A Derandomized Approach to Self Adaptation of Evolution Strategies

Andreas Ostermeier, Andreas Gawelczyk & Nikolaus Hansen

Technische Universität Berlin
Fachgebiet Bionik und Evolutionstechnik
Ackerstraße 71-76 (ACK1)
D-13355 Berlin
phone: 030 / 31472666
e-mail: ostermeier@fb10.tu-berlin.de

in Evol Comput 2(4) 1994

Abstract

Comparable to other optimization techniques, 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 have been evolved early in the development of ESs. These algorithms mostly work very efficiently as long as the scaling of the parameters to be optimized is known. If the scaling is not known, the strategy has to adapt individual step-sizes for all 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, and the algorithms known are not satisfactory.

The algorithm presented in this paper is based on the well known concept of mutative step-size control. Our investigations indicate that the adaptation by this concept declines due to an interaction of the random elements involved. We show that this weak point of mutative step-size control can be avoided by relatively small changes in the algorithm. The modifications may be summarized by the word "de-randomization". The derandomized scheme of mutative step-size control facilitates a reliable self-adaptation of individual step-sizes.

Keywords

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

1. Introduction: Step-size Adaptation in Evolution Strategies

In biology, mutation rates are of essential importance for evolutionary progress. Low mutation rates would not ensure a sufficiently high variability in the population. Too high mutation rates would result in a chaotic loss of genetic information. The evolutionary process is obviously able to adjust the mutation rates to sensible values, which vary very much for different species. Furthermore, the mutation rates are not constant for one species, but depend on the location in the genome (compare Eigen 1992). In the case of real valued continuous parameter optimization with evolution strategies (ESs), the biological mutation rate can be interpreted

as the size of mutation steps in the parameter space. Varying mutation rates on different locations in the genome can be interpreted as individual step-sizes for the different object parameters.

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 satisfactorily 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 and Schwefel (Rechenberg 1973, 1978; 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 treated similarly to the other (object-)parameters. They are incorporated in the genome of the individuals and thus affected by mutation and selection. In the following this is shown in a simple $(1, \lambda)$-ES algorithm with mutative step-size control of both general and individual step-sizes.

First, the object parameter vector $\vec{x}_E$ of the parent is mutated by adding a normally distributed random vector $\vec{z}$, which is scaled by the vector of individual step-sizes $\vec{\delta}$. Step-size variation is guaranteed by multiplying with the variation factors ξ and $\vec{\xi}$:

$$\vec{x}^{\,g}_{N_k} = \vec{x}^{\,g}_E + \xi^k \; \vec{\xi}^{\,k} \; \vec{\delta}^{\,g} \; \vec{z}^{\,k} \qquad (k = 1, \ldots, \lambda)$$

where all multiplications of vectors refer to components and used symbols are

- $\vec{x}^{\,g}_{E/N}$ object parameter vector of generation g
  E (elder): parent / N (newer): offspring
- $\vec{\delta}^{\,g}$ vector of (individual) step-sizes of generation g
- $\xi^k = \alpha; 1/\alpha$ with equal probability and $\alpha \in [1.1; 1.5]$ global step-size variation factor of offspring k
- $\vec{\xi}^{\,k} = (\xi_1, \ldots, \xi_n)$ individual step-size variation of offspring k, ξ_i distributed as ξ, often more sophisticated distributions are used for ξ_i
- $\vec{z}^{\,k} = (z_1, \ldots, z_n)$ with z_i (0, 1)–normally distributed

Second, the fitness of all λ offspring is evaluated. The best object parameter vector $\vec{x}_{N_{sel}}$ and the corresponding step-size vector are selected as parent of the next generation:

$$\begin{aligned} \vec{x}^{\,g+1}_E &= \vec{x}^{\,g}_{N_{sel}} \\ \vec{\delta}^{\,g+1} &= \xi^{sel} \; \vec{\xi}^{\,sel} \; \vec{\delta}^{\,g} \end{aligned}$$

where $sel \in \{1, \ldots, \lambda\}$ is the index of selected offspring of generation g.

