Inter- and Intragenerational Mutation Shape Adaptation

Stefan Berlik, Madjid Fathi, and Alexander Holland
Department of Electrical Engineering and Computer Science
Siegen University
57068 Siegen, Germany
{berlik, fathi, alex}@informatik.uni-siegen.de

Abstract. *Till now, only uncorrelated evolution strategies benefit from the strength of the directed mutation principle. It is thus a natural idea to enhance more powerful correlated evolution strategies with directed mutation, too. This work aims at porting this approach from the uncorrelated setting in classical evolution strategies to the correlated case as given in covariance matrix adaptation-evolution strategies. The main problem to be addressed here is the shape vector update. The shape vector controls the distribution's skewness and can be updated intergenerationally as well as intragenerationally. Starting with an analogue to the intergenerational parameter update mechanics used in CMA-ES, we argue that an additional intragenerational update is of greater benefit. An appropriate heuristic will be presented and some experimental data of several test functions is provided.*

Keywords: Intragenerational adaptation, intergenerational adaptation, directed covariance matrix adaptation, directed mutation, evolutionary algorithm

1 Introduction

Since the early work of Rechenberg [10] and Schwefel [12] the design of mutation operators turned out to be one of the most critical points in Evolution Strategies (ESs hereafter). These early works relied on just one single mutation strength, i.e. step-size, for all problem dimensions (isotropic mutation) and were concerned mainly with determining the optimal step-size for a faster search. To put it in a more general light, the covariance matrix of the mutation operator's distribution was considered to be the identity matrix. Soon Schwefel extended this approach and proposed to self-adapt one step-size per variable, i.e. to use a diagonal covariance matrix with positive entries. Consequently, as the most general case, he later suggested self-adapting of the whole covariance matrix (correlated mutation). A more detailed review of the field's history is given e.g. by Bäck et al. [3, 2].

However, all of these methods rely on normally distributed mutations and relatively little effort has been put into examining different distributions as mutation operators. One such example is the so-called Fast Evolution Strategy by Yao [14], where a Cauchy distribution is proposed as mutation operator. Nevertheless, Rudolph [11] later proofed that the order of local convergence is identical to that of normal mutations. Just to exchange the mutation distribution seems in general to be a questionable idea. The intention of directed mutation on the other hand is to introduce a different mutation *principle*. It will abandon the *random mutation hypothesis* - a fundamental tenet postulating that mutations occur at random, regardless of fitness consequences to the resulting offspring. This seems to be justified by the fact that the ES knows its optimization history and is thus able to extrapolate the evolution path to some extent. Under the assumption of a local similar objective function it is obviously reasonable to generate a bigger portion of offspring along the successful path.

2 The DCMA-ES Algorithm

So far directed mutation was realized with uncorrelated mutation models only. However, its usefulness has been shown for both, for test function optimization [4, 5] as well as in real-world scenarios [6]. Now, as already mentioned in the introduction, there are several even more powerful ES approaches that rely on the flexibility of correlated mutations. The performance of these EA depends obviously highly on the choice of the covariance matrix C, which has to be adjusted not only to the problem at hand, but also to the current state of the evolution process. Several methods have been proposed, from the self-adaptation of the mutation parameters in ES (SA-ES) [13] to the Covariance Matrix Adaptation-ES (CMA-ES) [9]. While the first removes the need to manually adjust the covariance matrix, the latter takes into account the history of evolution and deterministically adapts the covariance matrix from the last moves of the algorithm, thereby directing the search to use the most recent descent direction. In [8] an advanced version of the CMA-ES is presented, that is computationally more efficient.

All these approaches use symmetric normally distributed random numbers. The aim of the sequel is therefore to accommodate the CMA-ES with a multivariate skew-normal distribution, yielding the Directed Covariance Matrix Adaptation-ES (DCMA-ES). Especially the update of the distribution's shape vector will be investigated. While in classical ES with multi-recombination of less than the whole population an intergenerational update is inevitable, the CMA-ES setting allows to establish an intragenerational tuning of the shape vector. Here one distinct mean, i.e. the center of the parental population, is calculated and used subsequently to generate all descendants. This fixed point can be treated as reference for the whole offspring generation cycle. Recent studies have shown remarkable results. However, much further research is necessary and the results are in that sense preliminary.

A conceptually related approach, called LS-CMA-ES, was presented by Auger et al. [1]. It is of second-order and based on quasi-Newton techniques, i.e. relying on local curvature information to find out the next points to sample. Therefor it aims at learning the local Hessian matrix by solving a linear least-square minimization problem. The solution is then found by evaluating the pseudo-inverse of this linear system. The cost of the direct computation of this pseudo-inverse by standard numerical methods is scaling as n^6. In contrast the DCMA-ES is of zeroth-order (i.e. no derivatives are used) and computationally by far less expensive.

