# A Derandomized Approach to Self Adaptation of Evolution Strategies

Andreas Ostermeier, Andreas Gawelczyk, Nikolaus Hansen

TR-93-003
July 1993

**Technische Universität Berlin**

**Fachgebiet Bionik und Evolutionstechnik**
**Ackerstraße 71-76 (ACK1)**
**D-13355 Berlin**

## Abstract

In this paper two variants of Evolution-Strategies (ES) with an improved concept of self-adaptation are presented.

The first one enables a reliable adaptation of individual step-sizes. This means that the ES is able to treat problems which are badly scaled with respect to the sensitivity of the parameters much better than before.

The second version also allows correlated variations of the parameters. In addition, the adaptation scheme is extended by using information from more than just the one last generation.

The present paper is written to give the reader the opportunity to reimplement the new algorithms developed. Although the basic ideas will be discussed, it is not intended to give a deep insight to every detail of the algorithms. Therefore the reader will find motivation and discussion of the algorithms sometimes quite sketchy.

## Mutative Step-size Control

The adaptation scheme introduced here is based on the well-known concept of mutative step-size control [1, 2]. Mutative step-size control generally works very well for the adaptation of a general step-size. Applied to the adaptation of individual step-sizes, it fails (in general). Obviously this happens because the selection process does not (or with poor probability) select offspring with a good set of step-sizes. Two very important aspects of the possible reasons are the following ones:

First, the standard algorithm of mutative step size control does not ensure that the selection of large or small parameter variation leads to a corresponding change of step-size. For instance, a relatively large parameter variation may occur in spite of a small step-size because of a large random number **realization**. In such a case, the step-size will be reduced although a large mutation was selected. In general, the same parameter variations (mutations) can be generated by totally different sets of strategy parameters (step-sizes). This has to be avoided.

Second, in the standard adaptation algorithm, the step-size variation between competing offspring in one generation is identical to the step-size variation from one generation to another. This leads to a conflict because, on the one hand, offspring have to be produced with clearly different step-sizes – otherwise they are of no selection relevance. On the other hand, step-size adaptation rates in the generation sequence have to be much smaller to reduce random fluctuations.

## Derandomized Mutative Step-size Control

The modified mutative step-size control presented here avoids the two difficulties discussed above. The possibility of producing the same mutations with different strategy parameters (step-sizes) is prevented by using only the values $\pm 1$ as random numbers instead of $(0\,,\,1)$- normally distributed ones. Nevertheless, to produce normally distributed (isotropic) mutations, $(0\,,\,1)^{+}$- normally distributed[1] step-size changing factors $\xi$ have to be chosen instead of $\alpha, 1/\alpha$. This allows the absolute value of a realized mutation to equal the corresponding strategy parameter step-size. Thus, it is guaranteed that the selection of large (small) mutations leads to a corresponding adaptation of the step-sizes.

In order to prevent the conflict "step-size variation (in one generation) versus step-size adaptation (in the generation sequence)", the step-size adaptation rate in this algorithm

---

1 The $^{+}$ indicates the positive part of the normal distribution. The geometric mean value of this distribution is less than one. Without selection pressure, this would result in constantly decreasing step-sizes. To prevent this, the algorithm adapts the step-sizes by the factor $(\xi + b)$. For $b = 0.35$, the geometric mean value of this term is approximately one.

is not determined by the step-size changing factor $\xi$ but by $\xi^{\beta}$. Its effect is clearly that a step-size adaptation by a factor $\xi$ is not realized in **one** but in at least **1/$\beta$** generations. Thus values $0 < \beta < 1$ reduce the adaptation rate.

**So derandomization occurs three times:**

First, the absolute value of a mutation depends only on the random number $\xi$ (not $\xi$ and Z as before) and determines the new step-size immediately.

Second, values of $\xi$ significantly different from 1 can be chosen. This ensures that the different step-sizes of the offspring are of selection relevance. Otherwise step-sizes would be selected by chance and perform nearly random walks.

