Evolutionary Gradient Search Revisited

Dirk V. Arnold

Technical Report CS-2005-09

July 8, 2005

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

Evolutionary Gradient Search Revisited

Dirk V. Arnold

***Abstract*— Evolutionary gradient search is an approach to optimization that combines features of gradient strategies with ideas from evolutionary computation. Recently, several modifications to the algorithm have been proposed with the goal of improving its robustness in the presence of noise and its suitability for implementation on parallel computers. In this paper the value of the proposed modifications is studied analytically. A scaling law is derived that describes the performance of the algorithm on the noisy sphere model and allows comparing it with competing strategies. The comparisons yield insights into the interplay of mutation, multirecombination, and selection. Then, the covariance matrix adaptation mechanism originally formulated for evolution strategies is adapted for use with evolutionary gradient search in order to make the algorithm competitive on objective functions with a large condition numbers of their Hessians. The resulting strategy is evaluated experimentally on a number of convex quadratic test functions.**

***Index Terms*— Evolutionary gradient search, evolution strategies, quality gain analysis, noise, covariance matrix adaptation.**

I. Introduction

Evolution strategies [1], [2], [3], [4] and gradient search algorithms [5] are approaches to numerical optimization that differ fundamentally in several important aspects. Salomon's evolutionary gradient search (EGS) [6] attempts to blend features of the two in order to combine the efficiency of gradient search strategies with the relative robustness of evolutionary approaches. The contributions of the present paper are twofold. First, the basic iterative steps of EGS are studied in order to learn about the influence of the strategy's parameters on optimization performance and to be able to draw comparisons with related evolution strategies. Recommendations with regard to the setting of strategy parameters are also given. And second, in an effort to achieve competitive performance on functions with different eigenvalue spectra of their Hessians, the covariance matrix adaptation mechanism devised by Hansen and Ostermeier [7] is adapted for use in EGS. The resulting strategy could be called CMA-EGS. Numerical experiments on several objective functions show that the insights gained in the analysis of the basic iterative steps are indeed useful under much more general conditions than those that they have been obtained under.

A. Basic Iterative Steps

The most basic gradient-based optimization strategy for the minimization of functions $f : \mathbb{R}^N \rightarrow \mathbb{R}$ is the method of steepest descent. It iteratively updates a search point by stepping in the direction of the negative of the gradient of the objective function at the current location in search space. If gradient information cannot be obtained directly, gradient strategies compute a finite-difference approximation to the gradient [8]. That approximation is obtained by evaluating the objective function at a number of trial points close to the current search point, typically by stepping in direction of the coordinate axes. Depending on whether forward differences or central differences are employed, evaluation of the objective function at either $N + 1$ or $2N$ points is required to obtain a gradient estimate, where N is the search space dimensionality. Provided that the objective function is locally smooth, the gradient estimate can be made to match the gradient with any degree of accuracy by making the trial steps sufficiently small.

Evolution strategies differ from gradient search algorithms both in their use of a population of search points and in the fact that rather than employing a regular pattern of trial steps, trial points (often referred to as offspring) are generated randomly. A search direction is inferred implicitly by the selection of favorable trial points. Those trial points that are selected form the population of search points in the next time step. Typically, selection is based on rank within the set of offspring, and objective function values are used only to establish a ranking. Another fact that distinguishes evolution strategies from gradient search algorithms is that for the former, the number of trial points generated per time step is flexible. There is no requirement that the number of offspring be $N + 1$ or $2N$. Indeed, in the context of local optimization, evolution strategies often rely on far fewer than N trial points per time step. As a consequence, the direction in which the population of search points moves may differ substantially from the (negative of the) gradient direction. However, many "cheap" steps computed on the basis of a relatively small number of objective function evaluations may well be more effective in their sum than a few expensive ones. While Rechenberg [1] characterizes the resulting search process as a form of "stochastic gradient descent", Beyer [9] points out that the search behavior of evolution strategies can be qualitatively different from gradient descent in that it is much more exploratory in nature.

The EGS procedure introduced by Salomon [6] is a hybrid strategy that combines features of gradient-based algorithms and evolution strategies. As evolution strategies, EGS generates trial points by random mutations, and the number of trial points generated per time step is flexible. As gradient strategies, it operates with a single search point rather than with a population, and it uses objective function values rather than merely ordinal information for inferring search directions. Provided that the variance of the probability distribution used to generate trial points is sufficiently small, the gradient direction is approached with increasing accuracy as the number of trial points generated per time step increases. In an experimental evaluation presented in [6], EGS performed superior to several

D. V. Arnold is with the Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada B3H 1W5.

variants of evolution strategies on a number of test functions.

In [10], a close connection between EGS and the $(\mu/\mu, \lambda)$-ES, an evolution strategy that is popular both for its amenability to theoretical analysis and for its good performance, was pointed out. See [4] for a taxonomy of evolution strategies. The form of recombination used by the $(\mu/\mu, \lambda)$-ES leads to the population effectively being contracted to a single search point that new trial points are generated from in every time step. The basic iterative steps that EGS and the $(\mu/\mu, \lambda)$-ES perform differ only in the way that the information gained from evaluating the trial points is used. The $(\mu/\mu, \lambda)$-ES uses objective function values only in order to obtain a ranking of the λ trial points generated. It then discards all but the μ best of them, and it applies global intermediate recombination — a simple, unweighted averaging — to the remaining ones. EGS on the other hand uses a *weighted* average of *all* λ trial points in order to derive a new search point. The weight of a trial point is determined by the difference between its objective function value and the objective function value of the search point it was generated from, and it can be negative if the trial point is inferior to the search point.

The similarity between EGS and the $(\mu/\mu, \lambda)$-ES makes it possible to use the tools developed for the analysis of evolution strategies when studying properties of EGS. In [10] the performance of the basic EGS steps with isotropic mutations was analyzed on the infinite-dimensional sphere model and compared with that of the $(\mu/\mu, \lambda)$-ES. It was seen that while EGS can offer a significant performance advantage compared to the $(\mu/\mu, \lambda)$-ES, it does have deficiencies when implemented on parallel computers as well as in the presence of noise. In particular, while the serial efficiency of the $(\mu/\mu, \lambda)$-ES increases with increasing population size (see [3]), that of EGS peaks at $\lambda = 5$ and decreases for larger numbers of trial points generated per time step. The potential speed-up resulting from evaluating trial solutions in parallel is sublinear for EGS while it is (slightly) superlinear for the $(\mu/\mu, \lambda)$-ES. For finite N it can be seen in experiments that the parallel performance advantage of the $(\mu/\mu, \lambda)$-ES for large numbers of trial points generated per time step is not as pronounced as for $N \to \infty$, but it is nonetheless present.

Moreover, it was seen in [10] that the performance of EGS degrades with increasing amounts of noise present. The reason for the better search capabilities in the presence of noise of the $(\mu/\mu, \lambda)$-ES was found not to be that strategy's reliance on ordinal information rather than function values, but instead its better use of the genetic repair effect. As seen in [1], [3], [11], the $(\mu/\mu, \lambda)$-ES benefits from multirecombination on the sphere model in that the averaging of candidate solutions implicit in the recombination procedure reduces the "harmful" components of mutation vectors. As a consequence, the search steps made by the strategy are shorter than its trial steps, and the strategy is able to operate with mutation strengths larger than those of a comparable strategy that does not use multirecombination. In [12] it was seen that in the presence of noise, increased mutation strengths have the additional benefit of reducing the noise-to-signal ratio that the strategy operates under. EGS as introduced in [6] explicitly ensures that trial steps and search steps are of the same length. As a result, it is not capable of operating with mutation strengths as large as those of a comparable $(\mu/\mu, \lambda)$-ES, and it thus suffers from a higher noise-to-signal ratio.

In light of these deficiencies, in a recent paper Salomon [13] has proposed two important modifications to the evolutionary gradient search procedure. Recognizing that the inferior parallel performance of EGS is due to a negative bias resulting from the fact that typically, the majority of trial points is inferior to the search point they originate from, Salomon suggests the use of "inverse mutations". That is, for every trial step, a symmetric trial step is made in the opposite direction. In the subsequent averaging, the weight associated with the step is chosen proportional to the difference between the function values of the two trial points and is thus independent of the function value of the search point. This procedure is reminiscent of the use of central differencing instead of forward differencing in gradient strategies and has proven to be valuable in experiments reported in [13]. The second modification consists in the introduction of a second step size parameter. One parameter is used to control the length of the trial steps, the other one that of the search steps. It becomes thus possible to imitate the result of the genetic repair effect observed in evolution strategies, and to reap the resulting benefits in the presence of noise. Interestingly, it will be seen below that the use of a second step length is closely akin to the idea of using rescaled mutations in evolution strategies as proposed by Ostermeier in [1] and analyzed by Beyer [14], [15].

B. Step Length Control and Non-Isotropic Mutations

Of course, the basic iterative steps discussed thus far are but one part of a numerical optimization strategy. A further important issue to be addressed is that of step length control. Gradient algorithms, EGS, and evolution strategies each have their own mechanisms for that purpose. When using the method of steepest descent, a locally optimal step length is usually determined by conducting a one-dimensional line search along the direction of the (negative of the) gradient. In the presence of noise, experiments reported in [16] have shown that rapidly decreasing step lengths and convergence to a non-optimal point may result on functions as simple as the sphere model. The original EGS strategy described in [6] tries out two steps, one slightly longer than the previous search step, the other one slightly shorter, and settles for the better of the two. The situation is complicated by the introduction of a second step length as proposed in [13] and discussed above. One possibility inspired by the effects of genetic repair in evolution strategies as well as by the idea of using rescaled mutations is to choose the length of the search steps to be a constant fraction of the length of the trial steps, thus leaving only the former to be controlled. In the realm of evolution strategies, step length adaptation mechanisms including the 1/5th success rule [17], mutative self-adaptation [17], [18], and cumulative step length adaptation [19] are commonly used. In the experiments described in [16], cumulative step length adaptation has proven relatively robust with respect to the effects of noise. Analytical results regarding the properties of cumulative step length adaptation on the sphere model can be found in [20], [21].

Finally, it is well known that while using isotropic mutations (i.e., using a mutation covariance matrix that is a scalar multiple of the unity matrix) is appropriate for the sphere model, the performance of isotropic evolution strategies rapidly declines with increasing condition number of the Hessian matrix of the objective function. As discussed by Whitley et al. [22], in order to achieve good performance on functions that contain long valleys or ridges that need to be followed, it is necessary to adapt the shape of the mutation distribution to the local characteristics of the problem at hand. Interestingly, even though EGS uses a weighted sum of all its trial steps to determine a search step where the weights can vary widely, the same can be observed to hold true for EGS. Salomon [23] proposes to use individual step sizes for the N dimensions, and thus to operate with a mutation covariance matrix in which all off-diagonal entries are zero. Much improved performance on ellipsoidal objective functions the axes of which are aligned with the axes of the coordinate system can be achieved. However, the resulting algorithm's performance is not invariant with respect to the orientation of the coordinate system. Progress along valleys or ridges that are not parallel to one of the main axes is still very slow. Salomon [6] also suggests using a momentum term in order to improve the ridge following capabilities of EGS. However, as pointed out by Poland and Zell [24], while the use of a momentum term is useful for ridge functions, it is of limited use if the eigenvalues of the Hessian of the objective function are widely spread and not dominated by a single value that is much smaller than all of the others. For multirecombination evolution strategies, Hansen and Ostermeier [7] have introduced a covariance adaptation mechanism that has been found in experiments to reliably transform any convex quadratic function into the simple sphere function. In the present paper, that algorithm is adapted for use in EGS.

