Qualms Regarding the Optimality of Cumulative Path Length Control in CSA/CMA-Evolution Strategies

Hans-Georg Beyer beyer@LS11.cs.uni-dortmund.de
Department of Computer Science XI, University of Dortmund, D-44221 Dortmund, Germany

Dirk V. Arnold arnold@LS11.cs.uni-dortmund.de
Department of Computer Science XI, University of Dortmund, D-44221 Dortmund, Germany

Abstract

The cumulative step-size adaptation (CSA) based on path length control is regarded as a robust alternative to the standard mutative self-adaptation technique in evolution strategies (ES), guaranteeing an almost optimal control of the mutation operator. In this short paper it is shown that the underlying basic assumption in CSA – the perpendicularity of expected consecutive steps – does not necessarily guarantee optimal progress performance for $(\mu/\mu_I, \lambda)$ intermediate recombinative ES.

Keywords

covariance matrix adaptation, cumulative step-size adaptation, evolution strategies, mutative self-adaptation, progress rate

1 Introduction and Problem Description

The local optimization performance of evolution strategies (ES) in real-valued search spaces $\mathbb{R}^N$ relies heavily on the strength of the mutations used and the shape of their distribution. In order to keep pace with the evolution process and obtain maximal (local) performance (i.e., to obtain the greatest improvement in the next generation step), the general mutation strength σ and – considering arbitrary normally distributed mutations – the covariance matrix $\mathbf{C}$ as the endogenous strategy parameters must be adapted online during the evolution process. This adaptation process is usually realized by so-called self-adaptation (SA) techniques.

The standard SA is based on the mutative step-size control paradigm (also referred to as mutative step-size adaptation, MSA) proposed by Rechenberg and Schwefel (see Rechenberg, 1973; Schwefel, 1981; Bäck & Schwefel, 1993; Bäck, Hammel, & Schwefel, 1997): Each individual has its own set of endogenous strategy parameters subject to variation (mutation and recombination) and the entire genetic information is inherited according to the individual's fitness. That is, those strategy parameters that belong to the fittest individuals are likely to survive. As has been shown in Beyer (1996) (see also Beyer, 2001), this adaptation technique is able to realize optimal performance on the sphere model in the case of the $(1, \lambda)$-σSA-ES (for performance definition, see Sect. 2.1); furthermore, this behavior is insensitive to the σ-mutation rule and the learning parameter used. Unlike the $(1, \lambda)$-σSA-ES, newer investigations regarding $(\mu/\mu, \lambda)$-σSA strategies have revealed, however, a sensitive dependence of the performance on the learning parameter (Grünz & Beyer, 1999). That is, $(\mu/\mu, \lambda)$-σSA-ES exhibit optimal performance only when the learning parameter is tuned accordingly.

During the mid-1990s an alternative adaptation strategy was proposed by Ostermeier, Gawelczyk, and Hansen (1994, 1995), the so-called cumulative step-size adaptation (CSA), that promised an improved and more reliable adaptation behavior. Unlike the mutation strength adaptation by MSA that uses one-generation fitness ranking information only and neglects the effect of recombination,[1] the CSA relies on fitness related search space information gathered over a sequence of consecutive generations. In CSA strategies as well as in the CMA-ES (CMA – covariance matrix adaptation, not considered in detail here[2]) the length of so-called evolution paths is used to control the variance σ^2 of the object parameter mutation operator (therefore it is sometimes referred to as cumulative path length control).

The evolution path $\mathbf{s}^{(g)}$ (g – generation counter) is a weighted vector sum of the actually realized steps $\mathbf{z}^{(g)}$ in the object parameter search space $\mathbb{R}^N$ (see (L4) in Eq. (1), below). The basic idea of cumulative path length control is explained in Hansen and Ostermeier (1996):

> "The evolution path mainly reveals information on *correlations* between mutation steps successively selected in the generation sequence. If successively selected mutation steps are parallel correlated (scalar product greater zero), the evolution path will be comparatively long. If successively selected mutation steps are anti-parallel correlated (scalar product less than zero), the evolution path will be comparatively short. Roughly speaking, parallel correlation means that successive steps are going into the same direction, and thus the same distance could be covered by fewer but longer steps. Anti-parallel correlation means, that the steps cancel each other out. Both is inefficient with respect to the single mutation step. Consequently, to make single mutation steps most efficient, it is the best to have no correlation between the selected mutation steps in the evolution path."

