# Evolution Strategies with Cumulative Step Length Adaptation on the Noisy Parabolic Ridge

**Dirk V. Arnold**
**Hans-Georg Beyer**

Technical Report CS-2006-02

January 16, 2006

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

# Evolution Strategies with Cumulative Step Length Adaptation on the Noisy Parabolic Ridge

Dirk V. Arnold
*Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada B3H 1W5*

Hans-Georg Beyer
*Department of Computer Science, Research Center Process and Product Engineering, Vorarlberg University of Applied Sciences, Hochschulstr. 1, A-6850 Dornbirn, Austria*

**Abstract.** This paper presents an analysis of the performance of the $(\mu/\mu, \lambda)$-ES with isotropic mutations and cumulative step length adaptation on the noisy parabolic ridge. Several forms of dependency of the noise strength on the distance from the ridge axis are considered. Closed form expressions are derived that describe the mutation strength and the progress rate of the strategy in high-dimensional search spaces. It is seen that as for the sphere model, larger levels of noise present lead to cumulative step length adaptation generating increasingly inadequate mutation strengths, and that the problem can be ameliorated to some degree by working with larger populations.

**Keywords:** Evolutionary computation, evolution strategies, optimisation, noise, cumulative step length adaptation, ridge functions

## 1. Introduction

Evolution strategies are a type of evolutionary algorithm that is most commonly used for the optimisation of real-valued functions of the form $f : \mathbb{R}^N \to \mathbb{R}$. Typical features of evolution strategies include the use of truncation selection, normally distributed mutations, and some form of self-adaptation for step length control. See [13] for a comprehensive introduction. Much work has gone into the analysis of the behaviour of evolution strategies on simple objective functions, such as the sphere model [1, 5, 11, 23], ellipsoidal fitness landscapes [12], the corridor model [22], and the ridge function class [10, 20, 21, 19, 23]. While the sphere and the ellipsoids serve as models for fitness landscapes in the vicinity of local optima, both the corridor and the ridge strive to model features of such landscapes in greater distance from the optima. More specifically, they test the ability of a strategy to make progress in a particular direction in search space, where deviation from that direction is penalised. The goal of such analyses is to derive scaling laws that help reveal strengths and weaknesses of particular strategy

variants, to make recommendations with regard to the setting of the strategies' exogenous parameters, and to contribute to the continued improvement of existing and the design of new strategy variants.

Ridge functions are known to pose significant problems for optimisation strategies. Whitley, Lunacek, and Knight [26] point out 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". For evolutionary algorithms in particular, while long term progress is best achieved with large step lengths, mutation strength adaptation mechanisms are often shortsighted and generate step lengths much shorter than optimal. This deficiency on ridges is the cause of the premature convergence of evolution strategies on a unimodal objective function that has been observed by Salomon [25].

Several steps have been made toward a quantitative understanding of the behaviour of evolution strategies on ridge functions. Rechenberg [23] provides (without derivation) a formula for the progress rate of evolution strategies on the parabolic ridge. However, that formula contains the location in search space as a parameter, and no attempt is made to model the distribution of the population in search space. Oyman, Beyer and Schwefel [20, 21] study the performance of the $(1 \stackrel{+}{,} \lambda)$-ES with fixed mutation strength on the parabolic ridge and derive formulas both for the average distance from the ridge axis and for the progress rate of the algorithm. They find that the comma strategy is generally superior to the plus strategy. In a generalisation of that work, Oyman and Beyer [19] study the behaviour of the $(\mu/\mu, \lambda)$-ES with both intermediate and dominant recombination. Beyer [10] also considers the performance of the $(1, \lambda)$-ES on ridges other than the parabolic one and finds that qualitatively different behaviours can result on different ridge topologies. Insights with regard to the issue of step length adaptation that have been published so far are purely empirical and include the aforementioned paper by Salomon [25] as well as results provided by Herdy [17].

This paper studies the performance of evolution strategies with intermediate multirecombination and cumulative step length adaptation on the noisy parabolic ridge function class. It thus extends previous work [19, 21] in two directions: first, it considers the influence of noise; and second, it analytically studies the performance of cumulative step length adaptation. Its remainder is organised as follows. Section 2 describes the $(\mu/\mu, \lambda)$-ES with cumulative step length adaptation, summarises some results on expected values of order statistics that are used later in the paper, and briefly reviews the approach to the analysis of the sphere model that will be seen to form an important step

in the investigation of the ridge function class. Section 3 studies the performance of the $(\mu/\mu, \lambda)$-ES on the noisy parabolic ridge without considering step length adaptation. Scaling laws for the progress rate as well as for the average distance from the ridge axis are derived, and several forms of dependency of the noise strength on the distance from the ridge axis are considered. The results are obtained in the limit of infinite search space dimensionality and their significance is verified in experiments in finite-dimensional search spaces. Section 4 includes cumulative step length adaptation in the analysis. Section 5 concludes with a brief summary of the results and suggestions for future research.

## 2. Preliminaries

This section first describes the $(\mu/\mu, \lambda)$-ES with isotropic mutations and cumulative step length adaptation for the optimisation of functions $f : \mathbb{R}^N \to \mathbb{R}$. Next, terminology and two useful lemmas from the field of order statistics are introduced. Then, some important results with regard to the behaviour of the $(\mu/\mu, \lambda)$-ES on the quadratic sphere model are summarised. Those results form a cornerstone in the analysis of the behaviour of the $(\mu/\mu, \lambda)$-ES on the ridge function class presented in Sections 3 and 4.

### 2.1. The $(\mu/\mu, \lambda)$-ES

The strategy under consideration in this paper is the $(\mu/\mu, \lambda)$-ES with isotropic mutations and intermediate recombination. That strategy is popular due to both its good performance and its amenability to mathematical analysis. The following description of the algorithm is deliberately brief. See [13] for a more comprehensive discussion of evolution strategies and their naming conventions, and see [18] for a thorough motivation of cumulative step length adaptation.

In every time step the $(\mu/\mu, \lambda)$-ES computes the centroid of the population of candidate solutions as a search point $\mathbf{x} \in \mathbb{R}^N$ that mutations are applied to. For the purpose of adapting the mutation strength, a vector $\mathbf{s} \in \mathbb{R}^N$ that is referred to as the search path is used to accumulate information about the directions of the most recently taken steps. An iteration of the strategy updates the search point along with the search path and the mutation strength of the strategy in five steps:

1. Generate $\lambda$ offspring candidate solutions $\mathbf{y}^{(i)} = \mathbf{x} + \sigma\mathbf{z}^{(i)}$, $i = 1, \ldots, \lambda$, where mutation strength $\sigma > 0$ determines the step length

and the $\mathbf{z}^{(i)}$ are vectors consisting of $N$ independent, standard normally distributed components.

2. Determine the objective function values $f(\mathbf{y}^{(i)})$ of the offspring candidate solutions and compute the average

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

of the $\mu$ best of the $\mathbf{z}^{(i)}$. The index $k;\lambda$ refers to the $k$th best of the $\lambda$ offspring candidate solutions (i.e., the $k$th largest if the task is maximisation and the $k$th smallest if the task is minimisation). Vector $\mathbf{z}^{(\text{avg})}$ is referred to as the progress vector.

3. Update the search point according to

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

Clearly, the new search point is the arithmetic mean of the $\mu$ best of the offspring candidate solutions.

4. Update the search path according to

$$\mathbf{s} \leftarrow (1-c)\mathbf{s} + \sqrt{\mu c(2-c)}\mathbf{z}^{(\text{avg})}, \tag{3}$$

where the cumulation parameter $c$ determines how rapidly the direction information stored in $\mathbf{s}$ fades.

5. Update the mutation strength according to

$$\sigma \leftarrow \sigma \exp\left(\frac{\|\mathbf{s}\|^2 - N}{2DN}\right), \tag{4}$$

where $D$ serves as a damping factor in the adaptation process.

Following recommendations given by Hansen [15], the cumulation parameter $c$ and damping constant $D$ are set to $1/\sqrt{N}$ and $\sqrt{N}$, respectively. It is the goal of cumulative step length adaptation to adapt the mutation strength such that correlations between successive steps are eliminated. The coefficients in Eq. (3) are chosen such that the search path $\mathbf{s}$ consists of standard normally distributed components if selection is random. The strategy used here differs from the original one described in [18] in that in Eq. (4), adaptation is accomplished based on the squared length of the search path rather than on its length. This modification will simplify the analysis in Section 4 without significantly impacting the algorithm's performance.

## 2.2. Some Results on Expected Values of Order Statistics

Let $X_1, X_2, \ldots, X_\lambda$ be a random sample from some univariate probability distribution, and arrange the $X_i$ in nondecreasing order such that $X_{1:\lambda} \leq X_{2:\lambda} \leq \cdots \leq X_{\lambda:\lambda}$. The $k$th smallest of the $X_i$ is denoted by $X_{k:\lambda}$ and referred to as the $k$th order statistic of the sample. See Balakrishnan and Rao [7] for an introduction to the area of order statistics. The following lemma gives an expression for the expected value of the mean of the $\mu$ largest of the $X_i$ for the case that the sample members are independently drawn from a normal distribution.

LEMMA 1. *Let $X_1, X_2, \ldots, X_\lambda$ be $\lambda$ independent, standard normally distributed random variables. Then the expected value of the arithmetic mean of the $(\lambda + 1 - \mu)$th through $\lambda$th order statistics is*

$$\mathrm{E}\left[\frac{1}{\mu}\sum_{k=1}^{\mu} X_{\lambda+1-k:\lambda}\right] = c_{\mu/\mu,\lambda} \tag{5}$$

*where*

$$c_{\mu/\mu,\lambda} = \frac{\lambda-\mu}{2\pi}\binom{\lambda}{\mu}\int_{-\infty}^{\infty} \mathrm{e}^{-x^2}[\Phi(x)]^{\lambda-\mu-1}[1-\Phi(x)]^{\mu-1}\mathrm{d}x$$

*is the $(\mu/\mu, \lambda)$-progress coefficient defined in [11] and where $\Phi(x)$ denotes the cumulative distribution function of the standardised normal distribution.*

See [11] for a derivation of this result. Figure 1 illustrates how the $(\mu/\mu, \lambda)$-progress coefficient depends on the population size parameters $\mu$ and $\lambda$. Lemma 1 will be seen to be useful when there is a direct connection between a random variable characterising a component of a mutation vector and the fitness of the corresponding offspring candidate solution. However, both in the presence of noise and on the ridge function class, that connection is only indirect, and a generalisation of the lemma is required.

