Evolution Strategies with Adaptively Rescaled Mutation Vectors

Dirk V. Arnold

Technical Report CS-2005-04

May 13, 2005

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

Evolution Strategies with Adaptively Rescaled Mutation Vectors

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

Abstract— **Rescaled mutations have been seen to have the potential to significantly improve the performance of evolution strategies in the presence of noise. However, to make use of that potential, the rescaling factor that determines the ratio of the lengths of the trial and search steps needs to be set appropriately. Good settings depend on a multitude of parameters and may vary over time. In this paper, an adaptive approach to generating rescaling factors is proposed. In experiments involving fitness-proportionate noise on several ellipsoidal test functions is it seen that robust and nearly optimal performance is achieved across a range of noise strengths.**

I. Introduction

Practical optimization problems often suffer from noise. Potential sources of noise include measurement limitations, the use of randomized algorithms or Monte Carlo methods, and human-computer interaction. Noise typically negatively affects the local performance of optimization algorithms, leading to reduced rates of convergence or even to divergent behavior. It is thus desirable to devise optimization strategies the performance of which is relatively robust with regard to the effects of noise. Focus in the present paper is on multirecombination evolution strategies for continuous optimization problems. A comparison presented in [1] has shown that multirecombination evolution strategies with cumulative step length adaptation are more effective in the presence of noise than several other popular direct search strategies. Reasons for their good performance include both the rescaling implicit in multirecombination and the relative robustness of the step length adaptation mechanism.

When using evolutionary algorithms for noisy optimization, noise directly affects the selection process. Noisy fitness information can lead to good candidate solutions not being able to reproduce or survive while relatively poor candidate solutions are selected for reproduction or survival. In order to ensure that it is the good candidate solutions that prevail, it is desirable to reduce the noise-to-signal ratio that a strategy operates under. This can be achieved in at least two different ways:

1) The noise strength can be reduced by evaluating candidate solutions multiple times and averaging over the measurements. If the noise is Gaussian, averaging over k samples reduces its standard deviation by a factor of $\sqrt{k}$. Suggestions on how to choose or adapt k have been made by Stagge [2] as well as more recently by Branke and Schmidt [3].
2) Attempts can be made to boost the signal strength. The signal strength is given by the standard deviation of the (undisturbed) fitness values of the candidate solutions subject to selection. For the case of multirecombination evolution strategies, that standard deviation is closely related to the mutation strength employed by the strategy. An increase in mutation strength typically results in an increased signal strength.

Unfortunately, however, both approaches have their downsides. Reducing the noise strength by averaging over multiple fitness measurements comes at a high computational cost as typically, fitness evaluations are expensive. Unless candidate solutions can be evaluated in parallel, the computational costs are linear in the number of samples that are averaged. On the other hand, boosting the signal strength by increasing the mutation strength is useful only up to a certain point. If the mutation strength is increased too far, then almost all of the offspring candidate solutions that are generated will be inferior to their parents, effectively rendering the strategy useless.

A solution to this latter predicament was proposed by Ostermeier and described by Rechenberg [4]. The basic idea is to make large steps when generating offspring candidate solutions, but in the end to realize only part of the step that has been generated. More specifically, the steps used to generate offspring are chosen large enough in order for the strength of the signal to compare favorably with the noise strength, and thus to allow reliable selection of the best candidate solutions. As for large steps it is likely that all of the candidate solutions thus generated are inferior to their parents, the mutation vectors corresponding to the offspring candidate solutions that have been selected are then "rescaled" by division by a factor $\kappa > 1$. The idea is illustrated in Fig. 1 for the simple case of a $(1, \lambda)$-ES. Beyer [5] has coined the phrase "Mutate large, but inherit small!" for the idea of using rescaled mutations. Notice the similarity of rescaled mutations to the idea underlying the gradient-based implicit filtering algorithm by Gilmore and Kelley [6] that obtains gradient estimates using finite differencing with a relatively large step

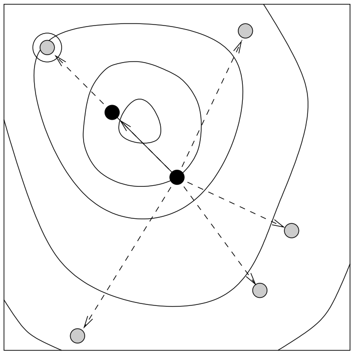

Fig. 1. A $(1, \lambda)$-ES using rescaled mutations. Five (gray) offspring candidate solutions are generated by mutating the (black) candidate solution in the center. All of the five are inferior to their parent. The parent of the next time step is generated by taking a step in direction of the best of the five (circled), but reducing the length of the step by a factor $1/\kappa$ in the process.

length in order to step over local minima.

An important question is of course how the rescaling factor κ should be chosen. Optimal values of κ depend on the objective function, parameters of the strategy, the strength of the noise present, and typically vary over time. A good setting at some point during an optimization process may be unfavorable at a later stage and vice versa. Analyses of the performance of evolution strategies with rescaled mutations that have been performed on the sphere model by Beyer [5], [7] are useful for providing a good understanding of the issues involved, but offer little practical advice with regard to the setting of κ.

The goal of the present paper is to propose and test a mechanism for the automatic adaptation of the rescaling factor κ. Its remainder is organized as follows. Section II gives a brief review of multirecombination evolution strategies and of the cumulative step length adaptation mechanism. In Section III, previous work with regard to the performance of evolution strategies on the sphere model is discussed. Links between the $(1, \lambda)$-ES using rescaled mutations, properties of the $(\mu/\mu, \lambda)$-ES, and the choice of weights in evolution strategies using weighted multirecombination are pointed out. Section IV proposes a strategy for adapting the rescaling factor κ. That strategy borrows ideas both from the self-adaptation mechanism proposed by Rechenberg [8] and Schwefel [9], and from the cumulative step length adaptation mechanism due to Ostermeier et al. [10]. The algorithm is evaluated experimentally in Section V on several test functions. Section VI concludes with a brief summary and directions for future work.

