Mutation Strength Control by Meta-ES on the Sharp Ridge

Hans-Georg Beyer
Research Center PPE
Vorarlberg University of Applied Sciences
Dornbirn, Austria
Hans-Georg.Beyer@fhv.at

Michael Hellwig
Research Center PPE
Vorarlberg University of Applied Sciences
Dornbirn, Austria
Michael.Hellwig@fhv.at

ABSTRACT

This paper investigates mutation strength control using Meta-ES on the sharp ridge. The asymptotical analysis presented allows for the prediction of the dynamics in ridge as well as in radial direction. Being based on this analysis the problem of the choice of population size λ and isolation parameter γ will be tackled. Remarkably, the qualitative convergence behavior is not determined by γ alone, but rather by the number of function evaluations $\lambda\gamma$ devoted to the inner ES.

Categories and Subject Descriptors

I.2.8 [**Problem Solving, Control Methods, and Search**]: Control theory

General Terms

Algorithms

Keywords

Adaption, Evolution Strategies, Meta-ES, Mutation Strength, Sharp Ridge Function

1. INTRODUCTION

Controlling the strategy parameters of an Evolutionary Strategy (ES) by hierarchically organized ESs, also referred to as Meta-ESs, is a common practice in the field of ESs. Formally Meta-ESs are described by the generalized ES bracket notation according to Rechenberg [8]

$$[\mu'/\rho', \lambda'(\mu/\rho, \lambda)^{\gamma}]\,. \tag{1}$$

The Meta-ES (1) runs λ' parallel inner $(\mu/\rho, \lambda)$-ESs over γ generations. Each of these ESs are equipped with different initial strategy parameters such as different initial mutation strengths or population sizes. Selection on the higher level then chooses the μ' best for recombination with respect to a previously defined fitness criterion. While there is experimented evidence that Meta-ES can tune the inner ES to optimal performance [5], the control investigations are still scarce. There are basically two papers on that topic: In [4] the performance of the $[1, 2(\mu/\mu_I, \lambda)^{\gamma}]$-Meta-ES on the sphere model has been analyzed. In [1] Arnold presented an analysis of the mutation strength adaption by $[1, 2(\mu, \mu_I, \lambda)^{\gamma}]$-Meta-ES on the family of ridge functions

$$F(\mathbf{x}) = -x_1 + d\left(\sum_{i=2}^{N} x_i^2\right)^{\frac{1}{2}\beta}\,, \quad \beta \geq 2. \tag{2}$$

His analysis did not include the sharp ridge. However, the sharp ridge is especially interesting since this fitness landscape poses a challenge on any mutation adaption mechanism including cumulative step-size adaption and self-adaption, see also [6]. Furthermore the sharp ridge can be regarded as a model for linear constraint optimization where standard methods are known to show a premature convergence behavior. Therefore, this paper will particularly investigate Meta-ES in the fitness environment defined by the sharp ridge function which is defined as

$$F(\mathbf{x}) = -x_1 + d\left(\sum_{i=2}^{N} x_i^2\right)^{\frac{1}{2}} = -x_1 + dR \tag{3}$$

for $\mathbf{x} = (x_1, \ldots, x_N) \in \mathbb{R}^N$. The parameter d is called the nonlinearity strength parameter of the ridge function.

The x_1-axis also referred to as the ridge axis is in this definition identified with an axis of the coordinate system. While this is a special choice, an arbitrary coordinate transformation would not have any effect to the strategy's performance. This is so because the inner ES uses isotropically distributed mutations.

The term $R := \sqrt{\sum_{i=2}^{N} x_i^2}$ describes the distance of the argument $\mathbf{x} \in \mathbb{R}^N$ to the ridge axis.

Without loss of generality we assume a minimization problem. The sharp ridge function has no finite optimum, thus minimization steadily decreases the F-value. Fitness improvement can be achieved in two

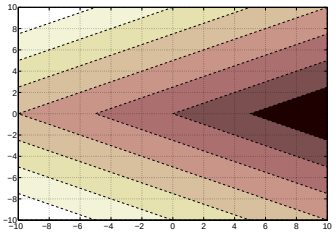

Figure 1: The contour plot of the 2-dimensional Sharp Ridge function with $d = 2$. Notice that darker regions correspond to lower function values.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
GECCO'12, July 7-11, 2012, Philadelphia, Pennsylvania, USA.
Copyright 2012 ACM 978-1-4503-1177-9/12/07 ...$10.00.

ways: either by reducing the distance R to the ridge axis or by increasing the x_1-value, i.e. making progress in direction of the ridge axis. In the short run this is realized fastest by reducing the distance to the ridge axis. However, the convergence to the ridge axis also results in a stagnation of the progress in direction of the ridge axis. Thus a long term improvement is achieved better by enlarging the x_1 value and accepting an increasing radius R, see [1].

This paper is organized as follows: Section 2 contains a description of the investigated $[1, 2(\mu/\mu_I, \lambda)^\gamma]$-Meta-ES algorithm. In Section 3 we work on the theoretical analysis of the inner ES for three cases depending on the initial distance to the ridge axis R. Resulting in a rule for the choice of the isolation length parameter γ which allows for the control of the mutation strength σ. The analysis of the Meta-ES behavior over multiple isolation periods is presented in Section 4. For sufficiently long isolation length γ we provide an asymptotic solution for the expected results. Section 5 investigates the impact of small dimensionalities N on the Meta-ES dynamics. Finally, Section 6 sums up our results and offers an outlook on future research.

2. CONTROLLING THE MUTATION STRENGTH BY [1, 2]-META-ES

This section recaps the simple $[1, 2(\mu/\mu_I, \lambda)^\gamma]$-Meta-ES which is used to control the mutation strength σ. The outer ES, see Fig. 2, generates two $(\mu/\mu_I, \lambda)$ populations. The inner ESs start from the same initial $\mathbf{y_p}$ but with different mutation strengths $\tilde{\sigma}$ which are kept constant during the isolation period of γ generations. As one

$[1, 2(\mu/\mu_I, \lambda)^\gamma]$-ES	line
Initialize($\mathbf{y_p}, \sigma_p, \alpha, \mu, \lambda, \gamma, N$);	1
$t = 0$;	2
Repeat	3
$\tilde{\sigma}_1 := \sigma_p \alpha$;	4
$\tilde{\sigma}_2 := \sigma_p \alpha^{-1}$;	5
$[\tilde{\mathbf{y}}_\mathbf{1}, \tilde{f}_1] := \mathrm{ES}(\mu, \lambda, \gamma, \tilde{\sigma}_1, \mathbf{y_p})$;	6
$[\tilde{\mathbf{y}}_\mathbf{2}, \tilde{f}_2] := \mathrm{ES}(\mu, \lambda, \gamma, \tilde{\sigma}_2, \mathbf{y_p})$;	7
$\sigma_p := \tilde{\sigma}_{1;2}$;	8
$\mathbf{y_p} := \tilde{\mathbf{y}}_{\mathbf{1};2}$;	9
$t := t + 1$;	10
Until(termination condition)	11

