Step Length Adaptation on Ridge Functions

Dirk V. Arnold
Alexander MacLeod

Technical Report CS-2006-08

October 19, 2006

Faculty of Computer Science
6050 University Ave., Halifax, Nova Scotia, B3H 1W5, Canada

Step Length Adaptation on Ridge Functions

Dirk V. Arnold dirk@cs.dal.ca
Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia,
Canada B3H 1W5

Alexander MacLeod amacleod@cs.dal.ca
Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia,
Canada B3H 1W5

Abstract

Step length adaptation is central to evolutionary algorithms in real-valued search spaces. This paper contrasts several step length adaptation algorithms for evolution strategies on a family of ridge functions. The algorithms considered are cumulative step length adaptation, a variant of mutative self-adaptation, two-point adaptation, and hierarchically organised strategies. In all cases, analytical results are derived that yield insights into scaling properties of the algorithms. The influence of noise on adaptation behaviour is investigated. Similarities and differences between the adaptation strategies are discussed.

Keywords

Evolution strategies, step length adaptation, ridge functions.

1 Introduction

Adaptation of the mutation strength is often a necessary requirement for successful evolutionary optimisation in real-valued search spaces. Mutation strength adaptation algorithms that have been proposed include cumulative step length adaptation (Ostermeier et al., 1994), mutative self-adaptation (Rechenberg, 1973; Schwefel, 1981), two-point adaptation (Salomon, 1998), and hierarchically organised strategies (Herdy, 1992; Rechenberg, 1994). Analytical knowledge with regard to properties of adaptation mechanisms is sparse, and few comparisons between the different approaches have been published. This paper analyses the behaviour of the aforementioned adaptation strategies on a family of ridge functions, revealing similarities as well as important differences.

The choice of ridge functions as a test bed for evaluating the performance of adaptation strategies has been made for two reasons. First, they are among the simplest nontrivial test functions that can be considered. The mutation strength of adaptive evolution strategies on a ridge converges to zero, diverges, or assumes a state that is characterised by a stationary probability distribution. Due to the symmetries inherent in the problem, that state is described by a small number of parameters analytical approximations to which can be readily obtained. The equations that are derived (approximately) hold for a wide range of parameter values and are thus more valuable than experimental results obtained for particular parameter settings. Initialisation conditions and termination criteria are irrelevant for the results obtained. And second, it has been conjectured that ridge following is a recurring task in numerical optimisation. (Whitley et al., 2004) state that while the difficulties of optimising ridges "are relatively well documented in the mathematical literature on derivative free minimization algorithms [. . .], there is little discussion of this problem in the heuristic search literature". Findings with regard to the suitability of adaptation strategies for ridge following can thus be expected to be of practical significance.

The past decade has seen an increasing interest in an analytical understanding of adaptation strategies. (Beyer, 1996) studies the performance of the $(1, \lambda)$-ES with mutative self-adaptation for the sphere model. (Meyer-Nieberg and Beyer, 2005) consider the more general $(\mu/\mu, \lambda)$-ES. (Hansen, 2006) shows that mutative self-adaptation is not without problems on linear fitness functions. See (Meyer-Nieberg and Beyer, 2006) for pointers to further work on mutative self-adaptation, including convergence results for spherically symmetric objective functions obtained using a Markov chain approach. (Beyer and Arnold, 2003; Arnold and Beyer, 2004) study the performance of the $(\mu/\mu, \lambda)$-ES with cumulative step length adaptation on the sphere model. The performance of (non-adaptive) evolution strategies on parabolic ridges is analysed by (Oyman et al., 1998; Oyman et al., 2000; Oyman and Beyer, 2000). The results are generalised for other ridge topologies by (Beyer, 2001a). Adaptive evolution strategies on ridge functions have first been studied by (Herdy, 1992). In that reference, it is seen in experiments that mutative self-adaptation performs unsatisfactorily on some ridges, and hierarchically organised strategies are proposed as an alternative. A first analytical investigation of the behaviour of hierarchically organised evolution strategies on parabolic ridges has recently been presented by (Arnold and MacLeod, 2006). Mutative self-adaptation on sharp ridges is the subject of the analysis presented by (Beyer and Meyer-Nieberg, 2006). Cumulative step length adaptation on parabolic ridges is studied by (Arnold and Beyer, 2006) for different forms of noise present. That analysis is generalised by (Arnold, 2006) for other ridge topologies.

The approach followed in this paper is similar to that pursued by (Beyer, 2001b) and in other work in that the one-step behaviour of evolution strategies is described by a set of nonlinear, stochastic difference equations. The goal of the analyses is to find a stationary state that is characterised by the time-invariance of the variables that describe the state of the system. Both the nonlinearity and the randomness in the evolution equations make it impossible to obtain an exact solution without introducing simplifications. Such simplifications include ignoring stochastic effects by replacing variables with their expected values, and the assumption of high search space dimensionality. Where the goals of simplicity and accuracy are conflicting, in this paper we opt for the former. Better approximations could be derived, but they would likely act to obscure the lessons learnt here. Computer experiments are used to provide a sense of the accuracy of the predictions and to show that the qualitative characteristics of the behaviour of the strategies is captured correctly in the analyses.

The remainder of this paper is organised as follows. Section 2 outlines single steps of the basic strategy as well as the class of ridge functions considered. Previously obtained results with regard to the performance of the strategy are summarised in as far as they are relevant in the present context. Sections 3 through 6 consider, in order, cumulative step length adaptation, a variant of mutative self-adaptation, two-point adaptation, and hierarchically organised strategies. In each case, a performance law that describes the strategy's behaviour is derived, and its accuracy is verified experimentally. The analysis of the behaviour of cumulative step length adaptation generalises that in (Arnold and Beyer, 2006) by including ridges other than parabolic ones, and it generalises that in (Arnold, 2006) by including noise. The approach to the analyses of mutative self-adaptation and two-point adaptation is new and will conceivably prove useful in other fitness environments in the future. The treatment of hierarchically organised evolution strategies generalises that in (Arnold and MacLeod, 2006) both by considering ridges other than parabolic ones and by including the effects of noise. Section 7 contrasts and compares the properties of the different adaptation algorithms, and it concludes with suggestions for future work.

2 Preliminaries

This section first describes the $(\mu/\mu, \lambda)$-ES with isotropically distributed mutations as the basic strategy used to generate single steps. Then, the ridge function class is introduced, and previously derived results that form the basis of the computations in later sections are summarised.

2.1 Basic Strategy

The $(\mu/\mu, \lambda)$-ES with isotropically distributed mutations is an evolution strategy used for the optimisation of functions $f : I\!R^N \to I\!R$. It is popular both due to its good local optimisation performance and its relative amenability to theoretical analysis. The $(\mu/\mu, \lambda)$-ES is an instance of the more general $(\mu/\rho \overset{+}{,} \lambda)$-ES where $\rho = \mu$ (i.e., the entire population is parent to every offspring candidate solution generated), and comma selection is used (i.e., the life span of an individual cannot exceed a single generation). See (Beyer and Schwefel, 2002) for a comprehensive introduction to evolution strategies and the $(\mu/\rho \overset{+}{,} \lambda)$-notation.

In every time step the $(\mu/\mu, \lambda)$-ES updates a search point $\mathbf{x} \in I\!R^N$ (the centroid of its population) using the following four steps:

1. A set of λ offspring candidate solutions $\mathbf{y}^{(i)} = \mathbf{x} + \sigma\mathbf{z}^{(i)}$, $i = 1, \ldots, \lambda$, is generated. Mutation strength $\sigma > 0$ determines the step length and the mutation vectors $\mathbf{z}^{(i)} \in I\!R^N$ consist of independent, standard normally distributed components.

2. The objective function values $f(\mathbf{y}^{(i)})$ of the offspring candidate solutions are determined.

3. Letting the index $k;\lambda$ refer to the kth best (i.e., the kth largest if the task is maximisation and the kth smallest if the task is minimisation) of the offspring candidate solutions, the average

$$\mathbf{z}^{(\text{avg})} = \frac{1}{\mu}\sum_{k=1}^{\mu} \mathbf{z}^{(k;\lambda)} \tag{1}$$

   of those μ mutation vectors that correspond to the μ best offspring candidate solutions is computed.

4. The search point is updated according to

$$\mathbf{x} \leftarrow \mathbf{x} + \sigma\mathbf{z}^{(\text{avg})} \tag{2}$$

   where "$\leftarrow$" denotes the assignment operator.

Different mechanisms for the adaptation of the mutation strength σ will be considered in Sections 3 through 6. Arguably the most significant limitation of the strategy outlined here is its reliance on isotropically distributed mutations. (Whitley et al., 2004) point out that significantly improved performance on ridge functions can be achieved if mutation vectors are generated using a general N-dimensional normal distribution

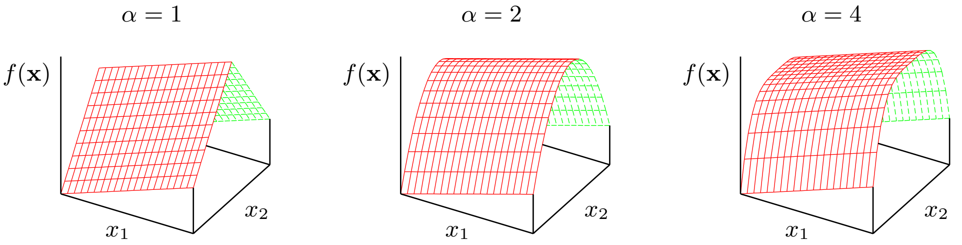

Figure 1: Plots of two-dimensional ridges with topology parameter $\alpha = 1$ (sharp ridge), $\alpha = 2$ (parabolic ridge), and $\alpha = 4$.

with appropriately chosen variances and covariances. Adapting the $N(N+1)/2$ parameters that constitute the covariance matrix is a more difficult task both to solve and to analyse than adapting a single step length. While strategies such as the covariance matrix adaptation evolution strategy (CMA-ES) proposed by (Hansen and Ostermeier, 2001) exist, the analysis of their performance remains as a task for future work. It is conceivable that the analyses presented here are of significance even when using non-isotropically distributed mutations as strategies such as the CMA-ES adapt their step length separately from the covariance matrix.

2.2 The Ridge Function Class

The class of objective functions considered throughout this paper is

$$f(\mathbf{x}) = x_1 - d\left(\sum_{i=2}^{N} x_i^2\right)^{\alpha/2}, \qquad \mathbf{x} = \langle x_1, \ldots, x_N \rangle \in I\!R^N \tag{3}$$

where $d > 0$ and $\alpha \geq 1$. The parameter α is referred to as the topology parameter. Ridges with $\alpha = 1$ and $\alpha = 2$ are referred to as sharp and parabolic ridges, respectively. Figure 1 shows plots of several two-dimensional ridges. The x_1-axis is referred to as the ridge axis. Notice that while in the definition used here the ridge axis is aligned with an axis of the coordinate system and the objective function is thus separable, that fact is irrelevant for a strategy that uses isotropically distributed mutations such as those considered in the present paper. The coordinate system could be subjected to an arbitrary rotation without affecting the strategies' performance.

While ridge functions as defined in Eq. (3) have no finite maximum, maximisation is still a meaningful task if increasing objective function values is considered the goal of optimisation. Candidate solutions with superior fitness can be achieved in two different ways: by making progress in the direction of the ridge axis (i.e., by increasing x_1) or by reducing the distance from the ridge axis. In the short term, the latter may be more successful; however, in the long term, only the former is a viable possibility as the distance from the ridge axis is always nonnegative. Therefore, as in (Oyman et al., 1998) and later work, the performance of evolution strategies on ridge functions is quantified by the progress rate

$$\varphi = \mathrm{E}\left[\sigma z_1^{(\mathrm{avg})}\right] \tag{4}$$

i.e., the expected progress of the search point in the direction of the ridge axis in a single time step.

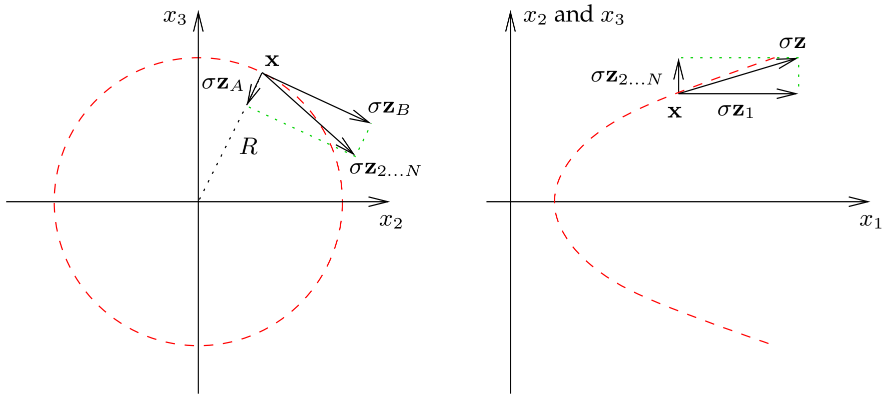

Figure 2: Decomposition of vector $\mathbf{z}$ into its axial component $\mathbf{z}_1$, central component $\mathbf{z}_A$, and lateral component $\mathbf{z}_B$ for $N = 3$. The dashed lines indicate locations of constant fitness.

