ELSEVIER

Theoretical Computer Science 289 (2002) 629–647

Theoretical Computer Science

www.elsevier.com/locate/tcs

Performance analysis of evolution strategies with multi-recombination in high-dimensional $\mathbb{R}^N$-search spaces disturbed by noise ☆

Dirk V. Arnold*, Hans-Georg Beyer[1]

Department of Computer Science XI, University of Dortmund, 44221 Dortmund, Germany

Received August 2000; received in revised form July 2001; accepted August 2001
Communicated by G. Rozenberg

Abstract

The presence of noise in real-world optimization problems poses difficulties to optimization strategies. It is frequently observed that evolutionary algorithms are quite capable of succeeding in noisy environments. Intuitively, the use of a population of candidate solutions alongside with some implicit or explicit form of averaging inherent in the algorithms is considered responsible. However, so as to arrive at a deeper understanding of the reasons for the capabilities of evolutionary algorithms, mathematical analyses of their performance in select environments are necessary. Such analyses can reveal how the performance of the algorithms scales with parameters of the problem—such as the dimensionality of the search space or the noise strength—or of the algorithms—such as population size or mutation strength. Recommendations regarding the optimal sizing of such parameters can then be derived.

The present paper derives an asymptotically exact approximation to the progress rate of the $(\mu/\mu_I, \lambda)$-evolution strategy (ES) on a finite-dimensional noisy sphere. It is shown that, in contrast to results obtained in the limit of infinite search space dimensionality, there is a finite optimal population size above which the efficiency of the strategy declines, and that therefore it is not possible to attain the efficiency that can be achieved in the absence of noise by increasing the population size. It is also shown that nonetheless, the benefits of genetic repair and an increased mutation strength make it possible for the multi-parent $(\mu/\mu_I, \lambda)$-ES to far outperform simple one-parent strategies.

Keywords: Evolution strategies; Darwinian evolution; Noise; Optimization

☆ This research was supported by the Deutsche Forschungsgemeinschaft (DFG) under grants Be 1578/4-1 and Be 1578/6-1.

* Corresponding author.

E-mail addresses: arnold@ls11.cs.uni-dortmund.de (D.V. Arnold), beyer@ls11.cs.uni-dortmund.de (H.-G. Beyer).

[1] Heisenberg Fellow of the DFG.

0304-3975/02/$ - see front matter
PII: S0304-3975(01)00384-X

1. Introduction

Evolutionary algorithms (EAs) are optimization strategies based on evolutionary principles. Starting from an initial population of candidate solutions, increasingly better candidate solutions are developed by means of selection and variation of existing candidate solutions. Industrial applications date back at least until the 1960s and today range from routing optimization in telecommunications networks to airline crew scheduling problems. In many instances, EAs have turned out to be robust and applicable to challenging problems where traditional methods are prone to failure, such as optimization problems with highly discontinuous objective functions or where only unreliable data is available. Major reasons for the widespread use of EAs are their universal applicability and their ease of implementation that often outweighs possible performance deficits as compared to specialized algorithms that require long times of development. The 1990s have seen not only a sharp increase in applications of EAs, but the field has received considerable attention also from a theoretical point of view, as witnessed by a special issue of *Theoretical Computer Science* on evolutionary computation. The article by Eiben and Rudolph [10] therein can serve as a starting point with references to a long list of theoretical work on EAs.

Alongside with genetic algorithms (GAs), evolutionary programming (EP), and genetic programming (GP), evolution strategies (ES) are one kind of EA that is both in widespread practical use and relatively amenable to theoretical investigations. In particular, the field of order statistics has proven to be a useful mathematical tool for the analysis of ES in real-valued search spaces. The goal of research such as that presented here is to understand how the performance of ES scales with parameters of the problem—such as the dimensionality of the search space or the noise strength—and of the optimization strategy—such as the population size or the mutation strength of the strategy. At the focus of interest are local performance measures, i.e. performance measures that describe the expected change of quantities such as objective function values from one time step to the next. Note that this differs from more traditional approaches in theoretical computer science that usually focus on run time complexities or proofs of convergence. Such studies do exist (see for example [9,20]), and we believe that the different approaches each have their own merits and should be pursued in parallel. The insights gained from the approach taken here include a quantitative understanding of the performance of ES that is of immediate usefulness to practitioners that are facing the problem of choosing appropriate values for the external parameters of their strategies.

The present paper focuses on the behavior of ES in noisy environments. Noise is a common phenomenon in many real-world optimization problems. It can stem from a variety of sources, including measurement limitations, the use of randomized algorithms, incomplete sampling of large spaces, and human-computer interaction. Reduced convergence velocity or even inability to approach the optimum are commonly observed consequences of the presence of noise on optimization strategies. EAs are frequently reported to be comparatively robust with regard to the effects of noise. In fact, noisy environments are considered a prime application domain for EAs. Empirical support for this contention has been provided by Nissen and Propach [15] who have presented an

empirical comparison of population-based and point-based optimization strategies. They contend that population-based strategies generally outperform point-based strategies in noisy environments.

The effects of noise on the performance of GAs have been investigated by, among others, Fitzpatrick and Grefenstette [11], Miller and Goldberg [14], and Rattray and Shapiro [17]. Their work has led to recommendations regarding population sizing and the use of resampling. Theoretical studies of the effects of noise on ES date back to early work of Rechenberg [18] who has analyzed the performance of a $(1+1)$-ES in a noisy corridor. An analysis of the performance of the $(1\overset{+}{,}\lambda)$-ES on a noisy sphere by Beyer [5] sparked empirical research by Hammel and Bäck [13] who concluded that the findings made for GAs do not always easily translate to ES. Arnold and Beyer [4] have addressed the effects of overvaluation of the parental fitness using a $(1+1)$-ES on a noisy sphere. For an overview of the status quo of ES research concerned with noisy environments see [1,7].

The performance of the $(\mu/\mu_I,\lambda)$-ES on a noisy sphere in the limit of infinite parameter space dimension has been analyzed in a recent paper [3]. The analysis has led to concise laws for fitness gain and progress rate of the strategy that have attractive implications. It has been demonstrated that for large populations noise is all but removed. Genetic repair has been shown to be the source of the improved performance. In addition to its beneficial effect of statistical error correction first described by Beyer [6], in noisy environments it has the additional effect of favorably influencing the signal-to-noise ratio by allowing for the use of higher mutation strengths. A comparison with the $(1+1)$-ES, which is the most efficient ES on the sphere in the absence of noise, has revealed that already for relatively moderate noise strengths the simple strategy is outperformed by the multi-parent strategy even for small population sizes. Moreover, based on these results, in [2] it has been shown that increasing the population size is always preferable to averaging over a number of independent fitness function evaluations. This is an encouraging result as it shows that ES are indeed able to cope with noise in that it is better to let them deal with it than to explicitly remove it.