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_\lambda, Y_\lambda)$ be a random sample from some bivariate probability distribution. If the sample is ordered by the $X_i$, then the $Y$-variate associated with the $k$th order statistic $X_{k:\lambda}$ is denoted by $Y_{[k:\lambda]}$ and referred to as the concomitant of the $k$th order statistic. See David and Nagaraja [14] for a treatment of concomitants of order statistics. The following lemma gives an expression for the expected value of the arithmetic mean of the concomitants of the $\mu$ largest order statistics for the case that $X = Y + Z$, where both $Y$ and $Z$ are normally distributed.

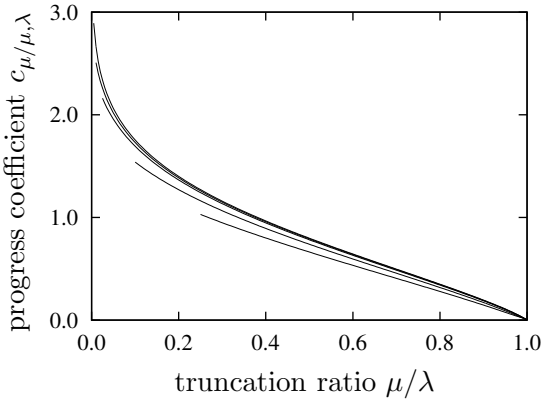


*Figure 1.* Progress coefficients $c_{\mu/\mu,\lambda}$ plotted against the ratio $\mu/\lambda$ for different values of $\lambda$. The curves correspond to, from bottom to top, $\lambda = 4$, 10, 40, 100, and the limit case $\lambda = \infty$. The curves are displayed in the range from $1/\lambda$ to 1. Note that only values $\mu/\lambda$ with integer $\mu$ are of interest.

LEMMA 2. *Let $Y_1, Y_2, \ldots, Y_\lambda$ be $\lambda$ independent, standard normally distributed random variables, and let $Z_1, Z_2, \ldots, Z_\lambda$ be $\lambda$ independent, normally distributed random variables with mean zero and with variance $\vartheta^2$. Then, defining $X_i = Y_i + Z_i$ for $i = 1, \ldots, \lambda$ and ordering the sample members by nondecreasing values of the $X$ variates, the expected value of the arithmetic mean of those $\mu$ of the $Y_i$ with the largest associated values of $X_i$ is*

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

*where the progress coefficient $c_{\mu/\mu,\lambda}$ was defined above.*

The derivation of this result is straightforward using the approach pursued in [3, 6]. The quantity $\vartheta$ is referred to as the noise-to-signal ratio of the selection process.

### 2.3. The Quadratic Sphere Model

Since the early work of Rechenberg [22], the performance of evolution strategies has extensively been studied on the quadratic sphere model with objective function

$$f(\mathbf{x}) = \sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \qquad\qquad \mathbf{x} \in I\!R^N$$

where $\hat{\mathbf{x}}$ is the optimiser and the task is minimisation. See [4] for a discussion of the usefulness of such considerations, and see [11] for

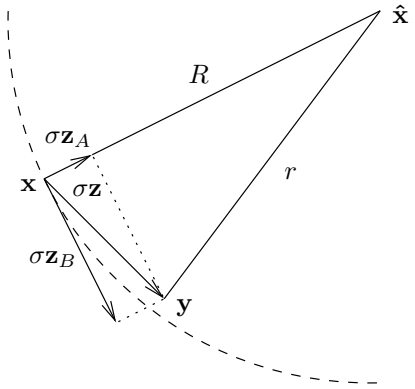


*Figure 2.* Decomposition of a vector $\mathbf{z}$ into central component $\mathbf{z}_A$ and lateral component $\mathbf{z}_B$. Vector $\mathbf{z}_A$ is parallel to $\hat{\mathbf{x}}-\mathbf{x}$, vector $\mathbf{z}_B$ is in the hyperplane perpendicular to that. The starting and end points, $\mathbf{x}$ and $\mathbf{y} = \mathbf{x}+\sigma\mathbf{z}$, of vector $\sigma\mathbf{z}$ are at distances $R$ and $r$ from the optimiser $\hat{\mathbf{x}}$, respectively.

comprehensive results and techniques of analysis for a great number of strategy variants.

Central to the quantitative characterisation of the performance of evolution strategies on the sphere model is the computation of the difference in fitness $f(\mathbf{x}) - f(\mathbf{x}+\sigma\mathbf{z})$ between candidate solutions (or search points) $\mathbf{x}$ and $\mathbf{y} = \mathbf{x} + \sigma\mathbf{z}$. That difference is referred to as the fitness advantage associated with vector $\mathbf{z}$. Depending on the context, $\mathbf{z}$ can be either a mutation vector or a progress vector. The commonly used approach to computing the fitness advantage relies on a decomposition of vector $\mathbf{z}$ that is illustrated in Fig. 2, where $R = \|\hat{\mathbf{x}}-\mathbf{x}\|$ and $r = \|\hat{\mathbf{x}}-\mathbf{y}\|$ are the distances of $\mathbf{x}$ and $\mathbf{y}$ from the optimiser. Using "$\cdot$" to denote the dot product, vector $\mathbf{z}$ can be written as the sum of two orthogonal vectors $\mathbf{z}_A$ and $\mathbf{z}_B$, where

$$\mathbf{z}_A = \frac{(\hat{\mathbf{x}}-\mathbf{x})\cdot\mathbf{z}}{R^2}(\hat{\mathbf{x}}-\mathbf{x})$$

is parallel to $\hat{\mathbf{x}}-\mathbf{x}$ and

$$\mathbf{z}_B = \mathbf{z} - \mathbf{z}_A$$

is in the $(N-1)$-dimensional hyperplane perpendicular to that. The vectors $\mathbf{z}_A$ and $\mathbf{z}_B$ are referred to as the central and lateral components of vector $\mathbf{z}$, respectively. The signed length

$$z_A = \frac{(\hat{\mathbf{x}}-\mathbf{x})\cdot\mathbf{z}}{R} \tag{6}$$

of the central component of vector $\mathbf{z}$ equals $\|\mathbf{z}_A\|$ if $\mathbf{z}_A$ points towards the optimiser and it equals $-\|\mathbf{z}_A\|$ if $\mathbf{z}_A$ points away from it.

Using elementary geometry, it can easily be seen from Fig. 2 that

$$r^2 = (R - \sigma z_A)^2 + \sigma^2 \|\mathbf{z}_B\|^2,$$

and therefore, rearranging terms and realising that $\|\mathbf{z}\|^2 = z_A^2 + \|\mathbf{z}_B\|^2$, that

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{x} + \sigma\mathbf{z}) &= R^2 - r^2 \\ &= 2R\sigma z_A - \sigma^2 \|\mathbf{z}\|^2. \end{aligned} \tag{7}$$

That is, the fitness advantage associated with vector $\mathbf{z}$ has two contributions: one that can be either positive or negative, depending on the signed length of the central component of that vector, and one that is always negative. It is of course desirable to achieve a positive fitness advantage associated with the progress vector in order to decrease the distance from the optimiser. Selection of good candidate solutions needs to ensure that the signed length of the central component of the progress vector is positive, and the mutation strength needs to be such that the negative contribution of the second term is outweighed.

Computation of the expected fitness advantage associated with the progress vector is best accomplished in the limit of infinite search space dimensionality. Computational experiments in finite-dimensional search spaces are then used to verify the accuracy of the predictions for finite $N$. Improved approximations for finite $N$ can often be derived, but are not subject of this paper.

Recall that the components of a mutation vector $\mathbf{z}$ are independently standard normally distributed. The squared length $\|\mathbf{z}\|^2$ of mutation vectors is thus $\chi^2$-distributed with $N$ degrees of freedom. Due to the properties of the $\chi^2$-distribution, in the limit $N \to \infty$, $\|\mathbf{z}\|^2/N$ asymptotically approaches 1 (its mean is 1 while its variance is $\mathcal{O}(1/N)$). As a consequence, as detailed in [11], in the range of mutation strengths where the fitness advantage associated with the progress vector is positive (and thus in the range of interest), variations in the negative contributions to the fitness advantage described by the second term on the right hand side of Eq. (7) can be ignored. The $k$th best offspring candidate solution is that with the $k$th largest signed length of the central component of the mutation vector that generated it. Thus, the signed length $z_A$ of the central component of the mutation vector that generates the $k$th best offspring candidate solution is the $(\lambda - 1 + k)$th order statistic in a sample of $\lambda$ independent, standard normally distributed random variates.

According to Eq. (1) the progress vector is the arithmetic mean of the $\mu$ best mutation vectors. It follows that the expected value of the signed length of its central component equals the expected value of the

arithmetic mean of the $(\lambda+1-\mu)$th through $\lambda$th order statistics. Letting $X_i = z_A^{(i)}$, Lemma 1 from Section 2.2 is thus immediately applicable and the expected signed length of the central component of the progress vector is $c_{\mu/\mu,\lambda}$. Moreover, it has been shown in [8] that

$$\frac{\|\mathbf{z}^{(\mathrm{avg})}\|^2}{N} \overset{N\to\infty}{=} \frac{1}{\mu}. \tag{8}$$

The reduction in the squared length by a factor of $1/\mu$ compared to that of the mutation vectors being averaged results from the fact that the lateral components of the latter have no influence on the fitness of the candidate solutions and are thus uncorrelated. Averaging $\mu$ uncorrelated random vectors of squared length 1 yields a random vector of squared length $1/\mu$. The averaging is beneficial as the negative term contributing to the fitness advantage of the progress vector is reduced. In [8], that reduction has been termed the genetic repair effect.