2.1 Shape Vector Control

As an ad hoc implementation we use the mechanics of step-size adaptation to adjust the shape vector. Shape control then reads

$$\boldsymbol{p}_\alpha^{(g+1)} = (1 - c_\alpha)\boldsymbol{p}_\alpha^{(g)} + \sqrt{c_\alpha(2 - c_\alpha)\mu_{\mathrm{eff}}}\,\boldsymbol{C}^{(g)^{-\frac{1}{2}}}\,\frac{\boldsymbol{m}^{(g+1)} - \boldsymbol{m}^{(g)}}{\alpha^{(g)}} \tag{1}$$

with learning rate

$$c_\alpha = \frac{\mu_{\mathrm{eff}} + 2}{n + \mu_{\mathrm{eff}} + 3} \tag{2}$$

and all other constants as given by Hansen and Kern [7]. However, although the learning rate has been altered over the whole $[0, \ldots, 1]$ range, no satisfying results have been obtained during the test runs. Therefore the *intergenerational* shape update is supplemented by an *intragenerational* one. We track the fitness of the generated offspring within every generation and adapt the shape vector accordingly.

One appropriate heuristic is given as follows: calculate the normalized direction vector from the mean of the current distribution to the actual offspring. If the fitness of this offspring is better than the mean fitness, then factor the direction vector into the shape vector. Otherwise take the opposite direction. Additionally, the fitness ratio is weighted exponentially and with the dimension. The definition of the update vector thus reads

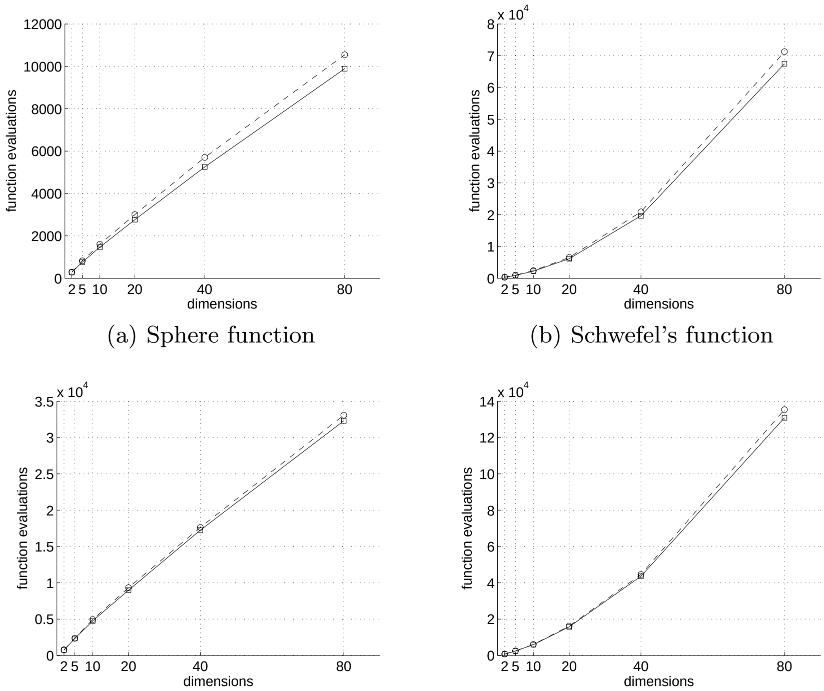
$$\boldsymbol{u} = n \exp\left(\frac{\mathrm{fit}\,\boldsymbol{x}}{\mathrm{fit}\,\boldsymbol{x}_{\mathrm{mean}}}\right)\frac{\boldsymbol{x} - \boldsymbol{x}_{\mathrm{mean}}}{|\boldsymbol{x} - \boldsymbol{x}_{\mathrm{mean}}|} \tag{3}$$

where n is the dimension, $\boldsymbol{x}$ and $\boldsymbol{x}_{\mathrm{mean}}$ are the actual individual and the mean used to generate the offspring, respectively, and the function $\mathrm{fit}(\cdot)$ gives the fitness of a sample. The intragenerational update of the shape vector $\boldsymbol{p}_\alpha^{[l]}$ depends on the actual individual's fitness.

$$\boldsymbol{p}_\alpha^{[l]} = \begin{cases} \boldsymbol{p}_\alpha^{[l-1]} + \boldsymbol{u} & \text{if } \text{fit}\, \boldsymbol{x} \succ \text{fit}\, \boldsymbol{x}_{\text{mean}} \\ \boldsymbol{p}_\alpha^{[l-1]} - \boldsymbol{u} & \text{else} \end{cases} \tag{4}$$

with $l \in [1, \ldots, \lambda]$ and $\boldsymbol{p}_\alpha^{[0]} = \boldsymbol{0}$.