Unfortunately, as will be seen in Section 4.1, for finite parameter space dimension the accuracy of predictions afforded by the progress rate law obtained in [3] turns out to be not very good. Substantial deviations of experimental data obtained in ES runs from predictions made under the assumption of infinite parameter space dimension can be observed for all but the smallest population sizes. Thus it is not clear whether the implications of the results from [3] hold in practical situations. Clearly, there is a need for a parameter space dimension dependent progress rate formula that provides a better approximation to the local performance of the $(\mu/\mu_I,\lambda)$-ES for large but finite-dimensional search spaces. In the present paper we derive such an approximation and discuss its implications. In particular, it is examined whether the results obtained in the limit of infinite parameter space dimension qualitatively hold for finite-dimensional search spaces.

In Section 2 of this paper we give a brief description of the $(\mu/\mu_I,\lambda)$-ES algorithm and outline the fitness environment for which its performance is analyzed. Local performance measures are discussed. In Section 3 an approximation to the progress rate is developed. In the course of the derivation, a number of simplifications need to be

introduced to arrive at analytically solvable expressions. Section 4 provides experimental evidence that the accuracy of the result is satisfactory by comparing its predictions with data generated in real ES runs. Then we discuss results from numerical evaluations of the approximation to the progress rate. In particular, the residual location error and the performance of the strategy in case of optimally adapted strategy parameters are investigated. It is shown that the superiority of recombinative strategies as opposed to single-parent strategies observed in the limit of infinite parameter space dimension also holds in finite-dimensional search spaces, but that in contrast to infinite-dimensional search spaces it is not possible to obtain the same efficiency as in the absence of noise by increasing the population size. Section 5 concludes with a brief summary.

2. Algorithm, fitness environment, and performance evaluation

Section 2.1 describes $(\mu/\mu_I,\lambda)$-ES with isotropic normal mutations applied to optimization problems with an objective function of the form $f:\mathbb{R}^N\to\mathbb{R}$. Adopting EA terminology, we also refer to the objective function as *fitness function*. Without loss of generality, it can be assumed that the task at hand is minimization, i.e. that high values of f correspond to low fitness and vice versa. Section 2.2 outlines the fitness environment for which the performance of the algorithm is analyzed in the succeeding sections, and in Section 2.3 local performance measures are discussed.

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

As an evolutionary algorithm, the $(\mu/\mu_I,\lambda)$-ES strives to drive a population of candidate solutions to an optimization problem towards increasingly better regions of the search space by means of variation and selection. Fig. 1 illustrates the evolution loop that is cycled through repeatedly. The parameters λ and μ refer to the number of candidate solutions generated per time step and the number of those retained after selection, respectively. Obviously, λ is also the number of fitness function evaluations required

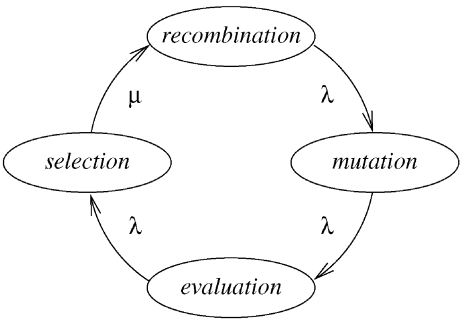

Fig. 1. The evolution loop. From a population of μ candidate solutions, $\lambda>\mu$ descendants are generated by means of recombination and subsequently subjected to mutation. Then, after evaluation of their fitness, μ of the descendants are selected to form the parental population of the succeeding time step.

per time step. We do not bother writing down initialization schemes and termination criteria as they are irrelevant for the analysis presented here.

Selection simply consists in retaining the μ best of the candidate solutions and discarding the remaining ones. The comma in $(\mu/\mu,\lambda)$ indicates that the set of candidate solutions to choose from consists only of the λ offspring, whereas a plus would indicate that selection is from the union of the set of offspring and the parental population. As shown in [4], plus-selection in noisy environments introduces overvaluation as an additional factor to consider, thus rendering the analysis considerably more complicated. Moreover, as detailed by Schwefel [21,22], comma-selection is preferable if the strategy employs mutative self-adaptation of the mutation strength, making it the more interesting variant to consider.

Variation is accomplished by means of recombination and mutation. As indicated by the second μ and the subscript I in $(\mu/\mu_I,\lambda)$, recombination is global intermediate. Mutations are isotropically normal. More specifically, let $\mathbf{x}^{(i)}\in\mathbb{R}^N$, $i=1,\ldots,\mu$, be the parameter space locations of the parent individuals. Recombination consists in computing the centroid

$$\langle\mathbf{x}\rangle=\frac{1}{\mu}\sum_{i=1}^{\mu}\mathbf{x}^{(i)}$$

of the parental population. For every descendant $\mathbf{y}^{(j)}\in\mathbb{R}^N$, $j=1,\ldots,\lambda$, a mutation vector $\mathbf{z}^{(j)}$, $j=1,\ldots,\lambda$, which consists of N independent, normally distributed components with mean 0 and variance σ^2, is generated and added to the centroid of the parental population. That is,

$$\mathbf{y}^{(j)}=\langle\mathbf{x}\rangle+\mathbf{z}^{(j)}.$$

The standard deviation σ of the components of the mutation vectors is referred to as the *mutation strength*.

2.2. The fitness environment

Finding analytical solutions describing the performance of ES is a hopeless task for all but the most simple fitness functions. In ES theory, a repertoire of fitness functions simple enough to be amenable to mathematical analysis while at the same time interesting enough to yield non-trivial results and insights has been established. The most commonplace of these fitness functions is the quadratic sphere

$$f(\mathbf{x})=\sum_{i=1}^{N}(\hat{x}_i-x_i)^2,$$

which maps vectors $\mathbf{x}=(x_1,\ldots,x_N)^{\mathrm{T}}\in\mathbb{R}^N$ to the square of their Euclidean distance to the optimum at parameter space location $\hat{\mathbf{x}}=(\hat{x}_1,\ldots,\hat{x}_N)^{\mathrm{T}}$. The sphere frequently serves as a model for fitness landscapes at a stage when the population of candidate solutions is already in close vicinity to the optimum. Other fitness landscapes such as the ridge analyzed by Oyman et al. [16] attempt to model features of fitness landscapes farther away from the optimum. Moreover, we contend that the fact that the sphere

scales uniformly in all directions in parameter space does not severely limit the value of the results derived below. According to results of Hansen and Ostermeier [12], an ES with non-isotropic mutations in combination with an advanced mutation strength adaptation algorithm such as the completely derandomized covariance matrix adaptation can transform arbitrary convex-quadratic functions into the sphere.