Figure 2: Pseudo code of the [1, 2]-Meta-ES. The Code of the inner ES is displayed in Fig. 3.

can see in Line 4 and 5 the two different $\tilde{\sigma}$ values are computed by increasing respectively decreasing the parental mutation strength by the factor $\alpha > 1$. As a result there is always one $(\mu/\mu_I, \lambda)^\gamma$-ES running with mutation strength $\tilde{\sigma}_1 = \alpha\sigma_p$ and one with $\tilde{\sigma}_2 = \sigma_p/\alpha$. Selection is explained in lines 8 and 9 using the standard notation "$m; \lambda'$" indicating the m-th best population out of all λ' populations. The populations are ordered by the function value returned by the respective inner standard ES, displayed in Fig. 3, after having evolved independently over γ generations. The inner ES generates a population of λ offspring by adding the product of the mutation strength σ and a vector of independent, standard normally distributed components to the centroid $\mathbf{y}$ of the parental generation. The μ best candidates (in terms of their function values $\tilde{F}_l$) out of these λ offsprings are used to build the new parental centroid $\mathbf{y}$. Proceeding this way over γ generations the inner ES returns the tuple $[\mathbf{y}, F(\mathbf{y})]$.

As a termination criterion for the outer ES one can for example choose an upper bound for the number of function evaluations or specify a fixed number of isolation periods.

Function: ES($\mu, \lambda, \gamma, \sigma, \mathbf{y}$)	line
g=1;	1
While $g \leq \gamma$	2
For $l = 1$ To λ	3
$\tilde{\mathbf{y}}_l := \mathbf{y} + \sigma\mathcal{N}_l(0, \mathbf{I})$;	4
$\tilde{F}_l := F(\tilde{\mathbf{y}}_l)$;	5
End For	6
$\mathbf{y} := \frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{\mathbf{y}}_{m;\lambda}$;	7
$g := g + 1$;	8
End While	9
Return $[\mathbf{y}, F(\mathbf{y})]$;	10

Figure 3: The inner $(\mu/\mu_I, \lambda)^\gamma$-ES

3. THEORETICAL ANALYSIS

In what follows, we examine the dynamics of the inner $(\mu/\mu_I, \lambda)$-ES over a period of γ generations. The theoretical analysis is based on mean value dynamics which is a common practice in the field of ES. Taking into account the selection mechanism of the Meta-ES we can derive directives for the mutation strength control depending on the population size and the fitness environment in terms of the dimensionality N and the nonlinearity strength d. The theoretical results are finally compared with the empirical analysis of the algorithm.

The analysis is based on the deterministic mutation rule that produces exactly two inner $(\mu/\mu_I, \lambda)$-ES, i.e. one with $\sigma_+ = \alpha\sigma^{(g)}$ and one with $\sigma_- = \sigma^{(g)}/\alpha$. Note that $\sigma_\pm$ corresponds to $\tilde{\sigma}_1$ and $\tilde{\sigma}_2$ mentioned in Section 2. We take a look at the expected value $F^{(g)}$ of the sharp ridge function at generation $g+1$ which is already constituted in [1] by the recurrence equation

$$F^{(g+1)} = -x^{(g+1)} + dR^{(g+1)}, \tag{4}$$

where the distance in direction of the ridge axis is given by

$$x^{(g+1)} = x^{(g)} + \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1 + d^2}} \tag{5}$$

and the distance to the ridge axis is constituted by

$$R^{(g+1)} = R^{(g)} + \frac{\sigma^2 N}{2\mu R^{(g)}} - \frac{d\sigma c_{\mu/\mu,\lambda}}{\sqrt{1 + d^2}}. \tag{6}$$

For the definition of the so called progress coefficients $c_{\mu/\mu,\lambda}$ we refer the reader to [3]. We can see that $x^{(g+1)}$ scales up linearly with the number of generations, thus we can transform it to

$$x^{(g+1)} = x^{(0)} + \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1 + d^2}}(g + 1). \tag{7}$$

In order to investigate the behavior of the $[1, 2(\mu/\mu, \lambda)^\gamma]$-Meta-ES we need a similar expression for $R^{(g+1)}$. Because of the nonlinearity of the recurrence equation (6) we are not able to find a closed analytical solution for $R^{(g+1)}$. Hence we switch to the continuous time limit to approximate the difference equation (6) by a differential equation which is formally obtained by expanding $R(g+1)$ into a Taylor series at g and breaking off after the linear part

$$R(g + 1) = R(g) + \frac{\mathrm{d}R}{\mathrm{d}g}1 + \dots . \tag{8}$$

Identifying $R(g + 1)$ with $R^{(g+1)}$ and $R(g)$ with $R^{(g)}$ in (6) we get the nonlinear differential equation

$$\frac{\mathrm{d}R}{\mathrm{d}g} = \frac{\sigma^2 N}{2\mu R} - \frac{d\sigma c_{\mu/\mu,\lambda}}{\sqrt{1 + d^2}}. \tag{9}$$

Equation (9) describes the basis of the further theoretical analysis. If the isolation period γ is sufficiently large, $\gamma \to \infty$, the expected distance reaches a steady state R denoted by R_∞. This is equivalent to $\frac{dR}{dg} \to 0$, i.e. $R^{(g)} = R^{(g-1)} = R_\infty$ for $g \to \infty$. That is, the inner ES with constant mutation strength σ approaches the expected residual steady state distance R_∞ to the sharp ridge axis which is obtained by resolving (9) for $\frac{dR}{dg} = 0$

$$R_\infty(\sigma) := \frac{N\sigma}{2\mu c_{\mu/\mu,\lambda}} \frac{\sqrt{1+d^2}}{d}. \tag{10}$$

Due to mathematical difficulties solving (9) for R we have to consider three special cases depending on the choice of the initial distance R of the starting point from the ridge axis:

$$\text{a) } R \approx R_\infty, \quad \text{b) } R \ll R_\infty, \quad \text{and c) } R \gg R_\infty.$$

3.1 Analysis of the $R \approx R_\infty$ Case

First we explore the situation in which the initial distance to the ridge axis lies in the vicinity of the residual distance R_∞. In this case we are able to replace the rhs of the differential equation (9) with its Taylor series around R_∞. With the derivation of (9) with respect to R

$$\frac{\mathrm{d}}{\mathrm{d}R}\Big(\frac{\mathrm{d}R}{\mathrm{d}g}\Big) = \frac{-\sigma^2 N}{2\mu R^2} \tag{11}$$

and by considering again only the linear terms we get

$$\begin{aligned} \frac{\mathrm{d}R}{\mathrm{d}g} &\approx \frac{\mathrm{d}R}{\mathrm{d}g}\Big|_{R=R_\infty} + \frac{\mathrm{d}}{\mathrm{d}R}\Big(\frac{\mathrm{d}R}{\mathrm{d}g}\Big)\Big|_{R=R_\infty}(R-R_\infty) && (12)\\ &= \frac{-2c^2_{\mu/\mu,\lambda}d^2\mu}{(1+d^2)N}(R-R_\infty) && (13)\end{aligned}$$