This philosophy culminates in (Hansen & Ostermeier, 1996, p.312):

> "The geometrical interpretation is, that successively selected mutation steps should be perpendicular to each other (apart from stochastic deviations)." and further in the:
>
> "fundamental adaptation principle . . .: Reasonable adaptation has to *reduce the difference between the distributions of the actual evolution path and an evolution path under random selection, . . .*" and:
>
> ". . . as substantiated by experiments, this [fundamental principle] leads to selected steps being uncorrelated and adapts optimal step size precisely."

Based on this fundamental principle an update rule for the $(\mu/\mu_I, \lambda)$-ES with standard CSA and isotropic mutations has been proposed (Hansen & Ostermeier, 1996) which

[1]Survival of the strategy parameter σ_l depends on the fitness of the lth offspring's object parameter set $\mathbf{y}_l$ which is generated by a mutation with strength σ_l. Recombination is applied *after* selection. Thus, MSA strategies cannot directly account for the effect of recombination.

[2]For an excellent introduction into that matter, Hansen and Ostermeier (2001) is recommended.

can be expressed as:

$$
\left.
\begin{aligned}
&\forall l = 1, \ldots, \lambda: \quad \mathbf{w}_l^{(g)} := \sigma^{(g)} \vec{\mathcal{N}}_l(0,1), && (L1)\\
&\mathbf{z}^{(g)} := \frac{1}{\mu} \sum_{m=1}^{\mu} \mathbf{w}_{m;\lambda}^{(g)}, && (L2)\\
&\mathbf{y}^{(g+1)} := \mathbf{y}^{(g)} + \mathbf{z}^{(g)}, && (L3)\\
&\mathbf{s}^{(g+1)} := (1-c)\mathbf{s}^{(g)} + \sqrt{c(2-c)} \frac{\sqrt{\mu}}{\sigma^{(g)}} \mathbf{z}^{(g)}, && (L4)\\
&\sigma^{(g+1)} := \sigma^{(g)} \exp\left[\frac{\|\mathbf{s}^{(g+1)}\| - \overline{\chi}_N}{D\overline{\chi}_N}\right]. && (L5)
\end{aligned}
\right\} \qquad (1)
$$

Here, the first three lines realize the standard $(\mu/\mu_I, \lambda)$-ES operations in object parameter space, i.e., produce λ normally distributed mutations $\mathbf{w} \in \mathbb{R}^N$ (L1), recombine the μ fittest mutations by a center of mass operation (L2), and update the parental state (L3).[3] The CSA is realized in the lines (L4) and (L5). In (L4), the actual evolution path is cumulated in a weighted fashion yielding the vector s. In (L5), the length difference of s and $\overline{\chi}_N$ ($\overline{\chi}_N$ is the expected value of the length χ_N of an N-dimensional random vector with $\mathcal{N}(0,1)$ standard normally distributed components) is used to change the mutation strength σ.[4] It is claimed by the designers of the CSA rules (L4), (L5) that – provided stationary conditions – this rule allows for an optimal control of the mutation strength, such that the progress rate φ on the sphere becomes nearly maximal (for the definitions, see below).

Even though the philosophy behind the "fundamental adaptation principle" is intuitively appealing, the claimed optimality of the perpendicularity condition between selected evolution steps still remained obscure. In (Hansen, 1998, p.5–7) a geometric explanation has been offered that relates this condition to the local progress performance of the ES. The aim of this short paper is to reconsider the arguments presented in Hansen (1998) and to show that the underlying assumptions are only valid for strategies where the selection information is *not* disturbed by random sources. The investigations to be presented here have been triggered by empirical observations reported in Arnold and Beyer (2000a): Where it was found that the $(\mu/\mu_I, \lambda)$-ES with CSA can fail when the fitness information is disturbed by a certain relative measuring error.

The remainder of this article is organized as follows. First, the geometric basis of the perpendicularity condition is investigated thoroughly. Secondly, the theoretical predictions are compared with simulations, and finally, conclusions are drawn and an outlook is given.

2 Performance Optimality vs. Perpendicularity

2.1 Progress Rate and Optimality Condition

The local performance of ES in search space $\mathbb{R}^N$ is usually measured by the progress rate φ as the expected value of the one-generation distance change to the optimum. Let

[3] As in standard ES, $\mathbf{y} \in \mathbb{R}^N$ refers to the object parameter vector; the fitness of which is given by $F = F(\mathbf{y})$. The λ offspring states are generated by $\tilde{\mathbf{y}}_l^{(g)} = \mathbf{y}^{(g)} + \mathbf{w}_l^{(g)}$. $\mathbf{w}_{m;\lambda}$ refers to the mutation that produced the mth best offspring w.r.t. its (measured, i.e., observed) fitness value $F_{m;\lambda}^{(g)} = F(\mathbf{y}^{(g)} + \mathbf{w}_{m;\lambda}^{(g)})$.