Third, the step-size adaptation rate ($\xi^{\beta}$) can be adjusted to values (by $\beta << 1$) that reduce disturbing random fluctuations of the step-sizes.

## A derandomized (1 , $\lambda$)- ES algorithm with individual step-sizes

(all multiplications, divisions and powers of vectors refer to components)

**Creation of $\lambda$ offspring:**

$$\vec{X}^{g}_{N_k} = \vec{X}^{g}_{E} + (\xi_k \delta^g) \cdot (\vec{\xi}_{scal_k} \vec{\delta}^{g}_{scal}) \cdot \vec{Z}_k \qquad (k = 1 \ldots \lambda)$$

**Selection / Adaptation:**

$$\vec{X}^{g+1}_{E} = \vec{X}^{g}_{N_{sel}}$$

$$\delta^{g+1} = (\xi_{sel})^{\beta} \cdot \delta^g$$

$$\vec{\delta}^{g+1}_{scal} = (\vec{\xi}_{scal_{sel}} + b)^{\beta_{scal}} \cdot \vec{\delta}^{g}_{scal}$$

**Symbols used:**

| | | |
|---|---|---|
| $\vec{X}^{g}_{E/N}$ | parameter vector of generation g (**E**: parent / **N**: offspring) | |
| $\delta^g$ | general step-size of generation g | |
| $\vec{\delta}^{g}_{scal}$ | individual step-sizes of generation g | |
| $\xi$ = | $\alpha$ , $1/\alpha$ | with equal probability $(\alpha = 1.4)$ |

$\vec{\xi}_{scal}$ = $(z_1, \ldots, z_n)$ with $z_i$ $(0, 1)^+$-normally distributed

$\vec{Z}$ = $(\pm 1, \ldots, \pm 1)$ +/- with equal probability

*sel* marks the quantities of the **sel**ected offspring of generation g

$\beta = \sqrt{1/n}$ (adaptation rates should decrease with increasing dimension n of the problem)

$\beta_{scal} = 1/n$

$b = 0.35$

## Correlated Mutations

The algorithm presented above works very effectively for problems that resemble axis-parallel ellipsoids or ridges. If the same problems are rotated, an effective adaptation is in general not possible. To deal effectively with such problems, the strategy has to allow (linearly) correlated variations of the parameters.

In biology, features such as pleiotropy and polygeny are expected to produce correlated variations. Thus, evolution has developed transformations from genotypic to phenotypic level that produce advantageously correlated variations on the phenotypic level. Mutations on the genotypic level are, of course, not thought to be correlated.

The algorithm proposed here tries to copy the effect of the biological example: Correlated variations on the phenotypic level. On the one hand, ES works on the object parameters, the changes of which are interpreted as mutations. So far, the object parameters are on the genotypic level. On the other hand, the object parameters have to be varied in a correlated manner. So far, they are on the phenotypic level. One way to overcome this problem might be to introduce an additional coordinate system in which ES works with uncorrelated mutations. In a second step, the object parameters would be changed by the adaptively transformed (rotated) mutations. Now the object parameters are on phenotypic level.

In order to prevent the need for an additional level between ES and object parameters, we would like to introduce correlated mutations. The following algorithm realizes mutation ellipsoids with a long axis of rotational symmetry that can be orientated arbitrarily. The same variations could be realized by rotating an axis-parallel ellipsoid with (n-1) similar short axes. But in contrast to this, the proposed algorithm allows a very simple and effective adaptation scheme. It just turns the long axis of the mutation ellipsoid in the direction of the selected offspring. A strategy parameter ($D_L$) determines to what extent the selected offspring influences the new direction.

## Adaptation by Accumulated Information

Step-size adaptation as described above refers to the information from the last generation. If the last selected step was larger than its expected mean value, the corresponding step-size will be increased in the next generation (and vice versa). This concept can be extended by considering not only the last step but the resulting variation of the last *m* steps. The adaptation scheme remains the same as before: The step-size is increased, if the resulting variation was larger than expected. All that must be known is the expected mean value of the resulting variation.

