Investigation of the (μ, λ)-ES in the Presence of Noise

Dirk V. Arnold and Hans-Georg Beyer
Department of Computer Science XI
University of Dortmund
44221 Dortmund, Germany
{arnold,beyer}@ls11.cs.uni-dortmund.de

Abstract-

While in the absence of noise no improvement in local performance can be gained from retaining but the best candidate solution found so far, it has been shown experimentally that, in the presence of noise, operating with a non-trivial population of candidate solutions can have a marked and positive effect on the local performance of evolution strategies (ES). In this paper, we attempt to shed some light on the reasons for the potential performance improvement. In particular, we derive a progress law for the (μ, λ)-ES on a noisy linear fitness function and both numerically and empirically study its implications. We then discuss the significance of the progress coefficients that have been obtained on the linear function for the quadratic sphere. Comparisons of the local performance of the (μ, λ)-ES and of the $(1 + 1)$-ES and the $(1, \lambda)$-ES are presented.

1 Introduction

Evolution strategies (ES) are nature-inspired search heuristics that iteratively apply operators of variation and selection to a population of candidate solutions to an optimization problem. There is evidence that such strategies are particularly effective and superior to many other common optimization schemes in the case that the objective function – in what follows referred to as *fitness function* – is disturbed by noise. Nissen and Propach [9] present empirical evidence that suggests that it is the use of a non-trivial population of candidate solutions that is responsible for the relative effectiveness of evolution strategies in the presence of noise.

In a number of papers, the local performance of the $(1 \dagger \lambda)$-ES [1, 4] as well as of the $(\mu/\mu_I, \lambda)$-ES [2, 3] have been studied in noisy, spherically symmetric fitness environments. In both cases, the analysis was simplified because there was no need to consider a non-trivial population of candidate solutions. For the $(1 \dagger \lambda)$-ES this is immediately obvious. For the $(\mu/\mu_I, \lambda)$-ES it is due to the fact that recombination in every time step computes the centroid of the parental population as the common starting point for all mutations.

The (μ, λ)-ES does not utilize recombination. If $\mu > 1$ its performance is considerably harder to analyze than that of the strategies that work on a trivial population of candidate solutions. It is not possible to consider single generations only. Instead, the population of candidate solutions that emerges in the course of evolution has to be modeled. In the absence of noise, Beyer [5] presented an analysis of the performance of the (μ, λ)-ES for spherically symmetric fitness function. A main result of his analysis, which was also stated by Rechenberg [11], is the observation that on the noise-free sphere the performance of the (μ, λ)-ES is never superior to that of the $(1, \lambda)$-ES, and thus no benefits can be gained from retaining any but the best candidate solution generated. However, Rechenberg [11] also provides empirical evidence that this is not true in the presence of noise. Simple numerical computer experiments can be used to demonstrate that in the very same fitness environment, significant speed-up factors over the $(1, \lambda)$-ES can be achieved by retaining more than just the (seemingly) best candidate solution if there is noise present.

In the present paper, we attempt to both shed some light on the reasons for the speed-up that can be achieved and to better understand its significance. For that purpose, in Section 2 we describe briefly the (μ, λ)-ES and outline and motivate the fitness environments for which its performance is analyzed. Local performance measures are introduced. In Section 3 a progress law for the (μ, λ)-ES in a linear fitness environment is derived that offers an intuitively appealing explanation for the speed-up. The population variance that appears as an important term in that progress law is studied empirically in Section 4. An attempt to obtain an analytical expression for the population variance based on a normal approximation proves to be rather inaccurate. In Section 5 we reduce the problem of obtaining a progress law for a spherically symmetric quadratic fitness function in high-dimensional search spaces to that of the linear fitness environment. Subsequently, the performance of the (μ, λ)-ES is compared with those of the $(1 + 1)$-ES and of the $(1, \lambda)$-ES. The benefits of maintaining a non-trivial population of candidate solutions are compared with those stemming from recombination. Finally, Section 6 concludes with a brief summary of the main results and with directions for future research.

2 Preliminaries

Evolution strategies together with fitness functions form iterated dynamical systems. The behavior of such systems is impossible to characterize analytically unless simplifying conditions regarding the strategies and/or the fitness environment hold true. Examples of such simplifying conditions are the assumption of linear or quadratic fitness functions, very high search space dimensionality, or the use of isotropic mutations or relatively simple forms of recombination.

Of course, real-world optimization problems rarely involve

0-7803-6657-3/01/$10.00 ©2001 IEEE

linear objective functions, and in many practical situations non-isotropic mutations are required for an ES to perform satisfactorily. However, even though the dynamical systems analyzed in theory are much simpler than those typically encountered in practical applications, theoretical analyses nonetheless have important implications for practitioners as they have the potential to reveal important scaling laws that can shed new light on design decisions that practitioners frequently face. Analytical considerations based on simple systems have led to prescriptions as useful as Rechenberg's 1/5th success rule [10], the discovery of the genetic repair principle [6], and to recommendations regarding the introduction of a selection threshold to improve the performance of the $(1 + 1)$-ES in the presence of noise [8]. We therefore feel justified in considering the following relatively simple strategy and fitness environment.

