Hierarchically Organised Evolution Strategies on the Parabolic Ridge

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

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

ABSTRACT

Organising evolution strategies hierarchically has been proposed as a means for adapting strategy parameters such as step lengths. Experimental research has shown that on ridge functions, hierarchically organised strategies can significantly outperform strategies that rely on mutative self-adaptation. This paper presents a first theoretical analysis of the behaviour of a hierarchically organised evolution strategy. Quantitative results are derived for the parabolic ridge that describe the dependence on the length of the isolation periods of the mutation strength and the progress rate. The issue of choosing an appropriate length of the isolation periods is discussed and comparisons with recent results for cumulative step length adaptation are drawn.

Categories and Subject Descriptors

G.1.6 [**Optimization**]: Unconstrained Optimization; I.2.8 [**Problem Solving, Control Methods, and Search**]; I.2.6 [**Learning**]: Parameter Learning

General Terms

Algorithms, Performance, Theory

Keywords

Hierarchically organised evolution strategies, step length adaptation, ridge functions

1. INTRODUCTION

Evolution strategies [5] are a type of evolutionary algorithm that is most commonly used for the optimisation of functions $f : \mathbb{R}^N \to \mathbb{R}$. In an attempt to achieve optimal or near optimal performance, they typically adapt their step lengths throughout the optimisation process. Step length adaptation mechanisms that have been proposed include mutative self-adaptation [3, 13, 16], cumulative step length adaptation [1, 9], and the use of hierarchically organised strategies [7, 14, 15].

A motivation for the use of hierarchically organised strategies is the insight that strategy parameter adaptation really is an optimisation problem. Consequently, evolutionary algorithms can be applied to solve it. 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. Variation and selection are used on both levels. Notice that mutative self-adaptation 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 [8] considers the problem of adapting the optimal number of offspring generated per time step and demonstrates empirically that near optimal values on the hyperplane and sphere models can be obtained.

While having been proposed a long time ago [14], there is not yet much knowledge — neither empirical nor theoretical — with regard to the capabilities and limitations of hierarchically organised evolution strategies, and relatively few publications have dealt with the issue. A notable exception is a paper by Herdy [7] in which the performance of strategies using a hierarchical organisation for adapting step lengths is compared empirically with that of strategies using mutative self-adaptation. Several objective functions are considered, including the sphere model as well as the sharp and parabolic ridges. It is found that isolation is detrimental to the performance of the strategies on the sphere. The sphere model requires fast adaptation of the step length, and isolation is neither necessary nor useful for successful adaptation. The situation is different on the ridges. On the parabolic ridge, mutative self-adaptation generates step lengths that are much smaller than optimal, resulting in slow progress. Short steps are likely to succeed in the short term, but yield inferior long

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.
Copyright 2006 ACM 1-59593-186-4/06/0007 ...$5.00.

term performance. Without isolation, opportunistic individuals that make short steps are rewarded, hampering long term progress. Herdy observes that hierarchically organised strategies with longer isolation periods generate much larger step lengths and thus significantly outperform strategies that use mutative self-adaptation. This situation is even more pronounced on the sharp ridge where strategies that use mutative self-adaptation drive their step lengths to zero and stagnate while hierarchically organised strategies are capable of tracking the ridge. As ridges are common features of many objective functions [17], the superior performance on ridges of hierarchically organised strategies is likely to be of practical significance.

Realising that isolation periods of different lengths are optimal in different environments, Herdy [7] 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 as well as on the 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.

Overall, the behaviour of hierarchically organised evolution strategies is not well understood. In particular, there is little knowledge with regard to the influence of strategy parameter settings on the upper level of the strategy, including the length of the isolation periods. A better understanding of the effects of design choices could presumably lead to the more widespread use and acceptance of hierarchically organised strategies. This paper makes a first step toward such an understanding by analysing the behaviour of a nontrivial hierarchically organised evolution strategy on the parabolic ridge. Its remainder is organised as follows. Section 2 describes hierarchically organised evolution strategies and introduces useful notation. Section 3 briefly summarises previously derived results with regard to the performance of (non-hierarchically organised) evolution strategies on the parabolic ridge. Those results are used in Section 4 which presents an analysis of the performance of hierarchically organised evolution strategies on the parabolic ridge. The approach relies on several simplifications and assumptions. The very concise results have the advantage of being easily understood and interpretable. However, they are not exact. In Section 5, it is verified that despite their simplicity, the theoretically obtained results qualitatively agree with experimental observations. Implications for the choice of the length of isolation periods are discussed, and comparisons with recently obtained results for cumulative step length adaptation are drawn. Section 6 concludes with a brief summary and suggestions for future research.

