The Correlation-Triggered Adaptive Variance Scaling IDEA

Jörn Grahl
Mannheim Business School
Dept. of Logistics
68131 Mannheim, Germany
joern.grahl@bwl.uni-mannheim.de

Peter A.N. Bosman
Centre for Mathematics and Computer Science
P.O. Box 94079
1090 GB Amsterdam
The Netherlands
Peter.Bosman@cwi.nl

Franz Rothlauf
Mannheim Business School
Dept. of Business Administration
and Information Systems
68131 Mannheim, Germany
rothlauf@uni-mannheim.de

ABSTRACT

It has previously been shown analytically and experimentally that continuous Estimation of Distribution Algorithms (EDAs) based on the normal pdf can easily suffer from premature convergence. This paper takes a principled first step towards solving this problem. First, prerequisites for the successful use of search distributions in EDAs are presented. Then, an adaptive variance scaling theme is introduced that aims at reducing the risk of premature convergence. Integrating the scheme into the iterated density–estimation evolutionary algorithm (IDEA) yields the correlation-triggered adaptive variance scaling IDEA (CT-AVS-IDEA). The CT-AVS-IDEA is compared to the original IDEA and the Evolution Strategy with Covariance Matrix Adaptation (CMA-ES) on a wide range of unimodal test-problems by means of a scalability analysis. It is found that the average number of fitness evaluations grows subquadratically with the dimensionality, competitively with the CMA-ES. In addition, CT-AVS-IDEA is indeed found to enlarge the class of problems that continuous EDAs can solve reliably.

Categories and Subject Descriptors

G.1.6 [**Numerical Analysis**]: Optimization—*Gradient methods*; I.2 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search

General Terms

Algorithms, Optimization, Performance, Scalability

Keywords

Evolutionary Algorithms, Estimation of Distribution Algorithms, Numerical Optimization, Adaptive Variance Scaling

1. INTRODUCTION

This paper is in line with recent work and ongoing discussion on strengths and limitations of continuous Estimation of Distribution Algorithms for real-valued function optimization. The probabilistic models used in continuous EDAs are often based on the normal pdf [1, 4, 6, 7, 9, 16, 17, 22, 23]. A major drawback of this approach is that, without precaution, the variance of the normal pdf decreases fast on slope-like regions of the search space, likely causing premature convergence against sub-optimal solutions. This drawback has been noticed experimentally [7] and has of late been proved theoretically [11].

Recently, studies have been carried out that indicate that the problems mentioned above may yet be coped with. Yuan and Gallagher [25] showed in an initial investigation that by artificially keeping the variance at a value of at least 1, certain problems could be solved by a continuous EDA that were previously intractable. Ocenasek *et al.* [20] used a self-adaptation approach adopted from evolution strategies to scale the normal kernels.

In this paper we discuss a scheme to solve the problem of premature convergence. First, it is assessed which requirements a probability distribution has to meet in order to function properly as a search distribution in EDAs. Second, these findings are exploited to develop a correlation-triggered adaptive variance scaling scheme that helps reducing the risk of premature convergence of continuous EDAs based on the normal pdf. The normal pdf is simple in its nature and its use in an EDA is well understood. We are therefore able to identify the exact problem at hand and provide a proper, well-tailored remedy. This remedy is integrated into the iterated density–estimation evolutionary algorithm (IDEA, see [6]), yielding the correlation-triggered adaptive variance scaling IDEA (CT-AVS-IDEA). To validate the applicability of the approach and to gain insights into the running time complexity of the algorithm, we investigate the scale-up behavior of CT-AVS-IDEA. Such a scale–up analysis of (variance–enhanced) continuous EDAs is novel in itself. The results are compared to those of both the IDEA without variance adaptation and the Evolution Strategy with Covariance Matrix Adaptation (CMA-ES, see [12, 13]) on a test bed of unimodal test-problems. The experimental results indicate that for all regarded algorithms the number of fitness evaluations that is required to reliably solve the problems grows subquadratically with respect to the dimensionality of the problems. However, CT-AVS-IDEA is capable of solving all problems, even in high dimensions, whereas the IDEA without variance scaling fails on some of these problems. The integration of adaptive vari-

GECCO '06, July 8–12, 2006, Seattle, Washington, USA

ance scaling thus enlarges the class of problems that continuous EDA can solve reliably and efficiently.

The remainder of this paper is organized as follows. Section 2 briefly summarizes results regarding convergence properties of continuous EDAs that have been obtained so far. In this Section it is also assessed which requirements a probability distribution has to meet in order to function properly as a search distribution in an EDA. Next, CT-AVS-IDEA is proposed in Section 3 and experimental results are obtained and interpreted in Section 4. The paper ends with concluding remarks in section 5.

2. PREMISES FOR SUCCESSFUL CONTINUOUS EDAS

2.1 A brief introduction into EDAs

Estimation of Distribution Algorithms [19] are descendants of Evolutionary Algorithms (EAs). Similar to EAs, EDAs are stochastic search methods that maintain a set of candidate solutions, called the population, throughout the search. Each individual has an associated fitness value that measures its quality. An individual consists of a genotype that is its genetic encoding, and a phenotype, that is the actual solution to the optimization problem at hand. Whereas the quality of the individuals is measured on basis of the phenotypes, new candidate solutions are constructed on basis of the genotypes. The goal of the EDA is to find the individual of highest quality.

Usually, the initial population is filled with randomly generated solutions. All individuals are evaluated and the better solutions are selected using a selection scheme (see [3]). Selection pushes the EDA into promising regions of the search space. What differentiates EDAs from other optimizers is that they now explicitly learn a density estimate from the genotypes of the selected individuals. Then, an EDA performs induction on the set of selected solutions by randomly sampling the density estimate. Thereby, new candidate solutions are generated. The new candidate solutions replace the old population partly or as a whole, advancing it to the next generation. EDAs execute an iterative process of evaluation, selection, model building, model sampling and replacement. This process is stopped when a predefined stopping criterion is met, like the convergence of the whole population against a single solution.