2.1 The (μ, λ)-ES

Given a fitness function $f : I\!R^N \to I\!R$, the (μ, λ)-ES at time step t maintains a population of μ candidate solutions $\mathbf{x}_i^{(t)} \in I\!R^N, i = 1, \ldots, \mu$. So as to obtain the population of candidate solutions at time step $t + 1$, λ new candidate solutions $\mathbf{y}_j^{(t)} \in I\!R^N$, $j = 1, \ldots, \lambda$, are generated by λ times independently picking one of the $\mathbf{x}_i^{(t)}$ and adding a *mutation vector* $\sigma^{(t)}\mathbf{z}_j^{(t)}$, where $\sigma^{(t)}$ is a strategy parameter referred to as the *mutation strength* and $\mathbf{z}_j^{(t)}$ is an realization of a random vector with N independent standard normally distributed components.

As indicated by the comma in "(μ, λ)" the population of candidate solutions at time step $t+1$ consists of those μ of the λ newly generated candidate solutions that score best in terms of the fitness function. A plus instead of the comma would indicate that the union of the parental and the offspring candidate solutions is used as the pool to select the population of the next time step from. Wherever possible without causing confusion, we will omit the time superscript in what follows.

2.2 The Fitness Environments

In this paper, we examine two separate fitness functions, a linear function

$$f_1 : I\!R^N \to I\!R$$
$$f_1(\mathbf{x}) = \mathbf{a}^{\mathrm{T}}\mathbf{x}, \tag{1}$$

where $\mathbf{a} \in I\!R^N$ is a parameter vector, and a spherically symmetric quadratic function

$$f_2 : I\!R^N \to I\!R$$
$$f_2(\mathbf{x}) = (\mathbf{x} - \hat{\mathbf{x}})^{\mathrm{T}}(\mathbf{x} - \hat{\mathbf{x}}) \tag{2}$$

that maps candidate solution $\mathbf{x}$ to the square of its Euclidean distance from the optimum at $\hat{\mathbf{x}}$. In both cases, without loss of generality the task at hand is minimization. In Section 5 we will see that the progress law for the quadratic sphere can, for high-dimensional search spaces, be derived from that for the linear function. The quadratic sphere is a model for unconstrained optimization problems at a stage where the search population is already in relatively close vicinity to the optimum.

Optimizing fitness functions as simple as f_1 and f_2 can be difficult if there is noise present. In this paper, we assume that exact fitness function values cannot be obtained, but that measuring fitness at search space location $\mathbf{x}$ yields a value that is normally distributed with mean $f(\mathbf{x})$ and with variance $\sigma_\epsilon^2(\mathbf{x})$. The standard deviation $\sigma_\epsilon(\mathbf{x})$ of the fitness measurements is referred to as the *noise strength*. For the linear fitness function we assume that the noise strength is independent of $\mathbf{x}$. For the quadratic sphere we consider the case that the noise strength is proportional to $f(\mathbf{x})$. Such relative errors of measurement are of great practical importance as they arise for example in connection with physical measurement devices that are accurate up to a certain percentage of the quantity they measure.

2.3 Measuring Performance

The local performance of ES can be measured in search space or in the space of fitness values. The corresponding performance measures are the *progress rate* and the *expected fitness gain* (or *quality gain*), respectively. The progress rate is defined as the expected change of the average distance to the optimum of the population at consecutive time steps. For fitness function f_1 that does not have an optimum it is defined as the expected distance traveled in direction of the gradient of the fitness function in a single time step. The expected fitness gain is the expected difference in average fitness of the population of candidate solutions at consecutive time steps.

3 Progress on the Linear Function

Determining the progress of a (μ, λ)-ES on the plane defined in Equation (1) really is a one-dimensional problem: changes in directions orthogonal to the gradient direction $\mathbf{a}$ are without influence on the fitness of the candidate solutions. Therefore, without loss of relevant information, all candidate solutions can be projected orthogonally onto a line defined by parameter vector $\mathbf{a}$. The parental population of candidate solutions is described appropriately by μ real numbers

$$x_i = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{x}_i}{\|\mathbf{a}\|}, \qquad i = 1, \ldots, \mu,$$

that indicate the positions of the projections of the $\mathbf{x}_i$ onto that line. Analogously, the set of offspring candidate solutions is appropriately described by projected variables

$$y_j = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{y}_j}{\|\mathbf{a}\|}, \qquad j = 1, \ldots, \lambda.$$

Given the parental $\mathbf{x}_i^{(t)}$ at time step t, the expected fitness gain is the expected value of

$$\begin{aligned} q &= \frac{1}{\mu}\sum_{i=1}^{\mu} f_1(\mathbf{x}_i^{(t)}) - \frac{1}{\mu}\sum_{i=1}^{\mu} f_1(\mathbf{x}_i^{(t+1)}) \\ &= \frac{\|\mathbf{a}\|}{\mu}\sum_{i=1}^{\mu} x_i^{(t)} - \frac{\|\mathbf{a}\|}{\mu}\sum_{i=1}^{\mu} x_i^{(t+1)} \\ &= \frac{\|\mathbf{a}\|}{\mu}\sum_{i=1}^{\mu} x_i^{(t)} - \frac{\|\mathbf{a}\|}{\mu}\sum_{k=1}^{\mu} y_{k;\lambda}^{(t)}, \end{aligned} \tag{3}$$

where $k;\lambda$ denotes the index of the offspring candidate solution with the kth highest perceived fitness. The progress rate is $\varphi = \mathrm{E}[q]/\|\mathbf{a}\|$.

