Step-Size Adaptation Based on Non-Local Use of Selection Information

Andreas Ostermeier, Andreas Gawelczyk, Nikolaus Hansen
Technische Universität Berlin, Fachgebiet Bionik und Evolutionstechnik
Ackerstraße 71-76, D-13355 Berlin
E-mail: {ostermeier,gawelczyk,hansen}@fb10.tu-berlin.d400.de
Phone: 030/31472666

Abstract. The performance of Evolution Strategies (ESs) depends on a suitable choice of internal strategy control parameters. Apart from a fixed setting, ESs facilitate an adjustment of such parameters within a self-adaptation process. For step-size control in particular, various adaptation concepts were evolved early in the development of ESs. These algorithms mostly work very efficiently as long as the relative sensitivities of the parameters to be optimized are known. If this scaling is not known, the strategy has to adapt individual step-sizes for the parameters. In general, the number of necessary step-sizes (variances) equals the dimension of the problem. In this case, step-size adaptation proves to be difficult.

The algorithm presented in this paper is a development based on the derandomized scheme of mutative step-size control. The new adaptation concept uses information accumulated from the preceding generations with an exponential fading of old information instead of using information from the current generation only. Compared to the conventional adaptation scheme, this enables a less locally determined step-size control and allows a much faster adaptation of individual step-sizes without increasing disturbing random effects and without additional evaluations of the fitness function. The adaptation of the general step-size can be improved as well.

Keywords *evolution strategy, adaptation, self-adaptation, mutative step-size control, step-size, individual step-size, scaling*

Introduction: Step-size Adaptation in ESs

In biology, mutation rates are of essential importance for evolutionary progress. In the case of real-valued continuous parameter optimization with ESs, the biological mutation rate can be interpreted as the standard deviation of mutation steps in the parameter space.

In ESs, there are two common ways of realizing a step-size adaptation. One is Rechenberg's 1/5-success-rule (Rechenberg 1973). This algorithm works satisfying in most cases but depends on the applicability of an external model of parameter space topology and is only able to adapt one general step-size but no individual step-sizes.

The other method is the mutative step-size control proposed by Rechenberg (1973, 1978) and Schwefel (1977, 1981). This adaptation scheme does not depend on an external model and in principle facilitates the adaptation of individual step-sizes. Here, the strategy parameters (step-sizes) are part of the parameter sets of the individuals and affected by mutation and selection.

Mutative step-size control generally works very well on the adaptation of a general step-size. A corresponding adaptation of individual step-sizes is not possible within simple ESs with small populations, as Schwefel (1987) pointed out. Schwefel favors the use of more complex ESs with larger populations. The problem is, that to enable a reliable individual

step-size adaptation the resulting population sizes have to be much larger[1] than necessary concerning the object parameter optimization.

The problem of individual step-size adaptation can be explained from a general point of view by interpreting step-size adaptation as a problem of disturbed optimization (Rechenberg 1994). "Disturbed" means that the fitness value is not exactly measurable.

Derandomized mutative step-size control

Derandomized mutative step-size control (cf. Ostermeier 1994) enables a reliable adaptation of individual step-sizes even in small populations. Basically, the selection of large or small[2] mutations in every generation results directly in a corresponding increase/decrease of the step-sizes. The second, more important difference to conventional mutative step-size control is that the step-size variations passed from one generation to the next are much smaller than the variations within one generation. This reduces the adaptation rate per generation without reducing the step-size variations within the populations. Therefore the information required for a certain step-size adaptation is not gathered in one generation of a large population but in the generation sequence of a smaller population.

Derandomized mutative step-size control using accumulated information

According to the previous section, a sensible adaptation of step-sizes for small populations is only possible in a generation sequence. The adaptation scheme proposed here makes use of this fact. It does not analyze the sizes of the mutations of the last generation only, but the sizes of the variations resulting from adding up the mutations selected in the preceding generations.

