Convergence Phases, Variance Trajectories, and Runtime Analysis of Continuous EDAs

Jörn Grahl
Department of Logistics
University of Mannheim
68131 Mannheim

joern.grahl@bwl.uni-mannheim.de

Peter A. N. Bosman
Center for Mathematics and Computer Science
P.O. Box 94079
1090 GB Amsterdam

Peter.Bosman@cwi.nl

Stefan Minner
Department of Logistics
University of Mannheim
68131 Mannheim

minner@bwl.uni-mannheim.de

ABSTRACT

Considering the available body of literature on continuous EDAs, one must state that many important questions are still unanswered, e.g.: How do continuous EDAs really work, and how can we increase their efficiency further? The first question must be answered on the basis of formal models, but despite some recent results, the majority of contributions to the field is experimental. The second question should be answered by exploiting the insights that have been gained from formal models. We contribute to the theoretical literature on continuous EDAs by focussing on a simple, yet important, question: How should the variances used to sample offspring from change over an EDA run? To answer this question, the convergence process is separated into three phases and it is shown that for each phase, a preferable strategy exists for setting the variances. It is highly likely that the use of variances that have been estimated with maximum likelihood is not optimal. Thus, variance modification policies are not just a nice add-on. In the light of our findings, they become an integral component of continuous EDAs, and they should consider the specific requirements of all phases of the optimization process.

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, Theory

Keywords

Evolutionary Algorithms, Estimation of Distribution Algorithms, Numerical Optimization, Predictive Models

GECCO'07, July 7–11, 2007, London, England, United Kingdom.

1. INTRODUCTION

Since the introduction of continuous EDAs for numerical optimization, this field has made significant progress, see [2] for a review of the state of the art. Current implementations solve complicated, non-linear problems reliably and efficiently. We note that over the years, the focus of published work has changed with respect to at least the following issues. 1) In first-stage EDAs the correct choice of the *structure* of probabilistic models was regarded as crucial for efficient optimization. This was largely motivated through the lessons that had been learned from the analysis of the dynamics and design of discrete EDAs. The results obtained in the continuous domain were by far not comparable to that of their discrete counterparts and these algorithms sometimes failed on problems where much simpler algorithms, even hill-climbers, did not fail. More recent results suggest that a sensible adaption of the *parameters* of the model (specifically the variances) boosts the performance of continuous EDAs, see [11], [12], and [4]. Simple adaptive variance scaling policies such as the one proposed in [4] have lead to a performance that is comparable to that of state-of-the-art Evolution Strategies. 2) While most of the initial work was experimental, researchers have started to model the dynamics of continuous EDAs, see [3], [5], and [13]. In addition to being interesting itself, formal analysis provides guidelines for design decisions.

The contributions of this paper are in line with these trends. We present a formal study of the hill-climbing behavior of continuous EDAs on the sphere function. The EDA is initialized with a mean that is far from the optimal solution and should minimize the sphere function to a pre-defined value to reach. The convex region that holds solutions that have a fitness below this value to reach is called the optimal region. The convergence process is artificially decomposed into three phases, see figure 1. (1) The mean is far from the optimal region and optimal solutions have a negligible chance of being sampled. (2) The mean is close to, but still outside, the optimal region and significant proportions of the sampled candidate solutions are in the optimal region. (3) The mean is positioned on the optimum. For phase (1) we recapitulate work that has been developed elsewhere and tie relations to the necessity of variance scaling. We show that once an EDA is in phase (2), a unique, optimal sampling variance exists, that can be obtained in a closed form. The optimal sampling variance maximizes the proportion of solutions that are optimal. For conver-

gence in phase (3) we derive a lower bound on the number of generations that is required until the optimal solution is sampled with, e.g., 99.5% chance. Such runtime analysis is available for, e.g., Evolution Strategies in [8] and [7], but is still missing for continuous EDAs.

The results are discussed with a special focus on how they influence design guidelines for continuous EDAs. It should be noted, that this article presents results that are obtained under limiting assumptions such as the use of normal distributions and quadratic functions. Relaxations of these assumptions and generalizations of results to, e.g., generally unimodal convex functions and more general classes of distributions will be rewarding. This paper discusses a basic approach that might serve as a starting point for future analysis.

2. NOTATION, ALGORITHM, AND CONVERGENCE PHASES

Numerous symbols are used throughout this paper, all of which are explained in Table 1. The subscript t denotes the state of a variable in generation t. As an example, a variance in generation t is denoted by σ_t^2.