For comprehensive overviews on EDA instances, we refer the interested reader to the literature [8, 17, 21]. In this paper, we focus on real-valued, continuous EDAs for numerical optimization where both the genotype and the phenotype are continuous, real-valued vectors.

Continuous EDAs mostly use the normal pdf as the basis of their probabilistic model because the normal pdf is a commonly–used and computationally tractable approach to estimating probability distributions in continuous spaces. The normal pdf $P^{\mathcal{N}}_{(\boldsymbol{\mu},\boldsymbol{\Sigma})}$ for an $n-$ dimensional random variable X is parameterized by a vector $\boldsymbol{\mu}$ of means and a symmetric covariance matrix $\boldsymbol{\Sigma}$ and is defined by

$$P^{\mathcal{N}}_{(\boldsymbol{\mu},\boldsymbol{\Sigma})}(X)(\boldsymbol{x}) = \frac{(2\pi)^{-\frac{|\boldsymbol{n}|}{2}}}{(\det \boldsymbol{\Sigma})^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{\Sigma})^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \quad (1)$$

The number of parameters to be estimated from data to fit the normal distribution to selected individuals equals $\frac{1}{2}|\boldsymbol{n}|^2 + \frac{3}{2}|\boldsymbol{n}|$. A maximum likelihood estimation for the normal pdf is obtained from a vector $\boldsymbol{\mathcal{S}}$ of samples if the parameters are estimated by the sample average and the sample covariance matrix [2, 24]:

$$\hat{\boldsymbol{\mu}} = \frac{1}{|\boldsymbol{\mathcal{S}}|}\sum_{j=0}^{|\boldsymbol{\mathcal{S}}|-1} \boldsymbol{\mathcal{S}}_j, \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{|\boldsymbol{\mathcal{S}}|}\sum_{j=0}^{|\boldsymbol{\mathcal{S}}|-1} (\boldsymbol{\mathcal{S}}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{\mathcal{S}}_j - \hat{\boldsymbol{\mu}})^T \quad (2)$$

On the basis of the normal pdf, different probabilistic models can be estimated from the selected individuals in EDAs, e.g., Bayesian factorizations [4, 6], or mixtures of normal pdfs [1, 7]. Since the number of parameters to be estimated grows quadratically with $|\boldsymbol{n}|$, estimating factorizations based on the normal pdf is relatively fast and efficient.

2.2 Limitations of UMDAc

Of late, there has been an ongoing discussion on the limitations of continuous EDAs based on the normal pdf in solving numerical optimization problems.

Analytical results on the convergence properties of the univariate marginal distribution algorithm in the continuous domain (UMDAc, see [16]) are available in [10] and [11]. UMDAc is a simple real-valued EDA that uses a univariate factorization of the normal density as a fixed-structure probabilistic model. The analysis of UMDAc revealed important peculiarities of continuous EDAs based on the normal pdf.

To be more precise, the performance of UMDAc depends heavily on the structure of the area of the fitness landscape that UMDAc is currently exploring. Continuous search spaces can be seen as arrangements of two elemental structures: peaks and slopes. At the beginning of the search, the EDA will in general be approaching a local or global optimum by exploring a region that has a slope-like function. Eventually the search focuses around an optimum (either local or global) in its final phases, i.e. the region to be explored is then is shaped like a peak.

It has been shown that UMDAc can only reach the optimum if the set of search points is close to the optimum [10, 11]. The reason for this is that the mean of the normal distribution that is estimated by UMDAc can only move a limited distance before converging due to shrinking of the estimated variance. This means that on slope-parts of the search space, UMDAc will perform extremely poorly whereas on peak-parts UMDAc will perform nicely. Both studies assume that UMDAc uses the estimated normal density to generate new candidate solutions with no modification whatsoever.

UMDAc is a simple EDA with a univariately factorized probabilistic model. It was found that using more flexible probabilistic models does not help to solve this problem. In fact, current continuous EDAs fail on some standard numerical optimization test problems where other continuous EAs or even classical gradient-based algorithms succeed. This was first noticed in [7] and confirmed in [15] and [25].

2.3 Elementary requirements for search distributions in EDAs

The limited success of continuous EDAs that directly sample new candidate solutions from probabilistic models based on the normal pdf raises (at least) two important questions:

1. Which properties render a probability distribution a good choice for use as a search distribution in an EDA?
2. Is the normal pdf, based on these properties, a reasonably good choice for use in continuous EDAs?

Generally speaking, the inductive bias of any search strategy has to fit the structure of the problem it attempts to solve. This has implications for the choice of search distributions in EDAs.

Estimating the contours of the fitness landscapes on the basis of a probability distribution, as done in any EDA, results in a probabilistic representation of the true fitness landscape. The induced bias of an EDA is based on this internal probabilistic representation. The restrictions of the model used and the estimation procedure however cause the representation to only be an *approximation* of the optimal distribution; the latter being a close representation of the contours of the fitness landscape. To be more specific, a probability distribution has to meet two essential requirements in order to function properly as a search-distribution in EDAs [5]:

1. The probability-distribution class must be *adequate*.
2. The estimation procedure must be *competent*.

A class of probability distributions is considered *adequate* with respect to a given optimization problem, if it is able to closely model the contours of the fitness function of that problem with arbitrary exactness. A class of probability distributions is considered *inadequate* with respect to a given optimization problem, if it is not able to model the contours of the fitness landscape of the problem without significant loss of precision. Then, the estimated probabilistic representation of the fitness landscape can be (but not necessarily has to be) misleading. This is especially the case, if it introduces additional basins of attraction. It should then carefully be assessed whether this density can be seen as a reliable source of information for guiding the search. However, this is not common practice in the current state of EDA research.