Apart from some averaging effects, this method would make no fundamental difference, if the selected mutations in successive generations are uncorrelated. In fact, successively selected mutations are correlated in general: In the case of step-sizes being too large, selected mutations tend to compensate preceding mutations. In effect, the selected mutations are correlated antiparallel in the generation sequence. Correspondingly, too small step-sizes cause parallel correlated mutations.

A parallel correlation of successive mutation steps increases the absolute value of the resulting sum and vice versa. The following algorithm utilizes these correlations - not simply the absolute values of the single mutation steps - by adding up successive mutations. Only the absolute values of the accumulated mutations are evaluated for step-size adaptation.

The adaptation scheme of the individual step-sizes remains formally the same as in the algorithm of derandomized step-size adaptation (Ostermeier 1994). Only the absolute values of the selected mutations have to be replaced by the absolute values of the accumulated selected mutations.

---

1 Our investigations suggest population sizes about 10*n (n: dimension of the problem). In smaller populations, individual step-sizes that have become small due to stochastic fluctuations perform almost random walks. This leads to long stagnation periods of the optimization if only individual step-sizes but no inclination angles (correlated mutations) are adapted. We interpret the reliable convergence of Schwefel's strategy variant with correlated mutations only partly as a result of a sensible adaptation process. Its uncritical behaviour mainly results from rotating the mutation ellipsoids almost randomly through parameter space. This ensures the variation of all parameters in spite of some individual step-sizes being arbitrarily small. The inclination angles are uncritical strategy parameters because of their cyclical characteristic: they cannot drift away.

2 "Large" or "small" refers to the mean variation of the underlying random distribution.

The weighting of the last generation and the lifespan of the information of preceding generations respectively is determined by the newly introduced constant $\mathbf{c} \in (\mathbf{0,1}]$. The factor $(c/(2\text{-}c))^{1/2}$ normalizes the mean variations of the resulting distributions to one (when no selection takes place). It results from the geometric series of the mean variations of the added mutations:

$$\lim_{m\to\infty} \sqrt{c^2 + \left(c\cdot(1-c)^1\right)^2 + \left(c\cdot(1-c)^2\right)^2 + \ldots + \left(c\cdot(1-c)^m\right)^2} \;=\; \sqrt{\frac{c}{2-c}}$$

The adaptation scheme of the general step-size uses the convergence of the χ-distribution: $|\vec{z}| = \sqrt{\sum z_i^2} \xrightarrow{n\to\infty} N(\sqrt{n},0.5)$.

(1 , λ)-ES algorithm with derandomized mutative step-size control using accumulated information

(all multiplications and powers of vectors refer to components)

Creation of λ offspring:

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

Selection:

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

Accumulation of selected mutations:

$$\vec{\bar{Z}}^{\,g} \;=\; (1-c)\,\vec{\bar{Z}}^{\,g-1} \;+\; c\,\vec{Z}_{sel} \qquad\qquad \vec{\bar{Z}}^{\,0} = \vec{0}$$

Adaptation of general and individual step-sizes:

$$\delta^{g+1} \;=\; \delta^g \cdot \left(\exp\left(\frac{|\vec{\bar{Z}}^{\,g}|}{\sqrt{n}\cdot\sqrt{\frac{c}{2-c}}} - 1 + \frac{1}{5n} \right) \right)^{\beta}$$

(absolute value of vector 1/(5n) is a correction for small dimensions n.)

$$\vec{\delta}^{\,g+1}_{scal} \;=\; \vec{\delta}^{\,g}_{scal} \cdot \left(\frac{|\vec{\bar{Z}}^{\,g}|}{\sqrt{\frac{c}{2-c}}} + 0.35 \;^{*} \right)^{\beta_{scal}}$$

(absolute value of components)

Symbols used:

n	number of parameters to be optimized (dimension of all vectors used)
$\vec{X}^{\,g}_{E/N}$	parameter vector of generation g (**E**: parent / **N**: offspring)

---

* $N(0\,,\,1)^{+}$-distributed step-sizes $|Z|$ would cause systematically decreasing step-sizes, because the geometric mean of this distribution is less than one. The geometric mean of $(|Z| + 0.35)$ approximately equals one.
It is also possible to transform $|Z|$ by an integral transformation into a logarithmic normal distribution. This solves the problem in an elegant but much more costly way and, corresponding to our tests, does not affect the performance of the algorithm.