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. As discussed in [1, 3], most frequently, noise is modelled as an additive Gaussian term with mean zero and with a standard deviation $\sigma_\epsilon$ that is referred to as the noise strength. The noisy objective function value of candidate solution $\mathbf{y} = \mathbf{x} + \sigma\mathbf{z}$ then is

$$f_\epsilon(\mathbf{y}) = f(\mathbf{x}) - 2R\sigma z_A + \sigma^2\|\mathbf{z}\|^2 + \sigma_\epsilon z_\epsilon$$

where $z_\epsilon$ is standard normally distributed and where the index in $f_\epsilon$ indicates the measurement of the fitness value is disturbed by noise. Variations in the third term on the right hand side again lose significance as $N \to \infty$. As both the second and fourth terms are normally distributed, so is the noisy fitness advantage associated with vector $\mathbf{z}$.

Noise has no influence on the squared length of the progress vector. However, it does have an influence on the signed length of that vector's central component. The candidate solutions selected to survive are those with the largest values of $2R\sigma z_A - \sigma_\epsilon z_\epsilon$. The signed lengths of the central components of the mutation vectors are thus concomitants of the order statistics that result from ranking the offspring candidate solutions by their noisy objective function values. Letting $Y_i = z_A^{(i)}$ and $Z_i = -(\sigma_\epsilon/2R\sigma)z_\epsilon^{(i)}$, Lemma 2 from Section 2.2 is applicable and the expected signed length of the central component of the progress vector is

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

where $\vartheta = \sigma_\epsilon / 2R\sigma$ is the noise-to-signal ratio that the strategy operates under. Notice that in general, the noise strength need not be constant, but instead it may vary across the search space.

## 3. The $(\mu/\mu, \lambda)$-ES on the Noisy Parabolic Ridge

In this section the performance of the $(\mu/\mu, \lambda)$-ES with isotropic mutations is studied on the parabolic ridge. The treatment of cumulative step length adaptation is deferred until Section 4. The results presented here generalise those derived in [19] by considering noise in the analysis. In contrast to that reference, no attempt is made to include $N$-dependent terms in the calculations. Numerical experiments are used to illustrate that the accuracy of the results is good provided that $N$ is sufficiently large.

### 3.1. Expected Progress Vector

Even though the parabolic ridge described by objective function

$$f(\mathbf{x}) = x_1 - \frac{d}{N} \sum_{i=2}^{N} x_i^2 \qquad \mathbf{x} \in I\!R^N,\ d > 0 \tag{9}$$

has no finite optimum, maximisation is still a meaningful task if progress along the ridge axis (i.e., in the $x_1$-direction) is considered as a performance measure. It is also worth pointing out that while in the definition used here the ridge axis is aligned with an axis of the coordinate system, that fact is irrelevant for a strategy that uses isotropic mutations such as those considered in the present paper. The coordinate system could be subjected to an arbitrary rigid transformation without affecting the strategy's performance.

Clearly, the parabolic ridge contains within it an $(N-1)$-dimensional quadratic sphere. Similar to the decomposition of vectors described in Section 2.3, a mutation or progress vector $\mathbf{z} = (z_1, z_2, \ldots, z_N)^{\mathrm{T}}$ on the ridge can be written as the sum of three mutually orthogonal vectors that are straightforward to obtain. Let $\mathbf{z}_1 = (z_1, 0, \ldots, 0)^{\mathrm{T}}$ and $\mathbf{z}_{2\ldots N} = (0, z_2, \ldots, z_N)^{\mathrm{T}}$ denote the projections of $\mathbf{z}$ onto the hyperspaces with $z_2 = \cdots = z_N = 0$ and $z_1 = 0$, respectively. Furthermore, decompose $\mathbf{z}_{2\ldots N}$ into vectors $\mathbf{z}_A$ and $\mathbf{z}_B$ as done in Section 2.3 with $(x_1, 0, \ldots, 0)^{\mathrm{T}}$ for $\hat{\mathbf{x}}$. Then $\mathbf{z} = \mathbf{z}_1 + \mathbf{z}_A + \mathbf{z}_B$ is a decomposition of $\mathbf{z}$ into mutually orthogonal components and $\mathbf{z}_1$, $\mathbf{z}_A$, and $\mathbf{z}_B$ are referred to as the axial, central, and lateral components of $\mathbf{z}$, respectively. See Fig. 3 for an illustration.

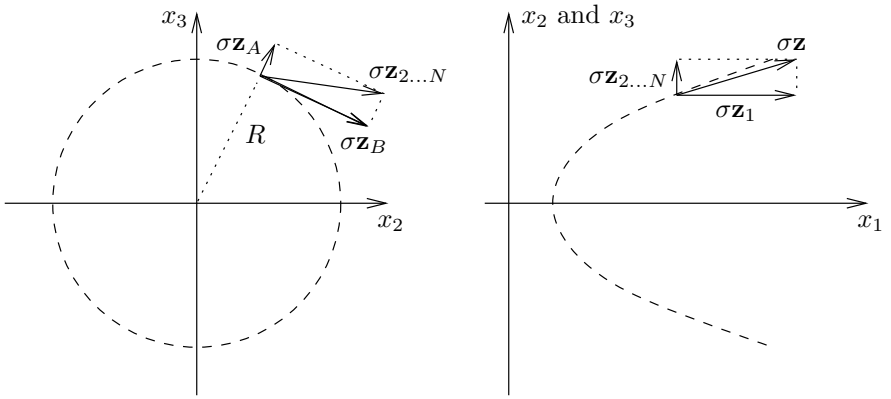


*Figure 3.* Decomposition of vector $\mathbf{z}$ into its axial component $\mathbf{z}_1$, central component $\mathbf{z}_A$, and lateral component $\mathbf{z}_B$ for $N = 3$. The dashed lines indicate locations of constant fitness.

Using the decomposition along with Eqs. (6) and (9), it follows that the fitness of candidate solution $\mathbf{y} = \mathbf{x} + \sigma\mathbf{z}$ is

$$\begin{aligned} f(\mathbf{y}) &= x_1 + \sigma z_1 - \frac{d}{N}\sum_{i=2}^{N}(x_i + \sigma z_i)^2 \\ &= x_1 - \frac{d}{N}\sum_{i=2}^{N} x_i^2 + \sigma z_1 - \frac{d}{N}\left(2\sigma\sum_{i=2}^{N} x_i z_i + \sigma^2\sum_{i=2}^{N} z_i^2\right) \\ &= f(\mathbf{x}) + \sigma z_1 + \frac{d}{N}\left(2\sigma R z_A - \sigma^2\|\mathbf{z}_{2\ldots N}\|^2\right) \end{aligned} \tag{10}$$

where $R = \|\mathbf{x}_{2\ldots N}\|$ denotes the distance of the search point from the ridge axis. If $\mathbf{z}$ is a mutation vector then $\|\mathbf{z}_{2\ldots N}\|^2/N \overset{N\to\infty}{=} 1$ holds as seen in the discussion of the sphere model in Section 2.3. Introducing the standardised distance from the ridge axis

$$\rho = \frac{2Rd}{N}$$

the noisy fitness of candidate solution $\mathbf{y}$ is thus

$$f_\epsilon(\mathbf{y}) \overset{N\to\infty}{=} f(\mathbf{x}) + \sigma z_1 + \rho\sigma z_A - \sigma^2 d + \sigma_\epsilon z_\epsilon \tag{11}$$

where $z_\epsilon$ is a standard normally distributed random variable reflecting the noise present in the evaluation process and where $\sigma_\epsilon$ denotes the noise strength.

For the purpose of selection, offspring candidate solutions are ranked according their noisy fitness values. According to Eq. (1), the mutation

vectors of those $\mu$ of the offspring with the highest noisy fitness values are averaged arithmetically to form the progress vector. The axial, central, and lateral components of the progress vector are thus the arithmetic means of the respective components of the selected mutation vectors. The signed lengths of the axial and central components of the mutation vectors are standard normally distributed. As the $k$th best offspring candidate solution is that with the $k$th largest value of $\sigma z_1 + \rho\sigma z_A + \sigma_\epsilon z_\epsilon$ (the other terms in Eq. (11) are identical for all offspring), the signed lengths of both the axial components $z_1$ and the central components $z_A$ of the mutation vectors are concomitants of the order statistics that result from ranking candidate solutions according to their noisy fitness. Moreover, as all random variables in Eq. (11) are normally distributed, and as the sum of two normally distributed random variables is again normally distributed, Lemma 2 from Section 2.2 is applicable. In particular, letting $Y_i = z_1^{(i)}$ and $Z_i = (\rho\sigma z_A^{(i)} + \sigma_\epsilon z_\epsilon^{(i)})/\sigma$, it follows from Lemma 2 that the expected value of the signed length of the axial component of the progress vector is

$$\mathrm{E}\left[z_1^{(\mathrm{avg})}\right] \overset{N\to\infty}{=} \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2+\rho^2}} \tag{12}$$

where $\vartheta = \sigma_\epsilon/\sigma$ denotes the noise-to-signal ratio that the strategy operates under. Similarly, letting $Y_i = z_A^{(i)}$ and $Z_i = (\sigma z_1^{(i)} + \sigma_\epsilon z_\epsilon^{(i)})/\rho\sigma$, it follows from Lemma 2 that the expected value of the signed length of the central component of the progress vector is

$$\begin{aligned}\mathrm{E}\left[z_A^{(\mathrm{avg})}\right] &\overset{N\to\infty}{=} \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+(\sigma^2+\sigma_\epsilon^2)/\rho^2\sigma^2}} \\ &= \frac{\rho c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2+\rho^2}}.\end{aligned} \tag{13}$$

Finally, for the squared length of the combined central and lateral components of the progress vector,

$$\frac{\|\mathbf{z}_{2\ldots N}^{(\mathrm{avg})}\|^2}{N} \overset{N\to\infty}{=} \frac{1}{\mu} \tag{14}$$

holds in analogy to the corresponding result in Eq. (8) for the sphere model.