This differential equation can be solved for g by separating the variables, in the first step yielding

$$\int_{R^{(0)}}^{R^{(g)}} \frac{1}{R'-R_\infty}\mathrm{d}R' = \int_0^g \frac{-2c^2_{\mu/\mu,\lambda}d^2\mu}{(1+d^2)N}\mathrm{d}g'. \tag{14}$$

The integration leads to the equation

$$\ln\Big(\frac{R^{(g)}-R_\infty}{R^{(0)}-R_\infty}\Big) = \frac{-2c^2_{\mu/\mu,\lambda}d^2\mu}{(1+d^2)N}g. \tag{15}$$

This can be solved for $R^{(g)}$

$$\begin{aligned} R^{(g)} &= R_\infty + (R^{(0)}-R_\infty)\mathrm{e}^{\frac{-2c^2_{\mu/\mu,\lambda}d^2\mu}{(1+d^2)N}g} && (16)\\ &= R^{(0)}\mathrm{e}^{-g/\tau} + R_\infty\left(1-\mathrm{e}^{-g/\tau}\right), && (17)\end{aligned}$$

where τ is defined as

$$\tau := \frac{(1+d^2)N}{2c^2_{\mu/\mu,\lambda}d^2\mu}. \tag{18}$$

Taking into account Eq. (4), (7), and (17) we obtain an expression for the expected value of the sharp ridge function at generation g

$$\boxed{F^{(g)} = F^{(0)} - \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}g - d\left(R^{(0)}-R_\infty\right)(1-\mathrm{e}^{-g/\tau}).} \tag{19}$$

The $F^{(g)}$-dynamics of the inner ES is governed by a linearly decreasing term and a term proportional to the initial $\left(R^{(0)}-R_\infty\right)$ difference which reaches a saturation value $d\left(R^{(0)}-R_\infty\right)$ exponentially fast. The latter is approached with the time constant τ, see Eq. (18). The approach is slower for decreasing d and increasing dimensionality N. The parental population size μ decreases the time constant (provided that $c_{\mu/\mu,\lambda} \approx const$).

Now we are able to investigate the γ-dependency of the Meta-ES. Starting with a random point at generation G with initial values $F^{(G)}$, $x^{(G)}$, $R^{(G)}$, and $\sigma^{(G)}$ the inner ES generates two σ-values from the parental $\sigma^{(G)}$

$$\sigma_+ = \alpha\sigma^{(G)} \qquad \text{and} \qquad \sigma_- = \sigma^{(G)}/\alpha. \tag{20}$$

Thus there are two expected function values $F^{(G+\gamma)}$ after γ generations of the inner $[(\mu/\mu_I,\lambda)^\gamma]$-ES

$$F_+^{(G+\gamma)} := F^{(G+\gamma)}(\alpha\sigma^{(G)}) \quad \text{and} \quad F_-^{(G+\gamma)} := F^{(G+\gamma)}(\sigma^{(G)}/\alpha). \tag{21}$$

Using (19) we obtain

$$F_+^{(G+\gamma)} = F^{(G)} - \frac{\alpha\sigma^{(G)}c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\gamma - d\left(R^{(G)} - R_\infty(\alpha\sigma^{(G)})\right)(1-\mathrm{e}^{-\gamma/\tau}) \tag{22}$$

$$F_-^{(G+\gamma)} = F^{(G)} - \frac{\sigma^{(G)}c_{\mu/\mu,\lambda}}{\alpha\sqrt{1+d^2}}\gamma - d\left(R^{(G)} - R_\infty(\sigma^{(G)}/\alpha)\right)(1-\mathrm{e}^{-\gamma/\tau}) \tag{23}$$

The sign of the function $\delta(\gamma)$ defined as the difference of the above constructed function values

$$\delta(\gamma) := F_+^{(G+\gamma)} - F_-^{(G+\gamma)} \tag{24}$$

determines whether to increase or to decrease the mutation strength in the outer $[1,2]$-ES. Since we are aiming at fast F-decrease we get the conditions

$$\begin{aligned} \delta(\gamma) < 0 \quad &\Rightarrow \quad \sigma^{(G+\gamma)} = \alpha\sigma^{(G)} \\ \delta(\gamma) > 0 \quad &\Rightarrow \quad \sigma^{(G+\gamma)} = \sigma^{(G)}/\alpha. \end{aligned} \tag{25}$$

Combining (22) and (23) one gets

$$\delta(\gamma) = \Delta(\gamma)\sigma^{(G)}\left(\alpha - \frac{1}{\alpha}\right) \tag{26}$$

with

$$\Delta(\gamma) := \frac{N\sqrt{1+d^2}}{2\mu c_{\mu/\mu,\lambda}}(1-\mathrm{e}^{-\gamma/\tau}) - \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\gamma. \tag{27}$$

Since $\alpha - \frac{1}{\alpha} > 0$ the condition (25) becomes

$$\begin{aligned} \Delta(\gamma) < 0 \quad &\Rightarrow \quad \sigma^{(G+\gamma)} = \alpha\sigma^{(G)} \\ \Delta(\gamma) > 0 \quad &\Rightarrow \quad \sigma^{(G+\gamma)} = \sigma^{(G)}/\alpha. \end{aligned} \tag{28}$$

Considering the long-term behavior of the F-evolution in (19) one sees that this behavior is mainly governed by the linear g term. Aiming at fast F-decrease, i.e. minimization, the mutation strength σ should be increased by the outer ES implying $\Delta(\gamma) < 0$ in (28). Thus, using (27) we obtain a rule for the choice of the γ-value

$$\begin{aligned} \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\gamma \quad &> \quad \frac{N\sqrt{1+d^2}}{2\mu c_{\mu/\mu,\lambda}}(1-\mathrm{e}^{-\gamma/\tau}) && (29)\\ \Leftrightarrow \qquad \gamma \quad &> \quad \frac{N(1+d^2)}{2\mu c^2_{\mu/\mu,\lambda}}\left(1-\mathrm{e}^{-\gamma/\tau}\right). && (30)\end{aligned}$$

This causes the Meta-ES to steadily increase the mutation strength σ. Since $(1-e^{-k\gamma}) < 1$ the inequality (30) is surely fulfilled for an isolation period $\gamma = \hat{\gamma}$

$$\hat{\gamma} = \left\lceil \frac{N(1+d^2)}{2\mu c^2_{\mu/\mu,\lambda}} \right\rceil. \tag{31}$$

