An Evolution Strategy with Coordinate System Invariant Adaptation of Arbitrary Normal Mutation Distributions Within the Concept of Mutative Strategy Parameter Control

Andreas Ostermeier & Nikolaus Hansen

Fachgebiet für Bionik und Evolutionstechnik

Technische Universität Berlin

Ackerstr. 71–76, 13355 Berlin, Germany

E-mail: {hansen,ostermeier}@bionik.tu-berlin.de

In: *W.Banzhaf, J.Daida, A.Eiben, M.H.Garzon, V.Honavar, M.Jakiela, R.E.Smith (Eds.). GECCO-99 Proceedings of the Genetic and Evolutionary Computation Conference (1999):* 902–909, San Francisco: Morgan Kaufmann Publishers.

http:/www.bionik.tu-berlin.de/user/niko/publications.html

ERRATUM: Equation (2) must be $\mathbf{C}_k^{(g+1)} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{z}_i \mathbf{z}_i^{\mathrm{t}}$

An Evolution Strategy with Coordinate System Invariant Adaptation of Arbitrary Normal Mutation Distributions Within the Concept of Mutative Strategy Parameter Control

Andreas Ostermeier
Technische Universität Berlin, Sekr. ACK 1
Fachgebiet für Bionik und Evolutionstechnik
Ackerstr. 71–76, 13355 Berlin, Germany
e-mail: ostermeier@bionik.tu-berlin.de
phone:+30 / 314 72666, fax:+30 / 314 72658

Nikolaus Hansen
Technische Universität Berlin, Sekr. ACK 1
Fachgebiet für Bionik und Evolutionstechnik
Ackerstr. 71–76, 13355 Berlin, Germany
e-mail: hansen@bionik.tu-berlin.de
phone:+30 / 314 72666, fax:+30 / 314 72658

Abstract

A self-adaptation of arbitrary normal mutation distributions within the concept of *mutative strategy parameter control* (MSC) using a newly formulated mutation operator is introduced. The coordinate system independent formulation ensures the invariance of the algorithm towards arbitrary linear transformations, which is a novelty for self-adaptation within the concept of MSC. To enable a sensible adaptation, the population size must scale quadratically with the problem size N – according to the number of adapted strategy parameters. Because the adaptation time (number of generations) also scales with N^2, the overall adaptation expense amounts to N^4 function evaluations.

1 Introduction

The essential feature of self-adaptation within evolution strategies (ESs) can be implemented using different basic concepts. Mostly common is the well known concept of *mutative* strategy parameter control (MSC) (Rechenberg 1973; Schwefel 1981; Rechenberg 1994; Schwefel 1995). A different approach is represented by the derandomized step-size control (DSC), suggested by Ostermeier et al. (1994). In both adaptation schemes the adaptation of arbitrary normal distributions has been realized. On the one hand Schwefel (1981) proposed an adaptation within the concept of MSC, which we will call rotation angle adaptation (RAA). On the other hand the covariance matrix adaptation (CMA) proposed by Hansen and Ostermeier (1996) utilizes the concept of DSC.

Apart from the different underlying adaptation concepts, RAA and CMA respond completely different to linear transformations of the object parameter space. Invariance against such linear transformations is of major importance, because it enables generalization of performance measurements obtained on test or real world objective functions. The CMA can be shown to be invariant towards arbitrary linear transformations of the object parameter space (apart from initialization). The RAA does not show this invariance.

The aim of this paper is to introduce the self-adaptation of arbitrary normal mutation distributions within the concept of MSC realizing invariance properties comparable to the CMA. This allows to compare the two different adaptation concepts in a more sensible way and offers answers to the questions:

1. Can self-adaptation with the mentioned invariance be successfully applied within the concept of MSC?
2. Can the mutative concept compete with the derandomized approach?

In this paper we mainly intend to answer these two scientifically interesting questions – we do not intend to introduce a new ES-algorithm to solve real world problems faster or more reliably.

2 The Concept of Mutative Strategy Parameter Control

The basic idea of the concept of mutative strategy parameter control (MSC) is to deal with the strategy

parameters that are controlling the mutation distribution in a comparable way as with the object parameters. Strategy parameters are added to the genome of the individuals and are subjected to mutation and recombination like the object parameters.

The mutation of an individual is carried out in two steps:

- First, the mutation distribution (concerning the strategy parameters) is mutated.
- Second, this variated (mutated) distribution is used to generate the mutation of the object parameters.

The fitness selection of the individuals of course only depends on the object parameter setting. Therefore the selection of better mutation distributions is only possible, if these actually produced the more successful object parameter variations.