### 3.2. Distance from the Ridge Axis and Progress Rate

The expected values of the signed lengths of the axial and central components of the progress vector computed in Eqs. (12) and (13) depend on the standardised distance $\rho$ of the search point from the ridge axis.

In the case that the mutation strength is constant, that distance varies with time and either has a time-invariant limit distribution or diverges to $\infty$. Steps 1, 2, and 3 of the algorithm outlined in Section 2.1 define an iterated stochastic mapping

$$\begin{aligned} {R^{(t+1)}}^2 &= \sum_{i=2}^{N} \left( x_i + \sigma z_i^{(\mathrm{avg})} \right)^2 \\ &= {R^{(t)}}^2 - 2R^{(t)} \sigma z_A^{(\mathrm{avg})} + \sigma^2 \| \mathbf{z}_{2\ldots N}^{(\mathrm{avg})} \|^2 \end{aligned}$$

of distances from the ridge axis, where Eq. (6) has been used and where superscripts indicate time. Multiplying by $4d^2/N^2$ in order to switch to standardised distances yields evolution rule

$$ {\rho^{(t+1)}}^2 = {\rho^{(t)}}^2 - \frac{4d}{N} \left( \rho^{(t)} \sigma z_A^{(\mathrm{avg})} - \frac{\sigma^2 d}{N} \| \mathbf{z}_{2\ldots N}^{(\mathrm{avg})} \|^2 \right) \tag{15} $$

for the dynamical system. Consider the case that $\rho$ does not diverge. In that case, iterating Eq. (15), the squared standardised distance to the ridge axis tends towards and then fluctuates around a stationary limit value. For given $\sigma$, $\sigma_\epsilon$, $\rho$, and $\lambda$, both the mean and the variance of the term in parentheses are in $\mathcal{O}(1)$. Due to the presence of the factor $4d/N$ that that term is multiplied with, for given mutation and noise strengths fluctuations are of order $\mathcal{O}(1/N)$ and thus decrease with increasing search space dimensionality. In the limit case $N \to \infty$, variances vanish altogether and all random variables can be replaced by their expected values. Using Eqs. (13) and (14) for the expected values of $z_A^{(\mathrm{avg})}$ and $\| \mathbf{z}_{2\ldots N}^{(\mathrm{avg})} \|^2$ and demanding that $\rho^{(t+1)} = \rho^{(t)}$, Eq. (15) yields

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

Introducing normalised quantities

$$ \sigma^* = \frac{\sigma d}{\mu c_{\mu/\mu,\lambda}} \qquad \text{and} \qquad \sigma_\epsilon^* = \frac{\sigma_\epsilon d}{\mu c_{\mu/\mu,\lambda}}, $$

squaring, and rearranging terms yields the equivalent condition

$$ \rho^4 - {\sigma^*}^2 \left( 1 + \rho^2 \right) - {\sigma_\epsilon^*}^2 = 0 \tag{16} $$

that can be used to obtain the stationary standardised distance from the ridge axis. The following three subsections consider three different types of dependency of the noise strength on the distance from the ridge axis. More specifically, the cases that the noise strength is uniform

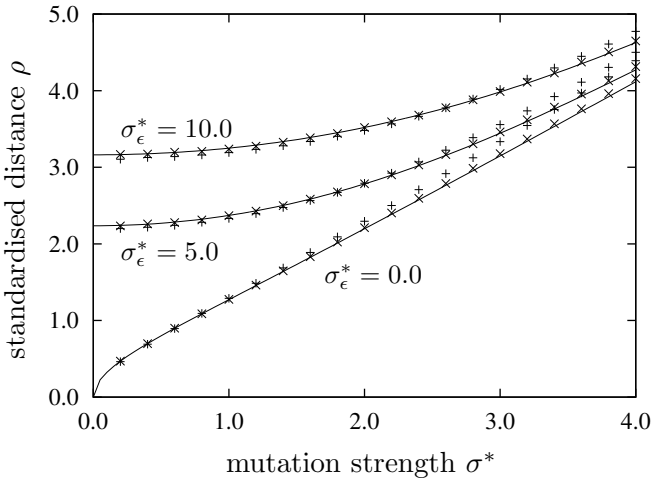


*Figure 4.* Standardised distance $\rho$ from the ridge axis plotted against normalised mutation strength $\sigma^*$ for the case of uniform noise strength. The solid lines have been obtained from Eq. (17). The points represent results measured in runs of the $(\mu/\mu, \lambda)$-ES on the parabolic ridge in search spaces with $N = 40$ (+) and $N = 400$ (×). In all cases, $\mu = 3$, $\lambda = 10$, and $d = 1.0$.

throughout the search space, that it increases quadratically with the distance from the ridge axis, and that the rate of increase is cubic are examined. The cases have been chosen as they can be handled analytically while at the same time exhibiting qualitatively different characteristics and posing widely differing demands to the step length control mechanism to be discussed in Section 4.

3.2.1. *Uniform Noise Strength*
Consider the case that the noise strength $\sigma_\epsilon$ is uniform throughout the search space. Solving Eq. (16) yields

$$\rho^2 \overset{N\to\infty}{=} \frac{\sigma^{*2}}{2} + \sqrt{\frac{\sigma^{*4}}{4} + \sigma^{*2} + \sigma_\epsilon^{*2}} \tag{17}$$

for the squared standardised distance from the ridge axis. Figure 4 illustrates how the accuracy of predictions made using Eq. (17) improves with increasing values of $N$ by comparing with values measured in runs of evolution strategies. Not shown here, greater values of $\mu$ and $\lambda$ generally require greater values of $N$ in order to achieve the same degree of accuracy. The quality of the approximation is largely independent of the strength of the noise present.

With the knowledge of the standardised distance from the ridge axis thus obtained, the performance of the $(\mu/\mu, \lambda)$-ES on the parabolic

ridge can now be quantified. Following [19, 20, 21] in defining the progress rate[1]

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

of the strategy as the expected distance in direction of the ridge axis that the search point moves per time step, it follows from Eq. (12) that

$$\varphi \overset{N\to\infty}{=} \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2+\rho^2}}. \tag{18}$$

Introducing normalisation

$$\varphi^* = \frac{\varphi d}{\mu c_{\mu/\mu,\lambda}^2}$$

and using Eq. (17) for the squared standardised distance from the ridge axis, the normalised progress rate of the $(\mu/\mu,\lambda)$-ES on the noisy parabolic ridge is

$$\begin{aligned}
\varphi^* &\overset{N\to\infty}{=} \frac{\sigma^*}{\sqrt{1+\vartheta^2+\rho^2}} \\
&\overset{N\to\infty}{=} \frac{\sigma^*}{\sqrt{1+\sigma_\epsilon^{*2}/\sigma^{*2}+\sigma^{*2}/2+\sqrt{\sigma^{*4}/4+\sigma^{*2}+\sigma_\epsilon^{*2}}}} \\
&= \frac{\sigma^{*2}}{\sigma^{*2}/2+\sqrt{\sigma^{*4}/4+\sigma^{*2}+\sigma_\epsilon^{*2}}}.
\end{aligned} \tag{19}$$

The last step can easily be verified by squaring the denominator in the last row. Figure 5 illustrates how the accuracy of predictions made using Eq. (19) improves with increasing values of $N$ by comparing with measurements made in runs of evolution strategies on the parabolic ridge. While for $N = 40$, substantial deviations between predictions and measured values exist, for larger values of $\sigma^*$ the agreement is generally good for $N = 400$.

As found in [19], in the absence of noise the progress rate of the $(\mu/\mu,\lambda)$-ES on the parabolic ridge monotonically increases with increasing mutation strength. It can be seen from Eq. (18) as well as from Fig. 5 that the same holds true in the presence of uniform noise.

[1] Alternatively, it is possible to use the quality gain, i.e., the expected change in the fitness value of the search point from one time step to the next, as a performance measure. In the stationary limit state, the (squared) distance from the ridge axis is unchanged in the mean, and the only contribution to the change in fitness values stems from the $x_1$-component. In that state, progress rate and quality gain thus agree in the limit $N \to \infty$.

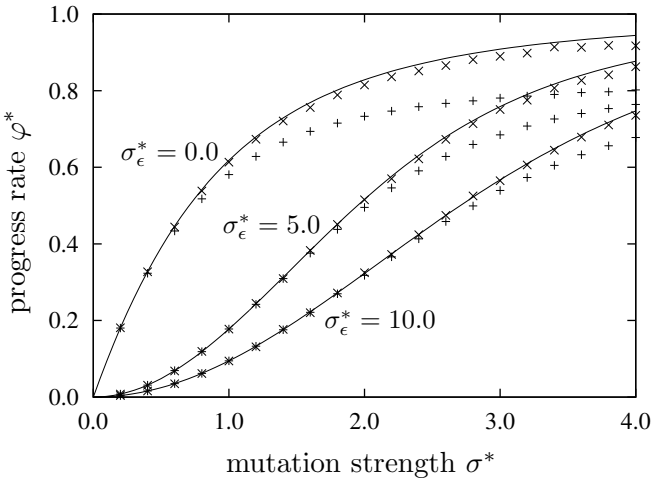


*Figure 5.* Normalised progress rate $\varphi^*$ plotted against normalised mutation strength $\sigma^*$ for the case of uniform noise strength. The solid lines have been obtained from Eq. (19). The points represent results measured in runs of the $(\mu/\mu, \lambda)$-ES on the parabolic ridge in search spaces with $N = 40$ (+) and $N = 400$ (×). In all cases, $\mu = 3$, $\lambda = 10$, and $d = 1.0$.

In addition to its beneficial effect for $\sigma_\epsilon = 0$, increasing $\sigma$ reduces the noise-to-signal ratio $\vartheta = \sigma_\epsilon/\sigma$ that the strategy operates under. For large mutation strengths that ratio tends to zero and according to Eq. (19) the effects of noise on the progress rate vanish.