C. Outlook

The remainder of this paper is organized as follows. In Section II, the EGS procedure with the modifications described in [13] is outlined and the sphere model is introduced as an important model for studying local search properties of optimization algorithms. In Section III, the performance of the modified EGS strategy with isotropic mutations is investigated on the sphere model. A number of simplifications are made in the calculations that hold in the limit $N \to \infty$, but that are seen to provide good approximations for moderately large values of N. Section IV generalizes the results from Section III by considering the effects of noise present in the observed objective function values. In Section V the performance of the modified EGS strategy is compared with that of the original EGS procedure as studied in [10] as well as with several variants of evolution strategies. Both the cases of fitness-proportionate and uniform noise strength are considered. In Section VI, the covariance matrix adaptation mechanism is formulated for EGS and tested on a number of objective functions. It is seen that while covariance matrix adaptation largely works for EGS as well as it does for evolution strategies, there are some problems with step length adaptation in low-dimensional search spaces. It is then seen that in the presence of noise, CMA-EGS is preferable to the $(\mu/\mu, \lambda)$-CMA-ES due to the additional flexibility afforded by its use of a rescaling parameter. Section VII concludes with a brief summary and suggestions for future research.

II. Preliminaries

This section outlines the modified EGS procedure for the minimization of functions $f : I\!R^N \to I\!R$ as described in [13] and motivates the changes that have been made to the original algorithm proposed in [6]. It then briefly summarizes some notational conventions and basic results for the sphere model.

A. Modified Evolutionary Gradient Search

An iteration of the modified EGS procedure with isotropic mutations updates the search point $\mathbf{x} \in I\!R^N$ of the strategy in three steps:

1) Generate 2λ trial points $\mathbf{y}_{\pm}^{(i)} = \mathbf{x} \pm \sigma\mathbf{z}^{(i)}$, $i = 1, \ldots, \lambda$, where mutation strength $\sigma > 0$ determines the step length and the $\mathbf{z}^{(i)}$ are vectors consisting of N independent, standard normally distributed components.
2) Determine the objective function values $f(\mathbf{y}_{\pm}^{(i)})$ of the trial points and compute the weighted sum
$$\mathbf{z}^{(\text{avg})} = \sum_{i=1}^{\lambda} \left(f(\mathbf{y}_{-}^{(i)}) - f(\mathbf{y}_{+}^{(i)}) \right) \mathbf{z}^{(i)} \tag{1}$$
of the $\mathbf{z}^{(i)}$ vectors.
3) Replace the search point $\mathbf{x}$ by $\mathbf{x} + \sigma\mathbf{z}^{(\text{prog})}$, where
$$\mathbf{z}^{(\text{prog})} = \frac{\sqrt{N}}{\kappa} \frac{\mathbf{z}^{(\text{avg})}}{\|\mathbf{z}^{(\text{avg})}\|} \tag{2}$$
is referred to as the EGS progress vector.

The use of non-isotropic mutations as well as the discussion of step length adaptation is deferred until Section VI. Notice that an iteration of the strategy thus described differs from one of the original EGS procedure in two important aspects. First, the modified EGS strategy generates *two* trial solutions for every mutation vector: one in direction of that vector, and a second one by taking a step of the same length in the opposite direction. The difference of the function values of the two points generated is used to weight the respective mutation vector in Eq. (1). The difference between this procedure and that of the original EGS algorithm is reminiscent of the difference between central differencing and forward differencing when using gradient strategies. It will be seen in Section III that for EGS, the use of central differencing has the desirable effect of eliminating a performance hampering bias term in the computation of $\mathbf{z}^{(\text{avg})}$. Second, the introduction of the factor $1/\kappa$ in Eq. (2) allows for making search steps that differ in length from the trial steps. While the length of the trial steps is roughly $\sigma\sqrt{N}$ for large N, that of the search steps is $\sigma\sqrt{N}/\kappa$. The modification is akin to the use of rescaled mutations in evolution strategies as suggested by Ostermeier and Rechenberg [1]. According to Beyer [14], [15],

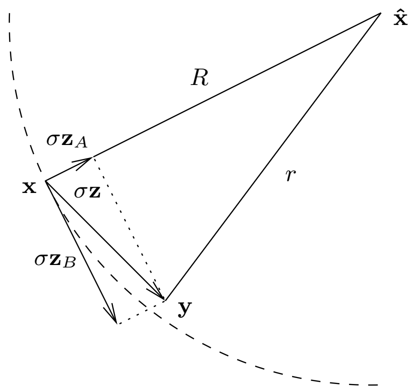

Fig. 1. *Decomposition of a vector $\mathbf{z}$ into central component $\mathbf{z}_A$ and lateral component $\mathbf{z}_B$. Vector $\mathbf{z}_A$ is parallel to $\hat{\mathbf{x}} - \mathbf{x}$, vector $\mathbf{z}_B$ is in the hyperplane perpendicular to that. The starting and end points, $\mathbf{x}$ and $\mathbf{y} = \mathbf{x} + \sigma\mathbf{z}$, of vector $\sigma\mathbf{z}$ are at distances R and r from the optimizer $\hat{\mathbf{x}}$, respectively.*

for evolution strategies the use of rescaled mutations can lead to greatly improved performance in the presence of noise. The investigations below confirm that the same holds true for EGS.

B. The Sphere Model

Since the early work of Rechenberg [17], the local performance of evolution strategies has commonly been studied on the quadratic sphere

$$f(\mathbf{x}) = (\hat{\mathbf{x}} - \mathbf{x})^{\mathrm{T}}(\hat{\mathbf{x}} - \mathbf{x}), \qquad \mathbf{x} \in I\!R^N,$$

where the task is minimization and where $\hat{\mathbf{x}} \in I\!R^N$ is the optimizer. The sphere serves as a model for objective functions in the vicinity of well behaved local optima. See [16] for a justification of the usefulness of such considerations. Using appropriate normalizations, most results arrived at for the quadratic sphere can be extended to general spherically symmetric functions $f(\mathbf{x}) = g(R)$, where $R = \|\hat{\mathbf{x}} - \mathbf{x}\|$ and where $g(R)$ is strictly monotonic.

In order to quantify the local performance of point-based search strategies, consider the effect of adding a vector $\sigma\mathbf{z}$ to the current search point $\mathbf{x}$. The EGS strategy does so both when generating trial points and when replacing the current search point at the end of an iteration. Denoting the respective distances of $\mathbf{x}$ and $\mathbf{y} = \mathbf{x} + \sigma\mathbf{z}$ from the optimizer by R and r, the difference $\delta(\mathbf{z}) = R^2 - r^2$ between objective function values $f(\mathbf{x}) = R^2$ and $f(\mathbf{y}) = r^2$ can be used to determine both the quality of trial points and the rate at which the strategy approaches the optimum. Keeping with the terminology established in connection with evolution strategies, $\delta(\mathbf{z})$ is referred to as the fitness advantage associated with vector $\mathbf{z}$.[1]

The commonly used approach to determining $\delta(\mathbf{z})$ on the sphere model relies on a decomposition of vector $\mathbf{z}$ that is illustrated in Fig. 1. A vector $\mathbf{z}$ originating at search 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 $(N-1)$-dimensional hyperplane perpendicular to that. The vectors $\mathbf{z}_A$ and $\mathbf{z}_B$ are referred to as the central and lateral components of vector $\mathbf{z}$, respectively. The signed length z_A of the central component of vector $\mathbf{z}$ is defined to equal $\|\mathbf{z}_A\|$ if $\mathbf{z}_A$ points towards the optimizer and to equal $-\|\mathbf{z}_A\|$ if it points away from it. Using elementary geometry, it can easily be seen from the figure that

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

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

$$\begin{aligned} \delta(\mathbf{z}) &= R^2 - r^2 \\ &= 2R\sigma z_A - \sigma^2 \|\mathbf{z}\|^2. \end{aligned} \tag{3}$$

Writing this in terms of normalized quantities

$$\sigma^* = \sigma \frac{N}{R} \quad \text{and} \quad \delta^* = \delta \frac{N}{2R^2}, \tag{4}$$

it follows

$$\delta^*(\mathbf{z}) = \sigma^* z_A - \frac{{\sigma^*}^2}{2N} \|\mathbf{z}\|^2. \tag{5}$$

for the normalized fitness advantage associated with vector $\mathbf{z}$. Equations (3) and (5) are used in two contexts. First, by considering mutation vectors they allow computing the fitness differences that occur in the definition of $\mathbf{z}^{(\text{avg})}$ in Eq. (1). Second, by considering progress vectors they help quantify the speed with which the strategy approaches the optimum.

In case $\mathbf{z}$ is a mutation vector, the signed length z_A of the central component of $\mathbf{z}$ is standard normally distributed as the distribution of $\mathbf{z}$ is isotropic. Moreover, for large N, the influence of the central component on the overall squared length of $\mathbf{z}$ becomes negligible, and thus z_A and $\|\mathbf{z}\|^2$ are asymptotically independent. Since $\|\mathbf{z}\|^2$ is the sum of squares of N independent, standard normally distributed random variables, it has expectation N and variance $2N$. For increasing N, $\|\mathbf{z}\|^2/N$ is thus increasingly well approximated by unity, and, using Eq. (5), it follows by considering both $\mathbf{z}$ and $-\mathbf{z}$ that

$$\delta^*(\pm\mathbf{z}) \overset{N\to\infty}{=} \pm\sigma^* z_A - \frac{{\sigma^*}^2}{2}. \tag{6}$$

Letting $r_\pm^{(i)}$ denote the distances from the optimizer to $\mathbf{y}_\pm^{(i)}$, respectively, and letting

$$\delta_c(\mathbf{z}) = f(\mathbf{y}_-^{(i)}) - f(\mathbf{y}_+^{(i)}), \tag{7}$$

it follows immediately that

$$\begin{aligned} \delta_c(\mathbf{z}) &= {r_-^{(i)}}^2 - {r_+^{(i)}}^2 \\ &= \left(R^2 - {r_+^{(i)}}^2\right) - \left(R^2 - {r_-^{(i)}}^2\right) \\ &= \delta(\mathbf{z}^{(i)}) - \delta(-\mathbf{z}^{(i)}) \end{aligned}$$

and therefore, using Eqs. (4) and (6), that

$$\frac{N}{2R^2} \delta_c(\mathbf{z}) \overset{N\to\infty}{=} 2\sigma^* z_A^{(i)}. \tag{8}$$

This result will be used below to determine the weights of the mutation vectors in the summation in Eq. (1).

[1]While the notation adopted here is deliberately brief and does not reflect that explicitly, it is important to keep in mind that the fitness advantage $\delta(\mathbf{z})$ depends not only on vector $\mathbf{z}$, but also on the mutation strength σ and, in case $\mathbf{z}$ is a progress vector, on the rescaling factor κ.