Finally, real-world optimisation problems often suffer from noise present in the process of evaluating the quality of candidate solutions. Such noise can be a consequence of factors as varied as the use of Monte Carlo techniques, physical measurement limitations, or human input in the selection process. Noise is most frequently modelled by an additive Gaussian term with mean zero and with a standard deviation σ_ϵ that is referred to as the noise strength. The noisy fitness $f_\epsilon(\mathbf{y}) = f(\mathbf{y}) + \sigma_\epsilon z_\epsilon$ of a candidate solution $\mathbf{y}$, where z_ϵ is a standard normally distributed random variate, is the fitness observed upon evaluation of the fitness function. In order to investigate the robustness of adaptation algorithms, we include noise of constant strength that is independent of the location in search space in the analyses. See (Arnold and Beyer, 2006) for other forms of noise that could be considered in future work.

2.3 Single Step Performance

For given mutation strength, after initialisation effects have faded the $(\mu/\mu, \lambda)$-ES assumes a state in which the ridge is tracked at a distance that fluctuates, but the distribution of which is stationary. It can be observed that the relative magnitude of the fluctuations decreases with increasing search space dimensionality. In the limit $N \to \infty$ an expression for the average distance at which the ridge is tracked can be obtained. The following derivation generalises that in (Arnold and Beyer, 2006) by considering ridges other than parabolic ones, and it extends that in (Arnold, 2006) by including the effects of noise in the calculations.

Letting $\mathbf{x}_{2\ldots N} = \langle 0, x_2, \ldots, x_N \rangle$ denote the projection of the search point onto the hyperplane with $x_1 = 0$, throughout this paper $R = \|\mathbf{x}_{2\ldots N}\|$ denotes the distance of the search point from the ridge axis. Central to the analysis of the performance of evolution strategies on ridge functions is a decomposition of mutation and progress vectors into three mutually orthogonal components $\mathbf{z}_1$, $\mathbf{z}_A$, and $\mathbf{z}_B$ that has been employed in (Arnold and Beyer, 2006; Arnold, 2006) and that is illustrated in Fig. 2. Vector $\mathbf{z}_1 = \langle z_1, 0, \ldots, 0 \rangle$ points in the direction of the ridge axis and is referred to as the axial component of $\mathbf{z}$. Letting $\mathbf{z}_{2\ldots N} = \langle 0, z_2, \ldots, z_N \rangle$, scalar quantity $z_A = -\mathbf{x}_{2\ldots N} \cdot \mathbf{z}_{2\ldots N}/R$ is the signed length of the central component $\mathbf{z}_A = -z_A \mathbf{x}_{2\ldots N}/R$ of vector $\mathbf{z}$ that points from the search point toward the ridge axis. Vector $\mathbf{z}_B$ equals $\mathbf{z}_{2\ldots N} - \mathbf{z}_A$ and is referred to as the lateral component of $\mathbf{z}$. Altogether, $\mathbf{z} = \mathbf{z}_1 + \mathbf{z}_A + \mathbf{z}_B$.

Consider the noisy fitness of an offspring candidate solution $\mathbf{y} = \mathbf{x} + \sigma\mathbf{z}$. Using the definitions of z_A and $\mathbf{z}_{2\ldots N}$, it follows from Eq. (3) that

$$\begin{aligned} f_\epsilon(\mathbf{y}) &= x_1 + \sigma z_1 - d\left(\sum_{i=2}^{N}(x_i + \sigma z_i)^2\right)^{\alpha/2} + \sigma_\epsilon z_\epsilon \\ &= x_1 + \sigma z_1 - d\left(R^2 - 2R\sigma z_A + \sigma^2\|\mathbf{z}_{2\ldots N}\|^2\right)^{\alpha/2} + \sigma_\epsilon z_\epsilon. \end{aligned} \tag{5}$$

As $\mathbf{z}$ is a mutation vector, $\|\mathbf{z}_{2\ldots N}\|^2$ is χ^2_{N-1}-distributed and has mean $N-1$ and variance $2(N-1)$. Because the distribution of mutation vectors is isotropic, z_A is standard normally distributed. Let us assume that

$$R \gg \sigma\sqrt{N}. \tag{6}$$

It will be seen below that for given mutation strength σ the resulting stationary distance R of the search point from the ridge axis is such that with increasing N, $\sigma\sqrt{N}/R$ tends to zero, thus providing an a posteriori justification for Eq. (6). Under the assumption, the first term in the parentheses in Eq. (5) dominates the other two. The power term can thus be expanded into a Taylor series with terms beyond the linear one ignored, yielding

$$\begin{aligned} f_\epsilon(\mathbf{y}) &\overset{N\to\infty}{=} x_1 + \sigma z_1 - d\left(R^\alpha - \frac{\alpha}{2}R^{\alpha-2}\left(2R\sigma z_A - \sigma^2\|\mathbf{z}_{2\ldots N}\|^2\right)\right) + \sigma_\epsilon z_\epsilon \\ &= f(\mathbf{x}) + \sigma z_1 + \alpha d R^{\alpha-1}\sigma z_A - \frac{\alpha d}{2}R^{\alpha-2}\sigma^2\|\mathbf{z}_{2\ldots N}\|^2 + \sigma_\epsilon z_\epsilon. \end{aligned} \tag{7}$$

Again using the assumption Eq. (6), for large N the variance of the term involving $\|\mathbf{z}_{2\ldots N}\|^2$ disappears relative to that of the term involving z_A, and it is possible to treat the former as a constant by replacing it with its expected value. The variable terms (those involving z_1, z_A, and z_ϵ) are all normally distributed due to the way that mutation vectors are generated and the assumption that noise is Gaussian. Selection ensures that those μ candidate solutions with the largest values of $\sigma z_1 + \alpha d R^{\alpha-1}\sigma z_A + \sigma_\epsilon z_\epsilon$ survive. The signed lengths z_1 and z_A of the axial and central components of the mutation vectors are thus concomitants of the order statistics that result from ranking offspring candidate solutions according to their noisy fitness. (See (David and Nagaraja, 1998) for an introduction to concomitants of order statistics.) According to Eq. (1), the axial, central, and lateral components $\mathbf{z}_1^{(\text{avg})}$, $\mathbf{z}_A^{(\text{avg})}$, and $\mathbf{z}_B^{(\text{avg})}$ of the progress vector are the averages of the respective components of the selected mutation vectors. The following lemma that has previously been used in (Arnold and Beyer, 2006; Arnold, 2006) is thus immediately applicable:

Lemma 1 *Let $X_i = Y_i + Z_i$ for $i = 1, \ldots, \lambda$, where the Y_i are independently standard normally distributed and the Z_i are independently normally distributed with mean zero and with variance σ_Z^2. Ordering the sample members by nondecreasing values of the X variates, the expected value of the arithmetic mean of those μ of the Y_i with the largest associated values of X_i is*

$$\mathrm{E}\left[\frac{1}{\mu}\sum_{k=1}^{\mu} Y_{\lambda+1-k;\lambda}\right] = \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+\sigma_Z^2}}$$

where $Y_{j;\lambda}$ denotes the concomitant of the jth order statistic and where $c_{\mu/\mu,\lambda}$ is the $(\mu/\mu,\lambda)$-progress coefficient defined in (Beyer, 2001b).

Specifically, with $Y = z_1$ and $Z = \alpha d R^{\alpha-1} z_A + \vartheta z_\epsilon$, where $\vartheta = \sigma_\epsilon/\sigma$, it follows from the lemma that

$$\mathrm{E}\left[z_1^{(\mathrm{avg})}\right] \overset{N\to\infty}{=} \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2+(\alpha d)^2 R^{2(\alpha-1)}}} \tag{8}$$

Similarly, with $Y = z_A$ and $Z = (z_1 + \vartheta z_\epsilon)/(\alpha d R^{\alpha-1})$, it follows that

$$\mathrm{E}\left[z_A^{(\mathrm{avg})}\right] \overset{N\to\infty}{=} \frac{\alpha d R^{\alpha-1} c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2+(\alpha d)^2 R^{2(\alpha-1)}}}. \tag{9}$$

Both equations are generalisations of the corresponding results for $\alpha = 2$ obtained in (Arnold and Beyer, 2006), and they generalise those in (Arnold, 2006) by including the effects of noise. Finally, as noted above, the influence of $\|\mathbf{z}_{2\ldots N}\|^2$ on the rank in the population of the corresponding offspring disappears as N increases. (The variance of the term involving $\|\mathbf{z}_{2\ldots N}\|^2$ in Eq. (7) disappears compared to that involving z_A.) For $N \to \infty$, $\mathbf{z}_B^{(\mathrm{avg})}$ is thus the average of μ uncorrelated random vectors. As seen in (Beyer, 2001b), averaging μ uncorrelated random vectors reduces the squared length of the vectors being averaged by a factor of $1/\mu$. Furthermore, the relative contribution of the central component $z_A^{(\mathrm{avg})}$ to $\|\mathbf{z}_{2\ldots N}^{(\mathrm{avg})}\|^2$ vanishes for large N. As a result,

$$\mathrm{E}\left[\frac{\|\mathbf{z}_{2\ldots N}^{(\mathrm{avg})}\|^2}{N}\right] \overset{N\to\infty}{=} \frac{1}{\mu}. \tag{10}$$

Altogether, Eqs. (8), (9), and (10) provide a description of the progress vector that is sufficient for obtaining a characterisation of the stationary state attained by an evolution strategy with stationary step length when tracking a ridge.

According to Eq. (2), the squared distance of the next time step's search point from the ridge axis is

$$\sum_{i=2}^{N}\left(x_i + \sigma z_i^{(\mathrm{avg})}\right)^2 = R^2 - 2R\sigma z_A^{(\mathrm{avg})} + \sigma^2 \|\mathbf{z}_{2\ldots N}^{(\mathrm{avg})}\|^2.$$

In order for stationarity to hold, the expected distance of the search point from the ridge axis must not change, yielding condition

$$2R\sigma \mathrm{E}\left[z_A^{(\mathrm{avg})}\right] = \sigma^2 \mathrm{E}\left[\|\mathbf{z}_{2\ldots N}^{(\mathrm{avg})}\|^2\right]. \tag{11}$$

Using Eqs. (9) and (10) and squaring both sides yields after some simple transformations condition

$$2\alpha d R^\alpha = \frac{N\sigma}{\mu c_{\mu/\mu,\lambda}}\sqrt{1+\vartheta^2+(\alpha d)^2 R^{2(\alpha-1)}}. \tag{12}$$

For given mutation strength, Eq. (12) can be used to determine the resulting average stationary distance of the search point from the ridge axis. For the sharp ridge with $\alpha = 1$, solving Eq. (12) yields

$$R = \frac{N\sigma}{\mu c_{\mu/\mu,\lambda}} \frac{\sqrt{1+\vartheta^2+d^2}}{2d}.$$

For the parabolic ridge with $\alpha = 2$ it follows

$$R = \sqrt{\frac{1}{8}\left(\frac{N\sigma}{\mu c_{\mu/\mu,\lambda}}\right)^2 + \sqrt{\frac{1}{64}\left(\frac{N\sigma}{\mu c_{\mu/\mu,\lambda}}\right)^4 + \left(\frac{N\sigma}{\mu c_{\mu/\mu,\lambda}}\right)^2 \frac{1+\vartheta^2}{16d^2}}}.$$

For other values of α, Eq. (12) needs to be solved numerically. After having obtained the distance R from the ridge axis, the progress rate can be computed as

$$\varphi \overset{N\to\infty}{=} \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2+(\alpha d)^2 R^{2(\alpha-1)}}} \tag{13}$$

as can be inferred from Eqs. (4) and (8).

Finally, it remains to establish that the assumption Eq. (6) that has been used in several places in the calculations holds. It follows from rearranging terms in Eq. (12) that

$$\begin{aligned}\frac{\sqrt{N}\sigma}{R} &\le \frac{2\mu c_{\mu/\mu,\lambda}}{\sqrt{N}} \frac{\alpha d R^{\alpha-1}}{\sqrt{1+\vartheta^2+(\alpha d)^2 R^{2(\alpha-1)}}} \\ &< \frac{2\mu c_{\mu/\mu,\lambda}}{\sqrt{N}}.\end{aligned}$$

For given population size parameters μ and λ, the limit case of high search space dimensionality implies $2\mu c_{\mu/\mu,\lambda} \ll \sqrt{N}$, and Eq. (6) thus holds. It can be observed in experiments that the results derived in this section are reasonably accurate even for moderate finite values of N.