From Eqs. (17) and (19) and undoing the normalisations, for large mutation strengths the strategy operates at a standardised distance of close to $\rho = \sigma d/\mu c_{\mu/\mu,\lambda}$ from the ridge axis and achieves a progress rate of nearly $\varphi_{\max} = \mu c^2_{\mu/\mu,\lambda}/d$. Increasing the population size parameters $\mu$ and $\lambda$ thus decreases the stationary standardised distance $\rho$ from the ridge axis and increases the optimal progress rate $\varphi_{\max}$. Similar to what has been found on the quadratic sphere [8, 11], a roughly linear increase in $\varphi_{\max}$ can be achieved as a result of increasing $\lambda$ provided that $\mu$ is increased such that the ratio $\mu/\lambda$ remains unchanged. For large values of $\lambda$, a setting of $\mu = 0.270\lambda$ is optimal as it maximises $\mu c^2_{\mu/\mu,\lambda}$. Larger values of $\lambda$ are beneficial and result in a linear speed-up if offspring candidate solutions can be evaluated in parallel. Notice however that Eq. (19) holds only in the limit $N \to \infty$. In finite-dimensional search spaces, the rate of increase of $\varphi_{\max}$ is sublinear in $\mu$ and $\lambda$, and the serial performance of the $(\mu/\mu, \lambda)$-ES suffers once the number of offspring generated per time step is too large. It is not possible to derive quantitative recommendations with regard to the choice of $\lambda$ from the

analysis presented here as all $N$-dependent terms have been left out of the calculations.

3.2.2. *Quadratic Noise Strength*
Next, consider the case that the noise strength increases quadratically with the distance from the ridge axis, i.e. that $\sigma_\epsilon = \zeta\rho^2$ for some $\zeta \geq 0$. We refer to $\zeta$ as the noise level. Equation (16) then reads

$$\rho^4 - \sigma^{*2}\left(1+\rho^2\right) - \zeta^{*2}\rho^4 = 0$$

where $\zeta^* = \zeta d/\mu c_{\mu/\mu,\lambda}$ is the normalised noise level. Solving for the square of the standardised distance from the ridge axis yields

$$\rho^2 \overset{N\to\infty}{=} \frac{\sigma^{*2}}{2(1-\zeta^{*2})} + \sqrt{\frac{\sigma^{*4}}{4(1-\zeta^{*2})^2} + \frac{\sigma^{*2}}{1-\zeta^{*2}}}. \tag{20}$$

For $\zeta^* \geq 1$, there is no real-valued solution and the strategy fails to track the ridge for any nonzero value of the mutation strength. In that case, the distance to the ridge axis diverges to $\infty$ and the resulting progress rate (but not the quality gain!) approaches zero. If $\zeta^* < 1$, then the resulting normalised progress rate is in close analogy to Eq. (19)

$$\begin{aligned}\varphi^* &\overset{N\to\infty}{=} \frac{\sigma^*}{\sqrt{1+\zeta^{*2}\rho^4/\sigma^{*2}+\rho^2}} \\ &\overset{N\to\infty}{=} \frac{\sigma^*(1-\zeta^{*2})}{\sigma^*/2+\sqrt{\sigma^{*2}/4+1-\zeta^{*2}}}.\end{aligned} \tag{21}$$

The dependence of the stationary distance $\rho$ from the ridge axis and of the normalised progress rate $\varphi^*$ on $\zeta^*$ are shown in Figs. 6 and 7. It can be seen that while the accuracy of the predictions is quite good for $N = 400$, the deviations in the lower dimensional search space with $N = 40$ are considerable. In contrast to the uniform noise case, the quality of the approximation also deteriorates with increasing levels of noise present. The influence of the $N$-dependent terms cannot be neglected and better approximations remain to be derived in future work. It is also worth noting that while Eqs. (20) and (21) suggest that increasing $\mu$ and $\lambda$ makes it possible to deal with any amount of noise present (by driving $\zeta^*$ to zero), this is an idealisation for $N \to \infty$ that does not hold for finite $N$.

It can be seen from Figs. 6 and 7 that as in the case of uniform noise strength, increasing the mutation strength is beneficial as it leads to an increase in progress rate. However, unlike in the situation where the noise strength is uniform, in the quadratic case it is not possible

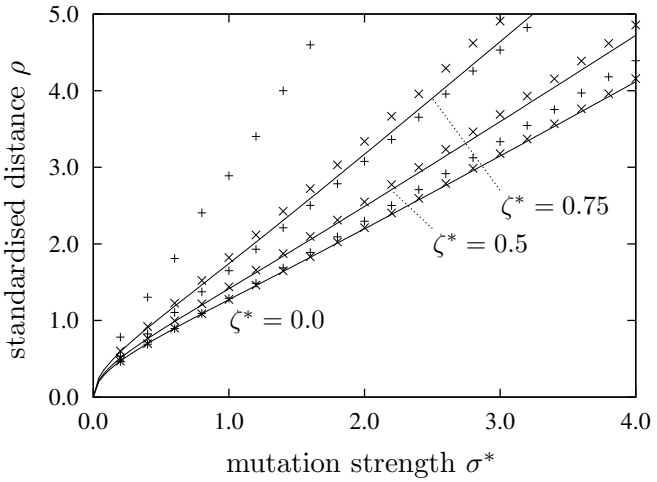


*Figure 6.* Standardised distance $\rho$ from the ridge axis plotted against normalised mutation strength $\sigma^*$ for the case that the noise strength increases quadratically with the distance from the ridge axis. The solid lines have been obtained from Eq. (20). The points represent results measured in runs of the $(\mu/\mu, \lambda)$-ES on the parabolic ridge in search spaces with $N = 40$ (+) and $N = 400$ (×). In all cases, $\mu = 3$, $\lambda = 10$, and $d = 1.0$.

to always achieve the same progress rate as in the absence of noise. In the uniform noise case, increasing the mutation strength results in a higher signal strength that serves to drive the noise-to-signal ratio to zero. In the case that the noise strength increases quadratically with the distance from the ridge axis however, the increased distance at which the ridge axis is tracked as a result of increasing $\sigma$ also results in an increase in the noise strength that leads to the noise-to-signal ratio tend to a non-zero limit value. More specifically, for large mutation strengths, the first term in the radicand in Eq. (20) dominates the second and the squared stationary standardised distance from the ridge axis is $\rho^2 = \sigma^{*2}/(1+\zeta^{*2})$. The corresponding normalised progress rate is $\varphi^*_{\max} = 1 - \zeta^{*2}$ and thus decreases with increasing $\zeta^*$. While the exact limit value is not well described by Eq. (21) unless $N$ is very large, the qualitative behaviour of $\rho^2$ and $\varphi^*$ is captured correctly by Eqs. (20) and (21).

### 3.2.3. *Cubic Noise Strength*

Finally, consider the case that the noise strength increases cubically with the distance from the ridge axis, i.e. that $\sigma_\epsilon = \zeta\rho^3$. Equation (16) then reads

$$\rho^4 - \sigma^{*2}\left(1+\rho^2\right) - \zeta^{*2}\rho^6 = 0 \tag{22}$$

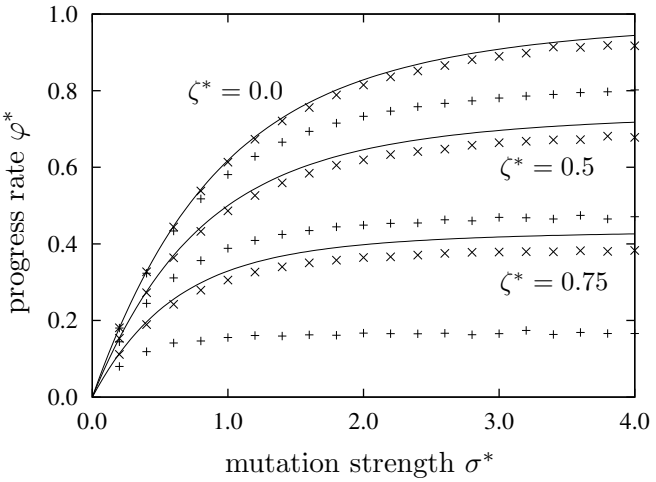


*Figure 7.* Normalised progress rate $\varphi^*$ plotted against normalised mutation strength $\sigma^*$ for the case that the noise strength increases quadratically with the distance from the ridge axis. The solid lines have been obtained from Eq. (21). The points represent results measured in runs of the $(\mu/\mu, \lambda)$-ES on the parabolic ridge in search spaces with $N = 40$ (+) and $N = 400$ (×). In all cases, $\mu = 3$, $\lambda = 10$, and $d = 1.0$.

and is thus no longer quadratic in $\rho^2$ but cubic instead. For a given value of $\zeta^*$ and small $\sigma^*$, Eq. (22) has two nonnegative real roots, the smaller of which corresponds to a stable fixed point of the cubic polynomial in Eq. (15). As $\sigma^*$ increases, the value of that fixed point grows until it coincides with the larger, unstable fixed point and subsequently disappears. The resulting effect on the standardised distance from the ridge axis is illustrated in Fig. 8. The solid lines in that figure have been obtained by numerically finding the root of Eq. (22) that corresponds to the stable fixed point in the mapping of squared standardised distances from the ridge axis. The line corresponding to $\zeta^* = 0.2$ abruptly ends at $\sigma^* \approx 2.4$ as the stable fixed point disappears. With mutation strengths beyond this point, no stable limit state exists and the distance from the ridge axis diverges to $\infty$. For $\zeta^* = 0.1$, the point where the stable limit state ceases to exist is beyond the range of mutation strengths shown. The measurements from runs of evolution strategies that are included in the figure show that the behaviour can indeed be observed in practice.