2. HIERARCHICALLY ORGANISED EVOLUTION STRATEGIES

The strategy considered in this paper is an instance of the general $[\mu'/\rho' \overset{+}{,} \lambda'(\mu/\rho \overset{+}{,} \lambda)^\gamma]$-ES described in [7, 15]. This section first describes the lower and then the upper level strategies.

2.1 Lower Level Strategy

The lower level strategy considered here is the $(\mu/\mu, \lambda)$-ES with isotropic mutations and intermediate recombination. It is popular both because it is relatively well understood and because of its good performance [5]. 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). More specifically, the $(\mu/\mu, \lambda)$-ES in every generation updates a search point $\mathbf{x} \in I\!R^N$ (the centroid of its population) using the following three 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 $\mathbf{z}^{(i)}$ are vectors consisting of N independent, standard normally distributed components.

2. The objective function values $f(\mathbf{y}^{(i)})$ of the offspring candidate solutions are determined. The index $k;\lambda$ is used to 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.

3. The average
$$\mathbf{x} = \frac{1}{\mu}\sum_{k=1}^{\mu}\mathbf{x}^{(k;\lambda)}$$
of the μ best of the offspring candidate solutions is computed and replaces the previous search point.

Notice that the mutation strength σ is constant throughout an entire run of the lower level strategy. Also notice that our lower level strategy differs from that considered by Herdy [7] only in that we use intermediate recombination rather than discrete recombination. The choice has been made as the former is easier to handle analytically.

2.2 Upper Level Strategy

The mutation strength σ is the single strategy parameter that is adapted by the upper level strategy. Thus, while the lower level strategy faces an N-dimensional optimisation problem, that of the upper level strategy is one-dimensional. As a consequence, a very simple algorithm can be used:

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

2. Parameter α is set to a value uniformly drawn from the interval $[1.1, 1.5]$.

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 $\sigma \cdot \alpha$, the other one uses σ/α.

4. The objective function values of the final search points generated in the two runs of the lower level strategy 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.

In the notation introduced in [7, 15], the overall strategy thus described is a $[1, 2(\mu/\mu, \lambda)^{\gamma}]$-ES. The purpose of step 2 is to generate two mutation strengths, one larger than the previous one and one smaller. The exact nature of the rule for doing so is of minor significance. We have randomised the choice of α rather than simply using $\alpha = 1.3$ as Herdy [7] did in order no to be confined to a discrete set of mutation strengths that would lead to artifacts in the performance graphs below. In general, admitting larger values of α allows for potentially faster adaptation while at the same time leading to stronger fluctuations in the adaptation process. For values of α very close to 1, the difference between the final search points of the two populations is dominated by random factors rather than being governed by the different mutation strengths used, and a random walk behaviour of the mutation strength may result. From our experience, constraining α to be in $[1.1, 1.5]$ avoids both strong fluctuations and random walk behaviour.

3. THE PARABOLIC RIDGE

The parabolic ridge is a commonly used function for testing the ability of optimisation strategies to make progress in one particular direction in search space, where deviation from that direction is penalised. It can be described by objective function

$$f(\mathbf{x}) = x_1 - \frac{d}{N}\sum_{i=2}^{N} x_i^2 \qquad (1)$$

where $\mathbf{x} = \langle x_1, \ldots, x_N \rangle \in I\!R^N$ and where the task is maximisation. Figure 1 shows a plot of the function for $N = 2$. Notice that while the ridge function has no finite maximum, maximisation still is a meaningful task if progress towards larger objective function values is considered the goal of optimisation. The x_1-axis is referred to as the ridge axis. It is important to realise that while in the definition used here the ridge is aligned with an axis of the coordinate system, that fact is irrelevant for strategies that rely on isotropic mutations such as those considered here and described in Section 2. Evolution strategies with isotropically distributed mutations do not exploit separability of the objective function. The coordinate system could be subjected to an arbitrary rotation without affecting the strategies' performance.

According to Eq. (1), 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

$$R = \sqrt{\sum_{i=2}^{N} x_i^2}$$

from the ridge axis. In the long term, only the former is a viable possibility as the first term on the right hand side of Eq. (1) can increase indefinitely while the second can never exceed zero. However, for the parabolic ridge, progress in the direction of the ridge axis enters the computation of the fitness linearly while the reduction of R makes a quadratic contribution. Small mutation strengths afford the short term advantage of reducing the distance from the ridge axis, but lead to slow long term progress.