Let $\langle x\rangle = \sum_{i=1}^{\mu} x_i^{(t)}/\mu$ and $D^2 = \sum_{i=1}^{\mu} {x_i^{(t)}}^2/\mu - \langle x\rangle^2$ denote the mean and the variance of the projected parental population of candidate solutions at time step t. As mutations are isotropic, the components in gradient direction of the mutations are normally distributed with mean zero and with variance σ^2. Thus, the mean of the set of projected offspring candidate solutions is $\langle x\rangle$ and its variance is $\sigma^2 + D^2$. Let us assume for the moment that the distribution of the projected offspring candidate solutions is normal. For $\mu = 1$ this is true as in that case $D^2 = 0$ and variation is solely due to mutations. In the general case, the distribution of the projected offspring candidate solutions is not exactly normal, but the results for the progress rate and the expected fitness gain that can be derived under the assumption of normality turn out to be quite accurate. For a discussion of the possibility of allowing for arbitrary distributions see Section 4.

Linear transformations preserve (or reverse) the order of a sample of observations of a random variable. Instead of selecting offspring candidate solutions y_j that are normally distributed with mean $\langle x\rangle$ and with variance $\sigma^2 + D^2$ at noise strength σ_ϵ we can introduce normalized variables

$$y_j^* = \frac{\langle x\rangle - y_j}{\sqrt{\sigma^2 + D^2}}, \qquad j = 1, \ldots, \lambda,$$

that are standard normally distributed and that are selected at noise strength $\sigma_\epsilon/\sqrt{\sigma^2 + D^2}$. It follows immediately from Equation (3) that the expected fitness gain of the strategy is the expected value of

$$q = \|\mathbf{a}\|\sqrt{\sigma^2 + D^2}\,\frac{1}{\mu}\sum_{k=1}^{\mu} y_{k;\lambda}^*, \tag{4}$$

where $k;\lambda$ is the index of the y_j^* with the kth highest perceived value.

According to a result derived in [2], the expected average of those μ of a sample of λ standard normally distributed values that have the highest perceived values, where noise of strength θ is interfering in the selection process, is

$$\frac{1}{\mu}\sum_{k=1}^{\mu}\int_{-\infty}^{\infty} x p_{k;\lambda}(x)\mathrm{d}x = \frac{e_{\mu,\lambda}^{1,0}}{\sqrt{1+\theta^2}}, \tag{5}$$

where $p_{k;\lambda}$ denotes the probability density function of the sample member with the kth highest perceived value and $e_{\mu,\lambda}^{1,0}$ is an instance of the general progress coefficients

$$e_{\mu,\lambda}^{\alpha,\beta} = \frac{\lambda-\mu}{\sqrt{2\pi}^{\alpha+1}}\binom{\lambda}{\mu}\int_{-\infty}^{\infty} x^{\beta}\mathrm{e}^{-\frac{\alpha+1}{2}x^2} [\Phi(x)]^{\lambda-\mu-1}\,[1-\Phi(x)]^{\mu-\alpha}\,\mathrm{d}x$$

introduced in [5] that can easily be obtained by numerical integration. It follows that

$$\mathrm{E}\left[\frac{1}{\mu}\sum_{k=1}^{\mu} y_{k;\lambda}^*\right] = \frac{e_{\mu,\lambda}^{1,0}}{\sqrt{1+\sigma_\epsilon^2/(\sigma^2+D^2)}}$$

and that according to Equation (4) the expected fitness gain and the progress rate can be written as

$$\begin{aligned} \mathrm{E}[q] &= \|\mathbf{a}\|\frac{\sqrt{\sigma^2+D^2}}{\sqrt{1+\sigma_\epsilon^2/(\sigma^2+D^2)}}e_{\mu,\lambda}^{1,0} \\ &= \|\mathbf{a}\|\sigma c_{\mu,\lambda}(\vartheta) \end{aligned} \tag{6}$$

and

$$\varphi = \sigma c_{\mu,\lambda}(\vartheta),$$

respectively. The progress coefficient in these relationships can be written as

$$c_{\mu,\lambda}(\vartheta) = \frac{1+\kappa_2}{\sqrt{1+\kappa_2+\vartheta^2}}e_{\mu,\lambda}^{1,0}, \tag{7}$$

where $\kappa_2 = D^2/\sigma^2$ and $\vartheta = \sigma_\epsilon/\sigma$ will be referred to as the *population variance* and the *noise level*, respectively.

For $\mu = 1$ the population variance is $\kappa_2 = 0$ and Equation (7) contains results from [4, 11] as a special case. For $\vartheta = 0$ it agrees with a result from [5]. Note that the population variance κ_2 remains as an unknown that is not determined easily. We will discuss an approach to computing it in Section 4 and content ourselves here with showing that it appropriately captures essential properties of the population distribution.

Figure 1 compares the progress coefficients computed using Equation (7) with empirical measurements. For the measurements, a (μ, λ)-ES was run at several noise levels with unit mutation strength on a linear function in one dimension. The mean progress of the population was averaged over 40,000 generations to obtain estimates of $c_{\mu,\lambda}(\vartheta)$. To verify the accuracy of the approach of assuming the population distribution to be normal, in those same experiments the population variance was averaged and used as an estimate for κ_2 in Equation (7). In general, it can be seen that the agreement between the two types of curves is quite good. The deviations are strongest for relatively high values of μ as well as for non-zero noise level in the range where the progress coefficients are maximal.