2.1 The (Normal) Mutation Distribution

Concerning the adaptation of one general step-size or N individual step-sizes (mean variations in the axes of the given coordinate system) the concept of MSC proves to work[1], not taking into account the generally observed phenomenon, that the overall variance of the mutation distribution is adapted systematically too small by MSC. The self-adaptation of arbitrary normal mutation distributions in the concept of MSC has failed up to now. The rotation angle adaptation (RAA) (Schwefel 1981) indeed is able to *produce* arbitrary normal mutation distributions (Rudolph 1992), but does not really adapt the mutation distribution to different target topologies of the fitness function. One reason for this is the coordinate system dependent formulation of the rotation procedure. As a result the algorithm shows for instance a totally different behavior on different linear transformations of the fitness function hyper-sphere.

Generating arbitrary normal distributions is equivalent to linear transformations of a given normal distribution and equivalent to a linear transformation of the object parameter space (and object parameter vector accordingly). An algorithm, that adapts arbitrary normal mutation distributions should be invariant towards arbitrary linear transformations of the object parameter space (apart from initialization).

[1] Choosing suitable population size and recombination scheme.

2.2 The Mutation Operator

To realize a self-adaptation of the mutation distribution within the concept of MSC, the mutation operator on the strategy parameters is of outstanding importance. If only one general step-size has to be adapted (the shape of the mutation distribution remains constant) the formulation of the mutation operator is quite simple. In this case the mutation of the mutation distribution is carried out by multiplying the variance of the mutation distribution with a properly chosen random factor. Multiplicative variation ensures the invariance of the algorithm towards linear transformation of the object parameter space of the form $\mathbf{x} \mapsto c\,\mathbf{x}$, where $c \in \mathbb{R}_{>0}$. An additive variation of the mutation variance would not ensure this invariance. Comparably the adaptation of individual step-sizes can be realized by multiplying the axes-parallel variances of the normal mutation distribution with independently generated random factors. Such algorithms will be invariant towards linear transformations of the object parameter space of the form $\mathbf{x} \mapsto \mathbf{D}\mathbf{x}$, where $\mathbf{D}$ is a $N \times N$ diagonal matrix.

At this point the interesting question is, how to formulate a mutation operator, that allows to produce and self-adapt an arbitrary normal mutation distribution in a way, that the resulting algorithm is invariant towards any linear transformation of the object parameter space. Of course, as mentioned before, the invariance depends on the choice of the initial mutation distribution. With different initializations usually a distinct adaptation phase will occur. In any case the optimization process should be completely independent of linear transformations if object parameter space *and* initialization of the algorithm are equally transformed.

A mutation operator that produces the features described above defines a sensible generalization of global and individual step-size adaptation because it yields the corresponding invariances.

3 The Coordinate System Independent Mutation Operator

In this section we describe a mutation operator for the mutation distribution, that produces arbitrary normal mutation distributions according to the invariance requirements mentioned above. Consider drawing a sample of N points from the mutation distribution. Such a sample of N vectors can be used to constitute a new, mutated distribution in a very simple manner: Multiply all the vectors with independently normally distributed random numbers and add up the resulting

random vectors. While this procedure is per se independent of any given coordinate system the chosen starting distribution can be interpreted as such dependency. Because the mutation process is repeated in an iterated sequence the dependency on the initial distribution vanishes with increasing generation number. This can be interpreted as the adaptation process.

Figure 1a and 1b illustrate the mutation operator. In the upper left corner a two-dimensional initial (parental) mutation distribution is shown. The following eleven distributions are generated using $N = 2$ random vector realizations of the parental distribution each time. In figure 1a the parental distribution is isotropic; figure 1b gives an example of an anisotropic parental distribution. In contrast to figure 1, that shows the mutation distributions of parent and descendants of one generation, **figure 2** shows a generation sequence. Here every distribution is generated using the preceding one. The degeneration of the distribution into one dimension is important to notice. In the situation shown here, this degeneration is inevitable. It happens in the ES, if random selection takes place. Under non-random selection this will be prevented by choosing a sufficiently large parent number μ combined with an intermediate recombination scheme.

4 Algorithm: The $(\mu/_{\mathrm{I}}\rho, \lambda)$-ES

We formulate the algorithm for the $(\mu/_{\mathrm{I}}\rho, \lambda)$-ES – an evolution strategy with μ parents, λ descendants and Intermediate multi-recombination of ρ out of μ parents. Every individual in the population consists of the object parameter vector $\mathbf{x} \in \mathbb{R}^N$ and the strategy parameters $\sigma \in \mathbb{R}_{>0}$ and $\mathbf{C} \in \mathbb{R}^N \times \mathbb{R}^N$, where σ can be interpreted as step-size, while $\mathbf{C}$ is a symmetric and positive definite covariance matrix. The transition from generation g to $g+1$ is defined by the following equations:

For descendant $k = 1, \ldots, \lambda$

$$\sigma_k^{(g+1)} = \sqrt[\rho]{\prod_{i \in I_{k,g}^{\mathrm{sel}}} \sigma_i^{(g)}} \cdot \exp(\xi), \quad (1)$$

where $\mathrm{P}(\xi = 0.4) = \mathrm{P}(\xi = -0.4) = 1/2$

$$\mathbf{C}_k^{(g+1)} = \sum_{i=1}^{m} \mathbf{z}_i \mathbf{z}_i^{\mathrm{t}}, \quad (2)$$

where $\mathbf{z}_i \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\rho} \sum_{j \in I_{k,g}^{\mathrm{sel}}} \mathbf{C}_j^{(g)}\right)$ i.i.d.

$$\mathbf{x}_k^{(g+1)} = \frac{1}{\rho} \sum_{i \in I_{k,g}^{\mathrm{sel}}} \mathbf{x}_i^{(g)} + \sigma_k^{(g+1)} \mathcal{N}\left(\mathbf{0}, \mathbf{C}_k^{(g+1)}\right) \quad (3)$$

$I_{k,g}^{\mathrm{sel}}$ is the index set of the ρ (parent-)individuals of generation g selected for recombination for descendant k. If $\rho = \mu$, it holds $I_{1,g}^{\mathrm{sel}} = \ldots = I_{\lambda,g}^{\mathrm{sel}}$. Equation (2) facilitates the mutation operator introduced above. m in (2) determines the mutation strength for $\mathbf{C}$. Large m means small mutation strength, because a large sample $\mathbf{z}_1, \ldots, \mathbf{z}_m$ gets closer to the original distribution than a small one. For $m < N$ the $\mathcal{N}\left(\mathbf{0}, \mathbf{C}_k^{(g+1)}\right)$ distribution becomes singular. We choose the "natural" value $m := N$ (see below).

Equation (1) provides an additional mutation of the overall variance, using the geometric mean of the ρ parent step-sizes. This additional adaptation of the global step-size can be very useful, because it facilitates a faster change of the overall variance than (2). The actual distribution of ξ in (1) is of lower relevance, if the given variance and its zero mean is retained. In view of this paper (1) can be removed without significant qualitative influence on our results.

5 Choice of Strategy Control Parameters

Using a $(\mu/_{\mathrm{I}}\mu, \lambda)$-ES with self-adaptation of arbitrary normal mutation distributions arises the question of how to choose the strategy control parameters. Of predominant importance are the population size and the mutation strength of the here newly formulated mutation operator.

The mutation operator described above has a defined mutation strength, because $m = N$. A natural way to vary its mutation strength is to use more than N random vectors to generate a mutated random distribution. The more vectors are used, the smaller is the variation of the mutation distribution. Larger variations can be realized by using less than N random vectors or by repeatedly applying the mutation operator. All these methods have practical disadvantages. We use a different method to vary the mutation strength:

Smaller variations are realized by mixing the mutated distribution with the parental distribution by averaging the corresponding covariance matrices. Larger variations are realized by multiplying the N generated random vectors each one with independent log-normal random numbers[2].

Firstly, we look at the loss of progress depending on the mutation strength to be tuned. **Figure 3** shows

[2]The square sum of these random numbers has to be normalized to N in order to leave the overall variance of the mutation distribution unchanged.