2.4 Normalisations

For $\alpha > 1$, the equations thus derived can be simplified by writing them in terms of normalised quantities

$$\begin{aligned}\sigma^* &= \frac{\sigma N(\alpha d)^{1/(\alpha-1)}}{\mu c_{\mu/\mu,\lambda}} & \varphi^* &= \frac{\varphi N(\alpha d)^{1/(\alpha-1)}}{\mu c^2_{\mu/\mu,\lambda}} \\ \sigma^*_\epsilon &= \frac{\sigma_\epsilon N(\alpha d)^{1/(\alpha-1)}}{\mu c_{\mu/\mu,\lambda}} & \varrho &= R(\alpha d)^{1/(\alpha-1)}.\end{aligned} \tag{14}$$

In particular, stationarity condition Eq. (12) becomes

$$2\varrho^\alpha = \sigma^* \sqrt{1+\vartheta^2+\varrho^{2(\alpha-1)}} \tag{15}$$

where $\vartheta = \sigma_\epsilon/\sigma = \sigma^*_\epsilon/\sigma^*$. Solving for the normalised mutation strength yields

$$\sigma^* = \sqrt{\frac{4\varrho^{2\alpha} - \sigma^{*2}_\epsilon}{1+\varrho^{2(\alpha-1)}}}. \tag{16}$$

Similarly, Eq. (13) reads in terms of normalised quantities

$$\varphi^* \overset{N\to\infty}{=} \frac{\sigma^*}{\sqrt{1+\vartheta^2+\varrho^{2(\alpha-1)}}}. \tag{17}$$

Notice that as a result of using normalised quantities these equations are independent of parameters N, μ, λ, and d. It can be observed in experiments (not shown here) that larger values of μ and λ generally require larger values of N in order for the approximations to hold with the same accuracy, and that the accuracy of the approximations is independent of d. Also notice that the sharp ridge with $\alpha = 1$ is not described by Eqs. (15) and (17) unless $d = 1$ as the normalisations in Eq. (14) are not applicable. If results are to be derived for the sharp ridge with $d \neq 1$, Eqs. (12) and (13) must be used instead. For simplicity, in the remainder of this paper whenever the sharp ridge is considered $d = 1$ is assumed.

Figure 3 compares predictions from Eqs. (15) and (17) with measurements from runs of evolution strategies. Shown are results for $\alpha \in \{1, 2, 4\}$ and for three different noise strengths. The measurements have been made with the search point of the evolution strategy initialised to lie on the ridge axis. The simulations have been run for $40N$ time steps in order to reach the state where the distance from the ridge axis is stationary on average. Then, ϱ and φ^* have been averaged over a period of 40000 time steps. It can be seen from the figure that the quality of the predictions is quite good and that it improves with increasing N. The most severe discrepancies between theory and experiment can be observed for the progress rate in the case that $\alpha = 4$, but those, too, disappear with increasing N.

It can be seen from Fig. 3 that independent of the topology parameter α, the distance from the ridge axis generally increases with both increasing mutation and noise strengths. Little surprisingly, the progress rate of the evolution strategy decreases with increasing noise strength. However, as noted by (Beyer, 2001a), the form of dependency of the progress rate on the mutation strength qualitatively depends on the topology parameter α. For $\alpha = 1$ (and indeed for any $\alpha < 2$) the progress rate increases indefinitely as the mutation strength increases. For ridges with $\alpha > 2$, tracking the ridge at too great a distance is ineffective and the progress rate peaks at a finite mutation strength. The parabolic ridge with $\alpha = 2$ lies in between those two cases. The progress rate of the evolution strategy on parabolic ridges increases with increasing mutation strength, but it tends to a finite limit value rather than increasing indefinitely. Consequently, a mutation strength adaptation algorithm should ideally generate mutation strengths in the vicinity of the progress rate maximum for $\alpha > 2$, and it should increase the mutation strength indefinitely for $\alpha \leq 2$.

2.5 Optimal Performance

Interestingly, in the absence of noise, optimal settings of the mutation strength and the resulting progress rate can be determined analytically even though Eq. (15) can generally only be solved numerically. Using Eq. (16) to eliminate σ^* in Eq. (17) yields for $\sigma^*_\epsilon = 0$

$$\varphi^* = \frac{2\varrho^\alpha}{1 + \varrho^{2(\alpha-1)}}.$$

Computing the derivative

$$\frac{\mathrm{d}\varphi^*}{\mathrm{d}\varrho} = \frac{2\alpha\varrho^{\alpha-1}(1 + \varrho^{2(\alpha-1)}) - 4(\alpha-1)\varrho^{3(\alpha-1)}}{(1 + \varrho^{2(\alpha-1)})^2}$$

and demanding that it be zero yields condition

$$\alpha \left(1 + \varrho^{2(\alpha-1)}\right) = 2(\alpha-1)\varrho^{2(\alpha-1)}$$

that must hold for maximal progress. Solving for ϱ yields

$$\varrho = \sqrt{\frac{\alpha}{\alpha-2}}^{1/(\alpha-1)} \tag{18}$$

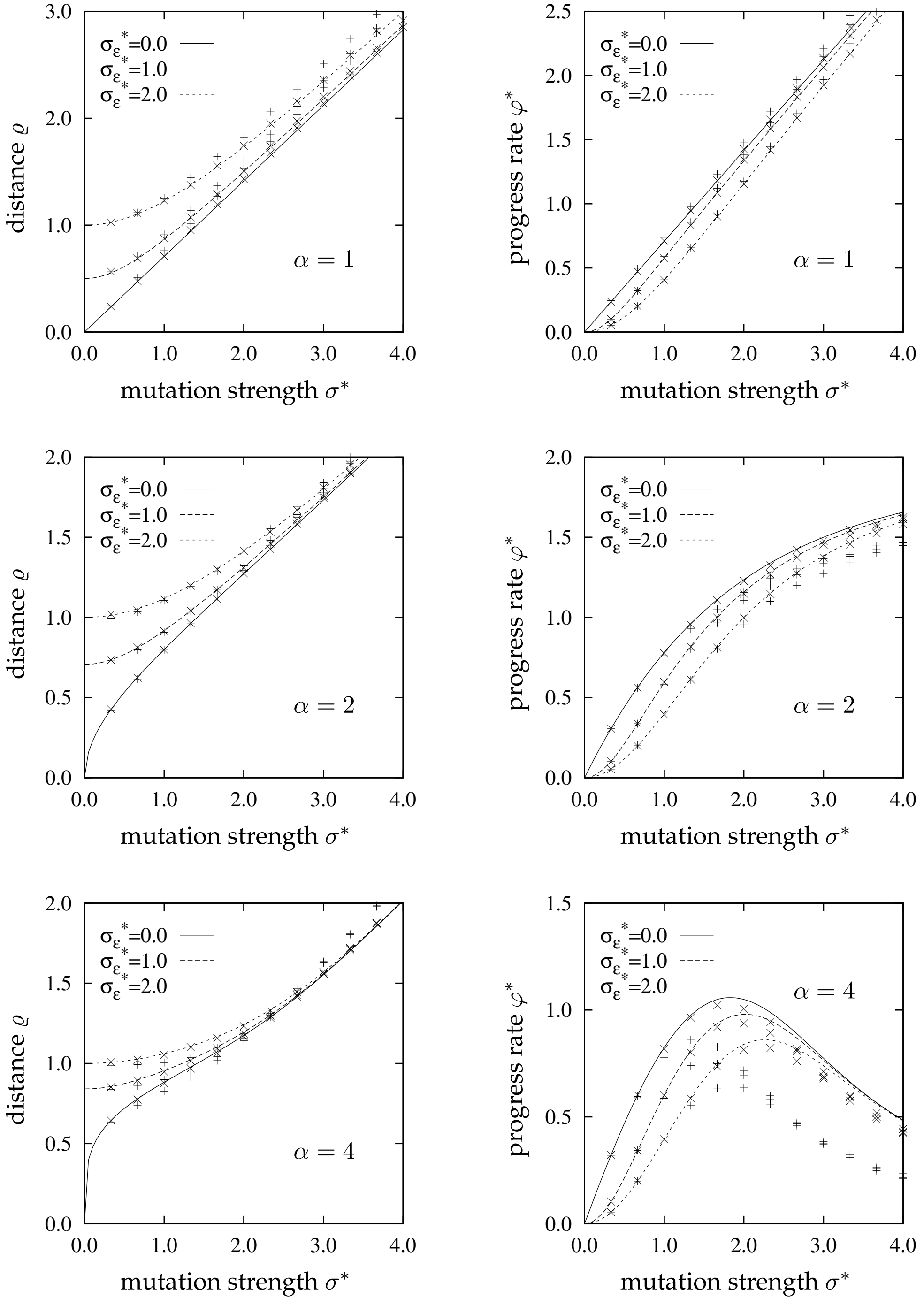

Figure 3: Normalised distance ϱ from the ridge axis and normalised progress rate φ^* plotted against normalised mutation strength σ^* for ridges with $\alpha \in \{1, 2, 4\}$ and $\sigma^*_\epsilon \in \{0.0, 1.0, 2.0\}$. The points mark measurements made in runs of the $(3/3, 10)$-ES for $N = 40$ (+) and $N = 400$ (×). The lines represent predictions obtained by numerically solving Eq. (15) for ϱ and using Eq. (17) to compute φ^*.

for the optimal normalised stationary distance from the ridge axis. Using Eqs. (15) and (17), the corresponding normalised mutation strength and progress rate are

$$\sigma^* = \sqrt{\frac{2}{\alpha-1}\left(\frac{\alpha^\alpha}{\alpha-2}\right)^{1/(\alpha-1)}} \tag{19}$$

and

$$\varphi^* = \frac{\sqrt{\alpha^\alpha(\alpha-2)^{\alpha-2}}^{1/(\alpha-1)}}{\alpha-1} \tag{20}$$

respectively. Eqs. (18), (19), and (20) confirm the observations made in Section 2.4. From Eq. (18), only for $\alpha > 2$ does a nonnegative solution exist for ϱ. For $\alpha \leq 2$ the optimal mutation strength is infinite, and the stationary distance of the search point from the ridge axis diverges. For $\alpha < 2$ the resulting progress rate increases indefinitely with increasing mutation strength as can be seen in Fig. 3 for the special case that $\alpha = 1$. Also seen in that figure, for $\alpha = 2$ a limit value of $\varphi^* = 2$ is approached as σ^* increases. For $\alpha > 2$, the optimal mutation strength is finite and results in a finite progress rate as witnessed by the curves for $\alpha = 4$ in Fig. 3. The same findings, albeit without analytical expressions for the optimal parameter settings, have been made by (Beyer, 2001a) for the special case of the $(1, \lambda)$-ES.

3 Cumulative Step Length Adaptation

Cumulative step length adaptation has been proposed by (Ostermeier et al., 1994) and is the standard step length adaptation method used in CMA-ES as described in (Hansen and Ostermeier, 2001). Its performance on the sphere model is analysed in (Arnold, 2002; Arnold and Beyer, 2004). The performance on the parabolic ridge is investigated for several forms of noise in (Arnold and Beyer, 2006) and on other ridges in the absence of noise in (Arnold, 2006). This section extends the work in those references by including the non-parabolic, noisy case.

3.1 Algorithm

Cumulative step length adaptation relies on the proposition that ideally, consecutive steps of the strategy should be uncorrelated. Positive correlations in the sequence of steps are taken to indicate that a larger step length should be used. Negative correlations suggest that the strategy steps back and forth, and that the step length should be reduced. In order to measure correlations, information about the most recently taken steps is accumulated in a search path $\mathbf{s} \in I\!R^N$ that is initialised to be zero and that is updated in every step according to

$$\mathbf{s} \leftarrow (1-c)\mathbf{s} + \sqrt{\mu c(2-c)}\mathbf{z}^{(\mathrm{avg})}. \tag{21}$$

The coefficient that the progress vector is multiplied with is chosen such that after initialisation effects have faded, the search path's components are standard normally distributed if selection is random (i.e., if the ranking of the candidate solutions in Section 2.1 is random rather than based on fitness). The cumulation parameter c determines how fast information about past steps fades. It must be chosen large enough in order not to hinder progress on functions such as the sphere model, and it must be chosen small enough in order to be able to reliably detect correlations. (Hansen, 1998) gives $c \in \Omega(1/N)$ and $c \in \mathcal{O}(1/\sqrt{N})$ as conditions. In this paper, $c = 1/\sqrt{N}$ is used.

In the presence of noise, better performance can be achieved by choosing c smaller; however, analytical results for $c \propto 1/N$ have not yet been obtained.

Finally, after the search path has been updated, the mutation strength is modified according to[1]

$$\sigma \leftarrow \sigma \exp\left(\frac{\|\mathbf{s}\|^2 - N}{2DN}\right). \tag{22}$$

The damping factor D is set to $1/c$. Recalling that the components of the search path are standard normally distributed if selection is random, the squared length of that vector is N if there are no correlations in the sequence of steps. Positive correlations increase the length of the search path compared to the random case, negative correlations decrease it. Eq. (22) thus acts to increase the step length in case of positive correlations, and it decreases the mutation strength if there are negative correlations.