It can also be seen from Figure 1 that in the presence of noise much higher progress coefficients than those of the

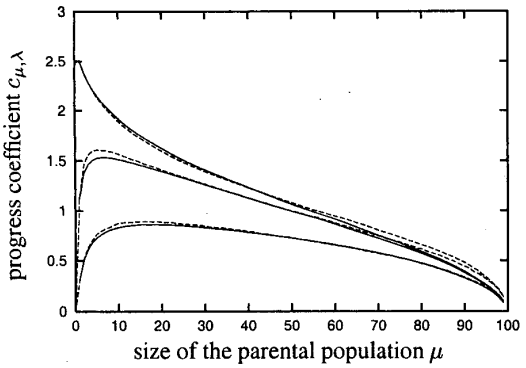

Figure 1: *Progress coefficients $c_{\mu,\lambda}(\vartheta)$ as a function of the size of the parental population μ for the (μ,λ)-ES on the plane with $\lambda = 100$. The solid lines correspond to, from top to bottom, measurements for noise levels $\vartheta = 0.0$, 2.0, and 8.0. The dashed lines mark the corresponding data obtained by measuring the variance of the parental population and using it in Equation (7) to obtain the progress coefficients.*

$(1,\lambda)$-ES can be achieved by choosing $\mu > 1$. For example, while the progress coefficient of the $(1,100)$-ES at noise level 8.0 is $c_{1,100}(8.0) \approx 0.31$, that of the $(22,100)$-ES is $c_{22,100}(8.0) \approx 0.87$. That is, progress rate and expected fitness gain are almost tripled by retaining more than just the seemingly best offspring candidate solution. Rechenberg [11] provides empirical evidence that at higher noise levels even greater speed up factors can be achieved.

Let us examine Equation (7) to determine the reason for the speed up that can be achieved. We will see in Section 4 that the population variance increases both with increasing size of the parental population μ and with increasing noise level ϑ. In the absence of noise we have the result $c_{\mu,\lambda}(0) = \sqrt{1+\kappa_2}e^{1,0}_{\mu,\lambda}$. When increasing the size of the parental population μ, the increase in $\sqrt{1+\kappa_2}$ due to an increase in population variance κ_2 is more than offset by the decrease in $e^{1,0}_{\mu,\lambda}$ if λ remains unchanged. The corresponding curve in Figure 1 is monotonically decreasing. In the presence of noise, however, not only are the population variances higher but also $1/\sqrt{1+\vartheta^2/(1+\kappa_2)}$ appears as an additional factor in Equation (7). While for the $(1,\lambda)$-ES $\kappa_2 = 0$ and the progress coefficient decreases with $1/\sqrt{1+\vartheta^2}$, for $\mu > 1$ the noise level is moderated by the non-zero population variance in the denominator under the square root. The strategy operates at a noise-to-signal ratio of $\vartheta/\sqrt{1+\kappa_2}$ as opposed to the noise-to-signal ratio of ϑ that strategies that have all mutations originate from a common parent operate at. As a consequence, for non-zero noise, the curves in Figure 1 have a maximum at intermediate values of μ. Optimal values of μ are typically in the range from about 0.1λ to 0.3λ.

4 Population Variance

The population variance κ_2 that appears in Equation (7) is a quantity that is hard to determine analytically. It is a product of the dynamical interaction between the (μ,λ)-ES and the fitness environment and is not the result of a single time step only but of the entire evolution history. Note that in [7] ($\sigma = 1$) it has been shown that $\mu - 1$ is an upper bound for the population variance that is attained if $\mu = \lambda$ or if $\vartheta \to \infty$.

An approach due to Beyer [5] attempts to find κ_2 by demanding that the expected population variance after a time step equals the population variance before that time step. The expected population variance after a time step is the expected value of

$$\begin{aligned}\frac{1}{\mu}\sum_{k=1}^{\mu}\left(y^*_{k;\lambda} - \frac{1}{\mu}\sum_{l=1}^{\mu}y^*_{l;\lambda}\right)^2 &= \frac{1}{\mu}\sum_{k=1}^{\mu}{y^*_{k;\lambda}}^2 - \frac{1}{\mu^2}\sum_{k=1}^{\mu}\sum_{l=1}^{\mu}y^*_{k;\lambda}y^*_{l;\lambda} \\ &= \frac{\mu-1}{\mu^2}\sum_{k=1}^{\mu}{y^*_{k;\lambda}}^2 - \frac{2}{\mu^2}\sum_{k=2}^{\mu}\sum_{l=1}^{k-1}y^*_{k;\lambda}y^*_{l;\lambda}, \quad (8)\end{aligned}$$

where $y^*_{k;\lambda}$ is the normalized and projected position of the offspring candidate solution with the kth highest perceived fitness.