In case $\mathbf{z} = \mathbf{z}^{(\text{prog})}$ is a progress vector, the expected normalized associated fitness advantage

$$\Delta^*_{\text{EGS}} = \mathrm{E}\left[\delta^*(\mathbf{z}^{(\text{prog})})\right] \tag{9}$$

quantifies the EGS quality gain, a commonly used performance measure as motivated in [3]. In order to determine the quality gain of EGS on the sphere model using Eqs. (5) and (9), expected values of the signed length of the central component and of the overall squared length of the EGS progress vector $\mathbf{z}^{(\text{prog})}$ need to be computed. The computation of the quality gain in the absence of noise is subject of Section III; the noisy case is considered in Section IV.

III. Performance of the Modified EGS Strategy on the Sphere Model

Due to the definition of the EGS progress vector in Eq. (2) it is clear that its squared length is N/κ^2, and therefore that

$$\Delta^*_{\text{EGS}} = \sigma^*\mathrm{E}\left[z_A^{(\text{prog})}\right] - \frac{\sigma^{*2}}{2\kappa^2}. \tag{10}$$

In order to compute $\mathrm{E}[z_A^{(\text{prog})}]$, notice that $\mathbf{z}^{(\text{avg})}$ enters the definition of $\mathbf{z}^{(\text{prog})}$ normalized to unit length, and that thus any vector parallel to $\mathbf{z}^{(\text{avg})}$ can be substituted instead provided that it is normalized as well. In particular, writing R for the distance from the search point $\mathbf{x}$ to the optimizer and defining

$$\mathbf{z}^* = \frac{N}{2R^2}\mathbf{z}^{(\text{avg})} \tag{11}$$

it follows from Eq. (2) that

$$z_A^{(\text{prog})} = \frac{\sqrt{N}}{\kappa}\frac{z_A^*}{\|\mathbf{z}^*\|}. \tag{12}$$

According to Eqs. (1), (4), (7), and (8),

$$\begin{aligned}\mathbf{z}^* &= \frac{N}{2R^2}\sum_{i=1}^{\lambda}\delta_c(\mathbf{z}^{(i)})\mathbf{z}^{(i)}\\ &\overset{N\to\infty}{=} 2\sigma^*\sum_{i=1}^{\lambda} z_A^{(i)}\mathbf{z}^{(i)}.\end{aligned}$$

Considering the signed length of the central component yields

$$z_A^* \overset{N\to\infty}{=} 2\sigma^*\sum_{i=1}^{\lambda} {z_A^{(i)}}^2. \tag{13}$$

For the overall squared length of $\mathbf{z}^{(\text{avg})}$ it follows from Eq. (1) using Eq. (7) that

$$\begin{aligned}\frac{\|\mathbf{z}^{(\text{avg})}\|^2}{N} &= \frac{1}{N}\sum_{k=1}^{N}\left(\sum_{i=1}^{\lambda}\delta_c(\mathbf{z}^{(i)})z_k^{(i)}\right)^2\\ &= \frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{\lambda}\left(\delta_c(\mathbf{z}^{(i)})z_k^{(i)}\right)^2 + \zeta,\end{aligned}$$

where of course the components of $\mathbf{z}^{(i)} = (z_1^{(i)}, \ldots, z_N^{(i)})^{\mathrm{T}}$ are independently standard normally distributed and where

$$\zeta = \frac{1}{N}\sum_{k=1}^{N}\sum_{i\neq j}\delta_c(\mathbf{z}^{(i)})\delta_c(\mathbf{z}^{(j)})z_k^{(i)}z_k^{(j)}$$

is a crosstalk term. That term has mean zero and a variance that can be shown to tend to zero as $N \to \infty$ and thus vanishes in the limit of infinite search space dimensionality. Furthermore making use of the fact $\|\mathbf{z}^{(i)}\|^2/N \to 1$ as $N \to \infty$, it follows from the definition of $\mathbf{z}^*$ using Eq. (8) that

$$\begin{aligned}\frac{\|\mathbf{z}^*\|^2}{N} &\overset{N\to\infty}{=} \left(\frac{N}{2R^2}\right)^2\frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{\lambda}\left(\delta_c(\mathbf{z}^{(i)})z_k^{(i)}\right)^2\\ &= \left(\frac{N}{2R^2}\right)^2\sum_{i=1}^{\lambda}\left(\delta_c(\mathbf{z}^{(i)})\right)^2\underbrace{\frac{1}{N}\sum_{k=1}^{N}{z_k^{(i)}}^2}_{\to 1}\\ &\overset{N\to\infty}{=} 4{\sigma^*}^2\sum_{i=1}^{\lambda}{z_A^{(i)}}^2.\end{aligned} \tag{14}$$

Combining the results from Eqs. (13) and (14), it follows using Eq. (12) that

$$z_A^{(\text{prog})} \overset{N\to\infty}{=} \frac{1}{\kappa}\sqrt{\sum_{i=1}^{\lambda}{z_A^{(i)}}^2}. \tag{15}$$

Recall that the $z_A^{(i)}$ are simply standard normally distributed random variables, and that the radicand thus is χ^2-distributed with λ degrees of freedom. Also notice that in contrast to the corresponding result derived in [10] for the original EGS procedure, but similar to evolution strategies, the signed length of the central component of the progress vector of the modified EGS algorithm (and indeed the progress vector itself) is independent of the length of the trial steps. The dependence of that quantity on the mutation strength in the original EGS procedure was a consequence of a negative bias of that strategy that results from the fact that the fitness advantage associated with mutation vectors is more often negative than it is positive. This is reflected in Eq. (5) by the negative term quadratic in $\|\mathbf{z}\|$. The modified EGS procedure eliminates that bias by evaluating mutation $-\mathbf{z}^{(i)}$ along with $\mathbf{z}^{(i)}$, and computing the difference between the corresponding objective function values in order to weight vector $\mathbf{z}^{(i)}$ in Eq. (1). It will be seen below that removing that bias has a beneficial effect on the performance of the strategy.

Writing E_λ for $\mathrm{E}[\sqrt{\chi^2_\lambda}]$ it follows from Eqs. (10) and (15) that the quality gain of the modified evolutionary gradient search procedure on the sphere model in the limit $N \to \infty$ is

$$\Delta^*_{\text{EGS}} \overset{N\to\infty}{=} \frac{1}{\kappa}\left[\sigma^* E_\lambda - \frac{{\sigma^*}^2}{2\kappa}\right]. \tag{16}$$

Due to the properties of the χ^2-distribution, E_λ is well approximated by $\sqrt{\lambda}$ (the error is below 5% for $\lambda \geq 5$ and below 1% for $\lambda \geq 25$). Substituting $\sqrt{\lambda}$ for E_λ yields approximation

$$\Delta^*_{\text{EGS}} \approx \frac{1}{\kappa}\left[\sigma^*\sqrt{\lambda} - \frac{{\sigma^*}^2}{2\kappa}\right] \tag{17}$$

for the quality gain of the modified EGS procedure on the high-dimensional sphere model. Figure 2 compares quality gain predictions made using Eqs. (16) and (17) with measurements from EGS runs on the sphere model. It can be seen that

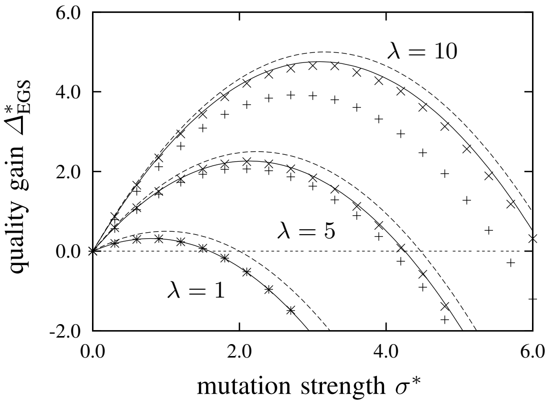

Fig. 2. EGS quality gain Δ^*_{EGS} plotted against the normalized mutation strength σ^* for $\lambda \in \{1, 5, 10\}$ and $\kappa = 1$. The points represent measurements from EGS runs on spheres with $N = 40$ (+) and $N = 400$ ($\times$). The solid lines represent asymptotically exact results from Eq. (16), the dashed lines those from the approximation Eq. (17).

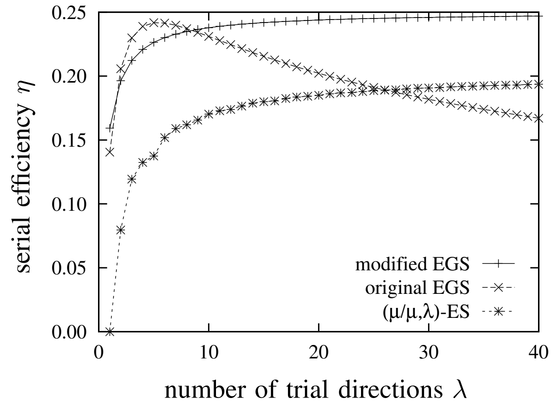

Fig. 3. Serial efficiencies η of strategies on the sphere model in the limit $N \to \infty$ plotted against the parameter λ that determines the number of trial points generated per time step. The data points for the $(\mu/\mu, \lambda)$-ES with optimally chosen μ and for the original EGS procedure have been computed as described in [3] and [10], respectively. The curve for the $(\mu/\mu, \lambda)$-ES is not smooth as μ can take on integer values only. The data points for the modified EGS procedure have been obtained from Eq. (16) with $\kappa = 1$.

while the analytical results were derived under the assumption of infinite search space dimensionality, the agreement with the measurements is good for N as small as 40 if λ is not too large. For $\lambda = 10$ the discrepancy between predictions and measurements is considerable for $N = 40$, but tolerable for most purposes for $N = 400$.

The result from Eq. (17) is to be compared with the corresponding approximation

$$\Delta^*_{\text{EGS}} \approx \sigma^* \sqrt{\frac{\lambda}{1 + \sigma^{*2}/4}} - \frac{\sigma^{*2}}{2} \tag{18}$$

for the original EGS procedure derived in [10]. Disregarding the parameter κ (that can without loss of performance be set to 1 in the noise-free case), the negative term that is quadratic in σ^* is unaffected. However, the switch from forward differencing to central differencing in the weighting of mutation vectors has led to the disappearance of the term quadratic in the mutation strength in the denominator of the positive term that hampered progress towards the optimum and limited the size of useful mutation strengths in the original EGS procedure. For any nonzero mutation strength, the term that contributes positively to the quality gain is larger for the modified EGS procedure than it is for the original algorithm.

Computing the derivative of Eq. (16) with respect to σ^* and finding the positive zero yields optimal mutation strength

$$\sigma^* = \kappa E_\lambda. \tag{19}$$

Reinserting this result into Eq. (16) results in maximum quality gain

$$\Delta^*_{\text{EGS}} = \frac{E_\lambda^2}{2}. \tag{20}$$

With increasing λ, the optimal quality gain of the modified EGS procedure thus approaches $\lambda/2$, independent of κ. In contrast to the original EGS procedure, but like the $(\mu/\mu, \lambda)$-ES, the modified EGS strategy thus exhibits a linear speed-up when increasing the number of trial directions generated per time step.