II. Multirecombination Evolution Strategies

Let $f : \mathbb{R}^N \to \mathbb{R}$ be a function to be minimized. Multirecombination evolution strategies with rescaled mutations repeatedly update a search point $\mathbf{x} \in \mathbb{R}^N$ using the following four steps:

1) Generate λ offspring candidate solutions
$$\mathbf{y}^{(i)} = \mathbf{x} + \kappa\sigma\mathbf{z}^{(i)} \qquad i = 1, \ldots, \lambda.$$
The steps used to generate the offspring candidate solutions are referred to as trial steps. The $\mathbf{z}^{(i)}$ are vectors consisting of N independent, standard normally distributed components and are referred to as mutation vectors. The nonnegative quantity σ is referred to as the mutation strength and, together with the rescaling factor κ, determines the length of the trial steps.

2) Determine the objective function values $f(\mathbf{y}^{(i)})$ of the offspring candidate solutions and order the $\mathbf{y}^{(i)}$ according to those values. After ordering, index $k;\lambda$ refers to the kth best of the λ offspring.

3) Compute the weighted sum
$$\mathbf{z}^{(\text{avg})} = \sum_{k=1}^{\lambda} w_{k;\lambda}\mathbf{z}^{(k;\lambda)} \tag{1}$$
of the mutation vectors. The $w_{k;\lambda}$ are weights that depend on the rank of the corresponding candidate solution in the set of all offspring. The vector $\mathbf{z}^{(\text{avg})}$ is referred to as the progress vector.

4) Replace the search point $\mathbf{x}$ by $\mathbf{x} + \sigma\mathbf{z}^{(\text{avg})}$. The resulting step is referred to as a search step.

Clearly, $\sigma\mathbf{z}^{(\text{avg})}$ connects consecutive search points. Depending on the choice of the weights $w_{k;\lambda}$ in Eq. (1), the definition of the multirecombination evolution strategy with rescaled mutations subsumes several variants of evolution strategies that can be found in the literature.

- Choosing
$$w_{k;\lambda} = \begin{cases} 1 & \text{if } k = 1 \\ 0 & \text{otherwise} \end{cases},$$
the strategy is a $(1, \lambda)$-ES with rescaled mutations as defined in [4] and [5][1] and illustrated in Fig. 1.
- If $\kappa = 1$ and
$$w_{k;\lambda} = \begin{cases} 1/\mu & \text{if } 1 \le k \le \mu \\ 0 & \text{otherwise} \end{cases}$$
and for some $\mu < \lambda$, then the strategy is a $(\mu/\mu, \lambda)$-ES as described in [11]. The search point $\mathbf{x}$ is the centroid of the population that consists of the μ best of the λ offspring candidate solutions generated.
- The choice
$$w_{k;\lambda} = E_{k;\lambda} \qquad k = 1, \ldots, \lambda,$$
where E_k is the expected value of the $(\lambda+1-k)$th order statistic of a sample of λ independent, standard normally distributed random variables yields the $(\lambda)_{\text{opt}}$-ES as introduced in [12].[2] The $(\lambda)_{\text{opt}}$-ES has the best performance of

[1] The definition used here differs from that used in [5] in that here, κ is used to multiply the trial steps rather than to divide the search steps. That is, here, $\kappa\sigma$ is the multiplier for the trial steps in step 1) of the algorithm and σ is the multiplier for the search steps in step 4). The respective multipliers used by Beyer are σ and σ/κ, respectively. The two strategies really are identical as the difference amounts to no more than a different parameterization of the mutation strength. The change is made as it is advantageous for the presentation of the mechanism for adapting κ that is outlined below.

[2] The same change in the parameterization of the mutation strength that has been described for the $(1, \lambda)$-ES in the previous footnote has been made for the $(\lambda)_{\text{opt}}$-ES.

any multirecombination evolution strategy on the infinite-dimensional sphere model in the absence of noise.

It is crucial for the performance of evolution strategies in real-valued search spaces that the mutation strength σ be adapted continually. Two mechanisms for updating the mutation strength are mutative self-adaptation due to Rechenberg [8] and Schwefel [9], and cumulative step length adaptation due to Ostermeier, Gawelczyk, and Hansen [10], [13]. In this paper, cumulative step length adaptation is employed, and the resulting strategies are referred to as CSA-ES. The goal of cumulative step length adaptation is to minimize correlations between successive steps. For that purpose, an exponentially fading record of the most recently taken steps is kept by accumulating progress vectors. Specifically, N-dimensional vector $\mathbf{s}$ is defined by $\mathbf{s}^{(0)} = \mathbf{0}$ and

$$\mathbf{s}^{(t+1)} = (1-c)\mathbf{s}^{(t)} + \sqrt{\frac{c(2-c)}{\chi}}\mathbf{z}^{(\text{avg})(t)}, \tag{2}$$

where t indicates time, and where $\chi = 1$ for the $(1,\lambda)$-ES, $\chi = 1/\mu$ for the $(\mu/\mu,\lambda)$-ES, and $\chi = \sum_{k=1}^{\lambda} E_{k;\lambda}^2$ for the $(\lambda)_{\text{opt}}$-ES. The mutation strength is then updated according to

$$\sigma^{(t+1)} = \sigma^{(t)} \exp\left(\frac{\|\mathbf{s}^{(t+1)}\|^2 - N}{2DN}\right). \tag{3}$$

The cumulation parameter c and damping constant D are set to $4/N$ and $N/4$, respectively, according to recommendations made by Hansen and Ostermeier [14].[3]

It is important to note that the use of isotropic mutations as described above is ineffective for many problems with widely differing eigenvalues of the Hessian matrix. For such problems, mutation vectors need to be generated with general covariance matrix $\mathbf{C}$, and $\mathbf{C}$ needs to be adapted to be roughly equal to the inverse of the Hessian matrix of the objective function. The CMA-ES introduced by Hansen and Ostermeier [13], [14] and studied experimentally by Kern et al. [16] accomplishes that task in that it has been found to effectively transform many objective functions into the sphere model considered in Section III.