The expected values of the sum and of the double sum can be computed in the same manner as the progress coefficient in [2]. As the calculations are rather lengthy we can only present the result here. Steps similar to those that have led to Equation (5) yield

$$\frac{1}{\mu}\sum_{k=1}^{\mu}\int_{-\infty}^{\infty}x^2 p_{k;\lambda}(x)\mathrm{d}x = 1 + \frac{e^{1,1}_{\mu,\lambda}}{1+\theta^2}$$

and

$$\frac{1}{\mu^2}\sum_{k=2}^{\mu}\sum_{l=1}^{k-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xy\,p_{kl;\lambda}(x,y)\mathrm{d}y\mathrm{d}x = \frac{\mu-1}{2\mu}\frac{e^{2,0}_{\mu,\lambda}}{1+\theta^2},$$

where again $p_{k;\lambda}$ is the probability density function of the sample member with the kth highest perceived value out of a sample of λ standard normally distributed values, and where $p_{kl;\lambda}$ is the joint probability density function of the member with the kth highest perceived value and that with the lth highest perceived value. It follows that

$$\mathrm{E}\left[\frac{1}{\mu}\sum_{k=1}^{\mu}{y^*_{k;\lambda}}^2\right] = 1 + \frac{e^{1,1}_{\mu,\lambda}}{1+\sigma_\epsilon^2/(\sigma^2+D^2)} \quad (9)$$

and

$$\mathrm{E}\left[\frac{1}{\mu^2}\sum_{k=2}^{\mu}\sum_{l=1}^{k-1}y^*_{k;\lambda}y^*_{l;\lambda}\right] = \frac{\mu-1}{2\mu}\frac{e^{2,0}_{\mu,\lambda}}{1+\sigma_\epsilon^2/(\sigma^2+D^2)}. \quad (10)$$

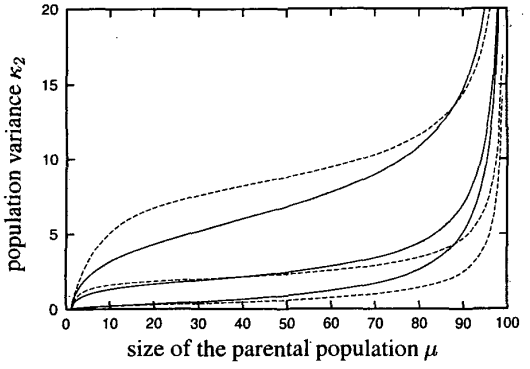

Figure 2: *Population variance κ_2 as a function of the size of the parental population μ for the (μ, λ)-ES on the plane with $\lambda = 100$. The solid lines correspond to, from bottom to top, measurements for noise levels $\vartheta = 0.0$, 2.0, and 8.0. The dashed lines mark the respective results from solving Equation (11) for the population variance.*

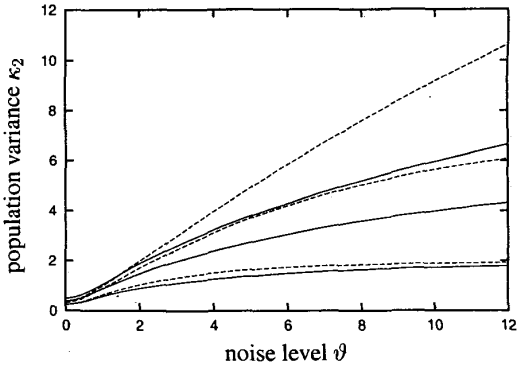

Figure 3: *Population variance κ_2 as a function of the noise level ϑ on the plane. The solid lines correspond to, from bottom to top, measurements for a $(3, 10)$-ES, a $(9, 30)$-ES, and a $(30, 100)$-ES. The dashed lines mark the respective results from solving Equation (11) for the population variance.*

The variance of the parental $\mathbf{x}_i$ is D^2. The variance of the selected offspring $\mathbf{y}_j$ is $(\sigma^2 + D^2)$ times the mean of the quantity from Equation (8). Demanding that the expected population variance after a time step equals the population variance before that time step thus means equating

$$D^2 = (\sigma^2 + D^2) \, \mathrm{E}\left[\frac{\mu - 1}{\mu^2} \sum_{k=1}^{\mu} {y^*_{k;\lambda}}^2 - \frac{2}{\mu^2} \sum_{k=2}^{\mu} \sum_{l=1}^{k-1} y^*_{k;\lambda} y^*_{l;\lambda}\right].$$

Using Equations (9) and (10) leads to the relationship

$$\kappa_2 = \frac{\mu - 1}{\mu}\left[1 + \kappa_2 - \frac{(1 + \kappa_2)^2}{1 + \kappa_2 + \vartheta^2}(e^{2,0}_{\mu,\lambda} - e^{1,1}_{\mu,\lambda})\right] \quad (11)$$

that can be solved for κ_2.

Figure 2 shows the dependency of the population variance κ_2 on the size of the parental population μ for $(\mu, 100)$-ES at different noise levels, Figure 3 illustrates its dependency on the noise level for different population sizes. It can be seen that there are considerable deviations between measured values for the population variance and those obtained from solving Equation (11) for κ_2. Generally, deviations seem to grow with increasing population variance. While predictions for low noise levels and for small population sizes are quite accurate, Equation (11) is of little value for high noise levels or if μ is too close to λ. However, results from Equation (11) agree with empirical measurements in that the population variance increases both with increasing size of the parental population and with increasing noise level.

The original approach pursued by Beyer [5] is more general in the sense that it does not assume normality of the distribution of the y^*_j. Instead, the unknown distribution is developed into a Gram-Charlier series. The coefficients in that series are related to the cumulants (the second of which is the variance) of the distribution. It is then demanded that the expected values of the cumulants after a time step agree with their values before that time step. The approach has been pursued for cumulants up to the third (the skewness of the distribution) and has been found to work reasonably well in the absence of noise. Even without noise, the calculations take up many pages; in the presence of noise, they become even more lengthy. Moreover, preliminary investigations suggest that in the presence of noise, considering only cumulants up to the third is not sufficient for obtaining a reasonably accurate approximation of the population parameters. At least the fourth cumulant (related to the kurtosis of the distribution) has to be included into the analysis in addition. It would be desirable to use an algebraic manipulation system to largely automate the process of finding the unknown cumulants, and we are currently pursuing that approach.