[4] The cumulation time parameter c, $0 \le c \le 1$, is usually chosen $c \propto 1/\sqrt{N}$, the damping parameter D, $D \propto \sqrt{N}$, and $\overline{\chi}_N = \sqrt{2}\,\Gamma\left(\frac{N+1}{2}\right) \Big/ \Gamma\left(\frac{N}{2}\right)$ (see Hansen & Ostermeier, 1997).

$\mathbf{R}$ be the vector from the optimum point $\hat{\mathbf{y}}$ to the parental center of mass at generation g and $\mathbf{r}$ the same vector at $g+1$ (see Fig. 1). The progress rate is defined as the expected value $\varphi := \mathrm{E}[R - r]$, where $R := \|\mathbf{R}\|$ and $r := \|\mathbf{r}\|$. Considering the sphere model, the fitness function F which is defined as $F(\mathbf{y}) := Q(\|\mathbf{y} - \hat{\mathbf{y}}\|) + \varepsilon_{\mathbf{y}} = Q(R) + \varepsilon_{\mathbf{y}}$ ($Q(R)$ – monotonic function), an asymptotically exact ($N \to \infty$) normalized progress rate expression for $(\mu/\mu_I, \lambda)$-ES including normally distributed fitness noise $\varepsilon_{\mathbf{y}}$, with standard deviation σ_ε and zero mean, can be derived (for details, see Arnold & Beyer, 2001)

$$\varphi^* = \sigma^{*2} \left[\frac{c_{\mu/\mu,\lambda}}{\sqrt{\sigma^{*2} + \sigma_\varepsilon^{*2}}} - \frac{1}{2\mu} \right] \tag{2}$$

($c_{\mu/\mu,\lambda}$ – progress coefficient which is defined as the expectation of the average over the first μ order statistics $x_{m:\lambda}$ of the $x \sim \mathcal{N}(0,1)$ random variate, μ – number of parents, λ – number of offspring) with the normalizations

$$\varphi^* := \varphi \frac{N}{R}, \quad \sigma^* := \sigma \frac{N}{R}, \quad \sigma_\varepsilon^* := \sigma_\varepsilon \frac{N}{Q'R}, \quad Q' = \frac{\mathrm{d}Q(R)}{\mathrm{d}R}, \tag{3}$$

where σ is the standard deviation of the mutation operator and σ_ε^* is the normalized standard deviation of the fitness noise source.[5] As one can see, local performance of the $(\mu/\mu_I, \lambda)$-ES on the sphere model depends on the fitness model $Q(R)$, the current distance R to the optimum, the fitness noise, and the mutation strength σ.

Given the $(\mu/\mu_I, \lambda)$-ES parameters, the fitness model, and the φ-measure, performance optimality w.r.t. the mutation operator is defined as

$$\hat{\varphi}^* := \max_{\sigma^*} \varphi^*(\sigma^*) \quad \text{with} \quad \hat{\sigma}^* := \arg\max_{\sigma^*} \varphi^*(\sigma^*). \tag{4}$$

That is, we are interested in the (locally) optimal mutation strength $\hat{\sigma}$ that provides the greatest improvement in the next generation step. The $\hat{\sigma}^*$ can be calculated from (2). In the case $\sigma_\varepsilon^* = 0$ one easily finds $\hat{\sigma}^* = \mu c_{\mu/\mu,\lambda}$, however, for $\sigma_\varepsilon^* > 0$ the optimal σ^* has no simple solution, a degree three algebraic equation in σ^{*2} must be solved instead. As known from algebra, such equations cannot be solved by simple geometric means (i.e., by ruler and circles). Therefore, it is in principle excluded that a general perpendicularity condition corresponds to the optimality condition (4) in the noisy case.

2.2 Perpendicularity Condition

Let us now have an alternative geometrically motivated view of performance optimality. In Fig. 1 a two-dimensional snapshot of the $\mathbb{R}^N$, spanned by the two parental center of mass vectors at generation g and $g+1$, $\mathbf{R}$ and $\mathbf{r}$, respectively, is displayed. The state change from g to $g+1$ is by the vector $\mathbf{z}$. Since $\mathbf{z}$ is generated by (L1) and (L2) in Eq. (1), the length of $\mathbf{z}$ depends on σ. Let α be the angle between $\mathbf{z}$ and direction to the optimum $-\mathbf{e}_R$. Provided that $\alpha = \mathrm{const.}$, i.e., that α does *not* depend on σ, optimal performance is obtained for that σ-value that yields $\mathbf{z}_{\mathrm{opt}}$ (see Fig. 1), because $\mathbf{z}_{\mathrm{opt}}$ yields the smallest $r = \|\mathbf{r}\|$ (recall, performance optimality is defined as the greatest improvement step minimizing the r of the next generation). Since α is assumed to be constant, by elementary geometry, optimality is equivalent to the orthogonality of $\mathbf{z}_{\mathrm{opt}}$ and $\mathbf{r}$, i.e., $\hat{\sigma}$ is alternatively determined by $\mathbf{r}^{\mathrm{T}}\mathbf{z}_{\mathrm{opt}} = 0$.[6] This condition transfers to the