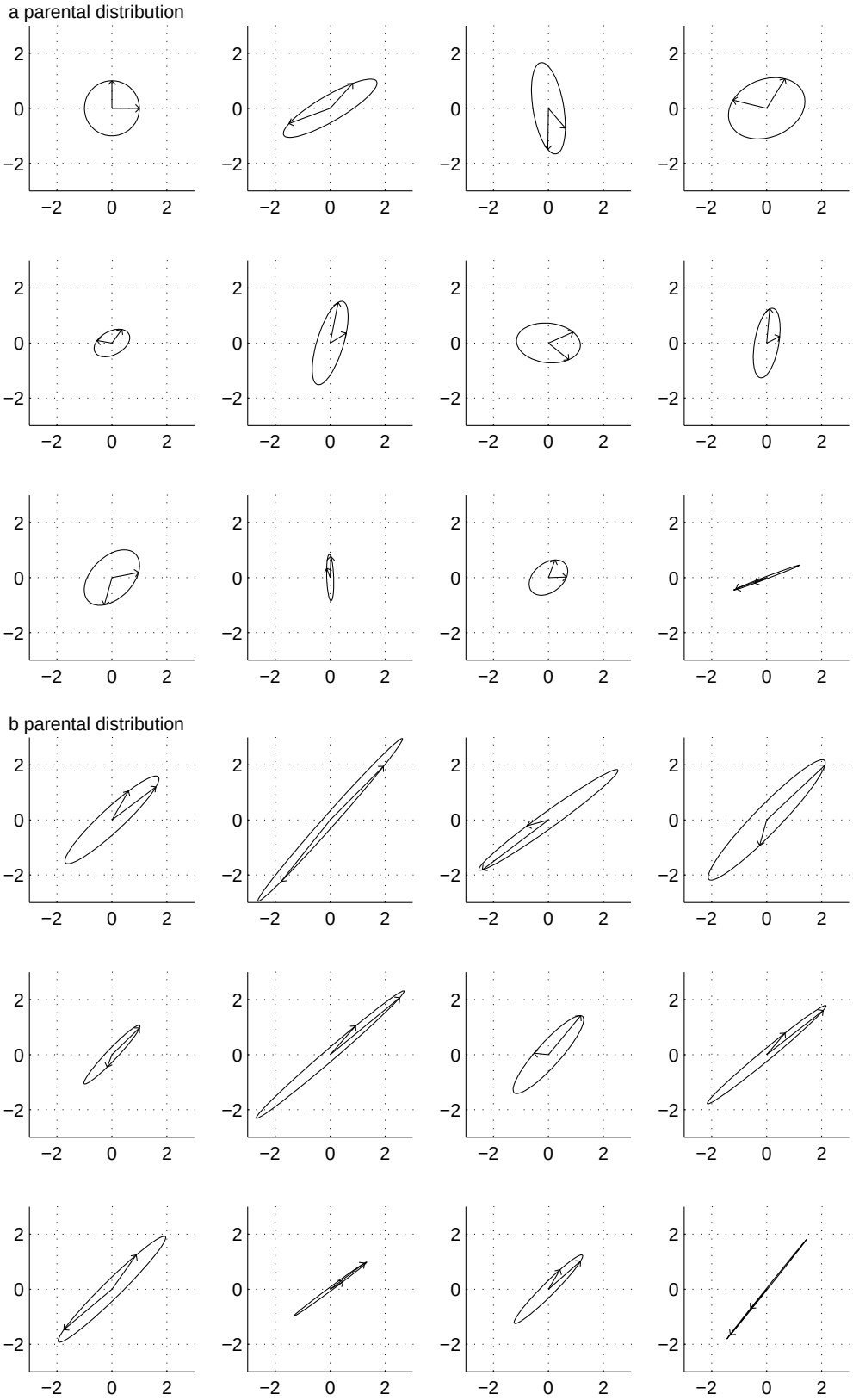

Figure 1: Parental (upper left) and eleven mutated distributions. The arrows are realizations from the parental distribution and are the vectors, the new distribution is constructed with. The ellipsoids indicate the one-σ line of the normal distributions.