For large values of N, the performance of the $(\mu/\mu, \lambda)$-ES on the parabolic ridge is relatively well understood. Oyman et al. [11, 12] have studied the behaviour of the $(1, \lambda)$-ES. Oyman and Beyer [10] have generalised the analysis for the $(\mu/\mu, \lambda)$-ES. All of these studies assume fixed mutation strengths. More recently, Arnold and Beyer [2] have investigated the influence of different forms of noise on the performance of the $(\mu/\mu, \lambda)$-ES on the parabolic ridge, and they have analysed the behaviour of cumulative step length adaptation. In Section 5, results from that study will be contrasted with those obtained in Section 4 for hierarchically organised strategies. The remainder of this section summarises the relevant insights gained in [10], adapted to conform to the somewhat different notation used here and simplified by dropping any terms that disappear in the limit $N \to \infty$.

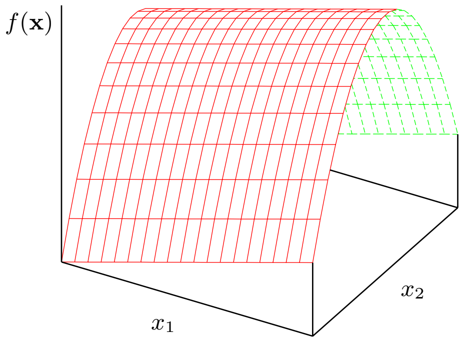

Figure 1: A plot of the two-dimensional parabolic ridge.

In all of what follows, R denotes the distance of the strategy's search point from the ridge axis. For fixed mutation strength σ, the $(\mu/\mu, \lambda)$-ES tracks the parabolic ridge at a varying distance R. After initialisation effects have faded, the distribution of R values is time invariant. The distance of the population centroid from the ridge axis fluctuates around a stationary average value while the value of the x_1-component increases. By considering the case that $N \to \infty$, results from the analysis of the sphere model can be used to derive an approximation for the average distance at which the ridge axis is tracked. Introducing for notational convenience $\varrho = 2Rd/N$ as the normalised distance of the population centroid from the ridge axis and $\sigma^* = \sigma d/\mu c_{\mu/\mu,\lambda}$ as the normalised mutation strength of the strategy, in [2, 10] it has been seen that

$$\varrho^2(\sigma^*) = \frac{\sigma^{*2}}{2} + \sqrt{\frac{\sigma^{*4}}{4} + \sigma^{*2}} \qquad (2)$$

can be used as an approximation for the average squared normalised distance from the ridge axis provided that N is sufficiently large. That is, the distance from the ridge axis increases monotonically with increasing mutation strength, and for large mutation strengths the dependence is nearly linear. Figure 2 illustrates this relationship and demonstrates that even though having been obtained for $N \to \infty$, Eq. (2) provides a reasonably good description of evolution strategy behaviour even for small values of N.

Furthermore, defining the progress rate φ of the strategy as the expected distance in direction of the x_1-axis that the strategy's search point travels per generation and introducing normalisation $\varphi^* = \varphi d/\mu c_{\mu/\mu,\lambda}^2$, it can be derived from

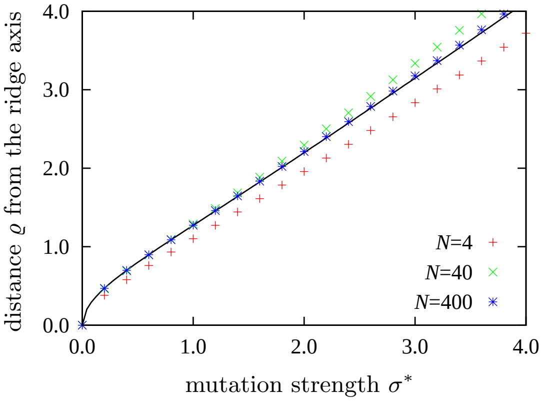

Figure 2: Normalised distance ϱ from the ridge axis plotted against normalised mutation strength σ^*. The solid line has been obtained from Eq. (2)**. The points represent results measured in runs of the** $(\mu/\mu,\lambda)$**-ES with** $\mu = 3$ **and** $\lambda = 10$ **in search spaces with** $N \in \{4, 40, 400\}$ **and** $d = 1.0$**.**