Instead of evaluating the last *m* generations, the implemented algorithm computes weighted averages of the entire generation sequence (see below). Information from previous generations declines exponentially with $(1\text{-}c)^{\Delta g}$. The number of considered generations *m* can be approximated by the lifespan 1/c of information. This algorithm is not only easy to implement, but also causes the influence of generations to decline continuously. This seems to be more sensible than a constant contribution for *m* generations which then vanishes suddenly.

It should be noted that the algorithm averages realized and selected **steps** and not only the **step-sizes**. By using the additional information of relative step-directions, adaptation becomes faster and reduces random fluctuations at the same time. For example, although only small mutation steps may be selected in every generation, a step-size can be increased if the selected steps are anisotropically distributed. On the other hand, a step-size can be decreased in spite of large selected steps if they are merely zig-zagging.

The factor $(c / (2\text{-}c))^{1/2}$ normalizes the random distributions. It results from the geometric series of the standard deviations:

$$\lim_{n\to\infty} \sqrt{c^2 + (c\cdot(1-c)^1)^2 + (c\cdot(1-c)^2)^2 + \ \ldots\ + (c\cdot(1-c)^n)^2} \quad = \quad \sqrt{\frac{c}{2-c}}$$

## A derandomized (1 , λ)- ES algorithm with individual step-sizes and correlated mutations

(all multiplications, divisions and powers of vectors component-wise)

**Creation of λ offspring:**

$$\vec{X}^{g}_{N_k} = \vec{X}^{g}_{E} \ + \ \vec{\delta}^{g}_{scal}\cdot(\xi_k \delta^{g}_{iso} \vec{Z}_k + \psi_k \delta^{g}_{ani} \vec{A}^{g}_{o}) \qquad (k = 1..\lambda)$$

**Selection / Adaptation:**

$$\vec{X}^{g+1}_{E} \ = \ \vec{X}^{g}_{N_{sel}}$$

**Calculation of the weighted averages:**

$$\vec{\overline{Z}}^{g} = (1-c_{iso})\,\vec{\overline{Z}}^{g-1} + c_{iso}\,\xi_{sel}\,\vec{Z}_{sel} \qquad \vec{\overline{Z}}^{0} = \vec{0}$$

$$\overline{\psi}^{g} = (1-c_{ani})\,\overline{\psi}^{g-1} + c_{ani}\,\psi_{sel} \qquad \overline{\psi}^{0} = 0$$

$$\vec{\overline{T}}^{g} = (1-c_{scal})\,\vec{\overline{T}}^{g-1} + \frac{c_{scal}\,(\vec{X}^{g}_{N_{sel}} - \vec{X}^{g}_{E})}{\vec{\delta}^{g}_{scal}\cdot\sqrt{(\xi_{sel}\cdot\delta^{g}_{iso})^{2} + (\delta^{g}_{ani}\cdot\vec{A}^{g}_{o})^{2}}} \qquad \vec{\overline{T}}^{0} = \vec{0}$$

(the **T**otal Mutation ($X_N$ - $X_E$) is divided by its standard deviation!)

**Step-size of the normally distributed (ISOtropic) mutations:**

$$\delta^{g+1}_{iso} = \delta^{g}_{iso}\cdot\left(\exp\left(\frac{\left|\vec{\overline{Z}}^{g}\right|}{\sqrt{n}\cdot\sqrt{\frac{c_{iso}}{2-c_{iso}}}} - 1 + \frac{1}{5n}\right)\right)^{\beta_{iso}}$$

(magnitude of vector; **1/5n** is a correction for small dimensions n.)