The density estimation procedure is considered *competent* if it is actually able to obtain an estimate for the probability distribution that closely models the fitness landscape and properly generalizes the sample set. This means that the probabilistic representation of the true landscape is correct. If enough and proper data are available, the estimation procedure should accurately model the fitness landscape.

We now assess briefly for the continuous problem domain, whether the normal pdf is competent and adequate for peaks and slopes; the two basic structures of continuous fitness landscapes.

- *Peaks:*
  The normal pdf can match contour-lines of a single peak nicely as it always concentrates the search around its mean and therefore can contract around a single peak with arbitrary exactness. If the search is initialized near the peak, selection can shift the mean of the pdf onto the peak. Thus, the normal pdf is adequate for search on a single peak. An estimation procedure based on the standard maximum-likelihood estimates is competent, because by using the maximum-likelihood estimates for the normal pdf, a properly generalizing estimate can be constructed from data in computationally tractable time. As a result, the UMDAc algorithm is able to converge on peaks, if it is initialized near it. This agrees with initial results on research into continuous EDAs [4, 17].
- *Slopes:*
  Things are different for slope-like regions of the search space. Contour-lines of slopes can not be matched with the normal pdf. The true structure is misrepresented using a maximum-likelihood estimation as the normal kernel introduces an additional basin of attraction around its mean. The probabilistic representation of the structure is different from the true structure. Estimates from the normal pdf are thus a much less reliable source of information for guiding the search compared to exploring a single peak. Relying the search on maximum-likelihood estimates of the normal pdf potentially misleads the EDA and can cause premature convergence on slope-like regions of the search space.

3. CORRELATION-TRIGGERED ADAPTIVE VARIANCE SCALING IDEA

3.1 Adaptive variance scaling

In order to solve the problem of premature convergence, a class of more involved probability distributions could theoretically be introduced for use as a search distribution in continuous EDAs. However, contours of continuous fitness landscapes can be of virtually any shape. As universal approximation in arbitrary exactness is computationally intractable, we develop a simple remedy that allows to use the normal pdf.

We propose a technique that modifies the estimation procedure of the normal pdf in continuous EDAs to make it more reliable when traversing a slope. The smaller the variance is in the estimated probability distribution, the smaller the area of exploration for the EDA. The variance in the normal pdf is explicitly stored in the covariance matrix $\mathbf{\Sigma}$. Hence, a straightforward manner to allow the EDA to increase the area of exploration and thereby increasing the probability of traversing a slope is to enlarge the variance beyond its maximum-likelihood estimate.

Therefore, an adaptive-variance-scaling coefficient c^{AVS} is maintained. Upon drawing new solutions from the probability distribution, the distribution is scaled by c^{AVS}. This means, that the covariance matrix used for sampling the normal pdf is $c^{\text{AVS}}\mathbf{\Sigma}$ instead of just $\mathbf{\Sigma}$. If the best fitness value improves in one generation, then the current size of the variance allows for progress. Hence, a further enlargement of the variance may allow further improvement in the next generation. To fight the variance diminishing effect of selection, the size of c^{AVS} is scaled by $\eta^{\text{INC}} > 1$. If on the other hand the best fitness does not improve, the range of exploration may be too large to be effective and the adaptive variance scaling coefficient should be decreased by a factor $\eta^{\text{DEC}} \in [0, 1]$. For symmetry, we set $\eta^{\text{INC}} = 1/\eta^{\text{DEC}}$.

We bound the magnitude of c^{AVS} from above by a predefined value $c^{\text{AVS-MAX}}$ and from below by a predefined value $c^{\text{AVS-MIN}}$. For symmetry, we set $c^{\text{AVS-MIN}} = 1/c^{\text{AVS-MAX}}$. Moreover, if $c^{\text{AVS}} < c^{\text{AVS-MIN}}$, we set c^{AVS} to $c^{\text{AVS-MAX}}$ in order to stimulate exploration.

3.2 Correlation triggering

In the above scheme, improving best fitness values automatically increase c^{AVS}. Improving fitness values however do not mean that the variance always needs to be enlarged. This is especially the case if the normal kernel is near the

optimum. In this case, the induced bias of the normal pdf already leads the EDA to the optimum. Increasing the variance will then only slow down convergence, as the EDA is forced to explore a larger area of the search space without necessity. We distinguish between two situations:

1. The EDA is traversing a slope (adaptive variance scaling is needed).
2. The EDA is searching around the optimum (adaptive variance scaling is not needed).

To identify which structure dominates in a generation, we exploit the relationship between normal density and the fitness of the selected solutions. If the selected solutions are clustered around an optimum, the density will strongly correlate with the fitness, as the normal density decreases if one moves away from the mean. This is desirable if the kernel surrounds a peak, as then better solutions are sampled with higher probability. If the selected solutions are spread on a slope, the density and fitness are not as strongly correlated with each other.

This motivates the triggering of adaptive variance scaling on the basis of a correlation coefficient r between the ranks of density and fitness. We use ranked correlation (see [14], pp. 338 and 400) because a larger density should be associated with a higher (lower) fitness in case of maximization (minimization). The exact form of the fitness landscape is less important. Assume now that we seek to minimize a cost function. We propose a threshold value θ^{corr} such that if $r \leq \theta^{\text{corr}}$, then the maximum-likelihood estimates are used in EDAs without modification. If $r > \theta^{\text{corr}}$, the adaptive variance scaling is used.

The principle of correlation-triggered adaptive variance scaling is EDA-independent. We integrated it into a continuous EDA based on Bayesian factorizations of normal pdfs, the iterated density- estimation evolutionary algorithm (IDEA, [6]). The resulting algorithm is called correlation-triggered adaptive variance scaling IDEA (CT-AVS-IDEA). Pseudocode for CT-AVS-IDEA is presented in Figure 1.