3.2 Analysis

The analysis of the performance of cumulative step length relies on the same ideas used for the analysis of the performance of the strategy with static step length above: the behaviour of the algorithm is expressed by a set of difference equations, simplifications are made by assuming large N and by replacing quantities with their expected values, and a fixed point of the resulting simplified mapping is determined that serves as an approximation for the average values of the state variables. In (Arnold and Beyer, 2006)

$$z_1^{(\mathrm{avg})^2} + z_A^{(\mathrm{avg})^2} = \frac{\sigma}{R} z_A^{(\mathrm{avg})} \|\mathbf{z}_{2\ldots N}^{(\mathrm{avg})}\|^2 \tag{23}$$

is derived as a stationarity condition for the parabolic case. The derivation is lengthy and holds without any changes for the general case as well. Using the expected values from Eqs. (8), (9), and (10) to replace $z_1^{(\mathrm{avg})}$, $z_A^{(\mathrm{avg})}$, and $\|\mathbf{z}_{2\ldots N}^{(\mathrm{avg})}\|^2$ it follows after substituting normalised quantities for σ and R that

$$1 + \varrho^{2(\alpha-1)} = \varrho^{\alpha-2}\sqrt{\left(1 + \varrho^{2(\alpha-1)}\right)\sigma^{*2} + \sigma_\epsilon^{*2}}.$$

Using Eq. (16) to eliminate the normalised mutation strength and solving the resulting equation yields

$$\varrho = 1 \tag{24}$$

for the normalised distance from the ridge axis. Using this result in Eqs. (16) and (17) yields

$$\sigma^* = \sqrt{2 - \frac{\sigma_\epsilon^{*2}}{2}} \tag{25}$$

for the normalised mutation strength generated by cumulative step length adaptation and

$$\varphi^* = 1 - \frac{\sigma_\epsilon^{*2}}{4} \tag{26}$$

for the corresponding normalised progress rate.

Figure 4 compares predictions from Eqs. (24), (25) and (26) with measurements made in runs of evolution strategies. The experimental setup is the same as that used

[1] Eq. (22) differs from the corresponding prescription used by (Ostermeier et al., 1994; Hansen, 1998) in that there, the mutation strength is modified based on the length of the search path rather than based on its squared length. The difference between the two update rules is insignificant for large enough values of N. The prescription used here simplifies the theoretical analysis.

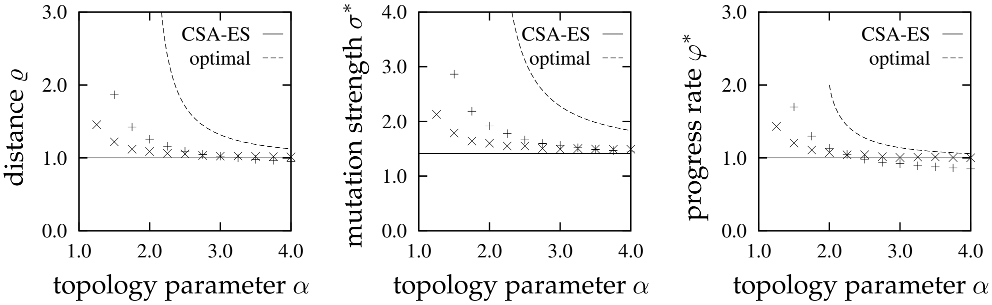

Figure 4: Normalised distance ϱ from the ridge axis, normalised mutation strength σ^*, and normalised progress rate φ^* of the $(\mu/\mu, \lambda)$-ES with cumulative step length adaptation plotted against the topology parameter α. The solid lines represent results from Eqs. (24), (25), and (26). The dashed lines show the optimal values from Eqs. (18), (19), and (20). The points mark measurements from runs of the strategy with $\mu = 3$ and $\lambda = 10$ in search spaces with $N = 40$ (+) and $N = 400$ (×).

to generate the data points in Fig. 3, except that now the mutation strength of the strategy is subject to cumulative step length adaptation. The noise strength is zero. It can be seen that while the behaviour of the strategy is quite well described by the analytical results for $\alpha > 2$, substantial deviations between analytically obtained curves and measurements exist for ridges near the sharp one unless N is very large. In order to understand the behaviour of cumulative step length adaptation on sharp ridges, Eqs. (11) and (23) with Eqs. (8), (9), and (10) can be used. Solving the equations shows that a fixed point of the mapping that describes the one-step behaviour of the strategy exists only for $d = 1$. Moreover, that fixed point is unstable. As a consequence, depending on the value of d, cumulative step length adaptation on sharp ridges either drives the mutation strength to zero or increases it indefinitely. In finite-dimensional search spaces, that instability can be observed not only for perfectly sharp ridges but for values of α in the vicinity of unity as well. A quantitative analysis of that behaviour would require taking N-dependent terms into account and it not attempted here.

Altogether, it can be seen from Fig. 4 that the mutation strengths generated by cumulative step length adaptation are generally smaller than optimal. With increasing values of α, the discrepancy between optimal values and values generated by cumulative step length adaptation decreases. As seen by (Arnold and Beyer, 2006), for the parabolic ridge and $N \to \infty$ cumulative step length adaptation achieves half of the optimal progress rate. For $\alpha > 2$ the fraction of the optimal progress rate that is achieved is higher. For $1 < \alpha \leq 2$ where infinite step lengths are optimal, the analytical results for $N \to \infty$ indicate that cumulative step length adaptation generates finite step lengths and thus achieves finite progress rates. However, for finite values of N the experimental measurements included in the figure show that the strategy performs better than predicted.

Figure 5 examines the performance of cumulative step length adaptation in the presence of noise. Included are results for the cases of $\alpha = 2$ and $\alpha = 4$. Data for the sharp ridge are not shown as for finite N the mutation strength of the strategy and with it the progress rate grow indefinitely. From Eq. (26), the progress rate of the strategy is positive up to a limit noise strength of

$$\sigma_\epsilon^* = 2. \tag{27}$$

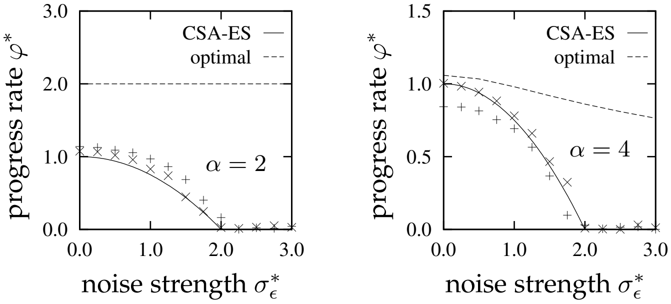

Figure 5: Normalised progress rate φ^* of the $(\mu/\mu, \lambda)$-ES with cumulative step length adaptation plotted against the normalised noise strength σ_ϵ^*. The solid lines represent results from Eq. (26). The dashed lines show values that would be achieved if the mutation strength were adapted optimally and that have been obtained numerically. The points mark measurements from runs of the strategy with $\mu = 3$ and $\lambda = 10$ in search spaces with $N = 40$ (+) and $N = 400$ (×).

Beyond that noise strength, cumulative step length adaptation drives the step length to zero even though positive progress could be achieved by increasing the mutation strength sufficiently. The figure shows that the behaviour of the strategy is predicted quite accurately.

4 Mutative Self-Adaptation

The roots of mutative self-adaptation go back to the early work of (Rechenberg, 1973). (Schwefel, 1981) developed the basic idea further and proposed to use it for adapting the entire mutation covariance matrix. Analyses of the behaviour of mutative self-adaptation, including proofs of convergence and the computation of the progress rate, have so far largely been restricted to the sphere model. See (Meyer-Nieberg and Beyer, 2006) for an overview. A notable exception is a recent first analysis of mutative self-adaptation on sharp ridge functions in (Beyer and Meyer-Nieberg, 2006). That analysis is complementary to the one presented here in that it represents a more accurate approach that remains to be extended to multi-parent strategies and to ridge functions beyond the sharp one.

4.1 Algorithm

The basic idea of mutative self-adaptation is to associate different step lengths with different individuals. When generating an offspring candidate solution, first an individual mutation strength is generated by applying recombination and mutation to the mutation strengths of its parents. Then, the newly generated mutation strength is used to generate the object component of the new candidate solution. The underlying proposition is that better adapted mutation strengths are more likely to generate successful offspring (i.e., offspring with high fitness) and therefore to survive selection.

Recombination of the parental mutation strengths can be either by arithmetic or by geometric averaging. In this paper, only geometric recombination is considered.[2]

[2] See (Hansen, 2006) for a discussion of the choice of recombination operator. The geometric choice is easier to handle analytically as like the mutation operator, it does not introduce a bias in the variation of the logarithm of the mutation strength. Arithmetic averaging generally yields larger values than geometric averaging. As the mutation strengths generated by mutative self-adaptation are smaller than optimal on

Mutation operators for the mutation strength are usually multiplicative rather than additive in that the mutation strength used to generate the object component of the ith offspring candidate solution is computed as $\sigma^{(i)} = \sigma \cdot \xi^{(i)}$, where σ is the result obtained from recombining the parental mutation strengths and $\xi^{(i)}$ is a positive random variate. According to (Rechenberg, 1994), common choices for the distribution of $\xi^{(i)}$ include:

log-normal: $\xi^{(i)} = \exp(\tau \mathcal{N}(0,1))$ where $\mathcal{N}(0,1)$ denotes a standard normally distributed random variate

two-point: $\xi^{(i)} = \beta > 1$ with probability 0.5 and $\xi^{(i)} = 1/\beta$ otherwise

deterministic two-point: $\xi^{(i)} = \beta > 1$ if $1 \leq i \leq \lambda/2$ and $\xi^{(i)} = 1/\beta$ otherwise

Parameters τ (for log-normal) and β (for two-point and deterministic two-point) control the rate at which the mutation strength is varied and are constant throughout a run of the strategy. As shown by (Beyer, 2001b) in the context of the sphere model, the log-normal and two-point operators can be made to behave very similarly if the parameters τ and β are chosen appropriately. In this paper, only the deterministic two-point operator is considered.

The purpose of mutation is to introduce variation that results in meaningful information for selection. Generally, larger values of τ or β afford more information in that the fitness values of the resulting candidate solutions tend to vary more widely. However, choosing τ or β too large leads to fluctuations in the mutation strength that are detrimental to the performance of the strategy. In order to reap the benefits of large variation without suffering from excessive fluctuations, (Rechenberg, 1994; Beyer, 1998) propose to dampen the update of the mutation strength using the update rule

$$\sigma \leftarrow \sigma \cdot \left(\prod_{k=1}^{\mu} \xi^{(k;\lambda)} \right)^{1/(\mu\kappa)} \tag{28}$$

that incorporates the effects of both (geometric) recombination and selection. As in Section 2.1, index $k;\lambda$ refers to the offspring candidate solution with the kth best fitness. The update rule Eq. (28) serves to compute a single value, referred to as the population's mutation strength, from which the individual mutation strengths are computed (using log-normal, two-point, or deterministic two-point mutations). For $\kappa = 1$, the population's mutation strength is the geometric mean of the mutation strengths of the μ candidate solutions that form the population. Using $\kappa > 1$ introduces damping and allows control over the speed with which the population's mutation strength is adapted. It is important for κ not to grow faster than linearly in N as otherwise linear convergence on functions such as the sphere model would not be possible.

4.2 Analysis

The task of characterising the stationary state of strategies that employ mutative self-adaptation is complicated by the fact that there is randomness on two levels. Random variation on the object parameter level is influenced by the outcome of random variation on the strategy parameter level. Selection on the strategy parameter level is indirect in that it is those mutation strengths that have the best associated object components that are selected. Nonetheless, for the simple case of geometric recombination

ridge functions, arithmetic averaging yields better performance. However, it can be observed in experiments that the gain is relatively minor. An analytical investigation of arithmetic recombination of the mutation strength on ridge functions remain as a task for future work.

of mutation strengths in connection with the deterministic two-point operator for their mutation, approximate results can be obtained quite easily on ridge functions. The approach presented here is similar in spirit to that proposed by (Lunacek and Whitley, 2006). It goes beyond that reference in that multi-parent strategies are considered and analytical results are derived.

In the absence of selection, the expectation of the logarithm of the population's mutation strength is stationary due to the definitions of the recombination and mutation operators that act on mutation strengths. With selection, assuming that λ is even and using the deterministic two-point operator for generating mutation strengths, the (logarithm of the) population's mutation strength is unchanged if half of the offspring that are selected to survive have a mutation strength of $\sigma \cdot \beta$ and the other half have mutation strength σ/β. Consequently, (Lunacek and Whitley, 2006) propose to solve for the value of σ for which the probability of an offspring candidate solution with mutation strength $\sigma \cdot \beta$ to be selected is 0.5. Computing that probability and solving for the stationary mutation strength is a difficult task that remains to be tackled in future work. Instead, as a much easier to obtain approximation, we assume that μ is odd and compute the value of σ such that the expected value of the (noisy) fitness of the $(\mu+1)/2$th best of those $\lambda/2$ offspring generated with mutation strength $\sigma \cdot \beta$ is equal to the expected value of the (noisy) fitness of the $(\mu+1)/2$th best of those $\lambda/2$ offspring generated with mutation strength σ/β. As a result, the probability that the μth best of the entire λ offspring candidate solutions generated is the $(\mu+1)/2$th best of those with mutation strength $\sigma \cdot \beta$ is roughly equal to the probability that it is the $(\mu+1)/2$th best of the offspring generated with mutation strength σ/β. Either case is as close as possible to a balance between the two competing mutation strengths. Though crude, the approximation will be seen to yield surprisingly accurate results that properly reflect qualitative properties of the strategies under consideration.