δ^g general step-size of generation g

$\vec{\delta}^g_{scal}$ individual step-sizes of generation g $\quad \vec{\delta}^0_{scal} = (1, \ldots, 1)$

$\vec{Z} = (z_1, \ldots, z_n)$ with z_i (0 , 1)-normally distributed

sel index of **sel**ected offspring of generation g

$c = \sqrt{1/n}$ The factor "**c**" determines how fast the contribution of former generations declines. The loss is about a factor **3** every **1/c** generations. For $n \to \infty$ ($c \to 0$) ; $n = 10$ ($c \cong 0.3$) resp. holds:
$\left(\lim_{c\to 0}(1-c)^{1/c} = \frac{1}{e} \approx 0.37\,;\ (1-0.3)^{1/0.3} \approx 0.3\right)$

$\beta = \sqrt{1/n}$

$\beta_{scal} = 1/n$ Adaptation speed and precision depend on these two exponents. 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. See next section, figure 3. In the case of very difficult problems, a reduction of β_{scal} might be necessary.

Simulations

Tests of the described algorithm have been performed with $\lambda = 10$. Thus, the number of function evaluations equals ten times the number of generations. Simulations have been done on axis-parallel hyper-ellipsoids, Schwefel's problem, a generalized Rosenbrock's function, on a sum of different powers and on a Steiner-net.

In order to assess the performance of the derandomized step-size adaptation using accumulated information, results from the derandomized step-size adaptation (without using accumulated information) and from a (8,50)-ES according to Schwefel (1981) are also presented. Schwefel's strategy with discrete global recombination adapts n individual step-sizes and n(n-1)/2 inclination angles. This means that not only the sizes of the axes of the mutation ellipsoid vary, but it can also be rotated arbitrarily with respect to the coordinate system. A strategy variant that adapts only individual step-sizes and no inclination angles would correspond better to the algorithm presented here, but works - according to our experiments - very unsatisfactorily (cf. also Hoffmeister & Bäck 1991). The simulations with Schwefel's (8,50)-ES have been carried out with the *Evolution Machine* developed by Voigt, Born and Treptow (1991). Except for the adaptation using accumulated information, all other results are taken from Ostermeier (1994).

Axis-parallel Hyper-Ellipsoids:

Objective function: (*Not to be confused with the considerable different fct.* $\sum i \cdot x_i^2$)

$$F_n(\vec{x}) = \sum_{i=1}^{n} (i \cdot x_i)^2 \qquad \Rightarrow \qquad minimum \ (=0)$$

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

The simulation results (see figures 1 and 2) show that optimization speeds up considerably with adaptation of individual step-sizes. The feasible speed-up factor (10 ... 100 here) increases with the ratios of the ellipsoid-axes.

The optimization runs shown in figure 1 demonstrate that the step-size adaptation using accumulated information is able to adjust the correct set of individual step-sizes by which the problem is transformed into a hypersphere. After about 3000 function evaluations, the step-sizes are adapted correctly and the convergence rate is as high as with fixed individual step-sizes that are preadjusted correctly.