If we consider (31) as well as the truncation ratio $\nu = \frac{\mu}{\lambda}$ and assuming $c_{\mu/\mu,\lambda} \approx 1$, Eq. (30) can be transformed to

$$\boxed{\gamma\lambda > \frac{N(1+d^2)}{2\nu}.} \tag{32}$$

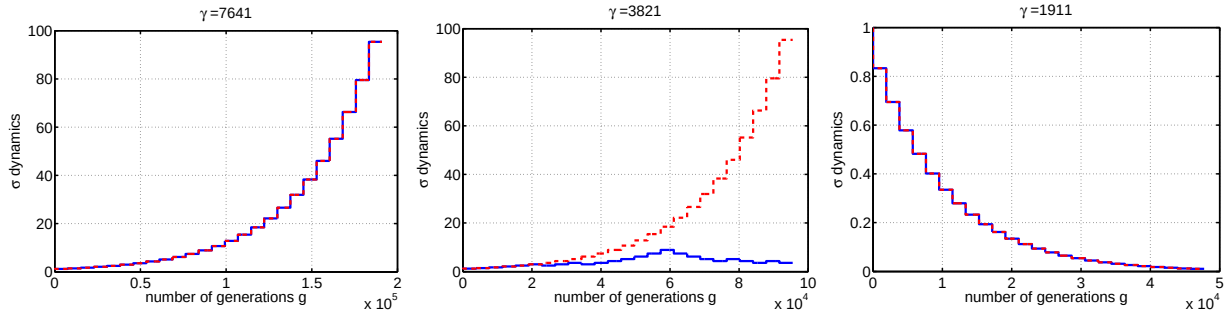

Figure 4: Illustration of the mutation strength behavior for three different choices of the isolation length. The σ-dynamics of a $[1, 2(3/3, 10)^{\gamma}]$-Meta-ES with $\gamma \approx 2\hat{\gamma}$, $\gamma = \hat{\gamma}$ and $\gamma \approx \hat{\gamma}/2$ are shown for $N = 1000$, $\sigma^{(0)} = 1$, $\alpha = 1.2$ on the sharp ridge with $d = 5$ over 25 isolation periods ($\hat{\gamma} = 3821$).

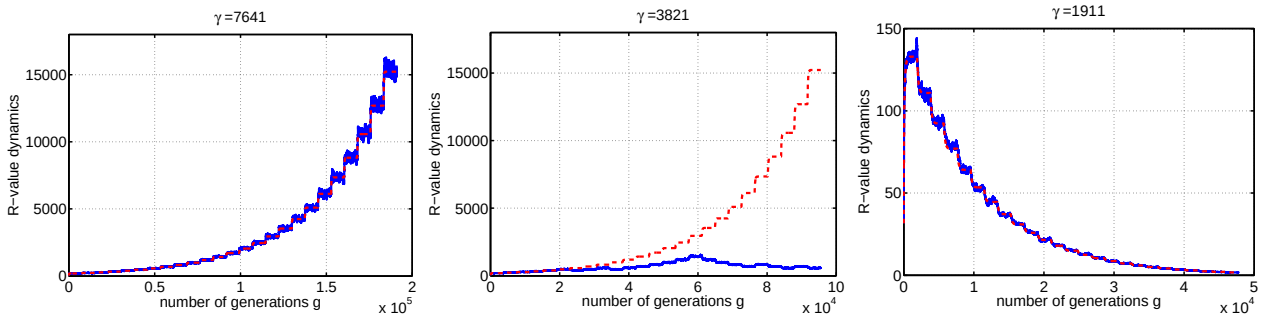

Figure 5: On the R-value dynamics of a $[1, 2(3/3, 10)^{\gamma}]$-Meta-ES with $\gamma \approx 2\hat{\gamma}$, $\gamma = \hat{\gamma}$ and $\gamma \approx \hat{\gamma}/2$ for $N = 1000$, $\sigma^{(0)} = 1$, $\alpha = 1.2$ on the sharp ridge with $d = 5$ over 25 isolation periods ($\hat{\gamma} = 3821$). The distance R to the ridge axis is displayed according to the σ dynamics presented in Fig. 4.

Note, $\gamma\lambda$ is the number of function evaluations during the inner ES. Equation (32) suggests that the divergence behavior of the Meta-ES can be controlled either by choosing a sufficiently large isolation length γ or by increasing the size of the population parameter λ (and μ proportionally).

Now we focus on the precision of our theoretical predictions from Eq. (6) and (7) compared to the results of a experimental run of the Meta-ES algorithm described in Section 2. For that reason the theoretical equations are iterated over γ generations in two runs with different constant mutation strengths $\alpha\sigma$ and σ/α. Then the resulting σ, R, and x values which lead to the best fitness function value $F = -x + dR$ are selected and used as starting points for the next iteration over γ generations. The σ dynamics and the R-value dynamics of the Meta-ES are illustrated for different choices of the isolation length parameter γ in Fig. 4, and Fig. 5 respectively. In each plot the experimental results of a run over 25 isolation periods, or 25γ generations, of the Meta-ES are represented by the solid blue lines whereas the iteratively computed analytical results are depicted as dashed red lines.

There is a good agreement between theoretical predictions and experiments. For long choices of the isolation length γ relative to $\hat{\gamma}$ in (31) we observe overlapping graphs of theoretical and experimental results for the σ dynamics in Fig. 4. That is, the Meta-ES steadily increases the mutation strength after each isolation period. According to the σ-dynamics the strategy increases the distance to the ridge axis with every isolation period, see Fig. 5 for the R dynamics, which leads the Meta-ES to gain constantly better fitness. Even for very short isolation periods $\gamma \ll \hat{\gamma}$ the experimental results suit our predictions. Thus we observe a permanent decrease of the mutation strength in Fig. 4 which is linked to the behavior of the R dynamics to get reduced to zero. This leads the Meta-ES to converge to the ridge axis. That is, choosing γ too small, the ES exhibits premature convergence behavior.

Only the vicinity of the critical γ-value we can observe distinct deviations between theory and experimental results. While the isolation length $\gamma = \hat{\gamma}$ is theoretically sufficient to cause a steady increase of the σ and R dynamics, the experimental run departs from the theoretical results. This does not come as surprise, since we replaced the stochastic dynamics by mean value dynamics.

3.2 Analysis of the $R \ll R_{\infty}$ Case

If the initial point of the Meta-ES lies near the ridge axis we again start the analysis with Eq. (9) and by using (10) one obtains

$$\frac{\mathrm{d}R}{\mathrm{d}g} = \frac{\sigma^2 N}{2\mu}\left(\frac{1}{R} - \frac{1}{R_{\infty}}\right) = \frac{-\sigma d c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\left(\frac{R - R_{\infty}}{R}\right). \tag{33}$$

By separating the variables this leads to the integral equation

$$\int_{R^{(0)}}^{R^{(g)}} \frac{R'}{R' - R_{\infty}}\mathrm{d}R' = \frac{-\sigma d c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}} \int_0^g \mathrm{d}g'. \tag{34}$$