In order to have a fair comparison that accounts for the differing computational costs of the various strategies, it is common to define the serial efficiency η of a strategy as as the quality gain per evaluation of the objective function for optimally set mutation strength. Clearly, the number of objective function evaluations per time step is λ for the $(\mu/\mu, \lambda)$-ES, it is $\lambda + 1$ for the original EGS procedure, and it is 2λ for the modified EGS strategy. Figure 3 contrasts the respective serial efficiencies of the various strategies. While the serial efficiency of the $(\mu/\mu, \lambda)$-ES approaches a value of 0.202 as λ increases and that of the original EGS procedure decreases once λ exceeds five, that of the modified EGS strategy asymptotically approaches a value of 0.25. The modified EGS procedure is superior to the $(\mu/\mu, \lambda)$-ES for any value of λ. Notice however that for small values of λ the original EGS procedure is somewhat more efficient than the modified one as it requires fewer objective function evaluations per time step.

IV. Performance in the Presence of Noise

The assumption that the objective function value of a candidate solution can be determined exactly is usually an idealization. Real-world optimization problems often suffer from noise that can stem from sources as different as — and not restricted to — measurement limitations, the use of Monte Carlo methods, and human computer interaction. Understanding how noise impacts the performance of optimization strategies is important for choosing appropriate strategy variants, for the sizing of strategy parameters, and for the design of new, more noise resistant algorithms.

The most commonly employed noise model assumes additive Gaussian noise. That is, it is assumed that when evaluating the objective function at some search space location $\mathbf{y}$, the value obtained is normally distributed with mean $f(\mathbf{y})$ and with some standard deviation $\sigma_\epsilon(\mathbf{y})$. In [10] it had been seen that the original EGS procedure is generally inferior to the $(\mu/\mu, \lambda)$-ES on the noisy sphere model. The $(\mu/\mu, \lambda)$-ES is able to make use of what has been termed the *genetic repair effect* by Beyer [3], [11]: the averaging of mutation vectors that is implicit in the multirecombination procedure together with the selection mechanism of the strategy leads to the lateral component of the $(\mu/\mu, \lambda)$-progress vector being shorter (by a factor of $\sqrt{\mu}$) than that of the mutation vectors. The length of the central component is not affected significantly. As a consequence, the $(\mu/\mu, \lambda)$-ES is able to use mutation strengths much larger than those useful for evolution strategies that do not employ multirecombination. As discussed in [12], in the presence of noise larger mutation strengths have the advantage of reducing the noise-to-signal ratio that the strategy operates under. In the limit of infinite search space dimensionality, increasing the population size parameters μ and λ makes it possible to increase the mutation strength to a degree that all but eliminates the noise. In finite-dimensional search spaces the potential benefits of genetic repair are more limited, but as seen in [25] still considerable.

It was seen in [10] that the original EGS procedure is not able to benefit from genetic repair to the degree that the $(\mu/\mu, \lambda)$-ES does for two reasons. First, the negative bias that leads to the occurrence of the mutation strength in the denominator in Eq. (18) puts a limit on useful mutation strengths. This has been addressed in Section III where it was seen that the use of the central differencing-like weighting of mutation vectors in Eq. (1) leads to the elimination of the bias and the disappearance of the performance hampering term from the quality gain law. Second, the ability of the original EGS procedure to use large mutation strengths is severely limited by the fact that it uses progress vectors that are of the same length as its mutation vectors are. It is for this reason that the modified strategy introduced above uses the rescaling factor κ in Eq. (2). That factor decouples the length of the progress vector from that of the mutation vectors, making it possible to emulate the effect that the $(\mu/\mu, \lambda)$-ES benefits from due to its use of multirecombination. The same idea has been suggested for the $(1, \lambda)$-ES by Ostermeier and Rechenberg [1] and analyzed by Beyer [14], [15]. The remainder of this section studies the potential benefits that result from the use of the rescaling factor in evolutionary gradient search.

As the weighting factors depend on objective function measurements, the presence of noise affects the weighting of the mutation vectors in Eq. (1). Assuming that the evaluation of all trial points in a time step is subject to the same noise strength σ_ϵ and introducing normalized noise strength

$$\sigma_\epsilon^* = \sigma_\epsilon \frac{N}{2R^2}, \tag{21}$$

Eq. (8) is to be replaced by

$$\frac{N}{2R^2}\delta_c(\mathbf{z}^{(i)}) \overset{N\to\infty}{=} 2\sigma^* z_A^{(i)} + \sqrt{2}\sigma_\epsilon^* z_\epsilon^{(i)} \tag{22}$$

in order to reflect the *measured* fitness difference, where $z_\epsilon^{(i)}$ is a standard normally distributed random variable that reflects the effects of noise. The factor $\sqrt{2}$ in the second term on the right hand side is a result of the fact that the modified EGS procedure evaluates two candidate solutions rather than a single one when weighting a mutation vector, and that the variance of the sum of the two noise terms equals the sum of the two variances.

Using Eq. (22) in the definition of vector $\mathbf{z}^*$ according to Eq. (11), it follows that in the presence of noise

$$\mathbf{z}^* \overset{N\to\infty}{=} \sum_{i=1}^{\lambda} \left(2\sigma^* z_A^{(i)} + \sqrt{2}\sigma_\epsilon^* z_\epsilon^{(i)}\right) \mathbf{z}^{(i)}.$$

Considering the signed length of the central component yields

$$z_A^* \overset{N\to\infty}{=} 2\sigma^* \sum_{i=1}^{\lambda} {z_A^{(i)}}^2 + \sqrt{2}\sigma_\epsilon^* \sum_{i=1}^{\lambda} z_\epsilon^{(i)} z_A^{(i)}. \tag{23}$$

As the $z_A^{(i)}$ and the $z_\epsilon^{(i)}$ are independently drawn from a standardized normal distribution, the expected values of the terms on the right hand side are $2\lambda\sigma^*$ and zero, respectively. For the overall squared length of $\mathbf{z}^*$ in the presence of noise it follows in close analogy to Eq. (13), but using Eq. (22) instead of Eq. (8), that

$$\begin{aligned}\frac{\|\mathbf{z}^*\|^2}{N} &\overset{N\to\infty}{=} \left(\frac{N}{2R^2}\right)^2 \sum_{i=1}^{\lambda} \left(\delta_c(\mathbf{z}^{(i)})\right)^2 \\ &\overset{N\to\infty}{=} \sum_{i=1}^{\lambda} \left(2\sigma^* z_A^{(i)} + \sqrt{2}\sigma_\epsilon^* z_\epsilon^{(i)}\right)^2 \\ &= 4{\sigma^*}^2 \sum_{i=1}^{\lambda} {z_A^{(i)}}^2 + 4\sqrt{2}\sigma^*\sigma_\epsilon^* \sum_{i=1}^{\lambda} z_A^{(i)} z_\epsilon^{(i)} \\ &\quad + 2{\sigma_\epsilon^*}^2 \sum_{i=1}^{\lambda} {z_\epsilon^{(i)}}^2. \end{aligned} \tag{24}$$

The expected values of the terms on the right hand side are $4\lambda{\sigma^*}^2$, zero, and $2\lambda{\sigma_\epsilon^*}^2$, respectively.

In order to compute the quality gain of the modified EGS procedure on the noisy sphere using Eqs. (10) and (12), $\mathrm{E}[z_A^*/\|\mathbf{z}^*\|]$ needs to be computed. Unfortunately, according to Eqs. (23) and (24), the quantities involved have distributions that make it impossible to compute the expectation in closed form. Instead, as done in [10] for the original EGS strategy, we make two simplifications. First, we ignore the terms in Eqs. (23) and (24) that have zero mean. The error introduced by this simplification is most noticeable for small values of λ as for larger numbers of search directions, the other terms in Eqs. (23) and (24) solidly dominate those with mean zero. And second, we assume that $\sum_{i=1}^{\lambda} z_A^{(i)} = \sum_{i=1}^{\lambda} z_\epsilon^{(i)}$. Notice that the two terms are identically distributed but independent. For increasing λ, due to the properties of the χ^2-distribution, the sums are dominated by their expectations, and the error introduced by the assumption of them being equal becomes less severe. Experiments show that for λ as small as ten, the resulting error is insignificant for most practical purposes. From the simplifications, it follows

$$\begin{aligned}\mathrm{E}\left[z_A^{(\mathrm{prog})}\right] &\approx \frac{1}{\kappa}\mathrm{E}\left[\frac{2\sigma^* \sum_{i=1}^{\lambda} {z_A^{(i)}}^2}{\sqrt{\left(4{\sigma^*}^2 + 2{\sigma_\epsilon^*}^2\right)\sum_{i=1}^{\lambda} {z_A^{(i)}}^2}}\right] \\ &= \frac{1}{\kappa}\frac{E_\lambda}{\sqrt{1 + {\sigma_\epsilon^*}^2/(2{\sigma^*}^2)}}\end{aligned}$$

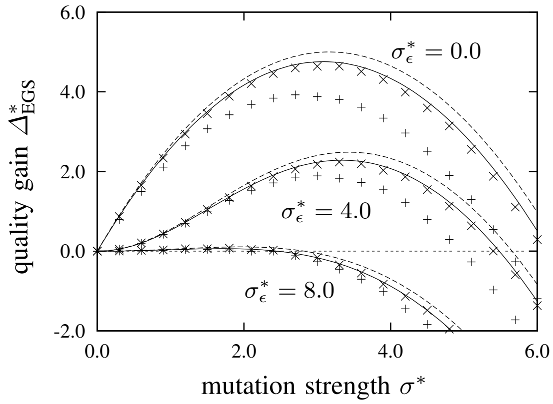

Fig. 4. Quality gain Δ^*_{EGS} plotted against normalized mutation strength σ^* for $\lambda = 10$, $\kappa = 1$, and normalized noise strengths $\sigma^*_\epsilon \in \{0.0, 4.0, 8.0\}$. The points represent measurements from EGS runs on spheres with $N = 40$ (+) and $N = 400$ (×). The solid and dashed lines represent the approximations Eqs. (25) and (26), respectively.