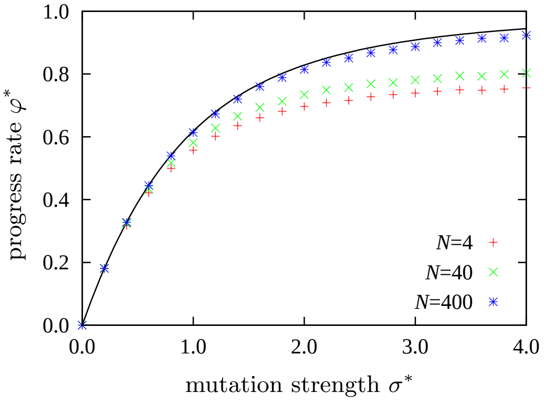

Figure 3: Normalised progress rate φ^* plotted against normalised mutation strength σ^*. The solid line has been obtained from Eq. (3)**. The points represent results measured in runs of the** $(\mu/\mu,\lambda)$**-ES with** $\mu = 3$ **and** $\lambda = 10$ **in search spaces with** $N \in \{4, 40, 400\}$ **and** $d = 1.0$**.**

the results in [10] that

$$\varphi^*(\sigma^*) = \frac{\sigma^{*2}}{\sigma^{*2}/2 + \sqrt{\sigma^{*4}/4 + \sigma^{*2}}} \tag{3}$$

can serve as an approximation for the normalised progress rate provided that N is sufficiently large. It is easy to see from Eq. (3) that the progress rate of the $(\mu/\mu,\lambda)$-ES increases monotonically with increasing mutation strength, and that for large values of σ^* the normalised progress rate tends toward a value of 1. Figure 3 illustrates this relationship. It can be seen how the quality of the approximation improves with increasing N.

4. PERFORMANCE ANALYSIS

This section uses the results describing the performance of the $(\mu/\mu,\lambda)$-ES on the parabolic ridge to derive a characterisation of the behaviour of the hierarchically organised strategy outlined in Section 2. The analysis assumes that the isolation periods are sufficiently long in order for several simplifications described below to be made. It will be seen in experiments that the accuracy of the predictions that can be obtained is good for large values of γ and N, but that the formulas derived also provide a good qualitative understanding of the behaviour of the hierarchically organised strategy for relatively small values of those parameters.

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 strategy for the duration of an isolation period. More specifically, given a population centroid $\mathbf{x}$ that has been arrived at with a mutation strength of σ, it is necessary to estimate the objective function value of the population centroid $\mathbf{x}'$ obtained after running the lower level strategy with a mutation strength of ς (which here is either $\sigma\cdot\alpha$ or σ/α) for a further γ time steps. 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 three assumptions are made.

1. At the end of an isolation period, the lower level strategy is in the stationary limit state described by Eq. (2).
2. 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.
3. For the purpose of comparing fitness values of population centroids, it is sufficient to consider their expected values; i.e., fluctuations can be ignored.

Clearly, validity of the second assumption implies validity of the first. Both of them hold if the length γ of the isolation periods is 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 assumptions to hold with a certain accuracy. As for the third 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 the qualitative agreement of results derived under the assumption with experimental measurements is good.

Assuming that γ is sufficiently large, the population centroid is at a normalised distance $\varrho(\sigma^*)$ from the ridge axis at the beginning and at a normalised distance $\varrho(\varsigma^*)$ at the end of the isolation period, where mutation strengths are normalised as outlined above and where the distances from the ridge axis are described by Eq. (2). From Eq. (1) with the normalisation of the distance from the ridge axis, the objective function values at the beginning and at the end of the isolation period are $f(\mathbf{x}) = x_1 - N\varrho^2(\sigma^*)/4d$ and $f(\mathbf{x}') = x'_1 - N\varrho^2(\varsigma^*)/4d$, respectively. The expected difference between the objective function values of population centroids $\mathbf{x}$ and $\mathbf{x}'$ is thus

$$\begin{aligned}\Delta f &= f(\mathbf{x}') - f(\mathbf{x}) \\ &= \gamma\varphi(\varsigma^*) - \frac{N}{4d}\left(\varrho^2(\sigma^*) - \varrho^2(\varsigma^*)\right).\end{aligned}$$

The first of the two terms on the right hand side is due to progress in the direction of the ridge axis and has been computed as the product of the expected progress per time step and the number of steps made. (Recall that progress is assumed to have been 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. With the normalised length $\gamma^* = \gamma\mu c^2_{\mu/\mu,\lambda}/N$ of the isolation periods, it thus follows

$$\Delta f(\gamma^*, \sigma^*, \varsigma^*) = \frac{N}{d}\left(\gamma^*\varphi^*(\varsigma^*) - \frac{\varrho^2(\sigma^*)}{4} + \frac{\varrho^2(\varsigma^*)}{4}\right) \quad (4)$$