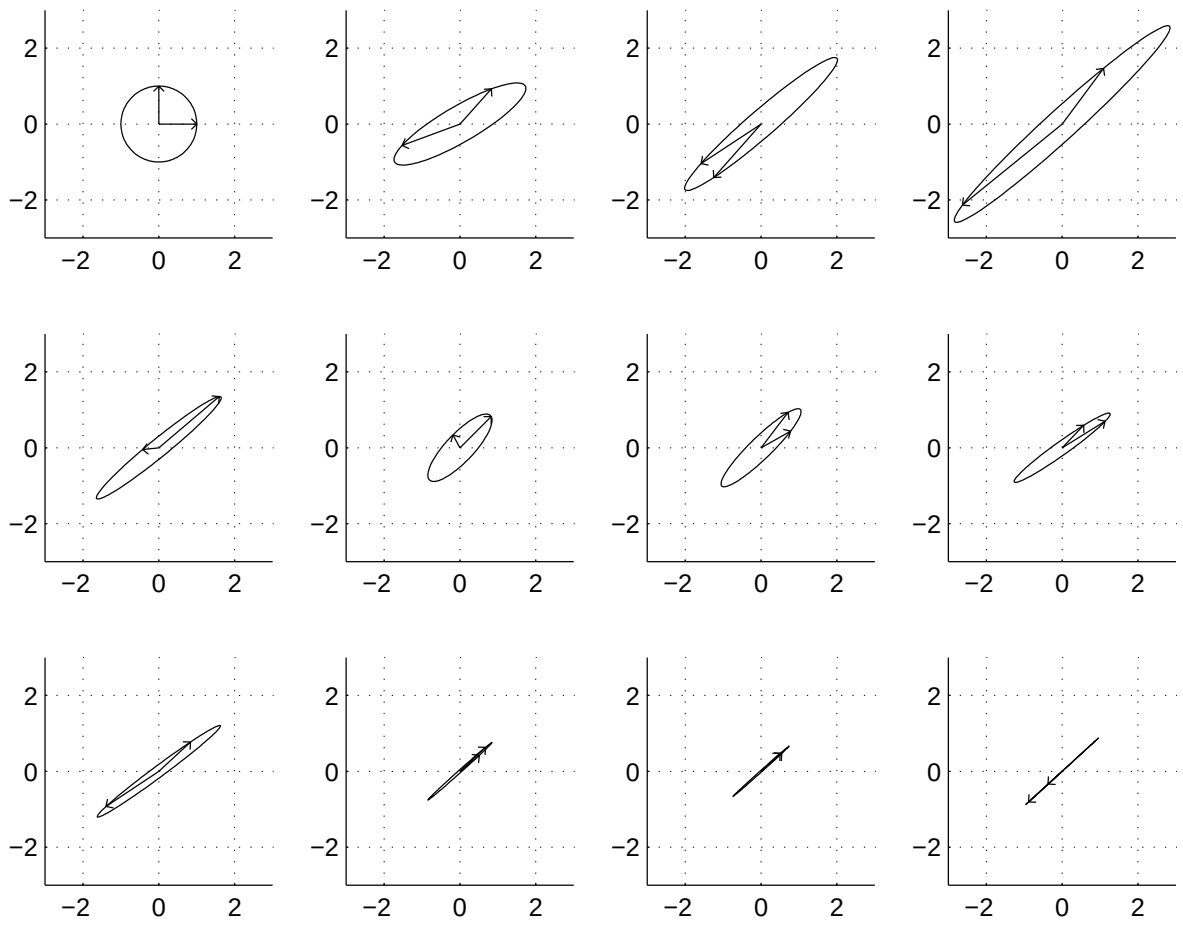

Figure 2: Sequence of distributions, respectively generated by realizations of the preceding distribution. The ellipsoids indicate the one-σ line of the normal distributions, the arrows the realizations due to the preceding distribution. The degeneration is inevitable.

serial progress rates[3] of the algorithm described above on the 5, 10 and 20-dimensional fitness function hypersphere, $\mathbf{x} \mapsto \sqrt{\sum_i x_i^2}$. The simulations are carried out with $\mu = 0.5N^2; N^2; 2N^2$. The population size is always chosen to be $\lambda = 4\mu$.

For values ≤ 1 the mutation strength on the abscissa of figure 3 corresponds to the mixing ratio between mutated and parental distribution in the form $(1-x)\mathbf{C}_{\text{parental}} + x\mathbf{C}_{\text{mutated}}$, where $x \leq 1$ and $\mathbf{C}$ is the covariance matrix of the distribution. For mutation strength one the mutation operator with $m = N$ vectors is solely used as described above. For mutation strength zero no variation of the mutation distribution takes place. For mutation strength values larger one no mixing takes place and the N vectors are multiplied with $\exp(\mathcal{N}(0, \sigma^2))$ i.i.d. random factors with $\sigma = 10^{x-1} - 1 \in [0, 9]$ for $x \in [1, 2]$.

The diagrams show continuously decreasing serial progress rates with increasing variations of the mutation distribution (mutation strength). This results from increasing deviations of the mutation distribution from the optimal isotropic shape. With increasing mutation strength progress becomes zero because of a degenerating mutation distributions. This result implies an upper bound of the mutation strength. On the other side variations of the mutation distribution should be as large as possible, because the correct selection depends on differences between descendant distributions and the speed of the adaptation process is principally limited by the mutation strength. In the case of changing topology of the fitness function or poor initialization of the mutation distribution, small variations of the mutation distribution will lead to long adaptation periods. This turns out to be a decisive factor with respect to the applicability of adaptation

[3]The term *serial* progress rate indicates the progress rate per generation divided by the number of descendants.