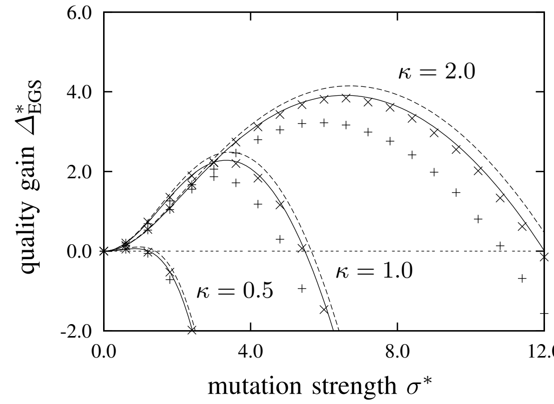

Fig. 5. Quality gain Δ^*_{EGS} plotted against normalized mutation strength σ^* for $\lambda = 10$, $\sigma^*_\epsilon = 4.0$, and $\kappa \in \{0.5, 1.0, 2.0\}$. The points represent measurements from EGS runs on spheres with $N = 40$ (+) and $N = 400$ (×). The solid and dashed lines represent the approximations Eqs. (25) and (26), respectively.

as an approximation for the expected signed length of the central component of the EGS progress vector. Using this approximation in Eq. (10) and writing $\vartheta = \sigma^*_\epsilon/\sigma^*$ for the noise-to-signal ratio that the strategy operates under yields approximations

$$\Delta^*_{\mathrm{EGS}} \approx \frac{1}{\kappa}\left[\sigma^* \frac{E_\lambda}{\sqrt{1+\vartheta^2/2}} - \frac{{\sigma^*}^2}{2\kappa}\right] \qquad (25)$$

$$\approx \frac{1}{\kappa}\left[\sigma^* \sqrt{\frac{\lambda}{1+\vartheta^2/2}} - \frac{{\sigma^*}^2}{2\kappa}\right] \qquad (26)$$

for the quality gain of the modified EGS strategy on the noisy sphere model. Figure 4 compares predictions made using Eqs. (25) and (26) with measurements from EGS runs with $\kappa = 1$. It can be seen that the quality of the approximation is good provided that N is not too small. The absolute error is smaller in the presence of noise than in its absence, thus providing a justification for the simplifications made in the derivation of the quality gain laws. Figure 5 illustrates the quality of approximations Eqs. (25) and (26) for varying values of κ. It can be seen that the accuracy of the predictions is best for small values of κ. Larger values of κ as well as of λ generally require higher search space dimensionalities in order for the approximations to be good. The same holds true for the corresponding quality gain approximation for the $(1,\lambda)$-ES with rescaled mutations derived by Beyer [14].

V. Evolution Strategies vs. Evolutionary Gradient Search

Using the results thus derived, the performance of the modified EGS procedure can now be compared with that of the original EGS strategy, the $(\mu/\mu,\lambda)$-ES, and the $(1,\lambda)$-ES with rescaled mutations. Comparisons are presented first by considering single time steps only, and then for the cases of fitness-proportionate and uniform noise strengths.

Table I compares the quality gain approximation Eq. (26) for the modified EGS procedure with those for the original EGS strategy derived in [10], the $(1,\lambda)$-ES with rescaled mutations analyzed in [14], and the $(\mu/\mu,\lambda)$-ES investigated in [12]. Common to all of the quality gain laws is the appearance of two terms: a term that contributes positively and that is due to the central components of the progress vectors, and a negative term that is due to the progress vectors' lateral components. The dependence of those terms on strategy and problem parameters differs from strategy to strategy. However, introducing standardized quantities

$$\bar{\sigma} = \frac{\sigma^*}{\hat{\sigma}^*}, \qquad \bar{\sigma}_\epsilon = \frac{\sigma^*_\epsilon}{\hat{\sigma}^*_\epsilon}, \qquad \text{and} \qquad \bar{\Delta} = \frac{\Delta^*}{\hat{\Delta}^*},$$

where the scaling parameters $\hat{\sigma}^*$, $\hat{\sigma}^*_\epsilon$, and $\hat{\Delta}^*$ for the various strategies are given in Table II, the quality gain laws of the $(1,\lambda)$-ES with rescaled mutations, the $(\mu/\mu,\lambda)$-ES, and the modified EGS procedure can all be written as

$$\bar{\Delta} = \frac{2\bar{\sigma}}{\sqrt{1+(\bar{\sigma}_\epsilon/\bar{\sigma})^2}} - \bar{\sigma}^2. \qquad (27)$$

The original EGS strategy does not fit into this pattern due to its negative bias discussed in Section III. For the other strategies, the dependence of optimal standardized mutation strength and corresponding standardized quality gain on the standardized noise strength is illustrated in Fig. 6. Clearly, $\hat{\sigma}^*$ is the optimal normalized noise strength in the absence of noise, and $\hat{\Delta}^*$ is the corresponding quality gain. The scaling parameter for the noise strength determines the normalized noise strength $2\hat{\sigma}^*_\epsilon$ up to which positive quality gain is possible. Altogether, the following conclusions can be drawn from Tables I and II and Fig. 6:

- The effect of the rescaling achieved by the inclusion of the factor κ in the definition of the strategy takes the same functional form for the modified EGS procedure as it does for the $(1,\lambda)$-ES with rescaled mutations. For the $(\mu/\mu,\lambda)$-ES, a similar effect is achieved implicitly by virtue of multirecombination as witnessed by the appearance of the factor μ in the denominator of the negative term in the respective quality gain law.
- As the strategies differ in how they combine the information obtained by evaluating their offspring, so do

TABLE I
QUALITY GAIN LAWS ON THE NOISY, INFINITE-DIMENSIONAL SPHERE.

$(1, \lambda)$-ES with rescaled mutations	$\frac{1}{\kappa}\left[\sigma^* \frac{c_{1,\lambda}}{\sqrt{1+\vartheta^2}} - \frac{\sigma^{*2}}{2\kappa}\right]$
$(\mu/\mu, \lambda)$-ES	$\sigma^* \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}} - \frac{\sigma^{*2}}{2\mu}$
original EGS	$\sigma^* \sqrt{\frac{\lambda}{1+\sigma^{*2}/4+\vartheta^2}} - \frac{\sigma^{*2}}{2}$
modified EGS	$\frac{1}{\kappa}\left[\sigma^* \sqrt{\frac{\lambda}{1+\vartheta^2/2}} - \frac{\sigma^{*2}}{2\kappa}\right]$

TABLE II
SCALING PARAMETERS $\hat{\sigma}^*$, $\hat{\sigma}^*_\epsilon$, AND $\hat{\Delta}^*$.

	$\hat{\sigma}^*$	$\hat{\sigma}^*_\epsilon$	$\hat{\Delta}^*$
$(1, \lambda)$-ES with rescaled mutations	$\kappa c_{1,\lambda}$	$\kappa c_{1,\lambda}$	$c^2_{1,\lambda}/2$
$(\mu/\mu, \lambda)$-ES	$\mu c_{\mu/\mu,\lambda}$	$\mu c_{\mu/\mu,\lambda}$	$\mu c^2_{\mu/\mu,\lambda}/2$
modified EGS	$\kappa\sqrt{\lambda}$	$\kappa\sqrt{2\lambda}$	$\lambda/2$

the positive terms in the respective quality gain laws. The effects of comma-selection (and, for the $(\mu/\mu, \lambda)$-ES, multirecombination) are captured by the progress coefficients $c_{1,\lambda}$ and $c_{\mu/\mu,\lambda}$. The corresponding term for evolutionary gradient search is $\sqrt{\lambda}$ and results from the weighted summation of mutation vectors, where weights are determined by differences between objective function values.

- The influence of the noise-to-signal ratio $\vartheta = \sigma^*_\epsilon/\sigma^*$ is similar for all of the strategies considered. It appears in the form of the $\sqrt{1+\vartheta^2}$ term in the denominator of the positive contribution to the quality gain for the evolution strategies. For the original EGS procedure, additionally, the mutation strength enters this term due to the negative bias of mutations discussed above. The modified EGS strategy fixes that bias by using central differences, and it divides the square of the noise-to-signal ratio by two as it relies on two objective function evaluations per mutation vector rather than on a single one.
- The $(1, \lambda)$-ES with rescaled mutations as well as the modified EGS procedure can achieve the respective maximal quality gains that they can attain in the absence of noise even if noise is present. Increasing κ allows the strategies to operate with larger mutation strengths and thus to reduce the noise-to-signal ratio to zero. Notice that this rescaling is possible for any value of λ. The $(\mu/\mu, \lambda)$-ES is capable of operating with large mutation strengths by choosing μ and λ large. The original EGS procedure is not able to benefit from the rescaling of mutation vectors and is thus generally inferior in the presence of noise.
- Of the strategies considered, the $(\mu/\mu, \lambda)$-ES and the modified EGS procedure exhibit a linear increase in quality gain when increasing λ (and, for the $(\mu/\mu, \lambda)$-ES, μ). Both the $(1, \lambda)$-ES and the original EGS strategy scale sublinearly.

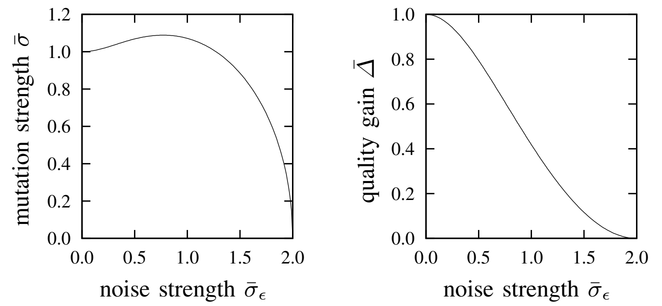

Fig. 6. Optimal standardized mutation strength $\bar{\sigma}$ and corresponding standardized quality gain $\bar{\Delta}$ of strategies on the noisy sphere model plotted against standardized noise strength $\bar{\sigma}_\epsilon$. The curves describe the behavior of the $(1, \lambda)$-ES with rescaled mutations, the $(\mu/\mu, \lambda)$-ES, and the modified EGS procedure and have been obtained from Eq. (27).

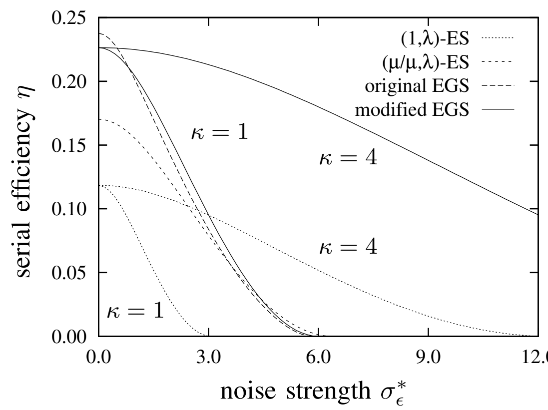

Fig. 7. Serial efficiency η plotted against normalized noise strength σ^*_ϵ for several evolution and evolutionary gradient search strategies. The strategies considered are the $(1, \lambda)$-ES with rescaled mutations ($\lambda = 10$, $\kappa \in \{1, 4\}$), the $(\mu/\mu, \lambda)$-ES ($\lambda = 10$, $\mu = 3$), the original EGS strategy ($\lambda = 9$), and the modified EGS procedure ($\lambda = 5$, $\kappa \in \{1, 4\}$).