Symbol	Description
$\chi^2_{\alpha,n}$	$(1-\alpha)$-quantile of the chi-square distribution with n degrees of freedom
$\boldsymbol{\mu}$	Mean
n	Dimensionality of the problem
ϕ	Probability density function (pdf) of the standard normal distribution
Φ	Cumulated density function (cdf) of the standard normal distribution
Φ^{-1}	Inverse of the standard normal cdf
Φ_{μ,σ^2}	Cumulated density function of the normal distribution with mean μ and variance σ^2
σ	Standard deviation
σ^2	Variance
Σ	Covariance matrix of order $(n \times n)$
t	Generation counter
τ	Percentage of selected solutions
v	A target fitness function value, the value to reach
$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$	A single solution

Table 1: List of symbols.

This paper analyzes a simple continuous EDA. The probabilistic model used is an n-dimensional normal distribution. New solutions are obtained by randomly sampling from it. In contrast to available practical implementations, the first population is not generated uniformly within a feasible region. Instead, the first population is a random sample from a n-dimensional normal distribution with an initial mean vector μ^0 and an initial covariance matrix Σ^0. All candidate solutions are evaluated with the sphere function. The sphere function assigns a single, $n-$dimensional solution $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ a fitness

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} x_i^2.$$

The best $\tau \cdot 100\%$ individuals are selected. From those, the mean vector and the covariance matrix are estimated using the known maximum likelihood estimators. The offspring replace the entire population, and the iterations starts over. Throughout the formal analysis, an infinite population size is assumed.

A value to reach is a real value v that denotes the maximal fitness that a solution may have to be considered optimal. The region that contains all optimal solutions is referred to as the optimal region.

Further, we refer to the estimated variance as the maximum-likelihood variance (ML-variance). A variance that is used to sample offspring from is referred to as a sampling variance. Note that in the above simple EDA, the ML-variance equals to the sampling variance. If the ML-variance is modified before sampling, e.g., because it is scaled, the sampling variance can be different.

We decompose the convergence process into three phases. It is assumed, that the EDA is initialized with a mean that is far from the optimal region. Phase (1) approximates the far from optimum dynamics. In this phase, the mean is far from the optimal region, and no significant portion of the sampled solutions is optimal, i.e., because the optimal region lies outside the 99.5% prediction ellipsoid of the normal distribution that is sampled from. As optimization progresses, the mean will shift towards the optimal region. In phase (2), the mean is still outside the optimal region, but a significant portion of offspring is optimal. In phase (3), the mean does not only lie inside the optimal region, but the EDA has successfully located the optimum as well. Not all sampled solutions must be optimal, because the variance might be large. Ideally, selection will reduce the variance far enough to sample solutions that are optimal with high probability.

The results for phase (1) and (2) are not exclusively valid for the sphere function only. Phase (1) approximates far from optimum search behavior in a region of the search space that has a slope-like function. Phase (2) generalizes to a situation in which solutions inside a convex region of interest are sampled with significant chance.

3. FAR FROM OPTIMUM DYNAMICS

In the following, we assume that the mean of the normal distribution is positioned far away from the optimum (phase (1)). "Far away" means that by sampling candidate solutions from the density, the chance of obtaining solutions that lie in the optimal region is virtually 0. Although the normal distribution is defined from $-\infty$ to $+\infty$ in arbitrary dimensionality, the probability of sampling solutions outside its 99.5% ellipsoid is negligible. This situation is modelled in [3] and [5]. Assume that a linear function $f(x) = x$ is maximized and x is unconstrained. Starting with an initial mean μ_0 and standard deviation σ_0, the simple EDA described in Section 2 is run under the assumption of an infinite population size. Since no optimal solution exists, no sensible search strategy will ever stop to enlarge x. [5] show that the distance that the mean can move is limited due to an exponentially fast decrease of the sampling variance. After