While (34) can be integrated in closed form, it is not possible to solve the resulting equation for $R^{(g)}$. Thus we consider the limit cases of small $R^{(0)}$ and large $R^{(0)}$ states separately. The former case allows for a neglection of R in the denominator of the lhs in (34)

$$\int_{R^{(0)}}^{R^{(g)}} \frac{R'}{R_{\infty}}\mathrm{d}R' = \frac{\sigma d c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}} \int_0^g \mathrm{d}g'. \tag{35}$$

by the use of the condition $R' \ll R_\infty$. Taking into account (10) and solving both integrals yields

$$\frac{1}{2R_\infty}\left(R^{(g)^2} - R^{(0)^2}\right) = \frac{2\sigma d c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}} g \quad (36)$$

$$\left(R^{(g)^2} - R^{(0)^2}\right) = \frac{\sigma^2 N}{\mu} g \quad (37)$$

$$R^{(g)} = \sqrt{R^{(0)^2} + \frac{\sigma^2 N}{\mu} g}. \quad (38)$$

Considering the case that the ES is initialized at the ridge axis, $R^{(0)} = 0$, one obtains

$$R^{(g)} = \sigma\sqrt{Ng/\mu} \propto \sqrt{g}. \quad (39)$$

That is, the ES departs with a random walk $\sqrt{g}$-law from the ridge axis. Using (4), (7), and (39) the fitness at generation g becomes

$$\boxed{F^{(g)} = -x^{(0)} - \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}} g + \frac{d\sigma\sqrt{N}}{\sqrt{\mu}}\sqrt{g}.} \quad (40)$$

For sufficiently large g (keeping $\sigma = const.$) $F^{(g)}$ decreases linearly. This is similar to (19), however, due to the $\sqrt{g}$ term it happens slower.

Considering the difference between $F_+^{(G+\gamma)}$ and $F_-^{(G+\gamma)}$, (21), the σ-learning behavior of the Meta-ES can be analyzed. Using (24) and (40) one gets

$$\delta(\gamma) = \left(\frac{d\sqrt{N}}{\sqrt{\mu}}\sqrt{\gamma} - \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\gamma\right)\sigma^{(G)}\left(\alpha - \frac{1}{\alpha}\right). \quad (41)$$

Now we can define

$$\tilde{\Delta}(\gamma) := \left(\frac{d\sqrt{N}}{\sqrt{\mu}}\sqrt{\gamma} - \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\gamma\right). \quad (42)$$

Similar to (28) we conclude that the Meta-ES increases the mutation strength σ if $\tilde{\Delta}(\gamma) < 0$

$$\tilde{\Delta}(\gamma) < 0 \quad \Leftrightarrow \quad \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\gamma > \frac{d\sqrt{N}}{\sqrt{\mu}}\sqrt{\gamma}. \quad (43)$$

Resolving for the isolation period duration parameter γ, we obtain

$$\gamma \geq \left\lceil \frac{d^2(1+d^2)N}{\mu c^2_{\mu/\mu,\lambda}} \right\rceil \quad (44)$$

$$\Rightarrow \quad \sigma^{(G+\gamma)} = \alpha\sigma^{(G)}. \quad (45)$$

Compared to $\hat{\gamma}$ from Eq. (31) one sees that the isolation period has to be chosen by a factor $2d^2$ larger in order to lead the strategy to diverge from the ridge axis in the case that the Meta-ES starts very close to the ridge axis. However, this result must be taken *cum grano salis*. The bound derived in (44) is based on the R-dynamics (39) that assumed $0 = R^{(0)} \ll R_\infty$. That is, $\sigma\sqrt{Ng/\mu} \ll R_\infty$ must hold. This puts a constraint on the number of generations g the approximation can be used. Taking (10) into account, we see

$$g \ll \frac{N}{4\mu c^2_{\mu/\mu,\lambda}}\left(1 + \frac{1}{d^2}\right) \quad (46)$$

If this condition is violated, the ES already enters the $R \approx R_\infty$ region. Comparing the rhs of (46) with the rhs of (44) we get $d \ll \frac{1}{\sqrt{2}}$. In other words the $R \ll R_\infty$ is left rather fast for sufficient large d and the bound (44) is rather pessimistic. Similar to Section 3.1 we are able to compute the R dynamics iteratively by the use of Eq. (39) and we compare them to the results of an experimental run

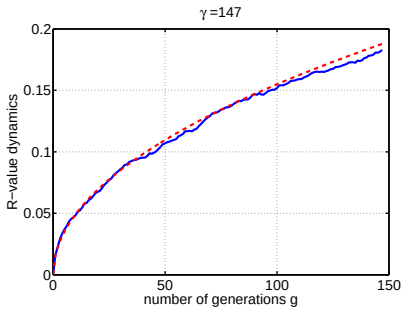

Figure 6: On the R dynamics of the inner strategy of the $[1, 2(3/3, 10)^\gamma]$-Meta-ES initialized on the ridge axis, $R^{(0)} = 0$, with $N = 500$, $d = \frac{1}{100}$, $\sigma^{(0)} = 0.001$ and $\alpha = 1.2$. The distance to the ridge axis is illustrated over the first isolation period of $\gamma = 147$ generations.

of the Meta-ES algorithm, see Fig. 6. Again the iterative result is depicted as dashed red line whereas the experiment is represented by the solid blue line and we observe a very good correspondence between the theory and experiments. Even for choices of small d and σ values the strategy is leaving the initial state on the ridge axis relatively fast. That is, after a few generations the special state of $R \ll R_\infty$ is left and the further dynamics can be described by the theory presented in Section 3.1.

3.3 Analysis of the $R \gg R_\infty$ Case

In the case that the radius R of the initial point is considerably larger than R_∞ we start just like in the previous section. But Eq. (34) now can be simplified to

$$\int_{R^{(0)}}^{R^{(g)}} 1 \mathrm{d}R' = \frac{-\sigma d c}{\sqrt{1+d^2}} \int_0^g \mathrm{d}g' \quad (47)$$

which leads directly to

$$R^{(g)} = R^{(0)} - \frac{d\sigma c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}} g. \quad (48)$$

Using (4), (7), and (48) the fitness dynamics of the inner ES become

$$\boxed{F^{(g)} = F^{(0)} - \sigma c_{\mu/\mu,\lambda}\sqrt{1+d^2} g.} \quad (49)$$

Considering the fitness difference in the Meta-ES (24), we immediately obtain

$$F_+^{(G+\gamma)} - F_-^{(G+\gamma)} = -\gamma c_{\mu/\mu,\lambda}\sqrt{1+d^2}\left(\alpha - \frac{1}{\alpha}\right)\sigma \quad (50)$$

which is equivalent to

$$\delta(\gamma) < 0 \quad \forall \gamma \in \mathbb{N}. \quad (51)$$

That is, each choice of γ is leading the Meta-ES to a steady increase of the mutation strength σ until the strategy reaches the steady state distance R_∞ and the dynamics described in Section 3.1 hold again.

Fig. 7 compares the R-value dynamics of two runs of the Meta-ES algorithm with the iteratively computed predictions of Eq. (9). In both cases we observe a good agreement of analytical and experimental results. The lhs of the illustration shows the strategy with a rather short isolation length γ, i.e. $\gamma \approx \hat{\gamma}/2$ whereas on the rhs of Fig. 7 a long isolation time of $\gamma \approx 2\hat{\gamma}$ generations has been chosen.