Figure 7 visualizes the serial efficiencies of several evolution and evolutionary gradient search strategies on the noisy sphere model. All of the strategies considered perform ten objective function evaluations per time step. It can be seen from the figure that the $(1, \lambda)$-ES exhibits relatively low performance in the absence of noise, but that its robustness with regard to the effects of noise can be increased by increasing κ. The $(\mu/\mu, \lambda)$-ES is more efficient than the $(1, \lambda)$-ES in the absence of noise, but does not offer the same flexibility in its rescaling of mutation vectors. An increase in population size is necessary in order for better performance in the presence of noise to be achieved. However, as seen in [25], choosing μ and λ large can lead to a loss in serial efficiency in low-dimensional search spaces. The original EGS procedure is more efficient than both types of evolution strategy in the absence of noise, and, for the case considered, also in its presence. However, it does not benefit from an increase in λ to the degree that the $(\mu/\mu, \lambda)$-ES does, and it does not offer the possibility of rescaling in the presence of noise. For values of λ larger than those considered here, the $(\mu/\mu, \lambda)$-ES would be the better strategy in the absence of noise, and the $(1, \lambda)$-ES with rescaled mutations in its presence. Finally,

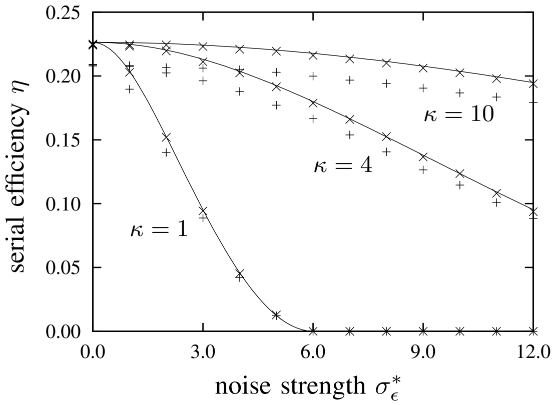

Fig. 8. Serial efficiency η of the modified EGS strategy with $\lambda = 5$ and $\kappa \in \{1, 4, 10\}$ plotted against normalized noise strength σ_ϵ^*. The points represent measurements from EGS runs on spheres with $N = 40$ (+) and $N = 400$ ($\times$). The solid lines have been obtained from Eq. (25).

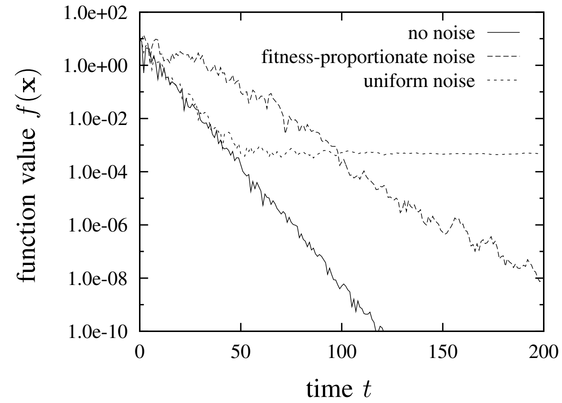

Fig. 9. Objective function value $f(\mathbf{x})$ of the search point on the sphere model plotted against time t. The curves represent measurements from typical EGS runs on the sphere model for the cases of no noise, fitness-proportionate noise, and noise of uniform strength. In all cases, the step length is controlled using cumulative step length adaptation as described in Section VI.

the modified EGS procedure combines the advantages of the two kinds of evolution strategy, while at the same time being nearly as efficient as the original EGS strategy in the absence of noise.

It is important to reemphasize that these results are idealizations in that they have been derived in the limit of infinite search space dimensionality. While they suggest that κ should be chosen arbitrarily large and that the efficiency of the noise-free case can be achieved for any level of noise present, in practice, finite search space dimensionalities limit useful values of κ. However, the results do provide a good understanding of the behavior of the modified EGS procedure and of the significance of its parameters if N is not too small. Figure 8 compares predictions from Eq. (25) with measurements of EGS runs in search spaces with $N = 40$ and $N = 400$. It can be seen that the quality of the approximation is very good for κ as large as ten if $N = 400$, and that it provides a good qualitative understanding for $N = 40$.

Finally, approximations like Eq. (25) describe the behavior of search strategies in single time steps. Different global behaviors of the algorithms can be observed depending on how the noise strength $\sigma_\epsilon(\mathbf{x})$ varies with the location in search space. Provided that the mutation strength is adapted successfully, without noise present EGS (like evolution strategies) exhibits stochastic linear convergence as illustrated in Fig. 9. The decrease in logarithmic function values of the search point is a random variable with a stationary distribution. When plotted against time, logarithmic objective function values fluctuate around a straight regression line the slope of which determines the speed of convergence. The frequently considered case of fitness-proportionate noise strength (i.e., $\sigma_\epsilon(\mathbf{x}) \propto f(\mathbf{x})$ and therefore $\sigma_\epsilon^*(\mathbf{x}) = const.$) leads to the same type of global behavior, but with a lower speed of convergence. Qualitatively different behavior results if the noise strength $\sigma_\epsilon(\mathbf{x})$ is independent of $\mathbf{x}$ and thus uniform throughout the search space. In that case, the strategy is not able to approach the optimum indefinitely but instead remains at a certain distance from it. As a consequence, objective function values of the search point approach a non-optimal limit value as indicated in Fig. 9. For the $(\mu/\mu, \lambda)$-ES, according to Beyer [3] that limit value is

$$f = \frac{N\sigma_\epsilon}{4\mu c_{\mu/\mu,\lambda}}. \tag{28}$$

For the modified EGS procedure, the corresponding result can be derived by using Eq. (26) to determine the point at which nonzero quality gain ceases to be possible. Assuming $\Delta^*_{\text{EGS}} = 0$, letting the mutation strength tend to zero, and using Eq. (21) yields

$$f = \frac{N\sigma_\epsilon}{4\kappa\sqrt{2\lambda}} \tag{29}$$

for the limit objective function value. Clearly, that value can be minimized by increasing λ or κ, where increasing κ has the stronger effect and the additional advantage of not increasing computational costs.

VI. Beyond the Sphere

In the realm of evolution strategies, the sphere model derives part of its significance from the existence of powerful mutation covariance matrix adaptation algorithms. Using a mutation covariance matrix that is adapted to the problem at hand can speed up the convergence of evolution strategies by several orders of magnitude. As recognized by Rudolph [26], ideally, the covariance matrix is the inverse of the Hessian matrix of the objective function at the current search point. In that case, the local performance of the evolution strategy is identical to that of an isotropic strategy on the sphere model. The covariance matrix adaptation algorithm of Hansen and Ostermeier [7] in particular has been found to successfully adapt the mutation covariance matrix for arbitrary convex quadratic objective functions. This section first briefly outlines that mechanism and describes the small modifications necessary to make it useful in connection with EGS. It then evaluates experimentally the performance of the resulting CMA-EGS strategy on a number of convex quadratic objective functions with greatly varying eigenvalue spectra.

A. Covariance Matrix Adaptation

The CMA-ES described in [7] accumulates consecutive search steps in order to provide information on the basis of which adaptation of the mutation covariance matrix is performed. Realizing that it may be advantageous to adapt the overall step length on a time scale shorter than that used for adapting the shape of the distribution, trial points are generated with covariance matrix $\sigma^2\mathbf{C}$, where step length parameter σ is adapted separately from symmetric, positive definite $N \times N$ matrix $\mathbf{C}$. Adaptation of the former uses the idea of cumulative step length adaptation introduced by Ostermeier et al [19]. Adaptation of the latter is done with the implicit goal of maximizing the probability of replicating successful steps. Somewhat inaccurately, $\mathbf{C}$ is referred to as the mutation covariance matrix. As the CMA-ES, CMA-EGS utilizes two N-dimensional vectors $\mathbf{s_C}$ and $\mathbf{s}_\sigma$ referred to as search paths that hold exponentially fading records of the most recently taken steps. An iteration of CMA-EGS updates the search paths along with the search point $\mathbf{x}$, the mutation strength σ, and matrix $\mathbf{C}$ using the following six steps (using "$\leftarrow$" to denote the assignment operator):

1) Compute an eigen decomposition $\mathbf{C} = \mathbf{BD}(\mathbf{BD})^{\mathrm{T}}$ of the mutation covariance matrix such that the columns of $N \times N$ matrix $\mathbf{B}$ are the normalized eigenvectors of $\mathbf{C}$ and $\mathbf{D}$ is a diagonal $N \times N$ matrix the diagonal elements of which are the square roots of the eigenvalues of $\mathbf{C}$.
2) Generate 2λ trial points
$$\mathbf{y}_{\pm}^{(i)} = \mathbf{x} \pm \sigma\mathbf{BDz}^{(i)}, \qquad i = 1, \ldots, \lambda,$$
where the $\mathbf{z}^{(i)}$ are mutation vectors consisting of N independent, standard normally distributed components.
3) Determine the objective function values $f(\mathbf{y}_{\pm}^{(i)})$ of the trial points and compute the weighted sum
$$\mathbf{z}^{(\mathrm{avg})} = \sum_{i=1}^{\lambda} \left(f(\mathbf{y}_{-}^{(i)}) - f(\mathbf{y}_{+}^{(i)}) \right) \mathbf{z}^{(i)}$$
of the $\mathbf{z}^{(i)}$ vectors.
4) Update the search point according to
$$\mathbf{x} \leftarrow \mathbf{x} + \sigma\mathbf{BD}\frac{\sqrt{N}}{\kappa}\frac{\mathbf{z}^{(\mathrm{avg})}}{\|\mathbf{z}^{(\mathrm{avg})}\|}.$$
5) Update the search paths according to
$$\mathbf{s_C} \leftarrow (1 - c_\mathbf{C})\mathbf{s_C} + \kappa\sqrt{c_\mathbf{C}(2 - c_\mathbf{C})}\mathbf{BDz}^{(\mathrm{avg})}$$
and
$$\mathbf{s}_\sigma \leftarrow (1 - c_\sigma)\mathbf{s}_\sigma + \kappa\sqrt{c_\sigma(2 - c_\sigma)}\mathbf{Bz}^{(\mathrm{avg})},$$
where $c_\mathbf{C} = c_\sigma = 4/(N + 4)$ as recommended in [7].
6) Update covariance matrix and step length according to
$$\mathbf{C} \leftarrow (1 - c_{\mathrm{cov}})\mathbf{C} + c_{\mathrm{cov}}\mathbf{s_C}\mathbf{s_C}^{\mathrm{T}}$$
and
$$\sigma \leftarrow \sigma \exp\left(\frac{\|\mathbf{s}_\sigma\|^2 - N}{2DN}\right),$$
where $c_{\mathrm{cov}} = 2/(N + \sqrt{2})^2$ and $D = 1 + 1/c_\sigma$ as recommended in [7].

Notice that steps 2) to 4) closely parallel steps 1) to 3) in Section II-A, except that there, mutations are isotropic. The adaptation procedure in steps 5) and 6) differs from that of the CMA-ES described in [7] only in two details: First, as in [20], [21], adaptation of the step length is performed on the basis of the squared length of the search path $\mathbf{s}_\sigma$ rather than based on its length. The change simplifies the formulas involved without significantly impacting performance. And second, the coefficients in the update rules for the search paths in step 5) have been modified in order to account for the difference between the lengths of the progress vectors of evolution strategies and EGS. Is is important to point out however that the settings of the cumulation and damping parameters have not been altered from the $(\mu/\mu, \lambda)$-CMA-ES, and that possibly better performance could be achieved if they were optimized for CMA-EGS.

Finally, realizing that the eigen decomposition in step 1) is expensive and that its cost may for large N outweigh the cost of evaluating the trial points, Hansen and Ostermeier [7] suggest to perform it only every $N/10$ steps, and to use slightly outdated matrices $\mathbf{B}$ and $\mathbf{D}$ in between. We have followed that suggestion for $N \geq 100$ without observing a significant loss in search performance.