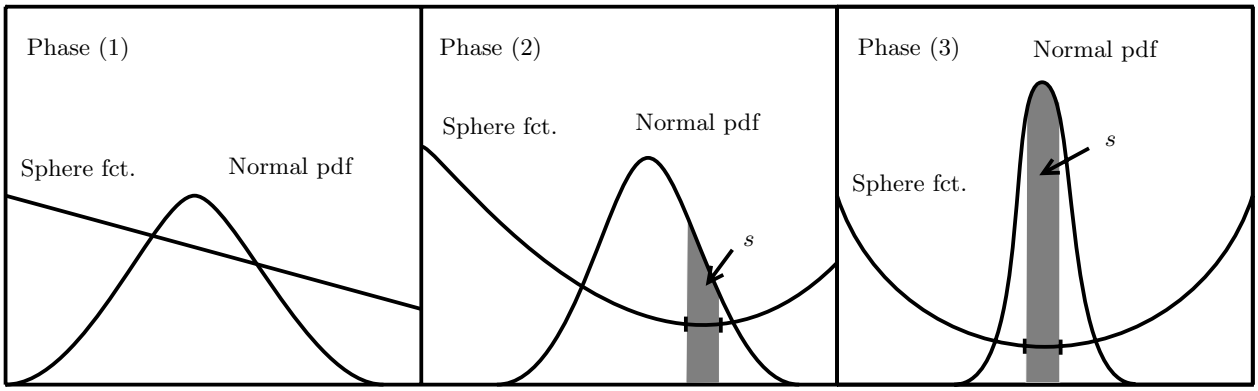

Figure 1: Decomposition of the overall process into three artificial phases. s denotes the success ratio, that is the mass of the normal pdf inside the optimal region.

an infinite number of generations, the mean μ_∞ will be

$$\mu_\infty = \mu_0 + \sigma_0 \cdot d(\tau) \cdot \frac{1}{1-\sqrt{c(\tau)}}, \text{with}$$
$$d(\tau) = \frac{\phi\left(\Phi^{-1}(\tau)\right)}{\tau}, \text{and}$$
$$c(\tau) = 1 + \frac{\Phi^{-1}(1-\tau)\phi\left(\Phi^{-1}(\tau)\right)}{\tau} - \left(\frac{\phi\left(\Phi^{-1}(\tau)\right)}{\tau}\right)^2.$$

The constants $c(\tau)$ and $d(\tau)$ can be computed numerically. This result explained premature convergence of EDA on standard test problems as observed in [1] and [9]. Consequently, variance adaption schemes were proposed in [11], [12] and [4]. All of these approaches modify the variance obtained from the known maximum likelihood estimators. The resulting sampling variance is used to generate offspring. [11] use an adaption scheme based on the $\frac{1}{5}-$th success-rule of Evolution Strategies. [13] suggest to amplify the variance by a constant factor. [4] propose a policy that bounds the variance from below and above. Variance enlargement is triggered, if the best solution found has improved. In order to suppress unnecessary scaling if the mean is located at the optimum (see phase (3)), the correlation between density and fitness of selected individuals is exploited.

An isolated analysis of phase (1) will come to the conclusion, that in order to prevent from premature convergence and to pass slope-like regions of the search space as fast as possible, the ML-variances are too small. They need to be enlarged towards a sampling-variance of maximum size.

4. OPTIMAL SAMPLING VARIANCES

According to the lessons learned from the last section, variance enlargement is crucial if a continuous EDA traverses a slope-like region of the search space. Relying on the maximum-likelihood estimators can easily lead to premature convergence. A fundamental question is, whether the sampling variance can be set arbitrarily high, or whether values exist that should preferably be chosen.

We seek to answer this question in this section. To this end, assume the simplified case that the one-dimensional sphere function $f(x) = x^2$ should be minimized to a value to reach v. All solutions x that lie inside an optimal region R have a fitness smaller than v. For the one-dimensional case, $R = [-\sqrt{v}; +\sqrt{v}]$. Consequently, we seek to find a variance that maximizes the chance to sample candidate solutions inside R. The success ratio s measures the overall probability that a solution sampled from a one-dimensional normal distribution with mean μ and variance σ^2 lies inside an optimal region $R = [a, b]$, with $\mu < a < b$, and is defined as

$$s(\mu, \sigma^2, a, b) = \Phi_{\mu,\sigma^2}(b) - \Phi_{\mu,\sigma^2}(a).$$

Without loss of generality we set $\mu = 0$, a and b are known parameters. The aim is to find a sampling variance $(\sigma^2)^\star$ that maximizes s:

$$(\sigma^2)^\star = \arg\max_{\sigma^2 \in R^+} s(0, \sigma^2, a, b)$$

The first order derivative of $s(\mu, \sigma^2, a, b)$ with respect to σ is given as

$$\frac{\mathrm{d}s}{\mathrm{d}\sigma} = -\frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{b^2}{2\sigma^2}} \cdot b + \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{a^2}{2\sigma^2}} \cdot a. \qquad (1)$$