2. Why Does this Adaptation Scheme Fail in Adapting Individual Step-Sizes in Small Populations?

Mutative step-size control generally works very well on the adaptation of a general step-size. Regardless of the distribution of ξ_i, which may also be continuously distributed, corresponding adaptation of individual step-sizes is not possible within small populations of simple ESs, as Schwefel (1987) pointed out. Schwefel favours the use of more complex ESs with larger populations. This works, but does not clarify the basic shortcomings, which are, in our opinion, the following:

- First, the standard algorithm of mutative step-size control presented above does not ensure that the selection of a large or small parameter variation leads to a corresponding increase or decrease of the individual step-size. For instance, a relatively large parameter variation may occur in spite of a small ξ_i due to a large random number instantiation of z_i. In such a case, the step-size will be reduced although a large mutation was selected. In general, the same parameter mutations (*random number instantiations*) can be generated with totally different sets of step-size variations. This problem is insignificant for the global step-size because $\|\vec{z}\|$ becomes virtually constant with increasing n, whereby ξ determines the length of the entire mutation step.

- Second, and most important, 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 variations in the generation sequence (i.e. between succeeding generations) have to be much smaller in order to reduce random fluctuations. Concerning the general step-size, this conflict has been analyzed extensively by Scheel (1985). Using intermediate recombination in large populations leads to significant reduction of these random fluctuations, which can produce long phases of stagnation in the optimization process.

3. Derandomized Mutative Step-Size Control

The modified mutative step-size control presented in the following section avoids the two difficulties discussed above. The possibility of producing the same mutations followed by different strategy parameter changes is prevented by using the *instantiation of* $\vec{z}$ for the step-size variation. Using the absolute values of the selected mutations as step-size adaptation factors ensures a variation of the individual step-sizes corresponding to the size of the selected mutations. Thus, it is guaranteed that the selection of large or small mutations leads to a corresponding adaptation of the step-sizes. The geometric mean of the χ_1-distributed absolute values of the mutations is less than one. Without selection, this causes systematically decreasing step-sizes, which can be prevented by a monotonic transformation $|z_i| \mapsto \xi_i$ generating a distribution with geometric mean equal to one. In the following algorithm, this is realized by $\xi_i = \exp\left\{|z_i| - \sqrt{2/\pi}\right\}$.

The conflict *step-size variation in one generation* versus *step-size adaptation in the generation sequence* cannot be resolved by choosing a good compromise for the step-size modification factor α. It appears promising, for instance, to determine the step-size adaptation rate not only by the step-size modification factor ξ, but by ξ^β, with $0 < \beta < 1$. The effect is clear: A step-size adaptation by a factor ξ is not realized in just one, but in at least $1/\beta > 1$ generations. Thus, the adaptation rate, and therefore the stochastic fluctuation, will be reduced *without* decreasing the variation between competing offspring. The information given by the selection of a large step-size is now interpreted as an indication to enlarge the step-size, while the selected step-size is *not* interpreted to be really the best one occurring in the population.

From a general point of view, step-size adaptation can be interpreted as a problem of disturbed optimization (Rechenberg 1994). *Disturbed* means that the quality/fitness value is not exactly measurable. Corresponding tests have shown that the concept of relatively large mutations within one generation, but passing only smaller variations to the next generation, is applicable successfully to parameter optimization superimposed with Gaussian noise.

4. $(1, \lambda)$-ES algorithm with derandomized mutative step-size

In this section we formally present the derandomized ES algorithm. All multiplications and powers of vectors refer to components.

1. Creation of λ offspring ($k = 1, \ldots, \lambda$):

$$\vec{x}^{\,g}_{N_k} = \vec{x}^{\,g}_{E} + \xi^k \; \vec{\delta}^{\,g} \; \vec{z}^{\,k}$$

2. Selection / Adaptation:

$$\begin{aligned} \vec{x}^{\,g+1}_{E} &= \vec{x}^{\,g}_{N_{sel}} \\ \vec{\delta}^{\,g+1} &= \left(\xi^{sel}\right)^{\beta} \left(\vec{\xi}_{\,\vec{z}^{\,sel}}\right)^{\beta_{\mathrm{scal}}} \vec{\delta}^{\,g} \end{aligned}$$