[5] Due to definition (3), σ_ε^* is basically proportional to the relative measuring error.

[6] A similar condition can be obtained for the CMA-ES with correlated mutations on the general quadratic fitness model. Let $\mathbf{C}$ be the covariance matrix of the mutation operator, the condition reads $\mathbf{r}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{z}_{\mathrm{opt}} = 0$.

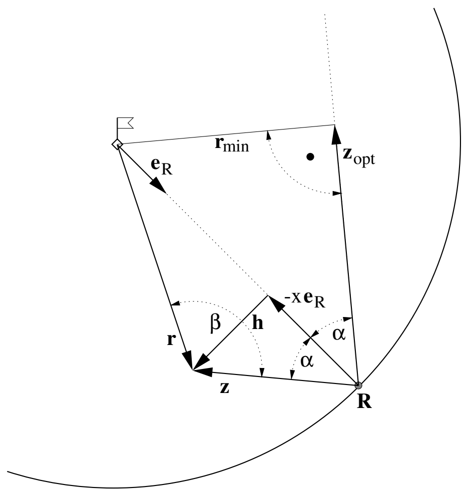

Figure 1: Snapshot of the search space $\mathbb{R}^N$. The two-dimensional projection is spanned by the optimum point (labeled by a flag) and the two parental center of mass states $\mathbf{R}$ and $\mathbf{r}$ at generation g and $g+1$, respectively.

expected next generation step because in the mean $\overline{\mathbf{z}^{(g+1)}}$ is (anti)parallel to $\mathbf{r}$ (due to the symmetry of the mutations used), i.e., $\mathbf{r} = -\kappa\overline{\mathbf{z}^{(g+1)}}$ does hold (κ - scalar factor). Thus, one ends up with the perpendicularity condition $\mathbf{z}_{\mathrm{opt}}^{(g)\mathrm{T}}\overline{\mathbf{z}^{(g+1)}} = \overline{\mathbf{z}^{(g)T}}\,\overline{\mathbf{z}^{(g+1)}} = 0$ (cf. Hansen, 1998, p.60.).

The crucial point in the derivation of the perpendicularity condition is in the $\alpha = \mathrm{const.}$ assumption. To see this, we consider the decomposition of the $\mathbf{z}$ vectors according to the *evolutionary progress principle* (Beyer, 1997) into a gain component x and a loss part $\mathbf{h}$: $\mathbf{z} = -x\mathbf{e}_R + \mathbf{h}$. That is, we decompose $\mathbf{z}$ into a component $-x\mathbf{e}_R$ parallel to optimum direction and in a perpendicular part $\mathbf{h}$. Thus, α can be calculated as $\alpha = \mathrm{arc}\cot(x/\|\mathbf{h}\|)$. Since x and $\mathbf{h}$ are random variates, α is a random variate, too. Its expected value $\overline{\alpha}$ can be approximated by $\overline{\alpha} \simeq \mathrm{arc}\cot(\overline{x}/\overline{\|\mathbf{h}\|})$. As can be shown by the method of stochastic differentials, this approximation is exact in the asymptotic limit $N \to \infty$. Since the proof is rather long and of technical interest only, we refrain from presenting it here. Using results derived in Arnold and Beyer (2001), the expected values of $\overline{x}$ and $\overline{\|\mathbf{h}\|}$ in the asymptotic limit case are

$$\overline{x} \simeq \sigma\frac{c_{\mu/\mu,\lambda}\sigma^*}{\sqrt{\sigma^{*2}+\sigma_\varepsilon^{*2}}} \quad \text{and} \quad \overline{\|\mathbf{h}\|} \simeq \sigma\sqrt{\frac{N}{\mu}}. \tag{5}$$