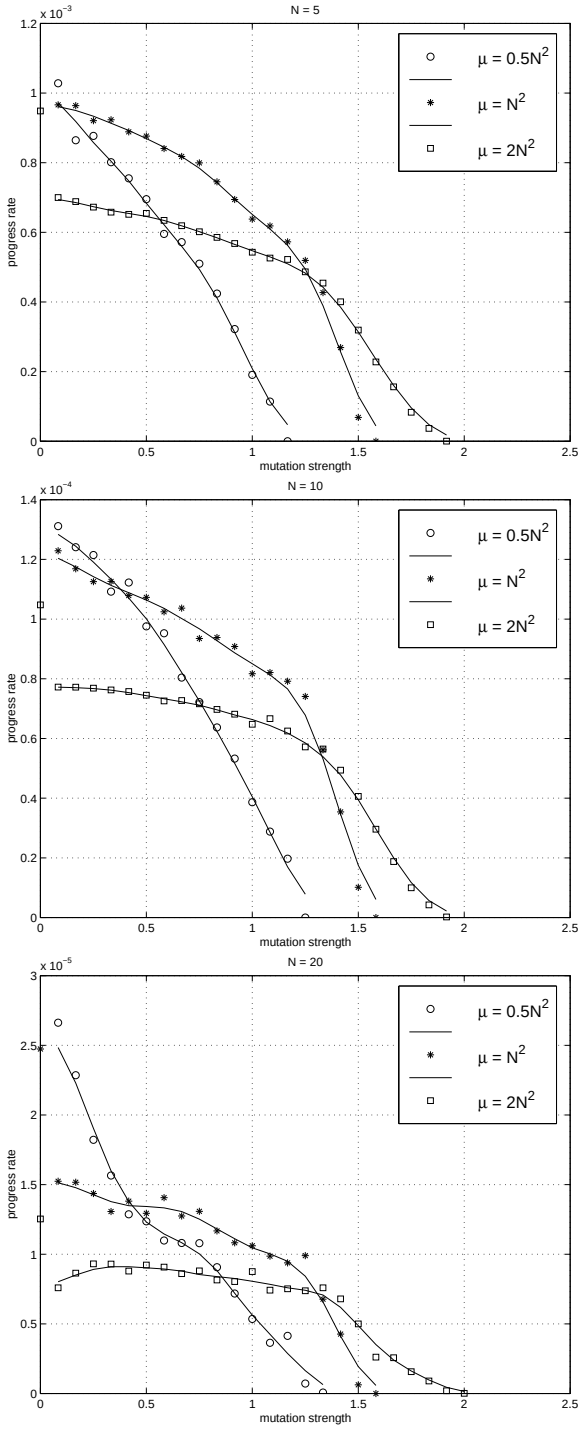

Figure 3: Serial progress rate on the hyper-sphere: $\mathrm{E}\left[\log(R^{(g-1)})/R^{(g)}\right]/\lambda$, where R denotes the distance to optimum, verse mutation strength. For mutation strength zero the serial progress rate is almost proportional to λ^{-1} (because of $\lambda \propto \mu \gg N$). The symbols indicate simulation results. The lines are smoothed interpolations.

algorithms.

According to figure 3 the mutation strength can be chosen 0.25 for $\mu = 0.5N^2$, one for $\mu = N^2$ and 1.5 for $\mu = 2N^2$ without substantially reducing the (serial) progress rate on the hyper-sphere. This choice leads to a progress reduction by a factor 0.6 compared to using very small variations of the distribution and by a factor 0.4 compared to using the optimal isotropic mutation distribution.

The choice of the population size $\lambda = 4\mu$ represents a comparable problem as the mutation strength: Small populations ($\mu = 0.5N^2$) facilitate larger serial progress rates for the constant topology of the hyper-sphere – but only without or with little variations of the mutation distribution. Large populations lead to smaller serial progress rates[4], but allow larger variations of the mutation distribution. **Figure 4** shows optimizations on the hyper-sphere with population sizes $\mu = 0.5N^2$, $\mu = N^2$ and $\mu = 2N^2$. The mutation strength (variation of the mutation distribution) is chosen according to the population size 0.25, 1 and 1.5 (see above). A very bad initialization of the mutation distribution with condition 10^6 (axis ratio 10^3) leads to a distinct adaptation phase. The diagrams show the number of generations on the abscissa. As expected, the adaptation phase reduces with increasing mutation strength. But taking the different population sizes into account the adaptation phase takes about

μ	feval
$0.5N^2$	$4.4 \cdot 10^6$
N^2	$1.0 \cdot 10^6$
$2N^2$	$1.4 \cdot 10^6$

With respect to function evaluations the medium population size $\mu = N^2$ facilitates the fastest adaptation.

The results of this section suggest to choose $\mu = N^2$ and mutation strength one. Mutation strength one denotes the "natural" mutation strength of the introduced mutation operator. That is: A variated mutation distribution is generated by N random vector realizations of the parental mutation distribution.

6 Testing Invariance and Scaling

The decisive aspect formulating the self-adaptation of arbitrary normal mutation distributions in the concept of MSC, introduced in this paper, is the invariance of

[4]Theoretical results (Beyer 1996) that the serial progress rate should not decrease with increasing population size are only valid for λ (considerably) smaller N and do not hold here.