III. Insights from the Sphere Model

The performance of evolution strategies is best understood on the sphere model, i.e. on objective function

$$f(\mathbf{y}) = \sum_{i=1}^{N} y_i^2.$$

The sphere model was introduced by Rechenberg [8] as a model for objective functions in the vicinity of local optima. Due to its scale invariance, provided that the step length is adapted appropriately, evolution strategies on the sphere model exhibit a stochastic form of linear convergence that is illustrated in Fig. 2 after initialization effects have faded. The speed of convergence can be quantified by the quality gain

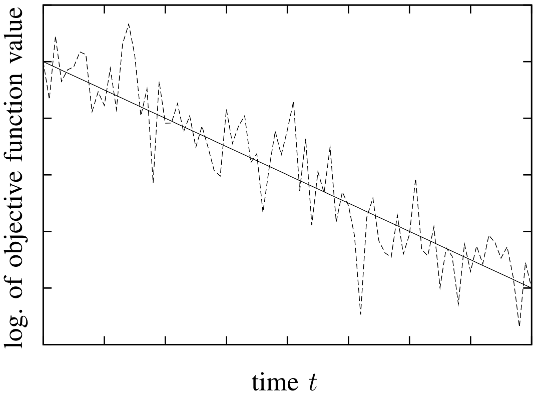

Fig. 2. Convergence behavior of evolution strategies on the sphere model. The dashed line shows the logarithm of the objective function value of the search point over time. The quality gain is determined by the slope of the solid regression line.

$$\varDelta = \mathrm{E}\left[\log\left(f(\mathbf{x}^{(t)})\right) - \log\left(f(\mathbf{x}^{(t+1)})\right)\right], \tag{4}$$

i.e. the expected improvement in the logarithmic objective function value of the search point from one time step to the next.[4] Clearly, the quality gain determines the slope of the regression line in Fig. 2. The greater the quality gain, the steeper the line and the faster the approach of the optimizer.

Noise is commonly modeled by an additive Gaussian term with mean zero. That is, it is assumed that evaluating a candidate solution $\mathbf{y}$ yields a value that is normally distributed with mean $f(\mathbf{y})$ and with a standard deviation $\sigma_\epsilon(\mathbf{y})$ that is referred to as the noise strength. If the noise strength varies with the location in search space such that it is proportional to the objective function value $f(\mathbf{y})$, then the scale invariance of the sphere model is preserved. As a consequence, provided that the step length adaptation mechanism can cope with it, evolution strategies will exhibit stochastic linear convergence in the presence of noise. The assumption of fitness-proportionate noise strength models relative errors of measurement that arise for example in connection with physical measurement devices that are accurate up to a certain percentage of the quantity they measure and will be adopted in all of what follows.

The performance of a variety of evolution strategies has been studied on the noisy sphere model in the limit of infinite

[3] Both the cumulation rule in Eq. (2) and the update rule in Eq. (3) differ somewhat from the respective rules given in [10], [13]. The cumulation rule has been adapted for use with the three types of strategies considered here. The update rule for the mutation strength differs in that it uses the squared length of the accumulated progress vector rather than its length. The change results in a more elegant formulation that makes the strategy amenable to analysis as shown in [12], [15]. It appears to not substantially affect the performance of the strategy in practice.

[4] The more common definition for the quality gain is

$$\varDelta = \mathrm{E}\left[f(\mathbf{x}^{(t)}) - f(\mathbf{x}^{(t+1)})\right]$$

and has been employed for example by Beyer [17]. In the limit of infinite search space dimensionality both measures agree on the sphere model if normalized appropriately. The definition employed here has previously been used in [1] and has the advantage of properly reflecting the slope of the regression line in Fig. 2.

search space dimensionality. Using appropriate normalizations

$$\sigma^* = \sigma\frac{N}{R}, \quad \sigma^*_\epsilon = \sigma_\epsilon\frac{N}{2R^2}, \quad \text{and} \quad \varDelta^* = \varDelta\frac{N}{2},$$

where $R = \|\mathbf{x}\|$ denotes the distance of the search point from the optimizer, the quality gain laws are concise and offer insights into the workings of the strategies as outlined in what follows. Notice that the assumption of fitness-proportionate noise strength implies that the normalized noise strength σ^*_ϵ is constant throughout the search space.

For the case of the $(1,\lambda)$-ES with rescaled mutations, Beyer [5] has derived the performance law[5]

$$\varDelta^* = \frac{c_{1,\lambda}\sigma^*}{\sqrt{1+(\sigma^*_\epsilon/\kappa\sigma^*)^2}} - \frac{{\sigma^*}^2}{2}. \tag{5}$$

The progress coefficient $c_{1,\lambda} = E_{1;\lambda}$ can be computed numerically and is tabulated in [17]. The quotient $\sigma^*_\epsilon/\kappa\sigma^*$ that appears in the denominator of the first term on the right hand side is the noise-to-signal ratio that the strategy operates under. In the absence of noise, that ratio equals zero and the setting of κ is without influence on the quality gain of the strategy. In the presence of noise, the use of a larger rescaling factor decreases the noise-to-signal ratio and leads to better performance. In fact, Eq. (5) suggests that the noise-to-signal ratio could be driven to zero by letting κ tend to infinity.

However, it is important to keep in mind that Eq. (5) has been derived under the assumption of infinite search space dimensionality, and that it represents merely an approximation in finite-dimensional search spaces. Beyer [7] has also derived a performance law for the $(1,\lambda)$-ES with rescaled mutations that is a better approximation for finite N. It has been seen that while operating with a rescaling factor $\kappa > 1$ is generally beneficial in the presence of noise, it is not true that κ can be increased indefinitely without negatively affecting the performance of the strategy.