B. Experimental Evaluation

In order to evaluate the performance of CMA-EGS and to compare it with that of the $(\mu/\mu, \lambda)$-ES, experiments have been run for a range of search space dimensionalities and noise levels on the sphere model as well as on the following three convex quadratic test functions:

$$\text{cigar:} \qquad f(\mathbf{y}) = y_1^2 + 10^6 \sum_{i=2}^{N} y_i^2$$

$$\text{discus:} \qquad f(\mathbf{y}) = 10^6 y_1^2 + \sum_{i=1}^{N} y_i^2$$

$$\text{ellipsoid:} \qquad f(\mathbf{y}) = \sum_{i=1}^{N} 10^{6(i-1)/(N-1)} y_i^2$$

Note that while the functions are separable and their respective optima are at the origin of the coordinate system, CMA-EGS (as the $(\mu/\mu, \lambda)$-CMA-ES) does not make use of these facts. Its behavior is invariant with regard to rigid transformations of the coordinate system, and the experiments could have been carried out using arbitrarily rigidly transformed coordinates without changing the results. The sphere was included in the comparison as it is the easiest of all convex quadratic test functions. On the sphere, there is no need for adapting the mutation covariance matrix (provided that it is initialized to the identity matrix). The behavior of CMA-EGS on the sphere is well described by the results derived for the isotropic strategy in Sections III through V. The remaining three test functions do require adaptation of the mutation covariance matrix in order for CMA-EGS to achieve good performance. After adaptation is complete, CMA-EGS achieves the same rate of convergence on any of the convex quadratic test functions as it

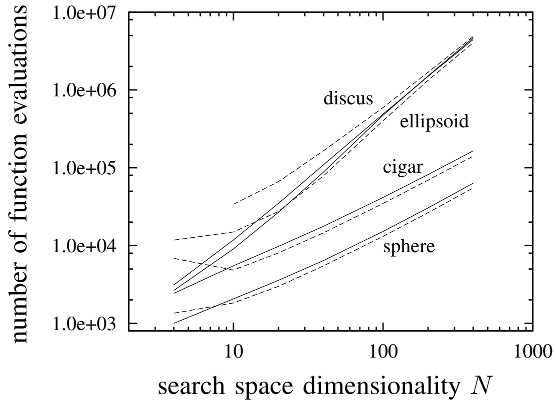

Fig. 10. Number of function evaluations required to reach $f_{\text{stop}} = 10^{-10}$ plotted against the search space dimensionality N. Solid lines represent results measured for the $(3/3, 10)$-ES, dashed ones those for the modified EGS strategy with $\lambda = 5$ and $\kappa = 1$.

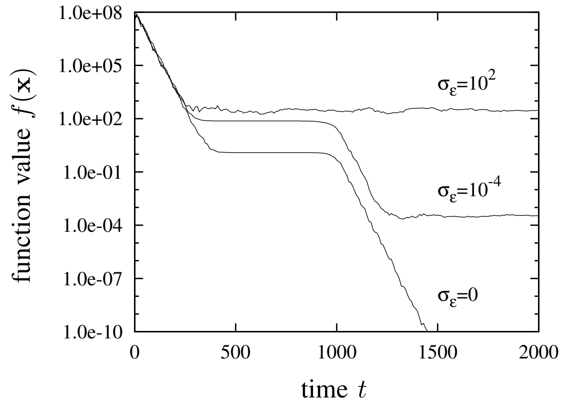

Fig. 11. Objective function value $f(\mathbf{x})$ of the search point plotted against time t for typical runs of CMA-EGS with $\lambda = 5$ and $\kappa = 1$ on the cigar with uniform noise strengths $\sigma_\epsilon \in \{0, 10^{-4}, 10^2\}$.

does on the sphere. The length of the adaptation phase depends on the eigenvalue spectrum of the objective function. All three of the functions considered have condition number 10^6, but they differ in their respective eigenvalue spectra. The cigar has one eigenvalue much smaller than all of the others and resembles a sphere that is stretched by a factor of 1000 in one dimension. Similarly, the discus has one eigenvalue that is much larger than all of the others and resembles a sphere that is squashed by a factor of 1000 in one dimension. The ellipsoid has an eigenvalue spectrum that is not dominated by any one value.

For all experiments reported below, the initial search point is $\mathbf{x} = (1, \ldots, 1)^{\mathrm{T}}$. The strategies' initial step length and mutation covariance matrix are $\sigma = 1$ and $\mathbf{C} = \mathbf{1}_{N\times N}$, respectively. Search paths $\mathbf{s}_{\mathbf{C}}$ and $\mathbf{s}_\sigma$ are zero initially. The experimental conditions are thus the same as in [7].

Figure 10 compares the performance of the CMA-EGS strategy with $\lambda = 5$ and $\kappa = 1$ with that of the $(3/3, 10)$-CMA-ES in the absence of noise. Notice that both strategies generate and evaluate ten trial points per time step and are therefore immediately comparable. The experiments serve as a test of the general ability of the covariance matrix adaptation mechanism to generate good mutation distributions for EGS. As a performance measure, the number of function evaluations required to reach an objective function value of $f_{\text{stop}} = 10^{-10}$ has been used. Tests have been conducted for search space dimensionalities $N \in \{4, 10, 20, 40, 100, 200, 400\}$ and results have been averaged over a number of test runs. The standard deviation of the measurements (not shown) is negligible except for $N = 4$. The performance advantage of CMA-EGS for $N \geq 10$ amounts to the strategy reaching f_{stop} using roughly 15% to 20% fewer function evaluations than the $(\mu/\mu, \lambda)$-CMA-ES. The speed-up is thus somewhat below the theoretical predictions derived for isotropic mutations and $N \to \infty$ in Section V but nonetheless significant. However, this does not hold true for $N = 4$ where the evolution strategy is the more efficient algorithm. Experiments with isotropic mutations and cumulative step length adaptation confirm that the reason for the inferior performance of EGS in low-dimensional search spaces is the step length adaptation component of the algorithm. Unfortunately, the analysis in Section III is not helpful for understanding the source of the imperfect step length adaptation in low-dimensional search spaces. It appears likely that an experimental approach will need to be used to shed light on the bad performance of cumulative step length adaptation when used in conjunction with EGS in low-dimensional search spaces. For the other test functions, the deficiency of cumulative step length adaptation when using EGS in low-dimensional search spaces is even more pronounced. For the cigar, the performance advantage of CMA-EGS is about the same as for the sphere for $N \geq 10$, but the strategy requires more function evaluations for $N = 4$ than it does for $N = 10$. For the ellipsoid, it is not until $N = 40$ that CMA-EGS is more efficient than the $(\mu/\mu, \lambda)$-CMA-ES. Finally, for the discus, while the curves suggest that this may change for larger values of N, the $(\mu/\mu, \lambda)$-ES is faster than CMA-EGS even for $N = 400$. For $N = 4$, CMA-EGS on the discus fails to reach f_{stop} altogether.

In order to compare the performance of CMA-EGS with that of the $(\mu/\mu, \lambda)$-CMA-ES in the presence of noise, it is important to consider a range of noise strengths. As can be seen from Fig. 11 for the special case of the cigar, in the absence of noise a run of CMA-EGS on functions other than the sphere consists of three distinct phases. Initially, the strategy converges linearly in spite of a badly adapted mutation distribution. At some point, progress begins to stagnate and the strategy spends a considerable amount of time "learning" a better mutation distribution. In the third phase, the mutation covariance matrix is proportional to the inverse of the Hessian of the objective function and the strategy again progresses as fast as it would on the sphere model. In the presence of noise of uniform strength, progress comes to a halt in either the first or the third of those phases, as illustrated in Fig. 11, depending on the strength of the noise present. In either case, objective function values of the search point fluctuate around a limit value that can serve as a performance measure for the strategies, where lower function values reflect better performance.

Figure 12 shows measured objective function values of the search point for both the $(3/3, 10)$-CMA-ES and for the CMA-EGS strategy with $\lambda = 5$ and $\kappa \in \{1, 20\}$. Measurements have been conducted for uniform noise strength and have been made after twice as many time steps as are required by the $(\mu/\mu, \lambda)$-ES to reach $f_{\text{stop}} = 10^{-10}$ in the absence of noise for the respective objective function. Except for the case of large noise strengths on the discus, this is sufficient to reach the state in which objective function values fluctuate around a stationary limit value. It can be seen from the figure that for the sphere, the limit value reached is virtually identical for the $(3/3, 10)$-CMA-ES and the CMA-EGS strategy with $\lambda = 5$ and $\kappa = 1$. It closely agrees with the values predicted by Eqs. (28) and (29). The CMA-EGS strategy with $\kappa = 20$ approaches the optimum closer by a factor of 20 as predicted by Eq. (29), underlining the significance of the theoretical insights gained with regard to the isotropic strategy. While the standard deviation of the measurements is very small for the sphere, relatively large standard deviations can be observed on the cigar and on the ellipsoid. As for the sphere, the performance of the $(3/3, 10)$-CMA-ES and of the CMA-EGS strategy with $\lambda = 5$ and $\kappa = 1$ are nearly identical while choosing $\kappa = 20$ substantially improves the performance of the latter strategy. For the discus, similar to the situation in the absence of noise, the performance of the CMA-EGS strategy with $\lambda = 5$ and $\kappa = 1$ is inferior to that of the $(3/3, 10)$-ES. However, as for the other objective functions, choosing $\kappa = 20$ makes the CMA-EGS strategy superior to the evolution strategy.

VII. Summary and Conclusions

To conclude, the present paper has presented an analysis on the infinite-dimensional sphere model of the behavior of the EGS strategy with the modifications proposed in [13]. It has been seen that the modifications are useful in that they both allow for an efficient parallelization and for the possibility of improved performance in the presence of noise of the strategy. The modified EGS procedure is about 25% more efficient than a comparable $(\mu/\mu, \lambda)$-ES in the absence of noise. On the noisy sphere, the flexibility that the rescaling parameter κ affords arguably makes EGS more powerful than the evolution strategy. Choosing κ large makes it possible to emulate the benefits of the genetic repair effect without the need to operate with a large number of trial points generated per time step. Numerical experiments have confirmed the value of the predictions and insights in finite-dimensional search spaces. Altogether, the modified EGS procedure combines the favorable scaling with λ of the $(\mu/\mu, \lambda)$-ES with the flexibility in the choice of κ of the $(1, \lambda)$-ES with rescaled mutations and the superior efficiency of the original EGS procedure.