Thus, we find

$$\overline{\alpha} \simeq \mathrm{arc}\cot\left(\sqrt{\frac{\mu}{N}}\frac{c_{\mu/\mu,\lambda}\sigma^*}{\sqrt{\sigma^{*2}+\sigma_\varepsilon^{*2}}}\right), \tag{6}$$

i.e., for $\sigma_\varepsilon^* > 0$ the expected α-angle depends on the (normalized) mutation strength. Even though Eq. (6) is exact for the asymptotic limit ($N \to \infty$, $\sigma^* < \infty$, $\sigma_\varepsilon^* < \infty$) only, it also can serve as an approximation formula provided that σ^* and σ_ε^* are sufficiently small: Fig. 2 shows an example of the $\overline{\alpha}(\sigma^*)$ dependency. Since we have seen

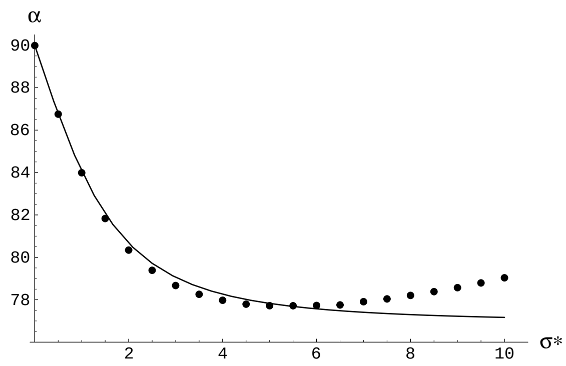

Figure 2: On the dependency of the $\overline{\alpha}$-angle on the normalized mutation strength σ^*. In the simulation, a $(4/4_I, 15)$-ES on the $N = 100$-dimensional quadratic sphere $Q = \mathbf{y}^2$ with $\sigma^*_\varepsilon = 2$ has been investigated. The data points (dots) are obtained by averaging over 40,000 independent one-generation experiments. The curve is a plot of Eq. (6) ($c_{4/4,15} \approx 1.1616$).

that the expected α-angle depends on the (normalized) mutation strength, we can now conclude that the validity of the perpendicularity condition cannot be based on the $\alpha = \text{const.}$ assumption.

2.3 Reconsidering the Perpendicularity Condition

So far we have shown that the orthogonality condition cannot be based on an $\alpha = \text{const.}$ assumption. Yet it might be possible that a (local) $\mathbf{r}^{\mathrm{T}}\mathbf{z}_{\text{opt}} = 0$ condition could guarantee performance optimality. In order to check this, the scalar product $\mathbf{r}^{\mathrm{T}}\mathbf{z}$ must be considered explicitly. Using information from Fig. 1 one finds $\mathbf{z} = -R\mathbf{e}_R + \mathbf{r}$ and therefore $\mathbf{r}^{\mathrm{T}}\mathbf{z} = -R\mathbf{r}^{\mathrm{T}}\mathbf{e}_R + ||\mathbf{r}||^2$. With $||\mathbf{r}||^2 = (R-x)^2 + ||\mathbf{h}||^2$, we obtain

$$\mathbf{r}^{\mathrm{T}}\mathbf{z} = -Rx + x^2 + ||\mathbf{h}||^2 = -Rx + ||\mathbf{z}||^2. \tag{7}$$

Taking the definition of the scalar product into account and neglecting deviations from the expected values, one gets

$$\cos\overline{\beta} \simeq \frac{\overline{\mathbf{r}^{\mathrm{T}}\mathbf{z}}}{\overline{||\mathbf{r}||}\,\overline{||\mathbf{z}||}} = \frac{-R\overline{x} + \overline{x^2} + \overline{||\mathbf{h}||^2}}{\overline{||\mathbf{r}||}\,\overline{||\mathbf{z}||}}. \tag{8}$$

Considering the asymptotic behavior ($N \to \infty$) and suppressing $\mathcal{O}(1/N^2)$ terms, the expected β value can be expressed by means of (5) as (note $\overline{||\mathbf{r}||} \simeq R$, $\overline{x^2} \simeq \overline{x}^2$)

$$\cos\overline{\beta} \simeq \sigma^* \sqrt{\frac{\mu}{N}} \left[\frac{1}{\mu} - \frac{c_{\mu/\mu,\lambda}}{\sqrt{\sigma^{*2} + \sigma_\varepsilon^{*2}}} \right]. \tag{9}$$

Non-trivial perpendicularity is obtained for vanishing brackets in (9). The mutation strength $\check{\sigma}^*$ at which this appears is

$$\check{\sigma}^* = \sqrt{\mu^2 c^2_{\mu/\mu,\lambda} - \sigma_\varepsilon^{*2}}. \tag{10}$$