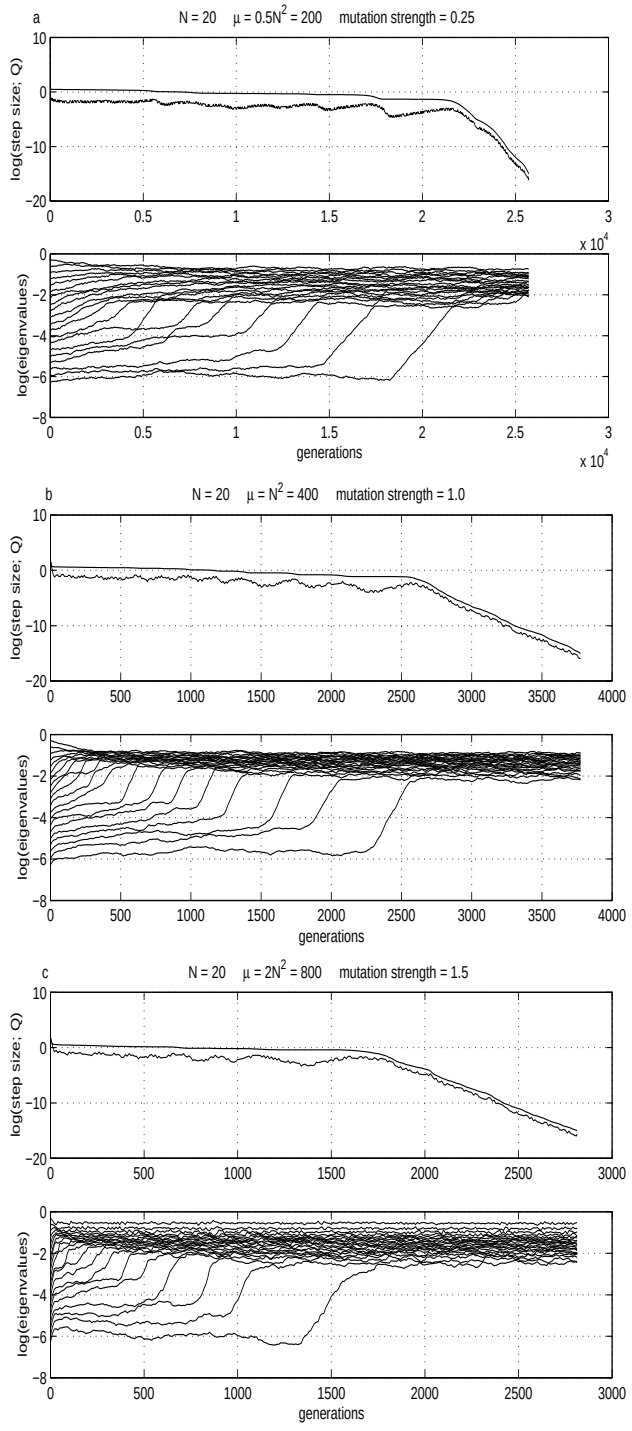

Figure 4: Simulation on the hyper-sphere fitness function with a "wrong" start distribution (condition 10^6) for $\mu = 0.5N^2$ (above), N^2 (middle) and $2N^2$ (below). Shown are fitness ('Q'), step-size (crinkled graph) and the variances of the principle components of the mutation distribution ellipsoid ('eigenvalues').

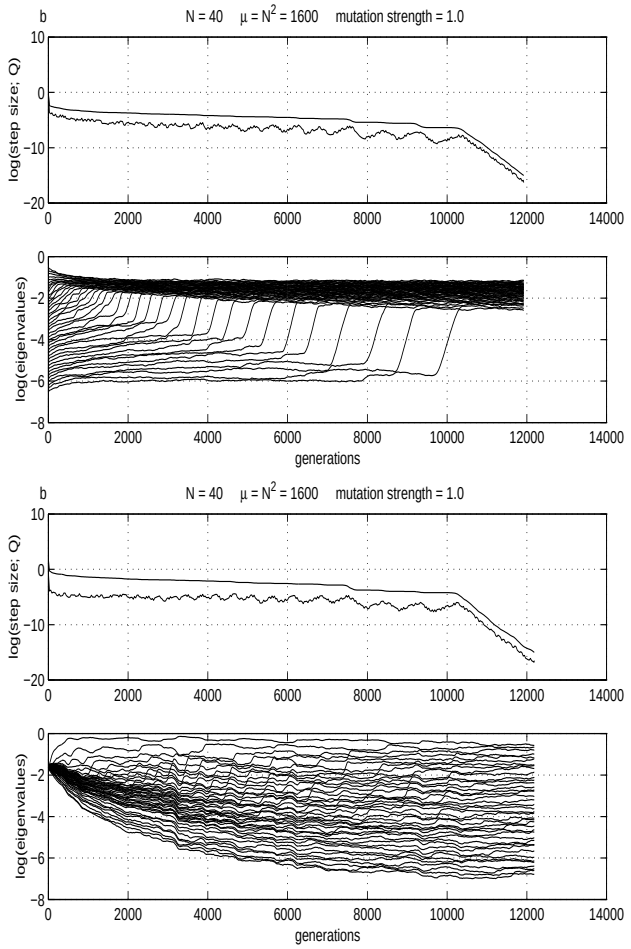

Figure 5: Simulation on the hyper-sphere with a wrong start distribution (above) and the same simulation with a linear transformation of both object parameter space and start distribution (below), which transforms the start distribution into an isotropic one. Shown are fitness ('Q'), step-size (crinkled graph) and the variances of the principle components of the mutation distribution ellipsoid ('eigenvalues').

