SDR: A Better Trigger for Adaptive Variance Scaling in Normal EDAs

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

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

Franz Rothlauf
Dept. of Information Systems
University of Mannheim
68163 Mannheim
Germany
rothlauf@uni-mannheim.de

ABSTRACT

Recently, advances have been made in continuous, normal–distribution–based Estimation–of–Distribution Algorithms (EDAs) by scaling the variance up from the maximum–likelihood estimate. When done properly, such scaling has been shown to prevent premature convergence on slope–like regions of the search space. In this paper we specifically focus on one way of scaling that was previously introduced as Adaptive Variance Scaling (AVS). It was found that when using AVS, the average number of fitness evaluations grows subquadratically with the dimensionality on a wide range of unimodal test–problems, competitively with the CMA-ES. Still, room for improvement exists because the variance doesn't always have to be scaled. A previously introduced trigger based on correlation that determines when to apply scaling was shown to fail on higher dimensional problems. Here we provide a new solution called the Standard–Deviation Ratio (SDR) trigger that is integrated with the Iterated Density–Estimation Evolutionary Algorithm (IDEA). Intuitively put, scaling is triggered with SDR only if improvements are found to be far away from the mean. SDR works even in high dimensions as a result of factorizing the decision rule behind the trigger according to the estimated Bayesian factorization. We evaluate SDR–AVS–IDEA on the same set of benchmark problems and compare it with AVS–IDEA and CMA–ES. We find that the addition of SDR gives AVS–IDEA an important extra edge for it to be used in future research and in applications both in single–objective optimization as well as in multi–objective and dynamic optimization. In addition, we provide practical rules of thumb for parameter settings for using SDR–AVS–IDEA that result in an asymptotic scale–up behavior that is sublinear for the population size ($\mathcal{O}(l^{0.85})$) and subquadratic ($\mathcal{O}(l^{1.85})$) for the number of evaluations.

GECCO'07, July 7–11, 2007, London, England, United Kingdom.
Copyright 2007 ACM 978-1-59593-697-4/07/0007 ...$5.00.

Categories and Subject Descriptors

G.1 [**Numerical Analysis**]: Optimization; I.2 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search

General Terms

Algorithms, Performance, Experimentation

Keywords

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

1. INTRODUCTION

Estimation–of–distribution algorithms (EDAs, [6, 15, 17, 19]) are a class of evolutionary algorithms (EAs) in which the main operator of variation is the estimation of a probability distribution from the selected solutions and the subsequent sampling from the estimated distribution. In this way, EDAs aim to induce and exploit structure from the optimization problem at hand. The probability distribution constitutes an explicit, probabilistic, search bias.

In general, for any optimization algorithm to be successful, the structure of the problem needs to match the bias of the algorithm. Recent studies have shown that the EDA approach in continuous spaces, specifically when based on the use of a maximum–likelihood normal distribution, is not always successful [4, 9, 10]. Also, it has been pointed out more clearly under which conditions an EDA *is* expected to be successful [9]. Summarizing, the probability–distribution class must be *adequate* and the estimation procedure must be *competent*. This means that the structure of the problem can be modeled by the probability distribution and the estimation procedure can do this modeling well. For the normal distribution however, this not always the case, especially if maximum–likelihood estimates are used.

As the normal distribution itself is a single peak, it can match the contour–lines of a single peak in the fitness landscape. Things are different for slope–like regions of the search space, i.e. when the optimum is outside the range of selected solutions. The true structure may then be misrepresented by a maximum–likelihood estimate because the normal kernel focuses search around its mean. Relying the search on maximum–likelihood estimates therefore potentially misleads the EDA and can cause premature convergence on slope–like regions of the search space.

Recently, a technique was introduced to remedy the problem of the prematurely vanishing variance in continuous, maximum–likelihood, EDAs, with promising results [9]. This technique, called adaptive variance scaling (AVS), is based on whether improvements were found in the previous generation. In case an improvement was found, the variance is increased beyond its maximum–likelihood estimate. The addition of variance scaling to the EDA brings about a different view on the way model–based search is performed with continuous EDAs. Originally, the covariance matrix was estimated using maximum likelihood directly from data (i.e. the selected solutions). With variance scaling, the covariance matrix is adapted according to additional sources of information. EDAs are not the only approach to model–guided search. Specifically, when regarding the use of the normal distribution, there are clear similarities with evolution strategies (ES) [2], or more recently, the CMA–ES [12, 13]. But also approaches like particle–swarm optimization (PSO) share a similar notion of maintaining, updating and adapting a model during search. All approaches have a different, but solid, rationale and background. As such, it is important and interesting to investigate and advance all these techniques. An advantage of the EDA approach that also holds for the AVS extension is that it is conceptually easy to understand and that its choices are well motivated and principled.

A drawback of the AVS scheme is that the variance is increased even in cases where it is not required. We will describe such cases in more detail later-on. A need for a trigger thus exists that identifies exactly when an (additional) increase in variance is called for. Such a trigger, based on correlation, was proposed with the introduction of the AVS scheme. This trigger was however found to be inadequate in higher dimensions. In this paper we will investigate this issue more deeply and propose a new trigger, called Standard–Deviation Ratio (SDR), that does not break up as the dimensionality is increased. We integrate this trigger in the iterated density–estimation evolutionary algorithm (ID$\mathbb{E}$A) framework, a framework that has previously often been used to design continuous EDAs. To validate the applicability of SDR and to gain further insight into the running–time complexity of the resulting EDA, we investigate the scale-up behavior of SDR–AVS–ID$\mathbb{E}$A. The results are compared to those of the AVS–ID$\mathbb{E}$A and the CMA–ES on a test bed of (mostly) unimodal test–problems. The experimental results indicate that for all regarded algorithms the required number of fitness evaluations that is required to reliably solve the problems grows subquadratically with respect to the dimensionality of the problems. SDR–AVS–ID$\mathbb{E}$A however has the additional benefit of further reducing the number of unnecessary variance scalings as the population size increases. The integration of the new SDR trigger with AVS thus results in a novel and competitive EDA for continuous function optimization.