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

The $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES described in Section 2 evolves two populations in parallel, one with mutation strength $\sigma \cdot \alpha$ and one with mutation strength σ/α. 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

$$g(\alpha) = \frac{d}{N}\left(f(\mathbf{x}'_1) - f(\mathbf{x}'_2)\right)$$

it is clear that the mutation strength used in the next iteration of the upper level strategy is $\sigma \cdot \alpha$ (the mutation strength that led to $\mathbf{x}'_1$) if $g(\alpha) \geq 0$ and σ/α (the mutation strength that led to $\mathbf{x}'_2$) otherwise. Function $g(\alpha)$ is referred to as the gain difference. With Eq. (4) it follows that

$$\begin{aligned} g(\alpha) &= \frac{d}{N}\left(\Delta f(\gamma^*, \sigma^*, \sigma^* \cdot \alpha) - \Delta f(\gamma^*, \sigma^*, \sigma^*/\alpha)\right) \\ &= \gamma^*\varphi^*(\sigma^* \cdot \alpha) - \gamma^*\varphi^*(\sigma^*/\alpha) \\ &\qquad - \frac{\varrho^2(\sigma^* \cdot \alpha)}{4} + \frac{\varrho^2(\sigma^*/\alpha)}{4} \end{aligned} \quad (5)$$

where ϱ and φ^* are given by Eqs. (2) and (3), respectively.

Rather than attempting to determine the distribution of the normalised mutation strength in the limit of large γ, we will see that an approximation of the average value of σ^* can be computed by relatively simple means. It is clear from Eq. (5) that $g(1) = 0$ independent of γ^* and σ^*. For sufficiently small values of α, the sign of $g(\alpha)$ in the vicinity of 1 is thus determined by the derivative $g'(1) = \partial g/\partial\alpha|_{\alpha=1}$. The mutation strength of the next iteration of the upper level strategy is $\sigma \cdot \alpha$ if $g'(1) > 0$ and it is σ/α if $g'(1) < 0$. That is, as $\alpha > 1$ by definition of the algorithm in Section 2, the mutation strength is increased if $g'(1) > 0$ and 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 $\sigma \cdot \alpha$ and σ/α 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.

Figure 4 plots the gain difference $g(\alpha)$ for $\gamma^* = 10.0$ and several values of σ^*. For $\sigma^* = 1.0$, the derivative $g'(1)$ is positive and the mutation strength will be increased. For $\sigma^* = 4.0$, the sign of $g'(1)$ is negative and the mutation strength will be decreased. For $\sigma^* = 2.0$, the curve is nearly flat at the origin and whether the mutation strength is increased or decreased is largely random. Thus, $\sigma^* = 2.0$ can be expected to be not far from the average normalised mutation strength of the $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES on the parabolic ridge with $\gamma^* = 10.0$. This is confirmed in experiments summarised below. Notice that the range of the horizontal axis in Fig. 4 is larger than the interval from which α-values are drawn, thus justifying the reliance on the sign of $g'(1)$ for determining the sign of $g(\alpha)$.

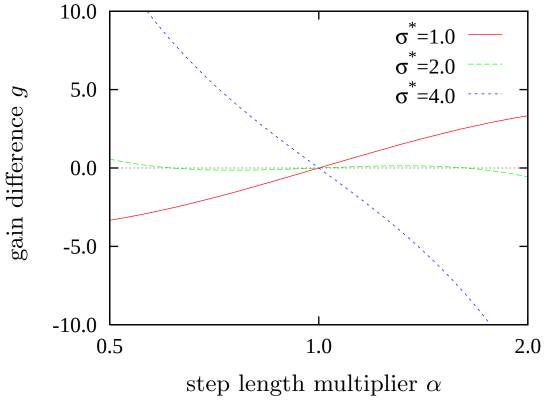

Figure 4: Gain difference $g(\alpha)$ plotted against the step length multiplier α for $\gamma^* = 10.0$ and several values of σ^*. Notice that the scale of the horizontal axis is logarithmic.