The performance of the $(\mu/\mu,\lambda)$-ES on the infinite-dimensional noisy sphere has been analyzed in [18], [19], yielding the quality gain law

$$\varDelta^* = \frac{c_{\mu/\mu,\lambda}\sigma^*}{\sqrt{1+(\sigma^*_\epsilon/\sigma^*)^2}} - \frac{{\sigma^*}^2}{2\mu}. \tag{6}$$

The progress coefficient $c_{\mu/\mu,\lambda} = \sum_{k=1}^{\mu} E_{k;\lambda}/\mu$ can be computed numerically and is tabulated in [17]. While the noise-to-signal ratio is $\sigma^*_\epsilon/\sigma^*$ and seems at first sight unaffected by the use of multirecombination, the appearance of the factor μ in the denominator of the second term on the right hand side signifies the presence of genetic repair (see Beyer [17]). As a consequence of genetic repair, the $(\mu/\mu,\lambda)$-ES is capable of operating with a mutation strength that is roughly μ-fold larger than that of the one-parent strategy. As a consequence of $(\mu/\mu,\lambda)$-type multirecombination, the squared length of the $(\mu/\mu,\lambda)$-search step is reduced by a factor of μ compared to the squared lengths of the respective trial steps. The $(\mu/\mu,\lambda)$-ES thus performs an *implicit* rescaling of mutation vectors that is equivalent in effect to the explicit rescaling of the $(1,\lambda)$-ES with rescaled mutations. Unlike the strategies that rescale explicitly, the amount of rescaling performed by the $(\mu/\mu,\lambda)$-ES is determined by the population size parameters μ and λ. As a consequence, it is not possible to reap the benefits of strong rescaling while operating with a small population.

In [20] the potential of rescaling mutation vectors was compared with the benefits resulting from the averaging over multiple evaluations of the objective function in an attempt to remove the noise as discussed in Section I. It was seen that the implicit rescaling that the $(\mu/\mu,\lambda)$-ES performs is more effective on the infinite-dimensional noisy sphere in that a k-fold increase in μ and λ yields a greater gain in performance than the averaging over k function evaluations does. While no such investigations have been conducted yet for strategies that rescale explicitly, it is expected that the advantage of rescaling over averaging is even more pronounced in that case as explicit rescaling comes without an increase in computational costs.

Equation (6) suggests that by increasing μ and λ indefinitely, the $(\mu/\mu,\lambda)$-ES can operate with arbitrarily large mutation strengths and thus reduce the noise-to-signal ratio to zero, effectively eliminating any amount of noise. However, as the recommendation to increase κ in the strategy that rescales mutation vectors explicitly, finite-dimensional search spaces place limits on useful population sizes. If μ and λ are increased too far, then the efficiency of the strategy begins to suffer.

Finally, for the $(\lambda)_{\text{opt}}$-ES the performance law[6]

$$\varDelta^* = W_\lambda\left(\frac{\sigma^*}{\sqrt{1+(\sigma^*_\epsilon/\kappa\sigma^*)^2}} - \frac{{\sigma^*}^2}{2}\right), \tag{7}$$

where $W_\lambda = \sum_{k=1}^{\lambda} E^2_{k;\lambda}$, has been derived in [12]. For large λ, the coefficient W_λ approaches λ. In the absence of noise, the $(\lambda)_{\text{opt}}$-ES is the most efficient of all multirecombination evolution strategies on the infinite-dimensional sphere model. As in the corresponding law for the $(1,\lambda)$-ES in Eq. (5), the noise-to-signal ratio is $\sigma^*_\epsilon/\kappa\sigma^*$ and is thus moderated by rescaling factors $\kappa > 1$. Also as for the strategies considered previously, finite search space dimensionalities place limits on useful values for the rescaling factor, and the accuracy of the equation decreases with increasing κ.

IV. Adaptively Rescaled Mutation Vectors

While the analyses of the performance of the $(1,\lambda)$-ES with rescaled mutations and of the $(\lambda)_{\text{opt}}$-ES on the infinite-dimensional sphere model resulting in Eqs. (5) and (7) place no bounds on the value of κ and suggest that it should be chosen as large as possible, this does not hold true in finite-dimensional search spaces. Figure 3 illustrates how the quality gain of the $(\lambda)_{\text{opt}}$-CSA-ES depends on κ for the sphere model

[5] To be exact, Beyer derived a law for the progress rate rather than for the quality gain. However, it is well known that in the limit $N \to \infty$ both performance measures agree if normalized appropriately. Also note that the law has been modified from [5] to reflect the change in parameterization of the mutation strength discussed in Section II.

[6] As for the $(1,\lambda)$-ES with rescaled mutations, this law was adapted to reflect the altered parameterization of the mutation strength discussed in Section II.

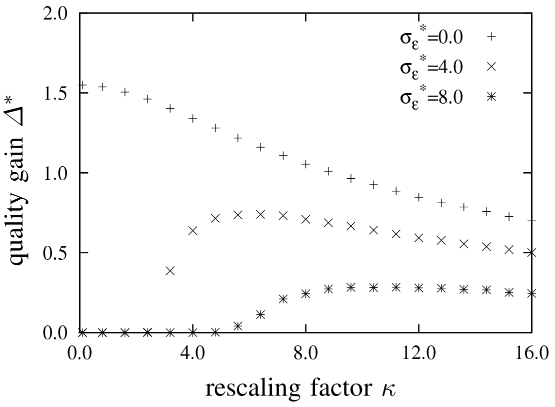

Fig. 3. Quality gain Δ^* on the sphere model of the $(\lambda)_{\mathrm{opt}}$-CSA-ES with $\lambda = 10$ plotted against the rescaling factor κ. The points indicate measurements made in runs of evolution strategies in a search space with dimensionality $N = 40$ and noise strengths $\sigma_\epsilon^* = 0.0$, 4.0, and 8.0 as indicated.