The remainder of this paper is organized as follows. In Section 2 we first briefly recall the scheme of AVS. Then, in Section 3 we introduce the addition of SDR. In Section 4 we provide experimental results. We propose guidelines for the use of SDR–AVS–ID$\mathbb{E}$A by a practitioner in Section 5 and also discuss future avenues of research. Finally, we conclude this paper in Section 6 with a summary and some final remarks.

2. ADAPTIVE VARIANCE SCALING

It has been shown that an EDA that uses maximum likelihood estimates for the mean and the variance can only reach the optimum if the set of search points is already close to the optimum [8, 11]. The reason for this is that the mean of the estimated normal distribution can only move a limited distance before convergence takes place because the variance shrinks exponentially fast. This means that on slope–parts of the search space, the EDA will perform extremely poorly whereas on peak–parts the EDA will perform nicely.

To remedy the problem of the prematurely vanishing variance, the variance can be scaled. This was first noted only recently [18]. One successful scheme for doing variance scaling in an adaptive fashion (i.e. during optimization) was recently introduced under the name adaptive variance scaling (AVS) [9]. This scheme significantly improves performance in the single–objective case and allows the EDA to solve problems that it couldn't solve without scaling the variance. We now briefly summarize AVS.

The smaller the variance, the smaller the area of exploration for the EDA. The variance in the normal distribution is stored in the covariance matrix $\mathbf{\Sigma}$. A variance multiplier c^{AVS} is maintained. Upon sampling new solutions, the distribution is scaled by c^{AVS}, i.e. the covariance matrix used for sampling 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 for further improvement in the next generation. The size of c^{AVS} is then 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 variance multiplier should be decreased by a factor $\eta^{\text{DEC}} \in [0, 1]$. For symmetry, $\eta^{\text{INC}} = 1/\eta^{\text{DEC}}$.

In this paper we propose a slight deviation from the original implementation of AVS. In the original implementation, the magnitude of c^{AVS} was bounded from above by a predefined value $c^{\text{AVS–MAX}} > 1$ and from below by $c^{\text{AVS–MIN}} < 1$. The upper bound is however not needed as the variance will automatically grow into the maximum variance for which improvements can still be obtained. The lower bound was introduced to allow the variance to shrink to less than its original size. This allows the algorithm to choose a niche in the case of a multimodal landscape. As we are in this paper only interested in unimodal landscapes, we simplify the scheme and never let the variance multiplier become smaller than 1. An overview of the integration of AVS in EDAs is given in Figure 1.

3. STANDARD–DEVIATION RATIO (SDR) TRIGGER

In the AVS scheme, improved fitness values automatically increase c^{AVS}. Improved fitness values however do not always mean that the variance 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 unnecessarily. In a previous paper [9], a trigger was formulated in an attempt to separate two cases: traversing a slope, and searching around an optimum. The relationship between the normal density and the shape of the function was exploited by computing the ranked correlation between the density of the selected solutions and their fitness values. If correlation is strong, then the search is focused around the optimum and no variance scaling is required.

Because the correlation coefficient is computed for all variables jointly, this approach doesn't always work, especially in higher dimensions. Suppose that all dimensions except a few do not require the scaling of variances. The contribution from the few non–correlated dimensions to the correlation measure becomes insignificant as the dimensionality increases. As a result, variance scaling is no longer triggered. Without variance scaling however, the maximum–likelihood EDA fails in the dimensions where scaling is required and hence, optimization fails altogether.

This motivates looking at the search directions of the EDA separately. We now focus specifically on the use of Bayesian factorizations as is done in the Iterated Density–Estimation Evolutionary Algorithm (ID𝔼A) [3]. To briefly recall Bayesian factorizations, we introduce a random variable X_i for each problem variable x_i, $i \in \{0, 1, \dots, l-1\}$ where l is called the problem dimensionality. We call the vector of random variables indicated by $X_{\boldsymbol{\pi}_i}$ on which X_i is conditioned in the Bayesian factorization, the vector of *parents* of X_i. A Bayesian factorization of the joint probability distribution of all involved random variables $\boldsymbol{\mathcal{X}} = (X_0, X_1, \dots, X_{l-1})$ can now be written as follows:

$$P(\boldsymbol{\mathcal{X}}) = \prod_{i=0}^{l-1} P(X_i | X_{\boldsymbol{\pi}_i}) \tag{1}$$

A greedy learning algorithm is (typically) used to compute the Bayesian factorization. For more details, we refer the interested reader to the relevant EDA literature [15, 16, 20]. The factorization imposes dependencies between the variables that are subject to search and hence allows rotation of the multivariate normal density, resulting in search directions that differ from the axis–parallel directions. This factorization can be used to design a trigger that tests in each search dimension separately whether AVS is required.