Using Eqs. (2), (3), and (5), the gain difference can be written as

$$g(\alpha) = \frac{\gamma^*\sigma^{*2}\alpha^2}{\varrho^2(\sigma \cdot \alpha)} - \frac{\gamma^*\sigma^{*2}}{\alpha^2\varrho^2(\sigma/\alpha)} - \frac{\varrho^2(\sigma \cdot \alpha)}{4} + \frac{\varrho^2(\sigma/\alpha)}{4}.$$

Computing the derivative with respect to α for $\alpha = 1$ results in

$$g'(1) = \frac{2\gamma^*\sigma^{*2}}{\varrho^4(\sigma^*)}\left(2\varrho^2(\sigma^*) - \sigma^*\frac{\mathrm{d}\varrho^2}{\mathrm{d}\sigma^*}\right) - \frac{\sigma^*}{2}\frac{\mathrm{d}\varrho^2}{\mathrm{d}\sigma^*}.$$

It is easily verified from Eq. (2) that

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

and that

$$\frac{\mathrm{d}\varrho^2}{\mathrm{d}\sigma^*} = \sigma^*\frac{\varrho^2(\sigma^*) + 1}{\varrho^2(\sigma^*) - \sigma^{*2}/2}.$$

It follows that

$$\begin{aligned} g'(1) &= \frac{2\gamma^*}{\varrho^2(\sigma^*) + 1}\left(2\varrho^2(\sigma^*) - \sigma^{*2}\frac{\varrho^2(\sigma^*) + 1}{\varrho^2(\sigma^*) - \sigma^{*2}/2}\right) \\ &\qquad - \frac{\sigma^{*2}}{2}\frac{\varrho^2(\sigma^*) + 1}{\varrho^2(\sigma^*) - \sigma^{*2}/2} \\ &= \frac{\sigma^{*2}}{\varrho^2(\sigma^*) - \sigma^{*2}/2}\left(\frac{2\gamma^*}{\varrho^2(\sigma^*) + 1} - \frac{\varrho^2(\sigma^*) + 1}{2}\right). \end{aligned}$$

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

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

Taking the square root results in

$$\varrho^2 = \sqrt{4\gamma^*} - 1 \quad (6)$$

as an approximation for the average squared normalised distance at which the $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES tracks the parabolic ridge.

Using Eq. (6) in Eq. (2) results in

$$\sqrt{4\gamma^*} - 1 - \frac{\sigma^{*2}}{2} = \sqrt{\frac{\sigma^{*4}}{4} + \sigma^{*2}}.$$

Squaring both sides and rearranging terms it follows that

$$\begin{aligned}\sigma^{*2} &= \sqrt{4\gamma^*} - 2 + \frac{1}{\sqrt{4\gamma^*}} \\ &= \frac{(\sqrt{4\gamma^*} - 1)^2}{\sqrt{4\gamma^*}}. \qquad (7)\end{aligned}$$

Taking the square root yields

$$\begin{aligned}\sigma^* &= \frac{\sqrt{4\gamma^*} - 1}{(4\gamma^*)^{1/4}} \\ &= (4\gamma^*)^{1/4} - (4\gamma^*)^{-1/4} \qquad (8)\end{aligned}$$

as an approximation for the average normalised mutation strength of the $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES on the parabolic ridge.

Finally, using Eqs. (2), (6), and (7) in Eq. (3) results in

$$\begin{aligned}\varphi^* &= \frac{\sigma^{*2}}{\varrho^2(\sigma^*)} \\ &= \frac{\sqrt{4\gamma^*} - 1}{\sqrt{4\gamma^*}} \\ &= 1 - \frac{1}{\sqrt{4\gamma^*}} \qquad (9)\end{aligned}$$

as an expression for the normalised progress rate of the $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES on the parabolic ridge. Notice the remarkable simplicity of Eqs. (6), (8), and (9).

5. DISCUSSION

Figures 5, 6, and 7 compare results from Eqs. (6), (8), and (9) with measurements made in runs of hierarchically organised evolution strategies on parabolic ridges of different dimensionalities. It can be seen that for $N = 400$ the accuracy of the predictions is good except for the smallest values of γ^*. Note that due to the normalisation of the length of the isolation periods, the smallest γ values represented in the graphs are indeed small; for $N = 4$, the leftmost data points in Figs. 5, 6, and 7 correspond to $\gamma = 1$ and thus no isolation at all. For the smaller search space dimensionalities, larger deviations of the measured results from those that have been obtained theoretically occur. However, except for small values of γ^*, the deviations that can be observed in the figures appear to be of the same order of magnitude as those in Figs. 2 and 3 that had been obtained for the $(\mu/\mu, \lambda)$-ES with fixed mutation strength. As the results for that strategy have been used in the derivation of the results for the hierarchically organised strategy, more accurate predictions could not have been expected. Any additional inaccuracies are due in part to the fact that the mutation strength is not constant but instead fluctuates, and that those fluctuations have not been considered in the analysis in Section 4. Altogether, Eqs. (6), (8), and (9) provide a useful qualitative description of the behaviour of the strategy even for N as small as 4.