While an explicit solution for the roots of the cubic equation exists, it is complicated and does not yield new insights. However, an upper bound on values of $\sigma^*$ that allow tracking the ridge can easily be derived. The two roots of the cubic polynomial in Eq. (22) are separated by a local maximum at $z = (1 + \sqrt{1 - 3\sigma^{*2}\zeta^{*2}})/3\zeta^{*2}$. That maximum

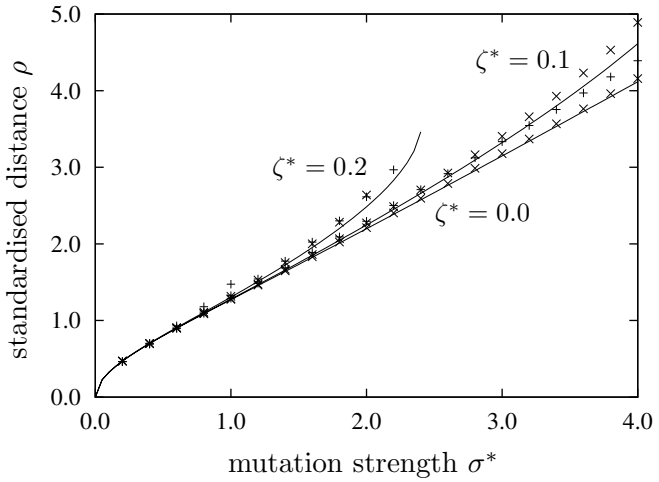


*Figure 8.* Standardised distance $\rho$ from the ridge axis plotted against normalised mutation strength $\sigma^*$ for the case that the noise strength increases cubically with the distance from the ridge axis. The solid lines have been found by numerically finding the stable root of Eq. (22). The points represent results measured in runs of the $(\mu/\mu, \lambda)$-ES on the parabolic ridge in search spaces with $N = 40$ (+) and $N = 400$ (×). In all cases, $\mu = 3$, $\lambda = 10$, and $d = 1.0$.

and with it the two roots do not exist if $1 - 3\sigma^{*2}\zeta^{*2} < 0$, making

$$\sigma^* \leq \frac{1}{\sqrt{3}\zeta^*}$$

a necessary (though not sufficient) condition for being able to track the ridge at a finite distance.

The case of noise that increases cubically with the distance from the ridge axis is interesting as it presents a situation that is qualitatively different from those considered so far. If the noise strength increases superquadratically with $\rho$, then the strategy is forced to track the ridge more closely than it would in the cases considered above and thus cannot use arbitrarily large mutation strengths. The larger the level of noise present, the smaller the mutation strength needs to be in order to be able to track the ridge. Figure 9 illustrates the dependence of the progress rate on the mutation strength for several values of $\zeta^*$. It can be seen that for $\zeta^* \neq 0$, increasing the mutation strength is beneficial up to some point. Beyond that point, the progress rate starts to decline before the strategy abruptly starts to lose its ability to track the ridge at all. As for the case of quadratic noise, the accuracy of the predictions is quite good for $N = 400$. For $N = 40$ it is merely the qualitative dependence that is described correctly, and $N$-dependent terms will need to be taken into account in order to derive recommendations with

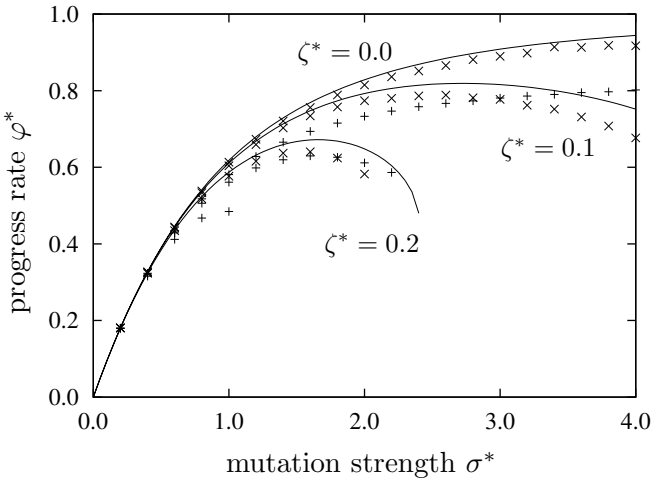


*Figure 9.* Normalised progress rate $\varphi^*$ plotted against normalised mutation strength $\sigma^*$ for the case that the noise strength increases cubically with the distance from the ridge axis. The solid lines have been obtained by inserting the numerically obtained stable root of Eq. (22) into Eq. (18). The points represent results measured in runs of the $(\mu/\mu, \lambda)$-ES on the parabolic ridge in search spaces with $N = 40$ (+) and $N = 400$ (×). In all cases, $\mu = 3$, $\lambda = 10$, and $d = 1.0$.

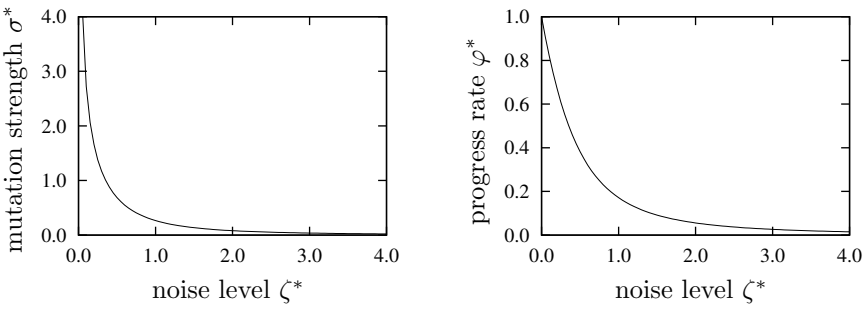


*Figure 10.* Optimal normalised mutation strength $\sigma^*$ and resulting normalised progress rate $\varphi^*$ plotted against the normalised noise level $\zeta^*$ for the case of noise that increases cubically with the distance from the ridge axis. The graphs have been obtained numerically using Eqs. (18) and (22).

regard to the choice of $\lambda$ in finite-dimensional search spaces. Finally, Fig. 10 illustrates the dependence of the optimal mutation strength and the resulting progress rate on $\zeta^*$. That graphs in that figure have been obtained by numerically optimising Eq. (18) using the stable fixed point of Eq. (22) for $\rho^2$. It can be seen that both the optimal mutation strength and the resulting progress rate decrease rapidly with increasing levels of noise present.

## 4. Cumulative Step Length Adaptation

The calculations presented in Section 3 have considered the mutation strength as a constant, exogenous quantity. In contrast, the algorithm outlined in Section 2.1 adapts the mutation strength using the cumulative step length adaptation mechanism [18]. Cumulative step length adaptation on the noisy sphere model has been studied in [1, 5]. This section presents an analysis of the behaviour of cumulative step length adaptation of the noisy parabolic ridge. The analysis allows handling all three forms of noise considered above.

### 4.1. Accumulated Search Path

It has been seen in Section 3.1 that the behaviour of the $(\mu/\mu, \lambda)$-ES with static step length is described on the noisy parabolic ridge by an iterated stochastic mapping with the (standardised) distance from the ridge axis as its only state variable and with Eq. (15) as its evolution rule. Using cumulative step length adaptation introduces as further state variables the axial and central lengths $s_1$ and $s_A$ of the accumulated progress vector, the squared length $\|\mathbf{s}\|^2$ of that vector, and the mutation strength $\sigma$. That is, the algorithm described in Section 2.1 defines a stochastic mapping

$$(\rho^{(t)}, s_1^{(t)}, s_A^{(t)}, \|\mathbf{s}^{(t)}\|^2, \sigma^{(t)}) \mapsto (\rho^{(t+1)}, s_1^{(t+1)}, s_A^{(t+1)}, \|\mathbf{s}^{(t+1)}\|^2, \sigma^{(t+1)})$$

that determines the behaviour of the strategy. The exact form of the mapping can be inferred from Eqs. (3) and (4). The approach to computing stationary values is the same as that used in [1, 5] for the sphere model and in Section 3 for the case of static step length: replace all quantities by their mean values and demand stationarity in that no change occurs between time steps $t$ and $t + 1$. Any terms that vanish in the limit $N \to \infty$ are dropped from the calculations. The approach yields useful approximations for large enough $N$ as it can be observed that fluctuations (quantified by the variation coefficients of the state variables) decrease with increasing search space dimensionality. As in Section 3, computer experiments will be used to verify the quality of the approximations.

From Eq. (3), the evolution rule for the signed length of the axial component of the accumulated progress vector reads

$$s_1^{(t+1)} = (1 - c)s_1^{(t)} + \sqrt{\mu c(2 - c)}z_1^{(\mathrm{avg})}.$$

Demanding stationarity yields

$$s_1 \overset{N\to\infty}{=} \sqrt{\frac{\mu(2 - c)}{c}}z_1^{(\mathrm{avg})} \tag{23}$$

for the mean value of that quantity, where $z_1^{(\mathrm{avg})}$ is given by Eq. (12).

Determining the signed length of the central component of the accumulated progress vector is complicated by the fact that the direction of that component changes from one time step to the next. Similar to Eq. (6), the signed length of the central component of the accumulated progress vector can be computed as

$$s_A = \frac{\mathbf{s} \cdot (\hat{\mathbf{x}} - \mathbf{x})}{R}$$

where $\hat{\mathbf{x}} = (x_1, 0, \ldots, 0)$ and where $\cdot$ denotes the inner product of two vectors. Using the assumption that the distance $R$ from the ridge axis does not change along with Eq. (3) and the fact that according to Step 3 of the algorithm in Section 2.1

$$\hat{\mathbf{x}}^{(t+1)} - \mathbf{x}^{(t+1)} = \hat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)} - \sigma^{(t)} \mathbf{z}_{2\ldots N}^{(\mathrm{avg})}$$

it follows that

$$\begin{aligned} s_A^{(t+1)} &= \frac{1}{R} \left( (1-c)\mathbf{s}^{(t)} + \sqrt{\mu c(2-c)}\mathbf{z}^{(\mathrm{avg})} \right) \cdot \left( \hat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)} - \sigma^{(t)} \mathbf{z}_{2\ldots N}^{(\mathrm{avg})} \right) \\ &= (1-c) \left( s_A^{(t)} - \frac{\sigma^{(t)}}{R} \mathbf{s}^{(t)} \cdot \mathbf{z}_{2\ldots N}^{(\mathrm{avg})} \right) \\ &\qquad + \sqrt{\mu c(2-c)} \left( z_A^{(\mathrm{avg})} - \frac{\sigma^{(t)}}{R} \|\mathbf{z}_{2\ldots N}^{(\mathrm{avg})}\|^2 \right). \end{aligned}$$

Considering the first pair of parentheses on the right hand side, as the direction of the lateral component of the progress vector $\mathbf{z}^{(\mathrm{avg})}$ is random, the product $\mathbf{s}^{(t)} \cdot \mathbf{z}_{2\ldots N}^{(\mathrm{avg})}$ in the mean equals $s_A^{(t)} z_A^{(\mathrm{avg})}$. Moreover, from Eq. (17) with $\rho = 2Rd/N$ it follows that $\sigma/R \leq 2\mu c_{\mu/\mu,\lambda}/N$. Thus, as $2\mu c_{\mu/\mu,\lambda} z_A^{(\mathrm{avg})}/N \ll 1$, the second term in the pair of parentheses disappears compared to the first in the limit $N \to \infty$ and can thus be dropped from the calculations. As a consequence, from the stationarity requirement on the signed length of the accumulated progress vector's central component it follows that

$$s_A \overset{N\to\infty}{=} \sqrt{\frac{\mu(2-c)}{c}} \left( z_A^{(\mathrm{avg})} - \frac{\sigma}{R} \|\mathbf{z}_{2\ldots N}^{(\mathrm{avg})}\|^2 \right) \tag{24}$$

where $z_A^{(\mathrm{avg})}$ and $\|\mathbf{z}_{2\ldots N}\|^2$ are given by Eqs. (13) and (14), respectively.