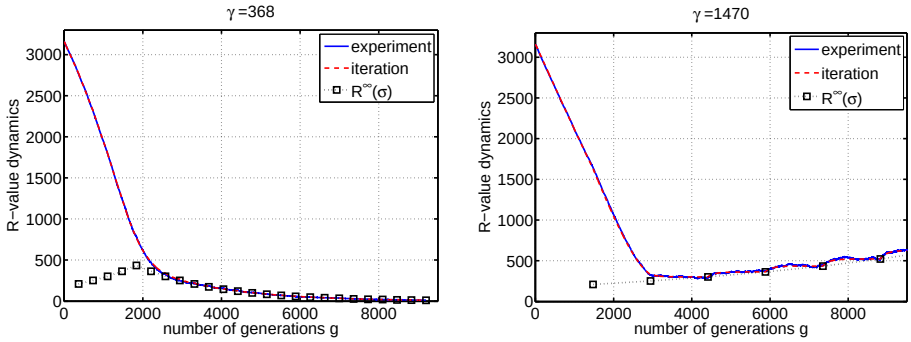

Figure 7: On the R dynamics of the $[1, 2(3/3, 10)^{\gamma}]$-Meta-ES initialized for $R^{(0)} \gg R_{\infty}$, with $N = 1000$, $d = 2$, $\sigma^{(0)} = 1$ and $\alpha = 1.2$. The distance to the ridge axis is displayed over 9500 generations with an isolation length of $368 = \gamma < \hat{\gamma}$ generations on the lhs and of $1470 = \gamma > \hat{\gamma}$ on the rhs of the figure. The black squares symbolize the location of the steady state distance R_{∞} of the respective isolation period which depends on the previously selected mutation strength σ, see also Fig. 8 for the respective σ dynamics.

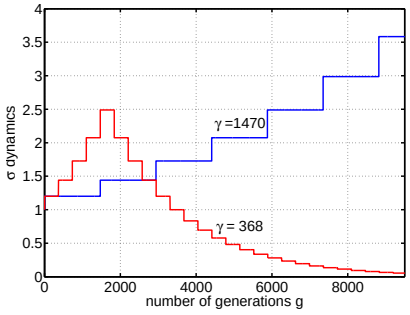

Figure 8: On the σ dynamics of the $[1, 2(3/3, 10)^{\gamma}]$-Meta-ES initialized with $R^{(0)} \gg R_{\infty}$, with $N = 1000$, $d = 2$, $\sigma^{(0)} = 1$ and $\alpha = 1.2$. For the respective R dynamics see Fig. 7. The blue line represents iteratively generated results of the σ-dynamics for long isolation time $\gamma = 1470$ whereas the red line depicts these results for the choice of a short isolation length parameter ($\gamma = 368$). Because the experimental results correspond to the iterative ones for both choices of γ, in this figure we disregard them for reasons of clarity.

As we can see in the respective σ dynamics, Fig. 8, the strategy increases the mutation strength until a sufficient vicinity to the steady state distance is reached in a certain isolation period. That is, we observe a great compliance with the theoretical results, see Eq. (51). Afterwards the Meta-ESs behaviors continue as described in Section 3.1, i.e., for short isolation time the strategy converges to the ridge axis and decreases the mutation strength. Conversely it diverges and steadily increases the mutation strength for long isolation periods.

4. BEHAVIOR OVER MULTIPLE ISOLATION PERIODS

Since in Section 3 the dynamics of the inner ES were already illustrated over several isolation periods and compared to the experimental results of the algorithm we are now going to theoretically investigate the behavior of the Meta-ES over multiple isolation periods. Taking $G + \gamma$ instead of g, defining

$$b := \frac{N\sqrt{1+d^2}}{2\mu c_{\mu/\mu,\lambda} d} \tag{52}$$

and remembering (18), Eq. (17) becomes

$$R^{(G+\gamma)} = R^{(G)} \mathrm{e}^{-\gamma/\tau} + b\sigma^{(G)} \left(1 - \mathrm{e}^{-\gamma/\tau}\right). \tag{53}$$

Our analysis is based on the results derived for the $R \approx R_{\infty}$ case in Section 3.1. That is sufficient because in both other cases, i.e. $R \ll R_{\infty}$ or $R \gg R_{\infty}$, the initial state is left rather fast with the strategy making progress towards the steady state distance R_{∞}. Assuming the choice of γ big enough, i.e. $\gamma > \hat{\gamma}$ in (31), the Meta-ES is supposed to increase the mutation strength after each isolation period. Thus, after the first isolation over γ generations the algorithm picks the parental $\mathbf{y}_p$ out of the inner ES with the distance to the ridge axis given by

$$R^{(G+\gamma)} = R^{(G)} \mathrm{e}^{-\gamma/\tau} + b\alpha\sigma^{(G)} \left(1 - \mathrm{e}^{-\gamma/\tau}\right). \tag{54}$$

After the second isolation period with

$$\sigma_p := \sigma^{(G+\gamma)} = \alpha\sigma^{(G)} \tag{55}$$

again the inner ES with increased mutation strength is chosen

$$\begin{aligned} R^{(G+2\gamma)} &= R^{(G+\gamma)} \mathrm{e}^{-\gamma/\tau} + b\alpha\sigma^{(G+\gamma)} \left(1 - \mathrm{e}^{-\gamma/\tau}\right) && (56) \\ &= R^{(G)} \mathrm{e}^{-2\gamma/\tau} + b\alpha\sigma^{(G)} \left(1 - \mathrm{e}^{-\gamma/\tau}\right)\left(\mathrm{e}^{-\gamma/\tau} + \alpha\right). && (57) \end{aligned}$$

Continuing this way we are able to predict the expected distance to the ridge axis $R^{(G+t\gamma)}$ after t isolation periods

$$R^{(G+t\gamma)} = R^{(G)} \mathrm{e}^{-t\gamma/\tau} + b\alpha\sigma^{(G)} \left(1 - \mathrm{e}^{-\gamma/\tau}\right) \mathrm{e}^{-(t-1)\gamma/\tau} \sum_{i=0}^{t-1} \alpha^i \mathrm{e}^{i\gamma/\tau}. \tag{58}$$

Applying the geometric series formula this equation becomes

$$R^{(G+t\gamma)} = R^{(G)} \mathrm{e}^{-t\gamma/\tau} + b\alpha\sigma^{(G)} \left(1 - \mathrm{e}^{-\gamma/\tau}\right) \mathrm{e}^{-(t-1)\gamma/\tau} \frac{\alpha^t \mathrm{e}^{t\gamma/\tau} - 1}{\alpha \mathrm{e}^{\gamma/\tau} - 1}. \tag{59}$$