According to Eq. (7) and the discussion in Section 2.3, the expected noisy fitness value of the mth best of l offspring candidate solutions generated with mutation strength $\sigma \cdot \beta$ is

$$E_{m,l}(\beta) \overset{N\to\infty}{=} f(\mathbf{x}) + \frac{\mu c_{\mu/\mu,\lambda}}{N(\alpha d)^{1/(\alpha-1)}} \cdot \left(\sqrt{\sigma^{*2}\left(1+\varrho^{2(\alpha-1)}\right)\beta^2 + \sigma_\epsilon^{*2}}\, e_{m,l} - \frac{\mu c_{\mu/\mu,\lambda}}{2} \varrho^{\alpha-2}\sigma^{*2}\beta^2 \right) \qquad (29)$$

where

$$e_{m,l} = \frac{1}{\sqrt{2\pi}} \frac{\Gamma(l+1)}{\Gamma(l-m+1)\Gamma(m)} \int_{-\infty}^{\infty} x \mathrm{e}^{-x^2/2} \left[\Phi(x)\right]^{l-m} \left[1-\Phi(x)\right]^{m-1} \mathrm{d}x$$

denotes the expectation of the $(l-m+1)$th order statistic of a sample of l independent standard normally distributed random variables and $\Phi(\cdot)$ is the cdf of the standardised normal distribution (see (David and Nagaraja, 1998)). Demanding stationarity as described above amounts to requiring that

$$E_{(\mu+1)/2,\lambda/2}(\beta) = E_{(\mu+1)/2,\lambda/2}(1/\beta).$$

Assuming that β is sufficiently close to unity in order to linearise, this condition translates into

$$\left.\frac{\mathrm{d}E_{(\mu+1)/2,\lambda/2}}{\mathrm{d}\beta}\right|_{\beta=1} = 0. \qquad (30)$$

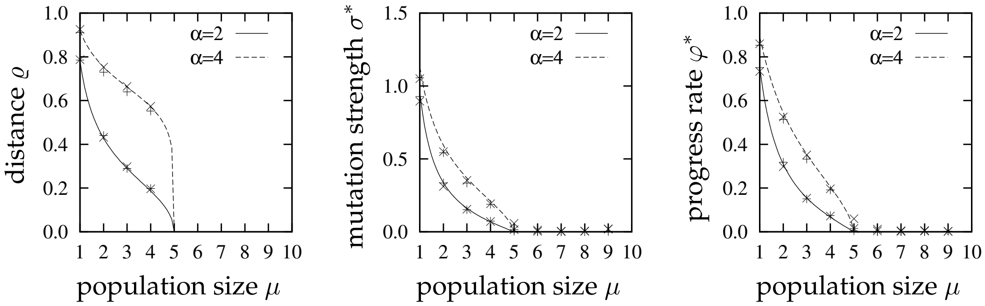

Figure 6: Normalised distance ϱ from the ridge axis, normalised mutation strength σ^*, and normalised progress rate φ^* of the $(\mu/\mu, \lambda)$-ES with mutative self-adaptation and $\lambda = 10$ plotted against population size μ. The solid and dashed lines represent results from (31) with Eqs. (16) and Eqs. (17) for $\alpha = 2$ and $\alpha = 4$. The points mark measurements from runs of the strategy in search spaces with $N = 40$ (+) and $N = 400$ (×).

Using Eq. (29) to compute the derivative and again assuming that β is close to unity in order to be able to use Eq. (16) to eliminate the normalised mutation strength yields after some simple transformations

$$\left(1 + \varrho^{2(\alpha-1)}\right) e_{(\mu+1)/2,\lambda/2} = 2\mu c_{\mu/\mu,\lambda} \varrho^{2(\alpha-1)}.$$

Finally, solving for the normalised distance from the ridge axis results in

$$\varrho = \sqrt{\frac{e_{(\mu+1)/2,\lambda/2}}{2\mu c_{\mu/\mu,\lambda} - e_{(\mu+1)/2,\lambda/2}}}^{\,1/(\alpha-1)}. \tag{31}$$

The corresponding mutation strength and resulting progress rate can be obtained from Eqs. (16) and (17).

It can be seen from Eq. (31) that in contrast to the strategy that uses cumulative step length adaptation, the distance from the ridge axis (and thus the normalised mutation strength and progress rate) of the strategy that employs mutative self-adaptation is not independent of the population size parameters μ and λ. Figure 6 illustrates how the performance of the strategy depends on the population size μ for $\lambda = 10$ and $\alpha \in \{2, 4\}$. Notice that normalised mutation strengths and progress rates for different values of μ are not immediately comparable as the population size parameters enter the normalisation described by Eq. (14). The experimental values have been obtained with $\beta = 1.3$ and $\kappa = N/4$. However, the data points are largely insensitive to the setting of those parameters as long as they are chosen in accordance with the criteria outlined in the discussion above. Not included in the figure is the case of the sharp ridge for which it will be seen below that mutative self-adaptation fails in that it drives the mutation strength to zero, resulting in no progress being made. A number of observations can be made from the figure. First, the accuracy of the analytically obtained predictions is quite good despite the simplifying assumptions made in their derivation. In particular, while Eq. (31) was derived for odd integer μ, the accuracy for even μ is as good as that for μ odd. Second, it can be seen that mutative self-adaptation fails to generate useful mutation strengths for $\mu \geq \lambda/2$. The same result was found by (Hansen, 2006) on linear fitness functions. If the selection pressure is too low, mutative self-adaptation

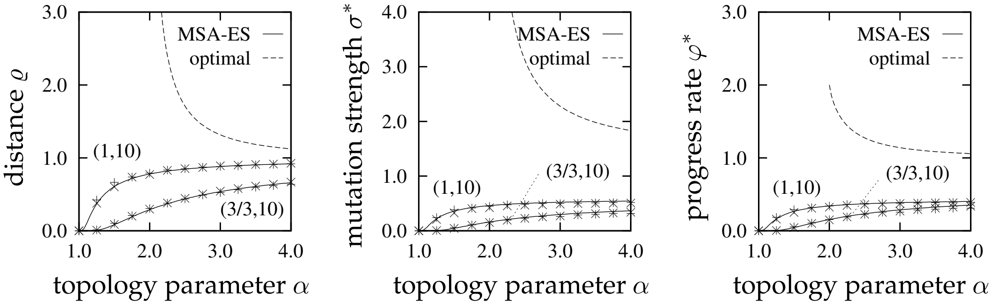

Figure 7: Normalised distance ϱ from the ridge axis, normalised mutation strength σ^*, and normalised progress rate φ^* of the $(\mu/\mu, \lambda)$-ES with mutative self-adaptation plotted against the topology parameter α. The solid lines represent results from Eq. (31) with Eqs. (16) and (17) for $(1, 10)$-ES and $(3/3, 10)$-ES. See the text for a discussion of the normalisations. The dashed lines show the optimal values from Eqs. (18), (19), and (20). The points mark measurements from runs of the strategies in search spaces with $N = 40$ (+) and $N = 400$ (×).

systematically drives the step length to zero, resulting in stagnation. And third, it can be confirmed from Eq. (17) with Eq. (31) that the highest progress rates are achieved with $\mu = 1$. Thus, when using geometric recombination in combination with mutative self-adaptation on ridge functions, point based strategies are superior to population based ones. As has also been pointed out by (Hansen, 2006), mutative self-adaptation fails to make use of the opportunity of using larger mutation strengths afforded by global intermediate recombination.

Figure 7 illustrates the dependence of the distance from the ridge axis, the mutation strength, and the progress rate of the $(\mu/\mu, \lambda)$-ES with mutative self-adaptation on the topology parameter α. Shown are results for the $(3/3, 10)$-ES both from Eq. (31) with Eqs. (16) and (17) and from computer experiments as well as the corresponding optimal values from Eqs. (18), (19), and (20). Also included in the figure are results for the $(1, 10)$-ES as it is superior to the $(3/3, 10)$-ES if mutative self-adaptation is used. In order to make it possible to immediately compare the performances of the two strategies, the mutation strength and progress rate of the $(1, 10)$-ES have been normalised using the same factors as those used for the $(3/3, 10)$-ES. It can be seen from the figure that as for cumulative step length adaptation, the mutation strengths generated using mutative self-adaptation are generally below their optimal values and that non-optimal progress rates result. The performance of mutative self-adaptation is especially inadequate for ridges near the sharp one where both the mutation strength and the progress rate are driven to zero. This deficiency of mutative self-adaptation has been observed experimentally by (Herdy, 1992) and has led to the development of the hierarchically organised strategies discussed in Section 6. A more complete discussion of the performance of the $(1, \lambda)$-ES on sharp ridges (and in particular of its dependence on the ridge parameter d) can be found in (Beyer and Meyer-Nieberg, 2006).

Finally, while the results derived above predict that the stationary mutation strength generated by mutative self-adaptation is unaffected by the presence of noise, it can be observed in experiments that in practice, noise leads to even smaller mutation strengths than those observed in the noise free case. The attractor described by Eq. (30) becomes increasingly unstable, and for small mutation strengths the pressure toward larger steps lengths is weak. As a consequence, the mutation strength performs a random walk at small values for much of the time, resulting in long periods of near stagnation. The present approach does not consider fluctuations and is unsuitable for capturing this type of behaviour.

5 Two-Point Adaptation

Two-point adaptation (not to be confused with the two-point mutation operator from Section 4) as described below is a simple variation of the standard step length adaptation operator used in evolutionary gradient search as introduced in (Salomon, 1998). It explicitly compares two steps with differing lengths in the same direction, and it settles for the better of the two. Two-point adaptation is thus reminiscent of a rudimentary line search.

5.1 Algorithm

Two-point adaptation performs steps 1 through 4 as described in Section 2.1. In addition to making a step of length σ, two steps of lengths $\sigma \cdot \beta$ and σ/β are tried, and σ is updated according to

$$\sigma \leftarrow \begin{cases} \sigma \cdot \beta^{1/\kappa} & \text{if } f(\mathbf{x} + (\sigma \cdot \beta)\mathbf{z}^{(\text{avg})}) > f(\mathbf{x} + (\sigma/\beta)\mathbf{z}^{(\text{avg})}) \\ \sigma/\beta^{1/\kappa} & \text{otherwise} \end{cases} \tag{32}$$

That is, the mutation strength is increased if the larger of the two steps is more successful; it is decreased if the smaller step yields the better fitness value. Notice that two-point adaptation requires two additional fitness function evaluations per time step, bringing the total to $\lambda + 2$. Parameter $\beta > 1$ must be chosen sufficiently different from unity to ensure that the difference between the two fitness values is significant. Too large a β is detrimental to the performance of the strategy. Parameter κ is as a damping parameter and helps avoid fluctuations of the mutation strength. As in mutative self-adaptation, κ should not be chosen to grow faster than linearly in N as otherwise linear convergence on the sphere model would not be possible. The procedure outlined here differs from that described by (Salomon, 1998) by virtue of the damping and by considering only two points rather than three.

5.2 Analysis

As in the previous sections on cumulative step length adaptation and mutative self-adaptation, we strive to determine the value of σ for which the mutation strength update rule of two-point adaptation affects no change in the mean. Letting $E(\beta)$ denote the expected value of the (noisy) fitness of a candidate solution generated by making a step of length $\sigma \cdot \beta$ in the direction of the vector $\mathbf{z}^{(\text{avg})}$, it follows from Eq. (7) with Eqs. (8), (9), and (10) and the normalisations from Eq. (14) that

$$E(\beta) = f(\mathbf{x}) + \frac{\mu c_{\mu/\mu,\lambda}}{N(\alpha d)^{1/(\alpha-1)}} \left(\frac{\sigma^{*2}(1+\varrho^{2(\alpha-1)})}{\sqrt{\sigma^{*2}(1+\varrho^{2(\alpha-1)}) + \sigma_\epsilon^{*2}}} \beta - \frac{\varrho^{\alpha-2}\sigma^{*2}}{2}\beta^2 \right).$$

The mutation strength of the $(\mu/\mu, \lambda)$-ES employing two-point adaptation is stationary if $E(\beta) = E(1/\beta)$. Assuming that β is sufficiently close to unity in order to linearise, this condition translates into

$$\left.\frac{\mathrm{d}E}{\mathrm{d}\beta}\right|_{\beta=1} = 0. \tag{33}$$

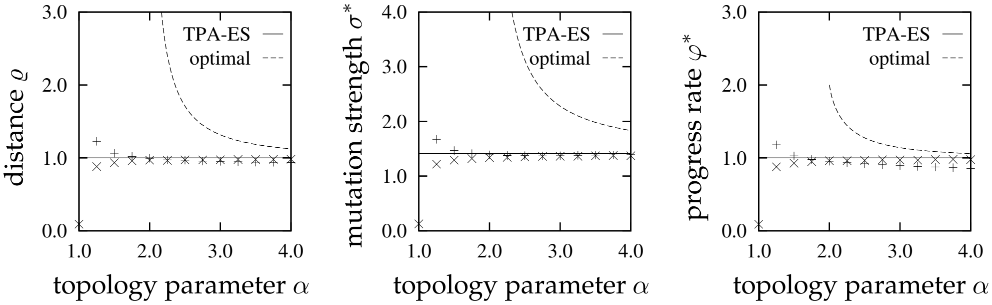

Figure 8: Normalised distance ϱ from the ridge axis, normalised mutation strength σ^*, and normalised progress rate φ^* of the $(\mu/\mu, \lambda)$-ES with two-point adaptation plotted against the topology parameter α. The solid lines represent results from Eq. (34) with Eqs. (16) and (17). The dashed lines show the optimal values from Eqs. (18), (19), and (20). The points mark measurements from runs of the strategy with $\mu = 3$ and $\lambda = 10$ in search spaces with $N = 40$ (+) and $N = 400$ (×).