While finding the optimum of a convex-quadratic function such as the sphere is about the most easy task an optimization algorithm can face, this is no longer true if there is noise involved. In what follows it is assumed that evaluating the fitness of a candidate solution at parameter space location $\mathbf{x}$ is noisy in that its *perceived fitness* differs from its *ideal fitness* $f(\mathbf{x})$. This form of noise has been termed *fitness noise*. It deceives the selection mechanism as it can lead to inferior candidate solutions being selected based on their perceived fitness while superior ones are discarded. Fitness noise is commonly modeled by means of an additive, normally distributed random term with mean zero. That is, in a noisy environment, evaluation of the fitness function at parameter space location $\mathbf{x}$ yields perceived fitness $f(\mathbf{x}) + \sigma_\varepsilon z_\varepsilon$, where z_ε is a standard normally distributed random variate. Quite naturally, σ_ε is referred to as the *noise strength*.

2.3. Measuring performance

The local performance of the $(\mu/\mu_I, \lambda)$-ES can be measured either in parameter space or in fitness space. The corresponding performance measures are the *progress rate* and the *expected fitness gain*, sometimes referred to as *quality gain*, respectively. Let $(k;\lambda)$ denote the index of the offspring individual with the kth highest perceived fitness, and define the *progress vector* as

$$\langle \mathbf{z} \rangle = \frac{1}{\mu} \sum_{k=1}^{\mu} \mathbf{z}^{(k;\lambda)}.$$

Then, the centroid of the selected offspring and therefore of the parental population of the following time step is $\langle \mathbf{x} \rangle + \langle \mathbf{z} \rangle$. The expected fitness gain is the expected difference in fitness between the centroids of the population at consecutive time steps and therefore the expected value of

$$\Delta_f(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(\lambda)}) = f(\langle \mathbf{x} \rangle) - f(\langle \mathbf{x} \rangle + \langle \mathbf{z} \rangle).$$

The progress rate, denoted by φ, is the expected distance in direction of the location of the optimum traveled in parameter space by the population's centroid from one generation to the next and therefore the expected value of

$$\Delta_{\mathbf{x}}(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(\lambda)}) = \|\langle \mathbf{x} \rangle - \hat{\mathbf{x}}\| - \|\langle \mathbf{x} \rangle + \langle \mathbf{z} \rangle - \hat{\mathbf{x}}\|.$$

As detailed in [3], for the sphere the two performance measures can be made to agree in the limit $N \to \infty$ by introducing appropriate normalizations and can be computed exactly. For finite parameter space dimension N, the two performance measures differ and approximations need to be made to arrive at concise results. In what follows,

only the progress rate is considered as it is the more commonly considered quantity in previous studies of ES performance.

While local performance measures such as the expected fitness gain and the progress rate only describe expected rates of change from one generation to the next, under certain circumstances information regarding longer time spans can be derived. Assuming a mechanism for the adaptation of the mutation strength that assures a constant progress rate, Beyer [8] gives the law

$$R^{(g)} = R^{(0)} \exp\left(-\frac{\varphi^*}{N} g\right),$$

where g is the generation number, $R^{(g)}$ denotes the expected distance to the location of the optimum of the centroid of the population at generation g, and φ^* is the normalized progress rate to be computed below, for the sphere.

3. Performance

To obtain an approximation to the progress rate of the $(\mu/\mu_I, \lambda)$-ES on the noisy sphere, we proceed in three steps. In Section 3.1, we introduce a decomposition of mutation vectors that has proven useful in previous analyses. In Section 3.2, an approximation to the expected progress vector is computed. Finally, in Section 3.3, the approximation to the expected progress vector is used to obtain an approximation to the progress rate.

3.1. Decomposition of mutation vectors

The approximation to the progress rate of the $(\mu/\mu_I, \lambda)$-ES to be derived relies on a decomposition of mutation vectors suggested in both [5] and [19] and illustrated in Fig. 2. A mutation vector $\mathbf{z}$ originating at parameter space location $\mathbf{x}$ can be written as the sum of two vectors $\mathbf{z}_A$ and $\mathbf{z}_B$, where $\mathbf{z}_A$ is parallel to $\hat{\mathbf{x}} - \mathbf{x}$ and $\mathbf{z}_B$ is in the plane normal to that. In what follows, $\mathbf{z}_A$ and $\mathbf{z}_B$ are referred to as the A- and B-components of vector $\mathbf{z}$, respectively. Due to the isotropy of mutations, it can without loss of generality be assumed that $\mathbf{z}_A = \sigma(z_1, 0, \ldots, 0)^{\mathrm{T}}$ and $\mathbf{z}_B = \sigma(0, z_2, \ldots, z_N)^{\mathrm{T}}$, where the z_i, $i = 1, \ldots, N$, are independent, standard normally distributed random variates. Using elementary geometry and denoting the respective distances of $\mathbf{x}$ and $\mathbf{x} + \mathbf{z}$ to the location of the optimum by R and r, it can be seen from Fig. 2 that

$$\begin{aligned} r^2 &= (R - \sigma z_1)^2 + \|\mathbf{z}_B\|^2 \\ &= R^2 - 2R\sigma z_1 + \sigma^2 z_1^2 + \sigma^2 \sum_{i=2}^{N} z_i^2. \end{aligned} \tag{1}$$

At this point, let us make the following two assumptions:

(1) The summand $\sigma^2 z_1^2$ on the right-hand side of Eq. (1) can be neglected for performance calculations.

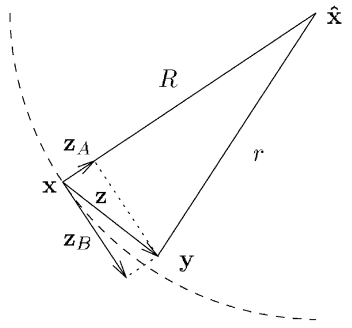

Fig. 2. Decomposition of a mutation vector $\mathbf{z}$ into two components $\mathbf{z}_A$ and $\mathbf{z}_B$. Vector $\mathbf{z}_A$ is parallel to $\hat{\mathbf{x}}-\mathbf{x}$, vector $\mathbf{z}_B$ is in the hyper-plane perpendicular to that. The starting and end points, $\mathbf{x}$ and $\mathbf{y}$, of the mutation are at distances R and r, respectively, from the location of the optimum.

(2) The summand $\|\mathbf{z}_B\|^2$ on the right-hand side of Eq. (1) can be modeled by a normal variate with mean $N\sigma^2$ and variance $2N\sigma^4$.