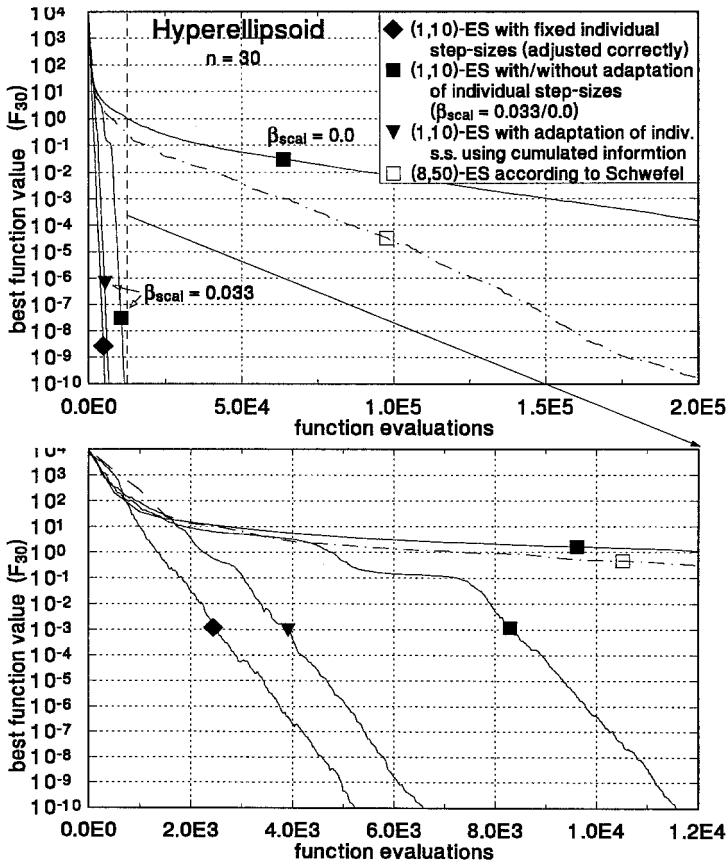

Figure 1
Convergence plots of optimization runs with the hyperellipsoid. The figure below is an enlarged detail of the first 12000 function evaluations of the same optimization runs shown above.

In order to assess the quality of step-size adaptation, an optimization run with the optimal set of (fixed) individual step-sizes is also shown.

In order to clarify the adaptation process, figure 2 - additionally to the fitness values - shows plots of the average ratio of the individual step-sizes to their correct values ($\pi(\delta_{scal})$). In figure 2a optimization runs with and without using accumulated information are compared. The plots of $\pi(\delta_{scal})$ show an acceleration of the step-size adaptation by about a factor three, using accumulated information. Additionally, the adaptation of the correct step-sizes is kept more precisely. The value of $\pi(\delta_{scal})$ stagnates at approximately 1.25 compared to 1.35 without using accumulated information. Figure 2b demonstrates the effect of varying β_{scal}. Reducing β_{scal} (= 0.01) facilitates a more precise but time consuming adaptation. Increasing β_{scal} (= 0.1) causes more stochastic fluctuations of the individual step-sizes.

To find out how to choose β_{scal}, the number of function evaluations needed to reach F_{stop} were measured for different values of β_{scal} (see figure 3). The minima result from the

conflict of fast versus precise adaptation. Large values of β_{scal} provoke such stochastic fluctuations that no sensible adaptation is possible. The acceleration of optimization using accumulated information is mainly caused by the faster individual step-size adaptation (cf. also figures 1 and 2). The improvement revealed for $\beta_{scal} = 0$ (no adaptation of individual step-sizes) is caused by the general step-size adaptation. Using accumulated information, the adaptation process acts less locally. In the case of varying curvatures of the quality surface (narrow valleys), this effect increases the general step-size and therefore accelerates the optimization.

According to figure 3, the optimal values of β_{scal} depend on the dimension n. Additional simulations have shown that this dependency does not change significantly with different ratios of the ellipsoid-axes. Thus, the value $\beta_{scal} = 1/n$ seems to be a good choice for a wide range of different problems. Compared to the simple derandomized step-size adaptation, the use of accumulated information does not change the range of sensible values for β_{scal}. All following simulations have been carried out with $\beta_{scal} = 1/n$, that is no special adjustment to the different test problems has been done.

