values. The rectangular corridor model has lower sr(10) values even when one doubles its corridor width.*

values are obtained by dividing the number of successful mutations (i.e. the descendants having a quality function value larger than their parent) by the number of offspring actually generated (λ_{actual}).

12 Summary and Outlook

Several experiments were performed in this work in order to understand the progress behavior of the $(1 \dagger \lambda)$–ES at the ridge functions and corridor models. At the ridge functions, the plus selection strategy does not give better progress rate results for *smaller* α values ($\alpha \lessapprox 4$). The $P_{s\ 1\dagger\lambda}$ and $R^{(\infty)}$ values are also investigated for both strategies, which are smaller for the $(1+\lambda)$–ES. The $\varphi^*_{1,\lambda}$–versus–σ^* curves for the comma strategy are explained analytically by a simple model which uses just first order partial derivatives.

The *parabolic ridge* function was revisited for some tests with larger λ values. The tests yield similar shapes for the progress rate of the $(1 \dagger 10)$–ES, where one observed that the success probability values are more reliable than the $R^{(\infty)}$ values in explaining the progress behavior. The $\hat{\varphi}^*_{1+\lambda}/\hat{\varphi}^*_{1,\lambda}$ ratio and also $\hat{\varphi}^*_{1,\lambda}/c^2_{1,\lambda}$ increases for increasing λ. The lower limit for the $P_{s\ 1,\lambda}$ is almost the *same* for any λ.

The performance of $(1+\lambda)$–ES at the parabolic ridge was tested with small $r^{(0)}$ values. These simulations showed us how bad the progress performance of the $(1+\lambda)$–ES can be if the initial condition is not selected appropriately.

The results on the *sharp ridge* were also summarized in this paper. The progress rate φ for this function was derived using the same local model approximation derived for the general case of ridge functions. The resulting formula is astonishingly simple, and accords to the simulation results. The $P_{s\ 1,10}$–versus–σ^* curve was explained almost exactly using an analytically derived formula. The $(1+10)$–ES was tested on the sharp ridge, and it has been shown that it has a much worse progress rate (and success probability) than the $(1,10)$–ES for larger values of d. For the parabolic ridge, this was observed for any d value. Therefore, the plus strategy can perform worse than the comma strategy on some members of the ridge function family.

The limit behavior of the ridge functions at $(1 \dagger \lambda)$–ES was analyzed for large α, and it was compared to the corridor models'. The rectangular corridor model and the cylindrical corridor model do *not* seem to be the limit case of the ridge functions. Furthermore, the progress rate curves for these two corridor models are compared for the same cross section size. One observes that the $(1,\lambda)$–ES attains significantly larger progress rates at the cylindrical corridor model because of its spherical constraint structure.

Population models (i.e. (μ,λ)–ES) and the effect of the recombination (i.e. $(\mu/\rho,\lambda)$–ES) will be analyzed in a future work. These simulations will show the effect of having a *parental* population and of the recombination on the behavior, especially on the progress rate.

13 Acknowledgment

The first author was partially supported by a PhD scholarship of DAAD (German Academic Exchange Service) and by the Deutsche Forschungsgemeinschaft (DFG), grant Be1578/6-1. The second author acknowledges support as Heisenberg Fellow

of the DFG under grant Be1578/4-1.

References

[Bey93] H.-G. Beyer. Toward a Theory of Evolution Strategies: Some Asymptotical Results from the $(1 \dagger \lambda)$–Theory. *Evolutionary Computation*, 1(2):165–188, 1993.

[Bey94] H.-G. Beyer. Towards a Theory of 'Evolution Strategies'. Progress Rates and Quality Gain for $(1 \dagger \lambda)$–Strategies on (Nearly) Arbitrary Fitness Functions. In Y. Davidor, H.-P. Schwefel, and R. Männer, editors, *Parallel Problem Solving from Nature — PPSN III International Conference on Evolutionary Computation*, volume 866 of *Lecture Notes in Computer Science*, pages 58–67. Springer, 1994.

[Bey96] H.-G. Beyer. *Zur Analyse der Evolutionsstrategien.* Habilitationsschrift. University of Dortmund, Department of Computer Science, 1996. In German.

[BS93] T. Bäck and H.-P. Schwefel. An Overview of Evolutionary Algorithms for Parameter Optimization. *Evolutionary Computation*, 1(1):1–23, 1993.

[Fog95] D. B. Fogel. *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence.* IEEE Press, Piscataway, NJ, 1995.

[FOW66] L. J. Fogel, A. J. Owens, and M. J. Walsh. *Artificial Intelligence through Simulated Evolution.* Wiley, New York, 1966.

[Gol89] D. E. Goldberg. *Genetic Algorithms in Search, Optimization & Machine Learning.* Addison-Wesley, Reading, 1989.

[Hol75] J. H. Holland. *Adaptation in Natural and Artificial Systems.* Univ. of Michigan, NN, 1975.

[OBS97] A. I. Oyman, H.-G. Beyer, and H.-P. Schwefel. Analysis of a Simple ES on the "Parabolic Ridge". Technical Report SyS–2/97, University of Dortmund, Department of Computer Science, Systems Analysis Research Group, August 1997. `http://ls11-www.cs.uni-dortmund.de/~oyman/TR.ps.gz`.

[OBS98] A. I. Oyman, H.-G. Beyer, and H.-P. Schwefel. Where Elitists Start Limping: Evolution Strategies at Ridge Functions. In A. E. Eiben, Th. Bäck, M. Schoenauer, and H.-P. Schwefel, editors, *Parallel Problem Solving from Nature – PPSN V, International Conference on Evolutionary Computation*, volume 1498 of *Lecture Notes in Computer Science*, pages 34–43. Springer, Berlin, September 27-30 1998.
`http://ls11-www.cs.uni-dortmund.de/ ~oyman/TR/ppsn5.ps.gz`.

[Rec73] I. Rechenberg. *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution.* Verlag Frommann-Holzboog, Stuttgart-Bad Cannstatt, 1973. ISBN: 3-7728-0373-3, in German.

[Rec94] I. Rechenberg. *Evolutionsstrategie'94.* Band 1, Werkstatt Bionik und Evolutionstechnik. Frommann–Holzboog, Stuttgart, 1994. ISBN: 3-7728-1642-8, in German.

[Rud96] G. Rudolph. *Convergence Properties of Evolutionary Algorithms.* Dissertation, University of Dortmund, Department of Computer Science, November 1996.

[Sal96] R. Salomon. Re-evaluating genetic algorithm performance under coordinate rotation of benchmark functions. A survey of some theoretical and practical aspects of genetic algorithms. *BioSystems*, 39(3):263–278, 1996.

[Sch95] H.-P. Schwefel. *Evolution and Optimum Seeking.* Sixth–Generation Computer Technology Series. Wiley, New York, 1995. ISBN: 0-471-57148-2.

[Smi77] W. I. Smirnow. *Lehrgang der höheren Mathematik, Teil 2.* VEB Deutscher Verlag der Wissenschaften, Berlin, 1977.