A few comments are in order to explain why these assumptions can be considered plausible. Let us take a look at the second assumption first. The summand $\|\mathbf{z}_B\|^2$ is the product of σ^2 and a sum that can for large N according to the Central Limit Theorem of Statistics be approximated by a normal variate. We choose to model the sum by means of a normal variate with mean N and variance $2N$ rather than with mean $N-1$ and variance $2(N-1)$ as assuming N rather than $N-1$ degrees of freedom is of little influence for large N and compensates for part of the error resulting from the first assumption. Note that taking the variance of $\|\mathbf{z}_B\|^2$ into account is what distinguishes the analysis below from the infinite-dimensional case considered in [3]. The accuracy of the predictions for finite-dimensional search spaces turns out to be greatly increased by this measure.

As for the first assumption, by neglecting the term quadratic in z_1 we overestimate the fitness of an individual as the term linear in z_1 favors individuals with a large positive z_1-component while the term quadratic in z_1 favors those individuals with a small absolute z_1-component. For small mutation strengths ($\sigma \ll R$), the linear term outweighs the quadratic one and the error introduced by the assumption is minor. For large mutation strengths ($\sigma \gg R$) this is not true. However, note that the mean of the first order statistic of λ independent realizations of a standard normal random variate grows no faster than the square root of the logarithm of λ. Therefore, the overall influence of the z_1-component on the fitness advantage associated with mutation vector $\mathbf{z}$ is rather minor for high mutation strengths and grows only slowly with λ. Therefore, the assumption can be expected not to introduce too large an error. We feel justified in making the assumptions by the good agreement of results obtained on the basis of the assumptions and data generated in ES runs detailed below.

Under the two assumptions, the square of the distance to the location of the optimum of point $\mathbf{x}+\mathbf{z}$ can be written as

$$r^2 \simeq R^2 - 2R\sigma z_1 + N\sigma^2 - \sqrt{2N}\sigma^2 z_B, \tag{2}$$

where

$$z_B \simeq \frac{N - \|\mathbf{z}_B\|^2/\sigma^2}{\sqrt{2N}} \tag{3}$$

is standard normally distributed and the relation is asymptotically exact in the limit $N \to \infty$ as proven in [3]. Let the *fitness advantage* associated with vector $\mathbf{z}$ be the difference in fitness

$$q(\mathbf{z}) = f(\mathbf{x}) - f(\mathbf{x} + \mathbf{z}) = R^2 - r^2.$$

Then, introducing normalizations

$$q^*(\mathbf{z}) = q(\mathbf{z})\frac{N}{2R^2} \quad \text{and} \quad \sigma^* = \sigma\frac{N}{R},$$

the normalized fitness advantage associated with vector $\mathbf{z}$ can be written as

$$q^*(\mathbf{z}) \simeq \sigma^* z_1 + \frac{\sigma^{*2}}{\sqrt{2N}} z_B - \frac{\sigma^{*2}}{2}. \tag{4}$$

That is, the normalized fitness advantage is for large N approximately normally distributed with mean $-\sigma^{*2}/2$ and variance $\sigma^{*2}(1 + \sigma^{*2}/2N)$. The benefit of the normalizations introduced above is now obvious: the distribution of the normalized fitness advantage is independent of the location in parameter space.

Selection is performed on the basis of perceived fitness rather than ideal fitness. With the definition of fitness noise from Section 2.2 and normalization

$$\sigma_\varepsilon^* = \sigma_\varepsilon \frac{N}{2R^2}, \tag{5}$$

the perceived normalized fitness advantage associated with mutation vector $\mathbf{z}$ is

$$q_\varepsilon^*(\mathbf{z}) = q^*(\mathbf{z}) + \sigma_\varepsilon^* z_\varepsilon \simeq \sigma^* z_1 + \frac{\sigma^{*2}}{\sqrt{2N}} z_B - \frac{\sigma^{*2}}{2} + \sigma_\varepsilon^* z_\varepsilon, \tag{6}$$

where z_ε is a standard normal random variate.

3.2. Computing the expected progress vector

The progress vector $\langle \mathbf{z} \rangle$ can be written as the sum of two vectors $\langle \mathbf{z}_A \rangle$ and $\langle \mathbf{z}_B \rangle$ in very much the same manner as shown above and as illustrated in Fig. 2 for mutation vectors. The A-component of the progress vector is the average of the A-components of the selected mutation vectors, and the B-component is the average of the B-components of the selected mutation vectors. The expected progress vector can be computed using a result derived in [3] that is quoted here for reference:

Theorem 1. *Let x_i, $i = 1, \ldots, \lambda$, be independent realizations of a standard normally distributed random variable, and let y_i, $i = 1, \ldots, \lambda$, be independent realizations of a normal random variable with mean zero and with variance θ^2. Then, letting $p_{k;\lambda}$*

denote the probability density function of the x_i with the kth largest value of $x_i + y_i$, the mean of the average of those μ of the x_i with the largest values of $x_i + y_i$ is

$$\overline{\langle x \rangle} = \frac{1}{\mu} \sum_{k=1}^{\mu} \int_{-\infty}^{\infty} x\, p_{k;\lambda}(x)\, \mathrm{d}x = \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+\theta^2}},$$

where

$$c_{\mu/\mu,\lambda} = \frac{\lambda-\mu}{2\pi} \binom{\lambda}{\mu} \int_{-\infty}^{\infty} \mathrm{e}^{-x^2} [\Phi(x)]^{\lambda-\mu-1} [1-\Phi(x)]^{\mu-1}\, \mathrm{d}x \tag{7}$$

is the $(\mu/\mu,\lambda)$-progress coefficient known from Beyer [6].

Here as in what follows, overlined quantities indicate expected values.

To compute the expected length of the A-component of the progress vector, clearly, the perceived normalized fitness advantage from Eq. (6) can be written as

$$q_\varepsilon^*(\mathbf{z}) \simeq \sigma^* \left(z_1 + \frac{\sigma^*}{\sqrt{2N}} z_B + \vartheta z_\varepsilon \right) - \frac{\sigma^{*2}}{2}$$

where $\vartheta = \sigma_\varepsilon^*/\sigma^*$ denotes the noise-to-signal ratio. As the selection mechanism is indifferent to the linear transformation, and as $\sigma^* z_B/\sqrt{2N} + \vartheta z_\varepsilon$ is normally distributed with mean zero and variance $\sigma^{*2}/2N + \vartheta^2$, it follows from Theorem 1 that the expected average of the z_1-components of the selected offspring and therefore the expected length of the A-component of the progress vector is

$$\overline{\|\langle \mathbf{z}_A \rangle\|} = \overline{\langle z_1 \rangle} = \frac{1}{\mu} \sum_{k=1}^{\mu} \int_{-\infty}^{\infty} x\, p_1^{(k;\lambda)}(x)\, \mathrm{d}x \simeq \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2+\sigma^{*2}/2N}}, \tag{8}$$

where $p_1^{(k;\lambda)}$ denotes the probability density function of the z_1 with the kth largest value of $z_1 + \sigma^* z_B/\sqrt{2N} + \vartheta z_\varepsilon$. Compared with the result obtained in [3] in the limit $N \to \infty$, the expected average of the z_1-components of the selected offspring is reduced by the additional third summand under the square root that can be traced back to the variance of $\|\mathbf{z}_B\|^2$.