Rewriting the second term,

$$\begin{aligned} & b\alpha\sigma^{(G)} \left(1 - \mathrm{e}^{-\gamma/\tau}\right) \mathrm{e}^{-(t-1)\gamma/\tau} \frac{\alpha^t \mathrm{e}^{t\gamma/\tau} - 1}{\alpha \mathrm{e}^{\gamma/\tau} - 1} && (60) \\ = \quad & b\sigma^{(G)} \left(1 - \mathrm{e}^{-\gamma/\tau}\right) \alpha^t \frac{1 - \alpha^{-t} \mathrm{e}^{-t\gamma/\tau}}{1 - \alpha^{-1} \mathrm{e}^{-\gamma/\tau}} && (61) \end{aligned}$$

and using (10), the R-dynamics becomes

$$R^{(G+t\gamma)} = R^{(G)} \mathrm{e}^{-t\gamma/\tau} + R_{\infty}^{(G)} \alpha^t \left(1 - \mathrm{e}^{-\gamma/\tau}\right) \frac{1 - \alpha^{-t} \mathrm{e}^{-t\gamma/\tau}}{1 - \alpha^{-1} \mathrm{e}^{-\gamma/\tau}}, \tag{62}$$

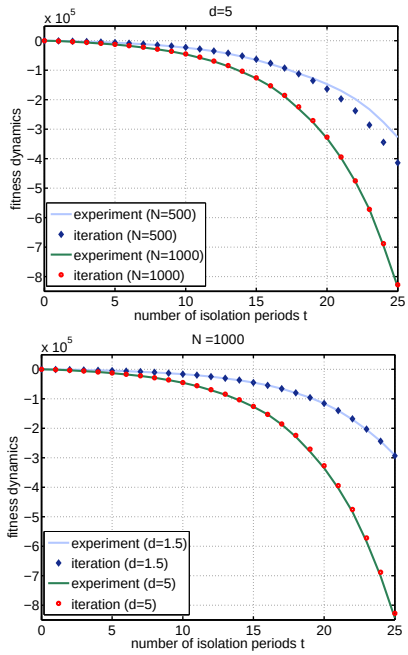

Figure 9: The upper graphs show the fitness dynamics of an $[1, 2(3/3, 10)^{2\hat{\gamma}}]$-Meta-ES with $\sigma^{(0)} = 1$, $\alpha = 1.2$ and two different choices of the dimension N and the nonlinearity strength d each. The dynamics are illustrated over 25 isolation periods of $2\hat{\gamma}$ generations with $\hat{\gamma}$ from (31) depending on N and d.

with $R^{(G)}_\infty = \frac{N\sqrt{1+d^2}}{2\mu c_{\mu/\mu,\lambda} d}\sigma^{(G)}$.
Equation (62) simplifies when taking into account $e^{-\gamma/\tau} \ll 1$

$$R^{(G+t\gamma)} \simeq R^{(G)}_\infty \alpha^t. \tag{63}$$

In order to have this dynamics, we have to demand $e^{-\gamma/\tau} \ll 1$, i.e. $\gamma/\tau \gg 1$. Using an isolation period (31) and τ given by (18), this transfers to the condition

$$\gamma/\tau = d^2 \gg 1 \tag{64}$$

which is approximately fulfilled for sufficiently large nonlinearity parameter d in (4). In the same manner we can derive a formula for the progress of the Meta-ES in direction of the ridge axis. Using (7) and (55) one gets

$$x^{(G+\gamma)} = x^{(G)} + \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\alpha\sigma^{(G)}\gamma \tag{65}$$

$$x^{(G+2\gamma)} = x^{(G+\gamma)} + \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\gamma\alpha\sigma^{(G+\gamma)} \tag{66}$$

$$= x^{(G)} + \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\gamma\sigma^{(G)}\alpha(1+\alpha). \tag{67}$$

After t isolation periods we have

$$x^{(G+t\gamma)} = x^{(G)} + \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\gamma\sigma^{(G)}\alpha\sum_{i=0}^{t-1}\alpha^i. \tag{68}$$

By use of the geometric series formula this can be simplified to

$$x^{(G+t\gamma)} = x^{(G)} + \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\gamma\sigma^{(G)}\alpha\frac{\alpha^t-1}{\alpha-1}. \tag{69}$$

Using (63), (68), and (4) the approximate F-dynamics of the Meta-ES reads ($x^{(G)} = 0$)

$$F^{(G+t\gamma)} \simeq -\frac{c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\gamma\sigma^{(G)}\alpha\frac{\alpha^t-1}{\alpha-1} + dR^{(G)}_\infty\alpha^t. \tag{70}$$

This leads to an asymptotic dynamics

$$F^{(G+t\gamma)} \sim -\alpha^t\sigma^{(G)}\frac{N\sqrt{1+d^2}}{2\mu c_{\mu/\mu,\lambda}}\left(\frac{2\mu c^2_{\mu/\mu,\lambda}}{N(1+d^2)}\frac{\gamma\alpha}{\alpha-1} - 1\right). \tag{71}$$

That is, the Meta-ES diverges exponentially fast at a rate of $\ln\alpha$.

The fitness value dynamics predicted in Eq. (70) compared to three different Meta-ES runs are displayed in Fig. 9 for $N = 500$ and $N = 1000$. We observe an increasing compliance of iteratively computed and experimental results by increasing the dimensionality N. On the lower depiction of Fig. 9 the influence of the nonlinearity strength parameter d is shown, i.e. the fitness dynamics of two Meta-ESs on different fitness landscapes are illustrated. The use of a bigger nonlinearity strength d causes longer isolation periods, see (31), letting the Meta-ES reach better fitness. For long isolation periods and high dimensionalities the predictions agree with the experiments.

5. FINITE SEARCH SPACE DIMENSIONALITY EFFECTS

The inequality (32) offers an astonishing interpretation: The progress on the ridge function only depends on the number of function evaluations being larger than the constant term on the rhs of (32). That is, in the asymptotic limit case ($N \to \infty$) isolation periods > 1 are unnecessary. This conclusion calls for an examination of the Meta-ES dynamics in low-dimensional spaces. From the progress rate analysis of the sphere model and also the ridge function we know that the progress rate does not linearly scale with the population size. That is, the increase of μ does only make sense if $\mu \ll N$. If this condition is violated, $\gamma > 1$ might be of use.

Therefore, we investigate the Meta-ES behaviors for a constant number of function evaluations but with different choices of population size parameters (μ, λ) and isolation length parameter γ under the condition of small N. Using (32) with $\nu = 0.25$ we consider

$$\frac{N(1+d^2)}{2} < \mu\gamma = 2^\beta = 2^{\beta_1}2^{\beta_2} = const. \tag{72}$$

with $\mu := 2^{\beta_1}$, $\gamma := 2^{\beta_2}$ and $\beta = \beta_1 + \beta_2 \in \mathbb{N}$. As we can see in the illustrations for small dimensional spaces, see Fig. 10 ($\beta = 8$, $N = 30$, $d = 0.9$) and Fig. 11 ($\beta = 8$, $N = 30$, $d = 3$), the use of isolation within the Meta-ES leads to a great improvement in the case of small search space dimensions. In Fig. 10 the inequality in (32) is relatively high because of the small choice of d which results in similar mutation strength dynamics and therefore overlapping graphs of the illustrated five Meta-ES settings with $(\mu = 1, \gamma = 256)$, $(\mu = 4, \gamma = 64)$, $(\mu = 16, \gamma = 16)$, $(\mu = 64, \gamma = 4)$ and $(\mu = 256, \gamma = 1)$. While the σ dynamics are equal we can observe that the Meta-ESs with longer isolation periods and smaller populations outperform the conversely constructed ones with respect to their R-value and fitness dynamics compared against the same number of function evaluations (or isolation periods respectively). The situation illustrated in Fig. 11 describes a low inequality in Eq. (32). In contrast to Fig. 10 the R-value dynamics of the Meta-ES settings with short isolation periods and larger population sizes tend to converge to the ridge axis. This leads to premature convergence behavior while longer isolation periods still lead to increase the distance to the ridge axis and gain greater progress of the fitness values. Regarding the σ dynamics in Fig. 11, we see that no Meta-ES setting steadily increases the mutation strength under the