(1) has two roots, one of which is infeasible due to negativity. The positive root

$$(\sigma^2)^\star = \left(\frac{2\ln\frac{b}{a}}{b^2 - a^2}\right)^{-1}$$

is a possible feasible maximizer for s. The second order derivative of $s(\mu, \sigma^2, a, b)$ with respect to σ is

$$\frac{\mathrm{d}^2 s(\mu, \sigma, a, b)}{\mathrm{d}\sigma^2} = \frac{e^{-\frac{b^2}{2\sigma^2}}}{\sqrt{\pi}} \cdot \left(\frac{\sqrt{2}}{\sigma^3} - \frac{b^3}{\sqrt{2}}\right) - \frac{e^{-\frac{a^2}{2\sigma^2}}}{\sqrt{\pi}} \cdot \left(\frac{\sqrt{2}}{\sigma^3} - \frac{a^3}{\sqrt{2}}\right). \qquad (2)$$

It can easily be shown, that (2) is $< 0 \ \forall \ a < b$. Thus, $(\sigma^2)^\star$ is the unique maximizer for the success probability. If a converges towards the mean $\mu = 0$, then $(\sigma^2)^\star \to 0$.

The existence of a unique maximizer for the success probability is an interesting and novel result with some important consequences for EDA design. If R is known, a sampling variance of $(\sigma^\star)^2$ maximizes the number of solutions that are sampled in R and hence maximizes convergence speed. If the used sampling variance deviates from $(\sigma^\star)^2$, less individuals will be sampled in R. The result applies only for one-dimensional search spaces, and we do not provide an extension to multi-dimensional spaces in this paper. We conjecture, that a similar maximizer exists for the multi-dimensional case as well, although it might be more difficult to obtain.

An isolated analysis of phase (2) will lead to setting the sampling variance to $(\sigma^2)^\star$.

5. RUNTIME WITH A STABLE MEAN

In this section, we analyze phase (3) of the search. The mean is positioned inside the optimal region. Not all solutions are optimal, as the success ratio depends also on the sampling variance. We derive a lower bound on the number of generations that a continuous EDA utilizing truncation selection needs in order to solve the sphere function to a given precision. Provided parameters are an initial variance σ_0^2 for each dimension of the normal distribution and a value to reach v. We consider the simpler one-dimensional case first and extend the results to n dimensions.

5.1 Runtime on x^2

The sphere function in a single dimension is $f(x) = x^2$. We consider the case that the EDA has already located the optimal solution $x^\star = 0$ and that the mean μ is $\mu = x^\star$. All following calculations assume an infinite population size. Under an infinite population size, μ will not move away from $x^\star$ in an EDA run. Truncation selection selects solutions point-symmetrically to $x^\star$. The consequence is that $\mu_t = 0 \; \forall \; t$.

5.1.1 Change from σ_t^2 to σ_{t+1}^2

Given a variance in period t denoted by σ_t^2 and the fraction of selected individuals τ, we seek to derive σ_{t+1}^2 – the variance after truncation selection. Since $\mu = x^\star = 0$, all individuals that lie inside the unique interval $[-w, w]$ satisfying

$$\int_{-w}^{w} \phi_{0,\sigma_t^2}(x)\mathrm{d}x = \tau$$

are selected. Selection equals a double truncation of the normal distribution. The variance of a doubly truncated normal distribution can be expressed in simple terms for the special case of truncation limits $A, B, A < B$ that are symmetric around the mean, cf. [6] p. 158. If $A - \mu = -(B - \mu) = -k\sigma$, then the mean of the truncated distribution is, again, μ. The variance $\sigma^{2'}$ of the truncated normal is

$$\sigma^{2'} = \sigma^2 \cdot \left(1 - \frac{2k\phi(k)}{2\Phi(k) - 1}\right). \tag{3}$$

This special case applies here. The upper bound B can be determined by

$$B = \Phi^{-1}(0.5 + 0.5\tau) = k.$$

The variance after selection can thus be written as

$$\begin{aligned}\sigma_{t+1}^2 &= \sigma_t^2 \left(1 - \frac{2\Phi^{-1}(0.5 + 0.5\tau)\phi\left(\Phi^{-1}(0.5 + 0.5\tau)\right)}{\tau}\right) \\ &= \sigma_t^2 \cdot b(\tau), \text{with} \\ b(\tau) &= \left(1 - \frac{2\Phi^{-1}(0.5 + 0.5\tau)\phi\left(\Phi^{-1}(0.5 + 0.5\tau)\right)}{\tau}\right)\end{aligned}$$

Thus, the variance is decreased by a constant factor that solely depends on the selection intensity. The term $b(\tau)$ can easily be computed numerically.