For symmetry reasons, the orientation of the B-component of the progress vector is random in the plane with normal vector $(1,0,\ldots,0)^{\mathrm{T}}$. To compute its expected squared length, the perceived normalized fitness advantage from Eq. (6) can be written as

$$q_\varepsilon^*(\mathbf{z}) \simeq \frac{\sigma^{*2}}{\sqrt{2N}} \left(z_B + \frac{\sqrt{2N}}{\sigma^*} (z_1 + \vartheta z_\varepsilon) \right) - \frac{\sigma^{*2}}{2}.$$

As $\sqrt{2N}(z_1 + \vartheta z_\varepsilon)/\sigma^*$ is normally distributed with mean zero and with variance $2N(1+\vartheta^2)/\sigma^{*2}$, it follows from Theorem 1 that the expected average of the z_B values

of the selected offspring is

$$\overline{\langle z_B \rangle} = \frac{1}{\mu} \sum_{k=1}^{\mu} \int_{-\infty}^{\infty} x\, p_B^{(k;\lambda)}(x)\, \mathrm{d}x \simeq \frac{c_{\mu/\mu,\lambda}}{\sqrt{1 + 2N/\sigma^{*2} + 2N\vartheta^2/\sigma^{*2}}} = \frac{\sigma^*}{\sqrt{2N}} \frac{c_{\mu/\mu,\lambda}}{\sqrt{1 + \vartheta^2 + \sigma^{*2}/2N}},$$

where $p_B^{(k;\lambda)}$ is the probability density function of the z_B with the kth largest value of $z_B + \sqrt{2N}(z_1 + \vartheta z_\varepsilon)/\sigma^*$. Therefore, according to Eq. (3), the average accepted $\mathbf{z}_B$ vector has an expected squared length of

$$\langle \overline{\|\mathbf{z}_B\|^2} \rangle = \frac{1}{\mu} \sum_{k=1}^{\mu} \overline{\|\mathbf{z}_B^{(k;\lambda)}\|^2} \simeq \sigma^2 \left(N - \frac{c_{\mu/\mu,\lambda}\sigma^*}{\sqrt{1 + \vartheta^2 + \sigma^{*2}/2N}} \right).$$

The B-component of the progress vector is the average of the B-components of the selected offspring individuals. Its expected squared length is

$$\begin{aligned}
\overline{\|\langle \mathbf{z}_B \rangle\|^2} &= \frac{1}{\mu^2} \sum_{i=2}^{N} \overline{\left(\sum_{k=1}^{\mu} z_i^{(k;\lambda)} \right)^2} \\
&= \frac{1}{\mu^2} \sum_{i=2}^{N} \sum_{k=1}^{\mu} \overline{(z_i^{(k;\lambda)})^2} + \underbrace{\frac{1}{\mu^2} \sum_{i=2}^{N} \sum_{j \neq k} \overline{z_i^{(j;\lambda)} z_i^{(k;\lambda)}}}_{=0} \\
&= \frac{1}{\mu^2} \sum_{k=1}^{\mu} \overline{\|\mathbf{z}_B^{(k;\lambda)}\|^2} \\
&= \frac{1}{\mu} \langle \overline{\|\mathbf{z}_B\|^2} \rangle.
\end{aligned}$$

All summands in the second sum in the second line are zero as the B-components of the selected offspring individuals are independent. Thus, the expected squared length of the B-component of the progress vector is

$$\overline{\|\langle \mathbf{z}_B \rangle\|^2} \simeq \frac{\sigma^2}{\mu} \left(N - \frac{c_{\mu/\mu,\lambda}\sigma^*}{\sqrt{1 + \vartheta^2 + \sigma^{*2}/2N}} \right). \tag{9}$$

Compared to the result obtained in the limit $N \to \infty$, it is reduced by the second term by taking the variance of $\|\mathbf{z}_B\|^2$ into account. However, it will be seen that this is more than offset by the reduction of the expected length of the A-component outlined above.

The expected squared length of the B-component of the progress vector is reduced by a factor of μ as compared to the expected squared lengths of the B-components of the selected offspring. Beyer [6] has coined the term *genetic repair* for this phenomenon. Global intermediate recombination acts to dampen the "harmful" B-component of

mutation vectors with increasing μ while leaving the "beneficial" A-component virtually unchanged. As a result, the strategy can be run at much higher mutation strengths. The present analysis in combination with the discussion below shows that this is still true in the presence of noise.

3.3. Approximating the progress rate

The progress rate is the expected distance covered by the centroid of the population towards the location of the optimum in parameter space within a generation. That is, the progress rate φ is the expected value of $R - r$, where R and r are the distances to the location of the optimum of the centroid of the parental population and the centroid of the selected offspring, respectively. Let us introduce normalization

$$\varphi^* = \varphi \frac{N}{R}.$$

To obtain an approximation to the normalized progress rate, we make one more assumption:

(3) The progress rate can be approximated as the progress associated with the expected progress vector.

That is, we assume that fluctuations of the progress vector are of little importance for the result or even out. Under this assumption, using Eq. (1) for computing r and Eqs. (8) and (9) for the expected length of the A-component and the expected squared length of the B-component therein, and by making use of the simplifications provided by Assumptions (1) and (2) in Section 3.1, the normalized progress rate is

$$\begin{aligned}
\varphi^*_{\mu/\mu,\lambda} &\simeq N \left[1 - \sqrt{1 - \frac{2\sigma}{R}\langle z_1 \rangle + \frac{\overline{\|\langle \mathbf{z}_B \rangle\|^2}}{R^2}} \right] \\
&\simeq N \left[1 - \sqrt{1 + \frac{\sigma^{*2}}{\mu N} - \frac{2c_{\mu/\mu,\lambda}\sigma^*(1 + \sigma^{*2}/2\mu N)}{N\sqrt{1 + \vartheta^2 + \sigma^{*2}/2N}}} \right] \\
&= N \left[1 - \sqrt{1 + \frac{\sigma^{*2}}{\mu N}} \sqrt{1 - \frac{2c_{\mu/\mu,\lambda}\sigma^*(1 + \sigma^{*2}/2\mu N)}{N(1 + \sigma^{*2}/\mu N)\sqrt{1 + \vartheta^2 + \sigma^{*2}/2N}}} \right]. \qquad (10)
\end{aligned}$$

Linearizing the second square root yields

$$\varphi^*_{\mu/\mu,\lambda} \simeq N \left[1 - \sqrt{1 + \frac{\sigma^{*2}}{\mu N}} \left(1 - \frac{c_{\mu/\mu,\lambda}\sigma^*(1 + \sigma^{*2}/2\mu N)}{N(1 + \sigma^{*2}/\mu N)\sqrt{1 + \vartheta^2 + \sigma^{*2}/2N}} \right) \right],$$

and after rearranging terms, we obtain

$$\varphi^*_{\mu/\mu,\lambda} \simeq \frac{c_{\mu/\mu,\lambda}\sigma^* \left(1 + \sigma^{*2}/2\mu N\right)}{\sqrt{1 + \sigma^{*2}/\mu N}\sqrt{1 + \vartheta^2 + \sigma^{*2}/2N}} - N\left[\sqrt{1 + \frac{\sigma^{*2}}{\mu N}} - 1\right] \tag{11}$$

as an approximation to the progress rate of the $(\mu/\mu_I, \lambda)$-ES on the noisy sphere. Note the formal agreement of the result for $\vartheta = 0$ with the result obtained by Beyer [6] for the noise-free case.