As seen in [10] and illustrated in Fig. 3 above, on the parabolic ridge it is always beneficial for the long term success of evolution strategies to increase the mutation strength. As for the hierarchically organised strategy longer isolation periods result in larger mutation strengths, the graph in

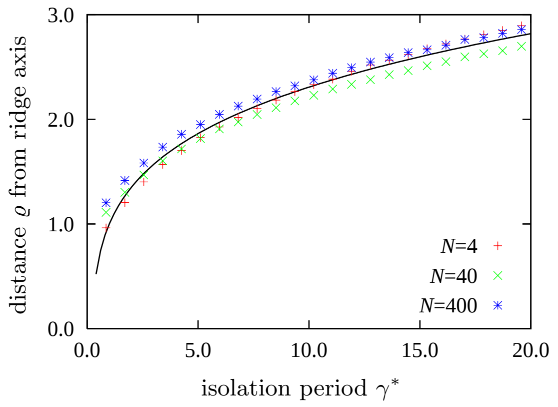

Figure 5: Average normalised distance ϱ from the ridge axis plotted against normalised length γ^* of the isolation periods. The solid line has been obtained from Eq. (6)**. The points have been measured in runs of the** $[1, 2(3/3, 10)^\gamma]$**-ES with** $d = 1.0$**.**

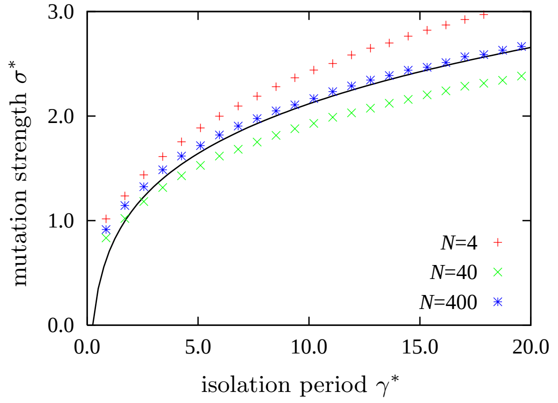

Figure 6: Average normalised mutation strength σ^* plotted against normalised length γ^* of the isolation periods. The solid line has been obtained from Eq. (8)**. The points have been measured in runs of the** $[1, 2(3/3, 10)^\gamma]$**-ES with** $d = 1.0$**.**

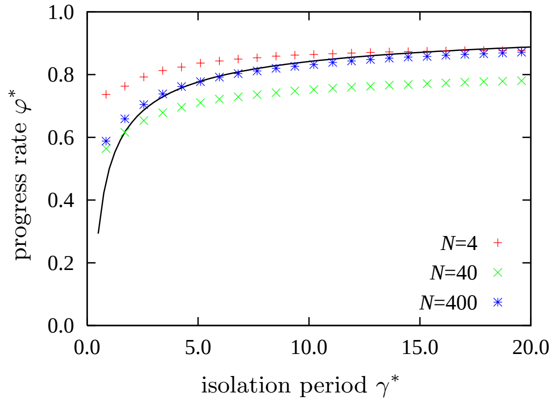

Figure 7: Normalised progress rate φ^* plotted against normalised length γ^* of the isolation periods. The solid line has been obtained from Eq. (9)**. The points have been measured in runs of the** $[1, 2(3/3, 1)^\gamma]$**-ES with** $d = 1.0$**.**

Fig. 7 is monotonically increasing. On other objective functions, such as the sphere model, long isolation periods prevent fast adaptation of the mutation strength and hamper progress. Adapting the length of the isolation periods as suggested by Herdy [7] is a possibility, but it adds to the computational costs of the strategy. It is thus desirable to give a recommendation for the length of the isolation periods that yields satisfactory performance on both the sphere and the ridge (and, hopefully, on other functions as well). It can be seen from Fig. 7 that a large proportion of the maximal progress rate is achieved already with relatively small values of γ^*. Equation (9) suggests that for $\gamma^* = 1$, the $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES achieves 50% of its maximal progress; for $\gamma^* = 4$ it achieves 75%. Figure 7 shows that for finite N, the proportions of the maximal progress rate that are achieved with those values of the normalised length of the isolation periods are even higher. For the parabolic ridge, choosing

$$\gamma = \frac{\beta N}{\mu c^2_{\mu/\mu,\lambda}} \tag{10}$$

where $\beta \in [1.0, 4.0]$ thus guarantees that a substantial proportion of the maximal progress rate is achieved. Increasing γ further would yield a speed-up of at most 33%. The length of the isolation periods should thus be chosen proportional to the dimensionality of the search space. Equation (10) also suggests that using larger values of μ and λ (and thus increasing the denominator) allows operating with shorter isolation periods. It is important to keep in mind however that the results from Section 4 are not sufficiently accurate in order to determine optimal settings of the population size parameters, and more work will need to be done in order to confirm the value of Eq. (10) for the choice of the length of the isolation periods. It also remains to be seen what proportion of the optimal progress rate on the sphere model can be achieved with that recommendation.