Computing the derivative and rearranging terms yields

$$1 + \varrho^{2(\alpha-1)} = \varrho^{\alpha-2}\sqrt{\sigma^{*2}(1 + \varrho^{2(\alpha-1)}) + \sigma_\epsilon^{*2}}.$$

Finally, using Eq. (16) to eliminate the normalised mutation strength results in

$$\varrho = 1 \tag{34}$$

for the normalised distance from the ridge axis. The corresponding mutation strength and resulting progress rate can be obtained from Eqs. (16) and (17). Interestingly, a comparison with Eq. (24) shows that in the limit of infinite search space dimensionality, two-point adaptation generates the same behaviour as cumulative step length adaptation.

Figure 8 compares predictions from Eqs. (34), (16), and (17) with measurements from runs of the $(\mu/\mu, \lambda)$-ES with two-point adaptation. Parameter settings $\beta = 1.3$ and $\kappa = N/4$ have been used in the experiments, but as for mutative self-adaptation, the influence of the choice of those parameters is minor as long as the settings are reasonable. It can be seen from the figure that the accuracy of the predictions is quite good except for the smallest values of α. The mutation strength and progress rate generated by two-point adaptation are consistently below optimal values, with optimal behaviour being approached as the topology parameter α increases. Comparing Fig. 8 with Fig. 4 suggests that, at least for the population size parameter values considered, cumulative step length adaptation performs somewhat better in finite-dimensional search spaces that two-point adaptation does. For the sharp ridge, as for mutative self-adaptation, the fixed point described by Eq. (33) turns unstable and the present approach is unsuitable for describing the strategy's behaviour.

Finally, Fig. 9 examines the performance of two-point adaptation in the presence of noise. Shown are results for $\alpha = 2$ and $\alpha = 4$. As for cumulative step length adaptation the distance from the ridge axis is not influenced by the presence of noise and Eq. (34) holds independently of σ_ϵ^*. As a result, the normalised mutation strength and the normalised progress rate of the strategy that employs two-point adaptation are described by Eqs. (25) and (26). The figure shows that the behaviour of the strategy is predicted quite accurately.

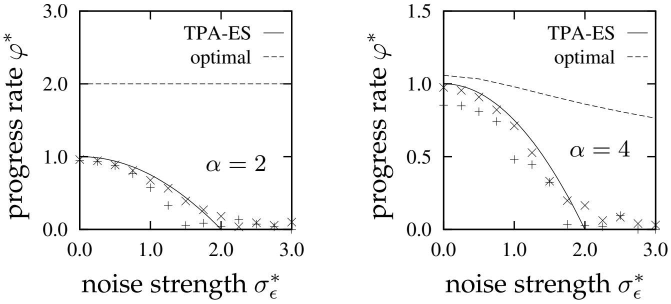

Figure 9: Normalised progress rate φ^* of the $(\mu/\mu, \lambda)$-ES with two-point adaptation plotted against the normalised noise strength σ_ϵ^*. The solid lines represent results from Eq. (34) with Eqs. (16) and (17). The dashed lines show values that would be achieved if the mutation strength were adapted optimally and that have been obtained numerically. The points mark measurements from runs of the strategy with $\mu = 3$ and $\lambda = 10$ in search spaces with $N = 40$ (+) and $N = 400$ (×).

6 Hierarchically Organised Strategies

The idea of organising evolution strategies hierarchically was born out of the insight that strategy parameter adaptation really is an optimisation problem, and that thus evolutionary algorithms can be applied to solve it. See (Rechenberg, 1978; Herdy, 1992; Rechenberg, 1994) for a motivation. Several populations (sometimes referred to as species) with differing strategy parameter settings evolve in isolation of each other. After some time, the amount of progress that has been made by the various populations is compared. The strategy parameter settings of the most successful populations are subjected to variation, and a new set of species is set up and run with those new strategy parameter settings. Thus, evolutionary optimisation happens on two levels: the search space of the lower level strategy is that of the optimisation problem at hand; that of the upper level strategy is the strategy parameter space of the lower level strategies.

(Herdy, 1992) empirically compares the performance of hierarchically organised strategies with that of strategies using mutative self-adaptation. Several objective functions are considered, including the sphere model as well as sharp and parabolic ridges. It is found that isolation can be detrimental to the performance of the strategies on the sphere model where fast adaptation is required. The situation is different on ridges where mutative self-adaptation often performs unsatisfactorily. Introducing isolation decreases the likelihood of opportunistic individuals that make short steps being rewarded. (Arnold and MacLeod, 2006) analytically study the performance of hierarchically organised strategies on parabolic ridges and derive expressions that describe the dependence of the strategies' performance on the length of the isolation periods. The derivation presented here generalises those results by considering ridge topologies other than parabolic ones, and by including the effects of noise in the calculations.

6.1 Algorithm

The lower level strategy employed in this paper is the $(\mu/\mu, \lambda)$-ES described in Section 2.1. The upper level strategy adapts the step length parameter σ of the lower level strategy by proceeding as follows:

1. The search point $\mathbf{x}$ and the mutation strength σ are initialised.

2. Parameter β is set to a value uniformly drawn from the interval $[1.2, 1.4]$.

3. Two runs of the lower level strategy are conducted in parallel. The runs last for γ generations each and both use $\mathbf{x}$ as their initial search point. One run uses mutation strength $\sigma \cdot \beta$, the other one uses σ/β.

4. The objective function values of the final search points generated in the two runs are compared. The search point $\mathbf{x}$ of the upper level strategy is set to the better of those two points; mutation strength σ is set to the mutation strength used in the more successful of the two runs.

5. The process is terminated if a prescribed number of steps has been made or otherwise continues with step 2.

Similar to mutative self-adaptation and two-point adaptation, the purpose of step 2 is to generate two mutation strengths, one larger than the previous one and one smaller. The two mutation strengths need to be sufficiently different to yield a reliable signal for selection; they should not be too different as if σ is nearly optimal, then neither $\sigma \cdot \beta$ nor σ/β are if β is too large. We have randomised the choice of β rather than simply using $\beta = 1.3$ as that choice would result in the strategy being confined to a discrete set of mutation strengths that would lead to artifacts in the performance graphs below. Notice that in contrast to the strategies studied above, the hierarchically organised strategy does not use damping as the use of isolation in combination with damping would make it impossible to adapt the mutation strength sufficiently fast on functions such as the sphere model.

In the notation introduced in (Herdy, 1992; Rechenberg, 1994), the overall strategy thus described is a $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES. It is thus an instance of the general $[\mu'/\rho' \overset{+}{,} \lambda'(\mu/\rho \overset{+}{,} \lambda)^\gamma]$-ES where the population size parameters of the upper level strategy are $\mu' = \rho' = 1$, and $\lambda' = 2$ and where comma selection is used on both levels. The choice of population size parameter settings has been made as it is sufficient for the one-dimensional optimisation problem that the upper level strategy faces. Larger population sizes (i.e., a greater number of lower level strategies to be run) would be unnecessarily computationally expensive unless adequate parallel resources are available.

Notice that mutative self-adaptation as discussed in Section 4 can be interpreted as a special (trivial) case of hierarchically organised evolution strategies where each species consists of a single individual, and where isolation periods last for a single generation. Also notice that adaptation by means of hierarchically organised strategies is not limited to step lengths but can be applied to other strategy parameters as well. (Herdy, 1993) considers the problem of adapting the number of offspring generated per time step and demonstrates empirically that near optimal values can be obtained on the hyperplane and sphere models.

Finally, realising that isolation periods of different lengths are optimal in different environments, (Herdy, 1992) proposes adding yet another level to the hierarchy of evolutionary strategies, with the goal of optimising the length of the isolation periods. He shows empirically that in the long term (i.e., after many time steps), the strategy that adapts the length of its isolation periods performs well on the sphere model as well as on ridges. Of course, the process of adding higher levels with the goal of optimising parameters of the strategy one level below could be continued indefinitely. Practically, limitations on the number of objective function evaluations that can be performed before a result is expected typically lead to flat hierarchies being used. Throughout this paper, only two-level hierarchies are considered.

6.2 Analysis

Central to the analysis of the performance of hierarchically organised evolution strategies is the need to characterise the cumulative effect of running the lower level strategies for the duration of an isolation period. More specifically, given a search point $\mathbf{x}$ that has been arrived at with a mutation strength of σ, the objective function value of the search point $\mathbf{x}'$ obtained after running the lower level strategy with a mutation strength of ς (which is either $\sigma \cdot \beta$ or σ/β) for a further γ time steps needs to be estimated. The respective values of $f(\mathbf{x}')$ for the different populations that evolve in parallel determine the mutation strength used in the next iteration of the upper level strategy. It is particularly easy to obtain such an estimate if the following two assumptions are made.

1. At the end of an isolation period, the lower level strategy is in the stationary limit state described by Eq. (15). Moreover, that limit state is reached so early in the isolation period that it can be assumed that all of the progress in the direction of the ridge axis made during the isolation period is made in that limit state.

2. For the purpose of comparing fitness values of population centroids, it is sufficient to consider their expected values; i.e., fluctuations can be ignored.

The first assumption requires that the length γ of the isolation periods be sufficiently large, where what is sufficient depends on the mutation strengths σ and ς as well as on the population size parameters μ and λ and the search space dimensionality N. The more ς differs from σ, the larger γ needs to be in order for the assumption to hold with a certain accuracy. As for the second assumption, it is generally valid if ς is sufficiently different from σ. While again, quantifying what is sufficient is a difficult task and depends on, among other things, the search space dimensionality, it will be seen that for the choice of β described above the qualitative agreement of results derived under the assumption with experimental measurements is good unless γ is too small.

Assuming that γ is sufficiently large, the population centroid is at a distance $R(\sigma)$ from the ridge axis at the beginning and at a distance $R(\varsigma)$ at the end of an isolation period. From Eq. (3), the respective objective function values are $f(\mathbf{x}) = x_1 - dR^\alpha(\sigma)$ and $f(\mathbf{x}') = x_1' - dR^\alpha(\varsigma)$. The expected difference between the objective function values of population centroids $\mathbf{x}$ and $\mathbf{x}'$ is thus

$$\begin{aligned}\Delta f &= \mathrm{E}\left[f(\mathbf{x}') - f(\mathbf{x})\right] \\ &= \gamma\varphi(\varsigma) - d\left(R^\alpha(\varsigma) - R^\alpha(\sigma)\right).\end{aligned}$$

The first of the two terms on the right hand side is due to progress in the direction of the ridge axis and is computed as the product of the expected progress per time step and the number of steps made. (Recall that progress is assumed to be made in the limit state assumed at the end of the isolation period.) The second term on the right hand side is due to the change in distance from the ridge axis that results from the altered mutation strength. Using the normalisations from Eq. (14) and introducing the normalised length

$$\gamma^* = \frac{\gamma\mu c_{\mu/\mu,\lambda}^2}{N} \tag{35}$$

of the isolation periods, it thus follows

$$\Delta f(\gamma^*, \sigma^*, \varsigma^*) = \frac{1}{(\alpha^\alpha d)^{1/(\alpha-1)}}\left(\alpha\gamma^*\varphi^*(\varsigma^*) - \varrho^\alpha(\varsigma^*) + \varrho^\alpha(\sigma^*)\right) \tag{36}$$

for the expected difference between the objective function values of $\mathbf{x}$ and $\mathbf{x}'$.

The $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES described above evolves two populations in parallel, one with mutation strength $\sigma \cdot \beta$ and one with mutation strength σ/β. After γ generations, the objective function values of the centroids $\mathbf{x}_1'$ and $\mathbf{x}_2'$ of the two populations are compared. The population with the larger objective function value of its centroid passes on its mutation strength to the next iteration of the upper level strategy. Letting

$$g(\beta) = (\alpha^\alpha d)^{1/(\alpha-1)} \left(f(\mathbf{x}_1') - f(\mathbf{x}_2') \right)$$

it is clear that the mutation strength used in the next iteration of the upper level strategy is $\sigma \cdot \beta$ (the mutation strength that led to $\mathbf{x}_1'$) if $g(\beta) \geq 0$ and σ/β (the mutation strength that led to $\mathbf{x}_2'$) otherwise. Function $g(\beta)$ is referred to as the gain difference. With Eq. (36) it follows that

$$\begin{aligned} g(\beta) &= \Delta f(\gamma^*, \sigma^*, \sigma^* \cdot \beta) - \Delta f(\gamma^*, \sigma^*, \sigma^*/\beta) \\ &= \alpha\gamma^*\varphi^*(\sigma^* \cdot \beta) - \alpha\gamma^*\varphi^*(\sigma^*/\beta) - \varrho^\alpha(\sigma^* \cdot \beta) + \varrho^\alpha(\sigma^*/\beta) \\ &= \frac{\alpha\gamma^*\sigma^{*2} \cdot \beta^2}{2\varrho^\alpha(\sigma^* \cdot \beta)} - \frac{\alpha\gamma^*\sigma^{*2}/\beta^2}{2\varrho^\alpha(\sigma^*/\beta)} - \varrho^\alpha(\sigma^* \cdot \beta) + \varrho^\alpha(\sigma^*/\beta) \end{aligned} \qquad (37)$$

where Eqs. (15) and (17) have been used in the last step.