4. Discussion

In Section 4.1 the accuracy of the result obtained in the previous section is tested by comparing with measured data from real ES runs. In Section 4.2 the conditions on noise strength and mutation strength under which positive progress can be expected are examined and the residual location error resulting from a fixed noise strength is discussed. Finally, in Section 4.3 the efficiency of the strategy in case of optimally adapted parameters is investigated.

4.1. Experimental verification

A number of assumptions and approximations that may cause doubts regarding the accuracy of the result have been made in the derivation of Eq. (11). In particular, Assumptions (1) through (3) in Sections 3.1 and 3.3 and the linearization of the square root in Section 3.3 have introduced errors. It is therefore necessary to evaluate the quality of the approximation by comparing with empirical measurements. Fig. 3 compares the results obtained from Eq. (11) with measurements of a $(8/8_I, 32)$-ES at parameter space dimension $N = 40$ for normalized noise strengths $\sigma^*_\varepsilon = 0.0$ and $\sigma^*_\varepsilon = 8.0$. All results have been obtained by averaging over 40,000 generations. The deviations

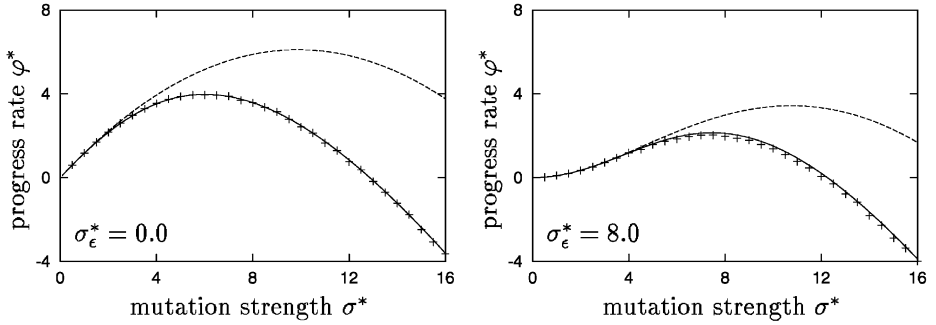

Fig. 3. Normalized progress rate φ^* as a function of normalized mutation strength σ^* for a $(8/8_I, 32)$-ES on a 40-dimensional noisy sphere. The noise strength is $\sigma^*_\varepsilon = 0.0$ in the left-hand graph and $\sigma^*_\varepsilon = 8.0$ in the right-hand graph. The solid lines display the result from Eq. (11), the crosses mark data generated in ES runs. The dashed lines represent the result for $N \to \infty$ obtained in [3].

Table 1
Absolute values of the relative error of Eq. (11). The four entries in each field of the table correspond to parameter space dimensions $N = 40$ (upper row) and $N = 400$ (lower row) and $(3/3_I, 10)$-ES (left column) and $(30/30_I, 100)$-ES (right column). The values have been obtained by averaging over 200,000 steps each.

	$\sigma_\varepsilon^* = 0.0$		$\sigma_\varepsilon^* = 4.0$		$\sigma_\varepsilon^* = 8.0$		$\sigma_\varepsilon^* = 16.0$	
$\sigma^* = 4.0$	0.039	0.001	0.027	0.017	0.040	0.018	0.032	0.007
	0.003	0.000	0.014	0.002	0.005	0.000	0.002	0.004
$\sigma^* = 8.0$	0.046	0.008	0.021	0.017	0.013	0.033	0.010	0.040
	0.015	0.000	0.004	0.002	0.001	0.004	0.000	0.004
$\sigma^* = 12.0$	0.019	0.018	0.015	0.023	0.011	0.034	0.006	0.052
	0.003	0.001	0.005	0.001	0.000	0.002	0.000	0.008
$\sigma^* = 16.0$	0.014	0.033	0.012	0.033	0.013	0.038	0.008	0.064
	0.002	0.000	0.003	0.001	0.001	0.005	0.001	0.010

between empirically observed values and computed values are minor. Note that as mentioned in the onset the agreement with results that have been obtained in the limit $N \to \infty$ in [3] is bad in the range of mutation strengths in which the performance is optimal.

Table 1 lists absolute values of the relative error of Eq. (11) for a number of parameter instances that are taken from across the spectrum of values that will be found to be of interest in the following section. It can be seen that errors of no more than 6.4% have been observed, and that in most instances the relative error is below 2%. Therefore, we do not expect qualitative differences between results derived on the basis of Eq. (11) and the behavior of the actual strategy.

4.2. Convergence properties

Positive progress towards the optimum can be expected only if the radicand in Eq. (10) is less than one. Straightforward calculation shows that this is the case if and only if

$$(2\mu c_{\mu/\mu,\lambda})^2 > \frac{\sigma_\varepsilon^{*2} + \sigma^{*2}(1 + \sigma^{*2}/2N)}{(1 + \sigma^{*2}/2\mu N)^2}. \tag{12}$$

Fig. 4 shows the maximal normalized mutation strength up to which positive progress can be expected as a function of normalized noise strength for a number of strategies and parameter space dimensionalities. In that figure, the expected progress is positive below the respective curves and negative above. It can be observed how increasing the population size increases the size of the region of positive expected progress. It can also be seen that the approximation obtained in [3] in the limit of infinite parameter space dimensionality becomes increasingly inaccurate with growing population size.