CT-AVS-IDEA(τ, n, η^{DEC}, $c^{\text{AVS-MAX}}$)

1. Set generation counter $t = 0$.
2. Initialize population $\mathcal{P}$ with n random individuals.
3. Assign $c^{\text{AVS-MIN}} = 1/c^{\text{AVS-MAX}}$.
4. Assign $\eta^{\text{INC}} = 1/\eta^{\text{DEC}}$.
5. Assign $c^{\text{AVS}} = 1$.
6. Evaluate solutions in $\mathcal{P}$.
7. Store best fitness found in $\mathcal{P}$ in b^t
8. Select best $\lfloor \tau n \rfloor$ individuals and store them in $\mathcal{S}$.
9. If $b^t = b^{t-1}$ then
   (a) assign $c^{\text{AVS}} = c^{\text{AVS}} \cdot \eta^{\text{DEC}}$.
   else
   (b) assign $c^{\text{AVS}} = c^{\text{AVS}} \cdot \eta^{\text{INC}}$.
10. If $c^{\text{AVS}} < c^{\text{AVS-MIN}}$ or $c^{\text{AVS}} > c^{\text{AVS-MAX}}$ then
    (a) assign $c^{\text{AVS}} = c^{\text{AVS-MAX}}$.
11. Estimate Bayesian factorization of normal pdf from $\mathcal{S}$.
12. Compute ranked correlation coefficient r.
13. If $r > \theta^{corr}$ then
    (a) Assign $\mathbf{\Sigma} = c^{\text{AVS}}\mathbf{\Sigma}$.
14. Sample $n - \lfloor \tau n \rfloor$ new candidate solutions from estimated normal pdf (with possibly scaled covariance matrix) and store new candidate solutions in $\mathcal{O}$.
15. Evaluate solutions in $\mathcal{O}$.
16. Replace worst $n - \lfloor \tau n \rfloor$ individuals in $\mathcal{P}$ with $\mathcal{O}$.
17. Update generation counter, i.e. assign $t = t + 1$.
18. If termination criterion is not met, go back to 7.

Figure 1: CT-AVS-IDEA pseudocode (minimization).

4. EXPERIMENTAL SECTION

4.1 Setup

We perform experiments on test functions listed in table 1 using CT-AVS-IDEA, the IDEA without adaptive variance scaling and the CMA-ES [12]. All functions are unimodal. The optimum for functions 1-7 is obtained by setting $x_i = 0$ for all i. For function 8 the optimum is obtained by setting $x_i = 1$ for all i. The optimum for functions 9 and 10 is obtained by setting $x_i = 0$ for all $i > 1$ and letting x_1 go to ∞. The initialization range used for all functions is $[-10, 5]$, i.e. asymmetric around the optimum and for functions 9 and 10 far away from the optimum for variable x_1.

Using a scalability analysis, the running time complexity of the algorithms is experimentally approximated. To be more specific, it is assessed how the total number of fitness evaluations e and the population size n required to solve the problems to optimality grows with the size of the problem l. Therefore, the dimensionality l was varied: $l \in \{2, 4, 8, 10, 20, 40, 80\}$. For each dimensionality we used a bisection method to obtain the minimally required population size for which the problem's value to reach was found in at least 95 out of 100 independent consecutive runs. The scalability analysis is important, as it allows us to predict whether CT-AVS-IDEA is a tractable approach for solving real-world problems that are often of much higher dimensionality.

For CT-AVS-IDEA we used $\eta^{\text{DEC}} = 0.9$, i.e. a small multiplication factor to allow for smooth adaptation of the variance multiplication factor. The correlation trigger threshold θ^{corr} was set to $\theta^{\text{corr}} = -0.55$ (see Section 4.2). The magnitude of c^{AVS} was bounded from above by $c^{\text{AVS-MAX}} = 10.0$. Following the rule of thumb from [18], the selection threshold τ was set to $\tau = 0.3$ for both CT-AVS-IDEA and the IDEA without variance adaptation.

4.2 Setting the correlation trigger threshold

In order to obtain a reasonable value for θ^{corr}, we tested when the ranked correlation coefficient between fitness and density actually triggers scaling of the variance on the sphere function. The sphere function is a single peak and can be solved by EDAs without variance scaling. We varied θ^{corr} from -1.0 to 1.0 in steps of 0.01. For each value of θ^{corr}, 100 independent runs of CT-AVS-IDEA on the sphere function in dimensionalities $l \in \{2, 4, 8, 10, 20, 40, 80\}$ were performed. Initial populations were drawn symmetrically around the optimal solution of 0 for all dimensions in a range of $[-7.5, 7.5]$. The population size that was used for a dimensionality l was equal to the minimally required population size for the IDEA to solve this problem optimally. In that case variance scaling is not required because the induced bias of the normal pdf itself suffices to locate the optimum.

Name	Definition	Value to reach
Sphere	$\sum_{i=1}^{l} x_i^2$	10^{-10}
Ellipsoid	$\sum_{i=1}^{l} 10^{6\frac{i-1}{l-1}} x_i^2$	10^{-10}
Cigar	$x_i^2 + \sum_{i=2}^{l} 10^6 x_i^2$	10^{-10}
Tablet	$10^6 x_1^2 + \sum_{i=2}^{l} x_i^2$	10^{-10}
Cigar Tablet	$x_1^2 + \sum_{i=2}^{l-1} 10^4 x_i^2 + 10^8 x_l^2$	10^{-10}
Two Axes	$\sum_{i=1}^{\lfloor l/2 \rfloor} 10^6 x_i^2 + \sum_{i=\lfloor l/2 \rfloor}^{n} x_i^2$	10^{-10}
Different Powers	$\sum_{i=1}^{l} \lvert x_i \rvert^{2+10\frac{i-1}{l-i}}$	10^{-15}
Rosenbrock	$\sum_{i=1}^{l-1} (100 \cdot (x_i^2 - x_{i+1})^2 + (x_i - 1)^2)$	10^{-10}
Parabolic Ridge	$-x_1 + 100 \sum_{i=2}^{l} x_i^2$	-10^{-10}
Sharp Ridge	$-x_1 + 100 \sqrt{\sum_{i=2}^{l} x_i^2}$	-10^{-10}

Table 1: Test functions and values to reach.