5 Progress on the Sphere

For very high parameter space dimension N the behavior of the (μ, λ)-ES on the sphere defined in Equation (2) can be characterized in terms of the progress coefficients determined in the previous sections. Unlike the situation on the plane defined in Equation (1), on the sphere changes in directions other than the local gradient direction are not without influence. However, they contribute a term that with increasing N tends to a constant.

5.1 A Progress Law

As mutations are isotropic and due to the spherical symmetry of the fitness environment and the fact that there is no interaction between different candidate solutions (such as recombination) other than through selection based on their fitness values, the problem of determining the expected fitness gain

or the progress rate on the sphere is again one-dimensional. The state of the population of parental candidate solutions at any point in time is described appropriately by μ real numbers R_i, $i = 1, \ldots, \mu$, the respective distances of the candidate solutions $\mathbf{x}_i$ from the optimum $\hat{\mathbf{x}}$.

The fitness gain q_j associated with a mutation vector $\sigma\mathbf{z}_j$ is the difference in fitness between the parental candidate solution $\mathbf{x}_i$ from which the mutation originates and that of the offspring candidate solution $\mathbf{y}_j = \mathbf{x}_i + \sigma\mathbf{z}_j$. As in [1, 2, 4, 11] we decompose mutation vectors into a component in direction of the local gradient and a perpendicular component. The length of the perpendicular component is for large N virtually independent of the particular mutation vector and contributes a term $-N\sigma^2$ to the fitness advantage. The component in direction of the local gradient contributes a term $-2\sigma\mathbf{z}_j^{\mathrm{T}}(\mathbf{x}_i - \hat{\mathbf{x}})$. As mutations are isotropic, $\mathbf{z}_j^{\mathrm{T}}(\mathbf{x}_i - \hat{\mathbf{x}})$ is normally distributed with mean zero and variance R_i^2, where $R_i = \|\mathbf{x}_i - \hat{\mathbf{x}}\|$. Therefore, the fitness gain associated with mutations applied to parent $\mathbf{x}_i$ is normally distributed with mean $-N\sigma^2$ and with standard deviation $2\sigma R_i$.

Let

$$R = \frac{1}{\mu}\sum_{i=1}^{\mu} R_i \quad \text{and} \quad D^2 = \frac{1}{\mu}\sum_{i=1}^{\mu} R_i^2 - R^2$$

denote mean and variance of the parental R_i. We make the assumption that $R \gg D$, i.e. that the distance from the optimum far exceeds the standard deviation of the population. This is usually the case unless the parameter space dimension is too small or the population size is too large. It then seems reasonable to assume that the fitness gain associated with single mutations is normally distributed with mean $-N\sigma^2$ and with standard deviation $2\sigma R$ independently of the parental candidate solutions to which they are applied. That is, the fitness advantage associated with mutation vector $\sigma\mathbf{z}_j$ is

$$q_j = 2\sigma R y_j^* - N\sigma^2,$$

where y_j^* is standard normally distributed. The fitness gain $q = \sum_{k=1}^{\mu} q_{k;\lambda}/\mu$ is the average fitness advantage associated with those offspring candidate solutions with the μ highest perceived values of y_j^*. The situation thus closely parallels that considered in Section 3, and in analogy to Equation (6) it follows

$$\mathrm{E}[q] = 2\sigma R \frac{1+\kappa_2}{\sqrt{1+\kappa_2+\vartheta^2}} e_{\mu,\lambda}^{1,0} - N\sigma^2$$

with $\kappa_2 = D^2/4\sigma^2R^2$ and $\vartheta = \sigma_\epsilon/2\sigma R$. Introducing normalized quantities

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

the expected normalized fitness gain can be written as

$$\mathrm{E}[q^*] = c_{\mu,\lambda}(\vartheta)\sigma^* - \frac{\sigma^{*2}}{2}, \tag{12}$$

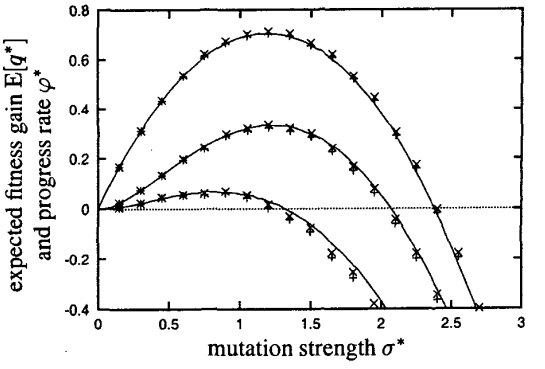

Figure 4: *Expected normalized fitness gain* $\mathrm{E}[q^*]$ *and normalized progress rate* φ^* *as functions of normalized mutation strength* σ^* *for a* $(3, 10)$*-ES on the quadratic sphere at, from top to bottom, normalized noise strengths* $\sigma_\epsilon^* = 0.0$, 2.0, *and* 4.0. *The solid lines mark results from Equation* (12), *the dots (*$\times$*: progress rate;* $+$*: expected fitness gain) data from real ES runs with search space dimension* $N = 40$.

where $\vartheta = \sigma_\epsilon^*/\sigma^*$ and the progress coefficient $c_{\mu,\lambda}$ was defined in Equation (7). Introducing the normalized progress rate $\varphi^* = \varphi N/R$ with slightly more effort it can be shown that $\varphi^* = \mathrm{E}[q^*]$.