If improvements mostly take place far away from the mean, then obviously, the mean needs to shift. As we know that mean–shift is problematic for maximum–likelihood normal EDAs, this is a situation in which AVS is called for. If however most of the improvements are obtained near the mean, then the EDA with maximum–likelihood parameters already has a good focus and no further variance enlargement is required. It is known (see, e.g. [1]) that for any value of the standard deviation σ, a fixed percentage of the density of the normal distribution is contained within $[\mu - c\sigma, \mu + c\sigma]$ where μ is the mean of the normal distribution and $c \geq 0$. Now, let $\overline{\boldsymbol{x}^{\text{IMP}}}(t)$ denote the average of all new samples drawn in generation t that were an improvement over the set of selected solutions in that same generation. We propose to use a threshold $\theta^{\text{SDR}} \in [0, \infty]$ and trigger the further enlargement of the variance multiplier in generation $t+1$ whenever $\overline{\boldsymbol{x}^{\text{IMP}}}(t)$ has a distance d to the estimated mean $\hat{\mu}(t)$ such that $d/\hat{\sigma}(t) > \theta^{\text{SDR}}$. Note that this trigger is independent of the sample range and has a fixed, predefined notion of being "close" to the mean.

Second, we note that this approach can easily be factorized according to the search distribution of the EDA by following the Bayesian factorization that was estimated from the selected solutions. To sample a new solution from the Bayesian factorized normal distribution, an ordering is constructed such that when sampling a new value for X_i from the corresponding factor $P(X_i | X_{\boldsymbol{\pi}_i})$ (a conditional normal), the parents $X_{\boldsymbol{\pi}_i}$ in that factor have already been sampled. Given the values $x_{\boldsymbol{\pi}_i}$ for these parent variables, the distribution to sample X_i from is again a normal distribution with mean $\breve{\mu}_i$ and standard deviation $\breve{\sigma}_i$ [3]:

$$\hat{P}^{\mathcal{N}}(X_i | X_{\boldsymbol{\pi}_i})(x_i, x_{\boldsymbol{\pi}_i}) = \frac{1}{(\breve{\sigma}_i \sqrt{2\pi})} e^{\frac{-(x_i - \breve{\mu}_i)^2}{2\breve{\sigma}_i^2}} \tag{2}$$

$$\text{where} \begin{cases} \breve{\sigma}_i = \frac{1}{\sqrt{\hat{\boldsymbol{W}}_{00}^{(i,\boldsymbol{\pi}_i)}}} \\ \breve{\mu}_i = \frac{\hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{W}}_{00}^{(i,\boldsymbol{\pi}_i)} - \sum_{j=0}^{|\boldsymbol{\pi}_i|-1} (x_{(\boldsymbol{\pi}_i)_j} - \hat{\boldsymbol{\mu}}_{(\boldsymbol{\pi}_i)_j}) \hat{\boldsymbol{W}}_{(j+1)0}^{(i,\boldsymbol{\pi}_i)}}{\hat{\boldsymbol{W}}_{00}^{(i,\boldsymbol{\pi}_i)}} \end{cases}$$

where $\boldsymbol{W}^{\boldsymbol{j}}$ is the inverse of the symmetric covariance matrix for variables $X_{\boldsymbol{j}}$, that is, $\boldsymbol{W}^{\boldsymbol{j}} = (c^{\text{AVS}} \boldsymbol{\Sigma}^{\boldsymbol{j}})^{-1}$

We can use this result to compute for each factor separately the standard–deviation ratio of $\overline{\boldsymbol{x}^{\text{IMP}}}(t)$:

$$\text{SDR}_i = \frac{|\overline{\boldsymbol{x}_i^{\text{IMP}}}(t) - \breve{\mu}_i(t)|}{\breve{\sigma}_i} \tag{3}$$

where $\breve{\mu}_i$ and $\breve{\sigma}_i$ are computed from $\overline{\boldsymbol{x}_i^{\text{IMP}}}(t)$.

To complete the trigger, we must make a decision based upon all SDR_i. To this end, we decide to trigger the further enlargement of the variance multiplier if the ratio in any direction is larger than the threshold. In other words, if there is any search direction that requires scaling (i.e. slope traversing), AVS is triggered. This is identical to computing a single SDR as the maximum of the SDR_i and comparing this value to θ^{SDR}:

$$\text{SDR} = \max_{i=0}^{l-1} \{\text{SDR}_i\} \tag{4}$$

An overview of the integration of SDR in the AVS–ID𝔼A is given in Figure 1.

```
$\boldsymbol{\mathcal{S}} \leftarrow$ Selection($\boldsymbol{\mathcal{P}}$)
$(\boldsymbol{\pi}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}}) \leftarrow$ EstimateDistribution($\boldsymbol{\mathcal{S}}$)
$\hat{\boldsymbol{\Sigma}} \leftarrow c^{\text{AVS}} \hat{\boldsymbol{\Sigma}}$
$\boldsymbol{\mathcal{O}} \leftarrow$ SampleNewSolutions($\boldsymbol{\pi}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}}$)
$n^{\text{IMP}} \leftarrow 0$
$\overline{\boldsymbol{x}^{\text{IMP}}} \leftarrow (0, 0, \dots, 0)$
for $i \leftarrow 0$ to $|\boldsymbol{\mathcal{O}}| - 1$ do
    if $\boldsymbol{\mathcal{O}}_i$ is an improvement then
        $n^{\text{IMP}} \leftarrow n^{\text{IMP}} + 1$
        $\overline{\boldsymbol{x}^{\text{IMP}}} \leftarrow \overline{\boldsymbol{x}^{\text{IMP}}} + \boldsymbol{\mathcal{O}}_i$
if $n^{\text{IMP}} > 0$ then
    $\overline{\boldsymbol{x}^{\text{IMP}}} \leftarrow \overline{\boldsymbol{x}^{\text{IMP}}} / n^{\text{IMP}}$
    $\text{SDR} \leftarrow \max_{j=0}^{l-1} \left\{ |\overline{\boldsymbol{x}_j^{\text{IMP}}}(t) - \breve{\mu}_j(t)| \ / \ \breve{\sigma}_j \right\}$
    if $\text{SDR} > \theta^{\text{SDR}}$ then
        $c^{\text{AVS}} \leftarrow \eta^{\text{INC}} c^{\text{AVS}}$
else
    $c^{\text{AVS}} \leftarrow \eta^{\text{DEC}} c^{\text{AVS}}$
if $c^{\text{AVS}} < 1$ then
    $c^{\text{AVS}} \leftarrow 1$
$\boldsymbol{\mathcal{P}} \leftarrow (\boldsymbol{S}, \boldsymbol{O})$
```

Figure 1: Standard–Deviation Ratio (SDR) triggering and Adaptive Variance Scaling (AVS) in the generational loop of the normal ID𝔼A. The gray lines are SDR–only.