Finally, again from Eq. (3), the overall squared length of the accumulated progress vector at time $t+1$ is

$$\begin{aligned} \|\mathbf{s}^{(t+1)}\|^2 = (1-c)^2 \|\mathbf{s}^{(t)}\|^2 \\ + 2(1-c)\sqrt{\mu c(2-c)}\mathbf{s}^{(t)} \cdot \mathbf{z}^{(\mathrm{avg})} + \mu c(2-c)\|\mathbf{z}^{(\mathrm{avg})}\|^2. \end{aligned}$$

As the direction of the lateral component of the progress vector $\mathbf{z}^{(\mathrm{avg})}$ is random, the product $\mathbf{s}^{(t)} \cdot \mathbf{z}^{(\mathrm{avg})}$ in the mean equals $s_1^{(t)} z_1^{(\mathrm{avg})} + s_A^{(t)} z_A^{(\mathrm{avg})}$, and, using Eqs. (23) and (24),

$$s_1^{(t)} z_1^{(\mathrm{avg})} + s_A^{(t)} z_A^{(\mathrm{avg})} \overset{N\to\infty}{=} \sqrt{\frac{\mu(2-c)}{c}} \left( {z_1^{(\mathrm{avg})}}^2 + {z_A^{(\mathrm{avg})}}^2 - \frac{\sigma^{(t)}}{R} z_A^{(\mathrm{avg})} \|\mathbf{z}_{2\ldots N}^{(\mathrm{avg})}\|^2 \right).$$

Demanding stationarity and solving for the squared length of the accumulated progress vector thus yields

$$\|\mathbf{s}\|^2 \overset{N\to\infty}{=} \frac{2(1-c)}{c} \mu \left( {z_1^{(\mathrm{avg})}}^2 + {z_A^{(\mathrm{avg})}}^2 - \frac{\sigma}{R} z_A^{(\mathrm{avg})} \|\mathbf{z}_{2\ldots N}^{(\mathrm{avg})}\|^2 \right) + \mu \|\mathbf{z}^{(\mathrm{avg})}\|^2. \quad (25)$$

Altogether, Eqs (23), (24), and (25) together with Eqs. (12), (13), and (14) characterise the accumulated progress vector of the $(\mu/\mu, \lambda)$-CSA-ES on the noisy parabolic ridge for high search space dimensionality.

### 4.2. Mutation Strength and Progress Rate

From Eq. (4), the stationarity requirement for the mutation strength is satisfied if and only if $\|\mathbf{s}\|^2 = N$. The rightmost term on the right hand side of Eq. (25) equals $\mu({z_1^{(\mathrm{avg})}}^2 + \|\mathbf{z}_{2\ldots N}^{(\mathrm{avg})}\|^2)$, where $z_1^{(\mathrm{avg})}$ and $\|\mathbf{z}_{2\ldots N}^{(\mathrm{avg})}\|^2$ are given by Eqs. (12) and (14), respectively. Due to the choice of the cumulation coefficient $c$, the term involving $z_1^{(\mathrm{avg})}$ thus disappears for $N \to \infty$ compared to both that involving $\|\mathbf{z}_{2\ldots N}^{(\mathrm{avg})}\|^2$ and the first term on the right hand side of Eq. (25). For large $N$, the stationarity condition thus requires that

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

and therefore, with Eqs. (12), (13), and (14) and the definition of $\rho$, that

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

Using the definition of $\sigma^*$ and squaring both sides, this condition can be written as

$$1 + 2\rho^2 + \rho^4 = 4{\sigma^*}^2 + 4{\sigma_\epsilon^*}^2 + 4{\sigma^*}^2\rho^2.$$

Using Eq. (16) to eliminate the normalised noise strength $\sigma_\epsilon^*$ finally yields condition

$$1 + 2\rho^2 - 3\rho^4 = 0$$

the only positive root of which is $\rho^2 = 1$. That is, independently of how the noise strength depends on the distance from the ridge axis, cumulative step length adaptation on the parabolic ridge generates mutation strengths such that the resulting standardised distance from the ridge axis is 1. If the standardised distance from the ridge axis exceeds 1, then Eq. (4) acts to reduce the mutation strength and vice versa.

Using $\rho^2 = 1$ in Eq. (16) and solving for $\sigma^*$ yields

$$\sigma^* \overset{N\to\infty}{=} \sqrt{\frac{1 - \sigma_\epsilon^{*2}}{2}} \tag{26}$$

for the stationary normalised mutation strength that the $(\mu/\mu, \lambda)$-CSA-ES employs on the noisy quadratic ridge in the limit $N \to \infty$. Notice that as $\rho = 1$, for the quadratic and cubic cases it follows that $\zeta^* \approx \sigma_\epsilon^*$. From Eq. (18) the resulting normalised progress rate is

$$\varphi^* \overset{N\to\infty}{=} \frac{1 - \sigma_\epsilon^{*2}}{2}. \tag{27}$$

Thus, in the absence of noise, the mutation strength generated by cumulative step length adaptation results in a progress rate that is half of the optimal progress rate that would be obtained if large values of $\sigma^*$ were used. While for uniform noise strength and $N \to \infty$, the same progress rate as in the absence of noise could be achieved for any level of noise present, cumulative step length adaptation instead generates smaller mutation strengths as the noise strength increases and fails to track the ridge for $\sigma_\epsilon^* \geq 1$. Similarly, in the case that the noise strength varies quadratically with the distance from the ridge axis, the $(\mu/\mu, \lambda)$-CSA-ES fails to track the ridge for $\zeta^* \geq 1$; however, this failure is unavoidable as it has been seen in Section 3 that for values of $\zeta^*$ greater than 1, a positive progress rate cannot be achieved with any mutation strength. Finally, for the case of a cubic dependence of the noise strength on the distance from the ridge axis, positive (albeit small) progress could be achieved for high levels of noise present, but cumulative step length adaptation fails to generate the small step lengths necessary for that purpose.

Figures 11, 12, and 13 compare predictions from Eqs. (26) and (27) that hold for $N \to \infty$ with measurements made in runs of the $(\mu/\mu, \lambda)$-CSA-ES on the noisy parabolic ridge for finite search space dimensionality. The figures illustrate that the accuracy of the predictions is not

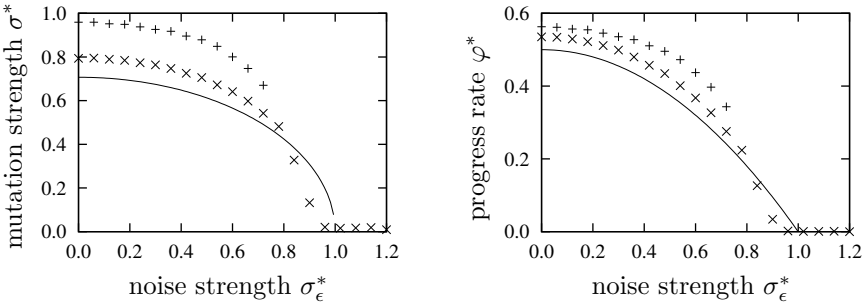


*Figure 11.* Average normalised mutation strength $\sigma^*$ and normalised progress rate $\varphi^*$ of the $(\mu/\mu, \lambda)$-CSA-ES plotted against normalised noise strength $\sigma^*_\epsilon$ for the case of uniform noise strength. The solid lines have been obtained from Eqs. (26) and (27), respectively. The points represent results measured in runs of the strategy on the parabolic ridge in search spaces with $N = 40$ (+) and $N = 400$ (×). Measurements have been made for $\mu = 3$, $\lambda = 10$, and $d = 1.0$.

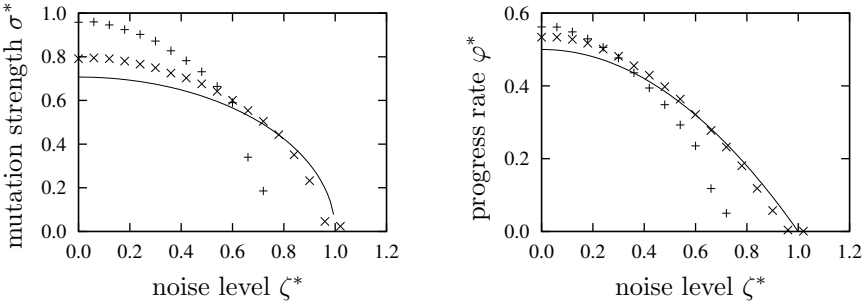


*Figure 12.* Average normalised mutation strength $\sigma^*$ and normalised progress rate $\varphi^*$ of the $(\mu/\mu, \lambda)$-CSA-ES plotted against normalised noise level $\zeta^*$ for the case that the noise strength increases quadratically with the distance from the ridge axis. The solid lines have been obtained from Eqs. (26) and (27), respectively. The points represent results measured in runs of the strategy on the parabolic ridge in search spaces with $N = 40$ (+) and $N = 400$ (×). Measurements have been made for $\mu = 3$, $\lambda = 10$, and $d = 1.0$.

as good as it is for the case of static step length. Especially for the case of cubic dependence of the noise strength on the distance from the ridge axis and $N = 40$ is the dynamic adaptation process unstable except for the smallest noise levels. However, it can also be seen that the accuracy of the predictions increases for increasing search space dimensionality as expected, and that the qualitative dependence on the mutation strength is described properly by Eqs. (26) and (27). Taking $N$-dependent terms into account in future work will allow making more accurate predictions and making quantitative recommendations with regard to the choice of $\mu$ and $\lambda$.