Figure 4 shows the dependency of the expected normalized fitness gain and of the normalized progress rate on the normalized mutation strength for a $(3, 10)$-ES on the noisy quadratic sphere at different normalized noise strengths. The dots mark the results from ES runs in a 40-dimensional search space. The lines represent the approximation Equation (12) with measured progress coefficients. The agreement is quite good, but slightly deteriorates with increasing noise strength. Generally, the quality of the approximation improves with increasing search space dimension.

5.2 Comparing Efficiencies

Let us define the *efficiency* of an evolution strategy as the expected normalized fitness gain per fitness function evaluation in case of optimally adapted mutation strength. That is, the efficiency is defined as

$$\eta = \frac{1}{\lambda}\max_{\sigma^*}\left[\mathrm{E}[q^*]\right]. \tag{13}$$

As expected normalized fitness gain and normalized progress rate agree, we could as well have defined the efficiency via the progress rate instead of the expected fitness gain.

As the progress coefficients $c_{\mu,\lambda}(\vartheta)$ can be determined in computer simulations in one dimension only, they can be computed in great numbers and with satisfactory accuracy. Equation (12) can then be used to numerically determine optimal population sizes and maximal efficiencies that can be achieved. Figure 5 shows the optimal number of offspring $\hat{\lambda}$

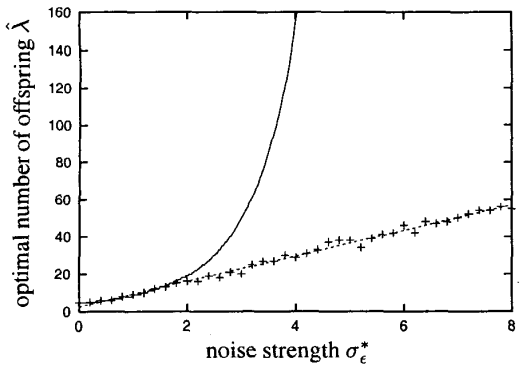

Figure 5: *Optimal number of offspring per generation $\hat{\lambda}$ as a function of normalized noise strength σ^*_ϵ. The solid line represents results for the $(1, \lambda)$-ES, the dots for the (μ, λ)-ES with optimally chosen μ. The straight dashed line is a least-squares fit to the data points for the (μ, λ)-ES.*

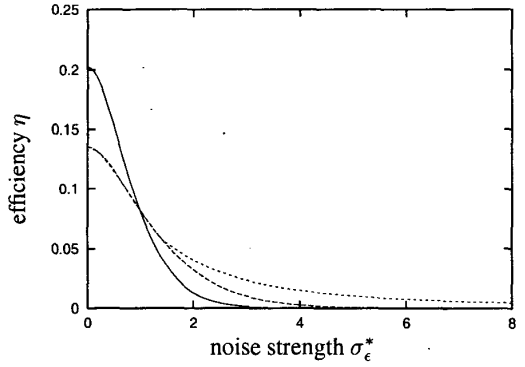

Figure 6: *Efficiency η as a function of normalized noise strength σ^*_ϵ. The solid line corresponds to the $(1 + 1)$-ES, the dashed line to the $(1, \lambda)$-ES with optimally chosen λ, and the dotted line to the (μ, λ)-ES with optimally chosen μ and λ.*

as a function of the normalized noise strength for the (μ, λ)-ES as well as for the $(1, \lambda)$-ES. It can be seen that, except for rather small noise strengths, the (μ, λ)-ES ideally operates with many fewer offspring candidate solutions than the $(1, \lambda)$-ES. The relationship between the optimal number of offspring candidate solutions of the (μ, λ)-ES and the normalized noise strength appears to be nearly linear.

Figure 6 compares the efficiency of the (μ, λ)-ES with optimally chosen population size parameters μ and λ with those of the $(1, \lambda)$-ES with optimally chosen λ and of the $(1 + 1)$-ES. The efficiency of the $(1 + 1)$-ES has been determined in [1]. It exceeds the performance of the other two strategies only up to a normalized noise strength of $\sigma^*_\epsilon \approx 1.0$ and is markedly inferior for higher noise strengths. Up to a normalized noise strength of about $\sigma^*_\epsilon \approx 1.4$ it is not useful to retain more than a single candidate solution and the curves for the (μ, λ)-ES and the $(1, \lambda)$-ES agree. Above this noise strength the efficiency of the (μ, λ)-ES can far exceed that of the $(1, \lambda)$-ES.