4. EXPERIMENTS

4.1 Setup

We perform experiments on test functions listed in table 1 using AVS–IDEA, SDR–AVS–IDEA and CMA–ES. All functions are unimodal with the exception of Rosenbrock's function, which has a single suboptimum [21]. 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_0 go to ∞. The initialization range used for all functions is $[-5, 5]$.

	Name	Definition	Value to reach
1	Sphere	$\sum_{i=1}^{l} x_i^2$	10^{-10}
2	Ellipsoid	$\sum_{i=1}^{l} 10^{6\frac{i-1}{l-1}} x_i^2$	10^{-10}
3	Cigar	$x_i^2 + \sum_{i=2}^{l} 10^6 x_i^2$	10^{-10}
4	Tablet	$10^6 x_1^2 + \sum_{i=2}^{l} x_i^2$	10^{-10}
5	Cigar Tablet	$x_1^2 + \sum_{i=2}^{l-1} 10^4 x_i^2 + 10^8 x_l^2$	10^{-10}
6	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}
7	Different Powers	$\sum_{i=1}^{l} \lvert x_i \rvert^{2+10\frac{i-1}{l-i}}$	10^{-15}
8	Rosenbrock	$\sum_{i=1}^{l-1}(100 \cdot (x_i^2 - x_{i+1})^2 + (x_i - 1)^2)$	10^{-10}
9	Parabolic Ridge	$-x_1 + 100 \sum_{i=2}^{l} x_i^2$	-10^{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.

Using a scalability analysis, the running time complexity of the algorithms was experimentally approximated. To be more concrete, it was assessed how the total number of fitness evaluations e and the population size n required to reliably solve the problems to optimality grows with the size of the problem l. The problem is said to be solved reliably if at least 95 out of 100 independent runs resulted in reaching the predefined value to reach. Therefore, the dimensionality l was varied: $l \in \{2, 4, 8, 10, 20, 40, 80\}$. In a recent study, a bisection method was used to obtain the minimally required population size for which the algorithms were successful [9]. However, it is important to realize that when using adaptive variance scaling, the minimum population size for which the problem can be solved is not the one that automatically minimizes the number of function evaluations. This is in contrast with well–known GA theory based on mixing which corresponds to searching inside a covered range [14, 22]. If the population size becomes very small, increasing the variance enough can still allow the EDA to solve the problem. However, because the variance is so large, more samples may be required than if a slightly larger population size is used. This phenomenon is experimentally shown in Figure 2 for the sphere function in 80 dimensions. Indeed, for the normal EDA, if the population size becomes too small, the problem can no longer be solved and the smallest population size that is successful is the one that leads to the least number of evaluations. Adding AVS or SDR–AVS allows the population size to become smaller, but the smallest population size for which the problem can be reliably solved does not correspond to the population size that leads to the minimum number of required evaluations.

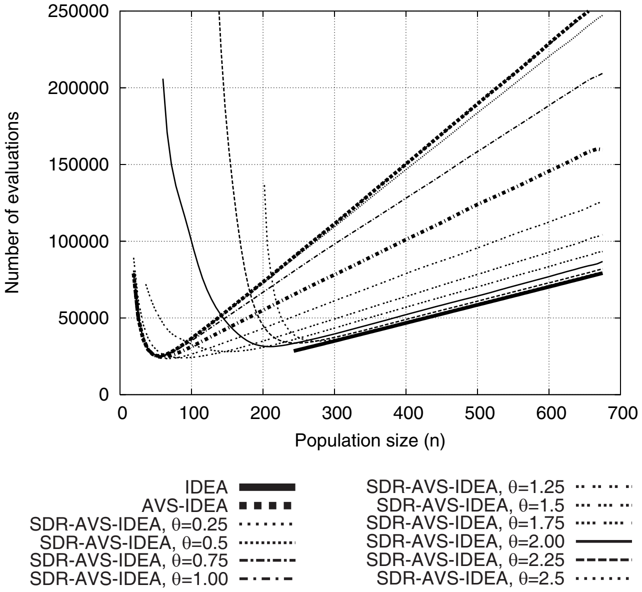


Figure 2: Average number of evaluations required to reach 10^{-10} on the sphere function in 80 dimensions in 95 out of 100 runs.

For (SDR)–AVS–IDEA we used $\eta^{\text{DEC}} = 0.9$, i.e. a small factor to allow for smooth adaptation of the variance multiplier, similar to earlier work [9]. The SDR threshold θ^{SDR} was set to $\theta^{\text{SDR}} = 1.0$ (see Section 4.2). Also following earlier work [3, 9] we set the selection percentile $\tau = 0.3$.