In the noise-free case, $\sigma_\varepsilon^* = 0$, $\check{\sigma}^* = \mu c_{\mu/\mu,\lambda} = \hat{\sigma}^*$ is indeed fulfilled (cf. Eq. (4)), i.e., the perpendicularity condition guarantees performance optimality asymptotically. However, for $\sigma_\varepsilon^* > 0$, (10) cannot maximize Eq. (2). Even more critical, (10) puts a constraint on maximal noise level above which the CSA-ES cannot work as an optimization algorithm:

$$\text{CSA evolution criterion:} \quad \sigma_\varepsilon^* < \mu c_{\mu/\mu,\lambda}. \tag{11}$$

This is a more restrictive criterion than the $\sigma_\varepsilon^* < 2\mu c_{\mu/\mu,\lambda}$ ES *evolution criterion* obtained from the progress rate formula (2). That is, although the ES algorithm (L1 – L3) in Eq. (1) can exhibit positive progress for $\mu c_{\mu/\mu,\lambda} \le \sigma_\varepsilon^* < 2\mu c_{\mu/\mu,\lambda}$ (provided that the mutation strength is chosen appropriately), $\overline{\beta} < 90°$ holds independently no matter how $\sigma^* > 0$ is chosen: The CSA (L4, L5) assumes mutation steps too long and reduces σ^* continuously to zero. This is what has been observed in Arnold and Beyer (2000a).

2.4 Finite N-Size Effects and Simulations

The results presented so far are based on asymptotically exact expressions. As approximations for $N < \infty$ their predictive power might be of limited value. It would be useful to have some N-dependent approximations that account for finite N-size effects. Recent results in N-dependent progress rate analysis can be used for this purpose. Using results from Arnold and Beyer (2000b) instead of (5)

$$\overline{x} \simeq \frac{R}{N} \frac{c_{\mu/\mu,\lambda}\sigma^{*2}}{\sqrt{\sigma^{*2} + \sigma_\varepsilon^{*2} + \sigma^{*4}/2N}} \tag{12}$$

and

$$\overline{\|\mathbf{h}\|^2} \simeq \frac{R^2}{N} \frac{\sigma^{*2}}{\mu} \left[1 - \frac{1}{N} \frac{c_{\mu/\mu,\lambda}\sigma^{*2}}{\sqrt{\sigma^{*2} + \sigma_\varepsilon^{*2} + \sigma^{*4}/2N}} \right], \tag{13}$$

the progress rate φ^* can be expressed by

$$\varphi^* \simeq N \left[1 - \sqrt{\left(1 - \frac{\overline{x}}{R}\right)^2 + \frac{\overline{\|\mathbf{h}\|^2}}{R^2}} \right]. \tag{14}$$

Similarly, one has $\overline{\|\mathbf{r}\|} \simeq \sqrt{(R - \overline{x})^2 + \overline{\|\mathbf{h}\|^2}}$ and furthermore $\overline{\|\mathbf{z}\|} \simeq \sqrt{\overline{x}^2 + \overline{\|\mathbf{h}\|^2}}$ (neglecting fluctuations around the mean values). These estimations can be inserted in $\cos\overline{\beta}$, Eq. (8). The resulting formula is rather long, but it predicts the (static) behavior of the CSA surprisingly well ($N \ge 40$). As an example the $(4/4_I, 15)$-ES on an $N = 100$-dimensional quadratic sphere is presented in Fig. 3 ($c_{4/4,15} \approx 1.1616$).

The N-dependent $\overline{\beta}$ and φ^* formulae can be used to investigate the optimality condition for $\overline{\beta}$ explicitly. By determining $\hat{\sigma}^*$, Eq. (4), depending on σ_ε^*, using (14) with (12) and (13), numerically and inserting this $\sigma^* = \hat{\sigma}^*$ in the $\overline{\beta}$ formula (8), one obtains the optimal $\overline{\beta}$-angle for which the progress rate would be maximal given a noise level σ_ε^*. For the $(4/4_I, 15)$-ES, $N = 100$, this $\beta_{\text{opt}}(\sigma_\varepsilon^*)$-dependence is displayed in Fig. 4. While the abrupt bending of the β_{opt} curve at a specific (high) σ_ε^*-value corresponds to the violation of the CSA evolution criterion (11) in the asymptotic case, the behavior at $\sigma_\varepsilon^* = 0$ is a finite N effect. Unlike the prediction of the asymptotic theory, governed by (9), the N-dependent β_{opt} curve shows that the perpendicularity condition is *not* optimal for $\sigma_\varepsilon^* = 0$. Instead, there is a strategy- and N-specific σ_ε^*-value for which perpendicularity implies optimality.