Figure 2 illustrates the percentage of generations in which variance scaling was nonetheless triggered (averaged over 100 runs). As a rule of thumb, we propose to set θ^{corr} to $\theta^{\text{corr}} = -0.55$. For this value, the number of unnecessary correlation triggers is rather constant and at most 25%. If a smaller value (i.e. closer to -1.0) is chosen, it can be seen from Figure 2 that the number of unnecessary correlation triggers will grow with increasing dimensionality. Although the value of -0.55 is rather robust, i.e. values between -0.6 and -0.4 lead to good results, the value for the correlation trigger should not become much larger. If a larger value (i.e. closer to 1.0) is chosen, the scaling of variances was observed from initial experimentation not to be triggered when it is required on slopes.

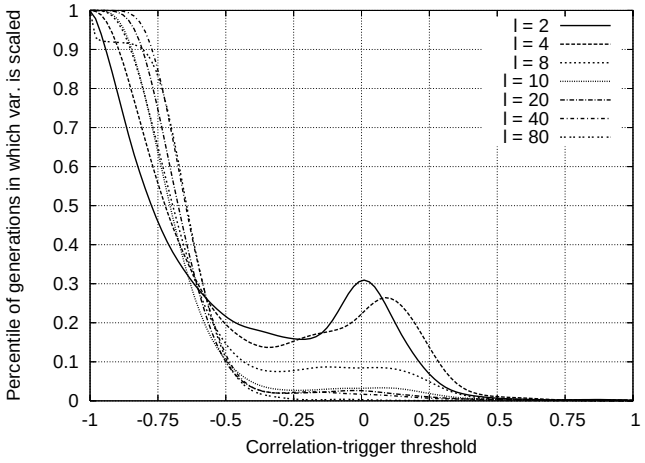


Figure 2: Correlation trigger thresholds.

4.3 Results and interpretation

AVS-IDEA, IDEA and CMA-ES

Plots that reveal the influence of problem dimensionality on the average number of evaluations and the minimal population size required to solve the problems are presented in figure 3. As the plots have a log-log scale, straight lines indicate polynomial scalability. Additionally, table 2 shows results from two linear least squares regressions on log-log-scaled data where the average number of evaluations e and the minimally required population size n depend on the dimensionality l of the problems as follows:

$$\log n = \log l^{\alpha} + \epsilon \quad \text{and} \quad \log e = \log l^{\beta} + \epsilon, \tag{3}$$

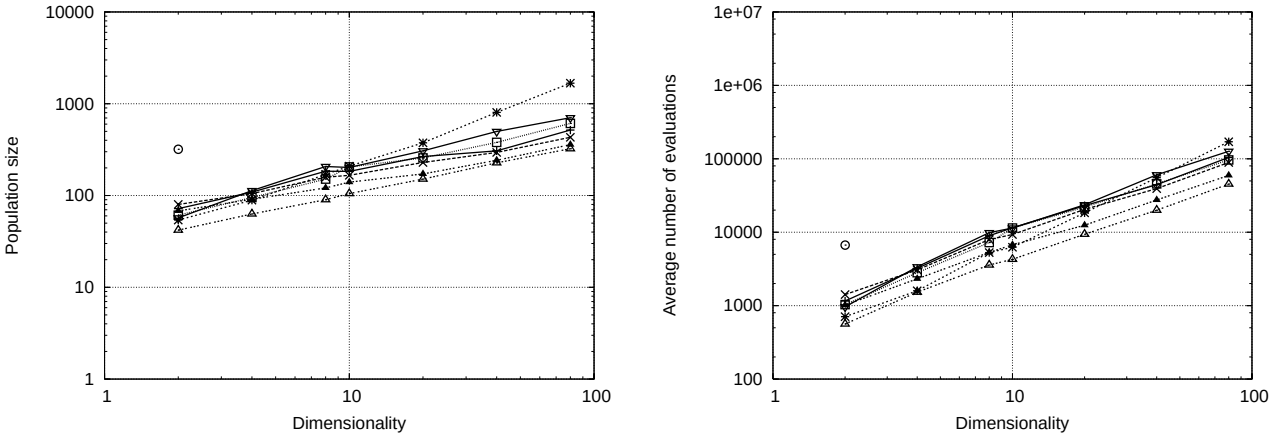
where ϵ is a standard-normally distributed error term.

Function	Algorithm	α	β
Sphere	IDEA	0.5541	1.1635
	AVS-IDEA	0.1994	1.6563
	CMA-ES	0.0000	0.9601
Ellipsoid	IDEA	0.6119	1.2171
	AVS-IDEA	0.1870	1.6870
	CMA-ES	0.0000	1.5183
Cigar	IDEA	0.5052	1.1865
	AVS-IDEA	0.2125	1.6976
	CMA-ES	0.0000	1.1093
Tablet	IDEA	0.4398	1.0860
	AVS-IDEA	0.2066	1.6397
	CMA-ES	0.0000	1.4178
Cigar Tablet	IDEA	0.4521	1.1142
	AVS-IDEA	0.1879	1.7155
	CMA-ES	0.0000	1.2431
Two Axes	IDEA	0.6603	1.2854
	AVS-IDEA	0.2177	1.6551
	CMA-ES	0.0000	1.7208
Different Powers	IDEA	0.9355	1.4983
	AVS-IDEA	0.8419	1.1692
	CMA-ES	0.0000	1.5845
Rosenbrock	IDEA	not solved	
	AVS-IDEA	0.7475	1.9154
	CMA-ES	0.6885	1.4872
Parabolic Ridge	IDEA	not solved	
	AVS-IDEA	0.1064	1.1160
	CMA-ES	0.0000	1.0853
Sharp Ridge	IDEA	not solved	
	AVS-IDEA	0.1678	0.8563
	CMA-ES	0.5228	1.4764