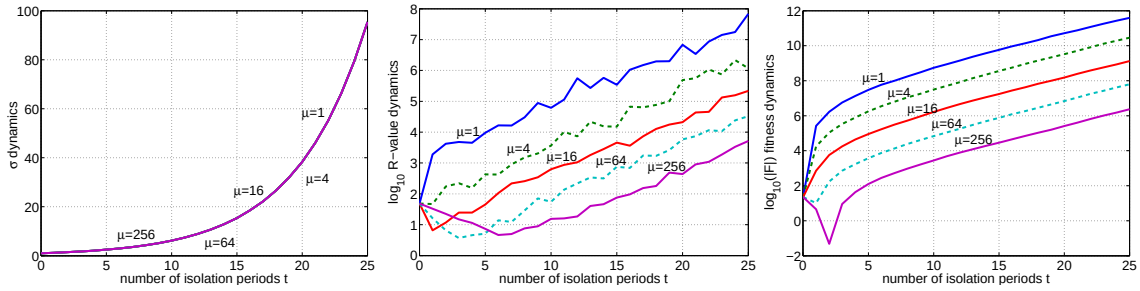

Figure 10: Illustration of the $[1, 2(\mu/\mu, \lambda)^\gamma]$-Meta-ES dynamics for dimension $N = 30$, nonlinearity strength parameter $d = 0.9$ and truncation ratio $\nu = 0.25$. The mutation strength σ, the distance to the ridge axis R and the fitness function value F are compared for five different combinations of $\gamma\mu = 2^8 = 256$ over a time of 25 isolation periods.

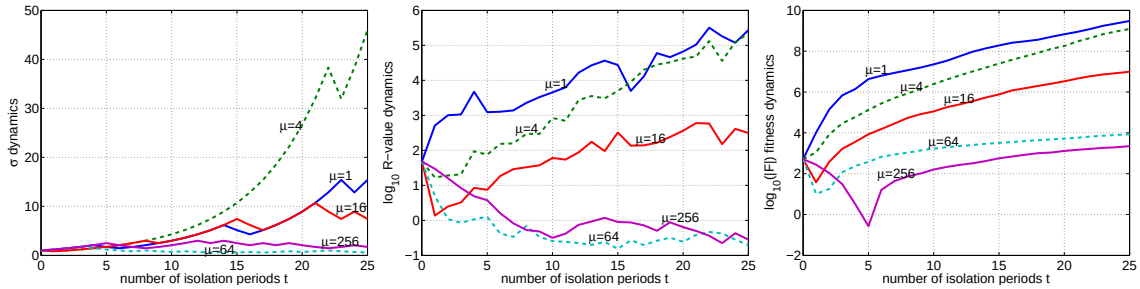

Figure 11: Depiction of the $[1, 2(\mu/\mu, \lambda)^\gamma]$-Meta-ES dynamics for dimension $N = 30$, $d = 3$ and $\nu = 0.25$. Again the dynamics of the five Meta-ES runs with $\gamma\mu = 2^8 = 256$ are compared over 25 isolation periods. Note that Fig. 10 only differs from Fig. 11 by the choice of the nonlinearity strength parameter d.

condition that the inequality (32) is not sufficiently satisfied. But again rather the settings with longer isolation periods and smaller population sizes tend to increase the mutation strength. That is, leaving the asymptotic limit case we can infer advantages from the use of a sufficiently long isolation period.

6. CONCLUSIONS AND OUTLOOK

In this paper we examined the σ-control of the $[1, 2(\mu/\mu, \lambda)^\gamma]$-Meta-ES on the sharp ridge. Our theoretical analysis allows for a deeper understanding of the influence of the isolation length γ on the performance of the Meta-ES. We derived an estimate for the choice of γ which depends on the population sizes λ and μ and the dimensionality N as well as d being the nonlinearity strength parameter determining the "hardness" of the sharp ridge. From earlier empirical investigations [5] it was conjectured that the isolation parameter γ has considerable influence on the performance of the Meta-ES. The asymptotical analysis done here, however, revealed that it is not γ alone that determines the qualitative behavior (whether there is premature convergence or not), but the number of function evaluations $\gamma\lambda$ devoted to the inner ES. That is, similar convergence effects can be obtained by either increasing the population size or the isolation period. This is a very remarkable results, since it predicts Meta-ES functioning even for $\gamma = 1$ provided that the population size has been chosen sufficiently large, see Inequality (32). The result has been obtained for $N \to \infty$ assuming that the population size is much smaller than the search space dimensionality N. This still leaves room for $\gamma > 1$ advantages. Considering small dimensionalities, we have found indeed experimental evidence for advantages of longer isolation periods.

The investigations should be extended to other test problems where the choice of the populations size versus the isolation time might have a much greater impact on the ES performance, as perhaps for noisy problems. The theoretical analysis performed will be also the basis for the design of theoretically motivated population size control rules.

7. ACKNOWLEDGMENTS

This work was supported by the Austrian Science Fund (FWF) under grant P22649-N23.

8. REFERENCES

[1] D. V. Arnold and A. MacLeod. Step length adaption on ridge functions. *Evolutionary Computation*, 16:151–184, 2008.

[2] H.-G. Beyer. On the Performance on $(1, \lambda)$-Evolution Strategies for the Ridge Function Class. *IEEE Transactions on Evolutionary Computation*, 5(3), June 2001.

[3] H.-G. Beyer. *The Theory of Evolution Strategies*. Natural Computing Series, Springer, Heidelberg, 2001.

[4] H.-G. Beyer, M. Dobler, C. Hämmerle, and P. Masser. On Strategy Parameter Control by Meta-ES. GECCO'09, 2009.

[5] M. Herdy. Reproductive Isolation as Strategy parameter in Hierarchically Organized Evolution Strategies. In *R. Männer and B. Manderick, editors, Parallel Problem Solving from Nature*, volume 2, pages 207–217. Elsevier, 1992.

[6] M. Lunacek and D. Whitley. Searching for Balance: Understanding Self-Adaption on Ridge Functions. PPSN, 2006.

[7] S. Meyer-Nieberg. *Self-Adaptation in Evolution Strategies*. PhD thesis, Dortmund, 2007.

[8] I. Rechenberg. *Evolutionsstrategie '94*. Frommann-Holzboog Verlag, Stuttgart, 1994.