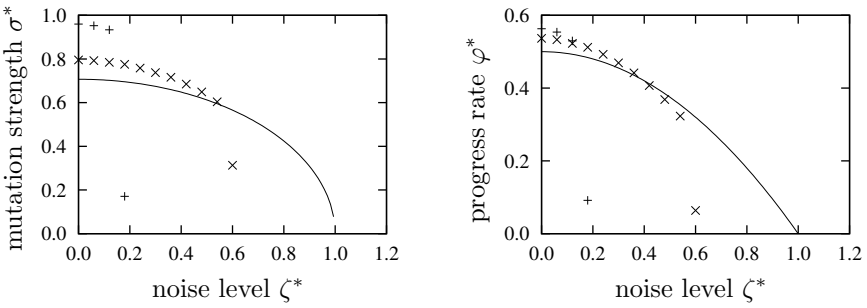


*Figure 13.* Average normalised mutation strength $\sigma^*$ and normalised progress rate $\varphi^*$ of the $(\mu/\mu, \lambda)$-CSA-ES plotted against normalised noise level $\zeta^*$ for the case that the noise strength increases cubically with the distance from the ridge axis. The solid lines have been obtained from Eqs. (26) and (27), respectively. The points represent results measured in runs of the strategy on the parabolic ridge in search spaces with $N = 40$ (+) and $N = 400$ (×). Measurements have been made for $\mu = 3$, $\lambda = 10$, and $d = 1.0$.

## 5. Conclusions and Outlook

This paper has examined the effects of noise on the performance of the $(\mu/\mu, \lambda)$-ES on the quadratic ridge function class. Three forms of noise have been considered: uniform noise that has the same strength throughout the search space, noise that increases quadratically with the distance from the ridge axis, and noise where that dependence is cubic. Quantitative results have been derived in the limit $N \to \infty$, and their accuracy has been tested experimentally in finite-dimensional search spaces. It has been seen that uniform noise can be effectively eliminated by working with a large mutation strength and tracking the ridge at a great distance. If the noise increases quadratically with the distance from the ridge, then it is no longer possible to drive the noise-to-signal ratio to zero by increasing the mutation strength, and above a certain noise level tracking the ridge becomes impossible unless $\mu$ and $\lambda$ are increased. If the increase of the noise strength with the distance from the ridge is cubic, then positive progress can always be achieved, albeit only with very small mutation strengths. Altogether, the three cases considered provide scenarios with widely differing characteristics that pose different demands to step length adaptation mechanisms.

Then, the performance of cumulative step length adaptation has been investigated for the three scenarios. It has been seen that independently of how the noise varies with the distance from the ridge axis, cumulative step length adaptation always generates mutation strengths that lead to the ridge being tracked at unit standardised distance. In the absence of noise, the progress rate that is achieved is half of the optimal

progress rate. If there is noise present, then cumulative step length adaptation fails to generate useful step lengths if the noise exceeds unit strength. The point where cumulative step length adaptation starts to fail can be deferred by increasing $\mu$ and $\lambda$.

This paper is but a first step toward an understanding of the behaviour of adaptive evolution strategies on the ridge function class. The directions in which the results presented here can be extended are numerous. First, it is desirable to obtain an improved understanding in finite-dimensional search spaces. Finite values of $N$ place limits on how far the population size parameters can beneficially be increased and are instrumental for deriving recommendations with regard to the choice of $\mu$ and $\lambda$. Such an understanding can be obtained by including some of the terms that have been dropped here in the analysis. The challenge is to determine what terms need to be considered, as well as the treatment of fluctuations (i.e., of quantities that cannot simply be replaced by their average values). Second, different forms of step length adaptation, such as mutative self-adaptation [9] or meta-ES [23] remain to be studied and compared with cumulative step length adaptation. For meta-ES, Herdy [17] provides empirical evidence for their usefulness for step length adaptation on the ridge. Third, other strategy variants, such as evolutionary gradient search strategies [24] or the $(\lambda)_{\text{opt}}$-ES studied in [2] on the sphere model remain to be considered. Of interest as well is the examination of ridge topologies other than the quadratic one. In the absence of noise and not considering step length adaptation, such an analysis has been presented in [10]. An finally, as pointed out by Whitley, Lunacek, and Knight [26], the ridge is a prime example for the usefulness of nonisotropic mutations. Strategies such as the CMA-ES [16] are capable of learning the direction of the ridge axis. After adaptation of the covariance matrix is complete, the CMA-ES can track the ridge by generating mutation vectors that have large components in direction of the ridge and much smaller components in other directions. This should prove especially useful if there is noise present that increases superquadratically with the distance from the ridge axis, and it will be interesting to see how noise affects the adaptation of the covariance matrix.

## Acknowledgements

This research has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

## References


1. D. V. Arnold. *Noisy Optimization with Evolution Strategies*. Genetic Algorithms and Evolutionary Computation Series. Kluwer Academic Publishers, Boston, 2002.
2. D. V. Arnold. Optimal weighted recombination. In A. H. Wright, M. D. Vose, K. A. De Jong, and L. M. Schmitt, editors, *Foundations of Genetic Algorithms 8*, pages 215–237. Springer Verlag, Heidelberg, 2005.
3. D. V. Arnold and H.-G. Beyer. Local performance of the $(\mu/\mu_I, \lambda)$-ES in a noisy environment. In W. N. Martin and W. M. Spears, editors, *Foundations of Genetic Algorithms 6*, pages 127–141. Morgan Kaufmann Publishers, San Francisco, 2001.
4. D. V. Arnold and H.-G. Beyer. A comparison of evolution strategies with other direct search methods in the presence of noise. *Computational Optimization and Applications*, 24(1):135–159, 2003.
5. 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.
6. D. V. Arnold and H.-G. Beyer. Expected sample moments of concomitants of selected order statistics. *Statistics and Computing*, 15(3):241–250, 2005.
7. N. Balakrishnan and C. R. Rao. Order statistics: An introduction. In N. Balakrishnan and C. R. Rao, editors, *Handbook of Statistics*, volume 16, pages 3–24. Elsevier, Amsterdam, 1998.
8. H.-G. Beyer. Toward a theory of evolution strategies: On the benefit of sex — the $(\mu/\mu, \lambda)$-theory. *Evolutionary Computation*, 3(1):81–111, 1995.
9. H.-G. Beyer. Toward a theory of evolution strategies: Self-adaptation. *Evolutionary Computation*, 3(3):311–347, 1996.
10. 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.
11. H.-G. Beyer. *The Theory of Evolution Strategies*. Natural Computing Series. Springer Verlag, Heidelberg, 2001.
12. H.-G. Beyer, D. V. Arnold, and S. Meyer-Nieberg. A new approach for predicting the final outcome of evolution strategy optimization under noise. *Genetic Programming and Evolvable Machines*, 6(1):7–24, 2005.
13. H.-G. Beyer and H.-P. Schwefel. Evolution strategies — A comprehensive introduction. *Natural Computing*, 1(1):3–52, 2002.
14. H. A. David and H. N. Nagaraja. Concomitants of order statistics. In N. Balakrishnan and C. R. Rao, editors, *Handbook of Statistics*, volume 16, pages 487–513. Elsevier, Amsterdam, 1998.
15. N. Hansen. *Verallgemeinerte individuelle Schrittweitenregelung in der Evolutionsstrategie*. Mensch & Buch Verlag, Berlin, 1998.
16. N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.
17. 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.
18. A. Ostermeier, A. Gawelczyk, and N. Hansen. Step-size adaptation based on non-local use of selection information. In Y. Davidor, H.-P. Schwefel, and

R. Männer, editors, *Parallel Problem Solving from Nature — PPSN III*, pages 189–198. Springer Verlag, Heidelberg, 1994.
19. 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.
20. A. I. Oyman, H.-G. Beyer, and H.-P. Schwefel. Where elitists start limping: Evolution strategies at ridge functions. In A. E. Eigen, T. Bäck, M. Schoenauer, and H.-P. Schwefel, editors, *Parallel Problem Solving from Nature — PPSN V*, pages 109–118. Springer Verlag, Heidelberg, 1998.
21. 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.
22. I. Rechenberg. *Evolutionsstrategie — Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Friedrich Frommann Verlag, Stuttgart, 1973.
23. I. Rechenberg. *Evolutionsstrategie '94*. Friedrich Frommann Verlag, Stuttgart, 1994.
24. R. Salomon. Evolutionary algorithms and gradient search: Similarities and differences. *IEEE Transactions on Evolutionary Computation*, 2(2):45–55, 1998.
25. R. Salomon. The curse of high-dimensional search spaces: Observing premature convergence in unimodal functions. In *Proc. of the 2004 IEEE Congress on Evolutionary Computation*, pages 918–923. IEEE Press,, Piscataway, NJ, 2004.
26. D. Whitley, M. Lunacek, and J. Knight. Ruffled by ridges: How evolutionary algorithms can fail. In K. Deb, R. Poli, W. Banzhaf, H.-G. Beyer, E. Burke, P. Darwen, D. Dasgupta, D. Floreano, J. Foster, M. Harman, O. Holland, P. L. Lanzi, L. Spector, A. Tettamanzi, D. Thierens, and A. Tyrell, editors, *Genetic and Evolutionary Computation — GECCO 2004*, pages 294–306. Springer Verlag, Heidelberg, 2004.