**Step-size of the correlated (ANIsotropic) mutations:**

$$\delta^{g+1}_{ani} = \delta^{g}_{ani}\cdot\left(\frac{\left|\overline{\psi}^{g}\right|}{\sqrt{\frac{c_{ani}}{2-c_{ani}}}} + b\right)^{\beta_{ani}}$$

**Individual step-sizes (SCALing factors):**

$$\vec{\delta}^{g+1}_{scal} = \vec{\delta}^{g}_{scal}\cdot\left(\frac{\left|\vec{\overline{T}}^{g}\right|}{\sqrt{\frac{c_{scal}}{2-c_{scal}}}} + b\right)^{\beta_{scal}}$$

(absolute value of components)

**Orientation of the anisotropic mutations:**

$$\vec{A}^{g+1} = \vec{A}^{g}_{o} + \left((\psi_{sel}-1)\cdot\vec{A}^{g}_{o} + \frac{\xi_{sel}\,\delta^{g}_{iso}\,\vec{Z}_{sel}}{\delta^{g}_{ani}}\right)\cdot D_{L}$$

$$\vec{A}^{g+1}_{o} = \vec{A}^{g+1}\cdot\frac{\sqrt{n}}{\left|\vec{A}^{g+1}\right|}$$

(magnitude of vector)

**Symbols used:**

| | |
|---|---|
| $\vec{X}^g_{E/N}$ | parameter vector of generation g (**E**: parent / **N**: offspring) |
| $\delta^g_{iso}$ | step-size of **iso**tropic mutations of generation g |
| $\delta^g_{ani}$ | step-size of **ani**sotropic (correlated) mutations of generation g |
| $\vec{\delta}^g_{scal}$ | **scal**ing factors of generation g ( $\vec{\delta}^0_{scal} = (1, \ldots, 1)$ ) (individual step-sizes) |
| $\vec{Z}$ | random vector with components (0 , 1)- normally distributed[2] |
| $\xi$ | (1 , 0.1)- normally distributed. $\xi$ is chosen like this, because the weighted averages approximate a normal distribution and not a logarithmic one (central limit theorem). |
| $\psi$ | (0 , 1)- normally distributed |
| $\vec{A}^g$ | orientation of long axis of mutation ellipse in genearation g $\vec{A}^g_0$ normed to $\sqrt{n}$ , $(\vec{A}^0_0 = \vec{0})$ |

$c_{iso} = \sqrt{1/n}$

$c_{scal} = \sqrt{1/n}$

$c_{ani} = \sqrt{1/n}$

The factors "**c**" determine how fast the contribution of former generations declines. The loss is about a factor **3** every **1/c** generations. $(\lim_{c\to 0}(1-c)^{1/c} = \frac{1}{e} \approx 0.37\,;\ (1-0.3)^{1/0.3} \approx 0.3\,)$ The given values of **c** are coarse estimations.

$\beta_{iso} = \sqrt{1/n}$

$\beta_{scal} = 1/n$

$\beta_{ani} = 0.5 \cdot \sqrt{1/n}$

$D_L = 0.5 \cdot \sqrt{1/n}$

Adaptation speed and precision depends on the exponents "β" and the direction learning rate "$D_L$". Sensible values are in the range (0 ; 1). Small values facilitate a precise but time-consuming adaptation and vice versa. The given values yield a good compromise. In the case of very hard problems (extremely poor success-probability) a reduction of $\beta_{scal}$ and / or $D_L$ might be necessary.

| | |
|---|---|
| *sel* | marks the quantities of the **sel**ected offspring of generation g |
| $b = 0.35$ | |

---

2 Here (0 , 1)- normally distributed random numbers are used again in order to simplify the algorithm. The adaptation of individual step-sizes works essentially in the same way as in the first algorithm because the step-size changing vector $\vec{\xi}$ is ommited.

## Simulations

Simulations on several optimization problems (brachistochrone, Steiner-net, neural-net) have shown that the presented algorithm with correlated mutations accelerates optimization considerably. The given values for the strategy parametes $c_{...}$ , $\beta_{...}$ and $D_L$ can be kept for all these functions.