Until now, neither the shape control nor the learning rate has been adapted satisfactorily to the special demands of shape vector control. However, even with this crude treatment of the shape vector some very promising results have been obtained.



(a) Sphere function

(b) Schwefel's function

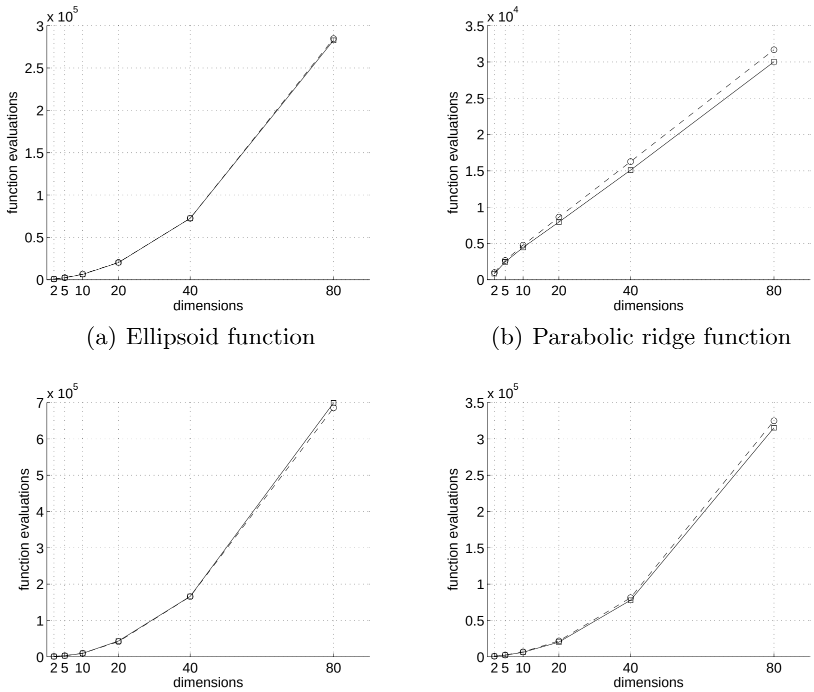
(c) Cigar function

(d) Tablet function

Fig. 1. Number of function evaluations versus the problem dimension for the functions f_{sphere}, f_{Schwefel}, f_{cigar}, and f_{tablet}. The Directed-CMA is plotted with a solid line, the original CMA with a dashed one.

3 Experimental Results

Two different CMA-ESs are experimentally investigated: the original variant as described by Hansen and Kern [7], using $\text{N}(0, \boldsymbol{C})$ distributed random



(a) Ellipsoid function (b) Parabolic ridge function

(c) Sharp ridge function (d) Rosenbrock's function

Fig. 2. Number of function evaluations versus the problem dimension for the functions f_{elli}, f_{parabR}, f_{sharpR}, and f_{Rosen}. The Directed-CMA is plotted with a solid line, the original CMA with a dashed one.

vectors and the DCMA-ES, using instead $\mathrm{SN_n}(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\alpha})$ distributed random vectors with intra- and intergenerational shape update. For the comparison of the two functions, a test suite consisting of eight well known functions is used. Initial values are set to $\boldsymbol{x}^{(0)} \in [-1, 1]^n$, $\sigma^{(0)} = 1$, and $\alpha^{(0)} = \mathbf{0}$ for all functions except for Rosenbrock's case where $\boldsymbol{x}^{(0)} = \mathbf{0}$, $\sigma^{(0)} = 0.1$, and again $\alpha^{(0)} = \mathbf{0}$. As stopping criterion for all functions but f_{parabR} and f_{sharpR} fitness better than 10^{-10} is demanded, for the two others fitness has to be less than -10^{10}. All functions, except for f_{sphere}, are highly nonseparable. Tests are carried out in $n = [2, 5, 10, 20, 40, 80]$ dimensions and for offspring numbers $\lambda = 4 + \lfloor 3 \log n \rfloor$ with parent numbers $\mu = \lceil \lambda/2 \rceil$. For each combination 25 runs are done. Depicted in Figures 1 and 2 are the results of each case, the corresponding numerical values are omitted for space reasons.