4.2 Setting the SDR trigger threshold

In order to obtain a reasonable value for θ^{SDR}, we ran tests on a problem that is notoriously hard for (normal) EDAs: Rosenbrock's function. Rosenbrock's function cannot be solved without variance scaling [10]. We varied θ^{SDR} and for each value of θ^{SDR}, we performed 100 independent runs of SDR–AVS–IDEA in dimensionalities $l \in \{10, 20, 40\}$. We determined the minimally required number of evaluations to solve each problem with dimensionality l reliably.

Figure 3 shows the ratio of the required number of evaluations of SDR–AVS–IDEA versus AVS–IDEA on Rosenbrock's function. As the threshold goes up, SDR–AVS-IDEA becomes less efficient on the Rosenbrock function because a larger threshold means less triggering of increasing the variance multiplier.

For the sphere function, θ^{SDR} should be set as large as possible. This can be seen in Figure 2. For population sizes larger than the population size that leads to the smallest number of evaluations, using less scaling leads to more efficient optimization. However, if the threshold becomes too large, the capacity for efficiently solving Rosenbrock's function diminishes. From the results we find that for values higher than 1.0 the results on Rosenbrock start to seriously deteriorate. We have therefore chosen to use a value of $\theta^{\text{SDR}} = 1.0$.

4.3 Results and interpretation

In Figure 4 the scalability of all tested algorithms is given on the benchmark problems. From the results we note that AVS–IDEA and SDR–AVS–IDEA have a slightly better scal-

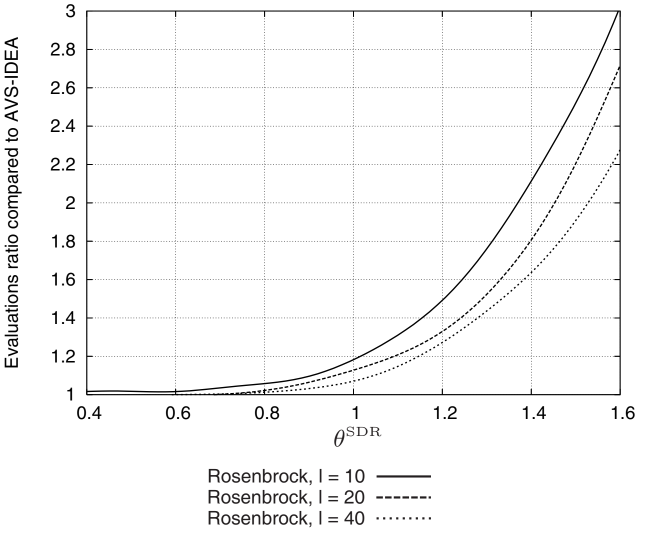


Figure 3: Minimum number of evaluations of SDR–AVS–ID$\mathbb{E}$A on Rosenbrock's function, relative to the minimum number of evaluations of AVS–ID$\mathbb{E}$A.

ability than CMA–ES in required number of evaluations (i.e. a less steep slope) on all problems except Rosenbrock's problem, for which the CMA–ES obtains better results. Hence, it can be concluded that the (SDR)–AVS–ID$\mathbb{E}$A is at least competitive with the CMA-ES for this benchmark.

There appears to be little to no difference in the results with and without the use of SDR. The reason for this is that these results reflect the choice of the population size that leads to the smallest number of evaluations. In Figure 2 it can be seen that indeed the smallest number of evaluations is virtually the same for all values of the threshold. However, if the population size becomes larger, the difference between the various values for threshold becomes clear. In practice, one typically chooses a population size and then runs the EA with it. In the next Section, we'll provide a guideline for choosing a population size. Here we first want to outline the benefits of the SDR trigger using additional results.

It is not to be expected in practice that one chooses a population size that is optimal. Instead, the selected population size is typically larger. The use of SDR triggering only gets better with larger population sizes. The most important contribution of SDR is therefore to make sure that AVS doesn't become very inefficient if suboptimal population-sizing is used. Figure 5 shows convergence plots for SDR–AVS–IDEA and AVS–ID$\mathbb{E}$A on the sphere function and Rosenbrock's function in $l = 20$ and $l = 80$ dimensions and a population size of $n = 1000$.

From the results in Figure 5 it becomes clear that SDR–AVS–ID$\mathbb{E}$A succeeds in reducing the number of evaluations and unnecessary variance scalings on the sphere function compared to AVS–ID$\mathbb{E}$A. The results in Figure 2 additionally show that the difference between AVS–ID$\mathbb{E}$A and SDR–AVS–ID$\mathbb{E}$A only become bigger if the population size is increased. The results also show that even on the Rosenbrock function, SDR–AVS–ID$\mathbb{E}$A becomes more efficient than AVS–ID$\mathbb{E}$A. Especially with a larger population size, even for this extreme function not all variance–multiplier scalings are truly required. Moreover, SDR is able to filter a good part of these unnecessary scalings out.