the algorithm towards linear transformations. This feature is ensured by the coordinate system independent formulation of the mutation operator. **Figure 5** gives an example of this invariance: Figure 5a shows an optimization on the hyper-sphere using a badly scaled initialization of the mutation distribution (condition 10^6). Transforming the fitness function and the initialization of the mutation distribution identically, leads to the situation shown in Figure 5b. Here an hyper-ellipsoid, $\mathbf{x} \mapsto \sqrt{\sum_{i=1}^{N} 10^{6\frac{i-1}{N-1}} x_i^2}$, is optimized using an isotropic initialization of the mutation distribution. The graphs of fitness and general step size are identical apart from stochastic effects. Eigenval-

ues and object parameters differ by the applied linear transformation. Arbitrary rotations of the fitness model do not influence the results.

The self-adaptation from an isotropic mutation sphere to an ellipsoid with condition 10^6, as shown in figure 5 for $N = 40$, has been carried out for dimensions 5, 10, 20, 40. The adaptation takes about 150, 600, 2500, 10000 generations respectively. The results clearly point out a quadratic dependency of the adaptation time (in generations) from the problem dimension. Taking into account, that the necessary population size also scales quadratically with the dimension N, the number of function evaluations needed for the self-adaptation of an arbitrary mutation ellipsoid here scales with N^4.

7 Conclusion

In this paper the concept of mutative strategy parameter control (MSC) is applied to the adaptation of all parameters of a N-dimensional normal mutation distribution with zero mean. As shown in simulations the presented algorithm properly rescales a linear transformation of the hyper-sphere fitness function yielding optimal progress after an adaptation phase. Decisive is the invariance of the new algorithm towards arbitrary linear transformations of the object parameter space (apart from initialization). The practical worth of the algorithm is limited because of its bad scaling with the problem dimension. In comparison, the covariance matrix adaptation (Hansen and Ostermeier 1996; Hansen and Ostermeier 1997), which utilizes the derandomized adaptation paradigm instead of MSC, yields comparable invariance qualities. The necessary function evaluations roughly scale with N^2 (or even better) there compared to N^4 for the algorithm introduced here. The comparison supports the observation that derandomization can be a very efficient mechanism to speed up self-adaptation. The advantage is mainly due to the small population size, which can be chosen independently of the problem size N in derandomized schemes. The other way around, the disadvantage of MSC is due to the fact that a large population size does not reduce the adaptation time significantly – while scaling of μ with the number of strategy parameters to be adapted seems to be inevitable for MSC. To conclude we answer the two questions raised at the beginning:

1. Self-adaptation of arbitrary normal mutation distributions within the concept of MSC can be done successfully and with invariance towards linear transformations of the object parameter space.
2. The concept of MSC cannot compete with the derandomized approach, mainly because the parent number μ must scale quadratically with the problem size.

Acknowledgments

This work was supported by the *Deutsche Forschungsgemeinschaft* under grant Re 215/12-1.

References

Beyer, H.-G. (1996). On the asymptotic behavior of multirecombinant evolution strategies. In H.-M. Voigt, W. Ebeling, I. Rechenberg and H.-P. Schwefel (Eds.), *Parallel Problem Solving from Nature—PPSN IV, Proceedings*, Berlin, pp. 122–133. Springer.

Hansen, N. and A. Ostermeier (1996). Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Proceedings of the 1996 IEEE International Conference on Evolutionary Computation*, pp. 312–317.

Hansen, N. and A. Ostermeier (1997). Convergence properties of evolution strategies with the derandomized covariance matrix adaptation: The $(\mu/\mu_{\mathrm{I}}, \lambda)$-CMA-ES. In *EUFIT'97, 5th Europ. Congr. on Intelligent Techniques and Soft Computing, Proceedings*, Aachen, pp. 650–654. Verlag Mainz, Wissenschaftsverlag.

Ostermeier, A., A. Gawelczyk and N. Hansen (1994). 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.

Rechenberg, I. (1994). *Evolutionsstrategie '94.* Stuttgart: frommann-holzboog.

Rudolph, G. (1992). On correlated mutations in evolution strategies. In R. Männer and B. Manderick (Eds.), *Parallel Problem Solving from Nature, 2, Proceedings*, Brüssel, pp. 105–114. North-Holland.

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

Schwefel, H.-P. (1995). *Evolution and Optimum Seeking.* Sixth-Generation Computer Technology Series. New York: John Wiley & Sons Inc.