Symbols used:

$\vec{x}^{\,g}_{E/N}$ object parameter vector of generation g
E (elder): parent / N (newer): offspring

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

$\xi^k = \alpha; 1/\alpha$ with equal probability and $\alpha = 1.4$ global step-size variation of offspring k

$\vec{\xi}_{\,\vec{z}^{\,sel}} = (\xi_1, \ldots, \xi_n)$ with $\xi_i = \exp\left\{|z_i^{sel}| - \sqrt{2/\pi}\right\}$ individual step-size adaptation, where $\sqrt{2/\pi}$ is the expectation of $|z_i|$. This is a simple way to prevent a systematic drift without selection. It is also possible to transform $|z_i|$ 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.

$\vec{z}^{\,k} = (z_1, \ldots, z_n)$ with z_i (0, 1)–normally distributed

sel index of selected offspring of generation g

n number of object parameters to be optimized (size of all vectors used)

$\beta = \sqrt{1/n}$
$\beta_{\mathrm{scal}} = 1/n$
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. In the case of very difficult problems, a reduction of β_{scal} might be necessary.

Derandomization is introduced at three places in the algorithm:

- First, the step-size adaptation rate can be adjusted by choosing $\beta \ll 1$ such that disturbing random fluctuations of the step-sizes are reduced.
- Second, the (0,1)-normally distributed z_i realize mutations that are of clearly different sizes $|z_i|$. This ensures selection relevance with respect to the individual step-size adaptation. Otherwise, step-sizes would be adapted by chance and perform nearly random walks.
- Third, the change of the individual step-sizes directly depends on the absolute value of a selected mutation. This ensures a corresponding adaptation of the step-sizes in the case of small or large selected mutations.

5. Simulations

Tests of the described agorithm 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 new step-size adaptation, simulation results with an (15/2,100)-ES according to Schwefel (1981), with adaptation of global and n individual step-sizes and intermediate recombination on object and strategy parameters, are also presented. Here, for reliable adaptation the number of parents has to be chosen — depending on n — clearly greater than one. Therefore in general, one may not be able to choose population size optimal according to the given objective problem.

The simulations with Schwefel's (15/2,100)-ES have been carried out with the *Evolution Machine* developed by Voigt, Born and Treptow (1991).

5.1 Axis-Parallel Hyper-Ellipsoids

Objective function:

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

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

The simulation results show that optimization speeds up considerably with derandomized adaptation of individual step-sizes, compared to the (1,10)-ES without individual step-size adaptation (see figure 1). The feasible speed-up factor (10 to 200 here) increases with n due to the increasing ratios of the ellipsoid-axes (see figure 2).

The optimization runs shown in figure 1 demonstrate clearly the ability of the derandomized step-size adaptation to adjust the correct set of individual step-sizes by which the problem is transformed into a hyper-sphere. After about 8000 function evaluations, the step-sizes are adapted correctly, and the convergence rate is nearly as high as with fixed individual step-sizes, that are preadjusted correctly. Schwefel's (15/2,100)-ES shows no distinct adaptation phase and is five to six times slower than the ES with correctly preadjusted individual step-sizes. Even taking into account the smaller progress rate due to the large population, this indicates that actually no complete adaptation of the correct scaling takes place, but some kind of subspace search is performed.

In order to find out how to choose β_{scal}, the number of function evaluations needed to reach F_{stop} are measured for different values of β_{scal}. The plots in figure 2 show clearly minima of function evaluations needed to reach F_{stop} (maxima of convergence speed). They result from the conflict of *fast* versus *precise* adaptation. For small values of β_{scal}, the adaptation process is slow, but finally approximates very precisely the correct set of individual step-sizes. Medium values of β_{scal} enable a faster adaption but cause more stochastic fluctuations of the individual step-sizes. Too large values of β_{scal} provoke such stochastic fluctuations that no sensible adaptation is possible. For $\beta_{\text{scal}} = 0$, no adaptation of individual step-sizes takes place and only one general step-size is adapted (symbols on the left). Choosing F_{stop} less / greater than 10^{-10} will somewhat move the minima to the left / right respectively.