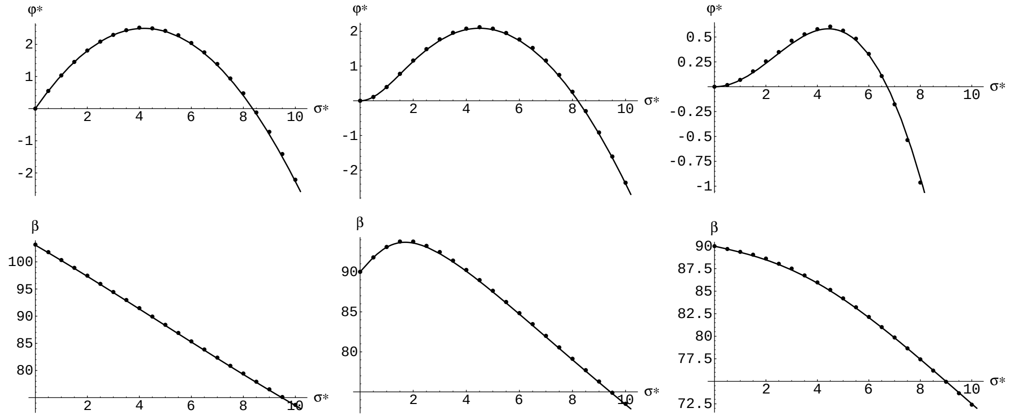

Figure 3: Normalized progress rates φ^* and $\overline{\beta}$-angles (measured in degrees) for a $(4/4_I, 15)$-ES on the $N = 100$-dimensional quadratic sphere $Q = \mathbf{y}^2$. The curves are the predictions obtained from Eqs. (8) and (14), respectively, using the N-dependent approximations (12) and (13). The dots represent simulation results each of which was obtained by averaging 40,000 independent one-generation experiments. The standard deviation of these mean values is smaller than the size of the radius of the dots. The left-hand graphs are obtained for noise-free fitness evaluations, i.e., $\sigma^*_\varepsilon = 0$, the graphs in the middle are for (normalized) fitness noise $\sigma^*_\varepsilon = 2$, and the right-hand graphs are for the case $\sigma^*_\varepsilon = 6$. As one can see, the perpendicularity condition ($\overline{\beta} = 90°$) does not exactly correspond to the performance optimum (maximum of φ^*). While there is only a small performance degradation for $\sigma^*_\varepsilon = 0$ and $\sigma^*_\varepsilon = 2$, the CSA must necessarily fail for the $\sigma^*_\varepsilon = 6$ case (i.e., $\sigma^* \to 0$).

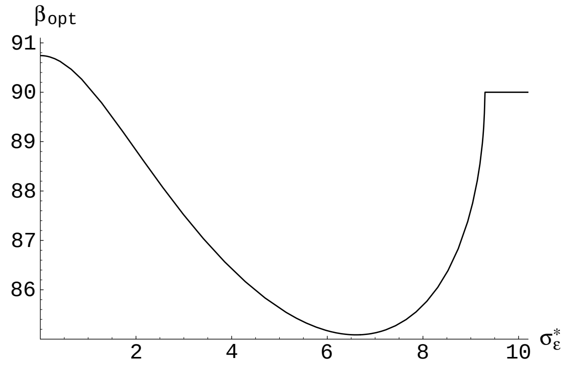

Figure 4: The expected *optimal* β (measured in degree) depending on the normalized noise strength σ^*_ε for the $(4/4_I, 15)$-ES, $N = 100$.

3 Conclusions and Outlook

In this paper we have shown that the philosophical basis of the fundamental adaptation principle the CMA and CSA strategies are based on, i.e., the perpendicularity condition, does not guarantee (static) performance optimality on the sphere model as claimed in Hansen and Ostermeier (1996). Especially in the noisy case, the perpendicularity condition may lead to a wrong adaptation behavior resulting in a premature convergence. There is experimental evidence that such behavior can occur in real CMA/CSA-ES implementations (see, e.g., Arnold & Beyer, 2000a). Users should be aware of this fact when applying such strategies.

In spite of the empirical evidence, the work presented is restricted in several aspects. First, it is a static analysis based on considerations on a static sphere model. While this may be regarded as a flaw – and it is a flaw – the investigations by Ostermeier and Hansen (see especially Hansen, 1998) use exactly the same model considerations. That is, dynamical aspects were not considered in this model neither for the σ^*-evolution nor for the search space dynamics, i.e., the r-evolution. Both aspects remain to be investigated. Second, from a much broader perspective, considering performance on quadratic models, as has been done here, might be too "far away" from performance aspects in real world optimization problems. While this holds necessarily for all theoretical performance investigations, the analysis presented makes a first step toward the incorporation of irregularities real world problems are faced with, by allowing for noisy fitness data. The usefulness of such a model becomes more clear when considering highly rugged fitness landscapes as a result of a noise process frozen in time. Using such a model of real world behavior might be a starting point for further theoretical and empirical investigations.