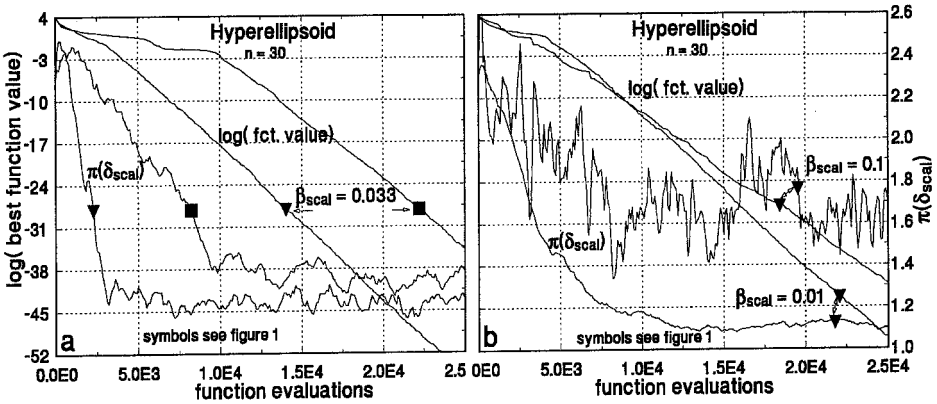

Figure 2 Convergence plots of optimization runs on the hyperellipsoid. The quantity $\pi(\delta_{scal})$ (the average deviation of the individual step-sizes) is defined as follows:

$\pi(\delta_{scal}) := \exp(\sigma(\ln(\delta_{scal_i} \cdot i)))$ **(i=1,...,n ; σ: mean variation)**

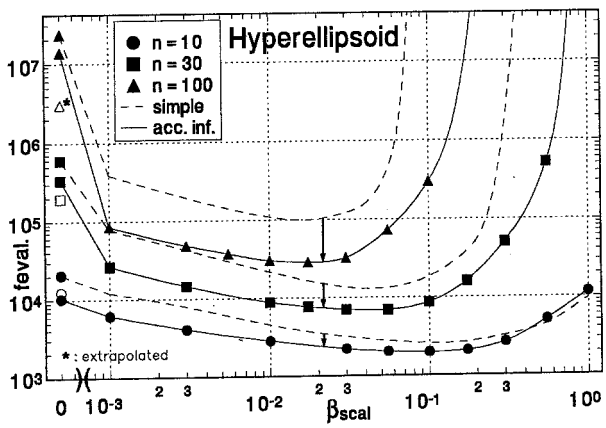

Figure 3 The symbols indicate the number of function evaluations to reach Fstop (average of 20 runs). For $\beta_{scal} \geq 0.5$; 0.1 (n=30; 100 resp.) the results of the simulations are influenced by the numerical precision of computation and thus are unreliable. The parameter settings of the derandomized ES with accumulated information are chosen as described above. Only β_{scal} varies. For $\beta_{scal} = 0$ no adaptation of individual step-sizes takes place. So only one general step-size is adapted (Symbols on the left). The dashed lines refer to step-size adaptation without accumulated information (simple). Schwefel's (8,50) ES (empty symbols) does not depend on the parameter β_{scal}. The results are shown here for comparison merely.

Schwefel´s problem

Objective function:

$$F(\vec{x}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{i} x_j \right)^2 \quad \Rightarrow \quad minimum \ (=0); \quad n = 20, \ -65 \le x_i^0 \le 65$$

This problem represents - with respect to the coordinate axes - rotated hyperellipsoids. Thus, correlated mutations should be superior to uncorrelated ones. The simulations (figure 4) show that the simple (1,10)-ES with adaptation of only one general step-size is about four times faster than Schwefel´s (8,50)-ES with correlated mutations. This suggests that no actual adaptation of the correlations to the topology of the problem takes place.

By the derandomized adaptation of individual step-sizes, optimization slows down by about 30 %. This is caused by the stochastic fluctuations of the individual step-sizes induced by the adaptation process. Because of the rotation of the ellipsoid axes, the initialization with identical individual step-sizes is optimal or nearly optimal. The acceleration of optimization using accumulated information is caused by an increased general step-size. This is comparable to the axis-parallel hyperellipsoids without adaptation of individual step-sizes.