According to figure 2, 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 $\beta_{\text{scal}} = 1/n$ seems to be a good choice for a wide range of different problems.

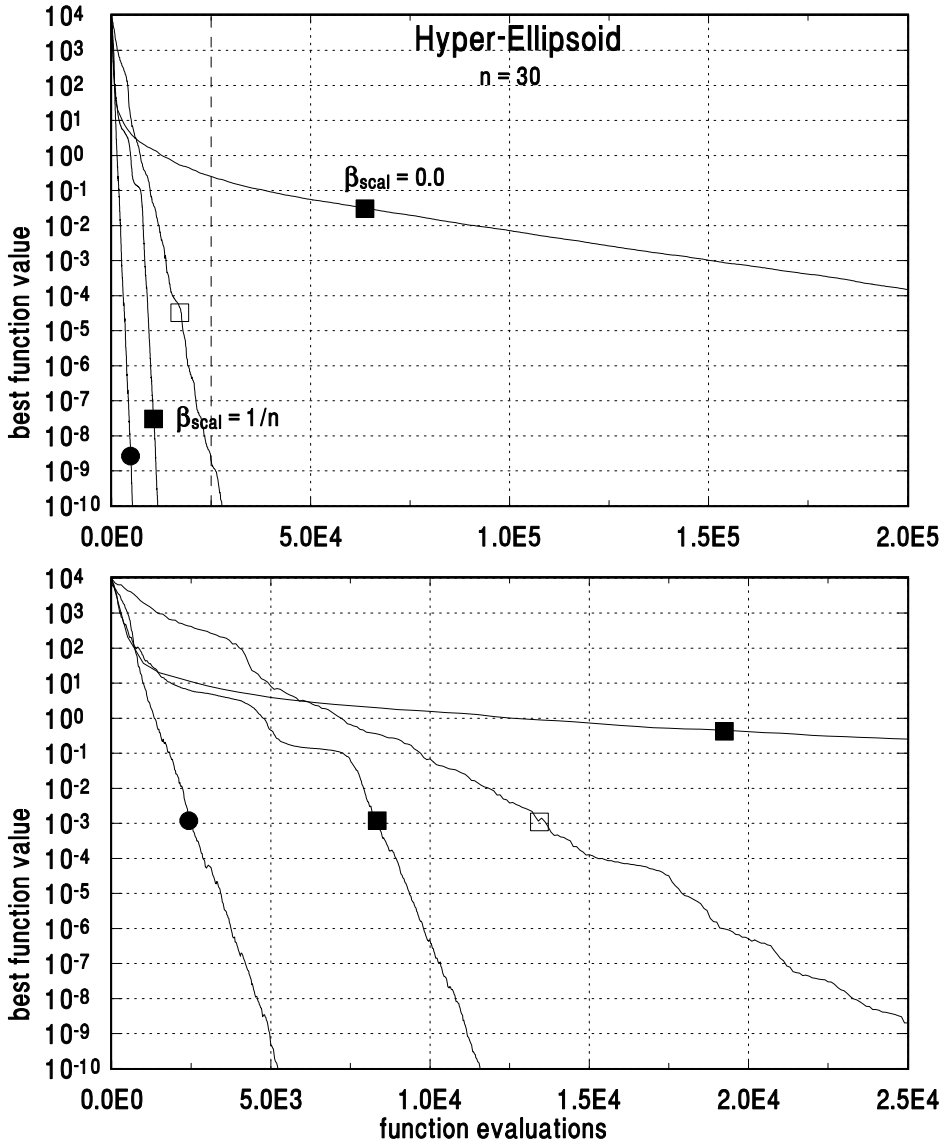

Figure 1: Convergence plots of optimization runs on the hyper-ellipsoid. ■ = (1,10)-ES with / without derandomized individual step-size adaptation, respectively ($\beta_{\text{scal}} = (1/n)$ / 0, respectively); □ = (15/2,100)-ES with mutative individual step-size adaptation according to Schwefel (1981); • = (1,10)-ES with mutative global step-size adaptation and correctly adjusted scaling of individual step-sizes. The lower figure is an enlarged detail of the first 25 000 function evaluations of the optimization runs shown above.