3.1 Discussion of the Results

The runs on f_{sphere} show a DCMA-ES that outperforms the CMA-ES relative constantly at about 8% for all but $n = 2$. On f_{Schwefel} it performs approximately 6% better for all dimensions, while the gain on f_{cigar} and f_{tablet} is only about 3%, and greater than 2%, repectively. On the function f_{elli} the outcome is somewhat irregular. The result is significant only for 5 and 10 dimensions, where the DCMA-ES is slightly superior. Interesting is the 2-dimensional case on f_{parabR}, where the DCMA-ES is 19% better, besides about 8%. The high performance of the DCMA-ES in low dimensions is also true on f_{sharpR}. But here its performance is rapidly decreasing with increasing dimensions. For $n > 10$ it is outperformed by the CMA-ES. The same tendency can be seen on . Here the gain of the DCMA-ES decreases from 10% for $n = 2$ to 3% for $n = 80$. In general, the DCMA-ES performs better on all functions except on f_{sharpR}. On average, there is a gain of a few percentage points. This must be seen against the background of the CMA-ES considered already as state-of-the-art in parameter optimization, the preliminary design of shape vector control, and the very small overhead caused by directed mutation. In fact, all that has to be done is to calculate one n-dimensional scalar product, generate one univariate random number, and do one comparison. Compared to a function evaluation in a real world application this can rather be neglected.

4 Conclusions

With directed mutation a promising new mutation principle for uncorrelated EAs has been presented, that has now also been ported to the correlated setting in a CMA-ES context. First results with the DCMA-ES left us optimistic about the potential of this approach. It is a zeroth-order algorithm causing only very small overhead. Regarding the presented results, it has to be kept in mind that intergenerational adaptation of the shape vector is still amendable, intragenerational adaptation is in an inadequate state, and the learning rate has to be further investigated. Thus, much work is left to be done to tune the DCMA-ES. Finally, more comprehensive experimental studies which should also include multi-objective problems are undone.

References

1. A. Auger, M. Schoenauer, and N. Vanhaecke. LS-CMA-ES: A second-order algorithm for covariance matrix adaptation. In *Parallel Problem Solving from Nature*, volume 3242 of *Lecture Notes in Computer Science*, pages 182–191. Springer-Verlag, Berlin Heidelberg, 2004.
2. T. Bäck, U. Hammel, and H.-P. Schwefel. Evolutionary Computation: Comments on the History and Current State. *IEEE Transactions on Evolutionary Computation*, 1(1):3–17, 1997.
3. T. Bäck, F. Hoffmeister, and H.-P. Schwefel. A survey of evolution strategies. pages 2–9, 1991.
4. S. Berlik. A step size preserving directed mutation operator. In K. Deb, editor, *Proc. of the Int. Genetic and Evolutionary Computation Conf., GECCO 2004,*

Jun. 26-30, 2004, Seattle, USA, volume 1 of *Lecture Notes in Computer Science*. Springer-Verlag Berlin Heidelberg, 2004.
5. S. Berlik. Directed mutation by means of the skew-normal distribution. In B. Reusch, editor, *Proc. of the Int. Conf. on Computational Intelligence, FUZZY DAYS, Sep. 29 - Oct. 1, 2004, Dortmund, Germany*, Advances in Soft Computing, pages 35–50. Springer-Verlag Berlin Heidelberg, 2005.
6. S. Berlik and M. Fathi. Multi-objective optimization using directed mutation. In H. R. Arabnia and R. Joshua, editors, *Proc. of the Int. Conf. on Artificial Intelligence, ICAI 2005, June 27-30, 2005, Las Vegas, USA*, volume II, pages 870–875. CSREA Press, USA, 2005.
7. N. Hansen and S. Kern. Evaluating the CMA Evolution Strategy on Multimodal Test Functions. In X. Yao, E. K. Burke, J. A. Lozano, J. Smith, J. J. M. Guervós, J. A. Bullinaria, J. E. Rowe, P. Tiño, A. Kabán, and H.-P. Schwefel, editors, *Parallel Problem Solving from Nature*, volume 3242 of *Lecture Notes in Computer Science*, pages 282–291. Springer-Verlag Berlin Heidelberg, 2004.
8. N. Hansen, S. D. Müller, and P. Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary Computation*, 11(1):1–18, 2003.
9. N. Hansen and A. Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *IEEE International Conference on Evolutionary Computation*, pages 312–317, 1996.
10. I. Rechenberg. *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog Verlag, Stuttgart, 1973.
11. G. Rudolph. Local convergence rates of simple evolutionary algorithms with Cauchy mutations. *IEEE Transactions on Evolutionary Computation*, 1(4):249–258, 1998.
12. H.-P. Schwefel. Adaptive Mechanismen in der biologischen Evolution und ihr Einfluß auf die Evolutionsgeschwindigkeit. Technical report, Technical University of Berlin, 1974.
13. H.-P. Schwefel. *Evolution and Optimum Seeking*. John Wiley & Sons, New York, 1995.
14. X. Yao and Y. Liu. Fast Evolution Strategies. *Control and Cybernetics*, 26(3):467–496, 1997.