5.1.2 Variance in Generation t

It is straightforward to derive the variance in a generation t if we know an initial variance σ_0^2. As the variance is decreased by $b(\tau)$ in each generation, the variance in generation t is

$$\sigma_t^2 = \sigma_0^2 \cdot b(\tau)^t. \tag{4}$$

5.1.3 Runtime

We define the runtime as the number of generations required until the variance has decreased so far, that solutions with a fitness that is smaller than a value to reach v are sampled with a probability of at least 99.5%.

It is required that $|x| < \sqrt{v}$, for x to be optimal. The above chance constraint can be expressed as

$$P(x \in [-\sqrt{v}; \sqrt{v}]) \geq 0.995. \tag{5}$$

It is known that 99.5% of the mass of the normal distribution lies within its 3σ-quantile. Thus, we can rewrite (5) as

$$\begin{aligned}3\sigma &< \sqrt{v} \\ \Leftrightarrow \sigma^2 &< \frac{v}{9}.\end{aligned}$$

As soon as the variance has decreased to a value smaller than $\frac{v}{9}$, optimal solutions are sampled with a probability of at least 99.5%. Using (4) , this will be the case, if

$$\sigma_0^2 b(\tau)^t \leq \frac{v}{9}.$$

Solving for t yields a runtime of

$$t \geq \frac{\log \frac{v}{9\sigma_0^2}}{\log b(\tau)}. \tag{6}$$

Equation (6) gives a lower bound on the runtime on the one-dimensional sphere function that depends on a value to reach, an initial variance, and the selection intensity.

5.2 Runtime on the n-dimensional Sphere Function

The n-dimensional sphere function is point-symmetric to 0 and has a unique global optimum at $\boldsymbol{x}^\star = 0$ with $f(\boldsymbol{x}^\star) = 0$. The result from Section 5.1 is generalized to n-dimensional spaces in the following. Therefore, $\boldsymbol{\mu}_t = \boldsymbol{\mu} \; \forall \; t$, the reasons being equal to the one-dimensional case.

5.2.1 The Multivariate Normal Distribution

The n-dimensional normal distribution is given by a mean vector $\boldsymbol{\mu}$ and a positive semidefinite covariance matrix Σ of order $(n \times n)$. The eigenvectors of Σ are denoted by $\boldsymbol{e}_i$, $i = 1, 2, \ldots, n$. The eigenvalues of Σ are denoted by $\boldsymbol{\lambda}_i$, $i = 1, 2, \ldots, n$. We are especially interested in some geometric properties of the multivariate normal. Points of equal density lie on hyper-ellipsoids. The half-axes of the ellipsoids point in direction of the eigenvectors of the normal distribution. The eigenvalues relate directly to the length of the associated half-axes. The covariances induce rotation of the ellipsoids around the mean. If all covariances are 0, the axis of the ellipsoids point exactly into the directions of the main axes of the reference (coordinate) system. Prediction ellipsoids are the smallest ellipsoidal regions in n-dimensional space that are centered around the mean and contain a certain percentage of the mass of the multivariate normal. Further details can be found in [10].

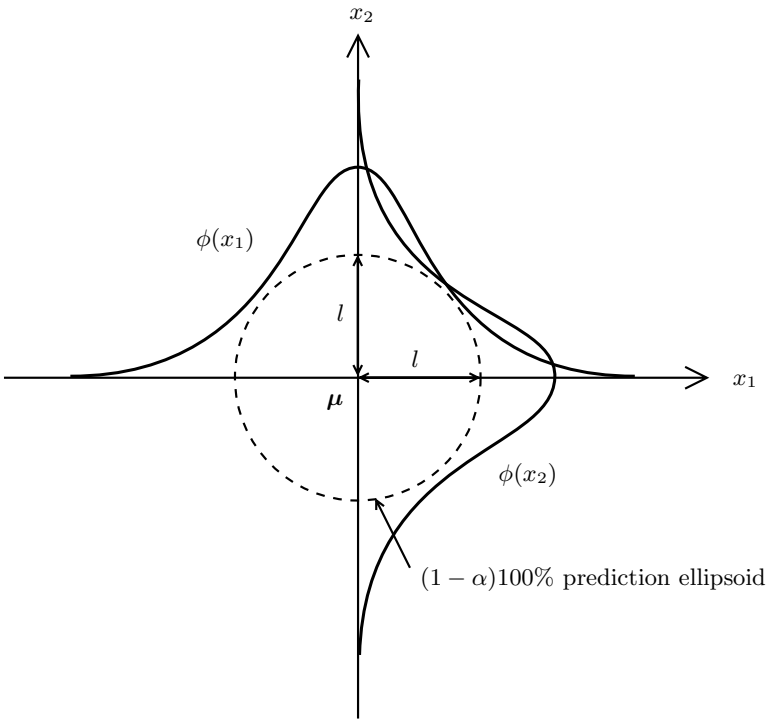

Figure 2: Univariate factorization of a two-dimensional normal pdf, prediction ellipsoid and half axes with lengths l. $\phi(x_1)$ and $\phi(x_2)$ denote the normal pdfs associated with x_1 and x_2.