5.2 Schwefel's Problem

Objective function:

$$F(\vec{x}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{i} x_j \right)^2 \qquad \rightarrow \quad \text{minimum } (= 0)$$

$$n = 20, \qquad -65 \leq x_i^0 \leq 65$$

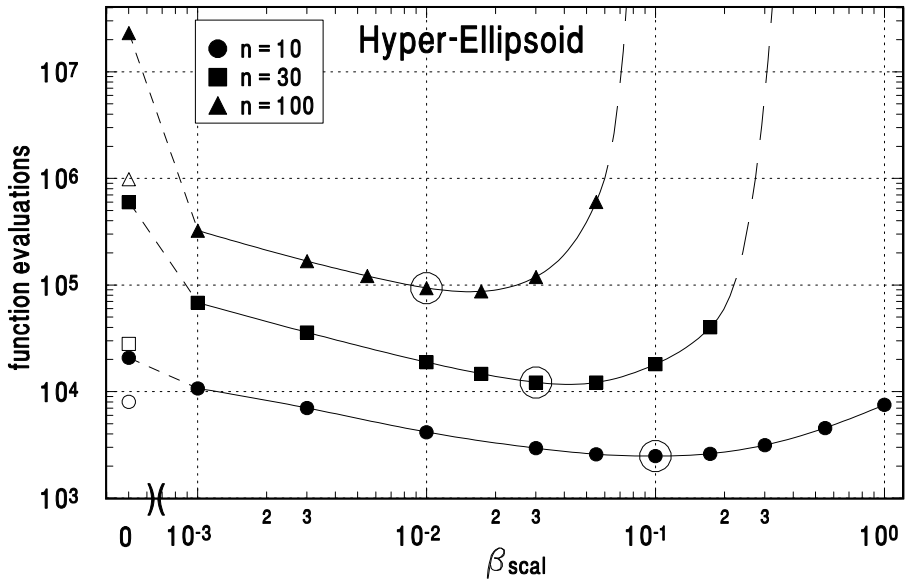

Figure 2: Number of function evaluations to reach F_{stop} with the derandomized ES at different values of β_{scal}. The circled symbols refer to $\beta_{scal} = 1/n$, as chosen in all further simulations. Schwefel's (15/2,100)-ES (empty symbols) is shown for comparison merely and does not depend on the parameter β_{scal}.

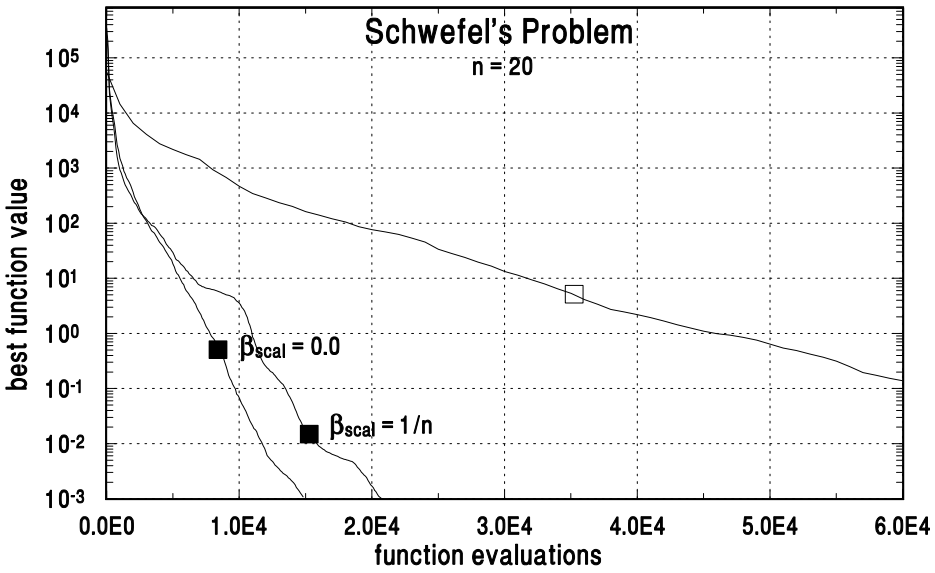

Figure 3: Convergence plots of optimizations on Schwefel's problem. ■ = (1,10)-ES with / without derandomized individual step-size adaptation, respectively (β_{scal} = $(1/n)$ / 0, respectively); □ = (15/2,100)-ES with mutative individual step-size adaptation according to Schwefel (1981).