An interesting quantity to consider is the *residual location error* R_∞. It is the steady state value of the distance to the optimum approached by the ES after an infinite number of generations provided that the noise strength σ_ε is constant. It can be obtained by

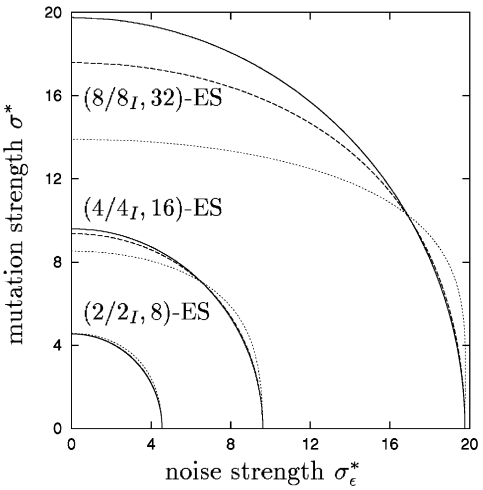

Fig. 4. Normalized mutation strength σ^* up to which the expected progress is positive as a function of normalized noise strength σ_ε^*. The curves correspond to a $(2/2_I, 8)$-ES, a $(4/4_I, 16)$-ES, and a $(8/8_I, 32)$-ES for parameter space dimensions $N = 40$ (dotted lines), $N = 400$ (dashed lines), and $N \to \infty$ (solid lines). The expected progress is positive for $(\sigma_\varepsilon^*, \sigma^*)$ combinations below the respective curves and negative for combinations above.

replacing the inequality operator in Eq. (12) by an equality, using Eq. (5), and solving for R, resulting in

$$R_\infty \simeq \sqrt{\frac{N\sigma_\varepsilon}{4\mu c_{\mu/\mu,\lambda}} \left[\left(1 + \frac{\sigma^{*2}}{2\mu N}\right)^2 - \left(\frac{\sigma^*}{2\mu c_{\mu/\mu,\lambda}}\right)^2 \left(1 + \frac{\sigma^{*2}}{2N}\right) \right]^{-1/2}}.$$

For vanishing normalized mutation strength, it follows

$$R_\infty \simeq \sqrt{\frac{N\sigma_\varepsilon}{4\mu c_{\mu/\mu,\lambda}}},$$

showing how the residual location error can be reduced by increasing the population size.

4.3. Optimizing the efficiency

Let $\hat{\varphi}$ denote the progress rate in case of optimally adapted mutation strength. Taking the number of fitness function evaluations as a measure for the computational costs of an optimization algorithm, the *efficiency* of the $(\mu/\mu_I, \lambda)$-ES is defined as the maximal expected progress per fitness function evaluation

$$\eta = \frac{\hat{\varphi}^*}{\lambda}.$$

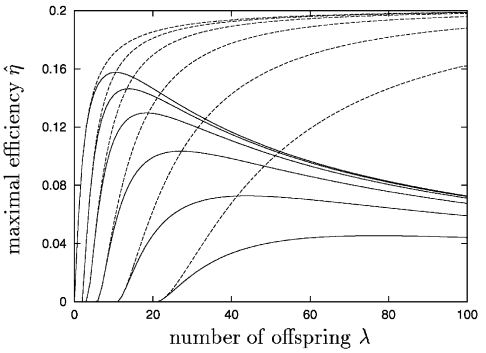

Fig. 5. Maximal efficiency $\hat{\eta}$ as a function of the number of offspring per generation λ for, from top to bottom, normalized noise strengths $\sigma_\varepsilon^* = 0.0,\ 1.0,\ 2.0,\ 4.0,\ 8.0$, and 16.0. The solid curves display results for parameter space dimension $N = 40$, the dashed curves are the corresponding results obtained for $N \to \infty$ in [3].

Increasing the number of offspring λ per generation is useful only if increased efficiency is a consequence. In what follows, the efficiency for an optimally chosen number of parents μ given a number of offspring λ is denoted by $\hat{\eta}$. Both μ and λ are treated as real-valued parameters where numerical optimization methods are used to find optimal values. Naturally, rounding to integer numbers is necessary to arrive at real strategies.

Fig. 5 shows the dependency of the optimal efficiency $\hat{\eta}$ on the number of offspring per generation λ. It has been obtained by numerically solving Eq. (11) for optimal values of the normalized mutation strength σ^* and of the size of the parental population μ. It can be seen that in contrast to the results obtained for $N \to \infty$ in [3], for finite parameter space dimension there is an optimal number of offspring above which the efficiency of the strategy declines. The choice of the parameter λ of the strategy becomes less critical with increasing noise strength as the maximum becomes less pronounced. It can also be seen from Fig. 5 that for finite N, in contrast to the infinite-dimensional case, the efficiency cannot be increased to the maximal value that can be achieved in the absence of noise by increasing the population size.

In Fig. 6, the optimal number of offspring per generation $\hat{\lambda}$ is shown a function of the normalized noise strength σ_ε^* and as a function of the search space dimension N. The values have been obtained by numerical optimization of Eq. (11). The curves show that the optimal number of offspring increases with increasing noise strength and with increasing search space dimension and demonstrate that overall, optimal values of λ are relatively small compared to the search space dimension N. Note that in the limit $N \to \infty$, increasing λ is always beneficial. An interesting aspect not shown in the figure is that for all search space dimensions and all noise strengths examined, the optimal truncation ratio μ/λ in case of optimally chosen λ is very close to 0.27, the value that is known to be optimal in the absence of noise in the limit $N \to \infty$.

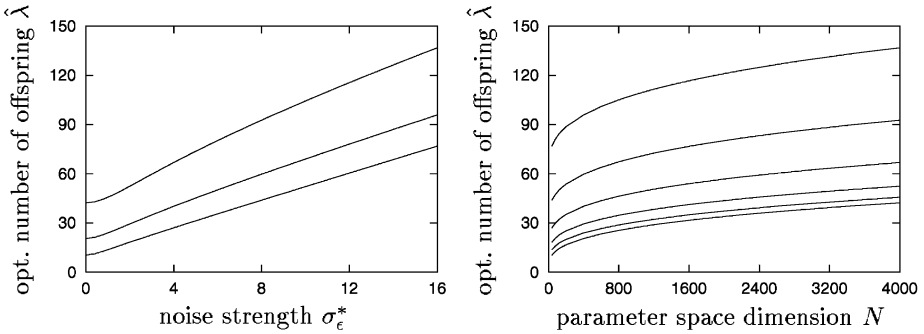

Fig. 6. Optimal number of offspring per generation $\hat{\lambda}$ as a function of normalized noise strength σ_ε^* and as a function of parameter space dimension N. The curves in the left-hand graph correspond to, from bottom to top, parameter space dimensions $N = 40$, $N = 400$, $N = 4000$. The curves in the right-hand graph correspond to, from bottom to top, normalized noise strengths $\sigma_\varepsilon^* = 0.0,\ 1.0,\ 2.0,\ 4.0,\ 8.0$, and 16.0.