5.2.2 *Change from Σ_t to Σ_{t+1}*

In order to solve the n-dimensional sphere function, the utilized EDA estimates and samples an n-dimensional normal distribution. It is assumed that in the estimation process no superfluous dependencies between the n random variables are introduced. Hence, the estimated model is a univariate factorization. This product of n univariate normals matches the separable structure of the sphere function perfectly. The general idea of our approach is that the modifications of the covariance matrix due to selection can be expressed fully in terms of modifications in the univariate normals. First it is analyzed which solutions in n-dimensional space are selected. Then, the impact of selection is modelled in one-dimensional space. This allows to reuse the basic idea from Section 5.1.

It is known that $(1-\alpha)\cdot 100\%$ of the mass of the multivariate normal lies within the so-called $(1-\alpha)$ prediction ellipsoid. This ellipsoid has half-axes with lengths of $l_i = \sqrt{\lambda_i \chi^2_{n,\alpha}}, i = 1, 2, \ldots, n$. Recall, that $\chi^2_{n,\alpha}$ denotes the $(1-\alpha)$ quantile of the chi-square distribution with n degrees of freedom. Truncation selection selects all individuals that lie in a prediction ellipsoid that covers exactly $\tau \cdot 100\%$ of the density. We are interested in the inner region of the ellipsoid, so $\alpha = 1-\tau$. Since selection is point-symmetric to the optimal solution all eigenvalues have equal values $\lambda_i = \lambda \;\forall\; i = 1, 2, \ldots, n$. The prediction ellipsoid that contains all selected individuals thus has axes with half lengths $l = \sqrt{\lambda \chi^2_{n,1-\tau}}$. If the probabilistic model consists of a factorization of univariate normals, the eigenvalues λ denote the variances into this direction. Hence the half-length l of the axes of the prediction ellipsoids covering the selected solutions is

$$l = \sigma \cdot \sqrt{\chi^2_{n,1-\tau}}. \tag{7}$$

A graphical illustration of this result is depicted in figure 2.

Knowing l we can reuse the approach from Section 5.1.1 to derive the variance in the next generation. Utilizing (3) leads to a variance after selection

$$\begin{aligned}\sigma^2_{t+1} &= \sigma^2_t \cdot \left(1 - \frac{2\chi^2_{n,1-\tau}\phi\left(\sqrt{\chi^2_{n,1-\tau}}\right)}{2\Phi\left(\sqrt{\chi^2_{n,1-\tau}}\right) - 1}\right) \\ &= \sigma^2_t \cdot \chi(n,\tau), \text{with}\end{aligned}$$

$$\chi(n,\tau) = \left(1 - \frac{2\chi^2_{n,1-\tau}\phi\left(\sqrt{\chi^2_{n,1-\tau}}\right)}{2\Phi\left(\sqrt{\chi^2_{n,1-\tau}}\right) - 1}\right).$$

Like in previous calculations of the variance after selection, the variance is reduced by a constant factor. The factor $\chi(n,\tau)$ solely depends on τ and n and can be computed numerically.

5.2.3 *Variance in Generation t*

Given an initial variance σ^2_0 in each dimension and the number of dimensions n, the variance in generation t can be computed as

$$\sigma^2_t = \sigma^2_0 \cdot \chi(n,\tau)^t.$$

5.2.4 *Runtime*

Section 5.2.2 showed that the overall modification of the covariance matrix can be expressed also through the mod-

ifications of the n variances assuming a univariate factorization and an initial isotropic normal distribution. In order to derive the runtime of a continuous EDA on the n-dimensional sphere function, we define a target standard deviation that, if used to sample from, will cause 99.5% of the solutions to have a fitness value smaller than a given value to reach v. Then we analyze how many generations selection must reduce the variance in order to reach this target value.