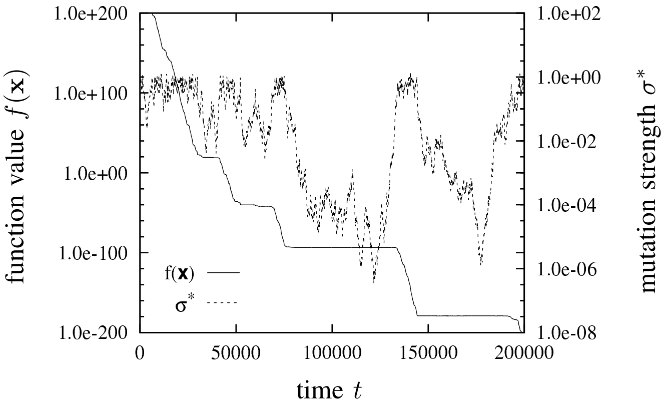

Fig. 4. Function value $f(\mathbf{x})$ and normalized mutation strength σ^* measured in a typical run of a $(\lambda)_{\mathrm{opt}}$-CSA-ES with $\lambda = 10$ and $\kappa = 1.0$ on the sphere model with $N = 40$ and $\sigma_\epsilon^* = 1.4$. Notice that the scales of both vertical axes are logarithmic.

with $N = 40$ and for several noise strengths. The figure shows that the choice of κ is crucial for the performance of the strategy in that for example, for $\sigma_\epsilon^* = 4.0$ positive quality gain is not achieved for rescaling factors $\kappa < 2.0$. Considerable progress *is* possible for larger values of κ. As a general rule, it is desirable to work with small values of κ if there is no noise, and to choose κ sufficiently large in the presence of noise. Higher noise strengths demand larger values of κ, but choosing κ too large is detrimental to the performance of the strategy.

The reasons for this behavior are easily understood. In the absence of noise, choosing κ small yields the best progress vector just as small steps yield the best approximation of the gradient when using finite differencing in a gradient strategy. In the presence of noise, larger values of κ reduce the noise-to-signal ratio that the strategy operates under as seen in Section III. If κ is chosen too small, then the noise outweighs the signal and search steps are essentially random and therefore uncorrelated. Cumulative step length adaptation, which tries to eliminate correlations in the sequence of steps taken, thus fails to see a need to increase the step length. The behavior of the $(\lambda)_{\mathrm{opt}}$-CSA-ES at the borderline between the regimes where positive quality gain is possible and where it is not is illustrated in Fig. 4. Phases of stochastic linear convergence (steep decrease in $f(\mathbf{x})$) alternate with phases where the strategy stagnates (plateaus in the $f(\mathbf{x})$ curve). In the stagnation phases, the noise-to-signal ratio is large and the logarithm of the mutation strength performs a random walk. If in the course of that random walk the step length grows large enough to result in a noise-to-signal ratio that allows progress towards the optimizer, then correlations reappear in the sequence of steps taken and cumulative step length adaptation acts to further increase the mutation strength, driving the strategy back into a phase of stochastic linear convergence.

It is conceivable that optimal values of κ could be derived as functions of N and σ_ϵ^* from quality gain laws for multirecombination evolution strategies in finite-dimensional search spaces. However, such results would be of little practical use as the conditions under which they have been derived are very particular. It seems unreasonable to assume that good settings for the rescaling parameter κ that have been derived on the sphere model with fitness-proportionate noise strength would be generally useful for other objective functions and noise models. Furthermore, the normalized noise strength is not a quantity that is measurable easily as it depends on the distance from the optimizer.

Instead, a mechanism is required that adapts κ to the particular situation at hand. Such a mechanism is proposed in Fig. 5. The underlying idea is to try two settings of the rescaling factor, and to settle for the better one. Cumulation and gradual adaptation are used to reduce harmful fluctuations and to make the algorithm robust in the face of noise. In order not to duplicate any work and incur a loss in efficiency if run on a single processor, the two settings of the rescaling factor are tried in an alternating sequence rather than in parallel. The algorithm thus combines elements of mutative self-adaptation (competition between different settings of a parameter) with the idea of "derandomization" propagated by Ostermeier et al. [10].

More specifically, the algorithm uses rescaling factors κ/α and $\kappa\cdot\alpha$, where κ is stored in a variable and where $\alpha > 1$ is a constant. Two further variables δ_- and δ_+ hold exponentially fading records of the gains achieved with the smaller and the larger settings of the rescaling factor, respectively. The **while** loop in Fig. 5 is the main loop of the evolutionary algorithm. In every iteration, the algorithm performs the two search steps and updates the records δ_- and δ_+. Function $\mathit{step}(\cdot)$ makes the search steps including mutation and multirecombination as well as cumulative step length adaptation as discussed in Section II. It returns the normalized difference

$$\frac{N}{2}\left[\log\left(f(\mathbf{x}^{(t)})\right) - \log\left(f(\mathbf{x}^{(t+1)})\right)\right] = \frac{N}{2}\log\left(\frac{f(\mathbf{x}^{(t)})}{f(\mathbf{x}^{(t+1)})}\right)$$

(compare Eq. (4) and the normalization in Section III), where

```
initialize $\kappa$
$\delta_- \leftarrow 0$
$\delta_+ \leftarrow 0$
while not done
    do $\delta_- \leftarrow (1 - c_\kappa)\delta_- + c_\kappa\, step(\kappa/\alpha)$
       $\delta_+ \leftarrow (1 - c_\kappa)\delta_+ + c_\kappa\, step(\kappa \cdot \alpha)$
       if $\delta_- < 0$
          then $\kappa \leftarrow \kappa \cdot \beta$
               $\sigma \leftarrow \sigma \cdot \beta$
       elseif $\delta_- > \delta_+$
          then $\kappa \leftarrow \kappa/\gamma$
       else $\kappa \leftarrow \kappa \cdot \gamma$
```

Fig. 5. An evolution strategy with adaptive rescaling of mutation vectors. The function $step(\cdot)$ performs a single search step as described in Section II and returns the resulting difference in logarithmic objective function values of the search point. Cumulative step length adaptation is also performed in $step(\cdot)$. See the text for a more detailed description of the entire algorithm and a discussion of the parameters.

the quotient $f(\mathbf{x}^{(t)})/f(\mathbf{x}^{(t+1)})$ is clamped to the interval $[1-\lambda/N, 1+\lambda/N]$ before taking the logarithm. The clamping is without consequences in the absence of noise, and it helps to remove outliers resulting from noisy objective function evaluations. The choice of the interval is inspired by theoretical considerations on the infinite-dimensional sphere, and it will be seen to be useful more generally in Section V.