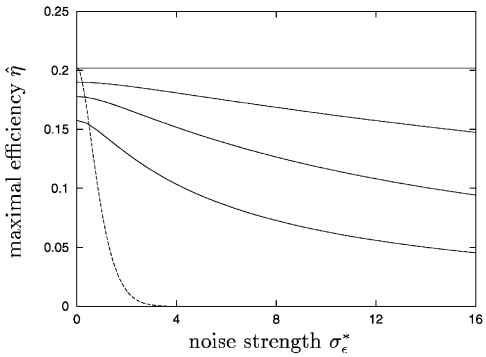

Fig. 7. Maximal efficiency $\hat{\eta}$ as a function of normalized noise strength σ_ε^*. The solid curves correspond to, from bottom to top, parameter space dimensions $N = 40$, $N = 400$, $N = 4000$, and the limiting case $N \to \infty$. The dashed curve represents the result for the $(1 + 1)$-ES derived in [4].

In Fig. 7, the maximal efficiency $\hat{\eta}$ is shown as a function of normalized noise strength σ_ε^*. The curves show a clear decline of the maximal efficiency with increasing noise strength if the parameter space dimensionality is finite. However, except for very low noise strength, the maximal efficiency of the $(1 + 1)$-ES that has been derived in [4] and that is included in the graph for reference is far exceeded. The reason for this gain in efficiency is the presence of genetic repair in the multi-parent strategy. In the noisy environment it not only results in statistical error correction but also affords the additional benefit of an increased signal-to-noise ratio by means of increased mutation strengths. The strong preference for recombinative multi-parent strategies over point-based strategies in noisy environments that has been found in [3] in the limit of an infinite-dimensional search space thus holds in finite-dimensional search spaces as well.

5. Conclusion

An asymptotically exact approximation to the progress rate of the $(\mu/\mu_I, \lambda)$-ES on a noisy sphere has been developed for large but finite parameter space dimensionality. The accuracy of the approximation has been verified numerically. It has been shown that the progress rate law derived in [3] in the limit of infinite parameter space dimensionality is insufficient to characterize the behavior of the strategy in finite-dimensional search spaces. While it has been demonstrated that in contrast to the infinite-dimensional case the efficiency attainable in the absence of noise cannot be achieved by increasing the population size if the parameter space dimensionality is finite, the superiority of recombinative multi-parent strategies over point-based strategies in noisy environments has been confirmed. It has been demonstrated that there is an optimal population size above which the efficiency of the algorithm declines, and that with increasing noise strength, the optimal population size increases while its choice becomes less critical.

Future work includes analyzing the effects of noise on self-adaptive mechanisms. Adaptation of the mutation strength is crucial for the performance of the ES, and the influence of noise on self-adaptive mechanisms is largely ununderstood. It is expected that the ground work laid in this article forms a corner stone in the analysis of self-adaptive ES in noisy environments.

References

[1] D.V. Arnold, Evolution strategies in noisy environments—a survey of existing work, in: L. Kallel, B. Naudts, A. Rogers (Eds.), Theoretical Aspects of Evolutionary Computing, Springer, Berlin, 2001, pp. 239–249.

[2] D.V. Arnold, H.-G. Beyer, Efficiency and self-adaptation of the $(\mu/\mu_I, \lambda)$-ES in a noisy environment, in: M. Schoenauer, et al. (Eds.), Parallel Problem Solving from Nature, Vol. 6, Springer, Heidelberg, 2000, pp. 39–48.

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

[4] D.V. Arnold, H.-G. Beyer, Local Performance of the $(1+1)$-ES in a noisy environment, IEEE Trans. Evolution. Comput. (2002), to appear.

[5] H.-G. Beyer, Toward a theory of evolution strategies: some asymptotical results from the $(1^+, \lambda)$-theory, Evolution. Comput. 1 (1993) 165–188.

[6] H.-G. Beyer, Toward a theory of evolution strategies: on the benefit of sex—the $(\mu/\mu, \lambda)$-theory, Evolution. Comput. 3 (1995) 81–111.

[7] H.-G. Beyer, Evolutionary algorithms in noisy environments: theoretical issues and guidelines for practice, Comput. Methods Mechan. Appl. Eng. 186 (2000) 239–267.

[8] H.-G. Beyer, The Theory of Evolution Strategies, Springer, Heidelberg, 2001.

[9] S. Droste, T. Jansen, I. Wegener, A rigorous complexity analysis of the $(1+1)$ evolutionary algorithm for separable functions with Boolean inputs, Evolution. Comput. 6 (1998) 185–196.

[10] A.E. Eiben, G. Rudolph, Theory of evolutionary algorithms: a bird's eye view, Theoret. Comput. Sci. 229 (1999) 3–9.

[11] J.M. Fitzpatrick, J.J. Grefenstette, Genetic Algorithms in Noisy Environments, in: P. Langley (Ed.), Machine Learning, Vol. 3, Kluwer, Dordrecht, 1988, pp. 101–120.

[12] N. Hansen, A. Ostermeier, Completely derandomized self-adaptation in evolution strategies, Evolution. Comput. 9 (2001) 159–195.

[13] U. Hammel, T. Bäck, Evolution strategies on noisy functions. How to improve convergence properties, in: Y. Davidor, R. Männer, H.-P. Schwefel (Eds.), Parallel Problem Solving from Nature, Vol. 3, Springer, Heidelberg, 1994, pp. 159–168.
[14] B.L. Miller, D.E. Goldberg, Genetic algorithms, selection schemes, and the varying effects of noise, Evolution. Comput. 4 (1997) 113–131.
[15] V. Nissen, J. Propach, Optimization with noisy function evaluations, in: A.E. Eiben, T. Bäck, M. Schoenauer, H.-P. Schwefel (Eds.), Parallel Problem Solving from Nature, Vol. 5, Springer, Heidelberg, 1998, pp. 159–168.
[16] A.I. Oyman, H.-G. Beyer, H.-P. Schwefel, Where elitists start limping: evolution strategies at ridge functions, in: A.E. Eiben, T. Bäck, M. Schoenauer, H.-P. Schwefel (Eds.), Parallel Problem Solving from Nature, Vol. 5, Springer, Heidelberg, 1998, pp. 34–43.
[17] M. Rattray, J. Shapiro, Noisy fitness evaluation in genetic algorithms and the dynamics of learning, in: R.K. Belew, M.D. Vose (Eds.), Foundations of Genetic Algorithms, Vol. 4, Morgan Kaufmann, San Mateo, 1997, pp. 117–139.
[18] I. Rechenberg, Evolutionsstrategie: Optimierung Technischer Systeme nach den Prinzipien der biologischen Evolution, Frommann-Holzboog, Stuttgart, 1973.
[19] I. Rechenberg, Evolutionsstrategie '94, Frommann-Holzboog, Stuttgart, 1994.
[20] G. Rudolph, Convergence Properties of Evolutionary Algorithms, Dr. Kovač, Hamburg, 1997.
[21] H.-P. Schwefel, Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie, Birkhäuser, Basel, 1977.
[22] H.-P. Schwefel, Evolution and Optimum Seeking, Wiley, New York, 1995.