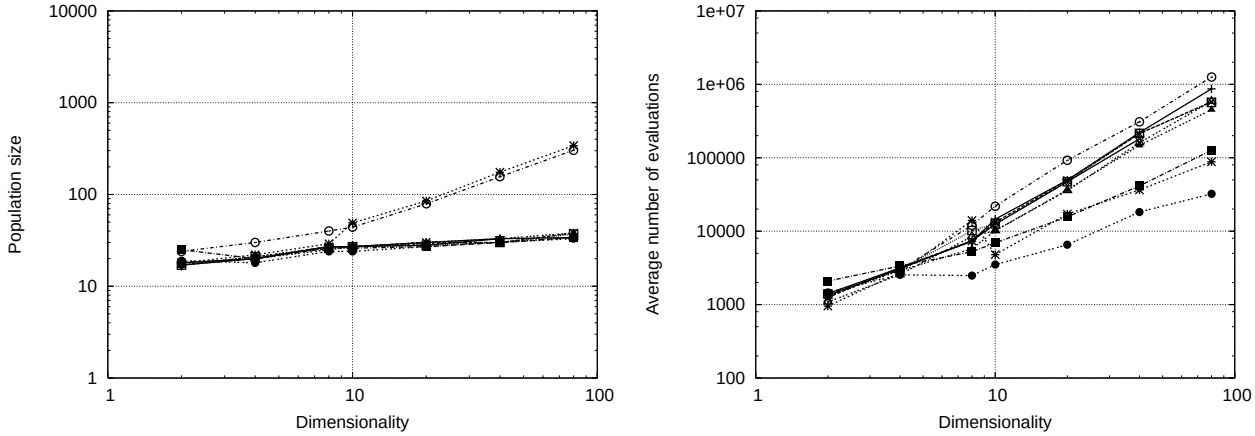
Table 2: Regression coefficients for scalability.

Results for α indicate that the population size n scales sublinearly with the problem size l for all regarded algorithms. For the IDEA without covariance adaptation, the population size n grows approximately with the square root of the dimensionality. For AVS-IDEA, the population size n grows even slower. For CMA-ES ([13]) , the population size needs not to be enlarged beyond the initial setting of $\mu = 2$ and $\lambda = 4$ for most functions, except for Rosenbrocks function and the Sharp Ridge function. The reason for this is that in the CMA-ES, the probability distribution used to guide the search is not entirely rebuilt from scratch using only the data in the current set of selected solutions. Instead, the distribution is weighted over a path of generations past and hence represents an accumulation of information.

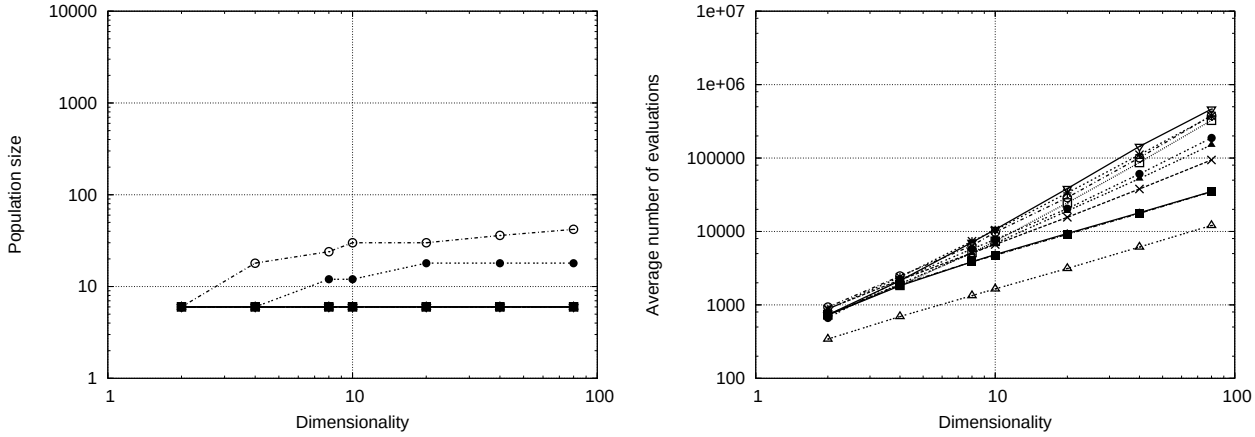
Results for β indicate that the average number of evaluations e for success grows subquadratically with l for all regarded algorithms. The average number of evaluations grows faster for the AVS-IDEA than for the IDEA without variance adaptation. However, AVS-IDEA is capable of solving *all* problems in high dimensionality which the IDEA without variance adaptation can not. The IDEA without variance adaptation is incapable of solving Rosenbrocks function, the Parabolic Ridge function and the Sharp Ridge function in higher dimensions. The reason for this is that to find the optimum for the latter two functions, the value for



(a) Normal IDEA without variance adaptation



(b) AVS-IDEA



(c) CMA-ES

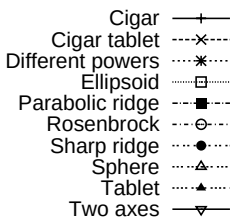


Figure 3: Scalability results for Normal IDEA, AVS IDEA and CMA-ES.