After updating the two records of gains made with the respective settings of the rescaling parameter, variable κ is updated for use in the next iteration of the algorithm. It is first tested whether the accumulated gain made with the smaller setting of the rescaling parameter is negative. This situation is indicative of a stagnation phase as illustrated in Fig. 4, and both the mutation strength and κ are increased by multiplication with a factor $\beta > 1$ in response. Otherwise, δ_- and δ_+ are compared, and κ is increased by multiplication with $\gamma > 1$ if a greater gain was achieved with the larger setting of the rescaling parameter, and κ is decreased by division by γ if the smaller setting was the more successful one. Not shown in Fig. 5, κ is clamped to the interval $[0.5, N/2]$ after being updated in order to avoid instabilities in the presence of excessive amounts of noise. Notice that no significant additional computational costs are incurred as a result of the mechanism for the adaptation of κ other than the need to evaluate the objective function value of the search point in every time step.

The algorithm in Fig. 5 introduces several new parameters into the optimization procedure:

- Cumulation parameter $c_\kappa > 0$ determines how quickly the information in the exponentially fading records δ_- and δ_+ is replaced with new information. The smaller c_κ, the more persistent is the memory and the more dampened are any fluctuations due to noise. Choosing c_κ too small prevents fast adaptation. A setting of $c_\kappa = 0.4/N$ results in κ being adapted on a time scale ten times longer than the scale that the mutation strength is adapted on and has been used in all of what follows.
- Parameter $\alpha > 1$ determines how much the two settings of the rescaling parameter differ from each other. It needs to be large enough in order to be able to reliably discriminate between the respective quality gains achieved with the two settings. If α is chosen too large, performance suffers as at least one of the two settings of the rescaling parameter must differ substantially from the optimal setting. In all of what follows, $\alpha = 1.5$ has been used.
- Parameters $\beta > 1$ and $\gamma > 1$ determine the rate at which the rescaling parameter is adapted. They need to be large enough in order to allow for fast adaptation of κ, but not so large as to introduce strong fluctuations in the adaptation process. In all of what follows, settings $\beta = \exp(0.15/N)$ and $\gamma = \exp(0.015/N)$ have been used.

The settings of all parameters are usually uncritical in that similar results are obtained with similar settings. Moreover, the settings suggested here have proven to be useful across a range of objective functions. Finally, it should be noted that initializing κ to a relatively large value is useful as it affords relatively reliable information on the basis of which the rescaling factor can be reduced quickly if necessary. In all of what follows, κ has been set to 10.0 initially.

V. Experimental Evaluation

In this section, the adaptation mechanism for the rescaling factor κ introduced in the previous section is evaluated experimentally in a number of fitness environments. In particular, its performance is measured in runs of the $(\lambda)_{\text{opt}}$-CSA-ES on the sphere model discussed in Section III as well as on the three objective functions

$$\text{ellipsoid I:} \qquad f(\mathbf{y}) = \sum_{i=1}^{N} i y_i^2$$

$$\text{ellipsoid II:} \qquad f(\mathbf{y}) = \sum_{i=1}^{N} i^2 y_i^2$$

$$\text{ellipsoid III:} \qquad f(\mathbf{y}) = N \sum_{i=1}^{N/2} y_i^2 + \sum_{i=N/2+1}^{N} y_i^2.$$

The contours of these objective functions are ellipsoids with widely varying ratios of the lengths of their principal axes. As the behavior of the CSA-ES does not depend on the orientation of the coordinate system, the coordinates could be subjected to an arbitrary orthogonal transformation (thus removing separability) without any impact on the results. Ellipsoids I and II have previously been used as test functions for example in [21], ellipsoid III is a special case of the class of functions studied by Jägersküpper [22].

As on the sphere model, the CSA-ES exhibits linear convergence on the ellipsoids after some initialization period in which the mutation strength and the rescaling parameter are adapted and the search point takes some default position on

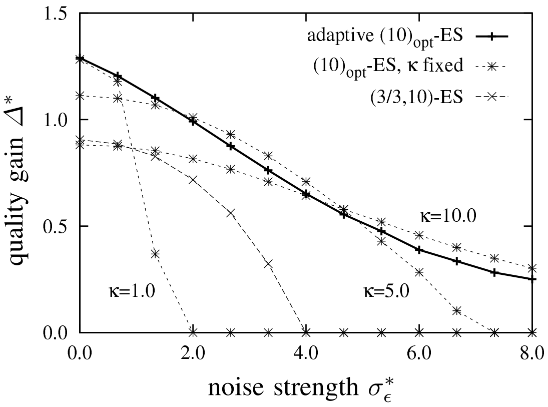

Fig. 6. Normalized quality gain $\varDelta^*$ of several CSA-ES on the noisy sphere model plotted against normalized noise strength σ_ϵ^*.

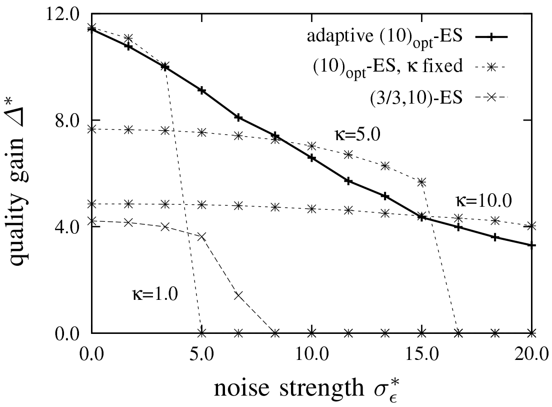

Fig. 7. Normalized quality gain $\varDelta^*$ of several CSA-ES on ellipsoid I plotted against normalized noise strength σ_ϵ^*.