As done above for mutative self-adaptation and two-point adaptation, we make the assumption that β is sufficiently close to unity in order to linearise. It is clear from Eq. (37) that $g(1) = 0$ independently of γ^* and σ^*. For sufficiently small values of β, the sign of $g(\beta)$ in the vicinity of unity is determined by the derivative $g'(1) = \mathrm{d}g/\mathrm{d}\beta|_{\beta=1}$. The mutation strength used in the next iteration of the upper level strategy is $\sigma \cdot \beta$ (i.e., it is increased) if $g'(1) > 0$ and it is σ/β (i.e., it is decreased) if $g'(1) < 0$. For $g'(1) = 0$, there is no strong pressure to either increase or decrease the mutation strength, and which one of $\sigma \cdot \beta$ and σ/β prevails is a matter of chance. Thus, the mutation strength for which $g'(1) = 0$ can be used as an approximation for the average mutation strength that the hierarchically organised strategy generates.

Computing the derivative of the gain difference from Eq. (37) results in

$$g'(1) = \frac{\alpha\gamma^*\sigma^{*2}}{\varrho^\alpha} \left(2 - \frac{\alpha\sigma^*}{\varrho} \varrho'(1) \right) - 2\alpha\sigma^*\varrho^{\alpha-1}\varrho'(1).$$

Demanding that $g'(1) = 0$ thus yields condition

$$2\gamma^*\sigma^*\varrho = \left(2\varrho^{2\alpha} + \alpha\gamma^*\sigma^{*2} \right) \varrho'(1). \qquad (38)$$

The derivative $\varrho'(1)$ of the normalised distance from the ridge axis can be obtained by squaring Eq. (15) and subsequently differentiating implicitly, resulting in

$$\varrho'(1) = \frac{\sigma^*(1 + \varrho^{2(\alpha-1)})}{4\alpha\varrho^{2\alpha-1} - (\alpha-1)\sigma^{*2}\varrho^{2\alpha-3}}.$$

Using this in Eq. (38) yields condition

$$2\gamma^* \left(4\alpha\varrho^{2\alpha} - (\alpha-1)\sigma^{*2}\varrho^{2(\alpha-1)} \right) = \left(1 + \varrho^{2(\alpha-1)} \right) \left(2\varrho^{2\alpha} + \alpha\gamma^*\sigma^{*2} \right).$$

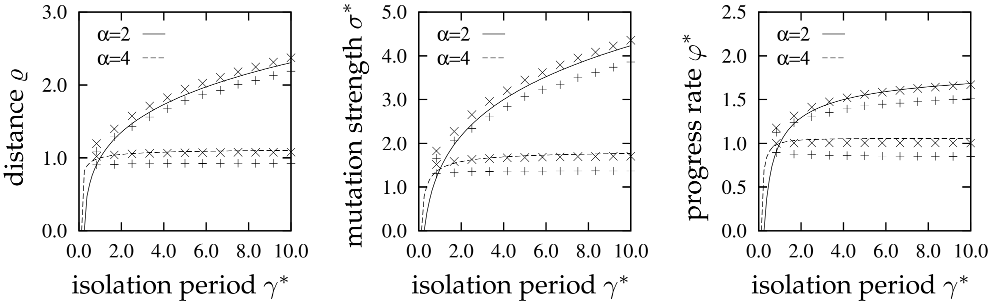

Figure 10: Normalised distance ϱ from the ridge axis, normalised mutation strength σ^*, and normalised progress rate φ^* of the hierarchically organised $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES plotted against the normalised length γ^* of the isolation periods. The solid and dashed lines represent results from Eq. (40) with Eqs. (16) and (17) for $\alpha \in \{2, 4\}$. The points mark measurements with $\mu = 3$ and $\lambda = 10$ in search spaces with $N = 40$ (+) and $N = 400$ (×).

Finally, using Eq. (16) in order to eliminate the normalised mutation strength yields after several simple transformations

$$\begin{aligned}\left(\alpha + (3\alpha - 2)\varrho^{2(\alpha-1)}\right)\gamma^*\sigma_\epsilon^{*2} = \\ 2\varrho^{2\alpha}\left(1 - 2\alpha\gamma^* + 2(1 + (\alpha-2)\gamma^*)\varrho^{2(\alpha-1)} + \varrho^{4(\alpha-1)}\right) \quad (39)\end{aligned}$$

as a condition that determines the normalised distance from the ridge axis that the $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES tracks ridges at. The corresponding mutation strength and resulting progress rate can be obtained from Eqs. (16) and (17).

In the absence of noise, Eq. (39) can be used to obtain a closed form solution for the normalised distance from the ridge axis. For $\sigma_\epsilon^* = 0$, dividing by $\varrho^{2\alpha}$ results in an equation that is quadratic in $\varrho^{2(\alpha-1)}$. Solving it yields

$$\varrho^{2(\alpha-1)} = \sqrt{(\alpha-2)^2\gamma^{*2} + 4(\alpha-1)\gamma^*} - (1 + (\alpha-2)\gamma^*)\,. \quad (40)$$

Figure 10 illustrates the dependence of the distance from the ridge axis, the mutation strength, and the progress rate on the length of the isolation periods. While some deviations between predictions and experimental measurements exist, the overall behaviour of the strategies is captured quite well. For any value of γ^*, the $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES assumes a finite average mutation strength, and it consequently tracks ridges at a finite distance. That distance as well as the mutation strength and the progress rate increase monotonically with the length of the isolation periods. Moreover, limit values approached as γ^* increases indefinitely can be obtained by Taylor expanding the square root in Eq. (40), resulting in

$$\begin{aligned}\varrho^{2(\alpha-1)} &= (\alpha-2)\gamma^*\left(\sqrt{1 + \frac{4(\alpha-1)}{(\alpha-2)^2\gamma^*}} - 1 - \frac{1}{(\alpha-2)\gamma^*}\right) \\ &\overset{\gamma^*\to\infty}{=} (\alpha-2)\gamma^*\left(1 + \frac{2(\alpha-1)}{(\alpha-2)^2\gamma^*} - 1 - \frac{1}{(\alpha-2)\gamma^*}\right) \\ &= \frac{\alpha}{\alpha-2}.\end{aligned}$$

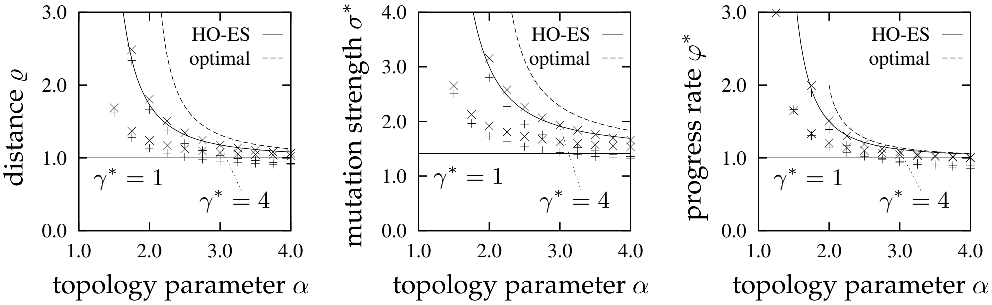

Figure 11: Normalised distance ϱ from the ridge axis, normalised mutation strength σ^*, and normalised progress rate φ^* of the hierarchically organised $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES plotted against the topology parameter α. The solid lines represent results for $\gamma^* = 1$ and $\gamma^* = 4$ obtained from Eq. (40) with Eqs. (16) and (17). The dashed lines show the optimal values from Eqs. (18), (19), and (20). The points mark measurements of the strategy with $\mu = 3$ and $\lambda = 10$ in search spaces with $N = 40$ (+) and $N = 400$ (×).

Comparison with Eq. (18) shows that that distance is in fact optimal, and that the $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES thus generates near optimal step lengths on ridge functions provided that the isolation periods are sufficiently long. As the second derivative of the normalised progress rate (with respect to the length of the isolation periods) is negative, the returns resulting from increasing the length of the isolation periods diminish with increasing γ^*. In practice, isolation periods should not be chosen too long as progress on objective functions that require fast adaptation of the step length, such as the sphere model, would be slow.

Figure 11 illustrates the dependence of the distance from the ridge axis, the mutation strength, and the progress rate on the topology parameter α for $\gamma^* = 1$ and $\gamma^* = 4$. The experimental setup is the same as that used to generate the data points in previous figures. Measurements from the first 100 isolation periods were discarded in order to reach the stationary limit state. The data points represent values that have been averaged over 10000 isolation periods. It can be seen that $\gamma^* = 1$ is generally insufficient to guarantee that the assumptions made in the calculations leading to Eq. (40) hold. The deviations between predictions and experiments are considerable especially for $\alpha < 2$, and they do not decrease significantly with increasing N. In contrast, relatively good agreement can be observed for $\gamma^* = 4$. Interestingly, for $\gamma^* = 1$, the analytically obtained stationary limit state of the $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES exactly agrees with that found for both cumulative step length adaptation and two-point adaptation in Sections 3 and 5, respectively. Increasing the length of the isolation periods improves both the accuracy of the results and the performance of the strategy.

Finally, Fig. 12 examines the behaviour of the $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES with $\gamma^* = 1$ and $\gamma^* = 4$ in the presence of noise. Shown are results for $\alpha = 2$ and $\alpha = 4$. At least for $N = 400$, the agreement of predictions and measurements is quite good unless σ^*_ϵ is too large. As for mutative self-adaptation, for high noise strengths fluctuations dominate the behaviour of the hierarchically organised strategy. In the face of low mutation strengths, which one of the two populations is successful is influenced by noise, and pressure toward larger mutation strengths is weak. For large σ^*_ϵ the strategy repeatedly performs a near random walk at low mutation strengths, resulting in low progress rates. Fluctuations of the mutation strength have not been considered in

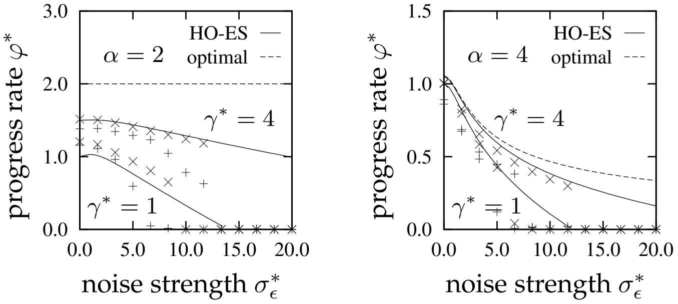

Figure 12: Normalised progress rate φ^* of the $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES plotted against the normalised noise strength σ_ϵ^*. The solid lines represent results for $\gamma^* = 1$ and $\gamma^* = 4$ obtained by solving Eq. (39) and using Eqs. (16) and (17). The dashed lines show the values that would be achieved if the mutation strength were adapted optimally. The points mark measurements of the strategy with $\mu = 3$ and $\lambda = 10$ in search spaces with $N = 40$ (+) and $N = 400$ (×).

the calculations. Therefore, the accuracy of the predictions is not satisfactory for larger values of σ_ϵ^*. However, the experimental results indicate that at least for large N, the hierarchically organised strategy is capable of significant progress beyond $\sigma_\epsilon^* = 2$. (Recall from Sections 3 and 5 that that value represents the noise strength beyond which cumulative step length adaptation and two-point adaptation systematically drive their step lengths to zero.) It is also clear both from Eq. (39) and from the measurements represented in Fig. 12 that longer isolation periods help improve the robustness of the hierarchically organised strategy in the presence of noise.

7 Summary and Conclusions

The previous sections have investigated the performance of several step length adaptation mechanisms on ridge functions. Equations have been derived that describe the average mutation strength as well as the progress rate achieved by the strategies in the limit $N \to \infty$. Figure 13 graphically summarises the analytically obtained results. Of the simplifications made in the analyses, the most significant are the assumption of infinite search space dimensionality and the decision not to model fluctuations of the state variables. Computer experiments in finite-dimensional search spaces have shown both the relative accuracy of the predictions in some parameter ranges and the failure of the simplified model underlying the analyses to accurately describe some of the strategies' behaviour in others (notably in the vicinity of the sharp ridge and for high levels of noise present). The main findings for the respective step length adaptation strategies are as follows:

Cumulative step length adaptation achieves at least 50% of the optimal progress rate for $\alpha \geq 2$, but generates finite step lengths (and thus achieves finite progress rates) for $1 < \alpha < 2$. In the vicinity of the sharp ridge, the quality of the analytically obtained predictions is not good unless the search space dimensionality is very high. On finite-dimensional ridges near the sharp one cumulative step length adaptation performs better than predicted. In the presence of noise, the step length is systematically driven to zero for $\sigma_\epsilon^* \geq 2$.