This problem represents — with respect to the coordinate axes — rotated hyper-ellipsoids. The simulations (see figure 3) show that the simple (1,10)-ESs are about five times faster than Schwefel's (15/2,100)-ES. 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,

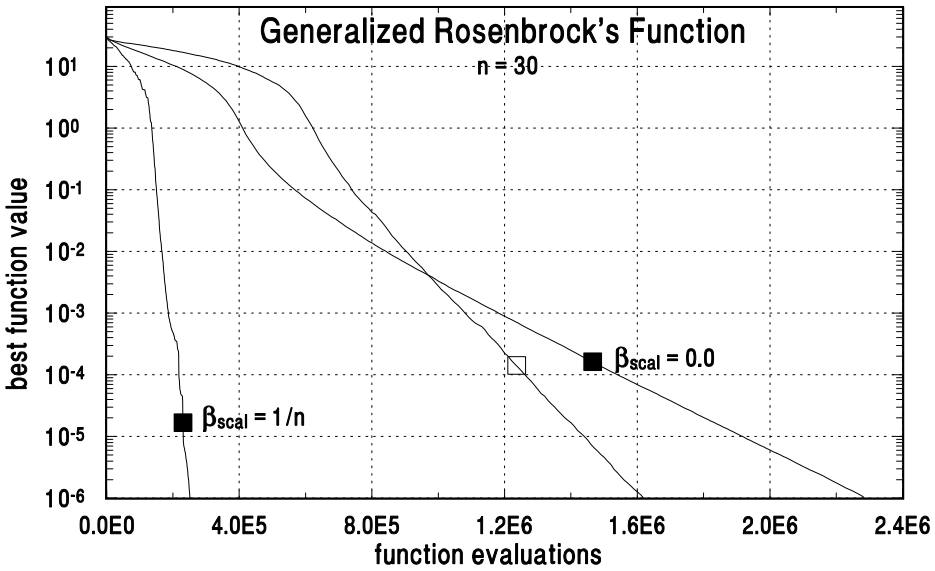

Figure 4: Convergence plots of optimizations on the generalized Rosenbrock function. ■ = (1,10)-ES with / without derandomized individual step-size adaptation, respectively ($\beta_{\mathrm{scal}} = (1/n)$ / 0, respectively); □ = (15/2,100)-ES with mutative individual step-size adaptation according to Schwefel (1981).

the initialization with identical individual step-sizes is optimal or nearly optimal.

5.3 Generalized Rosenbrock Function

Objective function:

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

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

This problem is characterized by the quadratic association of adjoining parameters. As can be seen in figure 4, 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. In the final stage, Schwefel's (15/2,100)-ES is about eight times slower than the derandomized ES.

5.4 Sum of Different Powers

Objective function:

$$F_n(\vec{x}) = \sum_{i=1}^{n} |x_i|^{i+1} \qquad \rightarrow \quad \text{minimum } (= 0)$$

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

This problem cannot be transformed into a hyper-sphere by an appropriate *constant* scaling. The sensitivity relations (partial deviations) of the parameters continuously worsen