The two following examples check the main features of the algorithms (scaling and correlations). Simulations have been carried out with $\lambda = \mathbf{10}$. So the number of function evaluations equals ten times the number of generations. The data given for ES-with scaling / scaling & correlation refer to the algorithms presented in this paper. The "simple" ES is the standard algorithm with mutative (general) step-size control.

Concerning the hyper-ellipsoids, the faster optimizations with the second algorithm do not result from the correlated mutations but from the cumulative adaptation scheme.

## Axis-parallel Hyper-Ellipsoids:

Objective function:

$$F_n(\vec{x}) \;=\; \sum_{i=1}^{n} (i \cdot x_i)^2 \qquad \Rightarrow \qquad \textit{minimum}$$

$$\vec{x}^0 \;=\; (1,\ \ldots\ ,1), \quad F_{stop} = 10^{-10}, \quad F_{10,30,100}(\vec{x}^0) = 385,\ 9455,\ 338350$$

**function evaluations to reach $F_{stop}$** (average of 20 runs)

| | **(1 , 10)- ES with ...** | | |
|---|---|---|---|
| | **scaling** | **scaling & correlations** | **simple** |
| n = 10 | 2.560 | 2.440 | 20.800 |
| n = 30 | 13.500 | 8.140 | 597.000 |
| n = 100 | 109.000 | 38.900 | $\approx 2.3 \cdot 10^7$ |

## Parabolic Ridge:

(directed in the main diagonal of the n-dimensional space)

Objective function:

$$F(\vec{x}) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i \;-\; \sum_{i=1}^{n} \left( \frac{1}{n} \sum_{k=1}^{n} (x_k - x_i) \right)^2 \qquad \Rightarrow \qquad \textit{maximize}$$

$$\vec{x}^0 \;=\; \vec{0}\ , \qquad F(\vec{x}^0) \;=\; 0\ , \qquad F_{stop} \;=\; 10^{10}$$

**function evaluations to reach $F_{stop}$** (average of 20 runs)

| | (1 , 10)- ES with ... | | |
|---|---|---|---|
| | **scaling** | **scaling & correlations** | **simple** |
| n = 10 | $\approx 1.4 \cdot 10^{12}$ | 1.850 | $\approx 9.3 \cdot 10^{11}$ |
| n = 30 | $\approx 5.1 \cdot 10^{12}$ | 4.700 | $\approx 3.4 \cdot 10^{12}$ |
| n = 100 | $\approx 1.7 \cdot 10^{13}$ | 16.000 | $\approx 1.1 \cdot 10^{13}$ |

## Conclusions

The efficiency of the adaptation scheme of mutative step-size control is reduced because of an interaction of the involved random numbers. The algorithms presented show that this weak point of mutative step-size control can be avoided by relatively small changes of the algorithm. The modifications may be summarized by the keyword "derandomization". Obviously it is of essential importance that the adaptation process is determined by the selected mutations of the offspring. The modified version with scaling and correlations makes the concept of mutative step-size control work very effectively and proves to be of universal applicability.

With the algorithms presented, the adaptation process causes only little computational expense, no additional function evaluations and proves to be very reliable. As a result, the optimization of functions such as the hyper-sphere - where the simple ES is adapted optimally - only slows down by a factor about two. This and the striking advantages in the case of difficult[3] problems predestine the proposed algorithms for standard use.

The concept of accumulated information enables a faster and less randomly disturbed adaptation at the same time. This process is very promising but relatively complex and needs more investigation, especially concerning the involved strategy parameters $c_{...}$ and $\beta_{...}$ .

## References


[1] I. Rechenberg, *Evolutionsstrategien*, in B. Schneider u. U. Ranft (Hrsg.), Simulationsmethoden in der Medizin und Biologie, Springer, Berlin, 1978.

[2] H.-P. Schwefel, *Numerical Optimization of Computer Models*, Wiley & Sons, Chichester, 1981.


---

3 "Difficult" refers to ridge-like or badly scaled problems but not to multimodal optimization.