Then, the covariance matrix adaptation mechanism originally proposed by Hansen and Ostermeier [7] for the $(\mu/\mu, \lambda)$-ES has been adapted for use in EGS. Numerical experiments have shown that the adaptation of the shape of the mutation distribution works for EGS as well as is does for evolution strategies, and that EGS generally maintains its performance advantage over the $(\mu/\mu, \lambda)$-ES. However, the experiments have also revealed deficiencies with regard to step length adaptation in low-dimensional search spaces. Those deficiencies are

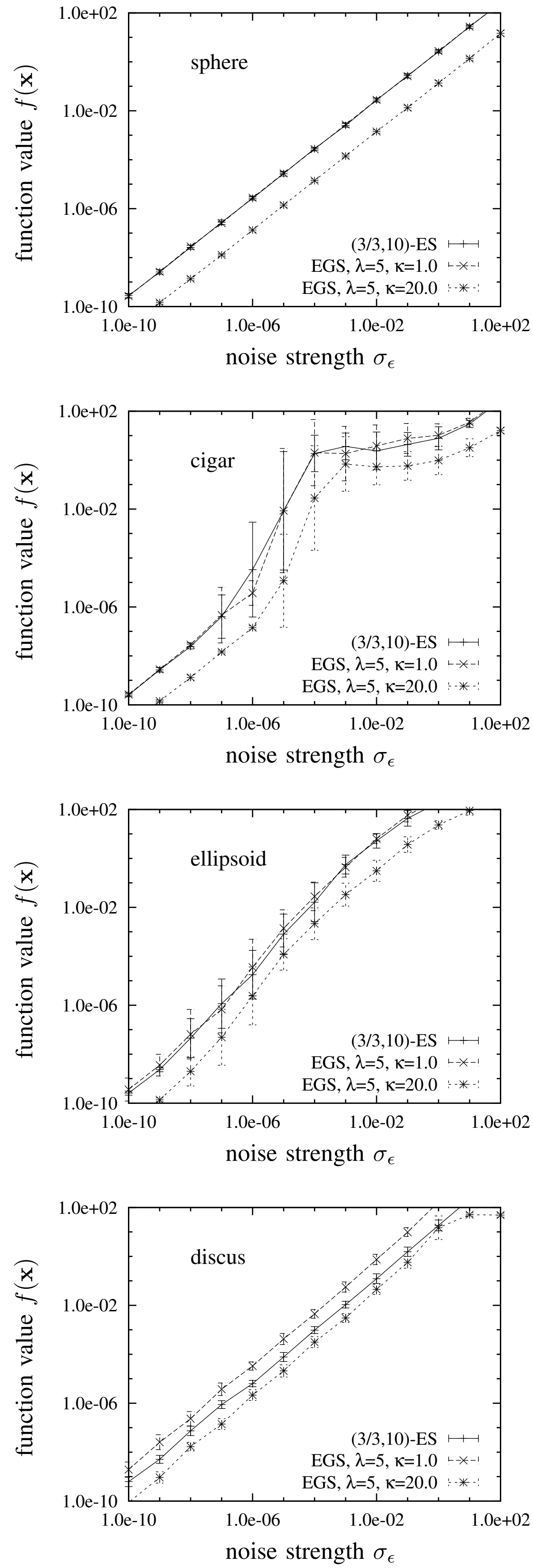

Fig. 12. Limit function value $f(\mathbf{x})$ of the search point plotted against noise strength σ_ϵ for the four test functions with $N = 40$. Both mean values and standard deviations of the measurements are shown.

more pronounced on some objective functions than they are on others, and more work will be required in order to understand their sources. However, despite those deficiencies, CMA-EGS proved to be generally superior to the $(\mu/\mu, \lambda)$-CMA-ES in case of uniform noise strength due to its ability to choose the rescaling factor κ independently of the number of trial solutions generated per time step.

Several questions remain to be addressed in future research. First, work needs to be done with the goal of understanding the reasons for the imperfect step length adaptation of EGS in low-dimensional search spaces. Second, the experimental evaluation of CMA-EGS and its comparison with the $(\mu/\mu, \lambda)$-CMA-ES should be extended beyond the convex quadratic test functions considered here. Hansen and Kern [27] have experimentally investigated global search properties of the latter strategy. It could be expected that the better local search properties of EGS come at the cost of inferior global search capabilities, and it remains to be seen if and to what degree the flexibility in the choice of κ can compensate. Third, Hansen et al. [28] have proposed a modification of the covariance matrix adaptation mechanism that enables much faster adaptation if the number of trial solutions generated per time step is large and if candidate solutions can be evaluated in parallel. It remains to be seen whether those modifications have the same beneficial effects in EGS as they have for evolution strategies. Fourth, recent work [30] has introduced weighted multirecombination evolution strategies. Those strategies differ from EGS in that they assign weights to mutation vectors based on the rank of the respective candidate solutions in the set of all offspring. A detailed comparison of the capabilities of weighted multirecombination evolution strategies and EGS remains to be done. And finally, in [29] a mechanism has been proposed for the adaptation of the rescaling factor κ in isotropic weighted multirecombination evolution strategies. It is of great interest to see whether that mechanism is useful in connection with CMA-EGS as well.

ACKNOWLEDGMENTS

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

REFERENCES

[1] I. Rechenberg, *Evolutionsstrategie '94*. Frommann-Holzboog Verlag, Stuttgart, 1994.
[2] H.-P. Schwefel, *Evolution and Optimum Seeking*. John Wiley & Sons, New York, 1995.
[3] H.-G. Beyer, *The Theory of Evolution Strategies*. Springer Verlag, Heidelberg, 2001.
[4] H.-G. Beyer and H.-P. Schwefel, "Evolution strategies — A comprehensive introduction," *Natural Computing*, vol. 1, no. 1, pp. 3–52, 2002.
[5] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer Verlag, Heidelberg, 2000.
[6] R. Salomon, "Evolutionary algorithms and gradient search: Similarities and differences," *IEEE Transactions on Evolutionary Computation*, vol. 2, no. 2, pp. 45–55, 1998.
[7] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, 2001.
[8] C. T. Kelley, *Iterative Methods for Optimization*. SIAM, Philadelphia, 1999.
[9] H.-G. Beyer, "On the "explorative power" of ES/EP-like algorithms," in *Evolutionary Propgramming VII*, A. E. Eiben, V. Porto, N. Saravanan, and D. Waagen, Eds. Springer Verlag, Heidelberg, 1998, pp. 323–334.
[10] D. V. Arnold, "An analysis of evolutionary gradient search," in *Proceedings of the IEEE Congress on Evolutionary Computation — CEC 2004*. IEEE Press, Piscataway, NJ, 2004, pp. 47–54.
[11] H.-G. Beyer, "Toward a theory of evolution strategies: On the benefit of sex – the $(\mu/\mu, \lambda)$-theory," *Evolutionary Computation*, vol. 3, no. 1, pp. 81–111, 1995.
[12] D. V. Arnold and H.-G. Beyer, "Local performance of the $(\mu/\mu, \lambda)$-ES in a noisy environment," in *Foundations of Genetic Algorithms 6*, W. N. Martin and W. M. Spears, Eds. Morgan Kaufmann Publishers, San Francisco, 2001, pp. 127–141.
[13] R. Salomon, "Inverse mutations: Making the evolutionary-gradient-search procedure noise robust," in *Proceedings of Artificial Intelligence and Applications — AIA 2005*, M. H. Hamza, Ed. Acta Press, Calgary, AB, 2005.
[14] H.-G. Beyer, "Mutate large, but inherit small! On the analysis of rescaled mutations in $(\tilde{1}, \tilde{\lambda})$-ES with noisy fitness data," in *Parallel Problem Solving from Nature — PPSN V*, A. E. Eiben, T. Bäck, M. Schoenauer, and H.-P. Schwefel, Eds. Springer Verlag, Heidelberg, 1998, pp. 109–118.
[15] ——, "Evolutionary algorithms in noisy environments: Theoretical issues and guidelines for practice," *Computer Methods in Mechanics and Applied Engineering*, vol. 186, pp. 239–267, 2000.
[16] D. V. Arnold and H.-G. Beyer, "A comparison of evolution strategies with other direct search methods in the presence of noise," *Computational Optimization and Applications*, vol. 24, no. 1, pp. 135–159, 2003.
[17] I. Rechenberg, *Evolutionsstrategie — Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Friedrich Frommann Verlag, Stuttgart, 1973.
[18] H.-P. Schwefel, *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie*. Birkhäuser Verlag, Basel, 1977.
[19] A. Ostermeier, A. Gawelczyk, and N. Hansen, "Step-size adaptation based on non-local use of selection information," in *Parallel Problem Solving from Nature — PPSN III*, Y. Davidor, R. Männer, and H.-P. Schwefel, Eds. Springer Verlag, Heidelberg, 1994, pp. 189–198.
[20] D. V. Arnold, *Noisy Optimization with Evolution Strategies*. Kluwer Academic Publishers, Boston, 2002.
[21] D. V. Arnold and H.-G. Beyer, "Performance analysis of evolutionary optimization with cumulative step length adaptation," *IEEE Transactions on Automatic Control*, vol. 49, no. 4, pp. 617–622, 2004.
[22] L. D. Whitley, M. Lunacek, and J. Knight, "Ruffled by ridges: How evolutionary algorithms can fail," in *Genetic and Evolutionary Computation Conference — GECCO 2004 (2)*, K. Deb, R. Poli, W. Banzhaf, H. G. Beyer, E. K. Burke, P. J. Darwen, D. Dasgupta, D. Floreano, J. A. Foster, M. Harman, O. Holland, P. L. Lanzi, L. Spector, A. Tettamanzi, D. Thierens, and A. M. Tyrrell, Eds. Springer Verlag, Heidelberg, 2004, pp. 294–306.
[23] R. Salomon, "Accelerating the evolutionary-gradient-search procedure: Individual step sizes," in *Parallel Problem Solving from Nature — PPSN V*, T. Bäck, A. E. Eiben, M. Schoenauer, and H.-P. Schwefel, Eds. Springer Verlag, Heidelberg, 1998, pp. 408–417.
[24] J. Poland and A. Zell, "Main vector adaptation: A CMA variant with linear time and space complexity," in *Proceedings of the Genetic and Evolutionary Computation Conference — GECCO 2001*, L. Spector, E. D. Goodman, A. Wu, W. B. Langdon, H.-M. Voigt, M. Gen, S. Sen, M. Dorigo, S. Pezeshk, M. H. Garzon, and E. Burke, Eds. Morgan-Kaufmann Publishers, San Francisco, 2001, pp. 1050–1055.
[25] D. V. Arnold and H.-G. Beyer, "Performance analysis of evolution strategies with multi-recombination in high-dimensional $\mathbb{R}^N$-search spaces disturbed by noise," *Theoretical Computer Science*, vol. 289, no. 1, pp. 629–647, 2002.
[26] G. Rudolph, "On correlated mutations in evolution strategies," in *Parallel Problem Solving from Nature — PPSN II*, R. Männer and B. Manderick, Eds. Elsevier, Amsterdam, 1992, pp. 105–114.
[27] N. Hansen and S. Kern, "Evaluating the CMA evolution strategy on multimodal test functions," in *Parallel Problem Solving from Nature — PPSN VIII*, X. Yao, E. Burke, J. A. Lozano, J. Smith, J. J. Merelo-Guervós, J. A. Bullinaria, J. Rowe, P. Tiňo, A. Kabán, and H.-P. Schwefel, Eds. Springer Verlag, Heidelberg, 2004, pp. 282–291.
[28] N. Hansen, S. D. Müller, and P. Koumoutsakos, "Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES)," *Evolutionary Computation*, vol. 11, no. 1, pp. 1–18, 2003.
[29] D. V. Arnold, "Evolution strategies with adaptively rescaled mutation vectors," in *Proceedings of the IEEE Congress on Evolutionary Computation — CEC 2005*. IEEE Press, Piscataway, NJ, 2005.

[30] ——, "Optimal weighted recombination," in *Foundations of Genetic Algorithms 8*, A. H. Wright, M. D. Vose, K. A. De Jong, and L. M. Schmitt, Eds. Springer Verlag, Heidelberg, 2005, pp. 215–237.