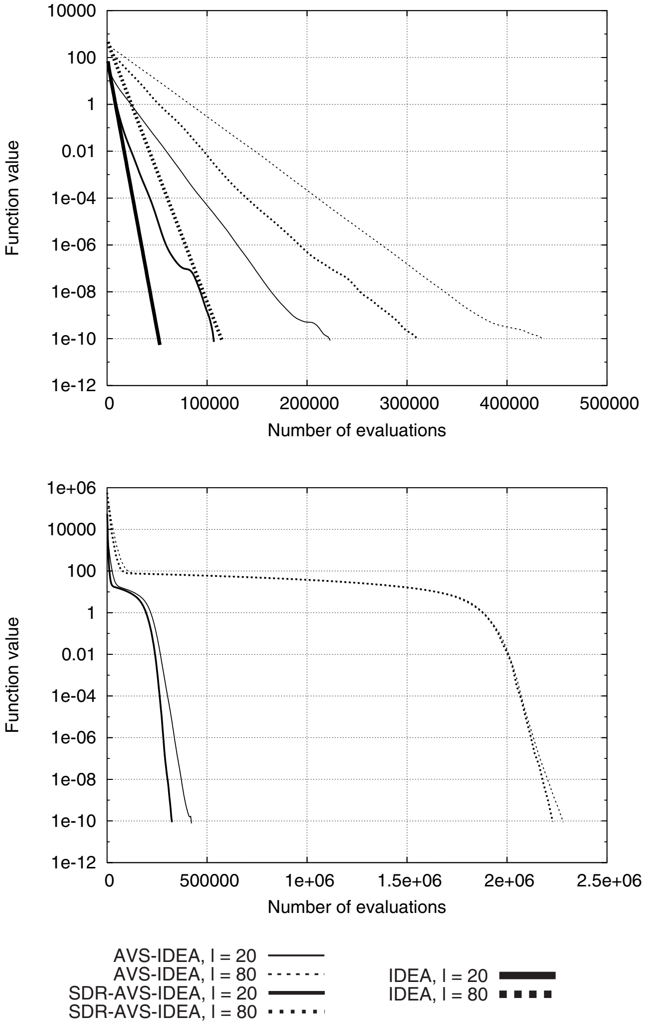
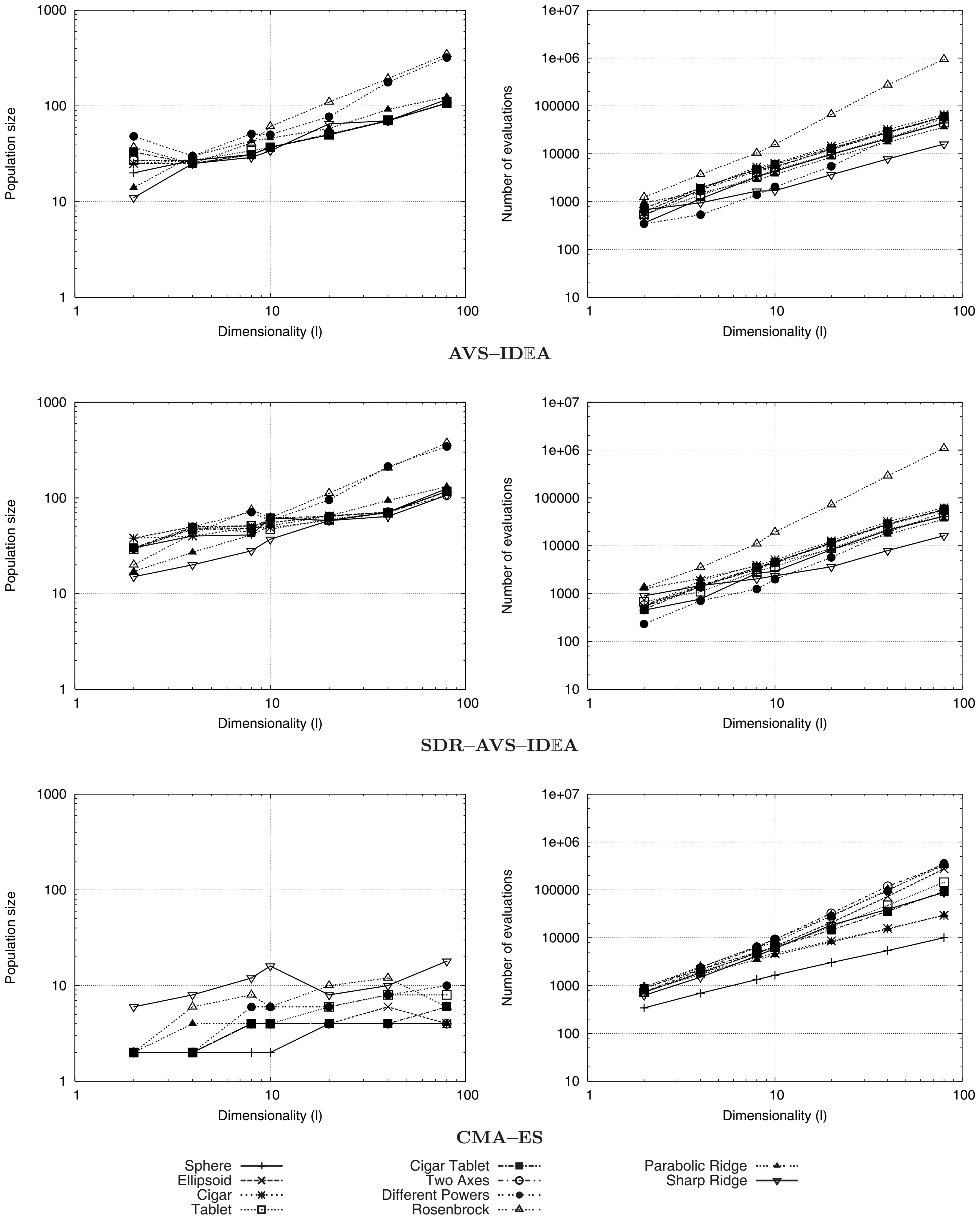


Figure 5: Convergence plots for a population size of $n = 1000$ and $l \in \{20, 80\}$ dimensions. Top: sphere, bottom: Rosenbrock.