It is furthermore of interest to compare the performance of the (μ, λ)-ES with that of the $(\mu/\mu_I, \lambda)$-ES that uses intermediate multirecombination. The efficiency of the latter strategy on the noisy sphere in a sufficiently high-dimensional search space was found in [2] to be

$$\eta_{\mu/\mu_I,\lambda} = \frac{1}{\lambda}\left[\sigma^* \frac{1}{\sqrt{1+\vartheta^2}} e^{1,0}_{\mu,\lambda} - \frac{\sigma^{*2}}{2\mu}\right] \qquad (14)$$

Comparing this with the efficiency

$$\eta_{\mu,\lambda} = \frac{1}{\lambda}\left[\sigma^* \frac{1+\kappa_2}{\sqrt{1+\kappa_2+\vartheta^2}} e^{1,0}_{\mu,\lambda} - \frac{\sigma^{*2}}{2}\right] \qquad (15)$$

of the (μ, λ)-ES makes clear that the two strategies incorporate two rather different approaches to coping with noise. The $(\mu/\mu_I, \lambda)$-ES benefits from genetic repair that reduces the negative term in the efficiency law by virtue of the factor μ in the denominator. As a consequence, the strategy can be run at higher mutation strengths that result in a reduced noise-to-signal ratio. The (μ, λ)-ES on the other hand reduces the noise-to-signal ratio by means of a non-zero population variance that adds to the effect of mutations.

It is difficult to compare the relative significance of the two effects. As of today, there is no simple analytical expression for the population variance κ_2. Moreover, in practical problems, the finite parameter space dimension N can have a decisive influence on the performance of the strategies and render Equations (14) and (15) highly inaccurate. At first sight the benefits from genetic repair seem to be stronger than those of a non-zero population variance. However, this is not true if a mutation strength adaptation scheme such as *mutative self-adaptation* [12] that cannot take full advantage of genetic repair is employed. Especially in that case, strategies like the $(\mu/2, \lambda)$-ES in which only two parents participate in recombination leading to an offspring candidate solution and that lead to a non-zero population variance while making limited use of genetic repair may well be the best.

6 Summary and Outlook

The expected fitness gain of the (μ, λ)-ES on the noisy plane has been investigated. It was possible to give a simple progress law that can serve as an intuitively appealing explanation for the speed-up that can be achieved by retaining $\mu > 1$ candidate solutions, but that unfortunately contains the population variance κ_2 as an unknown. Determining how κ_2 scales with the population size and the noise level is a difficult problem for which an easy answer may not exist. An empirical investigation of the population variance has shown that it increases with both population size and noise level.

Subsequently, it has been demonstrated that the progress coefficients obtained on the plane can be used in a reasonably exact approximation to the performance of the (μ, λ)-ES on

the quadratic sphere. The performance of the (μ, λ)-ES has been compared with those of the $(1+1)$-ES and of the $(1, \lambda)$-ES. It has been found that the (μ, λ)-ES ideally uses much smaller population sizes than the $(1, \lambda)$-ES, and that above a certain noise level both the $(1 + 1)$-ES and the $(1, \lambda)$-ES are outperformed.

Future research will focus on finding a more accurate approximation for the population variance than that afforded by Equation (11). The approach pursued in [5] may be a starting point.

Acknowledgements

Financial support by the Deutsche Forschungsgemeinschaft (DFG) under grants Be 1578/4-2 and Be 1578/6-3 is gratefully acknowledged. Hans-Georg Beyer is a Heisenberg Fellow of the DFG.

Bibliography

[1] D. V. Arnold and H.-G. Beyer, "Local Performance of the $(1 + 1)$-ES in a Noisy Environment", Technical Report CI-80/00, SFB 531, Universität Dortmund, submitted for publication, (2000).

[2] D. V. Arnold and H.-G. Beyer, "Local Performance of the $(\mu/\mu_I, \lambda)$-ES in a Noisy Environment", in W. Martin and W. M. Spears, editors, *Foundations of Genetic Algorithms* 6, (Morgan Kaufmann, San Mateo, 2001).

[3] D. V. Arnold and H.-G. Beyer, "Efficiency and Self-Adaptation of the $(\mu/\mu_I, \lambda)$-ES in a Noisy Environment", in M. Schoenauer et al., editors, *Parallel Problem Solving from Nature* 6, pages 39-48, (Springer, Heidelberg, 2000).

[4] H.-G. Beyer, "Toward a Theory of Evolution Strategies: Some Asymptotical Results from the $(1 \dagger \lambda)$-Theory", *Evolutionary Computation* 1(2), pages 165-188, (1993).

[5] H.-G. Beyer, "Toward a Theory of Evolution Strategies: The (μ, λ)-Theory", *Evolutionary Computation* 2(4), pages 381-407, (1995).

[6] H.-G. Beyer, "Toward a Theory of Evolution Strategies: On the Benefit of Sex – the $(\mu/\mu, \lambda)$-Theory", *Evolutionary Computation* 3(1), pages 81-111, (1995).

[7] H.-G. Beyer, "On the Dynamics of EAs without Selection", in W. Banzhaf and C. Reeves, editors, *Foundations of Genetic Algorithms* 5, pages 5-26, (Morgan Kaufmann, San Mateo, 1999).

[8] S. Markon, D. V. Arnold, T. Bäck, T. Beielstein, and H.-G. Beyer, "Thresholding – a Selection Operator for Noisy ES", *Congress on Evolutionary Computation, CEC 01*, to appear, (2001).

[9] V. Nissen and J. Propach, "Optimization with Noisy Function Evaluations", in A. E. Eiben et al., editors, *Parallel Problem Solving from Nature* 5, pages 159-168, (Springer, Heidelberg, 1998).

[10] I. Rechenberg, *Evolutionsstrategie: Optimierung Technischer Systeme nach den Prinzipien der biologischen Evolution*, (Frommann-Holzboog, Stuttgart, 1973).

[11] I. Rechenberg, *Evolutionsstrategie '94*, (Frommann-Holzboog, Stuttgart, 1994).

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