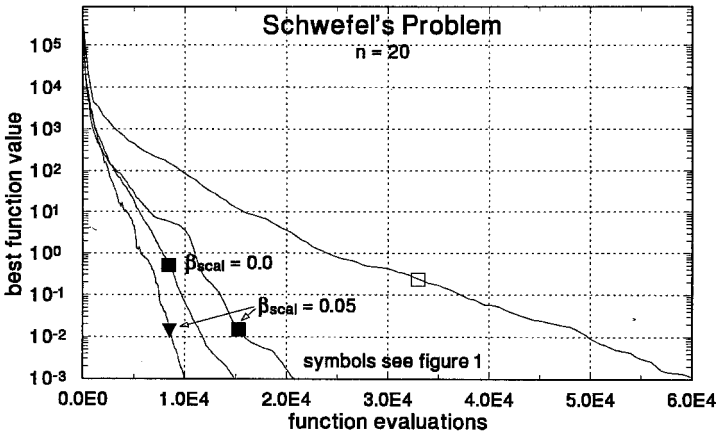

Figure 4
Convergence plots of optimizations with Schwefel´s problem

■ (1,10)-ES with ($\beta_{scal} = 0.05 = 1/n$) and without ($\beta_{scal} = 0$) adaptation of individual step-sizes

▼ (1,10)-ES with adaptation of individual step-sizes using accumulated information.

❑ (8,50)-ES according to Schwefel

Generalized Rosenbrock Function

Objective function:

$$F(\vec{x}) = \sum_{i=1}^{n-1} 100 \cdot \left(x_{i+1} - x_i^2 \right)^2 + (1 - x_i)^2 \quad \Rightarrow \quad minimum \ (=0)$$

$$n = 30, \ \vec{x}^0 = (0, \ \ldots \ , 0), \quad F(\vec{x}^0) = 29$$

This problem is characterized by the quadratic association of adjoining parameters. Thus correlated mutations of adjoining parameters should be superior. Schwefel´s (8,50)-ES facilitates a sensible adaptation only in the final stage of optimization (see figure 5). Derandomized adaptation of individual step-sizes accelerates the entire optimization cycle by increasing the step-sizes of adjoining parameters for which variations are of topical relevance.

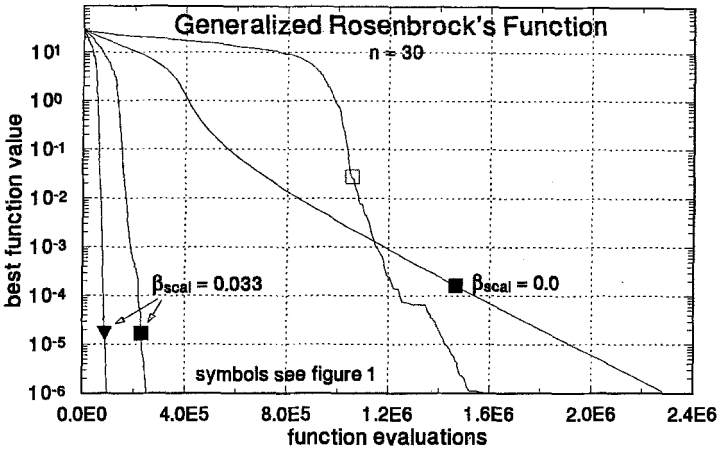

Figure 5
Convergence plots of optimizations with the generalized Rosenbrock function

■ (1,10)-ES with ($\beta_{scal} = 0.033 = 1/n$) and without ($\beta_{scal} = 0$) adaptation of individual step-sizes

▼ (1,10)-ES with adaptation of individual step-sizes using accumulated information.

❑ (8,50)-ES according to Schwefel

Sum of different Powers

Objective function:

$$F(\vec{x}) = \sum_{i=1}^{n} |x_i|^{(i+1)} \quad \Rightarrow \quad minimum \ (=0)$$

$$n = 30, \quad \vec{x}^0 = (1, \ldots, 1), \quad F(\vec{x}^0) = 30$$

This problem cannot be transformed into a hypersphere by an appropriate constant scaling. The sensitivity relations of the parameters (partial deviations of the quality fct.) continuously worsen when approaching the optimum. The derandomized ES is able to adapt the individual step-sizes according to the deteriorating scaling conditions. Its constant progress on the logarithmic scale is shown in figure 6. Schwefel's (8,50)-ES achieves a better quality than the (1,10)-ES without individual step-size adaptation but cannot deal with the deteriorating scaling conditions.