An intuitive way to see why SDR leads to more efficient normal EDAs is the following. In any generation, given an estimation of the mean, there is an optimal covariance matrix $\mathbf{\Sigma}^*(t)$ to use for sampling. Optimality here can be taken to be the highest probability of drawing a solution that is better compared to what has been encountered so far. Obviously, if the search is on a slope and the optimum is not enclosed within the region of the currently available solutions, $\mathbf{\Sigma}^*(t)$ corresponds to a wider distribution than the estimated $\mathbf{\Sigma}(t)$ and hence, for the optimal variance multiplier we have $c^{\mathrm{AVS},*}(t) > 1$. The AVS scheme increases the variance multiplier whenever an improvement is found. At first, especially in the slope–case, this means that the actual variance multiplier becomes closer to $c^{\mathrm{AVS},*}(t)$. However, AVS will then continue to increase the variance multiplier to a value of $c^{\mathrm{AVS},+}(t)$ which corresponds to a very small probability of finding an improvement. Thus, the AVS scheme will result in varying c^{AVS} roughly in the interval of $[c^{\mathrm{AVS},*}(t), c^{\mathrm{AVS},+}(t)]$. Note that for large values of the variance multiplier, any improvements are found relatively close to the mean in terms of standard–deviation ratio. This is


AVS–IDEA

SDR–AVS–IDEA

CMA–ES

Figure 4: Scalability results for AVS–IDEA, SDR–AVS–IDEA and CMA-ES.

exactly the case when the SDR trigger indicates that no further upscaling should be applied to the variance multiplier, preventing the variance multiplier from become excessively large. Thus, the SDR–AVS scheme will result in varying c^{AVS} in an interval that is more centered around $c^{\text{AVS},*}(t)$, resulting in more efficient optimization.

5. DISCUSSION AND GUIDELINES

After analysis of the scalability data obtained from the experiments, a guideline for setting the population size can be extracted. To this end, we have estimated, using least–squares analysis, the regression line that describes the fastest growing population size. To this we added a constant to account for the problems on which the population size grows slower, but is larger for the smaller dimensionalities. The further guidelines are in line with earlier research and earlier statements in this paper:

- $n \geq 30 + 10l^{0.85}$
- $\tau = 0.3$
- $\eta^{\text{DEC}} = 0.9$
- $\eta^{\text{INC}} = 1/\eta^{\text{DEC}}$
- $\theta^{\text{SDR}} = 1.0$

In Figure 6 the resulting scale–up behavior is shown for SDR–AVS–IDEA using the above guidelines. Again, the difference between the Rosenbrock function and the other functions can be seen. However, because the scalability of the population size is mainly dictated by the scalability as found for Rosenbrock's function and is thereby higher than the minimally required population size for most other problems, the number of evaluations scales asymptotically similar for all problems. Using least–squares we find that the number of evaluations is at most (i.e. for Rosenbrock's function) $e \approx 345 * l^{1.85}$, leading to the conclusion of a scalability of $\mathcal{O}(l^{1.85})$ for the number of evaluations if the guidelines are followed. In other words, for SDR–AVS–IDEA, the population size scales sublinearly and the number of evaluations subquadratically on the mostly unimodal problems in our test set, which is an important result.

The development of variance scaling techniques and specifically of SDR–AVS–IDEA has so far been tested only on unimodal test problems and Rosenbrock's function. Although this is important and establishes the asymptotic behavior to locate an optimum with precision, other types of problems exist and are arguably dominant in practice. To deal with issues such as multimodality, a restart strategy or a parallel strategy needs to be designed. Also the integration of local (gradient–based) search techniques is most likely vital for solving real–world problems. Such enhancements can be built on top of the research and guidelines in this paper.

Although the benefits of SDR have been identified for single–objective numerical optimization in this paper, one can expect additional benefits of SDR when using SDR–AVS–IDEA for other optimization problems such as multi–objective and dynamic problems. In both problems, the use of multiple normal distributions in parallel is advantageous. For multi–objective problems, one typically wants to distribute the search along the front [5]. In dynamic optimization, one typically wants to keep track of multiple optima simultaneously [7]. To this end, the SDR–AVS technique needs to be extended to multiple normal distributions.

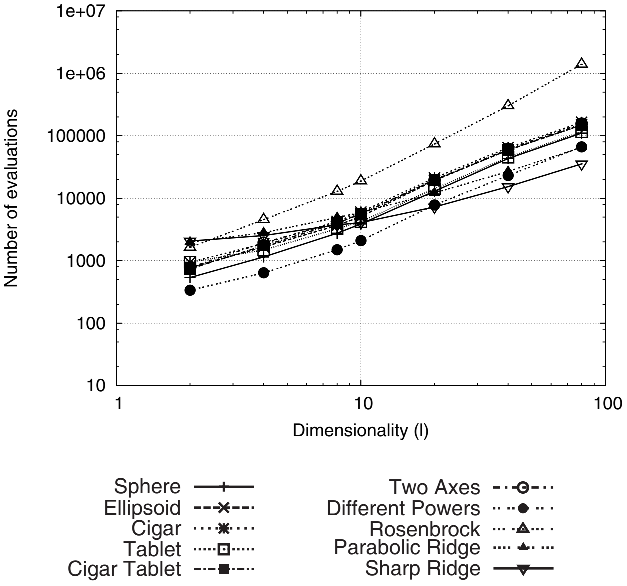


Figure 6: Scale–up of the number of required evaluations by SDR–AVS–IDEA using the guidelines.

In multi–objective problems, many improvements are made every generation as the Pareto–front is advanced. Many of these improvements are likely to be close to the means of the generating distributions along the front. Without SDR, the variance multipliers will grow very large, making further advancement much slower. In dynamic optimization the optimum keeps shifting away, making improvements more likely each generation. Depending on the speed of the movement however, the variance doesn't always need to be scaled. Without SDR, this would not be detected and further exploitation would again be slower. Summarizing, this paper lays important foundations and has given AVS–IDEA an important edge for subsequent steps to be made in continuous EDAs research as well as applications.