Mutative self-adaptation is the worst performing of the step length adaptation mech-

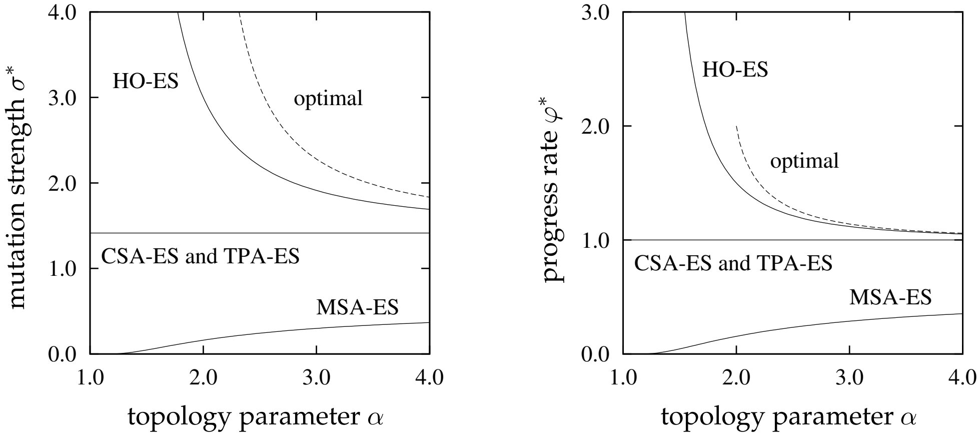

Figure 13: Comparison of adaptation strategies. The solid lines show theoretically obtained results valid in the limit $N \to \infty$ for strategies that use cumulative step length adaptation (CSA-ES), mutative self-adaptation (MSA-ES), and two-point adaptation (TPA-ES). Also represented is a hierarchically organised $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES with $\gamma^* = 4$ (HO-ES). For smaller values of γ^* the curve of the hierarchically organised strategy approaches those of CSA-ES and TPA-ES; for larger values of γ^* it approaches the optimal curves that have been obtained from Eqs. (19) and (20) and that are represented as dashed lines.

anisms considered. In contrast to the other strategies, it does not generate optimal step lengths even for large values of α. Moreover, it is unable to make efficient use of populations and performs best for $\mu = 1$. As noted in previous research, in the vicinity of the sharp ridge it may fail to generate useful step lengths altogether.

Two-point adaptation in theory generates step lengths that are identical for any value of α to those generated by cumulative step length adaptation. The quality of the approximation derived for two-point adaptation is better than that for cumulative step length adaptation. On finite-dimensional ridges near the sharp one, cumulative step length adaptation performs better than two-point adaptation. In the presence of noise the performance of the two adaptation mechanisms is very similar in that the step length is systematically driven to zero for $\sigma_\epsilon^* \geq 2$.

Hierarchically organised strategies offer the greatest potential on ridge functions of all of the adaptation strategies considered. The performance of the other step length adaptation mechanisms is matched for relatively small values of γ^*. With growing length of the isolation periods, nearly optimal performance is achieved for any value of α. In the presence of noise, too, do hierarchically organised strategies prove to be more robust than the other approaches.

However, two caveats need to be kept in mind when deploying hierarchically organised strategies. First, on objective functions that require fast, continual adaptation of the step length, long isolation periods are an impediment to progress. The realisation that isolation periods of different lengths are optimal in different settings has led to the suggestion made by (Herdy, 1992) to adapt γ by adding a third level to the hierarchy of evolutionary optimisation strategies. However, unless sufficient parallel processing capability is available, the computational costs may be prohibitive. On the positive side, it may be possible to set the length of the isolation periods statically such that satisfactory performance is achieved in a wide range of different settings. It has been seen above that good performance on ridges is typically achieved with relatively small settings of γ^* (such as $\gamma^* = 4$). According to Eq. (35), a constant value of γ^* can be achieved by choosing the length of the isolation periods proportional to N. Pending research on the performance of hierarchically organised strategies on the sphere model, it seems likely that that choice is such that it allows for linear convergence, opening up the possibility that a universally useful setting for the length of the isolation periods may be found. Moreover, Eq. (35) suggests that increasing the population size parameters μ and λ of the lower level strategy makes it possible to shorten the length of the isolation periods while at the same time maintaining the same value of γ^*.

The second caveat to be kept in mind are the higher computational costs of hierarchically organised strategies compared to the other algorithms. Hierarchically organised strategies require running several populations in parallel. For example, the $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES has twice the computational costs per time step of a strategy that uses cumulative step length adaptation if the same lower level population size parameters are used. Unless adequate parallel computational resources are available, the benefit of a better adapted step length needs to be weighed against the disadvantage of having to terminate after fewer time steps.

Clearly, this paper is but one step toward an improved understanding of the capabilities of different step length adaptation mechanisms. Arguably, the task of step length adaptation on ridge functions is a comparatively simple one. No continuous adaptation is required as optimal step lengths do not change over time. Evaluating the performance of step length adaptation mechanisms on ridge functions thus enables one to test the ability of a strategy to generate good static step lengths. It does not test the strategy's ability to generate those step lengths quickly, and to follow a moving target where optimal step lengths change as the optimisation progresses. That ability can be evaluated in fitness environments such as the sphere model. Some results with regard to mutative self-adaptation on the sphere model have been derived by (Beyer, 1996; Meyer-Nieberg and Beyer, 2005). Results for cumulative step length adaptation can be found in (Arnold, 2002; Arnold and Beyer, 2004). The performance of two-point adaptation and of hierarchically organised strategies on the sphere model remains to be studied. Especially results for the latter are of great interest as it remains to the seen whether the values for the length of the isolation periods that make the $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES perform well on ridges do not significantly impede its progress on the sphere.

A further task for future research is to improve the accuracy of the results in the vicinity of the sharp ridge, where the agreement of theoretical predictions with experimental measurements is often not good. Moreover, due to the normalisations adopted, any results that have been obtained for the sharp ridge only hold for $d = 1$. (Beyer and Meyer-Nieberg, 2006) have shown that for $\alpha = 1$, the performance of mutative self-adaptation depends qualitatively on the parameter d. The same can be found experimentally for other adaptation strategies. An analytically based explanation for this behaviour would contribute substantially to the understanding of those adaptation mechanisms.

It is furthermore desirable to extend the analyses to include N-dependent terms and to model fluctuations of the state variables, such as the distance from the ridge axis and the mutation strength. Modelling fluctuations would enable one to obtain more accurate results with regard to be behaviour of hierarchically organised strategies in the presence of noise. Considering N-dependent terms in the calculations would make it possible to infer recommendations with regard to the choice of population size parameters and allow making a more accurate comparison of the behaviour of various adaptation strategies in finite-dimensional search spaces.

Finally, it remains to study the behaviour of evolution strategies using nonisotropically distributed mutations, such as the CMA-ES. It has been pointed out by (Whitley et al., 2004) that strategies like the CMA-ES are potentially much more effective on ridge functions than those that use isotropically distributed mutations, and obtaining a quantitative understanding of their capabilities and limitations remains as a task for future work.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) though a Discovery Grant to D. V. Arnold and a USRA award to A. MacLeod.

References

Arnold, D. V. (2002). *Noisy Optimization with Evolution Strategies*. Genetic Algorithms and Evolutionary Computation Series. Kluwer Academic Publishers, Boston.

Arnold, D. V. (2006). Cumulative step length adaptation on ridge functions. In Runarsson, T. P. et al., editors, *Parallel Problem Solving from Nature — PPSN IX*, pages 11–20. Springer Verlag, Heidelberg.

Arnold, D. V. and Beyer, H.-G. (2004). Performance analysis of evolutionary optimization with cumulative step length adaptation. *IEEE Transactions on Automatic Control*, 49(4):617–622.

Arnold, D. V. and Beyer, H.-G. (2006). Evolution strategies with cumulative step length adaptation on the noisy parabolic ridge. Technical Report CS-2006-02, Faculty of Computer Science, Dalhousie University. Available at http://www.cs.dal.ca/research/techreports/2006/CS-2006-02.shtml.

Arnold, D. V. and MacLeod, A. (2006). Hierarchically organised evolution strategies on the parabolic ridge. In Cattolico, M. et al., editors, *GECCO '06: Proceedings of the 2006 Genetic and Evolutionary Computation Conference*, pages 437–444. ACM Press, New York.

Beyer, H.-G. (1996). Toward a theory of evolution strategies: Self-adaptation. *Evolutionary Computation*, 3(3):311–347.

Beyer, H.-G. (1998). Mutate large, but inherit small! On the analysis of rescaled mutations in $(\tilde{1}, \tilde{\lambda})$-ES with noisy fitness data. In Eiben, A. E. et al., editors, *Parallel Problem Solving from Nature — PPSN V*, pages 109–118. Springer Verlag, Heidelberg.

Beyer, H.-G. (2001a). On the performance of $(1, \lambda)$-evolution strategies for the ridge function class. *IEEE Transactions on Evolutionary Computation*, 5(3):218–235.

Beyer, H.-G. (2001b). *The Theory of Evolution Strategies*. Natural Computing Series. Springer Verlag, Heidelberg.

Beyer, H.-G. and Arnold, D. V. (2003). Qualms regarding the optimality of cumulative path length control in CSA/CMA-evolution strategies. *Evolutionary Computation*, 11(1):19–28.

Beyer, H.-G. and Meyer-Nieberg, S. (2006). Self-adaptation on the ridge functions class: First results for the sharp ridge. In Runarsson, T. P. et al., editors, *Parallel Problem Solving from Nature — PPSN IX*, pages 71–80. Springer Verlag, Heidelberg.

Beyer, H.-G. and Schwefel, H.-P. (2002). Evolution strategies — A comprehensive introduction. *Natural Computing*, 1(1):3–52.

David, H. A. and Nagaraja, H. N. (1998). Concomitants of order statistics. In Balakrishnan, N. et al., editors, *Handbook of Statistics*, volume 16, pages 487–513. Elsevier, Amsterdam.

Hansen, N. (1998). *Verallgemeinerte individuelle Schrittweitenregelung in der Evolutionsstrategie*. Mensch & Buch Verlag, Berlin.

Hansen, N. (2006). An analysis of mutative σ-self-adaptation on linear fitness functions. *Evolutionary Computation*, 14(3):255–275.

Hansen, N. and Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195.

Herdy, M. (1992). Reproductive isolation as strategy parameter in hierarchically organized evolution strategies. In Männer, R. and Manderick, B., editors, *Parallel Problem Solving from Nature — PPSN II*, pages 207–217. Elsevier, Amsterdam.

Herdy, M. (1993). The number of offspring as strategy parameter in hierarchically organized evolution strategies. *ACM SIGBIO Newsletter*, 13(2):2–9.

Lunacek, M. and Whitley, D. (2006). Searching for balance: Understanding self-adaptation on ridge functions. In Runarsson, T. P. et al., editors, *Parallel Problem Solving from Nature — PPSN IX*, pages 82–91. Springer Verlag, Heidelberg.

Meyer-Nieberg, S. and Beyer, H.-G. (2005). On the analysis of self-adaptive recombination strategies: First results. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*, pages 2341–2348. IEEE Press, Piscataway, NJ.

Meyer-Nieberg, S. and Beyer, H.-G. (2006). Self-adaptation in evolutionary algorithms. In Lobo, F. et al., editors, *Parameter Setting in Evolutionary Algorithms*. Springer Verlag, Heidelberg. In press.

Ostermeier, A., Gawelczyk, A., and Hansen, N. (1994). Step-size adaptation based on non-local use of selection information. In Davidor, Y. et al., editors, *Parallel Problem Solving from Nature — PPSN III*, pages 189–198. Springer Verlag, Heidelberg.

Oyman, A. I. and Beyer, H.-G. (2000). Analysis of the $(\mu/\mu, \lambda)$-ES on the parabolic ridge. *Evolutionary Computation*, 8(3):267–289.

Oyman, A. I., Beyer, H.-G., and Schwefel, H.-P. (1998). Where elitists start limping: Evolution strategies at ridge functions. In Eiben, A. E. et al., editors, *Parallel Problem Solving from Nature — PPSN V*, pages 109–118. Springer Verlag, Heidelberg.

Oyman, A. I., Beyer, H.-G., and Schwefel, H.-P. (2000). Analysis of the $(1, \lambda)$-ES on the parabolic ridge. *Evolutionary Computation*, 8(3):249–265.

Rechenberg, I. (1973). *Evolutionsstrategie — Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Friedrich Frommann Verlag, Stuttgart.

Rechenberg, I. (1978). Evolutionsstrategien. In Schneider, B. and Ranft, U., editors, *Simulationsmethoden in der Medizin und Biologie*, pages 83–114. Springer Verlag, Berlin.

Rechenberg, I. (1994). *Evolutionsstrategie '94*. Friedrich Frommann Verlag, Stuttgart.

Salomon, R. (1998). Evolutionary algorithms and gradient search: Similarities and differences. *IEEE Transactions on Evolutionary Computation*, 2(2):45–55.

Schwefel, H.-P. (1981). *Numerical Optimization of Computer Models*. Wiley, Chichester.

Whitley, D., Lunacek, M., and Knight, J. (2004). Ruffled by ridges: How evolutionary algorithms can fail. In Deb, K. et al., editors, *Genetic and Evolutionary Computation — GECCO 2004*, pages 294–306. Springer Verlag, Heidelberg.