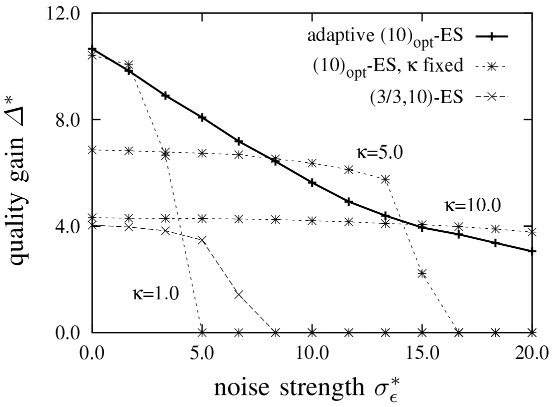

Fig. 8. Normalized quality gain $\varDelta^*$ of several CSA-ES on ellipsoid II plotted against normalized noise strength σ_ϵ^*.

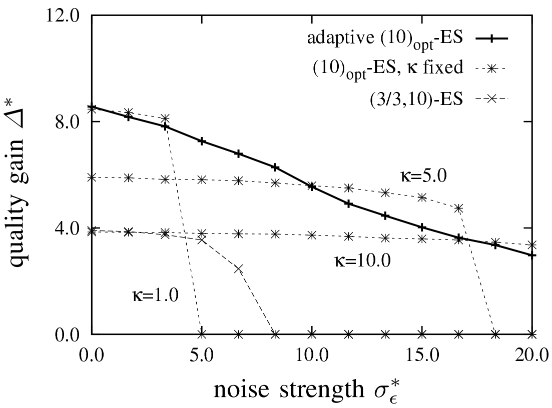

Fig. 9. Normalized quality gain $\varDelta^*$ of several CSA-ES on ellipsoid III plotted against normalized noise strength σ_ϵ^*.

the ellipsoid. Measurement of the quality gain starts after the initialization period is over and is measured until either 200,000 search steps have been made or the limit of numerical accuracy has been reached, whichever occurs first. Noise strength and quality gain are normalized as

$$\sigma_\epsilon^* = \sigma_\epsilon(\mathbf{y})\frac{\mathrm{Tr}(f)}{2f(\mathbf{y})} \quad \text{and} \quad \varDelta^* = \varDelta\frac{\mathrm{Tr}(f)}{2},$$

where $\mathrm{Tr}(f)$ denotes the trace of the Hessian matrix of the objective function. Keeping with the assumption of fitness-proportionate noise strength, the normalized noise strength σ_ϵ^* is thus constant. Notice that for the case of the sphere model, the normalizations agree with those used in Section III. The choice of the number of offspring λ generated per time step and the search space dimensionality N are uncritical for the qualitative behavior of the algorithm, and results will be reported only for the case that $\lambda = 10$ and $N = 40$.

Figures 6 to 9 contrast the normalized quality gain $\varDelta^*$ of the $(\lambda)_{\text{opt}}$-CSA-ES with adaptation of the rescaling factor κ with that of several instances of the $(\lambda)_{\text{opt}}$-CSA-ES with κ fixed and of the $(\mu/\mu, \lambda)$-CSA-ES on the four test functions. The most striking observation is the qualitative similarity of the four figures. In each case, it can be seen that the $(\lambda)_{\text{opt}}$-CSA-ES with $\kappa = 1.0$ is highly efficient in the absence of noise, but that its quality gain declines rapidly if there is noise present. (In fact, even better performance could be achieved in the absence of noise with rescaling factors $\kappa < 1$, but the difference is minor.) Choosing larger rescaling factors improves the performance of the $(\lambda)_{\text{opt}}$-CSA-ES in the presence of noise, but it does so at the cost of a reduced quality gain at low noise strengths. The adaptive strategy, the performance of which is indicated by bold lines, achieves about the same quality gain as the strategy that keeps $\kappa = 1.0$ fixed for $\sigma_\epsilon^* = 0.0$. In the noisy case, for every noise level there is an optimal setting for κ. The adaptive strategy does not quite achieve the optimal quality gain (in part because it uses two different settings of the rescaling parameter, in part as a result of fluctuations in the adaptation process), but it does achieve nearly optimal performance over the entire range of noise strengths considered. It is also markedly more efficient than the $(\mu/\mu, \lambda)$-CSA-ES both in the absence of noise and in its presence. The performance advantage of the adaptive $(\lambda)_{\text{opt}}$-ES over the $(\mu/\mu, \lambda)$-ES is more pronounced on the three ellipsoids than it is on the sphere model. By considering single runs of the $(\lambda)_{\text{opt}}$-CSA-ES (not shown here) it can also be seen that the adaptation mechanism effectively eliminates the long periods of stagnation observed in Fig. 4 as well as reflected by the steep drop in some of the curves for κ fixed in Figs. 6 to 9.

VI. Conclusions and Future Work

Rescaled mutations are a powerful means for evolutionary algorithms to deal with noisy information. In this paper, an algorithm has been proposed for adapting the rescaling factor that determines the ratio of the lengths of the trial and search steps in response to information gathered during the optimization process. It has been seen in experiments that that algorithm performs well on the noisy sphere model as well as on several other ellipsoidal fitness functions disturbed by noise. Assuming fitness-proportionate noise strength, the quality gain achieved in the stationary regime of the optimization process is not far below the optimal quality gain that can be achieved with any rescaling factor.