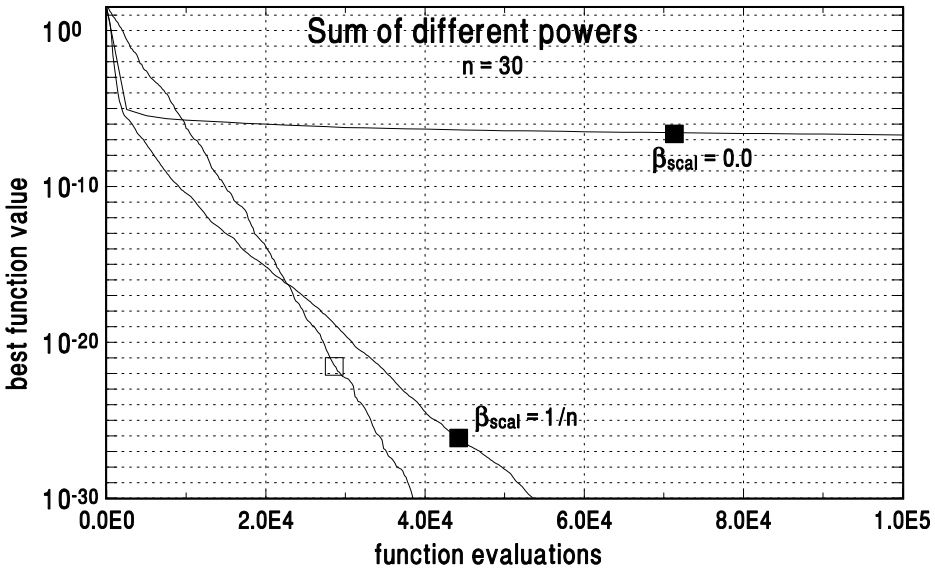

Figure 5: Convergence plots of optimizations on the sum of different powers. ■ = (1,10)-ES with / without derandomized individual step-size adaptation, respectively ($\beta_{\mathrm{scal}} = (1/n)$ / 0, respectively); □ = (15/2,100)-ES with mutative individual step-size adaptation according to Schwefel (1981).

when approaching the optimum. Both adaptation mechanisms can cope with the deteriorating scaling conditions. Their constant progress on the logarithmic scale is shown in figure 5. Schwefel's (15/2,100)-ES is about two times faster than the derandomized ES. This separable problem requires rather fast than precise adaptation.

5.5 The Steiner-Net (with fixed topology)

The optimization problem is to minimize the length of a Steiner-net by finding the optimal positions of the points of branching (Steiner-points). The points to be connected (house-points) and the topology of the net are fixed as shown in figure 6. Only the positions of the Steiner-points are subject to optimization. In the optimal solution, four of nine Steiner-points are located at house positions, while the others take positions with angles of 120° between their branches.

The difficulty with this problem is comparable to the sum of different powers. Worsening sensitivity relations of the parameters and premature step-size convergence of simple ESs are caused by the linear dependency of the net length on shifting of Steiner-points that are located on house positions. The corresponding partial deviations of the quality function remain constantly about $+1$ or -1, respectively, 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 does not find the optimal Steiner-point positions (see figure 7).

Tests with the individual step-size adaptation schemes have shown that they converge to the optimum without premature step-size convergence, while the derandomized scheme is slightly faster than Schwefel's (15/2,100)-ES, as can be seen in figure 7. To approximate the minimal net, the length of which is about 1229.40854, with an absolute precision of 10^{-3}, about 20000 function evaluations are needed. The (1,10)-ES without individual step-sizes mostly converge to nets that are 1 to 10 units longer.

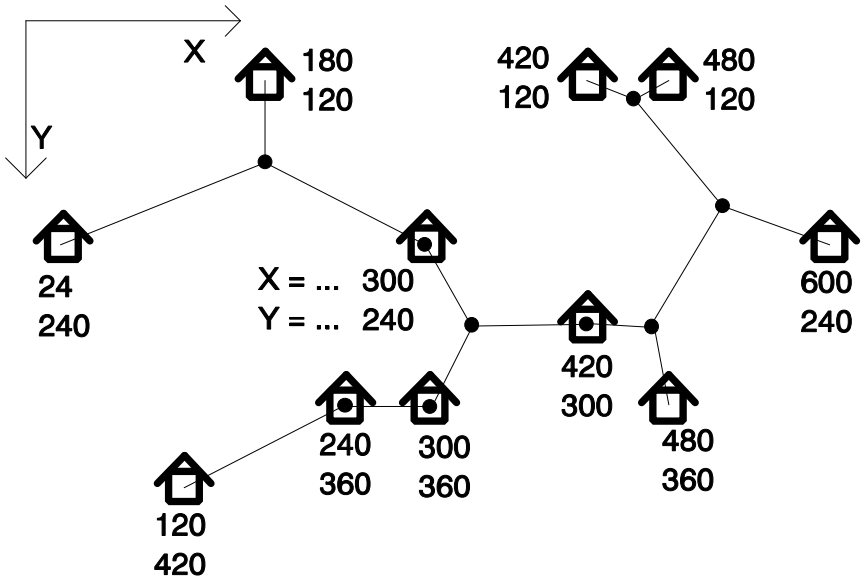

Figure 6: Steiner-net, used for optimization runs. The fixed points to be connected by the Steiner-net are symbolized by houses. The dots represent Steiner-points.