Optimal solutions are solutions whose fitness is smaller than v. For solutions to be optimal it is required that

$$\sum_{i=1}^{n} x_i^2 \leq v. \tag{8}$$

Under infinite population sizes the EDA behaves identically for every dimension. This has allowed the dimensionality reduction in the previous section and allows to rewrite (8) as

$$nx^2 \leq v$$
$$\Leftrightarrow x \leq \sqrt{\frac{v}{n}}.$$

In order to sample 99.5% optimal solutions, it is required that

$$3\sigma < \sqrt{\frac{v}{n}}.$$

Inserting the general variance in generation t leads to

$$3\sigma_0 \cdot \chi(n,\tau)^{\frac{t}{2}} \leq \sqrt{\frac{v}{n}},$$

which can be solved for t. The necessary number of generations t is at least

$$t > 2\frac{\log \frac{\sqrt{\frac{v}{n}}}{3\sigma_0}}{\log \chi(n,\tau)}. \tag{9}$$

The runtime depends on an initial variance, the selection intensity, the value to reach, and the number of dimensions.

It can easily be seen that - once search is in phase (3)- reducing the variance is preferable for reducing the runtime.

6. DISCUSSION OF THE RESULTS

How do continuous EDAs really work? Much of the currently available results are experimental. Although the effectiveness of current implementations is high, a thorough understanding of continuous EDAs can only be achieved on the basis of formal models. Taking together the current literature, many important questions are still unanswered. One of the most important questions is how the parameters of the probabilistic models (e.g., the covariance matrix and the mean vector) change over time due to selection. Such results are difficult to obtain if the underlying fitness landscape is complex. We have concentrated on the sphere function and have artificially decomposed the convergence process into the following three phases.

1. The mean is far from the optimum. The EDA traverses a region that has a slope-like function. The optimum is not sampled with significant probability.

2. Selection has shifted the mean towards the optimum. A significant portion of the sampled solutions lie in the optimal region, but the mean is still outside the optimal region.

3. The mean has moved onto the optimum and is relatively stable.

All three phases are characterized through ML-variance trajectories, that is a series of subsequent variances modified over generations through selection. In the first phase, variances estimated by maximum-likelihood estimators have been proven to lead to premature convergence. As a consequence, variance enlargement was introduced in some of the literature. This has lead to trajectories of sampling variances that are not equal to the estimated ones. In order to traverse slopes, a maximal increase of the ML-variance is beneficial as it increases progress.

It was an open question, whether this increase can come at a price in later phases of optimization. We have shown in Section 4 that too high a variance can indeed slow down progress if the optimal solution is coming "into sight" and is sampled with significant chance. For the one-dimensional case we have proved the existence of a sampling variance that is the unique maximizer of the success ratio (recall, that the success ratio was defined as the proportion of solutions sampled in the optimal region). This optimal sampling variance decreases with the distance between the mean and the region, and converges towards zero for the extreme case that the mean has reached the border of the region. We have not provided an extension of this result to the general multi-dimensional case, but it appears highly likely that a similar result applies, although it might be more difficult to obtain. Obviously, if the sampling variance that an EDA uses is very close or equal to the optimal sampling variance, progress of the search is maximized. If the sampling variance is too high, or too low, fewer sampled individuals are optimal.

In the third phase, selection has moved the mean onto the optimum and it is relatively stable. Until now, it was unknown how fast a continuous EDA can contract the distribution around a point of interest. We provided in Section 5.2 such a runtime result. The number of generations required until a value to reach is reliably sampled on the sphere function can be computed. By multiplying with a sensible estimate for the population size, the number of fitness evaluations can be approximated easily. A deeper analysis of these results will not only be interesting on its own, but also open the door for a principled comparison of continuous EDA with, e.g., evolution strategies.

From the available results on phase (1) dynamics, and the newly obtained results on phase (3) dynamics, we conjecture that the variance will decrease exponentially fast also if an EDA is in phase (2). This would mean that the variance goes down steadily over time throughout an EDA run. This leads to premature convergence on slope like regions of the search space. Furthermore, it is highly unlikely that such a decrease matches the optimal variances required for maximal progress in phase (2). However, it is a sensible approach in phase (3).

What can we learn from these results? Variance modification policies are not just a nice add-on. Optimal sampling variances are likely to exist, towards which modification policies should alter the estimated variances. This renders variance modification policies an integral component of continuous EDAs, at least, when maximum-likelihood normal distributions are used. These policies should not only

prevent from premature convergence in phase (1), but also use sampling variances coming close to the optimal ones in phase (2), and decrease rapidly in phase (3).