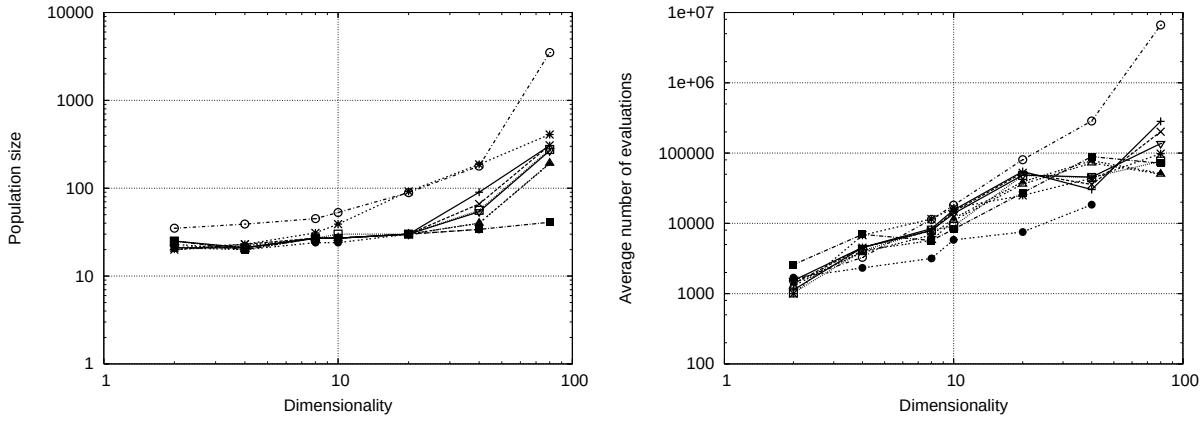


Figure 4: Scalability results for CT-AVS-IDEA (for legend, see Figure 3).

the first variable needs to be moved extremely far outside its initial range. Although the gradient along that direction is straightforward, i.e. it is a simple linear slope, the variance in the IDEA without variance adaptation shrinks too fast and the slope cannot be traveled. In Rosenbrocks function, again the variance shrinks too fast. Even though the optimum lies inside the initial range, the valley in which the optimum is contained is so narrow that the distribution quickly converges to a part inside the valley that is far from the optimum. The bottom of the valley, a curved slope, needs to be traveled to find the optimum. This slope cannot be traveled by the IDEA without variance adaptation.

Although the CMA-ES has a marginally better scalability than AVS-IDEA on the first half of the benchmark problems, this is not the case for all problems. Moreover, both algorithms scale sub–quadratically in the number of required evaluations to find the optimum.

CT-AVS-IDEA

In Figure 4 scalability results are shown for the CT-AVS-IDEA. From the results it can be seen that the addition of the correlation trigger indeed reduces the search effort of the AVS-IDEA. Up to 20 dimensions, although on the one hand the population size scales similarly to the AVS-IDEA, the number of evaluations scale more like those of the normal IDEA, indicating that less evaluations are required because variance scaling is not always required and is consequently correctly detected and signaled by the correlation trigger. However, for a dimensionality of 40 and 80, the correlation trigger reduces in efficiency. On the Sharp Ridge function, the correlation trigger even fails to trigger the scaling of variances altogether. The reason for this is that the correlation trigger and the scaling of variances is done globally for *all* directions, i.e. the entire covariance matrix. For the Sharp Ridge function, all dimensions except one do not require the scaling of variances. The signal obtained in the single non–correlated dimension becomes insignificant as the dimensionality increases and hence variance scaling is no longer triggered. Without variance scaling, the normal IDEA cannot solve the problem and hence, the CT-AVS-IDEA fails. The same will happen for the Parabolic Ridge function, albeit for even higher dimensions and similarly for Rosenbrocks function. For Rosenbrocks function, the problem can still be solved for $l = 80$, albeit clearly no longer in a polynomially scaling fashion, i.e. for even larger dimensionalities the CT-AVS-IDEA will start to behave more like the normal IDEA and hence fail. A solution to this problem may be to factorize the correlation trigger and the scaling of variances. In other words, to have various different variance scaling and correlation triggering mechanisms that are specialized in different directions.

5. CONCLUSIONS AND OUTLOOK

This paper contributed to the development of efficient and reliable EDAs for the continuous domain. It briefly discussed the defects of EDAs that directly sample the maximum-likelihood normal pdf. It then proposed the correlation-triggered adaptive variance scaling IDEA that scales the covariance matrix on slope-like regions of the search space. In order to identify the structure of the currently investigated region on the fly, we proposed the use of a ranked correlation coefficient between density and fitness.

AVS-IDEA was shown to be effective on a test bed of unimodal test functions. In comparison to the IDEA without variance adaptation, it solves all functions from the test bed and requires smaller populations. The total number of fitness evaluations grows faster for AVS-IDEA than for IDEA without variance adaptation. However, for both algorithms the average overall fitness evaluations still grows subquadratically with the number of dimensions. Adding the correlation trigger is effective for smaller problems. It does not always work well if the problem dimensionality is higher than 40.

It is an important goal of GEC research to enhance EAs such that they are able to solve an increasing array of problems. In this light, we have extended the class of problems that can be solved efficiently and reliably by continuous EDAs based on the normal pdf.

The correlation trigger and adaptive variance scaling need further research to ensure that scaling of the variances is performed only in the directions in which it is necessary. We also seek to expand the applicability of CT-AVS-IDEA to multimodal problems. Further, continuous EDAs and self-adaptive evolution strategies seem to be converging to algorithms with similar properties. It will be stimulating to investigate the similarities and differences between the two.