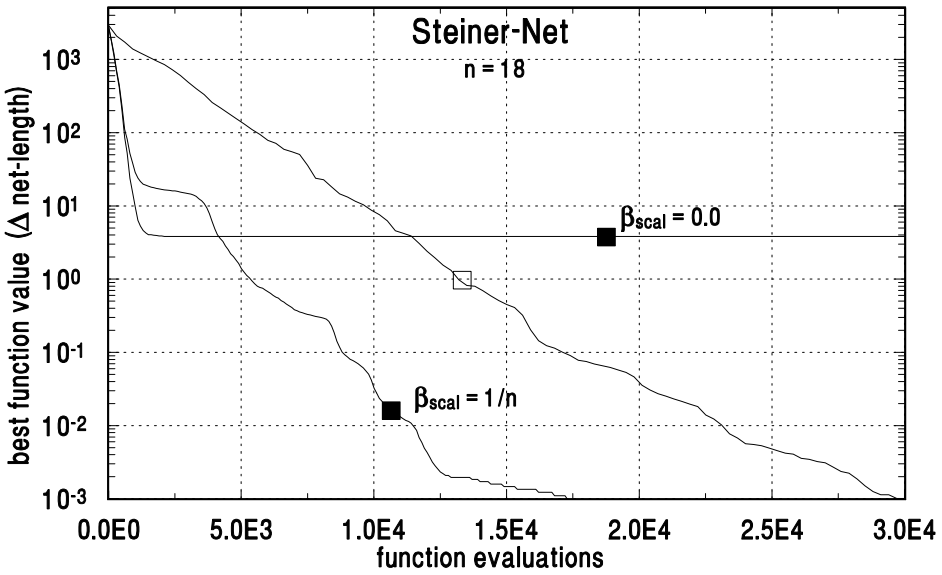

Figure 7: Convergence plots of optimizations runs on the Steiner-Net. ■ = (1,10)-ES with / without derandomized individual step-size adaptation, respectively (β_{scal} = $(1/n)$ / 0, respectively); □ = (15/2,100)-ES with mutative individual step-size adaptation according to Schwefel (1981). The function values plotted are the actual net lengths minus 1229.40854, which is the length of the minimal net.

6. 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 (e.g. Steiner-Net).

The attempts to use the concept of *mutative step-size control* to deal with the adaptation of individual step-sizes in small populations, have not been convincing up to now. Weak points of the mutative step-size control that are related to this shortcoming are demonstrated in this paper.

The new algorithm presented here demonstrates one possibility of overcoming these difficulties by relatively small modifications without changing the basic idea of mutative step-size control. *Derandomized mutative step-size control* allows a reliable adaptation of individual step-sizes even within quite simple ES-variants such as a (1,10)-ES. The population size needed for the adaptation process does not depend on the dimension of the problem. Furthermore, no extra function evaluations and only small computational expense are required.

Acknowledgement
Research for this paper was partly supported by the Bundesministerium für Forschung und Technologie (BMFT) under grant SALGON.

References

Eigen, M. (1992). Virus-Quasispezies oder die Büchse der Pandora. In: *Spektrum der Wissenschaft*, December 1992, 42–55.

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

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

Rechenberg, I. (1994). *Evolutionsstrategie '94.* Stuttgart: Frommann–Holzboog.

Scheel, A. (1985). *Beitrag zur Theorie der Evolutionsstrategie.* Doctoral thesis. Berlin: Technical University of Berlin.

Schwefel, H.–P. (1977). *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie.* Interdisciplinary Systems Research, *26.* 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, 2* (pp. 1025–1032). Budapest.

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