Finally, we note that most test functions used in this paper have either no dependencies (i.e. they are axis–parallel) or a low order of dependency between the problem variables (i.e. Rosenbrock's function). Although the SDR–AVS–IDEA approach works very well for these problems, future work will also focus on how well the approach holds up under rotations of the search space. Rotations leave the problem shape in tact, but introduce dependencies between the problem variables. Whereas the performance of the CMA–ES is known to be rotation–invariant, this is not yet known for the EDA approaches. Given the fact that the EDA approaches often do not by default use the entire covariance matrix, it is to be expected that the performance of EDA approaches is not rotation–invariant.

6. CONCLUSIONS

This paper presented ongoing research in the development of efficient and reliable EDAs for continuous single–objective optimization. Specifically it discussed an upgrade of the adaptive variance scaling (AVS) scheme that was recently proposed to improve the results of EDAs that use maximum–likelihood estimates of the normal distribution. To do so, the centroid of all improvements in a single generation is computed. This centroid is then compared to the contour lines of one standard deviation of the normal distribution. If the centroid of improvement lies outside the contour–lines, the covariance matrix is scaled further. Otherwise, the current covariance matrix covers the region where improvements can be found well enough and no further enlargement of the values in the covariance matrix is required. This enhancement of the adaptive variance scaling scheme is called standard–deviation ratio (SDR) triggering.

SDR–AVS–ID𝔼A was shown to be effective on a test bed of unimodal test functions. The population size grows sublinearly whereas the required number of evaluations grows subquadratically. Although the asymptotic behavior is similar to that of AVS–ID𝔼A, a reduction in evaluations is obtained, especially if the population size grows beyond the minimally–required size by using SDR. SDR ensures that the much required scaling of the variance by adding AVS to continuous EDAs is not exaggerated and prevents very inefficient optimization behavior if suboptimal population-sizing is used. In addition, we have argued that the addition of SDR will have additional, much required, benefits when applied to multi–objective optimization and dynamic optimization, both of which are often present in practice. Moreover, this trigger improves upon the earlier proposed trigger based upon correlation. The SDR trigger works in any dimensionality because it follows the estimated factorization whereas the previous trigger didn't work for all dimensionalities. With the introduction of SDR, we have improved the existing AVS–ID𝔼A. We believe the SDR–AVS–ID𝔼A approach to be ready for its application in practice and its transfer to other optimization problems such as multi–objective optimization and dynamic optimization. Moreover, we believe the SDR–AVS–ID𝔼A approach to be an easy–to–understand and principled approach.

Future work will extend this research to multi–objective, dynamic and multi–modal optimization problems and investigate the influence of landscape rotations. Also, we are currently establishing theoretically the efficiency of variance–scaling approaches and relating the SDR–AVS policy to this notion of efficiency.

7. REFERENCES

[1] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. Dover Publications, New York, New York, 1972.

[2] Th. Bäck and H.-P. Schwefel. An overview of evolutionary algorithms for parameter optimization. *Evolutionary Computation*, 1(1):1–23, 1993.

[3] 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.

[4] 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.

[5] P. A. N. Bosman and D. Thierens. Multi–objective optimization with diversity preserving mixture–based iterated density estimation evolutionary algorithms. *International Journal of Approximate Reasoning*, 31:259–289, 2002.

[6] 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.

[7] J. Branke, T. Kaußler, C. Schmidt, and H. Schmeck. A multi–population approach to dynamic optimization problems. In I. C. Parmee, editor, *Adaptive Computing in Design and Manufacture – ACDM 2000*, pages 299–308, Berlin, 2000. Springer Verlag.

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

[9] J. Grahl, P. A. N. Bosman, and F. Rothlauf. The correlation–triggered adaptive variance scaling idea. In M. Keijzer et al., editors, *Proceedings of the Genetic and Evolutionary Computation Conference – GECCO–2006*, pages 397–404, New York, New York, 2006. ACM Press.

[10] J. Grahl, S. Minner, and F. Rothlauf. Behaviour of UMDAc with truncation selection on monotonous functions. In *Proc. of the Congress on Evolutionary Computation – CEC–2005*, pages 2553–2559, Piscataway, New Jersey, 2005. IEEE Press.

[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] G. Harik, E. Cant-Paz, and D. E. Goldberg. The gambler's ruin problem, genetic algorithms, and the sizing of populations. In *Proceedings of the International Conference on Evolutionary Computation 1997 (ICEC '97)*, pages 7–12, Piscataway, NJ, 1997. IEEE Press.

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

[16] J. A. Lozano, P. Larrañaga, I. Inza, and E. Bengoetxea. *Towards a New Evolutionary Computation. Advances in Estimation of Distribution Algorithms.* Springer–Verlag, Berlin, 2006.

[17] H. Mühlenbein and G. Paaß. From recombination of genes to the estimation of distributions I. binary parameters. In A. E. Eiben et al., editors, *Parallel Problem Solving from Nature – PPSN V*, pages 178–187. Springer, 1996.

[18] 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.

[19] 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.

[20] M. Pelikan, K. Sastry, and E. Cantú-Paz. *Scalable Optimization via Probabilistic Modeling: From Algorithms to Applications*. Springer, Berlin, 2006.

[21] Y.-W. Shang. A note on the extended Rosenbrock function. *Evolutionary Computation*, 14(1):119–126, 2006.

[22] D. Thierens and D.E. Goldberg. Mixing in genetic algorithms. In S. Forrest, editor, *Proceedings of the fifth conference on Genetic Algorithms*, pages 38–45. Morgan Kaufmann, 1993.