4 Acknowledgements

The authors want to express their thanks to Nikolaus Hansen and Ingo Rechenberg for discussions concerning the philosophy of self-adaptation and the ideas of cumulative path-length adaptation. This work was supported by the Deutsche Forschungsgemeinschaft (DFG), grant Be 1578/6-1. The first author is DFG Heisenberg Fellow under grant Be 1578/4-2.

References

Arnold, D. V., & Beyer, H.-G. (2000a). Efficiency and Mutation Strength Adaptation of the $(\mu/\mu_I, \lambda)$-ES in a Noisy Environment. In M. Schoenauer (Ed.), *Parallel Problem Solving from Nature, 6* (pp. 39–48). Heidelberg: Springer.

Arnold, D. V., & Beyer, H.-G. (2000b). Performance Analysis of Evolution Strategies with Multi-Recombination in High-Dimensional $\mathbb{R}^N$-Search Spaces Disturbed by Noise. *Theoretical Computer Science*. (accepted for publication)

Arnold, D. V., & Beyer, H.-G. (2001). Local Performace of the $(\mu/\mu_I, \lambda)$-ES in a Noisy Environment. In W. Martin & W. Spears (Eds.), *Foundations of Genetic Algorithms, 6* (pp. 127–141). San Francisco, CA: Morgan Kaufmann.

Bäck, T., Hammel, U., & Schwefel, H.-P. (1997). Evolutionary computation: comments on the history and current state. *IEEE Transactions on Evolutionary Computation*, *1*(1), 3–17.

Bäck, T., & Schwefel, H.-P. (1993). An Overview of Evolutionary Algorithms for Parameter Optimization. *Evolutionary Computation*, *1*(1), 1–23.

Beyer, H.-G. (1996). Toward a Theory of Evolution Strategies: Self-Adaptation. *Evolutionary Computation*, *3*(3), 311–347.

Beyer, H.-G. (1997). An Alternative Explanation for the Manner in which Genetic Algorithms Operate. *BioSystems*, *41*, 1–15.

Beyer, H.-G. (2001). *The Theory of Evolution Strategies.* Heidelberg: Springer. (ISBN 3-540-67297-4)

Grünz, L., & Beyer, H.-G. (1999). Some Observations on the Interaction of Recombination and Self-Adaptation in Evolution Strategies. In P. Angeline (Ed.), *Proceedings of the CEC'99 Conference* (pp. 639–645). Piscataway, NJ: IEEE.

Hansen, N. (1998). *Verallgemeinerte individuelle Schrittweitenregelung in der Evolutionsstrategie.* Doctoral thesis, Technical University of Berlin, Berlin.

Hansen, N., & Ostermeier, A. (1996). Adapting Arbitrary Normal Mutation Distributions in Evolution Strategies: The Covariance Matrix Adaptation. In *Proceedings of 1996 IEEE Int'l Conf. on Evolutionary Computation (ICEC '96)* (pp. 312–317). IEEE Press, NY.

Hansen, N., & Ostermeier, A. (1997). Convergence Properties of Evolution Strategies with the Derandomized Covariance Matrix Adaptation: The $(\mu/\mu_I, \lambda)$-CMA-ES. In H.-J. Zimmermann (Ed.), *5th European Congress on Intelligent Techniques and Soft Computing (EUFIT'97)* (pp. 650–654). Aachen, Germany: Verlag Mainz.

Hansen, N., & Ostermeier, A. (2001). Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, *9*(2), 159–195.

Ostermeier, A., Gawelczyk, A., & Hansen, N. (1994). Step-Size Adaptation Based on Non-Local Use of Selection Information. In Y. Davidor, R. Männer, & H.-P. Schwefel (Eds.), *Parallel Problem Solving from Nature, 3* (pp. 189–198). Heidelberg: Springer-Verlag.

Ostermeier, A., Gawelczyk, A., & Hansen, N. (1995). A Derandomized Approach to Self-Adaptation of Evolution Strategies. *Evolutionary Computation*, *2*(4), 369–380.

Rechenberg, I. (1973). *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution.* Stuttgart: Frommann-Holzboog Verlag.

Schwefel, H.-P. (1981). *Numerical optimization of computer models.* Chichester: Wiley.