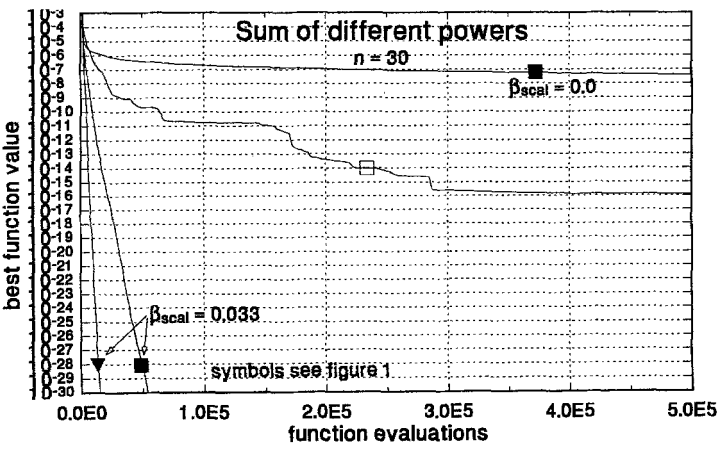

Figure 6
Convergence plots of optimizations with the sum of different powers.

■ (1,10)-ES with ($\beta_{scal} = 0.033 = 1/n$) and without ($\beta_{scal} = 0$) adaptation of individual step-sizes

▼ (1,10)-ES with adaptation of individual step-sizes using accumulated information.

❑ (8,50)-ES according to Schwefel

The Steiner-Net (with fixed topology)

The difficulty with this problem is comparable to the sum of different powers. The optimization problem is to minimize the length of a Steiner-net by finding the optimal positions of the Steiner-points (points of branching). The topology of the Steiner-tree is fixed (see figure 7).

The worsening sensitivity relations of the parameters and premature step-size convergence are caused by the linear dependency of the net-length on shiftings of Steiner-points that are located on "house" positions. The corresponding partial deviations of the quality function stay constantly about +1/-1 while the others converge to zero when approaching the optimum. As a result, the (1,10)-ES with mutative control of only one general step size and Schwefel's (8,50)-ES do not find the optimal Steiner-point positions.

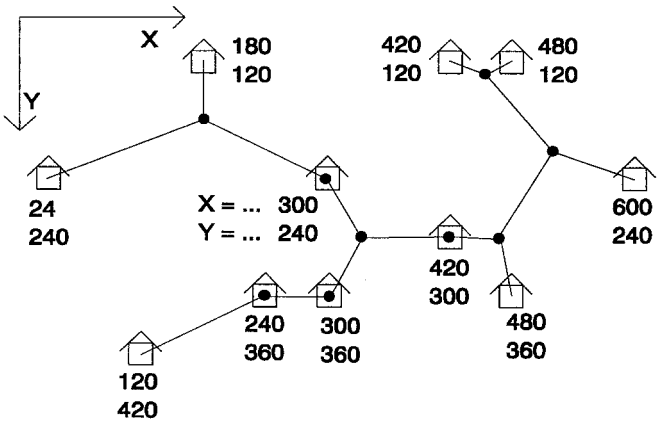

Figure 7
The points to be connected by the Steiner-net are symbolized by houses. The dots represent Steiner-points. The topology of the net is fixed as shown. Only the positions of the Steiner-points are subject to optimization. In the optimal solution, four of the nine Steiner-points are located at "house" positions.

Tests with the algorithm proposed here, have shown that it converges reliably to the optimum without premature step-size convergence. The (1,10)-ES without individual step-sizes and Schwefel's (8,50)-ES mostly converge to nets that are 1 to 10 units longer. Figure 8 shows some optimization runs.