Finally, it is interesting to compare the results for the hierarchically organised strategy with those for the $(\mu/\mu, \lambda)$-ES with cumulative step length adaptation. In [2] it has been seen that a $(\mu/\mu, \lambda)$-ES with cumulative step length adaptation on the parabolic ridge in the limit $N \to \infty$ employs an average normalised mutation strength of $\sigma^* = 1/\sqrt{2}$. With that step length, the average normalised distance from the ridge axis and the resulting normalised progress rate are $\varrho = 1$ and $\varphi^* = 1/2$, respectively. With long isolation periods, the hierarchically organised strategy can thus achieve nearly twice the progress rate of the strategy that uses cumulative step length adaptation. However, it is important to keep in mind that the hierarchically organised strategy evolves two populations in parallel. For the same value λ, its computational costs (quantified as the number of objective function evaluations) per time step are thus twice as high. The progress per unit of cost is thus roughly the same for both strategies. It is also interesting to note that the time scales on which the strategies adapt the mutation strength are similar. According to [6], for cumulative step length adaptation the cumulation parameter is commonly chosen to be inversely proportional to the search space dimensionality. It thus takes order N steps for the information accumulated in the search path to fade. For the hierarchically organised strategy, Eq. (10) suggests that the length of the isolation periods should be chosen proportional to N. Thus, for both strategies order N steps are required for the mutation strength to change by a constant factor.

6. SUMMARY AND CONCLUSIONS

To conclude, this paper has presented a first analysis of the behaviour or a hierarchically organised evolution strategy on the parabolic ridge. Equations have been derived that describe the average mutation strength as well as the progress rate achieved by the strategy. While several simplifications and assumptions have been made in the derivation of the results, numerical experiments suggest that the accuracy of the results is good for not too small values of the length of the isolation periods and of the search space dimensionality. It has been seen that both the average mutation strength of the hierarchically organised strategy and the average distance at which the ridge axis is tracked increase with the fourth root of the length of the isolation periods. The progress rate asymptotically approaches its optimal value (that is obtained for very large mutation strengths), and the deviation from the optimal value is inversely proportional to the square root of the length of the isolation periods. Choosing the length of the isolation periods according to Eq. (10) ensures that a substantial proportion of the maximal progress rate is realised. A comparison with the $(\mu/\mu, \lambda)$-ES that employs cumulative step length adaptation has shown that potentially, the $[1, 2(\mu/\mu, \lambda)^\gamma]$-ES can achieve twice the progress rate, albeit at twice the computational costs per time step.

Clearly, this paper is but a first step in the analysis of the behaviour of hierarchically organised evolution strategies. Numerous ways of generalising and extending its results are conceivable. First, it is desirable to obtain a more accurate description of the behaviour of the strategy for short isolation periods and for small values of N. Such an approximation would be useful for the task of computing optimal population size parameters for the lower level strategy. Second, it is interesting to study the influence of the choice of distribution used for generating the step length multiplier α on the performance of the strategy. It seems conceivable that larger populations on the lower level may allow using larger values of α, thus enabling faster adaptation. Third, other objective functions remain to be studied. Of particular interest is the sphere model as it requires relatively short isolation periods for efficient performance. The techniques used by Beyer [3] for the analysis of mutative self-adaptation may be useful for that task. Another interesting candidate for analysis is the general ridge function class, and in particular the sharp ridge. For fixed mutation strength, such an analysis has been presented by Beyer [4], and the approach pursued here should be easily adapted to that case. Also of interest is the case that there is noise present in the optimisation process. Results obtained in [2] indicate that cumulative step length adaptation performs less than optimally in the presence of noise, and it will be of great interest to derive corresponding results for hierarchically organised strategies and compare them. Finally, the potential of hierarchically organised strategies for step length adaptation in the CMA-ES described by Hansen and Ostermeier [6] remains to be explored.

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.

7. REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[15] I. Rechenberg. *Evolutionsstrategie '94*. Friedrich Frommann Verlag, Stuttgart, 1994.

[16] H.-P. Schwefel. *Numerical Optimization of Computer Models*. Wiley, Chichester, 1981.

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