6. REFERENCES

[1] C. W. Ahn, R. S. Ramakrishna, and D. E. Goldberg. Real-coded Bayesian optimization algorithm: Bringing the strength of BOA into the continuous world. In K. Deb et al., editors, *Proceedings of the GECCO–2004 Genetic and Evolutionary Computation Conference*, pages 840–851, Berlin, 2004. Springer–Verlag.

[2] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons Inc., New York, New York, 1958.

[3] T. Blickle and L. Thiele. A comparison of selection schemes used in evolutionary algorithms. *Evolutionary Computation*, 4(4):361–394, 1996.

[4] P. A. N. Bosman. *Design and Application of Iterated Density-Estimation Evolutionary Algorithms*. PhD thesis, University of Utrecht, Institute of Information and Computer Science, 2003.

[5] P. A. N. Bosman and J. Grahl. Matching inductive search bias and problem structure in continuous estimation-of-distribution algorithms. Technical Report 03/2005, Mannheim Business School, Dept. of Logistics, 2005.

[6] P. A. N. Bosman and D. Thierens. Expanding from discrete to continuous estimation of distribution algorithms: The IDEA. In M. Schoenauer et al., editors, *Parallel Problem Solving from Nature – PPSN VI*, pages 767–776, Berlin, 2000. Springer–Verlag.

[7] P. A. N. Bosman and D. Thierens. Advancing continuous IDEAs with mixture distributions and factorization selection metrics. In M. Pelikan and K. Sastry, editors, *Proceedings of the Optimization by Building and Using Probabilistic Models OBUPM Workshop at the Genetic and Evolutionary Computation Conference GECCO–2001*, pages 208–212, San Francisco, California, 2001. Morgan Kaufmann.

[8] P. A. N. Bosman and D. Thierens. Learning probabilistic models for enhanced evolutionary computation. In Y. Jin, editor, *Knowledge Incorporation in Evolutionary Computation*, pages 147–176. Springer–Verlag, Berlin, 2004.

[9] M. Gallagher, M. Frean, and T. Downs. Real–valued evolutionary optimization using a flexible probability density estimator. In W. Banzhaf et al., editors, *Proc. of the Genetic and Evolutionary Computation Conference GECCO–1999*, pages 840–846, San Francisco, California, 1999. Morgan Kaufmann.

[10] C. González, J. A. Lozano, and P. Larrañaga. Mathematical modelling of UMDAc algorithm with tournament selection. Behaviour on linear and quadratic functions. *International Journal of Approximate Reasoning*, 31(3):313–340, 2002.

[11] J. Grahl, S. Minner, and F. Rothlauf. Behaviour of UMDAc with truncation selection on monotonous functions. In *The 2005 IEEE Congress on Evolutionary Computation. IEEE CEC 2005*, 2005.

[12] N. Hansen, S. D. Müller, and P. Koumoutsakos. Reducing the time complexiy of the derandomized evolution strategy with covariance matrix adaption. *Evolutionary Computation*, 11(1):1–18, 2003.

[13] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[14] R. V. Hogg and A. T. Craig. *Introduction to Mathematical Statistics*. Macmillan, New York, 5th edition, 1995.

[15] S. Kern, S. Müller, N. Hansen, D. Büche, J. Ocenasek, and P. Koumoutsakos. Learning probability distributions in continuous evolutionary algorithms - a comparative review. *Natural Computing*, 3(1):77–112, 2004.

[16] P. Larrañaga, R. Etxeberria, J. A. Lozano, and J. M. Peña. Optimization in continuous domains by learning and simulation of Gaussian networks. In M. Pelikan et al., editors, *Proc. of the Optimization by Building and Using Probabilistic Models OBUPM Workshop at the Genetic and Evolutionary Computation Conference GECCO–2000*, pages 201–204, San Francisco, California, 2000. Morgan Kaufmann.

[17] P. Larrañaga and J. A. Lozano. *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation*. Kluwer Academic, London, 2001.

[18] H. Mühlenbein and T. Mahnig. FDA – a scalable evolutionary algorithm for the optimization of additively decomposed functions. *Evolutionary Computation*, 7(4):353–376, 1999.

[19] H. Mühlenbein and G. Paaß. From recombination of genes to the estimation of distributions I. Binary parameters. In *Lecture Notes in Computer Science 1411: Parallel Problem Solving from Nature - PPSN IV*, pages 178–187, 1996.

[20] J. Ocenasek, S. Kern, N. Hansen, and P. Koumoutsakos. A mixed Bayesian optimization algorithm with variance adaptation. In X. Yao et al., editors, *Parallel Problem Solving from Nature – PPSN VIII*, pages 352–361, Berlin, 2004. Springer–Verlag.

[21] M. Pelikan, D. E. Goldberg, and F. Lobo. A survey of optimization by building and using probabilistic models. *Computational Optimization and Applications*, 21(1):5–20, 2002. Also IlliGAL Report No. 99018.

[22] M. Sebag and A. Ducoulombier. Extending population–based incremental learning to continuous search spaces. In A. E. Eiben et al., editors, *Parallel Problem Solving from Nature – PPSN V*, pages 418–427, Berlin, 1998. Springer–Verlag.

[23] I. Servet, L. Trave-Massuyes, and D. Stern. Telephone network traffic overloading diagnosis and evolutionary computation technique. In J. K. Hao et al., editors, *Proceedings of Artificial Evolution '97*, pages 137–144, Berlin, 1997. Springer–Verlag.

[24] M. M. Tatsuoka. *Multivariate Analysis: Techniques for Educational and Psychological Research*. John Wiley & Sons Inc., New York, New York, 1971.

[25] B. Yuan and M. Gallagher. On the importance of diversity maintenance in estimation of distribution algorithms. In H.-G. Beyer et al., editors, *Proc. of the Genetic and Evolutionary Computation Conference GECCO-2005*, volume 1, pages 719–726, Washington DC, USA, 2005. ACM Press.