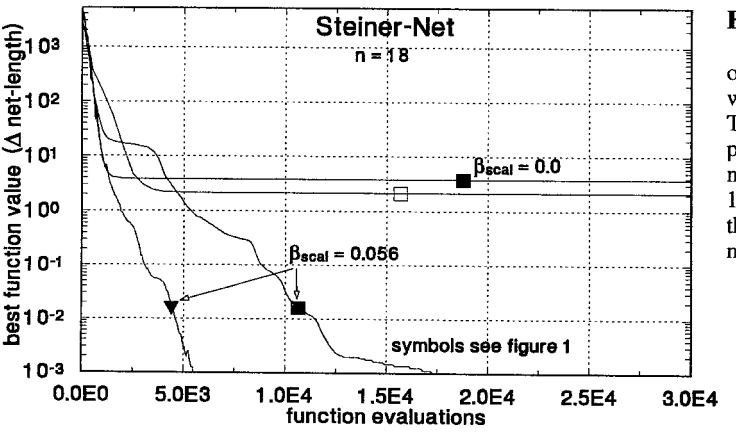

Figure 8
Convergence plots of optimizations runs with the Steiner-Net. The function values plotted are the actual net-length minus 1229.40854 which is the length of the minimal net.

Conclusions

A reliable adaptation of individual step-sizes is of essential importance for the applicability of the ES. Otherwise, the convergence rates can slow down by orders of magnitude for badly scaled problems. Even if the parameter-scaling seems not to be questionable, the lack of an appropriate adaptation of individual step-sizes can cause premature convergence of the general step-size.

Without modifications, mutative step-size control cannot be used for a reliable adaptation of individual step-sizes. Based on the concept of "Derandomized mutative step-size control", which enables a reliable step-size adaptation, the use of accumulated information decreases the locality of the adaptation process. Especially in small populations, the local character of step-size adaptation is disadvantageous because of the poor statistics involved. The additional information utilized by accumulation results from the generation sequence in a very simple way. Only the absolute values of the accumulated selected mutations have to be analyzed. The improvement achieved arises from the implicit use of correlations of the selected mutations in the generation sequence. Consequently, the step-sizes are adapted to such values that successive selected mutations tends to be orthogonal on average. This seems to be characteristical for optimal step-sizes in general.

Simulations show that the adaptation of individual step-sizes is accelerated considerably and becomes more precise and reliable at the same time. The adaptation of the general step-size can be improved as well. This occurs if the topology of the quality function resembles narrow valleys. In such cases the local character of step-size adaptation causes a too small step-size. Proceeding less locally, the use of accumulated information enables the adaptation of larger step-sizes and so accelerates the optimization.

References

Hoffmeister, F. & Bäck, T. (1991). *Genetic algorithms and evolution strategies: Similarities and differences.* In (Schwefel & Männer 1991), pages 455-470.

Ostermeier, A., Gawelczyk, A., Hansen, N. (1994). A Derandomized Approach to Self Adaptation of Evolution Strategies. In *Evolutionary Computation* (to be published).

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

Rechenberg, I. (1978). Evolutionsstrategien. In B. Schneider and U. Ranft (Eds.), *Simulationsmethoden in der Medizin und Biologie*, Berlin: Springer.

Rechenberg, I. (1994). *Evolutionsstrategie `94.* Stuttgart: Frommann-Holzboog (in print).

Schwefel, H.-P. (1977). *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie.* Volume 26 of *Interdisciplinary systems research.* Basel: Birkhäuser.

Schwefel, H.-P. (1981). *Numerical Optimization of Computer Models.* Chichester: Wiley.

Schwefel, H.-P. (1987). Collective phenomena in evolutionary systems. In *Preprints of the 31st Annual Meeting of the International Society for General System Research, Budapest,* 2: 1025-32.

Schwefel, H.-P. & Männer, R. (Eds.) (1991). *Parallel Problem Solving from Nature*, volume 496 of *Lecture Notes in Computer Science.* Berlin: Springer.

Voigt, H.-M., Born, J. & Treptow, J. (1991). *The Evolution Machine.* Manual. iir, Informatik, Informationen, Reporte.