7. FUTURE WORK

The main conclusion of this work is, that variance modification policies should be integral components of continuous EDAs using maximum likelihood for estimating the model parameters. From a theoretical analysis of variance trajectories, it has become clear that there is more to a good variance scaling policy than just preventing premature convergence. It is likely that optimal sampling variance trajectories exist that maximize progress throughout the search. In order to obtain practical approximations of such policies, further analysis of phase (2) is required. The separation of the convergence process into three phases is an artificial one that facilitates the understanding. The integration of all phases into one coherent framework is important future work. Furthermore, the obtained runtime results are a starting point for systematic comparisons of the effectiveness of continuous EDAs with other optimization techniques. Generalizations of the results to other distributions and classes of fitness functions will strengthen our understanding of how we can design more efficient EDAs.

Acknowledgements

The authors thank Jan Arnold and anonymous referees for helpful advice.

8. REFERENCES

[1] P. A. N. Bosman and D. Thierens. Exploiting gradient information in continuous iterated density estimation evolutionary algorithms. In B. Kröse, M. de Rijke, G. Schreiber, and M. van Someren, editors, *Proceedings of the Thirteenth Belgium–Netherlands Artificial Intelligence Conference BNAIC–2001*, pages 69–76, 2001.

[2] P. A. N. Bosman and Dirk Thierens. Scalable optimization via probabilistic modeling: From algorithms to applications. chapter Numerical Optimization with Real-Valued Estimation of Distribution Algorithms. Springer, Berlin.

[3] 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.

[4] J. Grahl, P. A. N. Bosman, and F. Rothlauf. The correlation-triggered adaptive variance scaling IDEA (CT-AVS-IDEA). In M. Keijzer, M. Cattolico, D. Arnold, V. Babovic, C. Blum, P. Bosman, M. V. Butz, C. Coello Coello, D. Dasgupta, S. G. Ficici, J. Foster, A. Hernandez-Aguirre, G. Hornby, H. Lipson, P. McMinn, J. Moore, G. Raidl, F. Rothlauf, C. Ryan, and D. Thierens, editors, *Proceedings of the Genetic and Evolutionary Computation Conference GECCO-2006*, pages 397–404, 2006.

[5] J. Grahl, S. Minner, and F. Rothlauf. Behaviour of UMDAc with truncation selection on monotonous functions. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*, volume 2, pages 2553–2559, Scotland, 2005. IEEE Press.

[6] N. L. Johnson, S. Kotz, and N Balakrishnan, editors. *Continuous Univariate Distributions*, volume 1 of *Wiley Series in Probability and Mathematical Statistics*. Wiley-Interscience, 2nd edition, 1994.

[7] J. Jägersküpper. Rigorous runtime analysis of the (1+1) es: 1/5-rule and ellipsoidal fitness landscapes. In A. H. Wright, M. D. Vose, K. A. De Jong, and L. M. Schmitt, editors, *Foundations of Genetic Algorithms 8*, pages 260–281. Springer, Berlin, 2005.

[8] J. Jägersküpper and C. Witt. Rigorous runtime analysis of a (μ+1)es for the sphere function. In H.-G. Beyer, U. M. O'Reilly, D. V. Arnold, W. Banzhaf, C. Blum, E. W. Bonabeau, E. Cantu-Paz, D. Dasgupta, K. Deb, J. A. Foster, E. D. de Jong, H. Lipson, X. Llora, S. Mancoridis, M. Pelikan, G. R. Raidl, T. Soule, A. M. Tyrrell, J.-P. Watson, and E. Zitzler, editors, *GECCO 2005: Proceedings of the 2005 conference on Genetic and evolutionary computation*, volume 1, pages 849–856, Washington DC, USA, 2005. ACM Press.

[9] S. Kern, S. D. 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.

[10] S. Kotz, N. Balakrishnan, and N. L. Johnson. *Continuous Multivariate Distributions*, volume 2 of *Wiley Series in Probability and Mathematical Statistics*. Wiley-Interscience, 2000.

[11] J. Ocenasek, S. Kern, N. Hansen, and P. Koumoutsakos. A mixed Bayesian optimization algorithm with variance adaptation. In J. A. Lozano J. Smith J. J. Merelo Guervós J. A. Bullinaria J. Rowe P. Tino A. Kaban X. Yao, E. Burke and H. P. Schwefel, editors, *Parallel Problem Solving from Nature – PPSN VIII*, pages 352–361, Berlin, 2004. Springer.

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

[13] B. Yuan and M. Gallagher. A mathematical modelling technique for the analysis of the dynamics of a simple continuous EDA. In *Proceedings of the 2006 Congress on Evolutionary Computation*, pages 1585–1591, Vancouver, Canada, 2006. IEEE Press.