In future work, the case of fitness-proportionate noise strength is but one of several scenarios that should be considered. While it has been seen that the stationary behavior of the proposed mechanism for the adaptation of the rescaling factor drives the strategy into the vicinity of its optimal working regime, it is unclear how fast and reliable the strategy can react to changing normalized noise strengths. Thus, the influence of other forms of noise, such as noise of constant strength as studied in [21] or actuator noise as considered by Beyer, Olhofer, and Sendhoff [23] remains to be studied.

Some preliminary experiments have been conducted with CSA-ES on ellipsoidal objective functions with an eigenvalue spectrum that is dominated by a single value. In those experiments it was observed that while more robust than the $(\mu/\mu,\lambda)$-CSA-ES in the presence of noise, the $(\lambda)_{\text{opt}}$-CSA-ES performed worse than the $(\mu/\mu,\lambda)$-CSA-ES in its absence. Presumably, it was this observation that led Hansen and Ostermeier [14] to advise against the use of negative weights in the multirecombination procedure. It seems likely that the performance of the $(\lambda)_{\text{opt}}$-CSA-ES on the ridge function class can be studied analytically, and that the results obtained may yield useful clues as to the behavior of the strategy on ellipsoidal functions dominated by a single eigenvalue. Also of interest is the influence of the rescaling parameter on optimization performance on such functions. In preliminary experiments, it appears to be qualitatively different from what is seen on the functions studied here.

Finally, it is important to keep in mind that due to its use of isotropic mutations, the CSA-ES is not an efficient strategy for the optimization of functions with widely differing eigenvalues of their Hessians. On such functions, the CMA-ES is to be preferred, and it is of great interest to investigate the potential of the use of rescaled mutations and of the proposed mechanism for the adaptation of the rescaling factor for those strategies.

Acknowledgments

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

References

[1] 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*, vol. 24, no. 1, pp. 135–159, 2003.

[2] P. Stagge, "Averaging efficiently in the presence of noise," in *Parallel Problem Solving from Nature — PPSN V*, A. E. Eiben *et al.*, Eds. Springer Verlag, Heidelberg, 1998, pp. 188–200.

[3] J. Branke and C. Schmidt, "Sequential sampling in noisy environments," in *Parallel Problem Solving from Nature — PPSN VIII*, X. Yao *et al.*, Eds. Springer Verlag, Heidelberg, 2004, pp. 202–211.

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

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

[6] P. Gilmore and C. T. Kelley, "An implicit filtering algorithm for optimization of functions with many local minima," *SIAM Journal on Optimization*, vol. 5, pp. 269–285, 1995.

[7] H.-G. Beyer, "Evolutionary algorithms in noisy environments: Theoretical issues and guidelines for practice," *Computer Methods in Mechanics and Applied Engineering*, vol. 186, pp. 239–267, 2000.

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

[9] H.-P. Schwefel, *Evolution and Optimum Seeking*. Wiley, New York, 1995.

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

[11] H.-G. Beyer and H.-P. Schwefel, "Evolution strategies: A comprehensive introduction," *Natural Computing*, vol. 1, no. 1, pp. 3–52, 2002.

[12] D. V. Arnold, "Optimal weighted recombination," in *Foundations of Genetic Algorithms 8*, K. De Jong *et al.*, Eds. Springer Verlag, Heidelberg, 2005.

[13] N. Hansen, *Verallgemeinerte individuelle Schrittweitenregelung in der Evolutionsstrategie*. Mensch & Buch Verlag, Berlin, 1998.

[14] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, 2001.

[15] D. V. Arnold and H.-G. Beyer, "Performance analysis of evolutionary optimization with cumulative step length adaptation," *IEEE Transactions on Automatic Control*, vol. 49, no. 4, pp. 617–622, 2004.

[16] S. Kern, S. D. Müller, N. Hansen, D. Büche, J. Ocenasek, and P. Koumoutsakos, "Learning probability distributions in continuous evolutionary algorithms — A comparative review," *Natural Computing*, vol. 3, no. 1, pp. 77–112, 2004.

[17] H.-G. Beyer, *The Theory of Evolution Strategies*, ser. Natural Computing. Springer Verlag, Heidelberg, 2001.

[18] D. V. Arnold and H.-G. Beyer, "Local performance of the $(\mu/\mu_I,\lambda)$-ES in a noisy environment," in *Foundations of Genetic Algorithms 6*, W. N. Martin and W. M. Spears, Eds. Morgan Kaufmann Publishers, San Francisco, 2001, pp. 127–141.

[19] D. V. Arnold, *Noisy Optimization with Evolution Strategies*, ser. Genetic Algorithms and Evolutionary Computation. Kluwer Academic Publishers, Boston, 2002.

[20] D. V. Arnold and H.-G. Beyer, "Efficiency and mutation strength adaptation of the $(\mu/\mu,\lambda)$-ES in a noisy environment," in *Parallel Problem Solving from Nature — PPSN VI*, M. Schoenauer *et al.*, Eds. Springer Verlag, Heidelberg, 2000, pp. 39–48.

[21] H.-G. Beyer and D. V. Arnold, "The steady state behavior of $(\mu/\mu_I,\lambda)$-ES on ellipsoidal fitness models disturbed by noise," in *Genetic and Evolutionary Computation — GECCO 2003*, E. Cantú-Paz *et al.*, Eds. Springer Verlag, Heidelberg, 2003, pp. 525–536.

[22] J. Jägersküpper, "Rigorous runtime analysis of the $(1+1)$-ES: 1/5-rule and ellipsoidal fitness landscapes," in *Foundations of Genetic Algorithms 8*, K. De Jong *et al.*, Eds. Springer Verlag, Heidelberg, 2005.

[23] H.-G. Beyer, M. Olhofer, and B. Sendhoff, "On the behavior of $(\mu/\mu_I,\lambda)$-ES optimizing functions disturbed by generalized noise," in *Foundations of Genetic Algorithms 7*, K. De Jong *et al.*, Eds. Morgan Kaufmann Publishers, San Francisco, 2003, pp. 307–328.